UnespDataLens-RM: Modelo de Referência para Engenharia Analítica de Dados com Governança, Qualidade, Proveniência e Reprodutibilidade

1. Título da proposta

UnespDataLens-RM: Modelo de Referência para Engenharia Analítica de Dados com Governança, Qualidade, Proveniência e Reprodutibilidade

2. Introdução

A crescente disponibilidade de dados em instituições públicas, educacionais, científicas e administrativas tem ampliado as possibilidades de análise, monitoramento, avaliação, planejamento e tomada de decisão baseada em evidências. Entretanto, a simples existência de grandes volumes de dados não garante, por si só, a produção de conhecimento confiável. Em muitos contextos institucionais, os dados encontram-se distribuídos em múltiplas fontes, produzidos por diferentes sistemas, armazenados em formatos heterogêneos, com distintos níveis de granularidade, periodicidade, qualidade, documentação, rastreabilidade e governança.

Esse cenário impõe desafios significativos à construção de ativos analíticos confiáveis. Bases de dados institucionais frequentemente apresentam inconsistências, duplicidades, ausência de metadados, mudanças de schema, diferentes regras de negócio, baixa interoperabilidade semântica, dados sensíveis sem classificação adequada, ausência de lineage, baixa reprodutibilidade e dificuldade de monitoramento ao longo do tempo. Como consequência, indicadores, dashboards, modelos preditivos, relatórios, APIs, grafos e demais produtos analíticos podem ser construídos sobre dados frágeis, pouco documentados ou de difícil auditoria.

No contexto da Ciência de Dados e da Inteligência Artificial aplicada, há uma tendência recorrente de enfatizar algoritmos, modelos e visualizações, enquanto as etapas de Engenharia de Dados que sustentam esses produtos permanecem pouco formalizadas. Entretanto, a qualidade das análises depende diretamente da qualidade dos dados, dos metadados, dos processos de transformação, da governança, do versionamento, da rastreabilidade e da reprodutibilidade dos pipelines. Sem uma estrutura técnico-metodológica que organize esse ciclo, os produtos analíticos tendem a ser pontuais, difíceis de manter, pouco reutilizáveis e frágeis diante de mudanças nas fontes ou nos requisitos institucionais.

Diante desse problema, esta proposta apresenta o **UnespDataLens-RM**, um modelo de referência para Engenharia Analítica de Dados voltado à especificação, implementação e avaliação de pipelines capazes de transformar dados heterogêneos, tabulares, temporais, espaciais, textuais, semânticos, relacionais, em grafos e multimodais em ativos analíticos confiáveis, qualificados, documentados, governados, versionados, rastreáveis e reprodutíveis. O modelo de referência é concebido como uma estrutura modular, composta por módulos técnico-operacionais e módulos transversais, permitindo organizar o ciclo

completo dos dados desde o inventário das fontes até a disponibilização e o consumo de produtos analíticos.

A presente versão amplia o escopo do modelo ao explicitar que a preparação dos dados deve considerar a natureza dos dados tratados. Diferentemente de abordagens tradicionais de ETL, frequentemente centradas em dados tabulares, o UnespDataLens-RM passa a organizar a preparação em pipelines especializados, contemplando dados estruturados, temporais, espaciais, textuais, semânticos, em grafos, multimodais e em fluxo contínuo. Essa organização reconhece que cada tipo de dado demanda operações, métricas, artefatos e cuidados próprios.
Nesse contexto, a preparação de dados textuais é incorporada como uma dimensão especializada do módulo de Transformação, Limpeza, Enriquecimento e Preparação dos Dados, e não como um módulo autônomo de análise de sentimentos. A análise de sentimentos é tratada como tarefa analítica posterior, enquanto a preparação textual envolve coleta, limpeza, normalização, anonimização, enriquecimento, representação, validação, documentação, versionamento e rastreabilidade de textos livres, comentários, fóruns, respostas abertas, mensagens, documentos e transcrições.

O UnespDataLens-RM também prevê mecanismos de apoio inteligente ao pipeline, por meio de agentes capazes de auxiliar usuários, analistas, pesquisadores, gestores e equipes técnicas na descoberta de dados, consulta a metadados, explicação de indicadores, interpretação de dashboards, identificação de dependências, análise de qualidade, apoio à documentação e manutenção dos ativos analíticos. Nesse contexto, mecanismos de Recuperação Aumentada por Geração, ou RAG, poderão ser incorporados como estratégia de apoio aos agentes inteligentes, permitindo que suas respostas sejam fundamentadas em evidências recuperadas dos artefatos produzidos pelo próprio modelo de referência, tais como catálogos, glossários, contratos de dados, registros de qualidade, lineage, versões, manifestos de execução, relatórios de monitoramento e documentação de produtos. O RAG, entretanto, não constitui o foco central da proposta, mas uma capacidade complementar associada aos agentes inteligentes do pipeline.

Assim, o UnespDataLens-RM busca contribuir para a formalização de um modelo técnico-metodológico de Engenharia Analítica de Dados que integre qualidade, governança, proveniência, rastreabilidade, reprodutibilidade, representações analíticas e apoio inteligente ao uso dos dados, fortalecendo a produção de evidências confiáveis para pesquisa, gestão e tomada de decisão.

3. Problema de pesquisa

Apesar dos avanços em Ciência de Dados, Engenharia de Dados, Inteligência Artificial, Business Intelligence, Learning Analytics, Mineração de Dados Educacionais e análise de políticas públicas, ainda há lacunas relevantes na forma como os dados são preparados, integrados, governados, documentados, versionados, monitorados e disponibilizados para uso analítico.

Em muitos projetos, os pipelines são construídos de maneira fragmentada, com foco na obtenção de resultados pontuais, mas sem uma estrutura formal que assegure qualidade, lineage, reprodutibilidade, governança e sustentabilidade. As transformações aplicadas aos dados nem sempre são registradas; os datasets nem sempre são versionados; as regras de integração podem permanecer implícitas; as variáveis derivadas podem não possuir documentação suficiente; os dashboards podem ser disponibilizados sem indicação clara de

origem, versão ou qualidade; e os modelos podem ser treinados sem rastreabilidade completa dos dados, features e rótulos utilizados.

Além disso, a crescente quantidade de fontes, atributos, produtos analíticos, versões, features, dashboards e relatórios pode gerar sobrecarga informacional, dificultando a descoberta e o uso adequado dos dados. Usuários frequentemente não sabem quais bases existem, qual versão deve ser utilizada, qual indicador é oficial, quais restrições se aplicam, qual regra de cálculo foi adotada ou quais produtos dependem de determinada fonte.

Nesse cenário, o problema central não é apenas técnico, mas também metodológico, semântico, organizacional e de governança. É necessário estruturar um modelo de referência que permita não apenas processar dados, mas qualificá-los, documentá-los, rastreá-los, governá-los, versioná-los, monitorá-los e disponibilizá-los como ativos analíticos confiáveis.

A questão central de pesquisa pode ser formulada da seguinte forma:

Como especificar, implementar e avaliar um modelo de referência para Engenharia Analítica de Dados capaz de transformar dados heterogêneos, tabulares, temporais, espaciais, textuais, semânticos, relacionais, em grafos e multimodais em ativos analíticos confiáveis, qualificados, documentados, governados, versionados, rastreáveis, equitativos e reprodutíveis, capazes de sustentar análises computacionais, modelos preditivos, indicadores, grafos, visualizações, APIs e processos de tomada de decisão baseada em evidências?

Além dessa questão geral, a proposta passa a enfrentar uma lacuna específica: como incorporar ao modelo de referência pipelines de preparação especializados por natureza do dado, especialmente para dados textuais e semânticos, sem transformar tarefas analíticas específicas, como análise de sentimentos, em módulos estruturais do modelo. Essa distinção é relevante porque a preparação textual pertence à Engenharia de Dados, enquanto a análise de sentimentos pertence à etapa de modelagem e análise.

4. Hipótese ou pressuposto da pesquisa

Parte-se do pressuposto de que a formalização de um modelo de referência modular para Engenharia Analítica de Dados, estruturado por módulos técnico-operacionais e transversais, pode aumentar a qualidade, rastreabilidade, governança, reprodutibilidade, reutilização e confiabilidade dos ativos analíticos produzidos a partir de dados heterogêneos.

Assume-se também que a integração entre metadados, catálogo, semântica, lineage, versionamento, qualidade, feature engineering, monitoramento e agentes inteligentes pode reduzir a fragmentação dos processos analíticos, melhorar a descoberta de dados, apoiar a interpretação dos produtos e fortalecer a tomada de decisão baseada em evidências.

5. Objetivo geral

O objetivo geral desta pesquisa é **propor, especificar, implementar e avaliar o UnespDataLens-RM, um modelo de referência modular para Engenharia Analítica de Dados com governança, qualidade, proveniência, linhagem, versionamento, reprodutibilidade, representações analíticas, feature engineering, monitoramento e apoio inteligente ao pipeline**, capaz de transformar dados heterogêneos em ativos

analíticos confiáveis e utilizáveis em análises computacionais, modelos preditivos, dashboards, indicadores, grafos, APIs, relatórios e processos de tomada de decisão baseada em evidências.

6. Objetivos específicos

São objetivos específicos da proposta:

1. especificar uma arquitetura modular para o UnespDataLens-RM, contemplando módulos técnico-operacionais e módulos transversais para Engenharia Analítica de Dados;
2. definir processos para inventário de fontes, extração, ingestão, armazenamento bruto, integração, harmonização, transformação, limpeza, enriquecimento, validação, armazenamento analítico e disponibilização de dados;
3. estabelecer mecanismos de avaliação da qualidade dos dados, incluindo completude, consistência, validade, unicidade, integridade, atualidade, conformidade e adequação ao uso;
4. incorporar mecanismos para identificação de desbalanceamento, vieses e riscos de equidade nos dados e nos produtos analíticos;
5. formalizar processos de metadados, catálogo, semântica, glossário, dicionário de dados e documentação dos ativos;
6. definir mecanismos de governança, segurança, privacidade, conformidade, classificação de sensibilidade e controle de acesso;
7. especificar processos de proveniência, linhagem e rastreabilidade dos dados ao longo do pipeline;
8. estruturar mecanismos de versionamento e reprodutibilidade para dados, scripts, regras, ambientes, parâmetros, features, modelos e produtos;
9. especificar representações analíticas tabulares, temporais e baseadas em grafos, permitindo diferentes formas de modelagem computacional dos dados;
10. estruturar processos de feature engineering e feature store, permitindo criação, documentação, validação, versionamento e reutilização de atributos analíticos;
11. definir mecanismos de monitoramento, drift, observabilidade e evolução dos ativos analíticos ao longo do tempo;
12. incorporar agentes inteligentes de apoio ao pipeline, capazes de auxiliar na consulta, explicação, documentação, recomendação, validação e manutenção dos ativos;
13. utilizar mecanismos de RAG como apoio aos agentes inteligentes, permitindo recuperação de evidências a partir dos artefatos internos do modelo de referência, sem deslocar o foco central da proposta;
14. avaliar o modelo de referência quanto à qualidade dos dados, rastreabilidade, reprodutibilidade, governança, utilidade, descoberta de dados, interpretabilidade, confiabilidade e adequação dos produtos analíticos gerados.

incorporar ao Módulo 4 uma dimensão de Preparação Especializada por Natureza do Dado, contemplando pipelines estruturados, temporais, espaciais, textuais e semânticos, em grafos, multimodais e em fluxo contínuo;
definir processos, artefatos e métricas específicos para Preparação de Dados Textuais e Semânticos, incluindo corpus bruto, corpus tratado, vocabulário, dicionários, regras de pré-processamento, embeddings, matrizes TF-IDF, metadados, registros de anonimização, lineage e pacotes de reprodutibilidade textual;
diferenciar conceitualmente preparação de dados textuais de tarefas analíticas posteriores, como análise de sentimentos, classificação de emoções, detecção de tópicos, sumarização, extração de entidades e detecção de toxicidade.

7. Justificativa

A justificativa da proposta está associada à necessidade de superar a visão fragmentada dos projetos de dados, nos quais cada etapa do pipeline é desenvolvida de maneira isolada, pouco documentada e dificilmente reprodutível. Em contextos institucionais, educacionais e de políticas públicas, decisões importantes podem ser apoiadas em indicadores, dashboards ou modelos produzidos a partir de dados cuja origem, qualidade, versão e transformação nem sempre são suficientemente conhecidas.

O UnespDataLens-RM justifica-se por propor uma estrutura técnico-metodológica capaz de organizar o ciclo completo dos dados, tratando-os como ativos analíticos que exigem governança, documentação, qualidade, rastreabilidade e monitoramento. A proposta não se limita à implementação de um pipeline específico, mas busca formalizar um modelo de referência reutilizável, extensível e avaliável, capaz de orientar a construção de pipelines analíticos em diferentes domínios.

Do ponto de vista científico, a proposta contribui para a área de Engenharia Analítica de Dados ao integrar dimensões frequentemente tratadas separadamente, como qualidade de dados, metadados, governança, lineage, versionamento, reprodutibilidade, feature store, drift, observabilidade, DataOps, MLOps e agentes inteligentes. Essa integração permite estruturar uma visão mais abrangente do ciclo de vida dos dados, conectando a preparação dos dados aos produtos analíticos e à tomada de decisão baseada em evidências.

Do ponto de vista tecnológico, o modelo de referência pode orientar a implementação de pipelines mais robustos, automatizados, rastreáveis e monitoráveis, reduzindo retrabalho, falhas silenciosas, duplicidade de esforços e dependência de conhecimento tácito. Do ponto de vista institucional, a proposta favorece transparência, governança, auditoria, reutilização de ativos e melhoria da confiança nos produtos analíticos.

A incorporação de agentes inteligentes, apoiados quando necessário por mecanismos de RAG, fortalece a capacidade de interação com o ecossistema de dados, permitindo que usuários consultem e compreendam ativos analíticos por meio de perguntas em linguagem natural. Essa capacidade, entretanto, é concebida como apoio à descoberta, explicação e governança, e não como substituição dos processos técnicos e metodológicos do modelo de referência.

8. Fundamentação conceitual

A proposta do UnespDataLens-RM articula conceitos de Engenharia de Dados, Ciência de Dados, Governança de Dados, Qualidade de Dados, DataOps, MLOps, Observabilidade, Proveniência, Reprodutibilidade, Feature Engineering, Representações Analíticas, Inteligência Artificial e Design Science Research.

A Engenharia de Dados fornece a base para construção de pipelines capazes de extrair, integrar, transformar, validar, armazenar e disponibilizar dados. A Ciência de Dados e a Inteligência Artificial fornecem o contexto de uso dos ativos analíticos, incluindo análise exploratória, modelagem preditiva, geração de indicadores, visualização, recomendação, classificação, explicabilidade e apoio à decisão.

A Governança de Dados estabelece papéis, responsabilidades, políticas, permissões, classificação de sensibilidade, finalidade de uso e conformidade. A Qualidade de Dados permite avaliar dimensões como completude, consistência, validade, unicidade, integridade, atualidade e adequação ao uso. A Proveniência, a Linhagem e a Rastreabilidade permitem reconstruir a origem, as transformações e os produtos derivados dos dados.

O Versionamento e a Reprodutibilidade permitem preservar versões de dados, regras, scripts, parâmetros, ambientes e resultados, garantindo que análises possam ser reexecutadas ou auditadas. A Feature Engineering e a Feature Store permitem formalizar atributos analíticos reutilizáveis, documentados, versionados e avaliados quanto à qualidade. O Monitoramento, o Drift e a Observabilidade permitem acompanhar a evolução dos dados, detectar mudanças e diagnosticar problemas.

A Design Science Research fornece a base metodológica para a construção e avaliação do UnespDataLens-RM como artefato científico. Nessa perspectiva, o modelo de referência é entendido como um artefato técnico-metodológico composto por constructos, modelos, métodos, instanciações e princípios de design. Sua validade deve ser avaliada tanto pela consistência interna de sua especificação quanto pela utilidade, aplicabilidade e capacidade de resolver problemas reais de Engenharia Analítica de Dados.

9. Estrutura geral do UnespDataLens-RM

O UnespDataLens-RM é organizado em módulos que representam o ciclo completo de Engenharia Analítica de Dados. O pipeline principal contempla os seguintes módulos:

1. **Fontes e Inventário de Dados**;
2. **Extração, Ingestão e Armazenamento Bruto**;
3. **Integração e Harmonização**;
4. **Transformação, Limpeza, Enriquecimento e Preparação dos Dados**;
5. **Qualidade e Validação de Dados**;
6. **Desbalanceamento, Vieses e Equidade**;
7. **Armazenamento Analítico**;
8. **Disponibilização e Consumo Analítico**.

A esses módulos somam-se módulos transversais:

9. **Metadados, Catálogo e Semântica dos Dados**;
10. **Governança, Segurança, Privacidade e Conformidade**;
11. **Proveniência, Linhagem e Rastreabilidade**;
12. **Versionamento e Reprodutibilidade**;
13. **Representações Analíticas: Tabular, Temporal, Espacial, Textual, Semântica e Grafos**;
14. **Feature Engineering e Feature Store**;
15. **Monitoramento, Drift e Evolução dos Ativos**;
16. **Agentes Inteligentes de Apoio ao Pipeline**.

Além desses módulos, o modelo de referência considera dimensões complementares que atravessam o ecossistema de dados, tais como Data Overload, Data Silos, Data Debt, Data Observability, Schema Drift, Data Contracts, Qualidade de Metadados, Data Discoverability, Data Minimization, Label Quality, Data Uncertainty, Interoperabilidade Semântica, DataOps, MLOps, Data Valuation e Preparação Especializada por Natureza do Dado.

10. Descrição sintética dos módulos

10.1 Módulo 1 — Fontes e Inventário de Dados

O Módulo 1 tem como finalidade identificar, mapear, caracterizar, classificar e documentar as fontes de dados relevantes para o ecossistema analítico. Inclui fontes internas, externas,

estruturadas, semiestruturadas, não estruturadas, tabulares, temporais, relacionais, documentais e provenientes de APIs, sistemas, planilhas ou repositórios públicos.

Esse módulo produz artefatos como inventário de fontes, ficha de fonte de dados, mapa de bases, dicionário preliminar, classificação de sensibilidade, registro de granularidade, matriz de relacionamento entre fontes e relatório de riscos e limitações.

10.2 Módulo 2 — Extração, Ingestão e Armazenamento Bruto

O Módulo 2 operacionaliza a entrada controlada dos dados no ambiente analítico. Envolve extração, ingestão, registro de logs, validação inicial, armazenamento bruto, preservação de snapshots, checksums, manifestos e controle de versões iniciais.

Esse módulo garante que os dados sejam preservados em seu estado original, permitindo auditoria, reprocessamento e rastreabilidade.

10.3 Módulo 3 — Integração e Harmonização

O Módulo 3 trata da integração de fontes heterogêneas, da identificação de chaves, entidades comuns, regras de correspondência, conflitos, equivalências semânticas e compatibilidade de granularidade. Seu objetivo é transformar fontes isoladas em bases integradas e coerentes.

Esse módulo é central para enfrentar silos de dados, inconsistências entre sistemas e divergências semânticas.

10.4 Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados

O Módulo 4 aplica regras de limpeza, padronização, tratamento de ausentes, remoção de duplicidades, tratamento de outliers, normalização, agregação, recodificação e criação de atributos derivados. Também registra as transformações realizadas, seus parâmetros, justificativas e impactos.

Esse módulo produz datasets tratados, enriquecidos e preparados para validação, armazenamento analítico e consumo.

10.5 Módulo 5 — Qualidade e Validação de Dados

O Módulo 5 define e executa regras de qualidade e validação, avaliando dimensões como completude, consistência, validade, unicidade, integridade, atualidade, conformidade, precisão, coerência semântica e adequação ao uso.

Esse módulo permite aprovar, rejeitar, sinalizar ou corrigir datasets antes de sua utilização analítica.

10.6 Módulo 6 — Desbalanceamento, Vieses e Equidade

O Módulo 6 identifica problemas de representatividade, desbalanceamento, vieses técnicos, vieses estatísticos, vieses de cobertura, perdas diferenciais e riscos de equidade. Avalia distribuições por grupos, métricas de fairness e limitações associadas ao uso de dados em modelos e indicadores.

Esse módulo é essencial para evitar que ativos analíticos reproduzam ou ampliem desigualdades existentes nos dados.

10.7 Módulo 7 — Armazenamento Analítico

O Módulo 7 organiza os dados tratados, integrados e validados em estruturas adequadas ao consumo analítico. Pode envolver data warehouses, data lakes, lakehouses, data marts, views, tabelas dimensionais, bancos relacionais, bancos de grafos e formatos colunares.

Seu objetivo é persistir ativos analíticos com organização, versionamento, controle de acesso, desempenho e adequação ao uso.

10.8 Módulo 8 — Disponibilização e Consumo Analítico

O Módulo 8 trata da publicação e do consumo dos ativos analíticos por usuários, sistemas e produtos. Inclui datasets curados, dashboards, APIs, relatórios, indicadores, views, bases para modelagem e produtos de dados.

Esse módulo garante que os ativos sejam disponibilizados com finalidade, versão, qualidade, restrição de acesso, público-alvo e limitações documentadas.

10.9 Módulo 9 — Metadados, Catálogo e Semântica dos Dados

O Módulo 9 organiza os metadados técnicos, semânticos, operacionais, analíticos, de qualidade, de governança e de linhagem. Inclui catálogo de dados, glossário, dicionário de dados, fichas de ativos, modelos semânticos e mapas de equivalência.

Esse módulo é fundamental para descoberta, compreensão, interoperabilidade semântica, reuso e governança dos dados.

10.10 Módulo 10 — Governança, Segurança, Privacidade e Conformidade

O Módulo 10 define políticas, papéis, responsabilidades, permissões, classificação de sensibilidade, finalidade de uso, restrições de acesso, privacidade, segurança e conformidade. Atua sobre fontes, datasets, features, produtos, dashboards, APIs e modelos.

Esse módulo garante que o uso dos dados seja controlado, responsável e compatível com requisitos institucionais, éticos e legais.

10.11 Módulo 11 — Proveniência, Linhagem e Rastreabilidade

O Módulo 11 registra a origem, trajetória, transformações, dependências e consumidores dos dados. Permite responder de onde veio determinado dado, por quais processos passou, quais regras foram aplicadas, quais versões foram usadas e quais produtos dependem dele.

Esse módulo sustenta auditoria, análise de impacto, reprodutibilidade e confiança nos ativos produzidos.

10.12 Módulo 12 — Versionamento e Reprodutibilidade

O Módulo 12 controla versões de dados, scripts, regras, parâmetros, ambientes, features, modelos, dashboards e produtos analíticos. Também define manifestos de execução e pacotes de reprodutibilidade.

Esse módulo permite reconstruir análises, comparar versões e auditar resultados.

10.13 Módulo 13 — Representações Analíticas: Tabular, Temporal, Espacial, Textual, Semântica e Grafos

O Módulo 13 especifica diferentes formas de representação dos dados, incluindo representações tabulares, temporais, longitudinais, sequenciais, relacionais, grafos estáticos, grafos heterogêneos, grafos bipartidos, grafos ponderados e grafos temporais.

Esse módulo permite adaptar os dados ao tipo de análise computacional necessária.

10.14 Módulo 14 — Feature Engineering e Feature Store

O Módulo 14 formaliza a criação, documentação, validação, versionamento, armazenamento, reutilização e monitoramento de features. Inclui atributos derivados, agregados, temporais, relacionais, baseados em grafos, estatísticos, semânticos e textuais.

Esse módulo permite transformar dados tratados em atributos analíticos reutilizáveis para modelos, indicadores, dashboards e análises.

10.15 Módulo 15 — Monitoramento, Drift e Evolução dos Ativos

O Módulo 15 acompanha a saúde, estabilidade, atualização, uso, qualidade, schema, distribuição, drift, obsolescência e evolução dos ativos analíticos. Inclui monitoramento de data drift, schema drift, feature drift, label drift, concept drift e mudanças no consumo dos produtos.

Esse módulo permite detectar degradações, falhas, mudanças e necessidades de revisão.

10.16 Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline

O Módulo 16 especifica agentes inteligentes capazes de apoiar consulta, explicação, documentação, recomendação, validação e manutenção dos ativos do UnespDataLens-RM. Esses agentes podem auxiliar usuários na descoberta de dados, interpretação de indicadores, consulta a metadados, explicação de lineage, análise de alertas e orientação sobre governança.

Nesse módulo, mecanismos de RAG podem ser incorporados como apoio à recuperação de evidências antes da geração de respostas. O RAG permite que os agentes consultem artefatos internos do modelo de referência, como catálogos, glossários, contratos, registros de qualidade, lineage, versões, manifestos e relatórios de monitoramento. Essa incorporação é complementar e não altera o foco central da proposta.

11. Dimensões complementares do modelo de referência

Além dos módulos principais, o UnespDataLens-RM considera dimensões complementares que ajudam a tratar desafios recorrentes em ecossistemas analíticos complexos.

A dimensão **Data Overload** trata da sobrecarga gerada pelo excesso de fontes, atributos, produtos, versões, dashboards, APIs, indicadores e features. Seu objetivo é apoiar curadoria, priorização, recomendação e descontinuação controlada de ativos.

A dimensão **Data Silos** trata do isolamento dos dados em sistemas, áreas, bases e repositórios distintos, buscando promover integração, interoperabilidade e compartilhamento controlado.

A dimensão **Data Debt** trata do acúmulo de passivos técnicos, metodológicos e organizacionais nos dados, metadados, pipelines e produtos analíticos.

A dimensão **Data Observability** trata da capacidade de observar, diagnosticar e explicar o comportamento dos dados, pipelines e produtos ao longo do tempo.

A dimensão **Schema Drift** trata das mudanças estruturais em atributos, tipos, domínios, chaves, categorias e granularidades.

A dimensão **Data Contracts** formaliza acordos entre produtores e consumidores de dados, definindo expectativas de schema, qualidade, periodicidade, acesso e mudança.

A dimensão **Qualidade de Metadados** avalia completude, precisão, consistência, atualidade, clareza, rastreabilidade e utilidade dos metadados.

A dimensão **Data Discoverability** trata da capacidade de localizar, compreender, avaliar e reutilizar ativos analíticos.

A dimensão **Data Minimization** orienta a coleta, armazenamento e uso apenas dos dados necessários à finalidade analítica.

A dimensão **Label Quality** trata da qualidade dos rótulos, variáveis-alvo, classes e desfechos utilizados em modelos e indicadores.

A dimensão **Data Uncertainty** trata da identificação, mensuração, documentação e comunicação de incertezas associadas aos dados, atributos, rótulos, features e produtos.

A dimensão **Interoperabilidade Semântica** trata da compatibilidade de significado entre fontes, atributos, categorias, indicadores e produtos.

A dimensão **DataOps** trata da automação, orquestração, validação, versionamento, monitoramento e melhoria contínua dos pipelines de dados.

A dimensão **MLOps** trata da gestão operacional dos modelos, incluindo registro de experimentos, versionamento, explicabilidade, fairness, monitoramento, drift e reentrenamento.

A dimensão Preparação Especializada por Natureza do Dado trata da organização de subpipelines de preparação conforme a natureza dos dados. Essa dimensão reconhece que dados estruturados, temporais, espaciais, textuais, semânticos, em grafos, multimodais e em fluxo contínuo exigem operações, métricas, artefatos, critérios de validação, registros de transformação, mecanismos de versionamento e relações de lineage próprias. No modelo, essa dimensão atua de forma articulada ao Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados, sem substituir o módulo nem criar uma etapa analítica específica. Sua contribuição é evitar que a preparação seja reduzida a procedimentos genéricos voltados apenas a dados tabulares, ampliando o alcance do

UnespDataLens-RM para cenários contemporâneos de Inteligência Artificial, PLN, análise semântica, mineração de textos, grafos, dados espaciais, dados temporais, dados multimodais e streaming.

Essas dimensões não substituem os módulos do modelo de referência, mas funcionam como capacidades complementares que ampliam sua robustez metodológica, técnica e institucional.

12. Integração de RAG como apoio aos agentes inteligentes

O UnespDataLens-RM poderá incorporar mecanismos de Recuperação Aumentada por Geração, ou RAG, como estratégia de apoio aos agentes inteligentes do pipeline. Essa incorporação não altera o foco central da proposta, que permanece voltado à Engenharia Analítica de Dados, com ênfase em qualidade, governança, proveniência, linhagem, versionamento, reprodutibilidade, representações analíticas, feature engineering e monitoramento dos ativos de dados.

No modelo de referência, o RAG será utilizado como mecanismo complementar para permitir que agentes inteligentes recuperem evidências a partir de artefatos produzidos pelo próprio UnespDataLens-RM, tais como metadados, catálogos, glossários, contratos de dados, registros de qualidade, lineage, versões, manifestos de execução, relatórios de monitoramento, feature store e documentação de produtos analíticos.

Com isso, os agentes poderão apoiar usuários na descoberta de dados, explicação de indicadores, interpretação de dashboards, identificação de dependências, consulta a regras de transformação, análise de limitações e recomendação de ações de manutenção ou governança.

As respostas geradas pelos agentes deverão ser fundamentadas nas evidências recuperadas, respeitando permissões de acesso, restrições de sensibilidade, versionamento dos artefatos e critérios de governança. Quando não houver evidência suficiente, o agente deverá explicitar a limitação, evitando respostas conclusivas. Assim, o RAG será tratado como uma capacidade de apoio à interação inteligente com o ecossistema analítico, sem constituir o núcleo da proposta nem deslocar o objetivo principal do UnespDataLens-RM.

13. Metodologia da pesquisa

A pesquisa será conduzida com base na abordagem de **Design Science Research**, adequada para pesquisas que envolvem a construção, formalização, instanciação e avaliação de artefatos técnico-metodológicos.

O UnespDataLens-RM será tratado como um artefato de pesquisa composto por constructos, modelos, métodos, instanciações e princípios de design. Os constructos correspondem aos conceitos centrais do modelo de referência, como fonte, ativo analítico, metadado, qualidade, lineage, versão, feature, contrato, drift e produto de dados. Os modelos representam a estrutura modular do modelo de referência, os fluxos de dados, as relações entre módulos, os modelos de metadados, lineage, governança e representações analíticas. Os métodos correspondem aos procedimentos de inventário, ingestão, integração, transformação, validação, versionamento, monitoramento e disponibilização. As instanciações correspondem à implementação experimental ou prototípica dos módulos em

cenários reais ou simulados. Os princípios de design correspondem às diretrizes gerais que orientam a construção de pipelines analíticos governados, rastreáveis e reprodutíveis.

A pesquisa poderá ser organizada nas seguintes etapas:

1. revisão e fundamentação teórica sobre Engenharia Analítica de Dados, DataOps, MLOps, Governança de Dados, Qualidade de Dados, Data Lineage, Reprodutibilidade, Feature Store, Observabilidade e agentes inteligentes;
2. especificação conceitual do UnespDataLens-RM, incluindo módulos, dimensões complementares, artefatos, métricas e relações entre componentes;
3. formalização dos artefatos técnico-metodológicos, contemplando finalidade, estrutura mínima, regras de construção, métricas associadas, critérios de validação e relações com os módulos;
4. implementação de uma instanciação experimental do modelo de referência, com pipelines, repositórios, catálogos, registros de qualidade, versionamento, lineage, feature store, monitoramento e produtos analíticos;
5. incorporação de agentes inteligentes de apoio ao pipeline, com possibilidade de uso de RAG para recuperação de evidências internas;
6. avaliação do modelo de referência por meio de estudos de caso, cenários experimentais, métricas de qualidade, rastreabilidade, reprodutibilidade, governança, utilidade, usabilidade e capacidade de descoberta;
7. análise dos resultados, identificação de limitações, refinamento do modelo de referência e sistematização das contribuições.

14. Artefatos esperados

A pesquisa deverá produzir um conjunto de artefatos técnico-metodológicos associados ao UnespDataLens-RM, entre eles:

- estrutura modular do modelo de referência;
- especificação dos módulos técnico-operacionais;
- especificação dos módulos transversais;
- modelo de metadados;
- catálogo de ativos;
- glossário e dicionário de dados;
- modelo de governança e controle de acesso;
- modelo de lineage e rastreabilidade;
- modelo de versionamento e reprodutibilidade;
- modelo de representações analíticas;
- especificação de feature store;
- plano de monitoramento e drift;
- modelo de agentes inteligentes;
- estratégia de uso de RAG como apoio aos agentes;
- métricas de avaliação dos módulos;
- artefatos de documentação, validação e auditoria;
- instanciação experimental do modelo de referência;
- relatórios de avaliação.

Esses artefatos serão especificados segundo a lógica da Design Science Research, com finalidade, estrutura, regras de construção, métricas, critérios de validação e contribuição.

15. Avaliação do modelo de referência

A avaliação do UnespDataLens-RM deverá considerar diferentes dimensões.

A **avaliação de qualidade dos dados** verificará completude, consistência, validade, unicidade, integridade, atualidade, conformidade e adequação ao uso.

A **avaliação de metadados e descoberta** verificará se os ativos são localizáveis, compreensíveis, documentados, classificados, reutilizáveis e semanticamente consistentes.

A **avaliação de governança** verificará se os dados possuem responsáveis, permissões, classificação de sensibilidade, finalidade de uso, restrições e conformidade.

A **avaliação de lineage e rastreabilidade** verificará se é possível reconstruir a origem, as transformações, as dependências e os consumidores dos dados.

A **avaliação de versionamento e reprodutibilidade** verificará se dados, scripts, regras, parâmetros, ambientes, features e produtos podem ser reconstruídos ou comparados entre versões.

A **avaliação de representações analíticas** verificará se os dados podem ser organizados de forma tabular, temporal ou relacional conforme as necessidades analíticas.

A **avaliação de features** verificará se os atributos derivados são documentados, versionados, reutilizáveis, rastreáveis e adequados ao uso.

A **avaliação de monitoramento e observabilidade** verificará se os ativos são acompanhados ao longo do tempo, com detecção de drift, falhas, obsolescência, alteração de schema e degradação de qualidade.

A **avaliação dos agentes inteligentes** verificará utilidade, correção, fundamentação, rastreabilidade, controle de acesso e capacidade de apoiar descoberta, explicação, documentação e manutenção dos ativos.

Quando mecanismos de RAG forem utilizados, sua avaliação deverá considerar se as respostas dos agentes são baseadas em evidências recuperadas, se indicam limitações, se respeitam permissões, se utilizam versões corretas dos artefatos e se contribuem efetivamente para o uso qualificado do ecossistema analítico.

16. Resultados esperados

Espera-se que a pesquisa resulte na especificação e validação de um modelo de referência modular para Engenharia Analítica de Dados capaz de orientar a construção de pipelines mais confiáveis, governados, rastreáveis, versionados e reprodutíveis.

Entre os principais resultados esperados, destacam-se:

- formalização do modelo de referência UnespDataLens-RM;
- definição de módulos técnico-operacionais e transversais;
- especificação de artefatos, métricas e critérios de validação;
- organização de processos de qualidade, governança, lineage, versionamento e reprodutibilidade;
- criação de mecanismos para representações tabulares, temporais, espaciais, textuais, semânticas e em grafos;
- estruturação de feature engineering e feature store;
- definição de processos de monitoramento, drift e observabilidade;

- incorporação de agentes inteligentes de apoio ao pipeline;
- uso complementar de RAG para recuperação de evidências pelos agentes;
- produção de uma instanciação experimental do modelo de referência;
- avaliação do modelo de referência em cenários de uso;
- sistematização de diretrizes para aplicação em diferentes domínios.

17. Contribuições esperadas

A proposta apresenta contribuições científicas, tecnológicas e institucionais.

Como contribuição científica, o UnespDataLens-RM propõe uma formalização integrada de Engenharia Analítica de Dados, articulando módulos, artefatos, métricas, critérios de validação e dimensões complementares. A proposta contribui para reduzir a fragmentação entre Engenharia de Dados, Ciência de Dados, Governança, DataOps, MLOps, Observabilidade e Inteligência Artificial aplicada.

Como contribuição tecnológica, o modelo de referência orienta a implementação de pipelines mais robustos, automatizados, rastreáveis, governados, versionados e monitoráveis, com potencial de aplicação em diferentes domínios.

Como contribuição institucional, a proposta favorece a produção de ativos analíticos confiáveis, reutilizáveis e adequados à tomada de decisão baseada em evidências. Também apoia a redução de silos de dados, data debt, sobrecarga informacional, baixa qualidade de metadados e dependência de conhecimento tácito.

A incorporação de agentes inteligentes, com uso complementar de RAG, contribui para tornar o ecossistema analítico mais acessível, explicável e navegável, sem deslocar o foco central da proposta.

18. Delimitação da proposta

A proposta não tem como objetivo principal desenvolver novos algoritmos de aprendizagem de máquina, novos modelos de linguagem ou uma arquitetura centrada em RAG. O foco está na especificação, implementação e avaliação de um modelo de referência para Engenharia Analítica de Dados.

Modelos preditivos, agentes inteligentes e mecanismos de RAG poderão ser utilizados como componentes de apoio ou instanciação do modelo de referência, mas não constituem o núcleo da contribuição. O núcleo da proposta está na organização técnico-metodológica dos dados, seus processos, artefatos, métricas, governança, rastreabilidade e reprodutibilidade.

19. Síntese da proposta

O **UnespDataLens-RM** é proposto como um modelo de referência modular para Engenharia Analítica de Dados voltado à transformação de dados heterogêneos em ativos analíticos confiáveis, qualificados, documentados, governados, versionados, rastreáveis e reprodutíveis. Sua estrutura contempla módulos técnico-operacionais, módulos transversais e dimensões complementares que abrangem desde o inventário de fontes até a disponibilização e o consumo de produtos analíticos.

A proposta busca fortalecer a qualidade, a governança, a proveniência, a linhagem, o versionamento, a reprodutibilidade, a interoperabilidade semântica, a observabilidade, o monitoramento e o uso responsável dos dados. Adicionalmente, incorpora agentes inteligentes de apoio ao pipeline, com possibilidade de uso de RAG como mecanismo complementar para recuperação de evidências internas do modelo de referência.

Dessa forma, o UnespDataLens-RM contribui para estruturar um ecossistema analítico capaz de sustentar análises, indicadores, modelos, dashboards, APIs, grafos, relatórios e decisões baseadas em evidências, preservando rigor técnico, transparência, rastreabilidade, governança e confiabilidade.

Módulo 1 — Fontes e Inventário de Dados

1. Finalidade do módulo

O **Módulo 1 — Fontes e Inventário de Dados** tem como finalidade identificar, mapear, caracterizar, classificar e documentar as fontes de dados que poderão compor o ecossistema analítico do **UnespDataLens-RM**. Esse módulo constitui a etapa inicial do modelo de referência, pois estabelece a base de conhecimento sobre quais dados existem, onde estão localizados, quem os produz, quem os mantém, qual sua finalidade original, quais restrições possuem, qual sua estrutura, qual sua periodicidade, qual sua cobertura temporal, qual sua granularidade, quais entidades representam e qual seu potencial de uso analítico.

No contexto do UnespDataLens-RM, uma fonte de dados não deve ser compreendida apenas como uma tabela, arquivo ou sistema isolado. Uma fonte pode ser um banco de dados institucional, uma API, uma planilha, um arquivo CSV, JSON ou Parquet, um sistema acadêmico, um ambiente virtual de aprendizagem, um sistema administrativo, uma base pública, um log de eventos, um fluxo de dados, um repositório documental, um painel já existente, um serviço web, um sensor, uma base governamental ou qualquer outro recurso capaz de fornecer dados para análise.

O Módulo 1 busca transformar fontes dispersas, heterogêneas, pouco documentadas e muitas vezes conhecidas apenas informalmente em um **inventário estruturado, rastreável, governável e analiticamente útil**. Esse inventário permite que os módulos posteriores atuem sobre dados previamente conhecidos, evitando extrações improvisadas, integrações frágeis, uso de fontes sem responsável, ausência de documentação, duplicidade de bases, desconhecimento de restrições de acesso e perda de rastreabilidade.

A importância deste módulo está no fato de que a qualidade de todo o pipeline analítico depende da compreensão inicial das fontes. Não é possível integrar, transformar, validar, governar, versionar, monitorar ou disponibilizar adequadamente dados cuja origem, estrutura, periodicidade, sensibilidade, granularidade, limitações e responsáveis não foram previamente identificados.

Assim, o Módulo 1 funciona como a porta de entrada metodológica do UnespDataLens-RM. Ele não executa ainda a ingestão dos dados, mas prepara as condições técnicas, semânticas, operacionais e de governança para que a extração e a ingestão ocorram de forma controlada no Módulo 2.

2. Objetivo geral

O objetivo geral do Módulo 1 é **construir um inventário estruturado das fontes de dados relevantes ao ecossistema UnespDataLens-RM**, documentando suas características técnicas, semânticas, temporais, operacionais, institucionais, analíticas, legais e de governança, de modo a orientar as etapas posteriores de extração, ingestão, integração, transformação, qualidade, armazenamento, disponibilização, lineage, versionamento, monitoramento e uso analítico.

3. Objetivos específicos

São objetivos específicos do Módulo 1:

1. identificar as fontes de dados disponíveis ou potencialmente relevantes para o projeto analítico;
2. classificar as fontes conforme domínio, tipo, formato, origem, responsável, periodicidade, granularidade, cobertura temporal e restrição de acesso;
3. registrar os responsáveis técnicos, institucionais ou operacionais por cada fonte;
4. documentar finalidade original, finalidade analítica potencial e limitações conhecidas de cada fonte;
5. mapear entidades, atributos, chaves, categorias, códigos e conceitos presentes nas fontes;
6. identificar a existência de dados históricos, dados temporais, logs, eventos, registros longitudinais, séries temporais ou fluxos de dados;
7. registrar periodicidade de atualização, latência esperada e possibilidade de ingestão batch, incremental, near real-time ou streaming;
8. avaliar sensibilidade, restrições de acesso, privacidade, finalidade de uso e riscos associados às fontes;
9. identificar fontes redundantes, complementares, isoladas ou pertencentes a silos de dados;
10. mapear relações potenciais entre fontes, entidades e domínios;
11. registrar documentação existente, dicionários prévios, contratos de dados, APIs, schemas ou glossários associados;
12. avaliar preliminarmente a qualidade, completude, atualidade e confiabilidade das fontes;
13. indicar o potencial de cada fonte para construção de indicadores, dashboards, modelos, features, grafos, grafos temporais ou agentes inteligentes;
14. produzir artefatos formais que orientem os módulos posteriores do modelo de referência;
15. reduzir dependência de conhecimento tácito sobre os dados institucionais.

4. Fundamentação conceitual do módulo

A etapa de identificação e inventário de fontes é uma prática essencial em Engenharia Analítica de Dados, pois permite compreender o ecossistema informacional antes de qualquer processo de extração, integração ou modelagem. Em ambientes institucionais complexos, os dados normalmente estão distribuídos em múltiplos sistemas, planilhas, arquivos, plataformas, bases históricas, registros operacionais e produtos já existentes.

Essas fontes podem apresentar diferenças significativas de formato, periodicidade, nomenclatura, granularidade, qualidade, cobertura temporal, semântica e restrição de uso. Uma base pode registrar estudantes, outra matrículas, outra interações, outra avaliações, outra indicadores agregados e outra eventos em tempo real. Em políticas públicas, uma fonte pode registrar atendimentos, outra territórios, outra famílias, outra equipamentos, outra serviços e outra indicadores socioeconômicos.

Sem inventário, essas fontes permanecem invisíveis ou parcialmente conhecidas. Isso gera problemas como silos de dados, duplicidade de esforços, uso de dados desatualizados, dificuldade de integração, ausência de governança, baixa rastreabilidade, incerteza sobre a origem dos dados e aumento da dívida de dados.

No UnespDataLens-RM, o inventário de fontes é tratado como um artefato técnico-metodológico. Ele não é apenas uma lista de bases, mas um modelo estruturado de conhecimento sobre o ecossistema de dados. Cada fonte deve ser descrita por atributos técnicos, temporais, semânticos, operacionais, institucionais e analíticos.

Esse módulo também deve contemplar explicitamente a dimensão temporal das fontes. Uma fonte pode conter dados históricos, dados atualizados periodicamente, eventos com timestamp, logs de interação, séries temporais, registros longitudinais ou fluxos dinâmicos. Essa caracterização é fundamental para orientar o Módulo 2, que definirá o tipo de ingestão mais adequado, e o Módulo 13, que definirá representações tabulares, temporais, de eventos, grafos ou grafos temporais.

À luz da **Design Science Research**, o Módulo 1 produz artefatos que representam constructos, modelos, métodos e instanciações. O inventário de fontes funciona como um modelo estruturado do ecossistema de dados; o mapa de entidades funciona como um modelo conceitual preliminar; a matriz de relacionamento entre fontes funciona como um modelo de dependências potenciais; e os registros de sensibilidade, granularidade e limitações funcionam como artefatos de governança e avaliação.

5. Tipos de fontes consideradas

O UnespDataLens-RM deve contemplar diferentes tipos de fontes de dados, considerando que ecossistemas analíticos reais são heterogêneos e dinâmicos.

Tipo de fonte	Descrição	Exemplo
Banco de dados relacional	Fonte estruturada baseada em tabelas, chaves e relações.	Sistema acadêmico, sistema administrativo, PostgreSQL, SQL Server.
Arquivo estruturado	Dados organizados em formato tabular ou hierárquico.	CSV, XLSX, JSON, XML, Parquet.
API ou serviço web	Dados acessados por endpoints ou serviços externos.	API governamental, API institucional, web service.
Planilha institucional	Arquivo mantido manualmente por equipes ou setores.	Planilha de acompanhamento de estudantes.
Sistema transacional	Sistema que registra operações ou eventos administrativos.	Matrícula, atendimento, protocolo, financeiro.

Ambiente virtual ou plataforma digital	Sistema com interações, logs e atividades de usuários.	AVA, LMS, Moodle, Google Classroom.
Log de eventos	Registro de ações com timestamp e contexto.	Acesso, clique, envio de atividade, consulta.
Fluxo contínuo de dados	Dados gerados continuamente ou com baixa latência.	Sensores, filas, tópicos, streaming.
Base pública	Dados disponibilizados por órgãos públicos ou instituições externas.	IBGE, INEP, DATASUS, bases municipais.
Dashboard ou relatório existente	Produto analítico que pode indicar fontes ou regras já utilizadas.	Painel institucional de evasão.
Repositório documental	Documentos que contêm dados ou informações semiestruturadas.	Relatórios, PDFs, formulários, atas.
Banco de grafos ou rede	Fonte estruturada em nós, arestas e propriedades.	Neo4j, ArangoDB, rede de relações.
Base textual	Registros textuais que podem ser analisados semanticamente.	Fóruns, respostas abertas, mensagens, chamados.
Fonte geoespacial	Dados associados a localização, território ou coordenadas.	Shapefiles, GeoJSON, bases territoriais.

Essa tipologia permite que o modelo de referência não fique restrito a dados tabulares tradicionais, ampliando sua capacidade para dados temporais, relacionais, textuais, geoespaciais, multimodais e dinâmicos.

6. Formalização sintética do módulo

Considerando o conjunto de fontes identificadas:

$F = \{f_1, f_2, ..., f_n\}$

cada fonte f_i pode ser descrita por um conjunto de atributos técnicos, semânticos, operacionais, temporais e de governança:

f_i = {id, nome, descrição, origem, responsável, domínio, tipo, formato, localização, periodicidade, cobertura_temporal, granularidade, volume, schema, entidades, atributos, chaves, sensibilidade, restrição_acesso, finalidade_original, finalidade_analítica, documentação_existente, limitações, qualidade_preliminar, potencial_analítico}

O conjunto de dimensões de caracterização das fontes pode ser representado por:

$D = \{d_1, d_2, ..., d_k\}$

em que cada dimensão corresponde a um aspecto de avaliação, como:

- identificação;
- origem;
- domínio;
- estrutura;
- formato;
- temporalidade;
- granularidade;
- volume;
- qualidade;
- sensibilidade;
- acesso;
- documentação;
- semântica;
- interoperabilidade;
- potencial analítico;
- risco;
- governança.

Também é possível representar relações potenciais entre fontes por uma matriz:

$R(F) = [r_{ij}]$

em que r_{ij} indica a possibilidade de relacionamento entre as fontes f_i e f_j, considerando entidades comuns, chaves, atributos equivalentes, granularidade, domínio, período ou finalidade analítica.

Essa formalização permite que o Módulo 1 produza um mapa estruturado do ecossistema de dados antes da ingestão e integração.

7. Entradas do módulo

As entradas do Módulo 1 são informações iniciais sobre o ecossistema de dados, provenientes de levantamentos, entrevistas, documentação institucional, sistemas existentes, repositórios, arquivos, painéis, APIs, bases públicas e conhecimento dos responsáveis.

As principais entradas incluem:

- lista preliminar de sistemas;
- bases de dados existentes;
- arquivos e planilhas disponíveis;
- APIs e serviços externos;
- documentação técnica;
- dicionários de dados existentes;
- relatórios institucionais;
- dashboards existentes;
- sistemas de origem;
- repositórios de dados;
- logs de sistemas;
- fluxos de eventos;
- bases históricas;
- dados públicos;
- bases geográficas;
- registros administrativos;
- registros acadêmicos;
- dados operacionais;
- informações de responsáveis;
- políticas de acesso;
- regras institucionais;
- glossários existentes;
- contratos de dados, quando houver;
- restrições legais ou institucionais;
- demandas analíticas dos usuários;
- problemas decisórios a serem investigados.

Além disso, podem ser utilizadas entrevistas, formulários, reuniões técnicas e análise documental para obter informações que não estão explicitamente registradas.

8. Processamento

O processamento do Módulo 1 envolve o levantamento, organização, classificação, validação preliminar e documentação das fontes.

8.1 Identificação das fontes

A primeira atividade consiste em identificar todas as fontes potencialmente relevantes ao projeto. Essa identificação pode ocorrer por levantamento documental, consulta a equipes técnicas, entrevistas com gestores, análise de sistemas, inspeção de repositórios, avaliação de dashboards existentes e busca por bases públicas ou externas.

O objetivo é evitar que fontes importantes permaneçam invisíveis ou sejam descobertas apenas em etapas posteriores.

8.2 Caracterização técnica

A segunda atividade consiste em registrar características técnicas da fonte, como tipo, formato, localização, mecanismo de acesso, tecnologia, volume, schema, periodicidade e forma de atualização.

Essa caracterização orienta o Módulo 2 na definição do método de extração e ingestão.

8.3 Caracterização temporal

A terceira atividade consiste em documentar a dimensão temporal da fonte. Devem ser registrados:

- período inicial e final dos dados;
- cobertura histórica;
- periodicidade de atualização;
- frequência de geração;
- latência esperada;
- existência de timestamps;
- existência de logs;
- existência de eventos;
- possibilidade de ingestão incremental;
- possibilidade de ingestão near real-time;
- existência de séries temporais;
- existência de dados longitudinais;
- existência de snapshots históricos.

Essa caracterização é essencial para análises temporais, monitoramento, detecção de drift, grafos temporais e modelos preditivos baseados em janelas de observação.

8.4 Caracterização semântica

A quarta atividade consiste em compreender o significado dos dados da fonte. Devem ser identificados conceitos, entidades, atributos, categorias, códigos, domínios, regras de negócio, termos ambíguos, sinônimos e possíveis conflitos semânticos.

Essa caracterização apoia os módulos de integração, harmonização, metadados, catálogo e interoperabilidade semântica.

8.5 Caracterização de granularidade

A quinta atividade consiste em identificar o nível de detalhe da fonte. Uma base pode estar no nível de estudante, matrícula, disciplina, turma, curso, atendimento, cidadão, família, município, território, evento ou período.

A granularidade deve ser documentada porque diferentes fontes só podem ser integradas adequadamente quando seus níveis de agregação são compreendidos.

8.6 Caracterização de qualidade preliminar

A sexta atividade consiste em avaliar, ainda de forma preliminar, possíveis problemas de qualidade da fonte, como dados ausentes, duplicidades, inconsistências, atualização irregular, baixa cobertura, campos sem documentação, categorias desconhecidas ou registros aparentemente obsoletos.

Essa avaliação não substitui o Módulo 5, mas fornece indícios iniciais para priorização e planejamento.

8.7 Classificação de sensibilidade e restrição de acesso

A sétima atividade consiste em identificar se a fonte contém dados pessoais, sensíveis, sigilosos, administrativos, restritos, públicos, anonimizados ou agregados.

Essa classificação orienta os módulos de governança, segurança, privacidade, minimização de dados e controle de acesso.

8.8 Identificação de relações potenciais entre fontes

A oitava atividade consiste em mapear possíveis conexões entre fontes. Devem ser identificadas entidades comuns, chaves de integração, atributos equivalentes, períodos compatíveis, categorias semelhantes e relações analíticas potenciais.

Essa atividade prepara o Módulo 3 — Integração e Harmonização.

8.9 Avaliação do potencial analítico

A nona atividade consiste em avaliar para quais finalidades a fonte pode ser útil. Uma fonte pode apoiar indicadores, dashboards, modelos preditivos, análises temporais, grafos, features, avaliação de políticas, monitoramento ou agentes inteligentes.

Essa avaliação contribui para priorização de fontes e planejamento do pipeline.

8.10 Registro no inventário

A décima atividade consiste em consolidar todas as informações em um inventário estruturado, com campos padronizados, responsáveis, versões, status e observações.

Esse inventário deve ser versionável e atualizável, pois novas fontes podem surgir e fontes existentes podem mudar ao longo do tempo.

9. Técnicas, métodos e instrumentos aplicáveis

O Módulo 1 pode utilizar diferentes técnicas e instrumentos para levantamento e documentação das fontes.

Entre as técnicas de levantamento, destacam-se:

- entrevistas com responsáveis pelos dados;
- questionários estruturados;
- análise documental;
- inspeção de sistemas;
- análise de repositórios;
- levantamento de dashboards existentes;
- análise de APIs;
- levantamento de bases públicas;
- reuniões técnicas;
- oficinas de mapeamento de dados;
- análise de fluxos institucionais.

Entre os métodos de documentação, podem ser utilizados:

- inventário de fontes;
- matriz de fontes por domínio;

- dicionário preliminar;
- catálogo preliminar;
- mapa de entidades;
- matriz de relacionamento entre fontes;
- ficha de fonte;
- registro de limitações;
- registro de sensibilidade;
- mapa de granularidade;
- mapa temporal da fonte;
- registro de responsáveis.

Entre as técnicas de análise preliminar, podem ser aplicadas:

- data profiling inicial;
- contagem de registros;
- inspeção de schema;
- identificação de chaves;
- análise de valores ausentes;
- identificação de duplicidades;
- análise de domínios;
- análise de periodicidade;
- análise de cobertura temporal;
- identificação de atributos sensíveis;
- análise de documentação existente.

Entre as ferramentas possíveis, podem ser consideradas:

- planilhas estruturadas;
- formulários eletrônicos;
- catálogos de dados;
- DataHub;
- OpenMetadata;
- Apache Atlas;
- CKAN;
- ferramentas de data profiling;
- SQL;
- Python;
- pandas;
- notebooks;
- ferramentas de documentação;
- sistemas de gestão de metadados;
- repositórios Git.

10. Métricas do módulo

As métricas do Módulo 1 avaliam a cobertura, completude, qualidade, documentação, governança e utilidade do inventário de fontes.

Métrica	Sigla	Descrição

Número de fontes identificadas	NFI	Quantidade total de fontes mapeadas.
Taxa de fontes documentadas	TFD	Proporção de fontes com ficha descritiva completa.
Taxa de fontes com responsável definido	TFRD	Proporção de fontes com responsável técnico ou institucional registrado.
Taxa de fontes com periodicidade documentada	TFPD	Proporção de fontes com frequência de atualização registrada.
Taxa de fontes com cobertura temporal documentada	TFCT	Proporção de fontes com período histórico conhecido.
Taxa de fontes com granularidade documentada	TFGD	Proporção de fontes com nível de detalhe identificado.
Taxa de fontes com schema conhecido	TFSC	Proporção de fontes com estrutura de atributos registrada.
Taxa de fontes com classificação de sensibilidade	TFCS	Proporção de fontes classificadas quanto à sensibilidade.
Taxa de fontes com restrição de acesso documentada	TFRA	Proporção de fontes com regras de acesso registradas.
Taxa de fontes com documentação existente	TFDE	Proporção de fontes que possuem documentação técnica ou semântica prévia.
Taxa de fontes com limitações documentadas	TFLD	Proporção de fontes com limitações conhecidas registradas.
Taxa de fontes com potencial	TFPAD	Proporção de fontes com usos analíticos

analítico definido		potenciais identificados.
Taxa de fontes temporais identificadas	TFTI	Proporção de fontes com dados históricos, temporais, eventos ou logs.
Taxa de fontes com atualização dinâmica	TFAD	Proporção de fontes com atualização incremental, near real-time ou streaming.
Grau de completude do inventário	GCI	Grau médio de preenchimento dos campos obrigatórios do inventário.
Índice de maturidade do inventário	IMI	Indicador composto da qualidade e completude do inventário.

10.1 Taxa de Fontes Documentadas — TFD

A **Taxa de Fontes Documentadas** mede a proporção de fontes identificadas que possuem informações mínimas registradas.

TFD = número de fontes documentadas / número total de fontes identificadas

Essa métrica indica o grau de formalização inicial do ecossistema de dados.

10.2 Taxa de Fontes com Responsável Definido — TFRD

A **Taxa de Fontes com Responsável Definido** mede a proporção de fontes que possuem responsável técnico, institucional ou operacional.

TFRD = número de fontes com responsável definido / número total de fontes identificadas

Essa métrica é relevante para governança, manutenção, validação e autorização de uso.

10.3 Taxa de Fontes com Cobertura Temporal Documentada — TFCT

A **Taxa de Fontes com Cobertura Temporal Documentada** mede a proporção de fontes cujo período histórico ou temporal foi identificado.

TFCT = número de fontes com cobertura temporal documentada / número total de fontes identificadas

Essa métrica é importante para análises longitudinais, séries históricas, janelas temporais, monitoramento e grafos temporais.

10.4 Grau de Completude do Inventário — GCI

O **Grau de Completude do Inventário** mede o preenchimento médio dos campos obrigatórios do inventário.

GCI = número de campos obrigatórios preenchidos / número total de campos obrigatórios esperados

Esse indicador pode ser calculado por fonte, domínio, área ou conjunto completo do inventário.

10.5 Índice de Maturidade do Inventário — IMI

O **Índice de Maturidade do Inventário** sintetiza diferentes aspectos do inventário.

Uma formulação possível é:

$$\mathbf{IMI = w_1TFD + w_2TFRD + w_3TFPD + w_4TFCT + w_5TFGD + w_6TFSC + w_7TFCS + w_8TFPAD}$$

em que:

- **TFD** representa a taxa de fontes documentadas;
- **TFRD** representa a taxa de fontes com responsável definido;
- **TFPD** representa a taxa de fontes com periodicidade documentada;
- **TFCT** representa a taxa de fontes com cobertura temporal documentada;
- **TFGD** representa a taxa de fontes com granularidade documentada;
- **TFSC** representa a taxa de fontes com schema conhecido;
- **TFCS** representa a taxa de fontes com classificação de sensibilidade;
- **TFPAD** representa a taxa de fontes com potencial analítico definido;
- w_i representa o peso atribuído a cada dimensão.

Quanto maior o IMI, maior a maturidade do inventário de fontes do UnespDataLens-RM.

11. Artefatos formalizados do módulo

Os artefatos do Módulo 1 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A1.1	Inventário de Fontes	Tabela ou catálogo	TFD, GCI	Registra fontes disponíveis e suas características principais.
A1.2	Mapa de Bases e Sistemas	Diagrama ou matriz	NFI	Visualiza sistemas, bases, áreas e domínios.

A1.3	Dicionário Preliminar de Dados	Tabela semântica	TFSC	Documenta atributos, tipos, domínios e significados iniciais.
A1.4	Matriz de Relacionamento entre Fontes	Matriz fonte × fonte	TFPAD	Identifica relações potenciais entre fontes.
A1.5	Registro de Granularidade	Tabela técnica	TFGD	Documenta nível de detalhe de cada fonte.
A1.6	Classificação de Sensibilidade	Tabela de governança	TFCS	Classifica dados quanto à privacidade, restrição e risco.
A1.7	Catálogo Preliminar de Fontes	Catálogo navegável	TFD	Organiza fontes para descoberta e consulta.
A1.8	Relatório de Riscos e Limitações	Relatório técnico	TFLD	Documenta limitações, lacunas, riscos e restrições.
A1.9	Mapa Preliminar de Entidades	Diagrama conceitual	TFPAD	Identifica entidades representadas nas fontes.
A1.10	Registro de Limitações Conhecidas	Tabela de limitações	TFLD	Consolida problemas, lacunas e incertezas das fontes.
A1.11	Mapa Temporal das Fontes	Tabela ou linha do tempo	TFCT, TFTI	Documenta cobertura histórica, periodicidade, eventos e atualização.
A1.12	Registro de Potencial Analítico	Matriz fonte × uso	TFPAD	Indica usos potenciais em indicadores, modelos, grafos e produtos.

12. Artefato A1.1 — Inventário de Fontes

12.1 Finalidade

O **Inventário de Fontes** é o principal artefato do Módulo 1. Ele registra de forma estruturada as fontes identificadas, reunindo informações técnicas, semânticas, temporais, operacionais, institucionais, analíticas e de governança.

12.2 Estrutura mínima

Campo	Descrição
ID da fonte	Identificador único da fonte.
Nome da fonte	Nome institucional ou técnico.
Descrição	Explicação resumida da fonte.
Origem	Sistema, área, órgão ou repositório de origem.
Responsável	Pessoa, equipe ou setor responsável.
Domínio	Área temática da fonte.
Tipo de fonte	Banco, API, arquivo, planilha, log, evento etc.
Formato	CSV, JSON, SQL, Parquet, API, XLSX etc.
Localização	Caminho, sistema, URL, banco ou repositório.
Periodicidade	Frequência de atualização.
Cobertura temporal	Período histórico disponível.

Granularidade	Nível de detalhe dos registros.
Volume estimado	Quantidade aproximada de registros ou tamanho.
Schema conhecido	Existência de atributos e tipos documentados.
Entidades representadas	Estudante, curso, território, atendimento etc.
Sensibilidade	Pública, restrita, pessoal, sensível etc.
Restrição de acesso	Condições de uso e acesso.
Finalidade original	Motivo pelo qual a fonte foi criada.
Finalidade analítica potencial	Possíveis usos no UnespDataLens-RM.
Documentação existente	Dicionários, manuais, contratos ou glossários.
Limitações conhecidas	Lacunas, problemas ou incertezas.
Status	Ativa, inativa, em avaliação, indisponível.

12.3 Métricas associadas

- taxa de fontes documentadas;
- grau de completude do inventário;
- taxa de fontes com responsável definido;
- taxa de fontes com cobertura temporal documentada;
- taxa de fontes com classificação de sensibilidade.

12.4 Critérios de validação

O inventário é considerado válido quando as fontes críticas possuem identificação, origem, responsável, formato, periodicidade, cobertura temporal, granularidade, sensibilidade, restrição de acesso, limitações e potencial analítico documentados.

12.5 Contribuição

Esse artefato cria a base estruturada para todo o pipeline analítico, reduzindo invisibilidade, improvisação, silos e dependência de conhecimento tácito.

13. Artefato A1.2 — Mapa de Bases e Sistemas

13.1 Finalidade

O **Mapa de Bases e Sistemas** representa visualmente ou matricialmente as fontes existentes, seus sistemas de origem, áreas responsáveis e domínios de dados.

13.2 Estrutura mínima

Campo	Descrição
Sistema ou base	Nome do sistema ou base.
Área responsável	Unidade ou setor responsável.
Domínio	Tema principal da fonte.
Tipo de dado	Acadêmico, administrativo, territorial, social etc.
Integrações conhecidas	Relações já existentes com outros sistemas.
Formato de acesso	API, banco, arquivo, planilha etc.
Status	Ativo, legado, descontinuado, em avaliação.

13.3 Métricas associadas

- número de fontes identificadas;
- taxa de sistemas mapeados;

- taxa de fontes por domínio;
- taxa de silos identificados.

13.4 Critérios de validação

O mapa é válido quando permite visualizar onde os dados estão, quem os mantém e como se organizam por domínio.

13.5 Contribuição

Esse artefato apoia a superação de silos e a priorização de fontes.

14. Artefato A1.3 — Dicionário Preliminar de Dados

14.1 Finalidade

O **Dicionário Preliminar de Dados** registra atributos, tipos, domínios, descrições e significados iniciais das fontes identificadas.

14.2 Estrutura mínima

Campo	Descrição
Fonte	Fonte associada.
Atributo	Nome do campo.
Tipo de dado	Texto, número, data, categoria etc.
Descrição	Significado preliminar.
Domínio de valores	Valores possíveis ou esperados.
Obrigatoriedade	Campo obrigatório ou opcional.
Exemplo	Valor ilustrativo.

Sensibilidade	Indicação de dado sensível ou pessoal.
Observações	Ambiguidades ou limitações.

14.3 Métricas associadas

- taxa de fontes com schema conhecido;
- taxa de atributos documentados;
- taxa de atributos com domínio definido;
- taxa de atributos sensíveis identificados.

14.4 Critérios de validação

O dicionário é válido quando atributos críticos possuem tipo, descrição, domínio e sensibilidade documentados.

14.5 Contribuição

Esse artefato apoia integração, qualidade, semântica, governança e feature engineering.

15. Artefato A1.4 — Matriz de Relacionamento entre Fontes

15.1 Finalidade

A **Matriz de Relacionamento entre Fontes** identifica relações potenciais entre fontes, considerando entidades comuns, chaves, períodos, domínios, atributos equivalentes e finalidades analíticas.

15.2 Estrutura mínima

Campo	Descrição
Fonte A	Primeira fonte.
Fonte B	Segunda fonte.
Entidade comum	Entidade compartilhada.

Chave potencial	Atributo ou regra de ligação.
Tipo de relação	Complementar, redundante, hierárquica, temporal etc.
Compatibilidade temporal	Períodos comparáveis ou não.
Compatibilidade de granularidade	Níveis compatíveis ou não.
Potencial de integração	Baixo, médio ou alto.
Riscos	Conflitos, duplicidades ou restrições.

15.3 Métricas associadas

- taxa de fontes com relação potencial identificada;
- número de relações mapeadas;
- taxa de fontes integráveis;
- taxa de conflitos potenciais.

15.4 Critérios de validação

A matriz é válida quando permite identificar quais fontes podem ser combinadas e quais cuidados são necessários.

15.5 Contribuição

Esse artefato prepara o Módulo 3 — Integração e Harmonização.

16. Artefato A1.5 — Registro de Granularidade

16.1 Finalidade

O **Registro de Granularidade** documenta o nível de detalhe de cada fonte, atributo ou registro.

16.2 Estrutura mínima

Campo	Descrição

Fonte	Fonte associada.
Unidade de registro	Entidade representada por linha ou evento.
Nível de agregação	Individual, turma, curso, território, período etc.
Periodicidade	Frequência dos registros.
Dimensão temporal	Data, mês, semestre, ano, timestamp etc.
Possibilidade de agregação	Como pode ser consolidada.
Possibilidade de desagregação	Se permite detalhamento maior.
Limitações	Restrições de interpretação.

16.3 Métricas associadas

- taxa de fontes com granularidade documentada;
- taxa de fontes com dimensão temporal definida;
- taxa de fontes compatíveis por granularidade.

16.4 Critérios de validação

O registro é válido quando permite compreender o nível de detalhe e o uso adequado da fonte.

16.5 Contribuição

Esse artefato evita integrações e análises em níveis incompatíveis.

17. Artefato A1.6 — Classificação de Sensibilidade

17.1 Finalidade

A **Classificação de Sensibilidade** identifica o grau de restrição, risco, privacidade e sensibilidade das fontes e atributos.

17.2 Estrutura mínima

Campo	Descrição
Fonte	Fonte associada.
Atributo ou conjunto de dados	Elemento classificado.
Tipo de sensibilidade	Público, restrito, pessoal, sensível etc.
Justificativa	Motivo da classificação.
Restrição de acesso	Condição de uso.
Finalidade permitida	Uso autorizado.
Necessidade de anonimização	Sim ou não.
Responsável pela classificação	Pessoa ou equipe.

17.3 Métricas associadas

- taxa de fontes com classificação de sensibilidade;
- taxa de atributos sensíveis identificados;
- taxa de fontes com restrição de acesso documentada.

17.4 Critérios de validação

A classificação é válida quando fontes e atributos críticos possuem sensibilidade e restrição de acesso documentadas.

17.5 Contribuição

Esse artefato sustenta governança, segurança, privacidade, conformidade e minimização de dados.

18. Artefato A1.7 — Catálogo Preliminar de Fontes

18.1 Finalidade

O **Catálogo Preliminar de Fontes** organiza as fontes identificadas em uma estrutura navegável, permitindo descoberta, consulta e priorização.

18.2 Estrutura mínima

Campo	Descrição
Nome da fonte	Nome de exibição.
Descrição	Resumo compreensível.
Domínio	Área temática.
Responsável	Pessoa ou setor.
Status	Ativa, inativa, em avaliação.
Tags	Termos de busca.
Sensibilidade	Classificação de acesso.
Potencial analítico	Indicadores, modelos, grafos, dashboards etc.
Link para ficha	Referência ao inventário completo.

18.3 Métricas associadas

- taxa de fontes catalogadas;
- taxa de fontes pesquisáveis;
- taxa de fontes com tags;
- taxa de fontes com responsável.

18.4 Critérios de validação

O catálogo é válido quando permite localizar e compreender rapidamente as fontes disponíveis.

18.5 Contribuição

Esse artefato apoia Data Discoverability e reduz Data Overload.

19. Artefato A1.8 — Relatório de Riscos e Limitações

19.1 Finalidade

O **Relatório de Riscos e Limitações** documenta problemas conhecidos que podem afetar o uso das fontes.

19.2 Estrutura mínima

Campo	Descrição
Fonte	Fonte avaliada.
Tipo de risco	Qualidade, acesso, privacidade, atualização, semântica etc.
Limitação	Descrição do problema.
Severidade	Baixa, média, alta ou crítica.
Impacto potencial	Módulos ou produtos afetados.
Ação recomendada	Mitigação ou cuidado necessário.
Responsável	Pessoa ou equipe.

19.3 Métricas associadas

- taxa de fontes com limitações documentadas;

- número de riscos críticos;
- taxa de fontes com risco de qualidade;
- taxa de fontes com risco de acesso.

19.4 Critérios de validação

O relatório é válido quando explicita riscos relevantes e orienta decisões de uso.

19.5 Contribuição

Esse artefato apoia governança, incerteza, qualidade e priorização de fontes.

20. Artefato A1.9 — Mapa Preliminar de Entidades

20.1 Finalidade

O **Mapa Preliminar de Entidades** identifica as entidades representadas pelas fontes e suas possíveis relações.

20.2 Estrutura mínima

Campo	Descrição
Entidade	Estudante, curso, atendimento, território etc.
Fontes associadas	Bases que registram a entidade.
Identificadores	Chaves ou atributos de identificação.
Atributos principais	Informações descritivas.
Relações potenciais	Relações com outras entidades.
Granularidade	Nível de detalhe.
Observações semânticas	Ambiguidades ou diferenças de significado.

20.3 Métricas associadas

- número de entidades identificadas;
- taxa de entidades com fonte associada;
- taxa de entidades com identificador conhecido;
- taxa de entidades com relação mapeada.

20.4 Critérios de validação

O mapa é válido quando permite compreender quais entidades existem e em quais fontes aparecem.

20.5 Contribuição

Esse artefato apoia integração, representação analítica e construção de grafos.

21. Artefato A1.10 — Registro de Limitações Conhecidas

21.1 Finalidade

O **Registro de Limitações Conhecidas** consolida restrições, lacunas, incertezas e problemas já identificados nas fontes.

21.2 Estrutura mínima

Campo	Descrição
Fonte	Fonte associada.
Limitação	Descrição da restrição.
Tipo	Temporal, semântica, qualidade, acesso, cobertura etc.
Evidência	Documento, relato, métrica ou observação.
Impacto	Como afeta o uso analítico.
Recomendaçã	Cuidado ou ação necessária.

o

Status	Aberta, mitigada, aceita, em análise.

21.3 Métricas associadas

- taxa de fontes com limitações documentadas;
- número de limitações críticas;
- taxa de limitações mitigadas.

21.4 Critérios de validação

O registro é válido quando permite que limitações sejam consideradas antes do uso analítico.

21.5 Contribuição

Esse artefato reduz incerteza e previne interpretações inadequadas.

22. Artefato A1.11 — Mapa Temporal das Fontes

22.1 Finalidade

O **Mapa Temporal das Fontes** documenta a dimensão temporal das fontes, incluindo cobertura histórica, periodicidade, atualização, latência, eventos e possibilidade de ingestão dinâmica.

22.2 Estrutura mínima

Campo	Descrição
Fonte	Fonte avaliada.
Data inicial disponível	Início da cobertura histórica.
Data final disponível	Fim da cobertura ou atualização atual.
Periodicidade	Diária, semanal, mensal, semestral etc.

Atributo temporal principal	Campo de data ou timestamp.
Tipo temporal	Histórico, longitudinal, série, evento, log, streaming.
Latência esperada	Tempo entre geração e disponibilidade.
Possibilidade incremental	Sim ou não.
Possibilidade near real-time	Sim ou não.
Lacunas temporais conhecidas	Períodos ausentes ou incompletos.

22.3 Métricas associadas

- taxa de fontes com cobertura temporal documentada;
- taxa de fontes temporais identificadas;
- taxa de fontes com atualização dinâmica;
- taxa de fontes com lacunas temporais conhecidas.

22.4 Critérios de validação

O mapa é válido quando permite compreender se a fonte suporta análises históricas, temporais, longitudinais, eventos ou fluxos dinâmicos.

22.5 Contribuição

Esse artefato fortalece a capacidade do UnespDataLens-RM de tratar dados temporais, eventos, logs, séries históricas e grafos temporais.

23. Artefato A1.12 — Registro de Potencial Analítico

23.1 Finalidade

O **Registro de Potencial Analítico** relaciona cada fonte às possíveis finalidades analíticas que ela pode apoiar.

23.2 Estrutura mínima

Campo	Descrição
Fonte	Fonte avaliada.
Indicadores possíveis	Indicadores que podem ser produzidos.
Dashboards possíveis	Painéis que podem ser alimentados.
Modelos possíveis	Modelos preditivos, classificatórios ou explicativos.
Features possíveis	Atributos derivados potenciais.
Grafos possíveis	Entidades e relações que podem ser modeladas.
Análises temporais	Séries, trajetórias, evolução ou eventos.
Agentes inteligentes	Usos em consulta, explicação ou recomendação.
Prioridade	Alta, média ou baixa.
Limitações	Restrições de uso.

23.3 Métricas associadas

- taxa de fontes com potencial analítico definido;
- número de usos analíticos por fonte;
- taxa de fontes prioritárias;
- taxa de fontes reutilizáveis.

23.4 Critérios de validação

O registro é válido quando permite priorizar fontes com base em sua utilidade para o modelo de referência.

23.5 Contribuição

Esse artefato conecta o inventário às finalidades analíticas do UnespDataLens-RM.

24. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Sistemas, bases, arquivos, APIs, planilhas, logs, eventos, fluxos, bases públicas, documentação, entrevistas, dashboards existentes, responsáveis e demandas analíticas.
Processamento	Identificação, caracterização técnica, temporal, semântica, de granularidade, qualidade preliminar, sensibilidade, relações potenciais e potencial analítico das fontes.
Artefatos	Inventário de fontes, mapa de bases, dicionário preliminar, matriz de relacionamento, registro de granularidade, classificação de sensibilidade, catálogo preliminar, relatório de riscos, mapa de entidades, mapa temporal e registro de potencial analítico.
Métricas	NFI, TFD, TFRD, TFPD, TFCT, TFGD, TFSC, TFCS, TFRA, TFDE, TFLD, TFPAD, TFTI, TFAD, GCI e IMI.
Saídas	Fontes documentadas, classificadas, avaliadas e priorizadas para extração, ingestão, integração, transformação, qualidade, governança, lineage, representação e consumo analítico.

25. Critérios gerais de validação do módulo

O Módulo 1 pode ser considerado adequadamente implementado quando:

1. as fontes relevantes foram identificadas;
2. fontes críticas possuem responsáveis definidos;
3. cada fonte possui descrição, origem, domínio e tipo documentados;
4. formatos, localização e mecanismos de acesso estão registrados;
5. periodicidade e cobertura temporal estão documentadas;
6. granularidade e unidade de registro foram identificadas;
7. schemas ou atributos principais foram mapeados;
8. entidades representadas foram documentadas;
9. restrições de acesso e sensibilidade foram classificadas;
10. limitações conhecidas foram registradas;

11. documentação existente foi associada à fonte;
12. relações potenciais entre fontes foram mapeadas;
13. fontes com dados temporais, eventos, logs ou fluxos dinâmicos foram identificadas;
14. potencial analítico das fontes foi avaliado;
15. o inventário está estruturado, versionado e disponível para os módulos posteriores.

26. Relação com os demais módulos do modelo de referência

O Módulo 1 fornece a base para todos os módulos posteriores do UnespDataLens-RM.

Com o **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, fornece informações sobre formato, localização, periodicidade, cobertura temporal, restrição de acesso, tipo de fonte, possibilidade de ingestão incremental e mecanismos de extração.

Com o **Módulo 3 — Integração e Harmonização**, fornece entidades, chaves, atributos equivalentes, granularidade e relações potenciais entre fontes.

Com o **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, fornece informações preliminares sobre atributos, domínios, formatos e limitações.

Com o **Módulo 5 — Qualidade e Validação de Dados**, fornece indícios iniciais de problemas de qualidade, documentação de limitações e características esperadas das fontes.

Com o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, identifica fontes com atributos sensíveis, grupos relevantes e riscos de representatividade.

Com o **Módulo 7 — Armazenamento Analítico**, orienta organização por domínio, fonte, granularidade e periodicidade.

Com o **Módulo 8 — Disponibilização e Consumo Analítico**, indica quais fontes podem originar produtos de dados, dashboards, APIs e relatórios.

Com o **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, fornece os metadados iniciais para catálogo, dicionário, glossário e modelo semântico.

Com o **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, fornece classificação de sensibilidade, restrições de acesso, finalidade e responsáveis.

Com o **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, fornece o ponto inicial da linhagem dos dados.

Com o **Módulo 12 — Versionamento e Reprodutibilidade**, orienta versionamento de fontes, schemas, snapshots e documentação.

Com o **Módulo 13 — Representações Analíticas**, fornece informações sobre entidades, atributos, temporalidade e relações que podem ser transformadas em tabelas, séries, eventos, grafos e grafos temporais.

Com o **Módulo 14 — Feature Engineering e Feature Store**, indica fontes e atributos com potencial para criação de features.

Com o **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, fornece periodicidade, baselines iniciais, cobertura temporal, schemas esperados e responsáveis.

Com o **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, fornece metadados, descrições e relações que podem ser consultadas por agentes para busca, explicação e recomendação.

27. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 1 pode identificar fontes como sistema acadêmico, base de matrícula, histórico escolar, ambiente virtual de aprendizagem, registros de frequência, notas, questionários socioeconômicos, dados de bolsas, registros de evasão, relatórios institucionais e dashboards existentes.

Cada fonte deve ser documentada quanto à origem, responsável, período disponível, periodicidade, granularidade, atributos, entidades, sensibilidade e potencial analítico.

Por exemplo, a base de matrícula pode possuir granularidade por estudante e período letivo, atualização semestral, cobertura histórica de vários anos e atributos como curso, forma de ingresso, situação acadêmica e período. Já o ambiente virtual pode conter logs de eventos com timestamp, permitindo análises de engajamento, trajetórias e grafos temporais de interação.

Com esse inventário, o UnespDataLens-RM pode selecionar fontes adequadas para diferentes análises institucionais, científicas e organizacionais, evitando o uso de dados sem documentação ou fora de contexto.

28. Exemplo de aplicação em políticas públicas

Em políticas públicas, o Módulo 1 pode identificar fontes de diferentes secretarias e órgãos, como saúde, educação, assistência social, habitação, trabalho, orçamento, território, atendimento, infraestrutura e indicadores socioeconômicos.

Cada fonte pode possuir formatos, periodicidades e granularidades distintas. Uma base de atendimento pode estar no nível de cidadão ou família; uma base territorial pode estar agregada por bairro; uma base de saúde pode ter registros mensais; uma base de educação pode estar organizada por escola ou estudante.

O inventário permite compreender quais fontes podem ser integradas, quais exigem harmonização semântica, quais possuem restrições de acesso, quais têm cobertura temporal suficiente e quais podem apoiar indicadores, dashboards, modelos ou análises intersetoriais.

29. Contribuição técnico-científica do módulo

A contribuição técnico-científica do Módulo 1 está em formalizar o conhecimento inicial sobre as fontes de dados como uma etapa explícita, mensurável, documentada e avaliável da Engenharia Analítica de Dados.

Em vez de iniciar o pipeline diretamente pela extração, o UnespDataLens-RM propõe que as fontes sejam previamente identificadas, caracterizadas, classificadas, avaliadas e

documentadas. Essa decisão metodológica reduz improvisação, aumenta governança, melhora rastreabilidade e permite maior qualidade nos módulos posteriores.

Essa formalização contribui para:

- reduzir invisibilidade das fontes;
- mapear silos de dados;
- documentar responsáveis;
- identificar fontes temporais, históricas, eventuais e dinâmicas;
- registrar granularidade e cobertura temporal;
- apoiar integração e interoperabilidade semântica;
- orientar extração e ingestão;
- reduzir riscos de privacidade e uso indevido;
- apoiar governança e conformidade;
- melhorar qualidade dos metadados;
- reduzir data debt;
- ampliar data discoverability;
- orientar representações analíticas, grafos e features;
- fortalecer a base metodológica do modelo de referência.

30. Síntese do módulo

O **Módulo 1 — Fontes e Inventário de Dados** constitui a etapa inicial do UnespDataLens-RM e tem como função transformar fontes dispersas, heterogêneas e pouco documentadas em um inventário estruturado, governável, rastreável e analiticamente útil.

Sua importância está em reconhecer que a qualidade do pipeline analítico começa antes da extração dos dados. É necessário compreender quais fontes existem, quem as mantém, o que representam, como são atualizadas, qual sua cobertura temporal, qual sua granularidade, quais restrições possuem, quais limitações apresentam e qual potencial analítico oferecem.

Ao produzir artefatos como inventário de fontes, mapa de bases, dicionário preliminar, matriz de relacionamento, registro de granularidade, classificação de sensibilidade, catálogo preliminar, relatório de riscos, mapa de entidades, mapa temporal e registro de potencial analítico, o Módulo 1 estabelece as condições necessárias para que os demais módulos do UnespDataLens-RM operem com maior rigor, rastreabilidade, governança e confiabilidade.

Módulo 2 — Extração, Ingestão e Armazenamento Bruto

1. Finalidade do módulo

O **Módulo 2 — Extração, Ingestão e Armazenamento Bruto** tem como finalidade operacionalizar a entrada controlada dos dados no ambiente analítico do modelo de referência **UnespDataLens-RM**, a partir das fontes previamente identificadas, caracterizadas e documentadas no **Módulo 1 — Fontes e Inventário de Dados**.

Esse módulo é responsável por definir, executar, registrar, validar e preservar os procedimentos de extração e ingestão de dados provenientes de múltiplas fontes, garantindo que os dados sejam capturados de forma controlada, mantidos em estado bruto, acompanhados de metadados técnicos e operacionais, versionados, auditáveis e preparados para as etapas posteriores de integração, harmonização, transformação, qualidade, governança, proveniência, linhagem, versionamento, reprodutibilidade, monitoramento e disponibilização analítica.

No contexto do UnespDataLens-RM, a extração e a ingestão não são tratadas apenas como operações técnicas de cópia, importação ou movimentação de dados. Elas são concebidas como processos formais, documentados, monitoráveis e auditáveis, nos quais cada execução deve registrar a origem dos dados, data e hora da extração, método utilizado, volume capturado, schema observado, parâmetros de conexão, formato dos arquivos, falhas, registros rejeitados, evidências de integridade, versão da fonte, responsável pela execução, restrições de acesso e condições de uso.

A preservação dos dados em estado bruto constitui um princípio central deste módulo. Ao armazenar uma cópia original, ou o mais próximo possível da origem, o modelo de referência cria condições para auditoria, reprocessamento, comparação entre versões, reconstrução de pipelines, análise de mudanças ao longo do tempo e avaliação de eventuais alterações estruturais, semânticas ou operacionais ocorridas nas fontes.

O Módulo 2 também deve contemplar diferentes formas de ingestão, incluindo ingestão em lote, ingestão incremental, ingestão por APIs, ingestão por arquivos, ingestão por conectores, ingestão orientada a eventos, ingestão quase em tempo real e, quando aplicável, ingestão por streaming. Essa ampliação é importante porque o UnespDataLens-RM não se limita a bases estáticas ou históricas; ele deve ser capaz de lidar com dados temporais, registros longitudinais, logs, eventos, séries históricas, atualizações periódicas e fluxos dinâmicos de dados.

Dessa forma, o Módulo 2 constitui a primeira materialização dos dados no pipeline analítico do UnespDataLens-RM. Ele transforma fontes documentadas em datasets brutos armazenados, com rastros de execução, evidências técnicas, metadados operacionais, registros de integridade, snapshots e permissões de acesso. Sua contribuição é essencial para garantir que os módulos seguintes trabalhem sobre dados cuja origem, forma de captura, versão, estrutura e condições de ingestão sejam conhecidas e verificáveis.

2. Objetivo geral

O objetivo geral do Módulo 2 é **extrair, ingerir e armazenar dados brutos de forma controlada, documentada, versionável, auditável e reprodutível**, preservando sua

origem, estrutura inicial, parâmetros de coleta, registros de execução, metadados operacionais, evidências de integridade e condições de acesso, de modo a apoiar os módulos posteriores do modelo de referência UnespDataLens-RM.

3. Objetivos específicos

São objetivos específicos do Módulo 2:

1. definir estratégias de extração adequadas às fontes identificadas no Módulo 1;
2. estabelecer procedimentos de ingestão compatíveis com diferentes formatos, periodicidades, volumes e restrições de acesso;
3. preservar os dados em estado bruto, evitando transformações irreversíveis no momento da ingestão;
4. registrar logs completos de execução, incluindo data, hora, responsável, status, falhas, volume extraído, tempo de execução e destino dos dados;
5. criar snapshots ou versões iniciais dos dados capturados;
6. registrar metadados técnicos e operacionais associados à ingestão;
7. validar preliminarmente estrutura, formato, volume, integridade, disponibilidade e conformidade mínima dos dados ingeridos;
8. detectar falhas de extração, divergências de schema, dados ausentes, arquivos corrompidos, registros rejeitados ou incompatibilidades de formato;
9. controlar diferentes modalidades de ingestão, como batch, incremental, near real-time e streaming;
10. registrar parâmetros de conexão, filtros, janelas temporais, endpoints, credenciais controladas, formatos e condições de acesso;
11. calcular evidências de integridade, como checksums, contagens de registros, tamanho de arquivos e comparação entre origem e destino;
12. classificar o armazenamento bruto conforme sensibilidade, restrição de acesso, finalidade e política de retenção;
13. fornecer evidências para os módulos de governança, lineage, versionamento, reprodutibilidade, observabilidade e DataOps;
14. preparar datasets brutos para integração, transformação, validação e armazenamento analítico;
15. permitir auditoria, reprocessamento e comparação entre extrações realizadas em diferentes momentos.

4. Fundamentação conceitual do módulo

O Módulo 2 fundamenta-se na ideia de que a entrada dos dados no ambiente analítico é uma etapa crítica do ciclo de Engenharia Analítica de Dados. A confiabilidade dos produtos finais depende da capacidade de demonstrar de onde os dados vieram, quando foram capturados, como foram capturados, em que formato foram preservados, quais parâmetros foram utilizados, quais falhas ocorreram e qual versão da fonte foi materializada.

A extração corresponde ao processo de obter dados a partir de uma fonte. Essa fonte pode ser um banco de dados, uma API, um arquivo, uma planilha, um sistema institucional, um log, um sensor, um repositório público, uma base governamental, uma aplicação web, um ambiente virtual de aprendizagem ou qualquer outro recurso capaz de fornecer dados ao projeto analítico.

A ingestão corresponde à entrada desses dados no ambiente controlado do modelo de referência. Essa entrada pode ocorrer por carga em lote, atualização incremental, captura

contínua de eventos, sincronização por API, envio de arquivos, leitura de logs ou consumo de fluxos de mensagens.

O armazenamento bruto corresponde à preservação dos dados em seu estado original ou em uma representação mínima de persistência, sem aplicar transformações analíticas definitivas. O objetivo é manter uma referência confiável daquilo que foi recebido ou capturado, permitindo rastreabilidade, auditoria e reprocessamento.

No UnespDataLens-RM, esse módulo deve contemplar diferentes padrões de ingestão:

Tipo de ingestão	Descrição	Exemplo
Batch	Extração periódica em lotes.	Carga mensal de base acadêmica.
Incremental	Captura apenas de registros novos ou alterados.	Matrículas atualizadas desde a última execução.
Full load	Cópia completa da fonte a cada execução.	Exportação completa de uma base institucional.
API-based	Captura por chamadas a serviços ou endpoints.	Consulta a API de dados governamentais.
File-based	Ingestão por arquivos enviados ou disponibilizados.	CSV, JSON, XML, Parquet ou planilha.
Event-based	Captura de eventos gerados por sistemas.	Logs de acesso em ambiente virtual.
Near real-time	Ingestão frequente com baixa latência.	Atualizações a cada poucos minutos.
Streaming	Captura contínua de eventos em fluxo.	Mensagens Kafka ou eventos de sensores.

A inclusão explícita dessas modalidades é importante porque o modelo de referência deve lidar tanto com dados históricos e administrativos quanto com dados temporais, logs, eventos, trajetórias e fluxos dinâmicos.

À luz da **Design Science Research**, os artefatos deste módulo são tratados como unidades técnico-metodológicas projetadas para resolver problemas concretos de ingestão, preservação, rastreabilidade e reprodutibilidade. Cada artefato possui finalidade, estrutura mínima, regras de construção, métricas associadas, critérios de validação e relação com os demais módulos do modelo de referência.

5. Formalização sintética do módulo

O Módulo 2 pode ser compreendido como um processo de transformação de fontes documentadas em datasets brutos preservados, acompanhados de registros operacionais e evidências de integridade.

Considerando o conjunto de fontes documentadas no Módulo 1:

$$\mathbf{F = \{f_1, f_2, ..., f_n\}}$$

em que cada $\mathbf{f_i}$ representa uma fonte de dados identificada, o Módulo 2 define um conjunto de processos de extração e ingestão:

$$\mathbf{E = \{e_1, e_2, ..., e_k\}}$$

em que cada processo $\mathbf{e_i}$ pode ser descrito por:

$$\mathbf{e_i = \{id, fonte, método, parâmetros, janela_temporal, data_execução, responsável, volume_extraído, schema_detectado, status, logs, erros, destino_bruto, versão, evidência_integridade\}}$$

O módulo produz um conjunto de ativos brutos:

$$\mathbf{B = \{b_1, b_2, ..., b_m\}}$$

em que cada ativo bruto $\mathbf{b_j}$ pode ser representado por:

$$\mathbf{b_j = \{id, fonte_origem, caminho, formato, tamanho, número_registros, número_atributos, data_ingestão, checksum, versão, periodicidade, janela_temporal, sensibilidade, restrição_acesso, log_execução, manifesto, status_integridade\}}$$

Dessa forma, os dados deixam de ser tratados apenas como arquivos copiados para um repositório e passam a ser representados como ativos brutos documentados, versionáveis, auditáveis e reutilizáveis no ciclo de Engenharia Analítica de Dados.

6. Entradas do módulo

As entradas do Módulo 2 incluem as fontes identificadas e documentadas no Módulo 1, bem como as informações técnicas e contextuais necessárias para sua extração e ingestão.

As principais entradas são:

- fontes de dados inventariadas;
- credenciais ou mecanismos autorizados de acesso;

- endpoints de APIs;
- conexões com bancos de dados;
- arquivos disponibilizados para ingestão;
- planilhas eletrônicas;
- logs de sistemas;
- fluxos de eventos;
- sensores;
- bases públicas;
- bases governamentais;
- bases históricas;
- registros operacionais;
- dados acadêmicos;
- dados administrativos;
- dados territoriais;
- dados socioeconômicos;
- dados educacionais;
- dados de saúde;
- dados de políticas públicas;
- dados de trabalho;
- documentação técnica das fontes;
- periodicidade de atualização;
- regras de acesso;
- classificação de sensibilidade;
- restrições institucionais;
- janelas temporais de extração;
- parâmetros de consulta;
- schemas esperados;
- contratos de dados, quando disponíveis.

Também são entradas relevantes os metadados produzidos no Módulo 1, como:

- identificador da fonte;
- nome da fonte;
- origem;
- responsável;
- formato;
- granularidade;
- cobertura temporal;
- periodicidade;
- restrição de acesso;
- sensibilidade;
- finalidade;
- documentação existente;
- limitações conhecidas;
- relações potenciais com outras fontes.

7. Processamento

O processamento do Módulo 2 envolve atividades de planejamento, extração, ingestão, validação preliminar, armazenamento bruto, registro de logs, geração de manifestos, controle de integridade e organização dos dados em zonas apropriadas.

7.1 Planejamento da extração

A primeira atividade consiste em definir como cada fonte será extraída. Devem ser considerados formato, volume, periodicidade, mecanismo de acesso, restrições, sensibilidade, janela temporal, necessidade de anonimização, latência esperada, custo operacional e impacto na fonte de origem.

O plano de extração deve indicar se a captura será completa, incremental, periódica, sob demanda, orientada a eventos, quase em tempo real ou contínua.

7.2 Configuração da conexão ou mecanismo de captura

A segunda atividade consiste em configurar os mecanismos técnicos necessários para acessar a fonte. Isso pode incluir conexões com bancos de dados, chamadas a APIs, leitura de arquivos, conectores, credenciais, tokens, parâmetros de autenticação, filas de mensagens, buckets de armazenamento, pastas monitoradas ou endpoints externos.

Essa configuração deve ser documentada e, quando possível, versionada.

7.3 Execução da extração

A terceira atividade consiste em executar a captura dos dados. Durante essa execução, devem ser registrados volume extraído, tempo de execução, data e hora, filtros utilizados, janela temporal, quantidade de registros, quantidade de atributos, status da operação, mensagens de erro e eventuais registros rejeitados.

7.4 Ingestão no ambiente controlado

A quarta atividade consiste em transferir os dados capturados para o ambiente analítico do UnespDataLens-RM. A ingestão deve preservar a forma original dos dados sempre que possível, mantendo arquivos, tabelas ou registros em uma zona bruta.

A ingestão pode ocorrer em:

- zona raw;
- staging area;
- object storage;
- banco relacional;
- banco NoSQL;
- data lake;
- lakehouse;
- repositório de arquivos;
- armazenamento de logs;
- fila ou tópico de eventos;
- estrutura particionada por data, fonte ou domínio.

7.5 Validação preliminar da ingestão

A quinta atividade consiste em realizar validações iniciais, sem ainda aplicar transformações analíticas profundas. Essas validações podem verificar:

- existência do arquivo ou tabela;
- legibilidade;
- formato esperado;
- schema observado;
- número de registros;

- tamanho do arquivo;
- presença de campos obrigatórios;
- conformidade mínima com contrato de dados;
- integridade por checksum;
- divergência entre origem e destino;
- duplicidade de arquivos;
- registros corrompidos;
- falhas de codificação;
- atualidade da carga.

7.6 Preservação de snapshots

A sexta atividade consiste em preservar snapshots ou versões dos dados ingeridos. Um snapshot representa uma materialização dos dados em determinado momento, com data, origem, versão, parâmetros e evidências associadas.

Snapshots são essenciais para reprocessamento, auditoria, comparação temporal, versionamento e reprodutibilidade.

7.7 Registro de logs e metadados operacionais

A sétima atividade consiste em registrar logs estruturados da execução. Esses logs devem permitir responder:

- quando a ingestão ocorreu?
- quem ou qual processo executou?
- qual fonte foi utilizada?
- quais parâmetros foram aplicados?
- quantos registros foram capturados?
- houve erro?
- houve registros rejeitados?
- qual foi o destino?
- qual versão foi gerada?
- quais evidências de integridade foram calculadas?

7.8 Geração do manifesto de ingestão

A oitava atividade consiste em gerar um manifesto de ingestão, isto é, um documento estruturado contendo as informações necessárias para reconstruir ou auditar a execução.

O manifesto deve registrar fonte, versão, parâmetros, ambiente, arquivos gerados, logs, checksums, schema observado, tempo de execução, status, responsável e destino.

7.9 Organização do armazenamento bruto

A nona atividade consiste em organizar os dados brutos conforme uma convenção de estrutura, nomenclatura, particionamento e retenção.

Exemplo de organização:

/raw/{dominio}/{fonte}/{ano}/{mes}/{dia}/{versao}/

Essa estrutura facilita descoberta, versionamento, controle de acesso, lineage, observabilidade e reprocessamento.

7.10 Tratamento de falhas de ingestão

A décima atividade consiste em registrar falhas, rejeições, exceções e inconsistências encontradas durante a extração ou ingestão. O módulo deve indicar severidade, causa provável, impacto, responsável e ação recomendada.

Falhas críticas podem bloquear a disponibilização dos dados para módulos posteriores.

8. Técnicas, métodos e tecnologias aplicáveis

O Módulo 2 pode empregar diferentes técnicas, métodos e tecnologias para extração, ingestão e armazenamento bruto.

Entre as técnicas de extração, destacam-se:

- extração batch;
- extração incremental;
- extração full load;
- captura por API;
- leitura de arquivos;
- consulta SQL;
- exportação de sistemas;
- coleta por conectores;
- captura de logs;
- ingestão orientada a eventos;
- captura near real-time;
- streaming de dados;
- change data capture;
- web services;
- filas e tópicos de mensagens.

Entre as técnicas de ingestão e persistência, podem ser utilizadas:

- staging area;
- zona raw;
- data lake;
- lakehouse;
- object storage;
- particionamento por data;
- particionamento por fonte;
- organização por domínio;
- armazenamento em CSV;
- armazenamento em JSON;
- armazenamento em XML;
- armazenamento em Parquet;
- armazenamento em Avro;
- armazenamento em ORC;
- tabelas relacionais;
- armazenamento de eventos;
- armazenamento de logs.

Entre os métodos de validação preliminar, destacam-se:

- validação de schema;

- validação de formato;
- contagem de registros;
- verificação de campos obrigatórios;
- comparação de volume com extrações anteriores;
- cálculo de checksum;
- verificação de duplicidade de arquivos;
- verificação de tamanho;
- validação de encoding;
- validação de data de referência;
- detecção de arquivos corrompidos;
- conformidade com contrato de dados;
- registro de divergências entre origem e destino.

Entre as tecnologias aplicáveis, podem ser consideradas:

- Apache Airflow;
- Prefect;
- Dagster;
- Apache NiFi;
- Apache Kafka;
- Spark;
- PySpark;
- pandas;
- SQL;
- dbt;
- DuckDB;
- PostgreSQL;
- object storage;
- Delta Lake;
- Apache Iceberg;
- Apache Hudi;
- LakeFS;
- DVC;
- MLflow;
- Great Expectations;
- Pandera;
- Soda Core;
- Deequ;
- OpenLineage;
- Marquez;
- DataHub;
- OpenMetadata.

9. Métricas do módulo

As métricas do Módulo 2 avaliam a efetividade, confiabilidade, completude, integridade, rastreabilidade e desempenho dos processos de extração, ingestão e armazenamento bruto.

Métrica	Sigla	Descrição

Número de fontes ingeridas	NFIg	Quantidade total de fontes efetivamente ingeridas no ambiente analítico.
Taxa de ingestão bem-sucedida	TIBS	Proporção de execuções de ingestão concluídas sem falhas críticas.
Volume de dados ingeridos	VDI	Quantidade de registros, arquivos ou bytes ingeridos.
Taxa de registros rejeitados	TRR	Proporção de registros não aceitos na ingestão.
Taxa de falhas de ingestão	TFI	Proporção de execuções com erro.
Tempo médio de ingestão	TMI	Tempo médio necessário para extrair e ingerir os dados.
Taxa de conformidade de schema	TCS	Proporção de ingestões compatíveis com o schema esperado.
Taxa de atualização no prazo	TAP	Proporção de ingestões executadas conforme periodicidade esperada.
Cobertura de logs de execução	CLE	Proporção de execuções com logs estruturados completos.
Cobertura de manifestos operacionais	CMO	Proporção de execuções com manifesto de ingestão.
Taxa de snapshots gerados	TSG	Proporção de ingestões que geraram snapshot preservado.
Taxa de fontes com armazenamento bruto	TFAB	Proporção de fontes ingeridas preservadas em zona bruta.

Taxa de rastreabilidade da origem	TRO	Proporção de datasets brutos associados à fonte original e metadados de ingestão.
Taxa de ativos com checksum preservado	TACP	Proporção de arquivos ou datasets brutos com evidência de integridade.
Taxa de ingestão incremental correta	TIIC	Proporção de ingestões incrementais sem duplicação ou perda de registros.
Latência de ingestão	LI	Tempo entre a geração do dado na fonte e sua disponibilidade no ambiente bruto.
Taxa de eventos processados	TEP	Proporção de eventos capturados e armazenados corretamente em fluxos contínuos.

9.1 Taxa de Ingestão Bem-Sucedida — TIBS

A **Taxa de Ingestão Bem-Sucedida** mede a proporção de execuções concluídas sem falhas críticas.

TIBS = número de ingestões bem-sucedidas / número total de ingestões executadas

Essa métrica indica a estabilidade operacional do processo de ingestão.

9.2 Taxa de Registros Rejeitados — TRR

A **Taxa de Registros Rejeitados** mede a proporção de registros descartados, inválidos, corrompidos ou não ingeridos.

TRR = número de registros rejeitados / número total de registros extraídos

Valores elevados indicam problemas de formato, schema, qualidade, encoding, integridade ou regra de ingestão.

9.3 Taxa de Conformidade de Schema — TCS

A **Taxa de Conformidade de Schema** mede se os dados ingeridos mantêm compatibilidade com o schema esperado.

TCS = número de ingestões compatíveis com schema esperado / número total de ingestões avaliadas

Essa métrica é importante para detectar schema drift logo na entrada dos dados.

9.4 Cobertura de Logs de Execução — CLE

A **Cobertura de Logs de Execução** mede a proporção de execuções com logs completos.

CLE = número de execuções com logs estruturados / número total de execuções

Logs completos são fundamentais para DataOps, observabilidade, auditoria e reprodutibilidade.

9.5 Cobertura de Manifestos Operacionais — CMO

A **Cobertura de Manifestos Operacionais** mede a proporção de ingestões acompanhadas de manifesto.

CMO = número de ingestões com manifesto / número total de ingestões

O manifesto registra dados, parâmetros, versões, ambiente, destino e evidências da execução.

9.6 Latência de Ingestão — LI

A **Latência de Ingestão** mede o tempo entre a geração ou atualização do dado na fonte e sua disponibilidade no ambiente bruto.

LI = horário de disponibilidade no ambiente bruto − horário de geração ou atualização na fonte

Essa métrica é especialmente relevante para ingestões near real-time, streaming, logs e dados operacionais com atualização frequente.

10. Artefatos formalizados do módulo

Os artefatos do Módulo 2 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A2.1	Plano de Extração e Ingestão	Documento técnico	TIBS	Define estratégia de captura e entrada dos dados.
A2.2	Configuração de Ingestão	Arquivo YAML, JSON, script ou tabela	CMO	Registra parâmetros, fonte, conexão e método de ingestão.
A2.3	Log de Execução	Log estruturado	CLE	Documenta execução,

	da Ingestão			falhas, volumes e status.
A2.4	Dataset Bruto	Arquivo, tabela ou objeto armazenado	TFAB	Preserva dados no estado original.
A2.5	Registro de Snapshot	Tabela ou manifesto versionado	TSG	Preserva versões temporais dos dados ingeridos.
A2.6	Manifesto de Ingestão	Documento estruturado	CMO	Permite auditoria e reprodutibilidade da ingestão.
A2.7	Relatório de Falhas e Rejeições	Relatório técnico	TRR, TFI	Registra erros, registros rejeitados e ações necessárias.
A2.8	Registro de Integridade	Checksums e contagens	TACP	Verifica integridade entre origem e destino.
A2.9	Mapa de Armazenamento Bruto	Diagrama ou tabela	TFAB	Organiza localização, zonas, formatos e partições.
A2.10	Registro de Permissões de Ingestão	Matriz de acesso	TACP	Controla acesso, credenciais, restrições e responsáveis.
A2.11	Registro de Janelas Temporais de Ingestão	Tabela temporal	LI, TAP	Documenta períodos extraídos, recortes e atualização.
A2.12	Registro de Eventos ou Streaming	Log de eventos ou tópico monitorado	TEP	Documenta ingestões orientadas a eventos ou fluxos contínuos.

11. Artefato A2.1 — Plano de Extração e Ingestão

11.1 Finalidade

O **Plano de Extração e Ingestão** define como cada fonte será capturada, com qual periodicidade, por qual método, com quais parâmetros, restrições, responsáveis, formato de destino e critérios de validação preliminar.

11.2 Tipo de artefato na DSR

Na perspectiva da Design Science Research, esse artefato é simultaneamente:

- um **método**, pois define o procedimento de extração e ingestão;
- um **modelo**, pois estrutura os elementos necessários ao planejamento da captura;
- uma **instanciação**, quando implementado em documentação, planilha, workflow, YAML, JSON ou ferramenta de orquestração.

11.3 Estrutura mínima

Campo	Descrição
Identificador da ingestão	Código único do processo.
Fonte	Fonte de dados associada.
Tipo de ingestão	Batch, incremental, full load, API, evento, streaming etc.
Método de extração	SQL, API, arquivo, conector, log, evento etc.
Periodicidade	Frequência esperada da ingestão.
Janela temporal	Período a ser capturado.
Formato de origem	CSV, JSON, API, tabela, log etc.
Formato de destino	CSV, Parquet, JSON, tabela, tópico etc.

Zona de destino	Raw, staging, landing ou outra.
Responsável	Pessoa, equipe ou sistema responsável.
Critérios de validação	Schema, volume, checksum, campos obrigatórios etc.
Restrições	Acesso, sensibilidade, privacidade ou limitação técnica.

11.4 Métricas associadas

- taxa de ingestão bem-sucedida;
- taxa de atualização no prazo;
- tempo médio de ingestão;
- taxa de falhas de ingestão;
- cobertura de manifestos operacionais.

11.5 Critérios de validação

O plano é considerado válido quando define fonte, método, periodicidade, responsável, destino, formato, critérios de validação e restrições de acesso.

11.6 Contribuição

Esse artefato reduz improvisação na entrada dos dados e transforma a ingestão em um processo planejado, controlável e avaliável.

12. Artefato A2.2 — Configuração de Ingestão

12.1 Finalidade

A **Configuração de Ingestão** registra os parâmetros técnicos necessários para executar a captura dos dados.

12.2 Estrutura mínima

Campo	Descrição
Identificador da configuração	Código único.
Fonte	Fonte associada.

Tipo de conexão	Banco, API, arquivo, fila, tópico etc.
Parâmetros	Host, endpoint, caminho, filtros, período etc.
Credenciais	Referência segura, nunca senha em texto aberto.
Formato esperado	CSV, JSON, XML, Parquet, tabela etc.
Schema esperado	Estrutura prevista dos dados.
Frequência	Agendamento ou gatilho.
Ambiente	Desenvolvimento, teste, homologação ou produção.
Versão da configuração	Controle de mudança.

12.3 Métricas associadas

- percentual de configurações versionadas;
- taxa de execuções com configuração identificada;
- taxa de falhas por configuração inválida;
- taxa de conformidade com schema esperado.

12.4 Critérios de validação

A configuração é válida quando permite executar a ingestão de forma controlada, segura, versionável e auditável.

12.5 Contribuição

Esse artefato permite automação, reprodutibilidade e governança operacional das ingestões.

13. Artefato A2.3 — Log de Execução da Ingestão

13.1 Finalidade

O **Log de Execução da Ingestão** registra os eventos operacionais associados a cada execução de extração e ingestão.

13.2 Estrutura mínima

Campo	Descrição
ID da execução	Código único da execução.
Fonte	Fonte ingerida.
Data e hora de início	Momento inicial.
Data e hora de término	Momento final.
Responsável ou executor	Usuário, sistema ou orquestrador.
Status	Sucesso, falha, parcial, cancelado.
Volume extraído	Registros, arquivos ou bytes.
Volume ingerido	Quantidade efetivamente armazenada.
Registros rejeitados	Quantidade descartada ou inválida.
Erros	Mensagens ou códigos de falha.
Destino	Local de armazenamento bruto.

Versão gerada	Versão ou snapshot associado.

13.3 Métricas associadas

- cobertura de logs de execução;
- taxa de falhas de ingestão;
- tempo médio de ingestão;
- taxa de registros rejeitados;
- taxa de ingestão bem-sucedida.

13.4 Critérios de validação

O log é válido quando permite reconstruir o que ocorreu durante a execução, incluindo status, volume, erro, origem, destino e responsável.

13.5 Contribuição

Esse artefato fornece evidências para auditoria, observabilidade, DataOps, lineage e reprodutibilidade.

14. Artefato A2.4 — Dataset Bruto

14.1 Finalidade

O **Dataset Bruto** representa a cópia preservada dos dados ingeridos, mantida em estado original ou minimamente adaptado para armazenamento.

14.2 Estrutura mínima

Campo	Descrição
Identificador do dataset	Código único.
Fonte de origem	Fonte associada.
Data de ingestão	Momento de entrada no ambiente.
Formato	CSV, JSON, Parquet, tabela etc.
Localização	Caminho, bucket, tabela ou

	repositório.
Schema observado	Estrutura detectada.
Número de registros	Total ingerido.
Número de atributos	Total de colunas ou campos.
Checksum	Evidência de integridade.
Versão	Identificador da versão ou snapshot.
Sensibilidade	Classificação herdada ou atribuída.
Restrição de acesso	Política aplicável.

14.3 Métricas associadas

- taxa de fontes com armazenamento bruto;
- volume de dados ingeridos;
- taxa de ativos com checksum preservado;
- taxa de rastreabilidade da origem.

14.4 Critérios de validação

O dataset bruto é válido quando está armazenado, localizado, identificado, associado à fonte de origem e acompanhado de metadados mínimos.

14.5 Contribuição

Esse artefato preserva a referência original dos dados, permitindo auditoria, reprocessamento e comparação entre versões.

15. Artefato A2.5 — Registro de Snapshot

15.1 Finalidade

O **Registro de Snapshot** documenta uma versão temporal dos dados ingeridos em determinado momento.

15.2 Estrutura mínima

Campo	Descrição
ID do snapshot	Código único.
Dataset associado	Dataset bruto correspondente.
Fonte	Fonte de origem.
Data de referência	Data dos dados na fonte.
Data de captura	Data da ingestão.
Janela temporal	Período abrangido.
Versão	Identificador da versão.
Checksum	Evidência de integridade.
Volume	Registros ou bytes.
Status	Válido, parcial, rejeitado, em revisão.

15.3 Métricas associadas

- taxa de snapshots gerados;
- número de versões por fonte;
- taxa de ingestões com janela temporal documentada;
- taxa de comparação entre snapshots.

15.4 Critérios de validação

O snapshot é válido quando possui fonte, data de captura, data de referência, versão, localização, volume e evidência de integridade.

15.5 Contribuição

Esse artefato permite reconstruir estados históricos dos dados e avaliar mudanças ao longo do tempo.

16. Artefato A2.6 — Manifesto de Ingestão

16.1 Finalidade

O **Manifesto de Ingestão** registra de forma estruturada todas as informações necessárias para auditar ou reproduzir uma execução de ingestão.

16.2 Estrutura mínima

Campo	Descrição
ID do manifesto	Código único.
ID da execução	Execução associada.
Fonte	Fonte ingerida.
Configuração usada	Parâmetros e versão.
Dados de entrada	Origem, formato e localização.
Dados de saída	Destino bruto gerado.
Ambiente	Sistema, ferramenta ou ambiente de execução.
Parâmetros	Filtros, datas, credenciais de referência,

	limites.
Schema observado	Estrutura detectada.
Métricas de execução	Volume, tempo, erros, rejeições.
Checksums	Evidências de integridade.
Logs associados	Referências aos logs.
Responsável	Executor ou sistema responsável.

16.3 Métricas associadas

- cobertura de manifestos operacionais;
- taxa de execuções reprodutíveis;
- taxa de rastreabilidade da origem;
- cobertura de logs de execução.

16.4 Critérios de validação

O manifesto é válido quando permite compreender, auditar e, quando possível, reproduzir a ingestão realizada.

16.5 Contribuição

Esse artefato conecta ingestão, versionamento, reprodutibilidade, DataOps e lineage.

17. Artefato A2.7 — Relatório de Falhas e Rejeições

17.1 Finalidade

O **Relatório de Falhas e Rejeições** documenta erros, registros rejeitados, arquivos inválidos, problemas de schema, falhas de conexão, inconsistências de formato e exceções ocorridas durante a ingestão.

17.2 Estrutura mínima

Campo	Descrição

Execução	ID da execução.
Fonte	Fonte associada.
Tipo de falha	Conexão, schema, formato, integridade, acesso etc.
Registros rejeitados	Quantidade ou lista de registros.
Severidade	Baixa, média, alta ou crítica.
Causa provável	Interpretação técnica.
Impacto	Módulos ou produtos afetados.
Ação recomendada	Correção, reprocessamento, bloqueio ou revisão.
Responsável	Pessoa ou equipe.
Status	Aberta, resolvida, aceita, em análise.

17.3 Métricas associadas

- taxa de falhas de ingestão;
- taxa de registros rejeitados;
- taxa de incidentes críticos;
- tempo médio de resolução de falhas.

17.4 Critérios de validação

O relatório é válido quando documenta a falha, sua severidade, causa provável, impacto e ação recomendada.

17.5 Contribuição

Esse artefato transforma falhas técnicas em evidências gerenciáveis, apoiando observabilidade, DataOps e qualidade.

18. Artefato A2.8 — Registro de Integridade

18.1 Finalidade

O **Registro de Integridade** documenta evidências que permitem verificar se os dados ingeridos correspondem ao que foi extraído da fonte.

18.2 Estrutura mínima

Campo	Descrição
Dataset	Dataset bruto associado.
Fonte	Fonte original.
Checksum	Hash ou assinatura de integridade.
Contagem de registros	Total extraído e total ingerido.
Tamanho do arquivo	Tamanho esperado e observado.
Data de verificação	Momento da validação.
Status	Íntegro, divergente, parcial, corrompido.
Observações	Comentários técnicos.

18.3 Métricas associadas

- taxa de ativos com checksum preservado;
- taxa de divergência entre origem e destino;

- taxa de datasets íntegros;
- número de falhas de integridade.

18.4 Critérios de validação

O registro é válido quando permite verificar integridade, detectar divergências e comprovar preservação dos dados.

18.5 Contribuição

Esse artefato fortalece confiabilidade, auditoria e reprodutibilidade.

19. Artefato A2.9 — Mapa de Armazenamento Bruto

19.1 Finalidade

O **Mapa de Armazenamento Bruto** documenta onde e como os dados brutos estão armazenados.

19.2 Estrutura mínima

Campo	Descrição
Dataset bruto	Identificador do ativo.
Fonte	Fonte de origem.
Zona	Raw, landing, staging etc.
Localização	Caminho, bucket, tabela ou repositório.
Formato	CSV, JSON, Parquet, tabela etc.
Particionamento	Data, fonte, domínio, versão etc.
Versão	Versão ou snapshot.

Restrição de acesso	Política aplicável.
Política de retenção	Prazo e condição de preservação.
Responsável	Pessoa ou equipe.

19.3 Métricas associadas

- taxa de fontes com armazenamento bruto;
- taxa de ativos com localização documentada;
- taxa de ativos com política de retenção;
- taxa de ativos com restrição de acesso definida.

19.4 Critérios de validação

O mapa é válido quando permite localizar, acessar, auditar e governar os datasets brutos.

19.5 Contribuição

Esse artefato organiza o ambiente bruto e reduz perda de rastreabilidade, duplicidade e descontrole operacional.

20. Artefato A2.10 — Registro de Permissões de Ingestão

20.1 Finalidade

O **Registro de Permissões de Ingestão** documenta quem pode executar, configurar, acessar ou consultar processos e ativos de ingestão.

20.2 Estrutura mínima

Campo	Descrição
Fonte	Fonte associada.
Processo de ingestão	Pipeline ou execução.

Usuário ou sistema	Entidade autorizada.
Tipo de permissão	Leitura, execução, configuração, administração.
Justificativa	Finalidade da permissão.
Validade	Período autorizado.
Restrição	Sensibilidade ou limitação.
Responsável pela autorização	Pessoa ou setor.
Log de acesso	Referência de auditoria.

20.3 Métricas associadas

- percentual de processos com permissões documentadas;
- taxa de acessos autorizados;
- taxa de acessos indevidos;
- taxa de permissões vencidas.

20.4 Critérios de validação

O registro é válido quando permissões, responsáveis, finalidades e restrições estão documentados.

20.5 Contribuição

Esse artefato conecta ingestão à governança, segurança, privacidade e conformidade.

21. Artefato A2.11 — Registro de Janelas Temporais de Ingestão

21.1 Finalidade

O **Registro de Janelas Temporais de Ingestão** documenta o período de dados capturado em cada execução, permitindo controle de dados históricos, incrementais, longitudinais e temporais.

21.2 Estrutura mínima

Campo	Descrição
Execução	ID da ingestão.
Fonte	Fonte associada.
Tipo de janela	Completa, incremental, evento, período fixo etc.
Data inicial	Início do período capturado.
Data final	Fim do período capturado.
Campo temporal usado	Atributo de referência temporal.
Critério de seleção	Regra de filtro temporal.
Registros capturados	Quantidade.
Lacunas temporais	Períodos ausentes ou incompletos.
Observações	Limitações temporais.

21.3 Métricas associadas

- taxa de ingestões com janela temporal documentada;
- latência de ingestão;
- taxa de atualização no prazo;
- número de lacunas temporais detectadas.

21.4 Critérios de validação

O registro é válido quando permite saber exatamente qual período foi capturado e quais lacunas temporais existem.

21.5 Contribuição

Esse artefato reforça a capacidade do UnespDataLens-RM de lidar com dados temporais, séries históricas, logs, eventos e trajetórias.

22. Artefato A2.12 — Registro de Eventos ou Streaming

22.1 Finalidade

O **Registro de Eventos ou Streaming** documenta fluxos de dados contínuos ou orientados a eventos, quando o projeto exigir ingestão near real-time ou streaming.

22.2 Estrutura mínima

Campo	Descrição
Fluxo ou tópico	Identificador do stream.
Fonte geradora	Sistema, sensor ou aplicação.
Tipo de evento	Categoria do evento capturado.
Timestamp	Data e hora do evento.
Latência	Tempo entre evento e ingestão.
Volume de eventos	Quantidade por período.
Status	Ativo, interrompido, parcial, falho.
Erros	Falhas de captura ou processamento.

Destino bruto	Local de armazenamento do evento.
Política de retenção	Prazo de preservação do fluxo.

22.3 Métricas associadas

- taxa de eventos processados;
- latência de ingestão;
- taxa de perda de eventos;
- volume de eventos por período;
- taxa de disponibilidade do fluxo.

22.4 Critérios de validação

O registro é válido quando permite acompanhar, auditar e monitorar fluxos contínuos de eventos.

22.5 Contribuição

Esse artefato explicita a capacidade do modelo de referência de tratar dados dinâmicos, eventos e atualizações em tempo quase real.

23. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Fontes documentadas, APIs, bancos, arquivos, planilhas, logs, eventos, conectores, credenciais autorizadas, parâmetros, schemas esperados, contratos e metadados do Módulo 1.
Processamento	Planejamento da extração, configuração da ingestão, execução, validação preliminar, armazenamento bruto, geração de logs, criação de snapshots, cálculo de checksums, registro de falhas e geração de manifestos.
Artefatos	Plano de extração, configuração de ingestão, logs, datasets brutos, snapshots, manifestos, relatórios de falhas, registros de integridade, mapa de armazenamento, permissões, janelas temporais e registros de eventos.

Métricas	TIBS, VDI, TRR, TFI, TMI, TCS, TAP, CLE, CMO, TSG, TFAB, TRO, TACP, TIIC, LI e TEP.
Saídas	Dados brutos preservados, versionados, rastreáveis, auditáveis e preparados para integração, transformação, qualidade, governança, lineage, versionamento e disponibilização analítica.

24. Critérios gerais de validação do módulo

O Módulo 2 pode ser considerado adequadamente implementado quando:

1. as fontes prioritárias identificadas no Módulo 1 possuem estratégia de extração definida;
2. os processos de ingestão possuem configuração documentada;
3. cada execução gera log estruturado;
4. os dados ingeridos são preservados em estado bruto;
5. os datasets brutos possuem localização, versão, formato e fonte de origem documentados;
6. snapshots são gerados para fontes críticas ou periódicas;
7. manifestos de ingestão são produzidos para execuções relevantes;
8. checksums ou evidências equivalentes de integridade são calculados;
9. falhas, rejeições e exceções são registradas;
10. janelas temporais de extração são documentadas quando aplicável;
11. ingestões incrementais evitam perda ou duplicação de registros;
12. ingestões near real-time ou streaming possuem controle de latência e eventos processados;
13. permissões de acesso aos dados brutos estão documentadas;
14. ativos sensíveis possuem restrições compatíveis com sua classificação;
15. os dados brutos podem ser auditados e reprocessados;
16. os artefatos produzidos são consumíveis pelos módulos posteriores.

25. Relação com os demais módulos do modelo de referência

O Módulo 2 conecta o inventário das fontes ao restante do pipeline analítico.

Com o **Módulo 1 — Fontes e Inventário de Dados**, utiliza informações sobre origem, formato, periodicidade, granularidade, restrição de acesso, sensibilidade e limitações das fontes.

Com o **Módulo 3 — Integração e Harmonização**, fornece datasets brutos, schemas observados, chaves preliminares, janelas temporais e registros de origem.

Com o **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, fornece os dados brutos que serão tratados, padronizados, recodificados, filtrados ou enriquecidos.

Com o **Módulo 5 — Qualidade e Validação de Dados**, fornece dados e metadados necessários para avaliação de completude, validade, consistência, atualidade, schema e integridade.

Com o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, preserva os dados originais para comparação de perdas, distribuição de grupos, representatividade e vieses introduzidos em etapas posteriores.

Com o **Módulo 7 — Armazenamento Analítico**, fornece insumos brutos que posteriormente serão organizados em zonas tratadas, integradas, curadas e de consumo.

Com o **Módulo 8 — Disponibilização e Consumo Analítico**, fornece evidências de origem e atualização dos produtos publicados.

Com o **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, produz metadados operacionais, schemas observados, formatos, logs, snapshots e registros de ingestão.

Com o **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, fornece registros de acesso, restrições, permissões, sensibilidade e evidências de uso autorizado.

Com o **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, fornece os primeiros eventos materiais de linhagem, conectando fontes originais a datasets brutos.

Com o **Módulo 12 — Versionamento e Reprodutibilidade**, fornece snapshots, manifestos, logs, checksums, configurações e versões necessárias para reconstrução de execuções.

Com o **Módulo 13 — Representações Analíticas**, preserva atributos temporais, eventos, entidades e relações que poderão ser usados em representações tabulares, temporais, espaciais, textuais, semânticas e em grafos.

Com o **Módulo 14 — Feature Engineering e Feature Store**, fornece dados brutos e históricos que permitem construir features reprodutíveis.

Com o **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, fornece logs, volumes, schemas, janelas temporais e dados de referência para detectar drift, atrasos, falhas e mudanças estruturais.

Com o **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, fornece metadados operacionais, logs e manifestos que podem ser consultados para explicações, recomendações e diagnóstico de problemas.

26. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 2 pode ser utilizado para extrair e ingerir dados de sistemas acadêmicos, ambientes virtuais de aprendizagem, registros de matrícula, histórico escolar, notas, frequência, questionários socioeconômicos, registros de evasão e logs de interação.

Por exemplo, uma ingestão mensal pode capturar dados de matrícula e situação acadêmica em lote, enquanto logs de acesso ao ambiente virtual podem ser capturados diariamente ou em tempo quase real. Cada ingestão deve registrar fonte, data, período letivo, volume, schema observado, registros rejeitados, dados ausentes, checksum, snapshot e destino bruto.

Se uma base de interações do ambiente virtual apresentar queda abrupta no volume de eventos, o log de ingestão e o registro de janelas temporais permitirão identificar se houve falha na fonte, interrupção na API, mudança de formato ou ausência real de eventos.

Essa estrutura permite que o UnespDataLens-RM trabalhe com dados tabulares, temporais, longitudinais, relacionais e de interação com maior confiabilidade e rastreabilidade.

27. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, o Módulo 2 pode capturar dados provenientes de secretarias municipais, bases públicas, sistemas de atendimento, cadastros territoriais, registros de saúde, educação, assistência social, habitação, trabalho e orçamento.

Algumas fontes podem ser atualizadas mensalmente, outras por arquivo enviado manualmente, outras por API e outras por fluxo de eventos. O módulo deve registrar a periodicidade de cada uma, o momento da ingestão, a janela temporal capturada, os dados ausentes, os registros rejeitados e as restrições de acesso.

Por exemplo, se uma secretaria envia uma base mensal de atendimentos sociais, o Módulo 2 deve preservar o arquivo original, registrar checksum, data de recebimento, mês de referência, volume de registros, campos disponíveis, responsável e eventuais problemas de schema. Caso a base não seja entregue no prazo, a falha deve ser registrada como incidente operacional e sinalizada aos módulos de observabilidade e monitoramento.

28. Contribuição técnico-científica do módulo

A contribuição técnico-científica do Módulo 2 está em formalizar a extração, ingestão e preservação dos dados brutos como uma etapa mensurável, auditável, rastreável e reprodutível da Engenharia Analítica de Dados.

Em vez de tratar a ingestão como uma operação técnica isolada, o UnespDataLens-RM propõe que cada processo de entrada de dados produza artefatos formais, como planos de ingestão, configurações versionáveis, logs, manifestos, snapshots, registros de integridade, relatórios de falhas, registros de janelas temporais e evidências de permissões.

Essa formalização contribui para:

- reduzir perda de rastreabilidade na origem dos dados;
- preservar dados brutos para auditoria e reprocessamento;
- aumentar a reprodutibilidade dos pipelines;
- documentar parâmetros e versões de ingestão;
- identificar falhas de extração de forma estruturada;
- apoiar governança e controle de acesso;
- fornecer evidências para lineage;
- preparar a base para qualidade, integração e transformação;
- permitir comparação entre extrações realizadas em diferentes momentos;
- contemplar dados históricos, temporais, incrementais, eventos e fluxos quase em tempo real;
- fortalecer DataOps, observabilidade e monitoramento desde a entrada dos dados.

29. Síntese do módulo

O **Módulo 2 — Extração, Ingestão e Armazenamento Bruto** operacionaliza a entrada dos dados no ambiente analítico do UnespDataLens-RM. Ele transforma fontes documentadas em datasets brutos armazenados, acompanhados de logs, metadados operacionais, manifestos, registros de integridade, snapshots, janelas temporais, registros de falhas e permissões de acesso.

Sua importância para o modelo de referência é central, pois estabelece a primeira materialização dos dados no pipeline analítico. Ao preservar a origem, registrar a execução e manter dados brutos auditáveis, o módulo fortalece a governança, a proveniência, a reprodutibilidade, o versionamento, a observabilidade e a confiabilidade dos processos analíticos.

Dessa forma, o Módulo 2 complementa o Módulo 1 e prepara as condições necessárias para integração, transformação, qualidade, governança, lineage, versionamento, monitoramento, representação temporal, feature engineering e disponibilização analítica dos dados.

Módulo 3 — Integração e Harmonização

1. Finalidade do módulo

O **Módulo 3 — Integração e Harmonização** tem como finalidade combinar, relacionar, compatibilizar e organizar dados provenientes de múltiplas fontes, previamente identificadas no **Módulo 1 — Fontes e Inventário de Dados** e ingeridas no **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**.

Esse módulo é responsável por transformar conjuntos de dados brutos, dispersos e heterogêneos em estruturas integradas e semanticamente coerentes, capazes de sustentar análises, indicadores, modelos preditivos, representações em grafos, dashboards, estudos longitudinais e processos de tomada de decisão baseada em evidências.

Em projetos de Engenharia Analítica de Dados, a integração de dados é uma das etapas mais críticas, pois as fontes frequentemente possuem diferenças de formato, nomenclatura, granularidade, periodicidade, chaves, categorias, codificações, escalas, domínios de valores e regras de negócio. Além disso, a mesma entidade pode aparecer em fontes distintas com identificadores diferentes, registros duplicados, atributos incompletos, variações de nomes, inconsistências temporais ou ausência de chaves explícitas.

No contexto do **UnespDataLens-RM**, o Módulo 3 tem papel estratégico porque formaliza a integração e a harmonização como processos documentados, avaliáveis e rastreáveis. Não se trata apenas de “juntar bases”, mas de definir critérios, regras, chaves, correspondências, compatibilizações semânticas, níveis de confiança, tratamentos de conflitos e evidências de integração.

Assim, esse módulo contribui para reduzir ambiguidades, evitar junções incorretas, explicitar decisões de harmonização, preservar a rastreabilidade das relações entre fontes e preparar os dados para os módulos posteriores de transformação, qualidade, governança, lineage, versionamento, representação analítica e disponibilização.

2. Objetivo geral

O objetivo geral do Módulo 3 é **integrar e harmonizar dados provenientes de múltiplas fontes**, produzindo estruturas analíticas consistentes, documentadas e rastreáveis, por meio da identificação de chaves, entidades compartilhadas, correspondências semânticas, regras de compatibilização, tratamento de conflitos e avaliação da qualidade das relações estabelecidas entre bases.

3. Objetivos específicos

São objetivos específicos do Módulo 3:

1. identificar fontes que podem ser integradas a partir de entidades, chaves, períodos, atributos ou conceitos compartilhados;
2. definir estratégias de integração adequadas ao tipo de dado, granularidade, periodicidade, estrutura e finalidade analítica;

3. estabelecer regras de correspondência entre registros, atributos, entidades e categorias de diferentes fontes;
4. harmonizar nomes de atributos, formatos, domínios de valores, códigos, categorias, unidades de medida e recortes temporais;
5. resolver ou registrar conflitos entre fontes, como divergência de valores, duplicidade, inconsistência de chaves ou diferença de granularidade;
6. aplicar técnicas de record linkage, entity resolution, deduplicação, normalização semântica e compatibilização temporal;
7. produzir datasets integrados, matrizes de correspondência, tabelas de mapeamento, regras de harmonização e registros de conflitos;
8. avaliar a qualidade da integração por meio de métricas de pareamento, cobertura, conflito, confiança, completude e consistência;
9. fornecer insumos para os módulos posteriores de transformação, qualidade, metadados, governança, lineage, versionamento, grafos e disponibilização analítica.

4. Fundamentação conceitual do módulo

O Módulo 3 é fundamentado na compreensão de que dados analíticos raramente estão prontos em uma única fonte. Em geral, respostas analíticas relevantes exigem combinar diferentes bases, oriundas de sistemas, setores, períodos e finalidades distintas. Essa combinação, entretanto, exige mais do que operações técnicas de junção. Ela demanda compreensão estrutural, semântica, temporal e institucional das fontes.

A integração de dados pode ser entendida como o processo de combinar registros, atributos ou entidades de diferentes fontes para produzir uma visão analítica mais completa. A harmonização, por sua vez, corresponde ao processo de compatibilizar formatos, conceitos, códigos, categorias, granularidades, escalas e significados, permitindo que os dados integrados sejam interpretados de forma coerente.

No UnespDataLens-RM, integração e harmonização são tratadas como atividades complementares. A integração estabelece vínculos entre fontes; a harmonização torna esses vínculos semanticamente e estruturalmente utilizáveis. Por exemplo, duas bases podem conter informações sobre estudantes, mas uma registra “situação acadêmica” como texto e outra como código numérico. A integração pode relacionar os registros por identificador; a harmonização precisa compatibilizar os significados dos valores.

Esse módulo enfrenta problemas frequentes, como:

- ausência de chaves únicas entre fontes;
- chaves inconsistentes ou incompletas;
- registros duplicados;
- nomes de atributos divergentes;
- categorias com significados diferentes;
- formatos temporais incompatíveis;
- granularidades distintas;
- sobreposição parcial de períodos;
- conflitos entre valores provenientes de diferentes sistemas;
- ausência de documentação sobre regras de integração;
- dificuldade de rastrear como um dataset integrado foi construído.

À luz da **Design Science Research**, os artefatos do Módulo 3 são tratados como modelos, métodos e instanciações voltados a formalizar a integração de dados como um processo técnico-metodológico avaliável. Cada artefato deve possuir finalidade, estrutura, regras de

construção, métricas, critérios de validação e relação com os demais módulos do modelo de referência.

5. Formalização sintética do módulo

O Módulo 3 pode ser compreendido como um processo de transformação de datasets brutos em estruturas integradas e harmonizadas.

Considerando o conjunto de datasets brutos produzido no Módulo 2:

$$\mathbf{B = \{b_1, b_2, ..., b_n\}}$$

o Módulo 3 identifica relações de integração entre esses datasets, produzindo um conjunto de vínculos:

$$\mathbf{L = \{l_1, l_2, ..., l_k\}}$$

em que cada vínculo $\mathbf{l_i}$ representa uma correspondência entre fontes, registros, atributos, entidades ou categorias.

Cada vínculo pode ser descrito por:

$$\mathbf{l_i = \{fonte_A, fonte_B, entidade, chave, tipo_relação, regra_correspondência, granularidade, período, grau_confiança, status\}}$$

O resultado do módulo é um conjunto de datasets integrados e harmonizados:

$$\mathbf{H = \{h_1, h_2, ..., h_m\}}$$

em que cada $\mathbf{h_j}$ representa um ativo analítico resultante da combinação controlada de duas ou mais fontes.

Cada dataset harmonizado $\mathbf{h_j}$ deve ser associado a metadados de integração:

$$\mathbf{h_j = \{id, fontes_origem, regras_integração, regras_harmonização, chaves_utilizadas, conflitos_resolvidos, conflitos_pendentes, versão, métricas_integração, lineage\}}$$

Dessa forma, o processo de integração deixa de ser uma operação implícita de junção e passa a ser tratado como uma construção documentada, mensurável, auditável e reprodutível.

6. Entradas do módulo

As entradas do Módulo 3 incluem dados brutos, metadados, artefatos de inventário e informações de relacionamento produzidas nos módulos anteriores.

As principais entradas são:

- datasets brutos gerados no Módulo 2;
- inventário de fontes de dados;
- mapa de bases;
- dicionário preliminar de dados;
- matriz de relacionamento entre fontes;
- registro de granularidade;

- classificação de sensibilidade;
- catálogo preliminar de fontes;
- logs de ingestão;
- manifestos de ingestão;
- registros de snapshots;
- documentação de chaves primárias e estrangeiras;
- domínios de valores;
- códigos institucionais;
- regras de negócio;
- tabelas auxiliares de mapeamento;
- vocabulários controlados;
- glossários preliminares;
- informações sobre cobertura temporal;
- políticas de acesso e restrições de uso.

As fontes podem apresentar diferentes formatos, como tabelas relacionais, arquivos CSV, JSON, planilhas, logs, APIs, dados temporais, bases territoriais, registros administrativos, bases educacionais, bases socioeconômicas e dados de políticas públicas.

7. Processamento

O processamento do Módulo 3 envolve atividades de preparação, identificação de correspondências, definição de regras de integração, harmonização semântica, compatibilização temporal, resolução de conflitos, geração de datasets integrados e documentação das decisões tomadas.

7.1 Seleção das fontes a integrar

A primeira atividade consiste em selecionar as fontes que serão integradas, com base no objetivo analítico, na matriz de relacionamento entre fontes, na disponibilidade de chaves, na compatibilidade de granularidade, na cobertura temporal e na relevância dos atributos.

Nem todas as fontes inventariadas ou ingeridas precisam ser integradas em uma mesma etapa. A seleção deve considerar finalidade, viabilidade técnica, qualidade preliminar e riscos de uso.

7.2 Identificação de entidades compartilhadas

A segunda atividade consiste em identificar entidades comuns entre fontes. Essas entidades podem ser pessoas, estudantes, cursos, disciplinas, municípios, territórios, famílias, empresas, trabalhadores, eventos, atendimentos, unidades administrativas ou qualquer outra unidade relevante.

A identificação de entidades compartilhadas permite estabelecer vínculos entre registros e orientar a definição das chaves de integração.

7.3 Identificação e validação de chaves

A terceira atividade consiste em identificar chaves primárias, chaves estrangeiras, identificadores diretos, identificadores compostos ou critérios indiretos de associação.

Quando não houver chaves explícitas, podem ser utilizadas estratégias de record linkage ou entity resolution, considerando combinações de atributos, similaridade textual, datas, localização, categorias ou outros critérios disponíveis.

As chaves identificadas devem ser documentadas e classificadas quanto ao grau de confiança.

7.4 Compatibilização de granularidade

A quarta atividade consiste em avaliar se as fontes possuem o mesmo nível de detalhe. Quando as granularidades forem distintas, o módulo deve definir estratégias de agregação, desagregação, replicação controlada ou restrição de integração.

Por exemplo, uma base no nível de estudante não deve ser integrada diretamente a uma base no nível de curso sem regras explícitas de agregação ou associação. Da mesma forma, dados por município não devem ser combinados com dados por indivíduo sem controle do nível analítico.

7.5 Harmonização de atributos e formatos

A quinta atividade consiste em padronizar nomes de atributos, tipos de dados, formatos de datas, codificações, unidades de medida, escalas e domínios de valores.

Essa atividade pode incluir conversão de tipos, normalização de textos, padronização de categorias, compatibilização de códigos, tratamento de encoding, padronização de datas e alinhamento de unidades.

7.6 Harmonização semântica

A sexta atividade consiste em verificar se atributos, categorias e conceitos possuem o mesmo significado entre fontes distintas. Essa etapa é essencial para evitar integrações tecnicamente corretas, mas semanticamente equivocadas.

A harmonização semântica pode envolver consulta a especialistas, análise de regras de negócio, glossários, vocabulários controlados, modelos semânticos e tabelas de equivalência.

7.7 Compatibilização temporal

A sétima atividade consiste em alinhar períodos, datas, janelas temporais, versões e frequências de atualização entre fontes. Essa etapa é necessária quando as bases possuem recortes temporais diferentes ou quando a análise depende de trajetórias e eventos ao longo do tempo.

A compatibilização temporal pode envolver agregação por mês, semestre ou ano, criação de janelas móveis, definição de períodos comuns, alinhamento por data de referência ou criação de snapshots integrados.

7.8 Resolução e registro de conflitos

A oitava atividade consiste em identificar e tratar conflitos entre fontes. Conflitos podem ocorrer quando duas bases apresentam valores diferentes para o mesmo atributo, registros duplicados, categorias divergentes, datas inconsistentes ou chaves contraditórias.

Os conflitos podem ser resolvidos por regras de precedência, validação com especialista, escolha da fonte mais confiável, agregação, votação, marcação de incerteza ou manutenção do conflito para análise posterior.

Independentemente da solução adotada, os conflitos devem ser registrados.

7.9 Geração de datasets integrados e harmonizados

Após a definição das regras de integração e harmonização, o módulo gera datasets integrados, preservando informações sobre fontes de origem, chaves utilizadas, regras aplicadas, conflitos identificados, conflitos resolvidos, versões e métricas de integração.

Esses datasets servirão como insumos para os módulos posteriores de transformação, qualidade, feature engineering, representação em grafos e disponibilização analítica.

7.10 Documentação da integração

Toda integração deve gerar artefatos documentais e técnicos que permitam reproduzir, auditar e compreender o processo. Devem ser registrados regras, chaves, parâmetros, versões, decisões, exceções, conflitos, métricas e responsáveis.

7.11 Preparação Especializada por Natureza do Dado

Além das operações gerais de transformação, limpeza e enriquecimento, o Módulo 4 organiza a preparação dos dados em subpipelines especializados conforme a natureza dos dados. Essa organização parte do pressuposto de que diferentes tipos de dados apresentam estruturas, granularidades, temporalidades, formatos, ruídos, dependências, requisitos de qualidade e formas de representação distintas. Assim, a preparação não deve ser compreendida como um conjunto único e homogêneo de procedimentos, mas como uma etapa metodologicamente orientada pelo tipo de dado e pela finalidade analítica pretendida.

No UnespDataLens-RM, essa preparação especializada é uma capacidade interna do Módulo 4 e deve ser articulada à dimensão complementar Preparação Especializada por Natureza do Dado, descrita na seção de dimensões complementares do modelo de referência. O Módulo 4 executa e documenta as operações de preparação; a dimensão complementar, por sua vez, estabelece os princípios, critérios, artefatos e métricas que orientam a especialização desses subpipelines ao longo do ecossistema analítico.

Cada subpipeline deve explicitar requisitos de entrada, operações de transformação, técnicas recomendadas, métricas de qualidade, artefatos produzidos, critérios de validação, registros de transformação, mecanismos de versionamento, relações de lineage e limitações de uso. Dessa forma, o modelo evita que dados temporais, espaciais, textuais, semânticos, em grafos, multimodais ou em fluxo contínuo sejam tratados apenas como tabelas convencionais, preservando suas características próprias e aumentando a confiabilidade dos ativos analíticos produzidos.

7.11.1 Subpipelines especializados previstos

O Módulo 4 contempla, de forma extensível, os seguintes subpipelines especializados de preparação: Preparação de Dados Estruturados; Preparação de Dados Temporais; Preparação de Dados Espaciais; Preparação de Dados Textuais e Semânticos; Preparação

de Dados em Grafos; Preparação de Dados Multimodais; e Preparação de Dados em Fluxo ou Streaming.
A Preparação de Dados Estruturados contempla registros organizados em tabelas, atributos, chaves, domínios e regras de integridade. A Preparação de Dados Temporais contempla séries históricas, eventos, logs, trajetórias, sequências e dados longitudinais. A Preparação de Dados Espaciais contempla coordenadas, territórios, regiões, camadas geográficas e relações espaciais. A Preparação de Dados em Grafos contempla nós, arestas, tipos de relação, direção, pesos e propriedades. A Preparação de Dados Multimodais contempla dados compostos por texto, imagem, áudio, vídeo, sensores, tabelas e eventos. A Preparação de Dados em Fluxo ou Streaming contempla dados produzidos continuamente ou com baixa latência.
Essa organização não cria módulos independentes para cada técnica analítica. Seu objetivo é reconhecer a diversidade metodológica da preparação dos dados e permitir que cada natureza de dado seja tratada por operações adequadas, mantendo a coerência do modelo de referência e sua reutilização em diferentes domínios de aplicação.

7.11.2 Preparação de Dados Textuais e Semânticos

A Preparação de Dados Textuais e Semânticos contempla os processos necessários para transformar textos livres, respostas abertas, comentários, avaliações, fóruns, mensagens, documentos, notícias, e-mails, chats, transcrições e demais registros linguísticos em ativos analíticos adequados para mineração de textos, Processamento de Linguagem Natural, análise semântica, uso de modelos de linguagem e tarefas analíticas posteriores.
Esse subpipeline é necessário porque dados textuais possuem características distintas dos dados estruturados tradicionais. Textos podem conter ambiguidade, ironia, negação, abreviações, erros ortográficos, variações linguísticas, emojis, URLs, ruídos, dados pessoais, termos de domínio, múltiplos idiomas, diferentes níveis de formalidade e forte dependência contextual. Por isso, sua preparação exige operações específicas de limpeza, normalização, enriquecimento, representação, validação, documentação e rastreabilidade.

7.11.3 Entradas do subpipeline textual e semântico

As entradas desse subpipeline podem incluir respostas abertas de questionários, comentários, avaliações institucionais, fóruns, mensagens de ambientes virtuais, redes sociais, e-mails, chats, chamados, documentos, relatórios, notícias, transcrições, atas, pareceres e registros qualitativos. Também podem ser utilizados metadados associados aos textos, tais como data, origem, autor ou perfil anonimizado, turma, curso, unidade, território, categoria, canal, idioma, sistema de origem, contexto de coleta e restrições de uso.
Antes da preparação, é necessário preservar o corpus bruto, registrar sua origem, documentar a forma de coleta, classificar restrições de acesso, identificar sensibilidade e registrar a finalidade analítica permitida. Essa preservação permite auditoria, reprocessamento, comparação de versões e reconstrução do percurso dos dados textuais ao longo do pipeline.

7.11.4 Processos do subpipeline textual e semântico

O processamento textual envolve um conjunto articulado de operações. Na etapa de organização e coleta, devem ser realizadas a unificação das fontes textuais, a consolidação

de registros, a remoção ou marcação de duplicidades, a preservação do texto original, a associação com metadados contextuais e a classificação de sensibilidade.
Na etapa de limpeza, podem ser aplicadas operações como remoção de HTML, URLs, caracteres especiais, ruídos de codificação, espaços excedentes, marcações técnicas, assinaturas automáticas, trechos irrelevantes e elementos de formatação. Emojis, emoticons, hashtags e menções devem ser removidos, normalizados ou preservados conforme a finalidade analítica, pois podem carregar informação semântica relevante em determinados contextos.
Na etapa de normalização, podem ser aplicadas padronização de caixa, tratamento de acentuação, correção de encoding, tokenização, segmentação de sentenças, stemming, lematização, tratamento de stopwords, preservação de negações, identificação de intensificadores, padronização de abreviações e normalização de termos de domínio. Essas decisões devem ser documentadas, pois impactam diretamente as representações e os resultados analíticos posteriores.
Na etapa de enriquecimento, podem ser incorporadas informações como idioma, detecção automática de linguagem, correção ortográfica, expansão de abreviações, dicionários de domínio, léxicos especializados, entidades nomeadas, categorias semânticas, tópicos preliminares e descritores contextuais. O enriquecimento deve preservar rastreabilidade entre texto original, texto tratado, recursos externos utilizados e atributos derivados.
Na etapa de representação, os textos podem ser convertidos em vocabulários, matrizes termo-documento, Bag of Words, TF-IDF, n-grams, Word2Vec, FastText, Doc2Vec, Sentence Transformers, embeddings de modelos de linguagem, representações semânticas, vetores contextuais ou estruturas híbridas texto-tabular. A escolha da representação deve ser compatível com o objetivo analítico, com o volume de dados, com o idioma, com a disponibilidade de rótulos e com os requisitos de interpretabilidade, desempenho e reprodutibilidade.

7.11.5 Qualidade, validação e governança textual

A preparação textual deve incluir avaliação de qualidade específica. Entre os aspectos a serem observados estão completude dos textos, taxa de textos vazios, taxa de duplicidade, taxa de ruído, cobertura lexical, diversidade textual, tamanho médio dos documentos, distribuição de idiomas, presença de dados pessoais, consistência dos metadados, taxa de anonimização, qualidade dos rótulos quando existirem e adequação da representação textual à finalidade analítica.
A governança textual é especialmente relevante porque textos livres podem conter informações pessoais, sensíveis ou identificáveis não previstas na estrutura original da base. Por isso, devem ser previstos mecanismos de anonimização, pseudonimização, minimização, controle de acesso, registro de finalidade, documentação de restrições e validação ética ou institucional quando aplicável.

7.11.6 Saídas e artefatos do subpipeline textual

As saídas incluem corpus bruto preservado, corpus limpo, corpus anonimizado, corpus normalizado, vocabulário, dicionários de domínio, lista de termos tratados, matriz termo-documento, matriz TF-IDF, embeddings persistidos, metadados textuais, relatório de qualidade textual, relatório de anonimização, pipeline NLP documentado, regras de pré-processamento, registros de parâmetros, manifesto de execução, registro de versionamento e lineage textual.

Esses artefatos devem permitir reconstruir o percurso texto original → corpus bruto → corpus tratado → representação textual → atributo derivado → modelo, indicador, dashboard, relatório ou API. Essa rastreabilidade é fundamental para auditoria, reprodutibilidade, explicação de resultados e avaliação de impacto das decisões de preparação.

7.11.7 Delimitação em relação à análise de sentimentos

A análise de sentimentos não é definida como etapa de preparação no UnespDataLens-RM. Ela é uma tarefa analítica posterior, situada no contexto de mineração de textos, modelagem analítica, Inteligência Artificial ou consumo analítico. A preparação textual oferece as condições para que tarefas como análise de sentimentos, classificação de emoções, detecção de tópicos, sumarização, detecção de toxicidade, classificação textual e extração de entidades sejam executadas sobre dados previamente preparados, documentados, governados, rastreáveis, versionados e reprodutíveis.
Essa distinção preserva o rigor metodológico do modelo de referência. O UnespDataLens-RM não propõe um módulo específico para análise de sentimentos, pois isso tornaria o modelo dependente de uma técnica ou tarefa particular. Sua contribuição está em formalizar a preparação de dados textuais e semânticos como parte da Engenharia Analítica de Dados, permitindo que diferentes tarefas analíticas sejam posteriormente instanciadas com maior confiabilidade.

8. Técnicas, métodos e algoritmos aplicáveis

O Módulo 3 pode empregar diferentes técnicas, métodos e algoritmos para integração e harmonização.

Entre as técnicas de integração, destacam-se:

- junção por chave primária;
- junção por chave estrangeira;
- integração por chave composta;
- integração por período;
- integração por território;
- integração por entidade;
- integração por evento;
- integração baseada em regras;
- integração incremental;
- integração por snapshots;
- integração federada;
- record linkage;
- entity resolution;
- deduplicação.

Entre as técnicas de harmonização, podem ser utilizadas:

- padronização de nomes de atributos;
- normalização de textos;
- compatibilização de domínios;
- conversão de tipos;
- padronização de datas;

- compatibilização de unidades de medida;
- mapeamento de categorias;
- recodificação de valores;
- alinhamento de granularidade;
- alinhamento temporal;
- construção de tabelas de equivalência;
- construção de vocabulários controlados.

Entre os algoritmos e métodos computacionais aplicáveis, destacam-se:

- comparação exata de chaves;
- distância de Levenshtein;
- similaridade Jaro-Winkler;
- similaridade por cosseno;
- TF-IDF para comparação textual;
- embeddings para similaridade semântica;
- blocking para redução do espaço de comparação;
- regras heurísticas;
- linkage determinístico;
- linkage probabilístico;
- Fellegi-Sunter;
- clustering de registros semelhantes;
- detecção de duplicidade;
- resolução de conflitos por regras de precedência;
- análise de consistência referencial;
- análise de sobreposição temporal;
- validação de cardinalidade.

Ferramentas possíveis incluem SQL, Python, pandas, PySpark, DuckDB, dbt, Apache Spark, ferramentas de ETL/ELT, bibliotecas de record linkage, bancos relacionais, data lakes, lakehouses e catálogos de dados.

9. Métricas do módulo

As métricas do Módulo 3 têm como objetivo avaliar a cobertura, a qualidade, a confiabilidade e a rastreabilidade da integração e da harmonização.

Métrica	Sigla	Descrição
Número de fontes integradas	NFI	Quantidade de fontes combinadas em um processo de integração
Taxa de pareamento	TP	Proporção de registros pareados entre fontes
Taxa de não pareamento	TNP	Proporção de registros sem correspondência

Taxa de chaves válidas	TCV	Proporção de registros com chaves válidas para integração
Taxa de duplicidade entre fontes	TDEF	Proporção de registros duplicados ou potencialmente duplicados
Taxa de conflito entre fontes	TCF	Proporção de registros com valores conflitantes
Taxa de conflitos resolvidos	TCR	Proporção de conflitos tratados por regras definidas
Cobertura temporal integrada	CTI	Período comum efetivamente coberto pelas fontes integradas
Compatibilidade de granularidade	CG	Grau de compatibilidade entre níveis de análise das fontes
Taxa de harmonização de atributos	THA	Proporção de atributos harmonizados em relação ao total necessário
Taxa de harmonização semântica	THS	Proporção de conceitos ou categorias semanticamente compatibilizados
Grau de confiança da integração	GCI	Avaliação do nível de confiabilidade das relações estabelecidas
Taxa de rastreabilidade da integração	TRI	Proporção de relações integradas com regras e origem documentadas
Taxa de relações auditáveis	TRA	Proporção de integrações que podem ser reconstituídas e verificadas

9.1 Taxa de Pareamento — TP

A **Taxa de Pareamento** mede a proporção de registros de uma fonte que encontraram correspondência em outra fonte.

TP = número de registros pareados / número total de registros candidatos ao pareamento

Essa métrica permite avaliar a efetividade do processo de relacionamento entre fontes.

9.2 Taxa de Não Pareamento — TNP

A **Taxa de Não Pareamento** mede a proporção de registros que não encontraram correspondência em outra fonte.

TNP = número de registros não pareados / número total de registros candidatos ao pareamento

Valores elevados podem indicar ausência de chaves, baixa qualidade de dados, diferenças de cobertura, incompatibilidade temporal ou granularidade inadequada.

9.3 Taxa de Conflito entre Fontes — TCF

A **Taxa de Conflito entre Fontes** mede a proporção de registros integrados que apresentam divergência de valores em atributos equivalentes.

TCF = número de registros com conflito / número total de registros integrados avaliados

Essa métrica permite identificar inconsistências entre sistemas e orientar regras de resolução.

9.4 Taxa de Conflitos Resolvidos — TCR

A **Taxa de Conflitos Resolvidos** mede a proporção de conflitos tratados por regras explícitas de harmonização.

TCR = número de conflitos resolvidos / número total de conflitos identificados

Essa métrica indica a capacidade do módulo de transformar conflitos em decisões documentadas.

9.5 Taxa de Rastreabilidade da Integração — TRI

A **Taxa de Rastreabilidade da Integração** avalia a proporção de relações integradas que possuem origem, regra, chave e justificativa documentadas.

TRI = número de relações integradas documentadas / número total de relações integradas

Essa métrica é essencial para governança, lineage e reprodutibilidade.

10. Artefatos formalizados do módulo

Os artefatos do Módulo 3 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, regras de construção, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A3.1	Plano de Integração	Documento técnico	TRI	Define estratégia, fontes, chaves e regras
A3.2	Matriz de Correspondência entre Fontes	Matriz fonte × fonte	TP, TNP	Registra relações entre fontes
A3.3	Tabela de Mapeamento de Atributos	Tabela estruturada	THA	Harmoniza atributos equivalentes
A3.4	Tabela de Equivalência Semântica	Tabela ou vocabulário	THS	Harmoniza conceitos, categorias e códigos
A3.5	Registro de Chaves e Entidades	Tabela técnica	TCV	Documenta chaves e entidades compartilhadas
A3.6	Registro de Compatibilidade de Granularidade	Tabela analítica	CG	Evita integrações inadequadas
A3.7	Registro de Conflitos de Integração	Relatório técnico	TCF, TCR	Documenta e trata conflitos
A3.8	Dataset Integrado e Harmonizado	Tabela, arquivo ou dataset	TP, TRI	Materializa a integração
A3.9	Manifesto de	Documento	TRI	Registra regras,

	Integração	estruturado		parâmetros e versões
A3.10	Relatório de Qualidade da Integração	Relatório técnico	TP, TNP, TCF	Avalia o resultado da integração

11. Artefato A3.1 — Plano de Integração

11.1 Finalidade

O **Plano de Integração** define a estratégia de integração e harmonização a ser adotada para um conjunto de fontes. Ele documenta quais fontes serão integradas, com qual finalidade, por quais chaves, em qual granularidade, com quais regras e sob quais critérios de validação.

11.2 Tipo de artefato na DSR

Na perspectiva da Design Science Research, o Plano de Integração é um **método**, pois orienta o procedimento de integração, e uma **instanciação**, quando documentado em planilha, documento técnico, arquivo YAML, JSON ou workflow computacional.

11.3 Estrutura mínima

Campo	Descrição
Identificador do plano	Código único do plano
Objetivo da integração	Finalidade analítica
Fontes envolvidas	Bases ou datasets a integrar
Entidades compartilhadas	Entidades que justificam a integração
Chaves previstas	Chaves diretas, compostas ou critérios indiretos
Tipo de integração	Determinística, probabilística, temporal, territorial etc.

Granularidade esperada	Unidade de análise após integração
Período de integração	Recorte temporal considerado
Regras de harmonização	Padronizações e compatibilizações previstas
Critérios de sucesso	Condições para validar a integração
Riscos	Riscos técnicos, semânticos ou de governança
Responsável	Pessoa ou equipe responsável

11.4 Métricas associadas

- percentual de integrações com plano documentado;
- taxa de rastreabilidade da integração;
- percentual de fontes com chaves definidas;
- percentual de planos com critérios de sucesso;
- percentual de planos com riscos registrados.

11.5 Critérios de validação

O plano é considerado válido quando as fontes, chaves, entidades, regras, granularidade, finalidade e critérios de sucesso estão documentados antes da execução da integração.

11.6 Contribuição

Esse artefato evita integrações improvisadas, explicita decisões metodológicas e cria base para auditoria, lineage e reprodutibilidade.

12. Artefato A3.2 — Matriz de Correspondência entre Fontes

12.1 Finalidade

A **Matriz de Correspondência entre Fontes** registra relações possíveis ou efetivas entre fontes integradas, indicando entidades compartilhadas, chaves, períodos comuns, grau de confiança e tipo de vínculo.

12.2 Estrutura formal

Considerando o conjunto de fontes:

$F = \{f_1, f_2, ..., f_n\}$

a matriz de correspondência pode ser representada como:

$M = [m_{ij}]$

em que cada elemento **m_{ij}** representa a correspondência entre as fontes **f_i** e **f_j**.

Cada correspondência pode ser descrita por:

m_{ij} = {entidade, chave, tipo_relação, período_comum, granularidade, confiança, status}

12.3 Métricas associadas

- taxa de pareamento;
- taxa de não pareamento;
- número de relações identificadas;
- grau médio de confiança;
- proporção de relações validadas;
- proporção de relações com chave definida.

12.4 Critérios de validação

Uma correspondência é considerada válida quando possui entidade compartilhada, critério de associação e justificativa documentada.

12.5 Contribuição

Esse artefato explicita as relações entre fontes e apoia integração, grafos, lineage e governança.

13. Artefato A3.3 — Tabela de Mapeamento de Atributos

13.1 Finalidade

A **Tabela de Mapeamento de Atributos** documenta correspondências entre atributos de diferentes fontes, indicando atributos equivalentes, semelhantes, derivados ou incompatíveis.

13.2 Estrutura mínima

Campo	Descrição
Fonte A	Primeira fonte

Atributo A	Atributo da primeira fonte
Fonte B	Segunda fonte
Atributo B	Atributo da segunda fonte
Tipo de correspondência	Equivalente, parcial, derivado, incompatível
Transformação necessária	Regra de conversão ou padronização
Grau de confiança	Alto, médio ou baixo
Justificativa	Motivo da correspondência
Status	Validado, pendente, rejeitado

13.3 Métricas associadas

- taxa de harmonização de atributos;
- percentual de atributos equivalentes;
- percentual de atributos incompatíveis;
- percentual de correspondências validadas;
- grau médio de confiança das correspondências.

13.4 Contribuição

Esse artefato reduz ambiguidade entre atributos e permite harmonização controlada de variáveis provenientes de diferentes fontes.

14. Artefato A3.4 — Tabela de Equivalência Semântica

14.1 Finalidade

A **Tabela de Equivalência Semântica** documenta a compatibilização de categorias, códigos, conceitos e valores utilizados em diferentes fontes.

14.2 Estrutura mínima

Campo	Descrição
Fonte	Base de origem
Atributo	Atributo ou campo analisado
Valor original	Código ou categoria original
Valor harmonizado	Valor padronizado adotado
Conceito associado	Significado do valor
Tipo de equivalência	Exata, aproximada, agregada, desagregada
Regra de harmonização	Critério aplicado
Validação	Técnica, documental ou por especialista
Observações	Cuidados interpretativos

14.3 Métricas associadas

- taxa de harmonização semântica;
- percentual de categorias harmonizadas;
- número de categorias ambíguas;
- percentual de equivalências validadas;
- número de conflitos semânticos.

14.4 Contribuição

Esse artefato evita que categorias de fontes diferentes sejam combinadas de forma equivocada, fortalecendo interoperabilidade semântica e qualidade analítica.

15. Artefato A3.5 — Registro de Chaves e Entidades

15.1 Finalidade

O **Registro de Chaves e Entidades** documenta as entidades compartilhadas entre fontes e os identificadores utilizados para relacioná-las.

15.2 Estrutura mínima

Campo	Descrição
Entidade	Estudante, município, trabalhador, evento etc.
Fonte	Fonte em que a entidade aparece
Chave direta	Identificador explícito
Chave composta	Combinação de atributos
Critério indireto	Regra alternativa de associação
Cardinalidade	Um-para-um, um-para-muitos, muitos-para-muitos
Cobertura	Proporção de registros com chave
Confiança	Grau de confiabilidade
Observações	Limitações ou riscos

15.3 Métricas associadas

- taxa de chaves válidas;
- percentual de entidades com chave definida;
- percentual de registros sem identificador;
- grau de cobertura da chave;

- taxa de cardinalidade inesperada.

15.4 Contribuição

Esse artefato formaliza as bases da integração entre entidades, reduzindo riscos de pareamento incorreto e apoiando construção de grafos e lineage.

16. Artefato A3.6 — Registro de Compatibilidade de Granularidade

16.1 Finalidade

O **Registro de Compatibilidade de Granularidade** avalia se as fontes possuem níveis de análise compatíveis para integração.

16.2 Estrutura mínima

Campo	Descrição
Fonte A	Primeira fonte
Granularidade A	Unidade de análise da primeira fonte
Fonte B	Segunda fonte
Granularidade B	Unidade de análise da segunda fonte
Compatibilidade	Compatível, parcialmente compatível, incompatível
Ajuste necessário	Agregação, desagregação, filtragem ou restrição
Risco	Risco de duplicação, perda ou interpretação indevida
Regra adotada	Procedimento de compatibilização

16.3 Métricas associadas

- percentual de fontes com granularidade compatível;
- número de integrações que exigem agregação;
- número de integrações com risco de duplicidade;
- percentual de incompatibilidades documentadas.

16.4 Contribuição

Esse artefato previne integrações metodologicamente inadequadas e explicita os ajustes necessários entre diferentes níveis de detalhe.

17. Artefato A3.7 — Registro de Conflitos de Integração

17.1 Finalidade

O **Registro de Conflitos de Integração** documenta divergências identificadas durante o processo de integração.

17.2 Tipos de conflito

- conflito de valor;
- conflito de chave;
- conflito de categoria;
- conflito temporal;
- conflito de granularidade;
- conflito de duplicidade;
- conflito de fonte prioritária;
- conflito semântico;
- conflito de formato.

17.3 Estrutura mínima

Campo	Descrição
Identificador do conflito	Código único
Fontes envolvidas	Bases em conflito
Registro ou entidade	Registro afetado
Atributo	Campo conflitante

Valor A	Valor na primeira fonte
Valor B	Valor na segunda fonte
Tipo de conflito	Valor, chave, categoria, tempo etc.
Regra de resolução	Critério adotado
Status	Resolvido, pendente, ignorado, mantido
Justificativa	Motivo da decisão

17.4 Métricas associadas

- taxa de conflito entre fontes;
- taxa de conflitos resolvidos;
- número de conflitos por atributo;
- número de conflitos por fonte;
- percentual de conflitos pendentes.

17.5 Contribuição

Esse artefato torna explícitas decisões de resolução de inconsistências, evitando perda de transparência no processo de integração.

18. Artefato A3.8 — Dataset Integrado e Harmonizado

18.1 Finalidade

O **Dataset Integrado e Harmonizado** é o ativo resultante da integração de múltiplas fontes, após aplicação de regras de correspondência, compatibilização e harmonização.

18.2 Estrutura formal

O conjunto de datasets integrados pode ser representado como:

$$\mathbf{H = \{h_1, h_2, ..., h_m\}}$$

em que cada $\mathbf{h_j}$ representa um dataset integrado.

Cada dataset integrado deve ser descrito por:

h_j = {id, fontes_origem, chaves, regras, atributos_harmonizados, granularidade, período, conflitos, versão, métricas, lineage}

18.3 Métricas associadas

- taxa de pareamento;
- taxa de não pareamento;
- taxa de conflito;
- taxa de harmonização;
- taxa de rastreabilidade;
- cobertura temporal integrada;
- número de fontes integradas.

18.4 Critérios de validação

Um dataset integrado é considerado válido quando possui fontes de origem documentadas, regras de integração, chaves utilizadas, granularidade definida, métricas calculadas e conflitos registrados.

18.5 Contribuição

Esse artefato materializa a integração e fornece base consistente para transformação, qualidade, modelagem, grafos, indicadores e dashboards.

19. Artefato A3.9 — Manifesto de Integração

19.1 Finalidade

O **Manifesto de Integração** registra os parâmetros, regras, fontes, versões e decisões aplicadas em uma execução de integração.

19.2 Estrutura mínima

Campo	Descrição
Identificador da integração	Código único
Fontes utilizadas	Datasets de origem
Versões das fontes	Snapshots ou versões usadas
Chaves aplicadas	Chaves ou critérios de pareamento

Regras de harmonização	Regras aplicadas
Parâmetros	Parâmetros da execução
Dataset resultante	Ativo gerado
Métricas	Métricas de integração
Conflitos	Resumo dos conflitos
Responsável	Executor ou equipe
Data da execução	Momento da integração
Lineage	Relações de origem e dependência

19.3 Métricas associadas

- percentual de integrações com manifesto;
- completude dos manifestos;
- taxa de rastreabilidade da integração;
- percentual de manifestos com métricas;
- percentual de manifestos com versões documentadas.

19.4 Contribuição

Esse artefato fortalece auditoria, lineage, versionamento e reprodutibilidade da integração.

20. Artefato A3.10 — Relatório de Qualidade da Integração

20.1 Finalidade

O **Relatório de Qualidade da Integração** avalia o resultado do processo de integração e harmonização, apresentando métricas, alertas, conflitos, limitações e recomendações.

20.2 Estrutura mínima

Campo	Descrição
Processo avaliado	Identificador da integração
Fontes integradas	Bases utilizadas
Taxa de pareamento	Resultado obtido
Taxa de não pareamento	Resultado obtido
Taxa de conflito	Resultado obtido
Conflitos resolvidos	Quantidade ou proporção
Conflitos pendentes	Quantidade ou proporção
Cobertura temporal	Período integrado
Compatibilidade de granularidade	Avaliação obtida
Limitações	Restrições ou cuidados
Recomendações	Ações sugeridas

20.3 Métricas associadas

- TP;
- TNP;
- TCF;
- TCR;
- TRI;
- THA;
- THS;

- cobertura temporal integrada;
- grau de confiança da integração.

20.4 Contribuição

Esse artefato permite avaliar objetivamente se a integração produziu um ativo confiável e adequado aos objetivos analíticos.

21. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Datasets brutos, inventário de fontes, matriz de relacionamento, dicionários preliminares, chaves, entidades, granularidade, logs e metadados de ingestão.
Processamento	Seleção de fontes, identificação de entidades, definição de chaves, compatibilização de granularidade, harmonização de atributos, harmonização semântica, resolução de conflitos e geração de datasets integrados.
Artefatos	Plano de integração, matriz de correspondência, tabelas de mapeamento, registro de chaves, registro de granularidade, conflitos, dataset integrado, manifesto e relatório de qualidade.
Métricas	TP, TNP, TCV, TCF, TCR, THA, THS, TRI, GCI e cobertura temporal integrada.
Saídas	Datasets integrados e harmonizados, com regras, chaves, conflitos, métricas e lineage documentados.

22. Critérios gerais de validação do módulo

O Módulo 3 pode ser considerado concluído quando:

1. as fontes a integrar foram selecionadas com base em finalidade analítica explícita;
2. as entidades compartilhadas foram identificadas;
3. as chaves diretas, compostas ou critérios de pareamento foram documentados;
4. a granularidade das fontes foi avaliada;
5. atributos equivalentes foram mapeados;
6. categorias e códigos foram harmonizados;
7. conflitos foram identificados e registrados;

8. regras de resolução foram documentadas;
9. datasets integrados foram gerados;
10. métricas de integração foram calculadas;
11. manifestos de integração foram produzidos;
12. os datasets integrados possuem rastreabilidade suficiente para serem auditados e reproduzidos.

23. Relação com os demais módulos do modelo de referência

O Módulo 3 depende diretamente do **Módulo 1 — Fontes e Inventário de Dados**, pois utiliza o inventário, o mapa de bases, a matriz de relacionamento, o registro de granularidade e o dicionário preliminar para planejar a integração.

Também depende do **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, pois utiliza os datasets brutos, logs, snapshots e manifestos de ingestão como insumos materiais.

Para o **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, fornece datasets integrados que serão tratados, limpos, padronizados e enriquecidos.

Para o **Módulo 5 — Qualidade e Validação de Dados**, fornece métricas de pareamento, conflito, completude, consistência e compatibilidade de granularidade.

Para o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, fornece informações sobre perdas de registros, grupos sub-representados, atributos sensíveis e possíveis vieses introduzidos pela integração.

Para o **Módulo 7 — Metadados, Catálogo e Semântica dos Dados**, fornece tabelas de mapeamento, equivalências semânticas, datasets integrados e regras de harmonização.

Para o **Módulo 8 — Governança, Segurança, Privacidade e Conformidade**, fornece registros sobre restrições, sensibilidade e riscos de integração entre fontes.

Para o **Módulo 9 — Proveniência, Lineage e Rastreabilidade**, fornece vínculos entre fontes, regras de integração, versões e manifestos necessários para reconstruir a trajetória dos dados.

Para o **Módulo 10 — Versionamento e Reprodutibilidade**, fornece versões de datasets integrados, parâmetros e manifestos de execução.

Para o **Módulo 11 — Representações Analíticas: Tabular, Temporal, Espacial, Textual, Semântica e Grafos**, fornece entidades, relações e vínculos que podem ser representados como nós, arestas, atributos e janelas temporais.

Para o **Módulo 12 — Feature Engineering e Feature Store**, fornece bases integradas a partir das quais atributos derivados e features podem ser construídos.

Para o **Módulo 13 — Disponibilização Analítica**, fornece datasets integrados e harmonizados para consumo por dashboards, APIs, relatórios e modelos.

24. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 3 poderia integrar dados de matrícula, histórico acadêmico, notas, frequência, ambiente virtual de aprendizagem, questionário socioeconômico e registros de evasão.

As entidades compartilhadas poderiam incluir estudante, curso, disciplina, turma e período letivo. As chaves poderiam envolver RA, código do curso, código da disciplina, ano, semestre e identificadores de turma. A harmonização poderia compatibilizar categorias como “ativo”, “trancado”, “cancelado”, “evadido”, “concluído” ou “matriculado”.

O módulo deveria registrar conflitos, como estudantes presentes em uma base de matrícula, mas ausentes no ambiente virtual; divergência entre situação acadêmica em fontes distintas; disciplinas com códigos alterados ao longo do tempo; ou diferenças de granularidade entre notas por avaliação e situação final por disciplina.

O resultado seria um dataset integrado de trajetória acadêmica, contendo informações harmonizadas sobre matrícula, desempenho, frequência, interações e situação acadêmica, com regras de integração, métricas de pareamento e conflitos documentados.

25. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, o Módulo 3 poderia integrar dados de saúde, educação, assistência social, território, trabalho e cadastro populacional.

As entidades compartilhadas poderiam incluir cidadão, família, território, município, equipamento público, atendimento e benefício. A integração poderia exigir compatibilização de códigos territoriais, datas de atendimento, unidades administrativas, categorias de vulnerabilidade e registros de serviços prestados.

Conflitos poderiam surgir quando diferentes secretarias registram a mesma família com identificadores distintos, quando endereços são preenchidos de forma inconsistente ou quando territórios possuem recortes administrativos diferentes.

O resultado seria uma base integrada para análise territorial, permitindo construir indicadores de vulnerabilidade, mapas de atendimento, redes de serviços e modelos de priorização de políticas públicas.

26. Contribuição técnico-científica do módulo

A contribuição do Módulo 3 está em formalizar a integração e a harmonização de dados como uma etapa mensurável, auditável e reprodutível da Engenharia Analítica de Dados.

Em vez de tratar a integração como uma simples junção de tabelas, o UnespDataLens-RM propõe que cada processo de integração seja acompanhado de artefatos formais, tais como plano de integração, matriz de correspondência, tabela de mapeamento de atributos, equivalência semântica, registro de chaves, registro de conflitos, dataset integrado, manifesto de integração e relatório de qualidade.

Essa formalização contribui para:

- reduzir integrações incorretas;
- explicitar regras de harmonização;
- documentar decisões técnicas e semânticas;
- avaliar a qualidade dos pareamentos;

- identificar perdas e conflitos entre fontes;
- preservar rastreabilidade entre fontes e datasets integrados;
- apoiar lineage e versionamento;
- preparar dados para transformação, qualidade, grafos e modelagem;
- permitir reprodutibilidade e auditoria dos datasets integrados.

27. Síntese do módulo

O **Módulo 3 — Integração e Harmonização** transforma datasets brutos dispersos em ativos integrados, coerentes e documentados. Sua função é estabelecer relações entre fontes, identificar entidades compartilhadas, definir chaves, compatibilizar granularidades, harmonizar atributos e categorias, resolver conflitos e gerar datasets integrados com métricas e rastreabilidade.

Sua importância para o UnespDataLens-RM é central, pois os dados analíticos raramente emergem de uma única fonte. A robustez dos modelos, indicadores, dashboards, grafos e análises depende diretamente da qualidade da integração realizada.

Ao formalizar a integração como um processo composto por artefatos, métricas e critérios de validação, o Módulo 3 fortalece a confiabilidade, a transparência, a governança, a proveniência e a reprodutibilidade dos ativos analíticos produzidos pelo modelo de referência.

Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados

1. Finalidade do módulo

O **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados** tem como finalidade preparar os dados integrados e harmonizados para uso analítico, por meio da aplicação controlada, documentada e rastreável de procedimentos de padronização, correção, tratamento de inconsistências, criação de atributos derivados, normalização, enriquecimento contextual e adequação estrutural dos dados.

Esse módulo recebe como entrada os datasets brutos, integrados ou parcialmente harmonizados produzidos pelos módulos anteriores, especialmente o **Módulo 2 — Extração, Ingestão e Armazenamento Bruto** e o **Módulo 3 — Integração e Harmonização**. Sua função é transformar esses dados em estruturas mais adequadas para análise exploratória, construção de indicadores, modelagem estatística, aprendizagem de máquina, mineração de dados, representação em grafos, dashboards e suporte à decisão.

No contexto do **UnespDataLens-RM**, a transformação e a limpeza de dados não são tratadas como tarefas meramente operacionais ou pontuais. Elas são concebidas como processos técnico-metodológicos formais, compostos por regras explícitas, parâmetros, registros de execução, justificativas, métricas de impacto, documentação das alterações e possibilidade de reprocessamento. Dessa forma, cada transformação aplicada aos dados deve ser compreensível, auditável, reproduzível e associada aos artefatos de origem.

A limpeza dos dados envolve identificar e tratar problemas como valores ausentes, duplicidades, inconsistências de formato, outliers, categorias inválidas, valores fora de domínio, registros contraditórios, erros de digitação, problemas de codificação, violações de regras de negócio e incompatibilidades estruturais. O enriquecimento, por sua vez, envolve a criação de novos atributos, agregações, indicadores, variáveis temporais, atributos territoriais, atributos semânticos, variáveis relacionais e informações derivadas de outras fontes ou regras de negócio.

Assim, o Módulo 4 é essencial para transformar dados apenas integrados em **dados analiticamente utilizáveis**, preservando a rastreabilidade entre os dados originais, as regras aplicadas e os ativos resultantes.

2. Objetivo geral

O objetivo geral do Módulo 4 é **transformar, limpar, padronizar e enriquecer dados brutos ou integrados**, produzindo datasets tratados, documentados, rastreáveis e adequados ao uso analítico, por meio de regras, procedimentos, métricas e artefatos que permitam avaliar o impacto das transformações realizadas.

3. Objetivos específicos

São objetivos específicos do Módulo 4:

1. identificar problemas estruturais, sintáticos, semânticos e operacionais nos dados integrados ou brutos;

2. aplicar procedimentos de limpeza para tratamento de valores ausentes, duplicidades, inconsistências, formatos inválidos, outliers e valores fora de domínio;
3. padronizar atributos, tipos de dados, formatos de datas, códigos, categorias, unidades de medida, escalas e textos;
4. documentar todas as transformações aplicadas, incluindo regra, parâmetro, justificativa, fonte de origem, atributo afetado e impacto produzido;
5. criar atributos derivados, agregações, indicadores, variáveis temporais, variáveis categóricas e atributos contextuais;
6. enriquecer datasets com informações complementares, como dados territoriais, temporais, socioeconômicos, educacionais, institucionais ou semânticos;
7. preservar a rastreabilidade entre os dados originais, os dados transformados e as regras aplicadas;
8. avaliar o impacto das transformações por meio de métricas de completude, consistência, duplicidade, validade, alteração de distribuição, perda de registros e ganho informacional;
9. produzir datasets tratados e documentados para os módulos posteriores de qualidade, desbalanceamento, metadados, governança, lineage, versionamento, representação analítica, feature engineering e disponibilização.

4. Fundamentação conceitual do módulo

O Módulo 4 está fundamentado na compreensão de que os dados raramente estão prontos para análise após a ingestão ou integração. Mesmo quando as fontes são corretamente identificadas e os datasets são tecnicamente integrados, podem persistir problemas que comprometem a validade dos resultados analíticos.

Esses problemas incluem atributos incompletos, inconsistências entre campos, formatos heterogêneos, valores duplicados, categorias ambíguas, registros conflitantes, erros de preenchimento, variações textuais, unidades incompatíveis, outliers, registros inválidos e ausência de variáveis derivadas necessárias à análise.

A transformação de dados corresponde ao conjunto de operações que modificam estrutura, formato, tipo, escala, granularidade ou representação dos dados para torná-los adequados a determinado uso analítico. A limpeza de dados corresponde ao conjunto de procedimentos destinados a corrigir, remover, sinalizar ou mitigar problemas de qualidade. O enriquecimento corresponde à incorporação ou derivação de novas informações que aumentam o valor analítico dos dados.

No UnespDataLens-RM, essas três dimensões são tratadas de forma integrada:

- **transformação**, para adequar estrutura e formato;
- **limpeza**, para reduzir erros e inconsistências;
- **enriquecimento**, para ampliar o valor analítico e contextual dos dados.

À luz da **Design Science Research**, o Módulo 4 formaliza artefatos que permitem transformar atividades normalmente informais de preparação de dados em processos documentados, avaliáveis e reprodutíveis. Cada transformação deixa de ser uma decisão implícita em scripts ou notebooks e passa a ser registrada como uma regra formal, com justificativa, escopo, parâmetro, impacto e evidência de execução.

5. Formalização sintética do módulo

O Módulo 4 pode ser compreendido como um processo de transformação de datasets brutos ou integrados em datasets tratados e enriquecidos.

Considerando o conjunto de datasets brutos ou integrados:

$$\mathbf{D = \{d_1, d_2, ..., d_n\}}$$

o módulo aplica um conjunto de transformações:

$$\mathbf{T = \{t_1, t_2, ..., t_k\}}$$

em que cada transformação $\mathbf{t_i}$ representa uma operação aplicada a um ou mais atributos, registros ou datasets.

Cada transformação pode ser descrita por:

$$\mathbf{t_i = \{id, tipo, atributo_alvo, regra, parâmetro, justificativa, entrada, saída, impacto, responsável, versão\}}$$

O resultado do módulo é um conjunto de datasets tratados:

$$\mathbf{C = \{c_1, c_2, ..., c_m\}}$$

em que cada $\mathbf{c_j}$ representa um dataset limpo, padronizado ou enriquecido.

Cada dataset tratado deve ser associado a metadados de transformação:

$$\mathbf{c_j = \{id, dataset_origem, transformações_aplicadas, atributos_afetados, atributos_criados, registros_afetados, métricas_antes, métricas_depois, versão, lineage\}}$$

Dessa forma, a transformação deixa de ser uma operação invisível e passa a ser um processo rastreável, mensurável e reprodutível.

6. Entradas do módulo

As entradas do Módulo 4 incluem datasets brutos, datasets integrados, metadados, regras de negócio, resultados de profiling e artefatos produzidos pelos módulos anteriores.

As principais entradas são:

- datasets brutos do Módulo 2;
- datasets integrados e harmonizados do Módulo 3;
- inventário de fontes;
- dicionário preliminar de dados;
- matriz de relacionamento entre fontes;
- tabelas de mapeamento de atributos;
- tabelas de equivalência semântica;
- registro de granularidade;
- classificação de sensibilidade;
- manifestos de ingestão;
- manifestos de integração;
- logs de execução;
- registros de conflitos;
- regras de negócio;

- domínios de valores;
- vocabulários controlados;
- parâmetros de transformação;
- metadados de qualidade preliminar;
- documentação institucional;
- conhecimento de especialistas do domínio.

7. Processamento

O processamento do Módulo 4 envolve atividades de diagnóstico, definição de regras, execução das transformações, registro de alterações, cálculo de métricas antes/depois e geração de datasets tratados.

7.1 Diagnóstico inicial dos dados

A primeira atividade consiste em analisar os datasets recebidos para identificar problemas de qualidade, inconsistências, lacunas, formatos divergentes, tipos incompatíveis, duplicidades, outliers, categorias inválidas e atributos que exigem padronização.

Esse diagnóstico pode ser realizado por data profiling, inspeção estatística, validação de schema, análise de domínios, análise de frequência, verificação de regras de negócio e comparação com dicionários ou glossários.

7.2 Definição das regras de transformação

A segunda atividade consiste em definir as regras que serão aplicadas aos dados. Cada regra deve indicar o atributo ou conjunto de atributos afetados, o tipo de transformação, a justificativa, os parâmetros, o escopo, a condição de aplicação e o resultado esperado.

Exemplos de regras incluem padronizar datas para formato ISO, converter categorias textuais para códigos, transformar valores monetários, normalizar textos, remover espaços, corrigir encoding, agrupar categorias raras, imputar valores ausentes ou criar variáveis derivadas.

7.3 Tratamento de valores ausentes

A terceira atividade consiste em identificar e tratar valores ausentes. O tratamento pode envolver manutenção do valor ausente, imputação, exclusão controlada, criação de categoria específica, sinalização por variável indicadora ou complementação a partir de outra fonte.

A escolha da estratégia deve considerar o tipo do atributo, o percentual de ausência, o mecanismo provável de ausência, a finalidade analítica, o risco de viés e a possibilidade de uso posterior em modelos.

7.4 Tratamento de duplicidades

A quarta atividade consiste em identificar registros duplicados ou potencialmente duplicados. O tratamento pode envolver remoção, consolidação, seleção do registro mais recente, combinação de informações, priorização de fonte ou manutenção documentada quando a duplicidade for legítima.

Essa etapa deve preservar evidências da decisão tomada, evitando perda de informação sem justificativa.

7.5 Padronização de formatos e tipos

A quinta atividade consiste em padronizar formatos de datas, números, textos, códigos, categorias, unidades de medida e tipos de dados. Essa padronização garante que os dados possam ser processados de maneira consistente em módulos posteriores.

7.6 Tratamento de outliers e valores fora de domínio

A sexta atividade consiste em identificar valores extremos, improváveis ou fora do domínio esperado. Nem todo outlier deve ser removido, pois valores extremos podem ser relevantes dependendo do domínio. O módulo deve classificar os outliers como erro, evento raro, valor legítimo ou caso a ser revisado.

As ações possíveis incluem manutenção, correção, winsorização, truncamento, remoção controlada, transformação de escala ou sinalização por atributo auxiliar.

7.7 Criação de atributos derivados

A sétima atividade consiste em criar novos atributos a partir dos dados existentes. Esses atributos podem representar indicadores, contagens, proporções, médias, agregações temporais, variáveis categorizadas, variáveis binárias, faixas, indicadores territoriais, métricas relacionais ou atributos semânticos.

A criação de atributos derivados deve ser registrada com fórmula, regra, fonte de origem, granularidade e justificativa.

7.8 Enriquecimento dos dados

A oitava atividade consiste em incorporar informações adicionais ao dataset, aumentando seu valor analítico. O enriquecimento pode ocorrer por integração com bases externas, agregação de indicadores contextuais, incorporação de dados territoriais, construção de atributos temporais, associação a classificações institucionais ou vinculação a conceitos semânticos.

7.9 Registro das transformações

Toda transformação deve ser registrada em artefatos específicos. O registro deve indicar o que foi alterado, por que foi alterado, qual regra foi aplicada, quais atributos foram afetados, quantos registros foram impactados, quais métricas foram modificadas e qual versão do dataset resultante foi gerada.

7.10 Avaliação do impacto das transformações

Após a execução, o módulo deve comparar métricas antes e depois da transformação. Essa comparação permite avaliar se houve melhoria na completude, redução de duplicidades, aumento de validade, redução de inconsistências, alteração de distribuição ou perda significativa de registros.

8. Técnicas, métodos e algoritmos aplicáveis

O Módulo 4 pode empregar diferentes técnicas, métodos e algoritmos para transformação, limpeza e enriquecimento.

Entre as técnicas de limpeza, destacam-se:

- tratamento de valores ausentes;
- remoção ou consolidação de duplicidades;
- padronização de categorias;
- correção de formatos;
- validação de domínios;
- normalização textual;
- correção de encoding;
- detecção de inconsistências;
- tratamento de outliers;
- validação de regras de negócio;
- sinalização de registros suspeitos.

Entre as técnicas de transformação, podem ser utilizadas:

- conversão de tipos;
- normalização e padronização;
- discretização;
- binarização;
- codificação de variáveis categóricas;
- agregação;
- desagregação controlada;
- transformação temporal;
- criação de janelas móveis;
- cálculo de indicadores;
- criação de variáveis derivadas;
- transformação logarítmica;
- escalonamento;
- redução de dimensionalidade, quando aplicável.

Entre as técnicas de enriquecimento, destacam-se:

- associação com dados territoriais;
- incorporação de indicadores socioeconômicos;
- enriquecimento temporal;
- vinculação a dados externos;
- criação de métricas relacionais;
- construção de atributos por grupo;
- enriquecimento semântico;
- associação a glossários, taxonomias ou categorias padronizadas;
- derivação de indicadores a partir de regras de negócio.

Entre os algoritmos e métodos computacionais aplicáveis, destacam-se:

- imputação por média, mediana, moda ou valor constante;
- imputação por KNN;
- imputação por modelos supervisionados;
- detecção de outliers por z-score;
- detecção de outliers por IQR;
- Isolation Forest;
- Local Outlier Factor;

- DBSCAN;
- deduplicação por regras;
- deduplicação por similaridade;
- distância de Levenshtein;
- similaridade Jaro-Winkler;
- expressões regulares;
- normalização Min-Max;
- padronização z-score;
- one-hot encoding;
- ordinal encoding;
- target encoding, quando apropriado;
- agregações temporais;
- rolling windows;
- cálculo de métricas estatísticas;
- classificação baseada em regras.

Ferramentas possíveis incluem SQL, Python, pandas, PySpark, DuckDB, dbt, Great Expectations, Pandera, Deequ, Soda Core, notebooks versionados, ferramentas ETL/ELT e ambientes de lakehouse.

9. Métricas do módulo

As métricas do Módulo 4 têm como objetivo avaliar o impacto das transformações, limpezas e enriquecimentos realizados.

Métrica	Sigla	Descrição
Taxa de valores ausentes antes/depois	TVA	Proporção de valores ausentes antes e depois do tratamento
Taxa de duplicidade antes/depois	TD	Proporção de registros duplicados antes e depois da limpeza
Taxa de registros alterados	TRA	Proporção de registros modificados por transformações
Taxa de atributos transformados	TAT	Proporção de atributos submetidos a transformação
Taxa de atributos enriquecidos	TAE	Proporção de atributos criados ou enriquecidos

Taxa de valores fora de domínio	TVFD	Proporção de valores que violam domínios definidos
Taxa de outliers identificados	TOI	Proporção de registros ou valores classificados como outliers
Taxa de outliers tratados	TOT	Proporção de outliers tratados ou sinalizados
Perda de registros	PR	Proporção de registros removidos no processo
Ganho de completude	GC	Diferença de completude antes e depois do tratamento
Índice de impacto da transformação	IIT	Grau de alteração causado pelas transformações
Taxa de regras documentadas	TRD	Proporção de regras de transformação documentadas
Taxa de transformações rastreáveis	TTR	Proporção de transformações com origem, regra e impacto registrados
Taxa de atributos derivados documentados	TADD	Proporção de atributos derivados com fórmula e justificativa documentadas
Variação de distribuição	VD	Mudança na distribuição de atributos após transformação

9.1 Ganho de Completude — GC

O **Ganho de Completude** mede a melhoria na proporção de valores preenchidos após o tratamento de ausências.

$$\mathbf{GC} = \mathbf{completude\ depois - completude\ antes}$$

Essa métrica permite avaliar se as ações de tratamento aumentaram a disponibilidade de informação.

9.2 Taxa de Transformações Rastreáveis — TTR

A **Taxa de Transformações Rastreáveis** mede a proporção de transformações que possuem registro completo de origem, regra, parâmetro, atributo afetado e impacto.

TTR = número de transformações documentadas / número total de transformações aplicadas

Essa métrica é essencial para lineage, governança e reprodutibilidade.

9.3 Índice de Impacto da Transformação — IIT

O **Índice de Impacto da Transformação** avalia o grau de alteração produzido por um conjunto de transformações sobre o dataset.

Uma formulação simples pode considerar:

IIT = registros alterados / total de registros

ou, de forma ampliada:

IIT = (registros alterados + atributos alterados + valores imputados + registros removidos) / elementos avaliados

Esse índice ajuda a identificar transformações com alto impacto, que exigem maior justificativa e validação.

9.4 Taxa de Atributos Derivados Documentados — TADD

A **Taxa de Atributos Derivados Documentados** mede a proporção de atributos criados no módulo que possuem regra, fórmula, fonte de origem, granularidade e justificativa documentadas.

TADD = atributos derivados documentados / total de atributos derivados criados

Essa métrica permite avaliar a reprodutibilidade dos atributos enriquecidos.

9.5 Métricas específicas para preparação textual

Para dados textuais e semânticos, o Módulo 4 deve considerar métricas adicionais, tais como taxa de textos válidos, taxa de textos vazios, taxa de duplicidade textual, taxa de anonimização textual, cobertura lexical, taxa de ruído textual, diversidade lexical, tamanho médio dos documentos, taxa de idioma identificado, percentual de textos tratados com regras rastreáveis, cobertura de vocabulário, qualidade do corpus e grau de reprodutibilidade do pré-processamento textual.
Essas métricas permitem avaliar se o corpus produzido é adequado para tarefas analíticas posteriores e se as decisões de preparação textual foram suficientemente documentadas, versionadas e rastreáveis.

10. Artefatos formalizados do módulo

Os artefatos do Módulo 4 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, regras de construção, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A4.1	Plano de Transformação e Limpeza	Documento técnico	TRD	Define regras e escopo das transformações
A4.2	Perfil Inicial de Qualidade	Relatório técnico	TVA, TD, TVFD	Diagnostica problemas antes da limpeza
A4.3	Registro de Regras de Transformação	Tabela estruturada	TTR	Documenta regras aplicadas
A4.4	Registro de Tratamento de Ausentes	Tabela técnica	GC, TVA	Formaliza decisões sobre ausências
A4.5	Registro de Tratamento de Duplicidades	Tabela técnica	TD	Documenta remoção ou consolidação
A4.6	Registro de Outliers e Valores Fora de Domínio	Relatório/tabela	TOI, TOT	Documenta valores extremos e inválidos
A4.7	Catálogo de Atributos Derivados	Tabela ou catálogo	TADD	Documenta atributos criados
A4.8	Dataset Tratado e Enriquecido	Dataset versionado	IIT, TTR	Materializa os dados preparados
A4.9	Manifesto de	Documento	TTR	Registra parâmetros,

	Transformação	estruturado		versões e impactos
A4.10	Relatório de Impacto das Transformações	Relatório técnico	IIT, VD	Avalia efeitos antes/depois

11. Artefato A4.1 — Plano de Transformação e Limpeza

11.1 Finalidade

O **Plano de Transformação e Limpeza** define quais procedimentos serão aplicados aos datasets recebidos, indicando problemas identificados, regras previstas, atributos afetados, critérios de decisão, responsáveis e métricas de avaliação.

11.2 Tipo de artefato na DSR

Na perspectiva da Design Science Research, esse artefato é um **método**, pois orienta o processo de transformação, e uma **instanciação**, quando documentado em planilha, arquivo estruturado, notebook, workflow ou documento técnico.

11.3 Estrutura mínima

Campo	Descrição
Identificador do plano	Código único
Dataset de entrada	Dataset bruto ou integrado
Objetivo da transformação	Finalidade analítica
Problema identificado	Ausência, duplicidade, formato, outlier etc.
Atributo afetado	Campo ou variável submetida à regra
Regra prevista	Procedimento a aplicar
Parâmetros	Valores, limiares ou critérios

Justificativa	Motivo técnico ou de domínio
Responsável	Pessoa, equipe ou sistema
Métrica de avaliação	Indicador de impacto ou qualidade
Critério de sucesso	Condição para aceitar a transformação

11.4 Métricas associadas

- taxa de regras documentadas;
- percentual de atributos com regra definida;
- percentual de transformações com justificativa;
- percentual de planos com métrica de avaliação;
- percentual de planos com critério de sucesso.

11.5 Critérios de validação

O plano é considerado válido quando as principais transformações possuem atributo alvo, regra, parâmetro, justificativa, métrica e critério de aceitação documentados.

11.6 Contribuição

Esse artefato evita transformações improvisadas e torna explícitas as decisões metodológicas adotadas na preparação dos dados.

12. Artefato A4.2 — Perfil Inicial de Qualidade

12.1 Finalidade

O **Perfil Inicial de Qualidade** registra o diagnóstico dos dados antes da transformação, permitindo compreender o estado inicial do dataset.

12.2 Estrutura mínima

Campo	Descrição
Dataset	Dataset analisado

Atributo	Atributo avaliado
Tipo detectado	Tipo de dado identificado
Valores ausentes	Quantidade e percentual
Valores únicos	Cardinalidade
Duplicidades	Quantidade ou percentual
Valores fora de domínio	Ocorrências identificadas
Outliers	Ocorrências detectadas
Distribuição	Estatísticas descritivas
Problemas observados	Observações iniciais

12.3 Métricas associadas

- taxa de valores ausentes;
- taxa de duplicidade;
- taxa de valores fora de domínio;
- taxa de outliers;
- cardinalidade;
- distribuição de frequências;
- completude inicial.

12.4 Contribuição

Esse artefato fornece a linha de base para avaliar o impacto das transformações e das limpezas aplicadas.

13. Artefato A4.3 — Registro de Regras de Transformação

13.1 Finalidade

O **Registro de Regras de Transformação** documenta todas as regras aplicadas aos dados.

13.2 Estrutura formal

Cada regra de transformação pode ser representada por:

r_i = {id, tipo, atributo_alvo, condição, ação, parâmetro, justificativa, entrada, saída, versão}

13.3 Estrutura mínima

Campo	Descrição
Identificador da regra	Código único
Tipo de transformação	Padronização, imputação, conversão, derivação etc.
Dataset	Dataset afetado
Atributo alvo	Campo afetado
Condição	Quando a regra é aplicada
Ação	Transformação executada
Parâmetro	Valor ou limiar utilizado
Justificativa	Motivo da regra
Registros afetados	Quantidade ou percentual

Versão	Versão da regra
Responsável	Pessoa, equipe ou sistema

13.4 Métricas associadas

- taxa de transformações rastreáveis;
- taxa de regras documentadas;
- número de registros afetados;
- índice de impacto da transformação;
- percentual de regras versionadas.

13.5 Critérios de validação

Uma regra é considerada válida quando possui atributo alvo, condição, ação, parâmetro, justificativa, versão e impacto documentados.

13.6 Contribuição

Esse artefato transforma regras implícitas em scripts em elementos explícitos, auditáveis e reutilizáveis.

14. Artefato A4.4 — Registro de Tratamento de Ausentes

14.1 Finalidade

O **Registro de Tratamento de Ausentes** documenta as estratégias utilizadas para lidar com valores ausentes.

14.2 Estrutura mínima

Campo	Descrição
Dataset	Dataset analisado
Atributo	Atributo com ausência

Percentual de ausência	Taxa antes do tratamento
Mecanismo provável	MCAR, MAR, MNAR ou desconhecido
Estratégia adotada	Manter, imputar, remover, sinalizar etc.
Método	Média, mediana, KNN, categoria, modelo etc.
Justificativa	Motivo da decisão
Registros afetados	Quantidade afetada
Risco de viés	Baixo, médio ou alto
Métrica após tratamento	Completude final

14.3 Métricas associadas

- taxa de valores ausentes antes/depois;
- ganho de completude;
- percentual de registros imputados;
- percentual de atributos com ausência tratada;
- risco de viés associado.

14.4 Contribuição

Esse artefato evita que valores ausentes sejam tratados sem justificativa, reduzindo riscos de distorção analítica e viés.

15. Artefato A4.5 — Registro de Tratamento de Duplicidades

15.1 Finalidade

O **Registro de Tratamento de Duplicidades** documenta registros duplicados ou potencialmente duplicados e as decisões adotadas.

15.2 Estrutura mínima

Campo	Descrição
Dataset	Dataset analisado
Chave ou critério	Critério de detecção
Tipo de duplicidade	Exata, parcial, semântica, temporal
Registros duplicados	Quantidade identificada
Estratégia adotada	Remover, consolidar, manter, revisar
Regra de precedência	Critério de escolha
Registros removidos	Quantidade removida
Registros consolidados	Quantidade consolidada
Justificativa	Motivo da decisão
Impacto	Efeito no dataset

15.3 Métricas associadas

- taxa de duplicidade antes/depois;
- número de registros removidos;
- número de registros consolidados;
- índice de impacto da deduplicação.

15.4 Contribuição

Esse artefato torna transparente a remoção ou consolidação de registros, evitando perda não documentada de informação.

16. Artefato A4.6 — Registro de Outliers e Valores Fora de Domínio

16.1 Finalidade

O **Registro de Outliers e Valores Fora de Domínio** documenta valores extremos, improváveis ou inválidos identificados no dataset.

16.2 Estrutura mínima

Campo	Descrição
Dataset	Dataset analisado
Atributo	Atributo avaliado
Valor identificado	Valor extremo ou inválido
Tipo de ocorrência	Outlier, erro, valor fora de domínio, evento raro
Método de detecção	IQR, z-score, regra, modelo etc.
Estratégia adotada	Manter, remover, corrigir, sinalizar
Justificativa	Motivo da decisão
Registros afetados	Quantidade afetada
Status	Resolvido, mantido, pendente

16.3 Métricas associadas

- taxa de outliers identificados;
- taxa de outliers tratados;
- taxa de valores fora de domínio;
- número de registros sinalizados;
- impacto da remoção ou correção.

16.4 Contribuição

Esse artefato permite distinguir erro de dado, evento raro e valor analiticamente relevante, preservando transparência nas decisões.

17. Artefato A4.7 — Catálogo de Atributos Derivados

17.1 Finalidade

O **Catálogo de Atributos Derivados** documenta variáveis criadas a partir dos dados originais, como indicadores, agregações, faixas, contagens, proporções, métricas temporais ou atributos contextuais.

17.2 Estrutura formal

Cada atributo derivado pode ser representado por:

v_i = {id, nome, descrição, fórmula, fontes_origem, atributos_origem, granularidade, periodicidade, justificativa, versão}

17.3 Estrutura mínima

Campo	Descrição
Nome do atributo	Nome da variável derivada
Descrição	Significado analítico
Fórmula ou regra	Como o atributo é calculado
Atributos de origem	Campos utilizados
Fontes de origem	Bases utilizadas
Granularidade	Unidade de análise

Periodicidade	Recorte temporal
Tipo de dado	Numérico, categórico, binário etc.
Justificativa	Motivo da criação
Limitações	Cuidados de uso
Versão	Versão da regra

17.4 Métricas associadas

- taxa de atributos derivados documentados;
- número de atributos derivados;
- percentual de atributos com fórmula;
- percentual de atributos com fonte de origem;
- percentual de atributos com granularidade documentada.

17.5 Contribuição

Esse artefato aumenta a reprodutibilidade das variáveis analíticas e prepara o caminho para feature engineering e feature store.

18. Artefato A4.8 — Dataset Tratado e Enriquecido

18.1 Finalidade

O **Dataset Tratado e Enriquecido** é o ativo resultante da aplicação das regras de limpeza, padronização, transformação e enriquecimento.

18.2 Estrutura formal

O conjunto de datasets tratados pode ser representado como:

$$\mathbf{C = \{c_1, c_2, \ldots, c_m\}}$$

em que cada $\mathbf{c_j}$ representa um dataset tratado.

Cada dataset tratado deve ser descrito por:

$$\mathbf{c_j = \{id, dataset_origem, regras_aplicadas, atributos_transformados, atributos_derivados, registros_afetados, métricas_antes, métricas_depois, versão, lineage\}}$$

18.3 Métricas associadas

- taxa de transformações rastreáveis;
- índice de impacto da transformação;
- ganho de completude;
- redução de duplicidade;
- taxa de atributos enriquecidos;
- variação de distribuição.

18.4 Critérios de validação

Um dataset tratado é considerado válido quando possui origem documentada, regras aplicadas, métricas antes/depois, atributos derivados documentados e versão identificada.

18.5 Contribuição

Esse artefato materializa os dados preparados para análise, preservando rastreabilidade entre origem, transformação e resultado.

19. Artefato A4.9 — Manifesto de Transformação

19.1 Finalidade

O **Manifesto de Transformação** resume os principais elementos de uma execução de transformação, registrando datasets, regras, parâmetros, versões, métricas e impactos.

19.2 Estrutura mínima

Campo	Descrição
Identificador da execução	Código único
Dataset de entrada	Dataset original ou integrado
Dataset de saída	Dataset tratado
Regras aplicadas	Lista de regras executadas
Parâmetros	Valores e limiares utilizados

Atributos afetados	Campos modificados
Atributos criados	Variáveis derivadas
Registros afetados	Quantidade ou percentual
Métricas antes	Indicadores prévios
Métricas depois	Indicadores posteriores
Responsável	Executor humano ou sistema
Data da execução	Momento da transformação
Versão	Versão do processo

19.3 Métricas associadas

- completude do manifesto;
- taxa de transformações rastreáveis;
- percentual de execuções com métricas antes/depois;
- percentual de manifestos com versão;
- percentual de manifestos com lineage.

19.4 Contribuição

Esse artefato apoia auditoria, lineage, versionamento e reprodutibilidade das transformações aplicadas.

20. Artefato A4.10 — Relatório de Impacto das Transformações

20.1 Finalidade

O **Relatório de Impacto das Transformações** avalia os efeitos das transformações sobre os dados.

20.2 Estrutura mínima

Campo	Descrição
Dataset avaliado	Dataset tratado
Transformações aplicadas	Regras executadas
Métricas antes	Indicadores prévios
Métricas depois	Indicadores posteriores
Registros alterados	Quantidade ou percentual
Registros removidos	Quantidade ou percentual
Atributos alterados	Campos modificados
Atributos criados	Variáveis derivadas
Alteração de distribuição	Mudança estatística observada
Riscos identificados	Possíveis distorções
Recomendações	Ações futuras

20.3 Métricas associadas

- índice de impacto da transformação;
- variação de distribuição;
- perda de registros;
- ganho de completude;

- redução de duplicidade;
- taxa de atributos enriquecidos.

20.4 Contribuição

Esse artefato permite avaliar se as transformações melhoraram os dados sem introduzir distorções excessivas ou perdas relevantes.

20.5 Artefatos específicos da Preparação de Dados Textuais e Semânticos

Além dos artefatos gerais de transformação, o subpipeline textual deve produzir artefatos específicos para garantir governança, rastreabilidade e reprodutibilidade da preparação de textos.
Corpus bruto: preserva os textos originais em seu estado inicial, com metadados de origem e coleta.
Corpus limpo e tratado: registra a versão preparada dos textos após limpeza, normalização e anonimização.
Regras de pré-processamento textual: documentam operações aplicadas, parâmetros, versões e justificativas.
Vocabulário e dicionários de domínio: registram termos, categorias, abreviações, sinônimos e conceitos relevantes.
Matriz textual: representa os textos por Bag of Words, TF-IDF ou outra representação vetorial.
Embeddings persistidos: armazenam vetores semânticos gerados por modelos específicos e suas versões.
Relatório de qualidade textual: avalia ruído, duplicidade, cobertura lexical, textos vazios, idioma e anonimização.
Registro de lineage textual: reconstrói o percurso texto original -> corpus tratado -> representação -> feature -> produto analítico.
Manifesto de reprodutibilidade textual: reúne corpus, scripts, parâmetros, versões, modelos, embeddings e resultados necessários para reexecutar o pipeline textual.

21. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Datasets brutos, datasets integrados, dicionários, regras de negócio, metadados, registros de conflitos, logs, manifestos e resultados de profiling.
Processamento	Diagnóstico inicial, definição de regras, tratamento de ausentes, duplicidades, outliers, padronização, criação de atributos derivados, enriquecimento e avaliação de impacto.

Artefatos	Plano de transformação, perfil inicial de qualidade, registro de regras, registros de ausentes, duplicidades, outliers, catálogo de atributos derivados, dataset tratado, manifesto e relatório de impacto.
Métricas	TVA, TD, TRA, TAT, TAE, TVFD, TOI, TOT, PR, GC, IIT, TRD, TTR, TADD e VD.
Saídas	Datasets tratados, limpos, padronizados e enriquecidos, com regras, métricas, versões, impactos e lineage documentados.

22. Critérios gerais de validação do módulo

O Módulo 4 pode ser considerado concluído quando:

1. os datasets de entrada foram diagnosticados por perfil inicial de qualidade;
2. as regras de transformação foram documentadas;
3. os valores ausentes foram tratados ou justificados;
4. duplicidades foram identificadas, tratadas ou justificadas;
5. outliers e valores fora de domínio foram analisados;
6. formatos, tipos, categorias e códigos foram padronizados;
7. atributos derivados foram documentados com fórmula, origem e justificativa;
8. datasets tratados e enriquecidos foram gerados;
9. métricas antes e depois foram calculadas;
10. o impacto das transformações foi avaliado;
11. os manifestos de transformação foram produzidos;
12. as transformações possuem rastreabilidade suficiente para auditoria e reprodutibilidade.

23. Relação com os demais módulos do modelo de referência

O Módulo 4 depende do **Módulo 1 — Fontes e Inventário de Dados**, pois utiliza o dicionário preliminar, a granularidade, a classificação de sensibilidade e as limitações conhecidas das fontes.

Depende do **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, pois utiliza datasets brutos, logs, snapshots e manifestos como insumos para transformação controlada.

Depende do **Módulo 3 — Integração e Harmonização**, pois recebe datasets integrados, regras de harmonização, conflitos e relações entre fontes.

Para o **Módulo 5 — Qualidade e Validação de Dados**, fornece datasets tratados e métricas antes/depois que permitem avaliação mais rigorosa da qualidade.

Para o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, fornece dados preparados, mas também registra transformações que podem alterar distribuições, grupos e representatividade.

Para o **Módulo 7 — Metadados, Catálogo e Semântica dos Dados**, fornece atributos derivados, regras de transformação, descrições, versões e limitações.

Para o **Módulo 8 — Governança, Segurança, Privacidade e Conformidade**, fornece registros sobre atributos sensíveis transformados, anonimizados, removidos ou derivados.

Para o **Módulo 9 — Proveniência, Lineage e Rastreabilidade**, fornece regras, manifestos e vínculos entre dados de entrada e saída.

Para o **Módulo 10 — Versionamento e Reprodutibilidade**, fornece versões de regras, parâmetros e datasets tratados.

Para o **Módulo 11 — Representações Analíticas: Tabular, Temporal, Espacial, Textual, Semântica e Grafos**, fornece dados padronizados e enriquecidos para diferentes formas de representação.

Para o **Módulo 12 — Feature Engineering e Feature Store**, fornece atributos derivados iniciais, variáveis tratadas e datasets preparados para construção de features.

Para o **Módulo 13 — Disponibilização Analítica**, fornece datasets limpos e enriquecidos para consumo por dashboards, APIs, relatórios e modelos.

Para o **Módulo 14 — Monitoramento, Drift e Evolução dos Ativos**, fornece métricas de distribuição e transformações que podem ser monitoradas ao longo do tempo.

24. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 4 poderia receber um dataset integrado contendo dados de matrícula, notas, frequência, interações em ambiente virtual, dados socioeconômicos e situação acadêmica.

As transformações poderiam incluir padronização de períodos letivos, normalização de códigos de curso, tratamento de notas ausentes, consolidação de registros duplicados de matrícula, padronização de categorias de situação acadêmica, criação de indicadores de desempenho, cálculo de frequência acumulada, criação de variáveis temporais e enriquecimento com dados socioeconômicos ou territoriais.

O módulo poderia criar atributos derivados como média de desempenho por semestre, número de reprovações acumuladas, tempo desde o ingresso, quantidade de disciplinas cursadas, indicador de baixa frequência, trajetória de matrícula e participação em atividades no ambiente virtual.

Cada transformação seria registrada com regra, justificativa, atributos afetados, impacto e versão, permitindo reproduzir a construção da base analítica educacional.

25. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, o Módulo 4 poderia tratar dados integrados de saúde, educação, assistência social, território e trabalho.

As transformações poderiam incluir padronização de endereços, compatibilização de códigos territoriais, tratamento de datas de atendimento, agregação por bairro ou município, criação de indicadores de vulnerabilidade, normalização de categorias de atendimento, tratamento de duplicidades de famílias e enriquecimento com indicadores socioeconômicos externos.

O resultado seria uma base tratada e enriquecida para análise territorial, construção de painéis, identificação de demandas, priorização de políticas e modelagem de risco social.

25.1 Exemplo de aplicação em análise de sentimentos

Em uma aplicação envolvendo análise de sentimentos, o Módulo 4 poderia receber textos provenientes de respostas abertas, fóruns, comentários, avaliações institucionais, mensagens ou documentos. O subpipeline textual realizaria remoção de ruídos, anonimização, normalização, tokenização, tratamento de stopwords, preservação de negações, correção de abreviações, geração de vocabulário, representação TF-IDF e criação de embeddings.
O resultado seria um corpus textual preparado, documentado, versionado e rastreável, adequado para ser utilizado em tarefas analíticas posteriores, como análise de sentimentos, classificação de emoções, extração de tópicos, sumarização ou detecção de entidades.
Nesse exemplo, a análise de sentimentos não é tratada como etapa de preparação, mas como técnica analítica que consome os artefatos produzidos pelo pipeline textual.

26. Contribuição técnico-científica do módulo

A contribuição do Módulo 4 está em formalizar a transformação, a limpeza e o enriquecimento de dados como processos mensuráveis, auditáveis e reprodutíveis da Engenharia Analítica de Dados.

Em vez de tratar a preparação de dados como uma sequência de scripts não documentados, o UnespDataLens-RM propõe que cada transformação produza artefatos formais, como plano de transformação, perfil inicial de qualidade, registro de regras, registros de tratamento de ausentes, duplicidades e outliers, catálogo de atributos derivados, manifesto de transformação e relatório de impacto.

Essa formalização contribui para:

- reduzir arbitrariedade na preparação dos dados;
- explicitar regras e justificativas de transformação;
- preservar rastreabilidade entre dados originais e tratados;
- avaliar impacto das transformações;
- reduzir riscos de distorção analítica;
- apoiar qualidade, governança e reprodutibilidade;
- documentar atributos derivados;
- preparar os dados para feature engineering, grafos, modelos e dashboards;
- permitir auditoria e reexecução dos processos.

27. Síntese do módulo

O **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados** prepara os dados para uso analítico por meio de procedimentos formais de limpeza,

padronização, transformação e enriquecimento. Ele atua sobre dados brutos ou integrados, produzindo datasets tratados, documentados, versionáveis e rastreáveis.

Sua importância para o UnespDataLens-RM é central, pois dados integrados ainda podem conter problemas que comprometem a qualidade das análises. Ao formalizar regras, métricas, impactos e artefatos, o módulo transforma a preparação dos dados em uma etapa avaliável e reprodutível.

Dessa forma, o Módulo 4 fortalece a confiabilidade dos ativos analíticos e prepara a base para qualidade, governança, desbalanceamento, fairness, metadados, lineage, versionamento, feature engineering, representação em grafos e disponibilização analítica.

Módulo 5 — Qualidade e Validação de Dados

1. Finalidade do módulo

O **Módulo 5 — Qualidade e Validação de Dados** tem como finalidade avaliar, mensurar, documentar e validar a qualidade dos dados produzidos ao longo do modelo de referência **UnespDataLens-RM**, considerando os dados brutos, integrados, harmonizados, transformados, limpos e enriquecidos nos módulos anteriores.

Esse módulo atua como um componente central de controle da confiabilidade dos ativos analíticos. Sua função é verificar se os dados atendem a critérios mínimos de completude, consistência, validade, unicidade, integridade, atualidade, conformidade, estabilidade, coerência semântica e adequação ao uso pretendido.

Em projetos de Engenharia Analítica de Dados, a qualidade dos dados não pode ser tratada apenas como uma inspeção pontual ou uma etapa final antes da modelagem. Problemas de qualidade podem surgir desde a origem, durante a ingestão, na integração entre fontes, na transformação, na criação de atributos derivados ou na disponibilização dos ativos. Por isso, o UnespDataLens-RM propõe que a qualidade seja tratada como um processo formal, contínuo, mensurável e rastreável.

No contexto do modelo de referência, o Módulo 5 recebe como insumos os artefatos dos módulos anteriores, como inventários, datasets brutos, manifestos de ingestão, datasets integrados, regras de harmonização, datasets tratados, registros de transformação e catálogos preliminares. A partir desses insumos, o módulo define regras de validação, executa testes, calcula métricas, registra não conformidades e produz relatórios de qualidade que orientam decisões sobre uso, correção, rejeição, reprocessamento ou liberação dos dados.

Assim, o Módulo 5 transforma a avaliação da qualidade dos dados em uma etapa técnico-metodológica formal, capaz de produzir evidências objetivas sobre a confiabilidade dos ativos analíticos gerados pelo UnespDataLens-RM.

2. Objetivo geral

O objetivo geral do Módulo 5 é **avaliar e validar a qualidade dos dados em diferentes etapas do pipeline analítico**, por meio da definição de dimensões de qualidade, regras de validação, métricas, testes automatizados, registros de não conformidade e critérios de aceitação dos ativos de dados.

3. Objetivos específicos

São objetivos específicos do Módulo 5:

1. definir dimensões de qualidade adequadas ao domínio e à finalidade analítica do projeto;
2. estabelecer regras de validação para fontes, atributos, registros, datasets, chaves, domínios de valores, relações e variáveis derivadas;

3. calcular métricas de qualidade antes e depois das etapas de integração, transformação, limpeza e enriquecimento;
4. identificar problemas de completude, inconsistência, duplicidade, invalidade, quebra de integridade referencial, desatualização, anomalias e violações de regras de negócio;
5. registrar não conformidades e classificá-las quanto à severidade, impacto, origem provável e possibilidade de correção;
6. produzir relatórios de qualidade para orientar decisões de correção, reprocessamento, aceitação ou rejeição de ativos analíticos;
7. documentar testes de validação e critérios mínimos de qualidade para cada tipo de ativo;
8. fornecer insumos para os módulos de governança, metadados, lineage, versionamento, monitoramento, feature engineering e disponibilização analítica;
9. criar evidências de qualidade que possam ser auditadas, reproduzidas e acompanhadas ao longo do ciclo de vida dos dados.

4. Fundamentação conceitual do módulo

A qualidade dos dados pode ser compreendida como o grau em que os dados são adequados para determinado propósito analítico, decisório, operacional ou científico. No contexto do UnespDataLens-RM, qualidade não significa apenas ausência de erros, mas envolve a capacidade dos dados de representar adequadamente entidades, eventos, atributos, relações e fenômenos relevantes ao problema investigado.

A avaliação da qualidade deve considerar múltiplas dimensões. Entre as principais, destacam-se:

Dimensão	Descrição
Completude	Grau em que os dados necessários estão presentes.
Consistência	Ausência de contradições entre campos, registros ou fontes.
Validade	Conformidade dos valores com domínios, formatos e regras esperadas.
Unicidade	Ausência de duplicidades indevidas.
Integridade	Preservação de relações válidas entre entidades, chaves e tabelas.
Atualidade	Grau em que os dados estão atualizados em relação ao fenômeno

	representado.
Conformidade	Aderência a padrões, schemas, contratos e regras institucionais.
Precisão	Grau em que os valores representam corretamente a realidade ou a fonte de referência.
Coerência semântica	Compatibilidade entre o significado dos dados, suas categorias e seu uso analítico.
Adequação ao uso	Grau em que o ativo atende ao propósito analítico declarado.

No UnespDataLens-RM, a qualidade dos dados deve ser avaliada de forma contextual. Uma base pode ser adequada para análise exploratória agregada, mas inadequada para modelagem preditiva individual. Da mesma forma, um atributo pode ser útil para descrever tendências gerais, mas não confiável para decisões sensíveis ou automatizadas.

À luz da **Design Science Research**, o Módulo 5 formaliza a qualidade dos dados como um conjunto de artefatos avaliáveis, composto por regras, métricas, testes, relatórios, registros de não conformidade e critérios de validação. Essa formalização permite transformar avaliações subjetivas ou informais em evidências mensuráveis e reproduzíveis.

5. Formalização sintética do módulo

O Módulo 5 pode ser compreendido como um processo de avaliação de qualidade aplicado a ativos de dados produzidos pelos módulos anteriores.

Considerando um conjunto de datasets ou ativos analíticos:

$$\mathbf{D = \{d_1, d_2, ..., d_n\}}$$

o módulo define um conjunto de dimensões de qualidade:

$$\mathbf{Q = \{q_1, q_2, ..., q_k\}}$$

em que cada dimensão $\mathbf{q_i}$ representa um aspecto avaliável, como completude, consistência, validade, unicidade, integridade ou atualidade.

Para cada dataset $\mathbf{d_j}$, são aplicadas regras de validação:

$$\mathbf{R = \{r_1, r_2, ..., r_m\}}$$

em que cada regra $\mathbf{r_l}$ pode ser descrita por:

r_l = {id, dimensão, atributo_alvo, condição, critério, limiar, severidade, ação_recomendada}

A avaliação produz um conjunto de resultados de qualidade:

$V = \{v_1, v_2, ..., v_p\}$

em que cada resultado **v_p** registra a execução de uma regra de validação sobre um dataset, atributo ou conjunto de registros.

Cada resultado pode ser descrito por:

v_p = {dataset, regra, status, valor_obtido, limiar, registros_afetados, severidade, data_execução, recomendação}

Dessa forma, a qualidade dos dados deixa de ser uma percepção subjetiva e passa a ser representada por métricas, regras, evidências e critérios de aceitação.

6. Entradas do módulo

As entradas do Módulo 5 incluem datasets, artefatos, metadados e registros produzidos pelos módulos anteriores.

As principais entradas são:

- datasets brutos do Módulo 2;
- datasets integrados e harmonizados do Módulo 3;
- datasets tratados e enriquecidos do Módulo 4;
- inventário de fontes;
- dicionário preliminar de dados;
- dicionário consolidado, quando disponível;
- manifestos de ingestão;
- manifestos de integração;
- manifestos de transformação;
- logs de execução;
- registros de conflitos;
- registros de valores ausentes;
- registros de duplicidades;
- registros de outliers;
- catálogo de atributos derivados;
- regras de negócio;
- domínios de valores;
- schemas esperados;
- restrições de integridade;
- contratos de dados;
- critérios de aceitação;
- parâmetros de validação;
- conhecimento de especialistas do domínio.

7. Processamento

O processamento do Módulo 5 envolve atividades de definição de dimensões, elaboração de regras, execução de testes, cálculo de métricas, identificação de não conformidades, classificação de severidade, recomendação de ações e documentação dos resultados.

7.1 Definição das dimensões de qualidade

A primeira atividade consiste em definir quais dimensões de qualidade serão avaliadas em cada dataset ou ativo analítico. Nem todas as dimensões precisam ser aplicadas da mesma forma a todos os dados. A seleção deve considerar domínio, finalidade, sensibilidade, granularidade, uso pretendido e riscos associados.

Por exemplo, em dados educacionais individuais, completude, consistência, integridade, atualidade e sensibilidade podem ser críticas. Em indicadores territoriais agregados, cobertura temporal, validade de domínio e coerência semântica podem ser prioritárias.

7.2 Definição das regras de validação

A segunda atividade consiste em transformar dimensões de qualidade em regras verificáveis. Uma regra de validação pode indicar que um atributo não deve possuir valores nulos, que uma data deve respeitar determinado intervalo, que uma chave deve ser única, que uma categoria deve pertencer a um conjunto válido ou que uma relação entre tabelas deve preservar integridade referencial.

Cada regra deve possuir descrição, atributo alvo, condição, limiar, severidade, justificativa e ação recomendada.

7.3 Execução dos testes de qualidade

A terceira atividade consiste em executar testes de qualidade sobre os datasets. Esses testes podem ser manuais, semiautomatizados ou automatizados. A execução deve gerar resultados documentados, indicando regras aprovadas, regras violadas, número de registros afetados, exemplos de problemas e impacto potencial.

7.4 Cálculo das métricas de qualidade

A quarta atividade consiste em calcular métricas associadas às dimensões avaliadas. Essas métricas podem ser calculadas por atributo, por dataset, por fonte, por período, por grupo, por entidade ou por ativo analítico.

Os resultados podem ser comparados com valores de referência, limiares mínimos ou métricas de execuções anteriores.

7.5 Identificação e classificação de não conformidades

A quinta atividade consiste em registrar problemas encontrados durante a validação. As não conformidades devem ser classificadas quanto à dimensão afetada, severidade, origem provável, impacto analítico e possibilidade de correção.

Essa classificação permite priorizar ações e evitar que problemas críticos sejam tratados da mesma forma que problemas de baixo impacto.

7.6 Recomendações de correção ou reprocessamento

A sexta atividade consiste em indicar ações recomendadas para tratar os problemas encontrados. Essas ações podem incluir retorno ao Módulo 4 para nova transformação, revisão da integração no Módulo 3, nova ingestão no Módulo 2, complementação documental no Módulo 7 ou restrição de uso do ativo no Módulo 8.

7.7 Geração de relatório de qualidade

A sétima atividade consiste em produzir um relatório consolidado de qualidade, contendo métricas, regras aplicadas, não conformidades, severidade, recomendações e decisão sobre o status do dataset.

O relatório deve permitir responder se o ativo está aprovado, aprovado com ressalvas, rejeitado ou pendente de correção.

7.8 Registro dos resultados para monitoramento

Os resultados das avaliações devem ser preservados para permitir comparação futura. Isso permite acompanhar evolução da qualidade, detectar degradação, monitorar melhorias e apoiar o Módulo 14 — Monitoramento, Drift e Evolução dos Ativos.

8. Técnicas, métodos e algoritmos aplicáveis

O Módulo 5 pode empregar diferentes técnicas, métodos e ferramentas de validação.

Entre as técnicas de avaliação de qualidade, destacam-se:

- data profiling;
- validação de schema;
- validação de domínio;
- validação de tipos;
- validação de chaves;
- validação de unicidade;
- validação de integridade referencial;
- validação de consistência entre campos;
- validação de regras de negócio;
- verificação de completude;
- análise de duplicidade;
- detecção de valores inválidos;
- detecção de outliers;
- comparação entre fontes;
- comparação temporal;
- análise de distribuição;
- detecção de anomalias;
- validação por amostragem;
- validação por especialista de domínio.

Entre os métodos e algoritmos aplicáveis, destacam-se:

- testes de constraints;
- regras booleanas;
- expressões regulares;
- validação por intervalo;
- validação por conjunto de valores permitidos;
- análise de cardinalidade;

- análise de frequência;
- comparação de distribuições;
- teste Kolmogorov-Smirnov;
- Population Stability Index;
- divergência de Kullback-Leibler;
- Jensen-Shannon divergence;
- z-score;
- IQR;
- Isolation Forest;
- Local Outlier Factor;
- regras de integridade relacional;
- validação de dependências funcionais;
- testes automatizados de dados.

Ferramentas possíveis incluem SQL, Python, pandas, PySpark, Great Expectations, Pandera, Deequ, Soda Core, dbt tests, ydata-profiling, Evidently, notebooks versionados e ferramentas de observabilidade de dados.

9. Métricas do módulo

As métricas do Módulo 5 têm como objetivo mensurar a qualidade dos dados em diferentes dimensões.

Métrica	Sigla	Descrição
Taxa de completude	TC	Proporção de valores preenchidos em relação ao total esperado
Taxa de valores ausentes	TVA	Proporção de valores nulos ou ausentes
Taxa de validade	TV	Proporção de valores que respeitam domínio, tipo ou formato
Taxa de inconsistência	TI	Proporção de registros que violam regras de consistência
Taxa de duplicidade	TD	Proporção de registros duplicados
Taxa de unicidade	TU	Proporção de valores únicos onde unicidade é esperada

Taxa de integridade referencial	TIR	Proporção de relações válidas entre chaves
Taxa de conformidade de schema	TCS	Proporção de atributos compatíveis com o schema esperado
Taxa de atualidade	TA	Grau de atualização dos dados em relação à periodicidade esperada
Taxa de registros válidos	TRV	Proporção de registros que passam em todas as regras críticas
Taxa de regras aprovadas	TRA	Proporção de regras de qualidade aprovadas
Taxa de regras violadas	TRVi	Proporção de regras de qualidade violadas
Índice geral de qualidade	IGQ	Indicador composto de qualidade do dataset
Severidade média das não conformidades	SMNC	Grau médio de severidade dos problemas encontrados
Taxa de ativos aprovados	TAA	Proporção de ativos aprovados para uso analítico

9.1 Taxa de Completude — TC

A **Taxa de Completude** mede a proporção de valores presentes em relação ao total esperado.

TC = número de valores preenchidos / número total de valores esperados

Essa métrica pode ser calculada por atributo, por registro, por dataset ou por grupo.

9.2 Taxa de Validade — TV

A **Taxa de Validade** mede a proporção de valores que respeitam domínios, formatos, tipos ou regras esperadas.

TV = número de valores válidos / número total de valores avaliados

Valores inválidos podem incluir datas fora do intervalo esperado, códigos inexistentes, categorias não permitidas ou formatos incompatíveis.

9.3 Taxa de Integridade Referencial — TIR

A **Taxa de Integridade Referencial** avalia a proporção de relações entre chaves que estão corretamente preservadas.

TIR = número de relações válidas / número total de relações avaliadas

Essa métrica é essencial para datasets integrados ou relacionais.

9.4 Taxa de Regras Aprovadas — TRA

A **Taxa de Regras Aprovadas** mede a proporção de regras de validação que foram atendidas.

TRA = número de regras aprovadas / número total de regras executadas

Essa métrica permite avaliar rapidamente o grau de conformidade do dataset.

9.5 Índice Geral de Qualidade — IGQ

O **Índice Geral de Qualidade** pode ser definido como uma composição ponderada das dimensões avaliadas.

Uma formulação possível é:

$$\mathbf{IGQ(d)} = \mathbf{w_1TC} + \mathbf{w_2TV} + \mathbf{w_3TIR} + \mathbf{w_4TU} + \mathbf{w_5TA} + \mathbf{w_6(1 - TI)}$$

em que:

- **TC** representa completude;
- **TV** representa validade;
- **TIR** representa integridade referencial;
- **TU** representa unicidade;
- **TA** representa atualidade;
- **TI** representa inconsistência;
- w_i representa o peso atribuído a cada dimensão.

Os pesos podem variar conforme o domínio, a finalidade e o risco do uso analítico.

10. Artefatos formalizados do módulo

Os artefatos do Módulo 5 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, regras de construção, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição

A5.1	Modelo de Dimensões de Qualidade	Documento/ taxonomia	Cobertura de dimensões	Define dimensões avaliáveis
A5.2	Plano de Validação de Dados	Documento técnico	Taxa de regras definidas	Formaliza critérios de validação
A5.3	Catálogo de Regras de Qualidade	Tabela estruturada	TRA, TRVi	Documenta regras aplicáveis
A5.4	Registro de Testes de Qualidade	Log ou tabela	TRA	Registra execução dos testes
A5.5	Relatório de Qualidade de Dados	Relatório técnico	IGQ	Consolida resultados da avaliação
A5.6	Registro de Não Conformidades	Tabela técnica	SMNC	Documenta problemas encontrados
A5.7	Matriz de Severidade e Impacto	Matriz de risco	Severidade média	Prioriza problemas de qualidade
A5.8	Registro de Ações Corretivas	Tabela de acompanhamento	Taxa de correção	Acompanha tratamento dos problemas
A5.9	Selo ou Status de Qualidade do Ativo	Classificação	TAA	Indica condição de uso do ativo
A5.10	Histórico de Qualidade	Série temporal ou tabela	Evolução do IGQ	Permite acompanhar qualidade ao longo

do tempo

11. Artefato A5.1 — Modelo de Dimensões de Qualidade

11.1 Finalidade

O **Modelo de Dimensões de Qualidade** define quais dimensões serão utilizadas para avaliar os dados no UnespDataLens-RM.

11.2 Tipo de artefato na DSR

Na perspectiva da Design Science Research, esse artefato é um **constructo** e um **modelo**, pois define os conceitos centrais utilizados para avaliar a qualidade dos dados.

11.3 Estrutura mínima

Dimensão	Definição	Métrica associada	Aplicação	Exemplo
Completude	Presença de valores esperados	TC, TVA	Atributos e registros	Campo renda preenchido
Validade	Conformidade com domínio	TV	Atributos	Código de curso válido
Consistência	Ausência de contradições	TI	Registros e fontes	Data de saída posterior à entrada
Unicidade	Ausência de duplicidade	TD, TU	Chaves e registros	Um RA por estudante
Integridade	Preservação de relações	TIR	Bases relacionais	Matrícula associada a curso existente
Atualidade	Dados atualizados	TA	Fontes periódicas	Base atualizada no mês esperado

Conformidade	Aderência a schema/contrato	TCS	Datasets	Campo com tipo esperado

11.4 Métricas associadas

- número de dimensões definidas;
- percentual de dimensões aplicadas por dataset;
- cobertura das dimensões de qualidade;
- percentual de dimensões com métricas associadas.

11.5 Contribuição

Esse artefato padroniza a avaliação da qualidade no modelo de referência, evitando que cada projeto utilize critérios diferentes e pouco comparáveis.

12. Artefato A5.2 — Plano de Validação de Dados

12.1 Finalidade

O **Plano de Validação de Dados** define como um dataset ou ativo analítico será avaliado quanto à qualidade.

12.2 Tipo de artefato na DSR

Esse artefato é um **método**, pois orienta o processo de validação, e uma **instanciação**, quando implementado em documento técnico, planilha, arquivo YAML, JSON, notebook ou ferramenta de validação.

12.3 Estrutura mínima

Campo	Descrição
Identificador do plano	Código único
Dataset avaliado	Ativo submetido à validação
Finalidade do dataset	Uso pretendido
Dimensões avaliadas	Completude, validade, consistência etc.

Regras aplicadas	Regras de qualidade
Limiar de aceitação	Valor mínimo aceitável
Severidade	Baixa, média, alta, crítica
Responsável	Pessoa ou equipe
Periodicidade	Frequência da validação
Critério de aprovação	Condição para liberar o ativo
Ações recomendadas	Correção, rejeição, reprocessamento etc.

12.4 Métricas associadas

- percentual de datasets com plano de validação;
- percentual de planos com limiares definidos;
- percentual de planos com critérios de aprovação;
- percentual de planos com responsáveis;
- taxa de regras executáveis.

12.5 Critérios de validação

O plano é considerado válido quando define dimensões, regras, limiares, severidades, responsáveis e critérios de aprovação.

12.6 Contribuição

Esse artefato transforma a validação dos dados em um processo explícito, padronizado e reprodutível.

13. Artefato A5.3 — Catálogo de Regras de Qualidade

13.1 Finalidade

O **Catálogo de Regras de Qualidade** registra as regras utilizadas para avaliar os datasets.

13.2 Estrutura formal

Cada regra de qualidade pode ser representada por:

r_i = {id, dimensão, dataset, atributo, condição, limiar, severidade, mensagem, ação_recomendada}

13.3 Estrutura mínima

Campo	Descrição
Identificador da regra	Código único
Dimensão de qualidade	Completude, validade, consistência etc.
Dataset	Ativo avaliado
Atributo	Campo avaliado
Condição	Regra lógica ou estatística
Limiar	Valor mínimo ou máximo aceitável
Severidade	Baixa, média, alta ou crítica
Justificativa	Motivo da regra
Ação recomendada	Correção, revisão, bloqueio, alerta
Status	Ativa, inativa, em revisão

13.4 Métricas associadas

- número de regras cadastradas;
- taxa de regras aprovadas;
- taxa de regras violadas;

- percentual de regras críticas;
- percentual de regras com limiar definido;
- percentual de regras com ação recomendada.

13.5 Critérios de validação

Uma regra é considerada válida quando possui dimensão, condição, limiar, severidade e ação recomendada documentados.

13.6 Contribuição

Esse artefato formaliza o conhecimento de validação, permitindo reutilização, auditoria e automação dos testes de qualidade.

14. Artefato A5.4 — Registro de Testes de Qualidade

14.1 Finalidade

O **Registro de Testes de Qualidade** documenta a execução das regras de validação sobre os datasets.

14.2 Estrutura mínima

Campo	Descrição
Execução	Identificador da execução
Dataset	Dataset avaliado
Regra aplicada	Regra de qualidade
Data da execução	Momento da validação
Resultado	Aprovado, violado, alerta
Valor obtido	Resultado calculado
Limiar esperado	Valor de referência

Registros afetados	Quantidade ou percentual
Exemplos	Amostra de registros problemáticos
Responsável	Executor humano ou sistema

14.3 Métricas associadas

- taxa de regras aprovadas;
- taxa de regras violadas;
- número de registros afetados;
- percentual de execuções documentadas;
- taxa de testes automatizados.

14.4 Contribuição

Esse artefato cria evidências auditáveis da validação dos dados, permitindo rastrear quando, como e com quais resultados cada teste foi executado.

15. Artefato A5.5 — Relatório de Qualidade de Dados

15.1 Finalidade

O **Relatório de Qualidade de Dados** consolida os resultados das avaliações realizadas sobre um dataset ou conjunto de ativos.

15.2 Estrutura mínima

Campo	Descrição
Dataset avaliado	Ativo analisado
Finalidade	Uso pretendido
Dimensões avaliadas	Completude, validade, consistência etc.

Métricas obtidas	Resultados calculados
Regras aprovadas	Quantidade ou percentual
Regras violadas	Quantidade ou percentual
Não conformidades	Problemas encontrados
Severidade	Classificação dos problemas
Recomendações	Ações sugeridas
Status final	Aprovado, aprovado com ressalvas, rejeitado, pendente

15.3 Métricas associadas

- índice geral de qualidade;
- taxa de regras aprovadas;
- taxa de registros válidos;
- severidade média das não conformidades;
- taxa de ativos aprovados.

15.4 Critérios de validação

O relatório é considerado completo quando apresenta dimensões avaliadas, regras aplicadas, métricas, não conformidades, severidade, recomendações e status final.

15.5 Contribuição

Esse artefato sintetiza a condição de qualidade do ativo e apoia decisões de uso, correção, rejeição ou disponibilização.

16. Artefato A5.6 — Registro de Não Conformidades

16.1 Finalidade

O **Registro de Não Conformidades** documenta problemas identificados durante a validação de qualidade.

16.2 Estrutura mínima

Campo	Descrição
Identificador	Código da não conformidade
Dataset	Ativo afetado
Dimensão	Dimensão de qualidade afetada
Regra violada	Regra não atendida
Descrição	Descrição do problema
Registros afetados	Quantidade ou percentual
Severidade	Baixa, média, alta ou crítica
Origem provável	Fonte, ingestão, integração, transformação etc.
Impacto	Consequência analítica potencial
Ação recomendada	Correção, alerta, bloqueio, revisão
Status	Aberta, em tratamento, resolvida, aceita

16.3 Métricas associadas

- número de não conformidades;
- severidade média;
- percentual de não conformidades críticas;
- percentual de não conformidades resolvidas;

- tempo médio de resolução.

16.4 Contribuição

Esse artefato permite priorizar problemas de qualidade e acompanhar seu tratamento ao longo do pipeline.

17. Artefato A5.7 — Matriz de Severidade e Impacto

17.1 Finalidade

A **Matriz de Severidade e Impacto** classifica problemas de qualidade conforme sua gravidade e consequência para o uso analítico.

17.2 Estrutura mínima

Severidade	Descrição	Exemplo	Ação recomendada
Baixa	Problema localizado e de baixo impacto	Pequena variação de formato	Corrigir ou documentar
Média	Problema relevante, mas controlável	Ausência moderada em atributo secundário	Corrigir antes de uso crítico
Alta	Problema com impacto analítico significativo	Chave incompleta em integração	Reprocessar ou restringir uso
Crítica	Problema que inviabiliza o uso	Variável-alvo inconsistente	Bloquear ou rejeitar ativo

17.3 Métricas associadas

- severidade média;
- número de problemas críticos;
- proporção de problemas por nível;
- taxa de ativos bloqueados por severidade;
- impacto potencial por dimensão.

17.4 Contribuição

Esse artefato evita que todos os problemas sejam tratados da mesma forma, permitindo priorização baseada em risco.

18. Artefato A5.8 — Registro de Ações Corretivas

18.1 Finalidade

O **Registro de Ações Corretivas** acompanha as ações propostas e executadas para corrigir problemas de qualidade.

18.2 Estrutura mínima

Campo	Descrição
Não conformidade associada	Problema que originou a ação
Ação proposta	Correção, reprocessamento, revisão etc.
Módulo responsável	Módulo em que a correção deve ocorrer
Responsável	Pessoa ou equipe
Prazo	Data prevista
Status	Planejada, em execução, concluída
Resultado	Efeito da ação
Evidência	Log, manifesto, relatório ou dataset corrigido

18.3 Métricas associadas

- taxa de ações concluídas;
- tempo médio de correção;
- percentual de problemas reabertos;
- taxa de redução de não conformidades;
- impacto das ações sobre o IGQ.

18.4 Contribuição

Esse artefato conecta avaliação de qualidade a melhoria efetiva dos dados, evitando que relatórios de qualidade não produzam ação concreta.

19. Artefato A5.9 — Selo ou Status de Qualidade do Ativo

19.1 Finalidade

O **Selo ou Status de Qualidade do Ativo** classifica o dataset quanto à sua condição de uso.

19.2 Estrutura mínima

Status	Interpretação	Condição de uso
Aprovado	Atende aos critérios mínimos	Liberado para uso
Aprovado com ressalvas	Possui limitações documentadas	Uso permitido com cautela
Pendente	Necessita correção ou revisão	Uso restrito
Rejeitado	Não atende critérios mínimos	Uso não recomendado
Bloqueado	Risco crítico identificado	Uso proibido até correção

19.3 Métricas associadas

- taxa de ativos aprovados;
- taxa de ativos aprovados com ressalvas;
- taxa de ativos rejeitados;
- taxa de ativos pendentes;
- taxa de ativos bloqueados.

19.4 Contribuição

Esse artefato traduz métricas complexas de qualidade em uma decisão operacional clara sobre uso do ativo.

20. Artefato A5.10 — Histórico de Qualidade

20.1 Finalidade

O **Histórico de Qualidade** registra a evolução das métricas de qualidade ao longo do tempo.

20.2 Estrutura mínima

Campo	Descrição
Dataset	Ativo monitorado
Data da avaliação	Momento da validação
Versão	Versão do dataset
Métricas	Indicadores de qualidade
Status	Classificação do ativo
Não conformidades	Problemas registrados
Ações corretivas	Ações associadas
Observações	Mudanças relevantes

20.3 Métricas associadas

- evolução do IGQ;
- variação da completude;
- variação da validade;
- variação da consistência;
- número de problemas recorrentes;
- tendência de melhoria ou degradação.

20.4 Contribuição

Esse artefato permite monitorar a qualidade dos dados ao longo do tempo, apoiando observabilidade, governança e melhoria contínua.

21. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Datasets brutos, integrados e tratados; dicionários; schemas; regras de negócio; contratos de dados; manifestos; logs; registros de transformação e metadados.
Processamento	Definição de dimensões, regras e limiares; execução de testes; cálculo de métricas; registro de não conformidades; classificação de severidade; recomendação de ações.
Artefatos	Modelo de dimensões, plano de validação, catálogo de regras, registros de testes, relatório de qualidade, não conformidades, matriz de severidade, ações corretivas, status de qualidade e histórico.
Métricas	TC, TVA, TV, TI, TD, TU, TIR, TCS, TA, TRV, TRA, TRVi, IGQ, SMNC e TAA.
Saídas	Ativos avaliados, relatórios de qualidade, status de aprovação, não conformidades documentadas e recomendações de correção ou uso.

22. Critérios gerais de validação do módulo

O Módulo 5 pode ser considerado concluído quando:

1. as dimensões de qualidade foram definidas para os ativos avaliados;
2. os datasets possuem plano de validação;
3. as regras de qualidade estão documentadas;
4. os testes foram executados e registrados;
5. as métricas de qualidade foram calculadas;
6. as não conformidades foram identificadas e classificadas;
7. problemas críticos possuem recomendação de ação;
8. os relatórios de qualidade foram produzidos;
9. o status de qualidade dos ativos foi definido;
10. os resultados podem ser utilizados pelos módulos posteriores;
11. o histórico de qualidade foi preservado para monitoramento futuro.

23. Relação com os demais módulos do modelo de referência

O Módulo 5 depende do **Módulo 1 — Fontes e Inventário de Dados**, pois utiliza informações sobre origem, estrutura, documentação, granularidade e limitações conhecidas das fontes.

Depende do **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, pois utiliza logs, manifestos, schemas iniciais, registros de integridade e datasets brutos para avaliar qualidade desde a entrada dos dados.

Depende do **Módulo 3 — Integração e Harmonização**, pois utiliza métricas de pareamento, registros de conflitos, chaves, entidades e regras de integração.

Depende do **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, pois utiliza registros de transformação, atributos derivados, datasets tratados e métricas antes/depois.

Para o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, fornece indicadores de qualidade que ajudam a identificar problemas que podem gerar vieses ou afetar grupos específicos.

Para o **Módulo 7 — Metadados, Catálogo e Semântica dos Dados**, fornece métricas, status de qualidade, regras e relatórios que devem ser incorporados ao catálogo.

Para o **Módulo 8 — Governança, Segurança, Privacidade e Conformidade**, fornece evidências sobre confiabilidade, restrições, problemas críticos e status de uso dos ativos.

Para o **Módulo 9 — Proveniência, Lineage e Rastreabilidade**, fornece registros de validação e não conformidades associados a cada etapa do pipeline.

Para o **Módulo 10 — Versionamento e Reprodutibilidade**, fornece histórico de qualidade por versão dos datasets.

Para o **Módulo 12 — Feature Engineering e Feature Store**, fornece critérios para selecionar atributos confiáveis e documentar qualidade das features.

Para o **Módulo 13 — Disponibilização Analítica**, define se o ativo pode ser disponibilizado, disponibilizado com ressalvas, corrigido ou bloqueado.

Para o **Módulo 14 — Monitoramento, Drift e Evolução dos Ativos**, fornece baseline de qualidade para acompanhamento contínuo.

24. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 5 poderia avaliar a qualidade de uma base integrada contendo dados de matrícula, notas, frequência, situação acadêmica, dados socioeconômicos e interações em ambiente virtual.

As regras poderiam verificar se todos os estudantes possuem identificador válido, se códigos de curso existem no catálogo institucional, se datas de matrícula são compatíveis com o período letivo, se notas estão no intervalo permitido, se frequência está entre 0% e

100%, se a situação acadêmica pertence a categorias válidas e se há inconsistências entre matrícula ativa e ausência total de registros acadêmicos.

As métricas poderiam indicar taxa de completude dos dados socioeconômicos, taxa de integridade entre estudante e curso, taxa de valores inválidos em notas, duplicidade de matrículas, consistência temporal da trajetória acadêmica e status geral da base para uso em modelos de evasão ou dashboards de acompanhamento.

25. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, o Módulo 5 poderia avaliar bases integradas de saúde, educação, assistência social e território.

As regras poderiam verificar se códigos territoriais são válidos, se registros de atendimento possuem data, se unidades públicas existem no cadastro oficial, se famílias não aparecem duplicadas em uma mesma base, se indicadores agregados respeitam intervalos plausíveis e se a periodicidade de atualização está adequada.

O relatório de qualidade poderia classificar determinados ativos como aprovados para uso em painéis públicos, aprovados com ressalvas para análise interna ou bloqueados para uso decisório devido a problemas críticos de consistência ou completude.

26. Contribuição técnico-científica do módulo

A contribuição do Módulo 5 está em formalizar a qualidade dos dados como um processo mensurável, auditável, reprodutível e integrado ao ciclo de Engenharia Analítica de Dados.

Em vez de tratar qualidade como uma percepção subjetiva ou uma inspeção manual isolada, o UnespDataLens-RM propõe que cada ativo seja avaliado por dimensões, regras, métricas, testes, relatórios, não conformidades, critérios de aprovação e histórico de qualidade.

Essa formalização contribui para:

- tornar explícitos os critérios de qualidade;
- avaliar objetivamente os ativos de dados;
- identificar problemas antes da modelagem ou disponibilização;
- reduzir risco de decisões baseadas em dados inadequados;
- apoiar governança e auditoria;
- documentar não conformidades e ações corretivas;
- acompanhar evolução da qualidade ao longo do tempo;
- integrar qualidade com metadados, lineage, versionamento e monitoramento;
- definir status de uso dos ativos analíticos.

27. Síntese do módulo

O **Módulo 5 — Qualidade e Validação de Dados** avalia a confiabilidade dos dados produzidos no UnespDataLens-RM por meio de dimensões, regras, métricas, testes, relatórios e critérios de aceitação. Ele transforma a qualidade dos dados em um componente formal, documentado, mensurável e auditável do modelo de referência.

Sua importância é central porque ativos analíticos só devem ser utilizados em indicadores, dashboards, modelos, grafos ou decisões se sua qualidade for conhecida, avaliada e documentada. Ao formalizar a validação dos dados, o módulo fortalece a confiabilidade, a transparência, a governança e a reprodutibilidade dos resultados produzidos pelo modelo de referência.

Dessa forma, o Módulo 5 cria a base para que o UnespDataLens-RM produza dados não apenas integrados e tratados, mas também avaliados, validados e adequados ao uso analítico pretendido.

Módulo 6 — Desbalanceamento, Vieses e Equidade

1. Finalidade do módulo

O **Módulo 6 — Desbalanceamento, Vieses e Equidade** tem como finalidade identificar, mensurar, documentar, mitigar e monitorar problemas relacionados à distribuição desigual dos dados, à representatividade insuficiente de grupos, à presença de vieses técnicos ou estatísticos e a possíveis impactos desproporcionais em análises, indicadores, modelos preditivos e processos de decisão baseados em dados.

Esse módulo atua sobre os dados produzidos nos módulos anteriores do **UnespDataLens-RM**, especialmente os dados inventariados, ingeridos, integrados, harmonizados, transformados, enriquecidos e avaliados quanto à qualidade. Sua função é verificar se os dados e os ativos analíticos produzidos representam adequadamente os grupos, classes, categorias, territórios, períodos, eventos ou entidades relevantes para o problema analisado.

Em projetos de Engenharia Analítica de Dados, problemas de desbalanceamento e viés podem surgir em diferentes etapas do pipeline. Eles podem estar presentes na própria fonte de dados, ser ampliados durante a ingestão, surgir na integração entre bases, ser introduzidos por exclusão de registros, imputação, transformação, criação de variáveis, seleção de atributos, definição da variável-alvo ou escolha de amostras para treinamento e avaliação de modelos.

No contexto do UnespDataLens-RM, o Módulo 6 tem papel estratégico porque impede que dados tecnicamente bem organizados, limpos e validados sejam utilizados sem análise crítica de representatividade, equidade e impacto. Um dataset pode apresentar boa completude, consistência e validade, mas ainda assim carregar distorções que afetam grupos específicos, classes minoritárias, territórios menos representados, períodos subobservados ou categorias relevantes para decisões públicas, educacionais ou institucionais.

Assim, o Módulo 6 formaliza desbalanceamento, vieses e equidade como dimensões técnico-metodológicas avaliáveis da Engenharia Analítica de Dados. Ele produz artefatos, métricas, relatórios e critérios de validação capazes de tornar explícitos os riscos de uso dos dados em contextos analíticos e decisórios.

2. Objetivo geral

O objetivo geral do Módulo 6 é **avaliar e tratar desbalanceamentos, vieses e problemas de equidade em dados analíticos**, por meio de métricas, regras, artefatos, métodos de mitigação e critérios de validação que permitam mensurar representatividade, distribuição de classes, impacto sobre grupos, efeitos de transformação e adequação dos dados para análises, modelos e decisões.

3. Objetivos específicos

São objetivos específicos do Módulo 6:

1. identificar desbalanceamentos em classes, grupos, categorias, períodos, territórios, entidades e variáveis-alvo;
2. avaliar a representatividade dos dados em relação ao fenômeno, população, domínio ou problema analítico;
3. identificar grupos sub-representados, classes minoritárias, categorias raras e padrões de exclusão ou ausência;
4. mensurar vieses técnicos ou estatísticos decorrentes de coleta, integração, transformação, imputação, amostragem, rotulagem, seleção de atributos ou definição da variável-alvo;
5. avaliar possíveis impactos desproporcionais sobre grupos sensíveis ou grupos analiticamente relevantes;
6. aplicar estratégias de mitigação de desbalanceamento, quando apropriado, preservando rastreabilidade e justificativa metodológica;
7. registrar efeitos das ações de balanceamento ou mitigação sobre distribuições, métricas de desempenho, variáveis e grupos;
8. produzir relatórios de equidade e representatividade para orientar decisões sobre uso, restrição, reprocessamento ou disponibilização dos dados;
9. fornecer insumos para módulos de metadados, governança, lineage, versionamento, feature engineering, modelagem, monitoramento e disponibilização analítica;
10. formalizar artefatos avaliáveis que permitam documentar e auditar decisões relacionadas a desbalanceamento, vieses e equidade.

4. Fundamentação conceitual do módulo

O Módulo 6 parte da compreensão de que dados não são neutros. Eles são produzidos por processos institucionais, sociais, técnicos e operacionais que podem refletir desigualdades, lacunas, restrições de coleta, regras administrativas, erros de registro, mudanças históricas, diferentes níveis de acesso a serviços ou assimetrias na representação de grupos.

O **desbalanceamento** ocorre quando determinadas classes, categorias ou grupos aparecem em proporções muito diferentes no conjunto de dados. Em problemas preditivos, isso pode fazer com que modelos aprendam melhor padrões associados às classes majoritárias e tenham baixo desempenho para classes minoritárias. Em análises descritivas ou painéis, o desbalanceamento pode levar a interpretações distorcidas se a distribuição observada não representar adequadamente o fenômeno.

O **viés** pode ser entendido como uma distorção sistemática nos dados, nos processos de preparação ou nos modelos que afeta a representação, a mensuração ou o tratamento de grupos, entidades ou fenômenos. No UnespDataLens-RM, o foco recai principalmente sobre vieses técnicos, estatísticos e algorítmicos associados ao ciclo de Engenharia Analítica de Dados, sem desconsiderar que esses vieses podem refletir processos sociais e institucionais mais amplos.

A **equidade** refere-se à preocupação com impactos diferenciados sobre grupos e à necessidade de avaliar se dados, indicadores ou modelos produzem resultados injustamente assimétricos, especialmente quando utilizados para apoiar decisões públicas, educacionais, administrativas ou institucionais. No contexto do modelo de referência, equidade não significa forçar igualdade artificial entre todos os grupos, mas tornar explícitas as diferenças de representação, desempenho, erro, cobertura e impacto, permitindo decisões analíticas mais responsáveis.

O Módulo 6 considera diferentes fontes de viés no pipeline:

Fonte de viés	Descrição	Exemplo
Viés de coleta	A fonte não representa adequadamente a população ou fenômeno.	Dados de estudantes com maior acesso ao AVA são mais completos.
Viés de cobertura	Certos grupos, territórios ou períodos aparecem pouco ou não aparecem.	Municípios menores têm menos registros administrativos.
Viés de integração	A junção entre bases perde registros de determinados grupos.	Estudantes sem chave válida não são pareados.
Viés de transformação	Limpezas, imputações ou filtros afetam grupos de forma desigual.	Remoção de registros incompletos exclui mais alunos vulneráveis.
Viés de rotulagem	A variável-alvo é construída com critérios incompletos ou inconsistentes.	Definir evasão apenas por ausência de matrícula no semestre seguinte.
Viés de seleção de atributos	Atributos usados reforçam desigualdades ou proxies sensíveis.	CEP como proxy de condição socioeconômica.
Viés de avaliação	Métricas globais escondem desempenho desigual entre grupos.	Acurácia alta com baixo recall para classe minoritária.
Viés de uso	Dados ou modelos são aplicados fora do contexto adequado.	Modelo treinado em um campus usado em outro sem validação.

À luz da **Design Science Research**, este módulo formaliza artefatos que permitem transformar a análise de desbalanceamento, vieses e equidade em um processo técnico-metodológico explícito, mensurável, auditável e reprodutível.

5. Formalização sintética do módulo

O Módulo 6 pode ser compreendido como um processo de avaliação de distribuições, grupos e impactos sobre ativos de dados produzidos pelos módulos anteriores.

Considerando um conjunto de datasets ou ativos analíticos:

$$\mathbf{D = \{d_1, d_2, ..., d_n\}}$$

e um conjunto de grupos, classes ou categorias relevantes:

$$\mathbf{G = \{g_1, g_2, ..., g_k\}}$$

o módulo avalia a distribuição dos registros, atributos, rótulos, erros, métricas ou decisões em relação a esses grupos.

Cada grupo $\mathbf{g_i}$ pode ser descrito por:

$$\mathbf{g_i = \{id, nome, atributo_associado, tipo, frequência, proporção, sensibilidade, relevância_analítica\}}$$

Cada avaliação de desbalanceamento ou equidade pode ser representada por:

$$\mathbf{e_i = \{dataset, grupo, variável_alvo, métrica, valor_obtido, limiar, severidade, recomendação\}}$$

Quando houver aplicação de uma técnica de mitigação, ela pode ser representada por:

$$\mathbf{m_i = \{id, dataset, problema_identificado, técnica, parâmetros, grupos_afetados, distribuição_antes, distribuição_depois, impacto, justificativa, versão\}}$$

O resultado do módulo é um conjunto de evidências, relatórios, métricas e datasets ajustados ou sinalizados para uso com ressalvas.

Dessa forma, desbalanceamento, viés e equidade deixam de ser tratados como preocupações genéricas e passam a ser descritos por artefatos formais, métricas e critérios de decisão.

6. Entradas do módulo

As entradas do Módulo 6 incluem datasets, metadados, variáveis, atributos sensíveis, registros de transformação e resultados de qualidade produzidos nos módulos anteriores.

As principais entradas são:

- datasets brutos;
- datasets integrados e harmonizados;
- datasets tratados e enriquecidos;
- relatórios de qualidade;
- registros de valores ausentes;
- registros de duplicidade;
- registros de outliers;
- registros de transformação;
- atributos derivados;
- dicionários de dados;
- classificação de sensibilidade;
- variáveis-alvo candidatas;

- atributos sensíveis ou protegidos;
- atributos socioeconômicos;
- atributos territoriais;
- atributos demográficos;
- atributos temporais;
- grupos analiticamente relevantes;
- regras de negócio;
- documentação sobre origem dos dados;
- registros de integração e pareamento;
- métricas de qualidade por grupo;
- dados de referência populacional, quando disponíveis.

7. Processamento

O processamento do Módulo 6 envolve atividades de identificação de grupos, análise de distribuição, mensuração de desbalanceamento, avaliação de vieses, aplicação de técnicas de mitigação, análise de impacto e documentação das decisões.

7.1 Identificação de variáveis sensíveis e grupos relevantes

A primeira atividade consiste em identificar os atributos que podem definir grupos sensíveis, protegidos ou analiticamente relevantes. Esses atributos podem incluir sexo, raça/cor, idade, deficiência, território, renda, escola de origem, curso, turno, modalidade, município, grupo socioeconômico ou outras categorias relevantes ao domínio.

A identificação deve considerar tanto atributos explicitamente sensíveis quanto atributos que possam funcionar como proxies de condições sociais, territoriais ou institucionais.

7.2 Análise de distribuição dos dados

A segunda atividade consiste em avaliar a distribuição dos registros entre classes, categorias, grupos, períodos e territórios. Essa análise permite identificar concentração excessiva, sub-representação, lacunas, categorias raras, classes minoritárias e desequilíbrios que possam comprometer análises ou modelos.

7.3 Avaliação de desbalanceamento da variável-alvo

Quando houver variável-alvo, como evasão, aprovação, risco, ocorrência, atendimento ou outra classe de interesse, o módulo deve avaliar sua distribuição. Essa avaliação é especialmente importante em problemas preditivos, nos quais classes minoritárias podem ser menos aprendidas pelos modelos.

7.4 Avaliação de representatividade

A quarta atividade consiste em comparar a distribuição observada nos dados com distribuições esperadas, quando houver referência disponível. Essa referência pode vir de censos, dados institucionais oficiais, estatísticas públicas, registros administrativos consolidados ou conhecimento de especialistas.

Quando não houver referência externa, a análise deve registrar essa limitação e avaliar representatividade de forma interna, considerando cobertura, períodos, grupos e fontes.

7.5 Avaliação de vieses introduzidos pelo pipeline

O módulo deve avaliar se etapas anteriores introduziram ou ampliaram vieses. Isso inclui verificar se a ingestão excluiu determinados registros, se a integração perdeu grupos por ausência de chaves, se a limpeza removeu registros de forma desigual, se a imputação afetou grupos específicos ou se a criação de variáveis derivadas gerou distorções.

7.6 Mensuração de métricas de equidade

Quando houver modelos, classificações, rankings, indicadores ou decisões baseadas em dados, o módulo deve calcular métricas de equidade entre grupos. Essas métricas podem avaliar diferenças de seleção, desempenho, erro, falsos positivos, falsos negativos, oportunidades ou impacto.

7.7 Aplicação de técnicas de mitigação

Quando identificado desbalanceamento ou viés relevante, podem ser aplicadas técnicas de mitigação. Essas técnicas podem atuar antes da modelagem, durante a modelagem ou após a modelagem. No escopo da Engenharia Analítica de Dados, o foco principal recai sobre ações pré-processamento e documentação dos efeitos.

As ações podem incluir reamostragem, ponderação, ajuste de limiares, estratificação, reavaliação de variáveis, revisão de rótulos, inclusão de fontes complementares, análise por subgrupos ou recomendação de restrição de uso.

7.8 Avaliação do impacto da mitigação

Toda ação de mitigação deve ser avaliada quanto ao impacto produzido. Isso inclui comparar distribuições antes e depois, verificar efeitos sobre métricas de desempenho, analisar perdas de informação, avaliar alterações em grupos e documentar riscos residuais.

7.9 Registro de decisões e limitações

O módulo deve registrar decisões tomadas, técnicas aplicadas, parâmetros, justificativas e limitações. Caso um problema de viés ou desbalanceamento não possa ser resolvido no pipeline, ele deve ser explicitamente documentado como limitação de uso.

8. Técnicas, métodos e algoritmos aplicáveis

O Módulo 6 pode empregar diferentes técnicas, métodos e algoritmos para análise e mitigação de desbalanceamento, vieses e equidade.

Entre as técnicas de diagnóstico, destacam-se:

- análise de distribuição de classes;
- análise de distribuição por grupo;
- análise de representatividade;
- análise de subgrupos;
- análise de cobertura;
- análise temporal de distribuição;
- análise territorial de distribuição;
- análise de perdas por etapa do pipeline;
- comparação antes/depois de transformações;
- análise da variável-alvo;
- análise de proxies;

- avaliação de métricas por grupo.

Entre as técnicas de tratamento de desbalanceamento, podem ser utilizadas:

- subamostragem da classe majoritária;
- sobreamostragem da classe minoritária;
- Random Under Sampling;
- Random Over Sampling;
- SMOTE;
- Borderline-SMOTE;
- ADASYN;
- Tomek Links;
- Edited Nearest Neighbors;
- SMOTEENN;
- SMOTETomek;
- class weights;
- cost-sensitive learning;
- estratificação;
- ajuste de limiar;
- amostragem temporal ou por grupo.

Entre as técnicas de mitigação de viés e fairness, destacam-se:

- reweighing;
- disparate impact remover;
- learning fair representations;
- adversarial debiasing, quando aplicável;
- equalized odds post-processing;
- calibrated equalized odds;
- reject option classification;
- análise contrafactual, quando viável;
- avaliação por subgrupos;
- restrição de atributos sensíveis;
- documentação de uso responsável;
- revisão de rótulos;
- validação com especialistas.

Entre os algoritmos e recursos computacionais aplicáveis, destacam-se:

- métricas de distribuição;
- análise de frequências;
- testes de diferença entre grupos;
- qui-quadrado;
- Kolmogorov-Smirnov;
- Population Stability Index;
- divergência de Kullback-Leibler;
- Jensen-Shannon divergence;
- árvores de decisão para identificação de subgrupos;
- modelos de classificação avaliados por grupo;
- matrizes de confusão estratificadas;
- curvas ROC e PR por grupo;
- análise de erro por grupo.

Ferramentas possíveis incluem Python, pandas, scikit-learn, imbalanced-learn, AIF360, Fairlearn, Evidently, notebooks versionados, ferramentas de profiling e bibliotecas estatísticas.

9. Métricas do módulo

As métricas do Módulo 6 têm como objetivo avaliar desbalanceamento, representatividade, viés, equidade, impacto de mitigação e risco residual.

Métrica	Sigla	Descrição
Razão de desbalanceamento	IR	Relação entre classe majoritária e classe minoritária
Proporção da classe minoritária	PCM	Percentual de registros da classe minoritária
Distribuição por grupo	DG	Proporção de registros por grupo relevante
Taxa de sub-representação	TSR	Grau em que um grupo aparece abaixo do esperado
Taxa de perda por grupo	TPG	Proporção de registros perdidos por grupo ao longo do pipeline
Diferença de completude por grupo	DCG	Diferença de completude entre grupos
Diferença de qualidade por grupo	DQG	Diferença de métricas de qualidade entre grupos
Statistical Parity Difference	SPD	Diferença na taxa de resultado positivo entre grupos
Disparate Impact	DI	Razão entre taxas de resultado positivo entre grupos

Equal Opportunity Difference	EOD	Diferença de true positive rate entre grupos
Average Odds Difference	AOD	Diferença média entre FPR e TPR entre grupos
Diferença de FPR	ΔFPR	Diferença de falsos positivos entre grupos
Diferença de FNR	ΔFNR	Diferença de falsos negativos entre grupos
Balanced Accuracy	BA	Média do desempenho entre classes
Macro-F1	MF1	Média do F1-score entre classes
PR-AUC	PRAUC	Área sob curva precisão-revocação
Índice de impacto da mitigação	IIM	Grau de alteração produzido pela mitigação
Risco residual de viés	RRV	Nível de risco após ações de mitigação

9.1 Razão de Desbalanceamento — IR

A **Razão de Desbalanceamento** mede a proporção entre a classe majoritária e a classe minoritária.

IR = número de registros da classe majoritária / número de registros da classe minoritária

Valores elevados indicam forte desbalanceamento e possível necessidade de tratamento, especialmente em problemas preditivos.

9.2 Taxa de Perda por Grupo — TPG

A **Taxa de Perda por Grupo** mede a proporção de registros de determinado grupo removidos ou perdidos ao longo de etapas como integração, limpeza ou transformação.

$TPG(g_i)$ = registros removidos do grupo g_i / registros iniciais do grupo g_i

Essa métrica permite identificar se o pipeline afeta grupos de forma desigual.

9.3 Statistical Parity Difference — SPD

O **Statistical Parity Difference** mede a diferença entre a taxa de resultado positivo de um grupo protegido e a taxa de resultado positivo de um grupo de referência.

SPD = P(ŷ = 1 | grupo protegido) − P(ŷ = 1 | grupo referência)

Valores próximos de zero indicam maior paridade estatística, embora a interpretação dependa do domínio e do problema.

9.4 Disparate Impact — DI

O **Disparate Impact** mede a razão entre a taxa de resultado positivo de um grupo protegido e a taxa de resultado positivo de um grupo de referência.

DI = P(ŷ = 1 | grupo protegido) / P(ŷ = 1 | grupo referência)

Valores muito abaixo de 1 podem indicar impacto desproporcional.

9.5 Equal Opportunity Difference — EOD

O **Equal Opportunity Difference** mede a diferença de taxa de verdadeiros positivos entre grupos.

$$\mathbf{EOD} = \mathbf{TPR(grupo\ protegido) - TPR(grupo\ referência)}$$

Essa métrica é relevante quando se deseja avaliar se os grupos recebem oportunidades semelhantes de classificação positiva correta.

9.6 Average Odds Difference — AOD

O **Average Odds Difference** mede a média das diferenças entre taxa de falsos positivos e taxa de verdadeiros positivos entre grupos.

$$\mathbf{AOD} = \mathbf{1/2\ [(FPR_protegido - FPR_referência) + (TPR_protegido - TPR_referência)]}$$

Essa métrica avalia diferenças nos padrões de erro entre grupos.

9.7 Índice de Impacto da Mitigação — IIM

O **Índice de Impacto da Mitigação** avalia o quanto uma intervenção alterou a distribuição dos dados ou o comportamento de um modelo.

Uma formulação simples pode considerar:

IIM = distância(distribuição antes, distribuição depois)

A distância pode ser estimada por diferença absoluta de proporções, PSI, divergência de Jensen-Shannon ou outra medida adequada ao tipo de dado.

10. Artefatos formalizados do módulo

Os artefatos do Módulo 6 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, regras de construção, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A6.1	Plano de Avaliação de Desbalanceamento e Equidade	Documento técnico	Cobertura de grupos	Define grupos, variáveis e critérios
A6.2	Registro de Grupos e Atributos Sensíveis	Tabela estruturada	DG, TSR	Formaliza grupos avaliados
A6.3	Perfil de Distribuição e Representatividade	Relatório técnico	IR, PCM, DG	Diagnostica desbalanceamentos
A6.4	Registro de Perdas por Etapa do Pipeline	Tabela analítica	TPG	Mede impacto do pipeline por grupo
A6.5	Relatório de Vieses Identificados	Relatório técnico	DCG, DQG, SPD, DI	Documenta riscos de viés
A6.6	Plano de Mitigação de Desbalanceamento	Documento técnico	IR, BA, MF1	Define ações de balanceamento
A6.7	Registro de Técnicas de Mitigação Aplicadas	Tabela técnica	IIM	Documenta técnicas, parâmetros e impacto
A6.8	Dataset Balanceado ou Sinalizado	Dataset versionado	IR, PCM	Materializa tratamento ou ressalva
A6.9	Relatório de Equidade e Fairness	Relatório técnico	SPD, DI, EOD, AOD	Avalia impacto entre grupos

A6.10	Registro de Risco Residual e Limitações	Relatório/ tabela	RRV	Documenta riscos remanescentes

11. Artefato A6.1 — Plano de Avaliação de Desbalanceamento e Equidade

11.1 Finalidade

O **Plano de Avaliação de Desbalanceamento e Equidade** define como o dataset será avaliado quanto à distribuição, representatividade, desbalanceamento, grupos relevantes, atributos sensíveis, variável-alvo e métricas de equidade.

11.2 Tipo de artefato na DSR

Na perspectiva da Design Science Research, esse artefato é um **método**, pois orienta o procedimento de avaliação, e uma **instanciação**, quando documentado em relatório, planilha, notebook, arquivo YAML, JSON ou ferramenta computacional.

11.3 Estrutura mínima

Campo	Descrição
Dataset avaliado	Ativo submetido à análise
Finalidade analítica	Uso previsto do dataset
Variável-alvo	Classe, evento ou resultado de interesse
Grupos avaliados	Grupos sensíveis ou analiticamente relevantes
Atributos sensíveis	Atributos relacionados aos grupos
Métricas previstas	IR, DG, SPD, DI, EOD, AOD etc.
Limiar de atenção	Valor que aciona alerta

Método de mitigação previsto	Quando aplicável
Responsável	Pessoa ou equipe
Critério de validação	Condição para aceitar ou restringir uso

11.4 Métricas associadas

- percentual de datasets com plano de avaliação;
- percentual de grupos definidos;
- percentual de variáveis sensíveis documentadas;
- percentual de métricas com limiar definido;
- cobertura de grupos avaliados.

11.5 Critérios de validação

O plano é considerado válido quando define dataset, finalidade, variável-alvo, grupos avaliados, atributos sensíveis, métricas, limiares e critérios de decisão.

11.6 Contribuição

Esse artefato impede que a avaliação de viés seja improvisada ou aplicada apenas depois da modelagem, inserindo a análise de equidade como parte formal do pipeline.

12. Artefato A6.2 — Registro de Grupos e Atributos Sensíveis

12.1 Finalidade

O **Registro de Grupos e Atributos Sensíveis** documenta quais grupos, atributos sensíveis, proxies e categorias serão considerados na avaliação de desbalanceamento, vieses e equidade.

12.2 Estrutura mínima

Campo	Descrição
Atributo	Campo que define grupo ou categoria
Tipo de atributo	Sensível, proxy, contextual, territorial, temporal

Grupos derivados	Categorias ou grupos analisados
Justificativa	Motivo da inclusão
Sensibilidade	Baixa, média, alta ou crítica
Uso permitido	Avaliação, mitigação, modelagem ou restrição
Restrição de acesso	Condição de uso
Observações	Cuidados éticos ou analíticos

12.3 Métricas associadas

- número de grupos definidos;
- percentual de grupos com justificativa;
- percentual de atributos sensíveis classificados;
- número de proxies identificados;
- cobertura de grupos relevantes.

12.4 Contribuição

Esse artefato torna explícito quais grupos serão avaliados, evitando invisibilidade de populações ou categorias relevantes.

13. Artefato A6.3 — Perfil de Distribuição e Representatividade

13.1 Finalidade

O **Perfil de Distribuição e Representatividade** apresenta a distribuição dos registros, classes, grupos e categorias presentes no dataset.

13.2 Estrutura mínima

Campo	Descrição

Dataset	Dataset avaliado
Variável analisada	Classe, grupo, categoria ou atributo
Grupo ou classe	Categoria específica
Frequência absoluta	Número de registros
Frequência relativa	Percentual no dataset
Referência esperada	Quando disponível
Diferença observada	Diferença em relação à referência
Classificação	Adequado, atenção, crítico
Observações	Limitações ou riscos

13.3 Métricas associadas

- razão de desbalanceamento;
- proporção da classe minoritária;
- distribuição por grupo;
- taxa de sub-representação;
- divergência em relação à distribuição esperada.

13.4 Contribuição

Esse artefato transforma a análise de distribuição em evidência documentada, apoiando decisões sobre balanceamento, amostragem, restrição de uso ou complementação de dados.

14. Artefato A6.4 — Registro de Perdas por Etapa do Pipeline

14.1 Finalidade

O **Registro de Perdas por Etapa do Pipeline** documenta quantos registros foram removidos, perdidos ou não pareados em cada etapa do pipeline, estratificando esses efeitos por grupo.

14.2 Estrutura mínima

Campo	Descrição
Dataset	Dataset analisado
Grupo	Grupo avaliado
Etapa do pipeline	Ingestão, integração, limpeza, transformação etc.
Registros antes	Quantidade antes da etapa
Registros depois	Quantidade após a etapa
Registros perdidos	Diferença entre antes e depois
Taxa de perda	Percentual de perda
Motivo	Causa da perda
Severidade	Baixa, média, alta ou crítica

14.3 Métricas associadas

- taxa de perda por grupo;

- diferença de perda entre grupos;
- percentual de grupos afetados;
- taxa de registros não pareados por grupo;
- severidade das perdas.

14.4 Contribuição

Esse artefato identifica se o próprio pipeline está produzindo exclusões desiguais, permitindo corrigir ou documentar esses efeitos.

15. Artefato A6.5 — Relatório de Vieses Identificados

15.1 Finalidade

O **Relatório de Vieses Identificados** documenta evidências de vieses técnicos, estatísticos ou algorítmicos observados nos dados, nas transformações ou nos resultados analíticos.

15.2 Tipos de viés registrados

- viés de coleta;
- viés de cobertura;
- viés de integração;
- viés de transformação;
- viés de imputação;
- viés de rotulagem;
- viés de seleção de atributos;
- viés de avaliação;
- viés de uso.

15.3 Estrutura mínima

Campo	Descrição
Dataset	Ativo analisado
Tipo de viés	Coleta, cobertura, integração etc.
Grupo afetado	Grupo ou categoria impactada
Evidência	Métrica ou observação que indica viés
Severidade	Baixa, média, alta ou crítica

Origem provável	Fonte ou etapa do pipeline
Impacto potencial	Consequência analítica ou decisória
Ação recomendada	Mitigar, monitorar, restringir, revisar
Status	Aberto, tratado, aceito, mitigado

15.4 Métricas associadas

- número de vieses identificados;
- severidade média;
- número de grupos afetados;
- diferença de qualidade por grupo;
- diferença de completude por grupo;
- métricas de fairness, quando aplicável.

15.5 Contribuição

Esse artefato torna visíveis distorções que poderiam permanecer ocultas em métricas globais de qualidade ou desempenho.

16. Artefato A6.6 — Plano de Mitigação de Desbalanceamento

16.1 Finalidade

O **Plano de Mitigação de Desbalanceamento** define as ações a serem adotadas para reduzir efeitos de desbalanceamento ou sub-representação.

16.2 Estrutura mínima

Campo	Descrição
Problema identificado	Desbalanceamento, sub-representação, classe rara etc.

Dataset	Dataset afetado
Grupo ou classe	Grupo ou classe envolvida
Técnica proposta	SMOTE, undersampling, ponderação etc.
Parâmetros	Configurações da técnica
Justificativa	Motivo da escolha
Risco	Possível distorção ou perda
Métricas antes	Distribuição original
Métricas esperadas	Resultado esperado
Critério de aceitação	Condição para validar a mitigação

16.3 Métricas associadas

- razão de desbalanceamento antes/depois;
- proporção da classe minoritária;
- macro-F1;
- balanced accuracy;
- PR-AUC;
- índice de impacto da mitigação.

16.4 Contribuição

Esse artefato evita que técnicas de balanceamento sejam aplicadas de forma automática e sem justificativa, documentando seus efeitos e riscos.

17. Artefato A6.7 — Registro de Técnicas de Mitigação Aplicadas

17.1 Finalidade

O **Registro de Técnicas de Mitigação Aplicadas** documenta as técnicas efetivamente utilizadas para tratar desbalanceamento ou viés.

17.2 Estrutura mínima

Campo	Descrição
Dataset	Dataset tratado
Técnica aplicada	Oversampling, undersampling, SMOTE, reweighing etc.
Parâmetros	Configurações utilizadas
Variável-alvo	Classe ou resultado de interesse
Grupos afetados	Grupos impactados
Distribuição antes	Frequência original
Distribuição depois	Frequência após mitigação
Impacto observado	Métricas antes/depois
Justificativa	Motivo da aplicação
Versão	Versão do dataset resultante

17.3 Métricas associadas

- índice de impacto da mitigação;
- alteração da razão de desbalanceamento;
- variação nas métricas de desempenho;
- variação nas métricas de fairness;
- taxa de registros sintéticos, quando aplicável.

17.4 Contribuição

Esse artefato preserva a rastreabilidade das intervenções aplicadas aos dados, permitindo auditoria e comparação entre cenários.

18. Artefato A6.8 — Dataset Balanceado ou Sinalizado

18.1 Finalidade

O **Dataset Balanceado ou Sinalizado** é o ativo resultante da aplicação de técnicas de mitigação ou da decisão de manter o dataset original com alertas e restrições de uso.

18.2 Estrutura formal

O dataset resultante pode ser representado por:

d' = {dataset_origem, técnica, parâmetros, distribuição_antes, distribuição_depois, grupos_afetados, métricas, versão, limitações}

18.3 Possíveis estados

Estado	Interpretação
Balanceado	Técnicas de reamostragem ou ponderação foram aplicadas
Sinalizado	O dataset foi mantido, mas possui alerta de desbalanceamento
Restrito	Uso permitido apenas para determinadas análises
Rejeitado	Uso não recomendado para finalidade prevista

18.4 Métricas associadas

- razão de desbalanceamento;
- proporção de classes;
- taxa de registros sintéticos;
- variação de distribuição;
- risco residual de viés.

18.5 Contribuição

Esse artefato materializa a decisão metodológica sobre o uso do dataset, evitando que dados desbalanceados sejam utilizados sem alerta.

19. Artefato A6.9 — Relatório de Equidade e Fairness

19.1 Finalidade

O **Relatório de Equidade e Fairness** consolida métricas de equidade entre grupos, especialmente quando o dataset é utilizado em modelos, classificações, rankings ou decisões.

19.2 Estrutura mínima

Campo	Descrição
Dataset ou modelo	Ativo avaliado
Grupos avaliados	Grupos comparados
Variável-alvo	Resultado analisado
Métricas calculadas	SPD, DI, EOD, AOD, FPR, FNR etc.
Resultados por grupo	Valores obtidos
Diferenças observadas	Distâncias entre grupos
Severidade	Baixa, média, alta ou crítica
Recomendações	Mitigar, revisar, restringir uso
Limitações	Restrições interpretativas

19.3 Métricas associadas

- SPD;
- DI;
- EOD;
- AOD;
- diferença de FPR;
- diferença de FNR;
- métricas de desempenho por grupo;
- risco residual de viés.

19.4 Contribuição

Esse artefato permite avaliar se dados, modelos ou indicadores produzem impactos desproporcionais entre grupos, apoiando decisões mais responsáveis.

20. Artefato A6.10 — Registro de Risco Residual e Limitações

20.1 Finalidade

O **Registro de Risco Residual e Limitações** documenta os riscos de viés, desbalanceamento ou inequidade que permanecem após análises e eventuais mitigações.

20.2 Estrutura mínima

Campo	Descrição
Dataset	Ativo avaliado
Risco residual	Risco remanescente
Grupo afetado	Grupo ou classe impactada
Evidência	Métrica ou observação
Severidade	Baixa, média, alta ou crítica
Mitigação aplicada	Técnica ou ação realizada
Limitação	O que não foi resolvido

Restrição de uso	Condição para uso responsável
Recomendação	Ação futura

20.3 Métricas associadas

- número de riscos residuais;
- severidade média;
- percentual de riscos mitigados;
- percentual de riscos aceitos;
- percentual de ativos com restrição de uso.

20.4 Contribuição

Esse artefato impede que limitações relevantes sejam omitidas, promovendo transparência, governança e uso responsável dos ativos analíticos.

21. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Datasets brutos, integrados e tratados; atributos sensíveis; variáveis-alvo; grupos relevantes; relatórios de qualidade; registros de transformação e metadados.
Processamento	Identificação de grupos, análise de distribuição, avaliação de desbalanceamento, mensuração de vieses, cálculo de fairness, mitigação, análise de impacto e registro de limitações.
Artefatos	Plano de avaliação, registro de grupos, perfil de distribuição, perdas por etapa, relatório de vieses, plano de mitigação, registro de técnicas, dataset balanceado ou sinalizado, relatório de equidade e risco residual.
Métricas	IR, PCM, DG, TSR, TPG, DCG, DQG, SPD, DI, EOD, AOD, ΔFPR, ΔFNR, BA, Macro-F1, PR-AUC, IIM e RRV.
Saídas	Evidências de desbalanceamento, relatórios de viés e equidade, datasets balanceados ou sinalizados, recomendações de uso e registros

de risco residual.

22. Critérios gerais de validação do módulo

O Módulo 6 pode ser considerado concluído quando:

1. os grupos relevantes foram identificados e documentados;
2. os atributos sensíveis ou proxies foram registrados;
3. a distribuição dos dados por classe, grupo, período ou território foi avaliada;
4. a variável-alvo foi analisada quanto ao desbalanceamento, quando aplicável;
5. perdas por etapa do pipeline foram calculadas, quando possível;
6. vieses potenciais foram identificados e classificados;
7. métricas de equidade foram calculadas, quando houver uso em modelos, rankings ou decisões;
8. técnicas de mitigação foram justificadas e documentadas, quando aplicadas;
9. impactos das mitigações foram avaliados;
10. riscos residuais e limitações foram registrados;
11. o dataset foi classificado como balanceado, sinalizado, restrito ou rejeitado conforme sua condição de uso.

23. Relação com os demais módulos do modelo de referência

O Módulo 6 depende do **Módulo 1 — Fontes e Inventário de Dados**, pois utiliza informações sobre origem, sensibilidade, grupos, restrições e limitações das fontes.

Depende do **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, pois problemas de cobertura e representatividade podem surgir já na ingestão dos dados.

Depende do **Módulo 3 — Integração e Harmonização**, pois perdas de pareamento, conflitos e ausência de chaves podem afetar grupos de forma desigual.

Depende do **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, pois imputações, remoções, filtros e atributos derivados podem introduzir ou ampliar vieses.

Depende do **Módulo 5 — Qualidade e Validação de Dados**, pois utiliza métricas de completude, validade, consistência e qualidade por grupo.

Para o **Módulo 7 — Metadados, Catálogo e Semântica dos Dados**, fornece informações sobre grupos, atributos sensíveis, riscos de viés, métricas de fairness e limitações de uso.

Para o **Módulo 8 — Governança, Segurança, Privacidade e Conformidade**, fornece evidências de sensibilidade, risco de impacto desproporcional e recomendações de uso responsável.

Para o **Módulo 9 — Proveniência, Lineage e Rastreabilidade**, fornece registros sobre onde vieses podem ter sido introduzidos ou mitigados no pipeline.

Para o **Módulo 10 — Versionamento e Reprodutibilidade**, fornece versões de datasets balanceados, técnicas aplicadas, parâmetros e métricas antes/depois.

Para o **Módulo 12 — Feature Engineering e Feature Store**, fornece alertas sobre atributos sensíveis, proxies, variáveis potencialmente enviesadas e qualidade por grupo.

Para o **Módulo 13 — Disponibilização Analítica**, fornece recomendações sobre uso, restrição, sinalização ou bloqueio de ativos.

Para o **Módulo 14 — Monitoramento, Drift e Evolução dos Ativos**, fornece métricas de distribuição e fairness que podem ser acompanhadas ao longo do tempo.

24. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 6 poderia avaliar uma base analítica voltada à predição de evasão ou acompanhamento de permanência estudantil.

A variável-alvo poderia ser evasão, permanência, aprovação ou reprovação. Os grupos analisados poderiam incluir sexo, idade, renda, escola de origem, modalidade de ingresso, curso, turno, campus, território, deficiência, situação socioeconômica ou participação em ações de permanência.

O módulo poderia identificar, por exemplo, que estudantes de baixa renda possuem maior proporção de valores ausentes em atributos socioeconômicos, que determinados cursos possuem baixa representação, que estudantes com menor acesso ao ambiente virtual são sub-representados em variáveis de interação ou que a integração entre bases removeu mais registros de determinado grupo.

Também poderia avaliar se um modelo de evasão apresenta desempenho desigual, por exemplo, alto desempenho global, mas baixo recall para estudantes vulneráveis. Nesse caso, métricas como FNR por grupo, Equal Opportunity Difference, Disparate Impact e Macro-F1 seriam relevantes.

As saídas poderiam incluir um relatório de equidade, um dataset sinalizado para uso com ressalvas, recomendações de mitigação e registro de limitações.

25. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, o Módulo 6 poderia avaliar dados integrados de saúde, educação, assistência social, território e trabalho.

Os grupos analisados poderiam incluir território, faixa etária, renda, condição de vulnerabilidade, tipo de atendimento, população rural ou urbana, região administrativa, escolaridade ou acesso a serviços públicos.

O módulo poderia identificar que determinados territórios possuem menor cobertura de registros, que bases de atendimento representam melhor famílias já cadastradas do que populações não atendidas, ou que decisões baseadas apenas em registros administrativos podem subestimar demandas invisíveis ao sistema.

A análise de equidade poderia orientar o uso responsável dos dados, evitando que ausência de registro seja interpretada como ausência de necessidade.

26. Contribuição técnico-científica do módulo

A contribuição do Módulo 6 está em formalizar desbalanceamento, vieses e equidade como componentes avaliáveis da Engenharia Analítica de Dados.

Em vez de tratar fairness e viés apenas como problemas da etapa de modelagem, o UnespDataLens-RM propõe sua análise ao longo do pipeline de dados, desde a origem, ingestão, integração, limpeza e transformação até a disponibilização dos ativos.

Essa formalização contribui para:

- identificar desbalanceamentos antes da modelagem;
- avaliar representatividade dos dados;
- detectar perdas desiguais produzidas pelo pipeline;
- documentar atributos sensíveis e proxies;
- mensurar diferenças de qualidade entre grupos;
- avaliar fairness em modelos, rankings ou indicadores;
- justificar técnicas de mitigação;
- registrar riscos residuais;
- apoiar governança e uso responsável;
- reduzir decisões baseadas em dados enviesados ou sub-representativos.

27. Síntese do módulo

O **Módulo 6 — Desbalanceamento, Vieses e Equidade** avalia se os dados e ativos analíticos produzidos pelo UnespDataLens-RM representam adequadamente classes, grupos, categorias, períodos, territórios e entidades relevantes ao problema analisado.

Sua importância é central porque dados podem ser tecnicamente válidos e ainda assim produzir impactos distorcidos, injustos ou desproporcionais. Ao formalizar métricas, artefatos, relatórios e critérios de validação, o módulo torna visíveis riscos que poderiam permanecer ocultos em métricas globais de qualidade ou desempenho.

Dessa forma, o Módulo 6 fortalece a responsabilidade, a transparência, a governança e a confiabilidade dos ativos analíticos, contribuindo para que o UnespDataLens-RM produza dados e modelos mais robustos, auditáveis, equitativos e adequados ao uso em contextos institucionais, educacionais e de políticas públicas.

Perfeito. Então o **Módulo 7 não deve ser “Metadados, Catálogo e Semântica dos Dados”**. Esse conteúdo deve ficar em outro módulo, provavelmente posterior. O **Módulo 7 correto**, pelo encadeamento técnico do UnespDataLens-RM, deve ser **Armazenamento Analítico**, responsável por organizar os dados tratados, integrados e validados em estruturas adequadas ao consumo analítico.

Módulo 7 — Armazenamento Analítico

1. Finalidade do módulo

O **Módulo 7 — Armazenamento Analítico** tem como finalidade estruturar, organizar, persistir e disponibilizar os dados tratados, integrados, harmonizados, validados e enriquecidos em ambientes adequados ao uso analítico no modelo de referência **UnespDataLens-RM**.

Esse módulo recebe como entrada os ativos produzidos pelos módulos anteriores, especialmente os dados extraídos e armazenados em estado bruto, os datasets integrados e harmonizados, os dados transformados e enriquecidos, os resultados de qualidade e as evidências relacionadas a desbalanceamento, vieses e equidade. A partir desses insumos, o Módulo 7 define como os dados serão organizados em estruturas persistentes, consultáveis, escaláveis, governáveis, versionáveis e preparadas para consumo por análises, modelos, dashboards, APIs, feature stores, grafos, indicadores e agentes inteligentes.

No contexto da Engenharia Analítica de Dados, o armazenamento analítico não se limita à escolha de um banco de dados, data warehouse, data lake ou lakehouse. Ele envolve decisões sobre modelagem física e lógica, organização por zonas, particionamento, indexação, formatos de arquivo, granularidade, compressão, versionamento, acesso, desempenho, retenção, atualização, consistência, segurança e interoperabilidade entre diferentes consumidores de dados.

No UnespDataLens-RM, o armazenamento analítico é concebido como uma etapa técnico-metodológica formal. Sua função é transformar datasets tratados em **ativos analíticos persistidos**, organizados segundo critérios de finalidade, qualidade, governança, rastreabilidade e reprodutibilidade. Assim, os dados deixam de ser apenas resultados intermediários de scripts ou notebooks e passam a compor uma base analítica estruturada, documentada e reutilizável.

Esse módulo é essencial porque dados bem integrados e tratados podem perder valor se forem armazenados de forma desorganizada, sem controle de versões, sem separação entre zonas, sem rastreabilidade, sem padronização de formatos ou sem mecanismos adequados de acesso. O armazenamento analítico, portanto, cria a infraestrutura lógica necessária para que os ativos de dados possam ser consumidos de maneira confiável, eficiente e auditável pelos módulos posteriores do modelo de referência.

2. Objetivo geral

O objetivo geral do Módulo 7 é **organizar e persistir os dados analíticos produzidos pelo UnespDataLens-RM em estruturas de armazenamento adequadas ao consumo, à governança, à rastreabilidade, ao desempenho, ao versionamento e à**

reprodutibilidade, permitindo sua utilização por análises, modelos, dashboards, indicadores, APIs, grafos e produtos de dados.

3. Objetivos específicos

São objetivos específicos do Módulo 7:

1. definir a arquitetura de armazenamento analítico adequada ao projeto, considerando data warehouse, data lake, lakehouse, data marts, bancos relacionais, bancos colunares, bancos de grafos ou estruturas híbridas;
2. organizar os dados em zonas ou áreas lógicas de armazenamento, como zona bruta, zona tratada, zona integrada, zona curada e zona de consumo;
3. persistir datasets tratados, integrados, harmonizados e validados em formatos adequados ao uso analítico;
4. definir estratégias de particionamento, indexação, compressão, organização física e otimização de consultas;
5. preservar versões, snapshots e estados dos datasets analíticos;
6. estabelecer padrões de nomenclatura, estruturação de diretórios, tabelas, schemas, views e data marts;
7. garantir que os dados armazenados mantenham vínculo com sua origem, regras de transformação, métricas de qualidade, restrições de acesso e versões;
8. organizar diferentes formas de representação dos dados, incluindo estruturas tabulares, temporais, multidimensionais, relacionais e orientadas a grafos;
9. preparar os dados para consumo por módulos posteriores de feature engineering, representação analítica, dashboards, modelos, APIs, agentes inteligentes e monitoramento;
10. definir métricas para avaliar cobertura, desempenho, disponibilidade, organização, rastreabilidade, versionamento e adequação do armazenamento analítico.

4. Fundamentação conceitual do módulo

O armazenamento analítico corresponde ao conjunto de decisões, estruturas e mecanismos destinados a persistir dados de forma adequada ao uso analítico. Diferentemente do armazenamento operacional, cujo foco está em registrar transações e apoiar processos cotidianos, o armazenamento analítico organiza dados para consulta, agregação, exploração, modelagem, visualização, auditoria, monitoramento e tomada de decisão.

No UnespDataLens-RM, o armazenamento analítico pode envolver diferentes paradigmas arquiteturais:

Paradigma	Característica	Uso no modelo de referência
Data Warehouse	Estrutura organizada, integrada e orientada a consultas analíticas.	Indicadores, relatórios, dashboards, data marts.

Data Lake	Armazenamento flexível de dados em diferentes formatos e níveis de estruturação.	Preservação de dados brutos, históricos e semiestruturados.
Lakehouse	Combina flexibilidade do data lake com governança, versionamento e desempenho analítico.	Dados curados, tabelas transacionais, reprocessamento e ciência de dados.
Data Mart	Subconjunto temático orientado a uma área, domínio ou finalidade.	Painéis educacionais, indicadores territoriais, análises institucionais.
Banco relacional analítico	Estruturas tabulares normalizadas ou dimensionais.	Consultas SQL, integração, relatórios e análises.
Banco colunar	Armazenamento otimizado para leitura e agregações.	Consultas analíticas de grande volume.
Banco de grafos	Estrutura orientada a nós, arestas e propriedades.	Relações entre entidades, grafos temporais e análises relacionais.
Feature Store	Armazenamento especializado de atributos para modelos.	Reuso, versionamento e consistência de features.

A escolha da estrutura depende do tipo de dado, volume, frequência de atualização, necessidade de consulta, granularidade, restrições de acesso, requisitos de desempenho e finalidade analítica.

O módulo também se fundamenta na ideia de **zonas de dados**. Uma arquitetura analítica robusta deve distinguir diferentes estados dos dados ao longo do pipeline. Por exemplo:

- **zona bruta**, que preserva dados ingeridos em estado original;
- **zona integrada**, que armazena dados combinados entre fontes;
- **zona tratada**, que contém dados limpos e padronizados;
- **zona curada**, que contém dados validados, enriquecidos e prontos para uso;
- **zona de consumo**, que disponibiliza dados para dashboards, APIs, modelos ou usuários.

Essa separação reduz confusão entre dados intermediários e dados prontos para uso, melhora a governança, facilita reprocessamento e permite rastrear a evolução dos ativos analíticos.

À luz da **Design Science Research**, o Módulo 7 formaliza o armazenamento analítico como um conjunto de artefatos projetados para resolver problemas de dispersão, baixa organização, dificuldade de reuso, baixa rastreabilidade, falta de versionamento e baixa eficiência no acesso aos dados analíticos.

5. Formalização sintética do módulo

O Módulo 7 pode ser compreendido como um processo de persistência estruturada dos datasets analíticos produzidos pelos módulos anteriores.

Considerando o conjunto de datasets tratados, integrados, validados ou sinalizados:

$$D = \{d_1, d_2, ..., d_n\}$$

o módulo define um conjunto de estruturas de armazenamento analítico:

$$S = \{s_1, s_2, ..., s_k\}$$

em que cada s_i representa uma tabela, arquivo, partição, view, data mart, schema, índice, grafo, feature table ou produto persistido.

Cada estrutura de armazenamento pode ser descrita por:

$$s_i = \{id, nome, tipo, dataset_origem, formato, local, zona, schema, granularidade, particionamento, versão, qualidade, sensibilidade, restrição_acesso, lineage\}$$

O módulo também define um conjunto de zonas:

$$Z = \{z_bruta, z_integrada, z_tratada, z_curada, z_consumo\}$$

Cada dataset d_j deve ser associado a uma ou mais zonas de armazenamento, de acordo com seu estado no pipeline e sua finalidade de uso.

O resultado do módulo é um conjunto de ativos analíticos armazenados:

$$A = \{a_1, a_2, ..., a_m\}$$

em que cada a_m representa um ativo persistido, consultável, versionado, rastreável e preparado para consumo.

Dessa forma, o armazenamento deixa de ser apenas um local físico ou lógico de gravação e passa a ser uma estrutura formal do modelo de referência, vinculada à qualidade, à governança, à linhagem, ao versionamento e ao consumo analítico.

6. Entradas do módulo

As entradas do Módulo 7 incluem datasets, metadados, regras, métricas e registros produzidos nos módulos anteriores.

As principais entradas são:

- datasets brutos preservados;
- datasets integrados e harmonizados;
- datasets tratados e enriquecidos;
- datasets validados;
- datasets balanceados ou sinalizados;
- manifestos de ingestão;
- manifestos de integração;
- manifestos de transformação;
- relatórios de qualidade;
- registros de não conformidade;
- relatórios de vieses e equidade;
- classificação de sensibilidade;
- registro de granularidade;
- dicionários de dados;
- regras de transformação;
- tabelas de mapeamento;
- registros de lineage preliminar;
- versões e snapshots;
- requisitos de consulta;
- requisitos de desempenho;
- requisitos de segurança;
- requisitos de consumo analítico;
- políticas de retenção;
- políticas de acesso;
- padrões de nomenclatura e organização.

7. Processamento

O processamento do Módulo 7 envolve decisões de arquitetura, modelagem, organização, persistência, otimização, versionamento e disponibilização interna dos dados analíticos.

7.1 Definição da arquitetura de armazenamento

A primeira atividade consiste em definir a arquitetura de armazenamento analítico mais adequada ao projeto. Essa decisão deve considerar volume de dados, variedade de formatos, frequência de atualização, tipos de consulta, necessidade de reprocessamento, requisitos de governança, custo, escalabilidade e integração com ferramentas analíticas.

Dependendo do projeto, pode-se adotar um data warehouse, data lake, lakehouse, banco relacional, banco colunar, banco de grafos, data marts ou uma arquitetura híbrida.

7.2 Definição das zonas de dados

A segunda atividade consiste em organizar os dados em zonas lógicas. Essa separação permite distinguir dados brutos, integrados, tratados, curados e prontos para consumo.

Cada zona deve possuir finalidade, regras de acesso, critérios de entrada, critérios de saída, política de retenção, padrão de nomenclatura e relação com os módulos do modelo de referência.

7.3 Modelagem lógica dos dados analíticos

A terceira atividade consiste em definir como os dados serão organizados logicamente. Essa modelagem pode envolver tabelas normalizadas, modelos dimensionais, tabelas fato e dimensão, views analíticas, data marts, tabelas agregadas, estruturas orientadas a eventos, séries temporais ou grafos.

A modelagem deve considerar a finalidade do consumo: análise exploratória, dashboards, modelos preditivos, análise temporal, representação relacional, grafos ou indicadores gerenciais.

7.4 Definição de formatos e padrões de persistência

A quarta atividade consiste em definir os formatos de armazenamento, como CSV, Parquet, Delta, Iceberg, Hudi, ORC, tabelas relacionais, documentos JSON, tabelas colunares ou estruturas de grafos.

A escolha do formato deve considerar eficiência de leitura, compressão, interoperabilidade, versionamento, suporte a schema evolution, governança e ferramentas disponíveis.

7.5 Particionamento, indexação e otimização

A quinta atividade consiste em definir estratégias para melhorar desempenho e organização. Isso pode incluir particionamento por data, fonte, domínio, território, curso, município, período letivo ou versão; criação de índices; organização por colunas; materialização de views; agregações pré-computadas; ordenação física e compactação de arquivos.

7.6 Persistência dos dados analíticos

A sexta atividade consiste em gravar os datasets nas estruturas definidas, preservando metadados mínimos sobre origem, versão, schema, qualidade, sensibilidade, data de atualização e lineage.

A persistência deve garantir que os dados possam ser localizados, consultados, auditados, versionados e utilizados por módulos posteriores.

7.7 Versionamento e snapshots analíticos

A sétima atividade consiste em preservar versões dos datasets analíticos, permitindo reconstrução de estados anteriores, comparação entre execuções e reprocessamento de análises.

O versionamento pode ocorrer por data, número de versão, hash, snapshot, tag semântica ou controle transacional em ambientes lakehouse.

7.8 Organização dos ativos para consumo

A oitava atividade consiste em preparar estruturas específicas para consumo analítico, como data marts, views, tabelas agregadas, datasets para treinamento, bases para dashboards, tabelas de features, estruturas para APIs ou grafos persistidos.

Essa organização não substitui o módulo de disponibilização analítica, mas cria as bases persistidas que serão posteriormente expostas para consumo.

7.9 Controle de acesso e sensibilidade no armazenamento

A nona atividade consiste em garantir que os dados armazenados respeitem classificações de sensibilidade, restrições de acesso, políticas de segurança, anonimização, pseudonimização ou segregação de dados, quando necessário.

Dados sensíveis devem possuir controles adequados de acesso, trilhas de auditoria e políticas de retenção.

7.10 Registro do armazenamento analítico

A décima atividade consiste em documentar onde cada ativo foi armazenado, em qual formato, em qual zona, com qual schema, versão, particionamento, tamanho, frequência de atualização, qualidade e restrição de acesso.

Esse registro será utilizado por módulos posteriores de catálogo, lineage, versionamento, monitoramento e agentes inteligentes.

8. Técnicas, métodos e tecnologias aplicáveis

O Módulo 7 pode empregar diferentes técnicas, métodos e tecnologias de armazenamento analítico.

Entre as técnicas de modelagem, destacam-se:

- modelagem relacional;
- modelagem dimensional;
- modelo estrela;
- modelo floco de neve;
- tabelas fato e dimensão;
- data marts temáticos;
- views materializadas;
- tabelas agregadas;
- modelagem orientada a eventos;
- modelagem temporal;
- modelagem orientada a grafos;
- modelagem híbrida;
- schema-on-read;
- schema-on-write;
- data vault, quando aplicável.

Entre as técnicas de organização física, podem ser utilizadas:

- particionamento;
- indexação;
- clustering físico;
- compressão;
- compactação de arquivos;
- ordenação por colunas;
- otimização de leitura;
- controle de schema;
- schema evolution;
- versionamento de tabelas;
- snapshots;

- materialização de consultas;
- separação por zonas;
- segregação por sensibilidade.

Entre os formatos e tecnologias possíveis, destacam-se:

- CSV;
- JSON;
- Parquet;
- ORC;
- Avro;
- Delta Lake;
- Apache Iceberg;
- Apache Hudi;
- PostgreSQL;
- SQL Server;
- MySQL;
- DuckDB;
- ClickHouse;
- BigQuery;
- Redshift;
- Snowflake;
- Databricks;
- Apache Hive;
- Spark SQL;
- Trino;
- Presto;
- Neo4j;
- ArangoDB;
- bancos orientados a grafos;
- object storage;
- data lakes;
- lakehouses.

A escolha das tecnologias deve ser justificada conforme os objetivos do projeto, os requisitos de desempenho, volume, governança, custo, interoperabilidade e capacidade de reprocessamento.

9. Métricas do módulo

As métricas do Módulo 7 têm como objetivo avaliar organização, cobertura, eficiência, rastreabilidade, versionamento, disponibilidade e adequação do armazenamento analítico.

Métrica	Sigla	Descrição
Cobertura de armazenamento analítico	CAA	Proporção de datasets tratados persistidos em estrutura analítica
Taxa de ativos armazenados	TAAS	Proporção de ativos com schema documentado

com schema

Taxa de ativos versionados	TAV	Proporção de ativos com versão ou snapshot
Taxa de ativos com localização documentada	TALD	Proporção de ativos com local de armazenamento registrado
Taxa de ativos com zona definida	TAZD	Proporção de ativos classificados por zona
Taxa de ativos com particionamento adequado	TAPA	Proporção de ativos particionados segundo critério definido
Tempo médio de consulta	TMC	Tempo médio para execução de consultas analíticas
Taxa de consultas bem-sucedidas	TCBS	Proporção de consultas concluídas sem erro
Taxa de ativos com controle de acesso	TACA	Proporção de ativos com regra de acesso definida
Taxa de ativos com lineage associado	TALA	Proporção de ativos armazenados com origem e dependências registradas
Taxa de duplicação de armazenamento	TDA	Proporção de ativos redundantes ou duplicados
Taxa de atualização conforme periodicidade	TACP	Proporção de ativos atualizados conforme periodicidade definida
Taxa de reprocessabilidade	TRP	Proporção de ativos que podem ser reconstruídos a partir de versões e metadados

Eficiência de armazenamento	EA	Relação entre volume armazenado, compressão e formato adotado
Disponibilidade do ativo	DA	Grau de disponibilidade do ativo para consulta autorizada

9.1 Cobertura de Armazenamento Analítico — CAA

A **Cobertura de Armazenamento Analítico** mede a proporção de datasets tratados, integrados ou validados que foram persistidos em estruturas analíticas.

CAA = número de datasets armazenados analiticamente / número total de datasets elegíveis para armazenamento analítico

Essa métrica indica se os ativos produzidos pelo pipeline estão sendo devidamente organizados para consumo posterior.

9.2 Taxa de Ativos Versionados — TAV

A **Taxa de Ativos Versionados** mede a proporção de ativos armazenados que possuem versão, snapshot ou mecanismo equivalente de identificação temporal.

TAV = número de ativos versionados / número total de ativos armazenados

Essa métrica é fundamental para reprodutibilidade e auditoria.

9.3 Taxa de Ativos com Localização Documentada — TALD

A **Taxa de Ativos com Localização Documentada** mede a proporção de ativos cujo local físico ou lógico de armazenamento foi registrado.

TALD = número de ativos com localização documentada / número total de ativos armazenados

Essa métrica evita perda de rastreabilidade e facilita descoberta e governança.

9.4 Tempo Médio de Consulta — TMC

O **Tempo Médio de Consulta** avalia o desempenho do armazenamento para consultas analíticas.

TMC = soma dos tempos de consulta / número total de consultas avaliadas

Essa métrica deve ser interpretada de acordo com volume, complexidade da consulta e requisitos de uso.

9.5 Taxa de Reprocessabilidade — TRP

A **Taxa de Reprocessabilidade** mede a proporção de ativos armazenados que podem ser reconstruídos a partir de suas fontes, versões, regras, parâmetros e metadados.

TRP = número de ativos reprocessáveis / número total de ativos armazenados

Essa métrica conecta o armazenamento analítico aos princípios de lineage, versionamento e reprodutibilidade.

10. Artefatos formalizados do módulo

Os artefatos do Módulo 7 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, regras de construção, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A7.1	Arquitetura de Armazenamento Analítico	Documento técnico/diagrama	CAA	Define a estrutura geral de armazenamento
A7.2	Modelo de Zonas de Dados	Especificação lógica	TAZD	Organiza estados dos dados no pipeline
A7.3	Modelo Lógico Analítico	Modelo relacional/dimensional/grafo	TAAS	Define organização lógica dos dados
A7.4	Mapa de Armazenamento Analítico	Tabela/catálogo técnico	TALD	Registra localização dos ativos
A7.5	Especificação de Schemas Analíticos	Dicionário/schema	TAAS	Formaliza estrutura dos ativos
A7.6	Registro de Particionamento e Indexação	Tabela técnica	TAPA, TMC	Otimiza consultas e organização
A7.7	Registro de Versionamento e	Tabela/manifesto	TAV, TRP	Preserva versões dos ativos

	Snapshots Analíticos			
A7.8	Data Mart ou Dataset Curado	Tabela/view/dataset	CAA	Materializa dados prontos para consumo
A7.9	Registro de Controle de Acesso ao Armazenamento	Matriz de acesso	TACA	Garante segurança e conformidade
A7.10	Relatório de Desempenho e Adequação do Armazenamento	Relatório técnico	TMC, DA	Avalia eficiência e adequação

11. Artefato A7.1 — Arquitetura de Armazenamento Analítico

11.1 Finalidade

A **Arquitetura de Armazenamento Analítico** define a estrutura geral utilizada para persistir, organizar e disponibilizar os dados analíticos no UnespDataLens-RM.

11.2 Tipo de artefato na DSR

Na perspectiva da Design Science Research, esse artefato é um **modelo**, pois representa a estrutura arquitetural de armazenamento, e uma **instanciação**, quando implementado em ambiente computacional.

11.3 Estrutura mínima

Elemento	Descrição
Paradigma adotado	Data warehouse, data lake, lakehouse, banco relacional, grafo etc.
Zonas de dados	Bruta, integrada, tratada, curada, consumo

Tecnologias	Ferramentas, bancos, formatos e plataformas
Critérios de organização	Domínio, data, fonte, versão, sensibilidade
Estratégia de versionamento	Snapshot, tabela transacional, tag, hash
Estratégia de acesso	Papéis, permissões, autenticação
Estratégia de desempenho	Particionamento, indexação, compressão
Estratégia de governança	Metadados, qualidade, lineage, retenção
Consumidores previstos	Dashboards, modelos, APIs, grafos, agentes

11.4 Métricas associadas

- cobertura de armazenamento analítico;
- taxa de ativos com zona definida;
- taxa de ativos com controle de acesso;
- taxa de ativos versionados;
- disponibilidade dos ativos.

11.5 Critérios de validação

A arquitetura é considerada válida quando permite armazenar os principais ativos analíticos do projeto de forma organizada, rastreável, segura, consultável e adequada aos usos previstos.

11.6 Contribuição

Esse artefato transforma decisões de armazenamento em uma estrutura arquitetural explícita, reduzindo improvisação e dispersão dos ativos.

12. Artefato A7.2 — Modelo de Zonas de Dados

12.1 Finalidade

O **Modelo de Zonas de Dados** define as áreas lógicas de armazenamento utilizadas para organizar os diferentes estados dos dados no pipeline.

12.2 Estrutura mínima

Zona	Finalidade	Tipo de dado	Critério de entrada	Critério de saída
Bruta	Preservar dados ingeridos	Dados originais	Ingestão concluída	Validação inicial
Integrada	Armazenar dados combinados	Dados relacionados entre fontes	Integração documentada	Harmonização validada
Tratada	Armazenar dados limpos	Dados padronizados	Transformação concluída	Validação de qualidade
Curada	Armazenar dados prontos para análise	Dados validados e enriquecidos	Qualidade aprovada	Consumo analítico
Consumo	Disponibilizar dados para uso	Data marts, views, APIs, dashboards	Ativo aprovado	Uso por consumidores

12.3 Métricas associadas

- taxa de ativos com zona definida;
- número de ativos por zona;
- taxa de ativos indevidamente classificados;
- tempo médio de permanência por zona;
- proporção de ativos curados.

12.4 Critérios de validação

O modelo é considerado válido quando cada ativo armazenado possui zona definida e quando os critérios de entrada e saída de cada zona estão documentados.

12.5 Contribuição

Esse artefato organiza o ciclo de vida dos dados no armazenamento analítico, evitando mistura entre dados brutos, intermediários e prontos para consumo.

13. Artefato A7.3 — Modelo Lógico Analítico

13.1 Finalidade

O **Modelo Lógico Analítico** define como os dados serão representados para consulta e análise.

13.2 Possíveis representações

Representação	Aplicação
Relacional	Consultas SQL, integração e consistência
Dimensional	Dashboards, indicadores e análise gerencial
Colunar	Consultas analíticas de grande volume
Temporal	Séries históricas e trajetórias
Evento	Logs, interações e processos dinâmicos
Grafo	Relações entre entidades
Híbrida	Projetos com múltiplas necessidades analíticas

13.3 Estrutura mínima

Campo	Descrição
Nome do modelo	Identificação
Dataset de origem	Ativo usado como base

Tipo de modelo	Relacional, dimensional, grafo, temporal etc.
Entidades ou tabelas	Estruturas principais
Chaves	Identificadores e relações
Granularidade	Unidade de análise
Atributos principais	Campos relevantes
Finalidade	Uso analítico esperado
Restrições	Cuidados ou limitações

13.4 Métricas associadas

- taxa de ativos com schema analítico;
- número de modelos definidos;
- taxa de modelos com granularidade explícita;
- taxa de modelos com chaves documentadas;
- adequação ao consumo previsto.

13.5 Contribuição

Esse artefato garante que o armazenamento analítico seja orientado por modelos de uso e não apenas por cópias físicas de arquivos.

14. Artefato A7.4 — Mapa de Armazenamento Analítico

14.1 Finalidade

O **Mapa de Armazenamento Analítico** registra onde cada ativo de dados está armazenado, em qual formato, zona, versão, schema e condição de acesso.

14.2 Estrutura mínima

Campo	Descrição

Identificador do ativo	Código único
Nome do ativo	Nome do dataset, tabela ou view
Zona	Bruta, integrada, tratada, curada ou consumo
Localização	Caminho, banco, schema, bucket ou endpoint
Formato	CSV, Parquet, Delta, tabela SQL etc.
Versão	Versão ou snapshot
Tamanho	Volume armazenado
Particionamento	Critério usado
Sensibilidade	Classificação
Restrição de acesso	Política de acesso
Atualização	Periodicidade ou última carga

14.3 Métricas associadas

- taxa de ativos com localização documentada;
- taxa de ativos com zona definida;
- taxa de ativos com versão;
- taxa de ativos com formato registrado;
- taxa de ativos com restrição de acesso definida.

14.4 Contribuição

Esse artefato permite localizar, auditar, consultar e governar os ativos armazenados.

15. Artefato A7.5 — Especificação de Schemas Analíticos

15.1 Finalidade

A **Especificação de Schemas Analíticos** documenta a estrutura dos datasets armazenados, incluindo atributos, tipos, domínios, chaves, constraints, granularidade e regras de atualização.

15.2 Estrutura mínima

Campo	Descrição
Ativo	Dataset ou tabela
Atributo	Nome do campo
Tipo	Tipo de dado
Domínio	Valores permitidos
Obrigatoriedade	Obrigatório ou opcional
Chave	Primária, estrangeira, composta ou ausente
Descrição	Significado
Granularidade	Unidade de análise
Sensibilidade	Classificação
Regra de atualização	Como o campo é atualizado

15.3 Métricas associadas

- taxa de ativos armazenados com schema;
- percentual de atributos documentados;
- percentual de chaves documentadas;
- taxa de conformidade de schema;
- taxa de schema evolution documentada.

15.4 Contribuição

Esse artefato reduz ambiguidade estrutural e apoia consulta, validação, governança, lineage e reprodutibilidade.

16. Artefato A7.6 — Registro de Particionamento e Indexação

16.1 Finalidade

O **Registro de Particionamento e Indexação** documenta as estratégias utilizadas para organizar fisicamente os dados e otimizar consultas.

16.2 Estrutura mínima

Campo	Descrição
Ativo	Dataset, tabela ou view
Critério de particionamento	Data, fonte, domínio, território etc.
Índices criados	Campos indexados
Justificativa	Motivo técnico
Consultas-alvo	Consultas beneficiadas
Impacto esperado	Redução de tempo, organização, filtragem
Data de criação	Momento da configuração

Observações	Limitações ou custos

16.3 Métricas associadas

- taxa de ativos com particionamento adequado;
- tempo médio de consulta;
- redução de tempo após indexação;
- taxa de consultas bem-sucedidas;
- custo de armazenamento adicional.

16.4 Contribuição

Esse artefato torna explícitas decisões de desempenho e organização física dos dados.

17. Artefato A7.7 — Registro de Versionamento e Snapshots Analíticos

17.1 Finalidade

O **Registro de Versionamento e Snapshots Analíticos** documenta versões dos ativos armazenados, permitindo reconstrução de análises e comparação entre estados temporais.

17.2 Estrutura mínima

Campo	Descrição
Ativo	Dataset ou tabela versionada
Versão	Número, tag, hash ou data
Snapshot	Identificador do estado preservado
Data de criação	Momento da versão
Origem	Dataset ou processo que gerou a

	versão
Alteração	Descrição da mudança
Responsável	Executor humano ou sistema
Localização	Caminho da versão
Status	Ativa, substituída, arquivada
Observações	Restrições ou limitações

17.3 Métricas associadas

- taxa de ativos versionados;
- número de versões por ativo;
- taxa de ativos reprocessáveis;
- percentual de snapshots íntegros;
- tempo de recuperação de versões anteriores.

17.4 Contribuição

Esse artefato fortalece reprodutibilidade, auditoria e rastreabilidade temporal dos dados analíticos.

18. Artefato A7.8 — Data Mart ou Dataset Curado

18.1 Finalidade

O **Data Mart ou Dataset Curado** representa um ativo analítico persistido e preparado para uma finalidade específica de consumo.

18.2 Estrutura formal

Um data mart ou dataset curado pode ser descrito por:

dm_i = {id, nome, domínio, finalidade, fontes_origem, tabelas, granularidade, período, métricas_qualidade, versão, restrição_acesso, consumidores}

18.3 Estrutura mínima

Campo	Descrição
Nome do ativo	Nome do data mart ou dataset curado
Domínio	Área temática
Finalidade	Uso analítico previsto
Fontes de origem	Bases utilizadas
Estrutura	Tabelas, views, arquivos ou grafos
Granularidade	Unidade de análise
Cobertura temporal	Período abrangido
Métricas de qualidade	Indicadores associados
Versão	Versão atual
Consumidores	Dashboards, modelos, APIs, analistas
Restrições de uso	Condições de acesso

18.4 Métricas associadas

- número de data marts criados;
- taxa de ativos curados;
- taxa de reutilização;
- tempo médio de consulta;

- status de qualidade dos ativos curados.

18.5 Contribuição

Esse artefato materializa dados prontos para uso analítico, reduzindo retrabalho e promovendo reutilização controlada.

19. Artefato A7.9 — Registro de Controle de Acesso ao Armazenamento

19.1 Finalidade

O **Registro de Controle de Acesso ao Armazenamento** documenta quem pode acessar cada ativo armazenado e sob quais condições.

19.2 Estrutura mínima

Campo	Descrição
Ativo	Dataset, tabela, view ou data mart
Sensibilidade	Classificação do dado
Perfil de acesso	Administrador, analista, pesquisador, gestor etc.
Permissão	Leitura, escrita, administração, consulta agregada
Restrição	Condição específica
Justificativa	Motivo da permissão
Responsável pela autorização	Pessoa ou setor
Data de autorização	Momento do registro

Validade	Prazo da autorização

19.3 Métricas associadas

- taxa de ativos com controle de acesso;
- número de acessos autorizados;
- número de acessos expirados;
- percentual de ativos sensíveis com controle explícito;
- taxa de conformidade de acesso.

19.4 Contribuição

Esse artefato conecta armazenamento analítico a governança, segurança, privacidade e conformidade.

20. Artefato A7.10 — Relatório de Desempenho e Adequação do Armazenamento

20.1 Finalidade

O **Relatório de Desempenho e Adequação do Armazenamento** avalia se as estruturas definidas atendem aos requisitos de consulta, volume, atualização, governança e consumo analítico.

20.2 Estrutura mínima

Campo	Descrição
Ativo avaliado	Dataset, tabela, data mart ou zona
Volume	Tamanho armazenado
Consultas avaliadas	Consultas ou usos testados
Tempo médio de consulta	Desempenho observado
Taxa de sucesso	Consultas concluídas

Gargalos	Problemas identificados
Adequação ao uso	Aprovado, ressalvas, inadequado
Recomendações	Otimização, particionamento, refatoração
Observações	Limitações ou riscos

20.3 Métricas associadas

- tempo médio de consulta;
- taxa de consultas bem-sucedidas;
- disponibilidade do ativo;
- eficiência de armazenamento;
- taxa de ativos adequados ao uso.

20.4 Contribuição

Esse artefato permite avaliar se o armazenamento analítico está efetivamente servindo aos consumidores de dados.

21. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Datasets brutos, integrados, tratados, validados e sinalizados; metadados; regras; qualidade; sensibilidade; versões; requisitos de consulta e consumo.
Processamento	Definição da arquitetura, zonas, modelo lógico, formatos, particionamento, indexação, persistência, versionamento, controle de acesso e avaliação de desempenho.
Artefatos	Arquitetura de armazenamento, modelo de zonas, modelo lógico analítico, mapa de armazenamento, schemas, particionamento, versionamento, data marts, controle de acesso e relatório de desempenho.

Métricas	CAA, TAAS, TAV, TALD, TAZD, TAPA, TMC, TCBS, TACA, TALA, TDA, TACP, TRP, EA e DA.
Saídas	Ativos analíticos persistidos, organizados, versionados, consultáveis, seguros, rastreáveis e preparados para consumo por análises, modelos, dashboards, APIs, grafos e agentes inteligentes.

22. Critérios gerais de validação do módulo

O Módulo 7 pode ser considerado concluído quando:

1. a arquitetura de armazenamento analítico foi definida;
2. as zonas de dados foram especificadas;
3. os datasets elegíveis foram persistidos em estruturas analíticas adequadas;
4. cada ativo armazenado possui localização documentada;
5. os schemas dos principais ativos foram especificados;
6. os ativos críticos possuem versão ou snapshot;
7. os ativos sensíveis possuem controle de acesso definido;
8. estratégias de particionamento ou indexação foram aplicadas quando necessário;
9. data marts ou datasets curados foram produzidos para consumo;
10. métricas de desempenho e adequação foram avaliadas;
11. os ativos armazenados mantêm vínculo com origem, qualidade, sensibilidade, versão e lineage;
12. os dados estão preparados para consumo pelos módulos posteriores.

23. Relação com os demais módulos do modelo de referência

O Módulo 7 depende do **Módulo 1 — Fontes e Inventário de Dados**, pois utiliza informações sobre origem, formato, domínio, sensibilidade, granularidade e periodicidade das fontes.

Depende do **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, pois incorpora dados brutos, snapshots e registros de ingestão à organização geral de armazenamento.

Depende do **Módulo 3 — Integração e Harmonização**, pois recebe datasets integrados, regras de integração e estruturas resultantes da combinação entre fontes.

Depende do **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, pois recebe dados tratados, enriquecidos e atributos derivados que precisam ser persistidos.

Depende do **Módulo 5 — Qualidade e Validação de Dados**, pois utiliza status de qualidade, métricas e relatórios para definir quais ativos podem ser curados ou disponibilizados.

Depende do **Módulo 6 — Desbalanceamento, Vieses e Equidade**, pois utiliza sinalizações, restrições e versões balanceadas ou não balanceadas de datasets.

Para o **Módulo 8 — Metadados, Catálogo e Semântica dos Dados**, fornece localização, estrutura, formato, schema, zona, versão e estado dos ativos armazenados.

Para o **Módulo 9 — Governança, Segurança, Privacidade e Conformidade**, fornece informações sobre sensibilidade, controle de acesso, retenção e restrições no armazenamento.

Para o **Módulo 10 — Proveniência, Lineage e Rastreabilidade**, fornece vínculos entre ativos armazenados, versões, zonas e datasets derivados.

Para o **Módulo 11 — Versionamento e Reprodutibilidade**, fornece snapshots, versões e estruturas persistidas necessárias à reconstrução de análises.

Para o **Módulo 12 — Representações Analíticas: Tabular, Temporal, Espacial, Textual, Semântica e Grafos**, fornece estruturas armazenadas que podem ser convertidas em diferentes representações analíticas.

Para o **Módulo 13 — Feature Engineering e Feature Store**, fornece datasets curados e estruturas persistidas para criação e armazenamento de features.

Para o **Módulo 14 — Disponibilização Analítica**, fornece data marts, views, tabelas, datasets e estruturas prontas para consumo externo ou interno.

Para o **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, fornece baseline de armazenamento, versões, volumes, schemas e histórico de atualização.

24. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 7 poderia organizar dados tratados sobre matrículas, histórico acadêmico, notas, frequência, evasão, interações em ambiente virtual, dados socioeconômicos e indicadores de permanência.

A arquitetura poderia prever uma zona bruta com dados originais, uma zona integrada com bases relacionadas por estudante, curso e período letivo, uma zona tratada com dados limpos e padronizados, uma zona curada com datasets validados para análise e uma zona de consumo com data marts para dashboards ou modelos preditivos.

Exemplos de ativos armazenados poderiam incluir:

- dataset curado de trajetória acadêmica;
- data mart de evasão por curso e período;
- tabela de indicadores de desempenho;
- view de estudantes ativos por semestre;
- dataset de interações no AVA por estudante e disciplina;
- tabela temporal de situação acadêmica;
- estrutura de grafo estudante-disciplina-curso.

Cada ativo teria schema, versão, granularidade, localização, qualidade, sensibilidade e restrição de acesso documentados.

25. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, o Módulo 7 poderia armazenar dados tratados e integrados de saúde, educação, assistência social, território, trabalho e habitação.

A arquitetura poderia organizar dados por zonas e domínios, criando data marts para análise territorial, vulnerabilidade social, atendimento a famílias, cobertura de serviços e indicadores municipais.

Exemplos de ativos armazenados poderiam incluir:

- data mart territorial;
- base curada de famílias atendidas;
- tabela de equipamentos públicos por território;
- dataset de atendimentos por serviço;
- indicadores socioeconômicos por município;
- grafo território-equipamento-atendimento;
- views agregadas para dashboards de gestão.

Cada ativo seria armazenado com controle de acesso, principalmente quando envolvesse dados pessoais ou sensíveis.

26. Contribuição técnico-científica do módulo

A contribuição do Módulo 7 está em formalizar o armazenamento analítico como uma etapa técnico-metodológica do modelo de referência, e não apenas como uma escolha de infraestrutura.

O módulo propõe que os dados analíticos sejam armazenados segundo critérios explícitos de arquitetura, zonas, modelagem lógica, formato, particionamento, versionamento, controle de acesso, desempenho, rastreabilidade e adequação ao consumo.

Essa formalização contribui para:

- evitar dispersão de datasets analíticos;
- reduzir duplicidade e inconsistência entre versões;
- organizar dados por estado e finalidade;
- melhorar desempenho de consultas;
- preservar versões e snapshots;
- apoiar governança e controle de acesso;
- facilitar reuso de dados curados;
- preparar dados para dashboards, modelos, grafos, APIs e agentes inteligentes;
- fortalecer lineage, reprodutibilidade e auditoria;
- transformar datasets tratados em ativos analíticos persistidos e reutilizáveis.

27. Síntese do módulo

O **Módulo 7 — Armazenamento Analítico** organiza e persiste os dados produzidos pelo UnespDataLens-RM em estruturas adequadas ao uso analítico. Ele define arquitetura, zonas, modelos lógicos, formatos, schemas, particionamento, versionamento, data marts, controle de acesso e critérios de desempenho.

Sua importância é central porque dados tratados e validados só se tornam efetivamente reutilizáveis quando armazenados de forma organizada, documentada, consultável, segura, versionável e rastreável.

Dessa forma, o Módulo 7 transforma os resultados intermediários do pipeline em ativos analíticos persistidos, preparados para consumo por análises, modelos, dashboards, APIs, grafos, feature stores, agentes inteligentes e processos de tomada de decisão baseada em evidências.

Perfeito. Então a estrutura fica mais coerente assim: os **Módulos 1 a 8** compõem o **pipeline técnico-operacional principal**, desde as fontes até o consumo dos ativos analíticos. A partir do **Módulo 9**, entram os módulos transversais, como metadados, governança, lineage, versionamento, monitoramento e agentes inteligentes.

Módulo 8 — Disponibilização e Consumo Analítico

1. Finalidade do módulo

O **Módulo 8 — Disponibilização e Consumo Analítico** tem como finalidade organizar, publicar, expor e tornar utilizáveis os ativos analíticos produzidos ao longo do pipeline principal do modelo de referência **UnespDataLens-RM**. Esse módulo representa a etapa final do fluxo técnico-operacional do pipeline, pois transforma dados armazenados, tratados, integrados, validados e organizados em produtos analíticos acessíveis a usuários, sistemas, dashboards, modelos, APIs, relatórios, aplicações, agentes inteligentes e processos de tomada de decisão.

Enquanto os módulos anteriores concentram-se na identificação, ingestão, integração, transformação, validação, análise de equidade e armazenamento analítico dos dados, o Módulo 8 concentra-se no momento em que os dados passam a ser efetivamente consumidos. Seu papel é garantir que os ativos analíticos sejam disponibilizados de forma controlada, documentada, segura, compreensível, rastreável e adequada às diferentes finalidades de uso.

No contexto da Engenharia Analítica de Dados, a disponibilização não deve ser confundida com simples exportação de tabelas ou geração de relatórios. Ela envolve decisões sobre forma de acesso, público-alvo, granularidade da informação, nível de agregação, autorização, formato de entrega, periodicidade de atualização, documentação, versionamento, rastreabilidade, indicadores de qualidade, restrições de uso e mecanismos de feedback.

No **UnespDataLens-RM**, o Módulo 8 encerra o pipeline principal ao transformar ativos armazenados em **produtos de dados consumíveis**, tais como datasets curados, data marts, dashboards, painéis de indicadores, APIs, relatórios analíticos, views, arquivos de exportação, bases para modelagem, bases para grafos, produtos de dados institucionais e insumos para agentes inteligentes.

Esse módulo é essencial porque um pipeline de dados só gera valor quando seus resultados podem ser utilizados de forma confiável por pessoas, sistemas ou processos decisórios. Dados bem preparados, mas não disponibilizados de maneira adequada, tornam-se ativos subutilizados. Por outro lado, dados disponibilizados sem documentação, controle de acesso, qualidade conhecida ou restrição de uso podem gerar interpretações equivocadas, riscos institucionais, problemas de privacidade e decisões inadequadas.

Assim, o Módulo 8 formaliza a etapa de disponibilização e consumo como um processo técnico-metodológico avaliável, composto por artefatos, métricas, critérios de validação e mecanismos de governança operacional.

2. Objetivo geral

O objetivo geral do Módulo 8 é **disponibilizar ativos analíticos produzidos pelo UnespDataLens-RM em formatos, interfaces e produtos adequados ao consumo**, assegurando controle de acesso, documentação mínima, qualidade conhecida, rastreabilidade, adequação ao uso, segurança, atualização e capacidade de reutilização por diferentes públicos e aplicações.

3. Objetivos específicos

São objetivos específicos do Módulo 8:

1. definir os públicos, sistemas, aplicações e processos que irão consumir os ativos analíticos;
2. organizar os ativos armazenados em formatos adequados ao consumo analítico, operacional, científico ou decisório;
3. disponibilizar datasets curados, views, data marts, APIs, dashboards, relatórios, indicadores, bases para modelagem, bases para grafos e produtos de dados;
4. estabelecer regras de acesso, restrição, agregação, anonimização ou filtragem conforme sensibilidade e finalidade de uso;
5. documentar a finalidade, origem, qualidade, versão, periodicidade, limitações e condições de uso de cada ativo disponibilizado;
6. definir mecanismos de atualização, publicação, republicação, descontinuação e substituição de ativos;
7. controlar versões dos ativos disponibilizados, evitando uso indevido de versões obsoletas;
8. registrar consumidores, formas de acesso, frequência de uso, dependências e feedback dos usuários;
9. avaliar a utilidade, desempenho, disponibilidade e aderência dos produtos analíticos às necessidades dos usuários;
10. encerrar o pipeline principal entregando ativos analíticos preparados para uso, ao mesmo tempo em que fornece insumos para os módulos transversais de metadados, governança, lineage, versionamento, monitoramento e agentes inteligentes.

4. Fundamentação conceitual do módulo

A disponibilização de dados corresponde ao processo de tornar ativos de dados acessíveis para consumo por pessoas, sistemas ou processos computacionais. No contexto analítico, essa disponibilização deve ocorrer de forma planejada, pois diferentes consumidores possuem necessidades distintas.

Um gestor pode necessitar de um painel com indicadores agregados. Um pesquisador pode precisar de um dataset anonimizado para análise estatística. Um cientista de dados pode precisar de uma base versionada para treinamento de modelos. Um sistema externo pode precisar acessar dados por API. Um agente inteligente pode consultar metadados e produtos de dados para responder perguntas ou recomendar ações. Um dashboard pode consumir views ou data marts atualizados periodicamente.

Assim, a disponibilização analítica deve considerar:

- quem irá consumir o ativo;
- para qual finalidade;
- em qual formato;
- com qual nível de granularidade;

- com quais restrições;
- com qual frequência de atualização;
- com qual documentação;
- com qual status de qualidade;
- com qual versão;
- com qual mecanismo de acesso;
- com quais limitações de interpretação.

No UnespDataLens-RM, essa etapa também é associada ao conceito de **produto de dados**. Um produto de dados é um ativo organizado, documentado, confiável, governado e preparado para consumo por determinado público ou finalidade. Ele pode assumir diferentes formas, como uma tabela curada, um data mart, um dashboard, uma API, um relatório, um indicador ou uma base analítica.

O Módulo 8 também marca a fronteira entre o pipeline principal e os módulos transversais. O pipeline principal produz os ativos; os módulos transversais garantem que esses ativos sejam permanentemente documentados, governados, rastreados, versionados, monitorados e apoiados por mecanismos inteligentes.

Dessa forma, a disponibilização e o consumo analítico constituem a etapa em que o valor produzido pelo pipeline se materializa em uso efetivo.

5. Formalização sintética do módulo

O Módulo 8 pode ser compreendido como um processo de transformação de ativos analíticos armazenados em produtos de dados consumíveis.

Considerando o conjunto de ativos analíticos armazenados no Módulo 7:

$$\mathbf{A = \{a_1, a_2, ..., a_n\}}$$

o Módulo 8 define um conjunto de produtos ou interfaces de consumo:

$$\mathbf{P = \{p_1, p_2, ..., p_k\}}$$

em que cada $\mathbf{p_i}$ representa um produto de dados, dashboard, API, view, relatório, indicador, dataset curado, base de modelagem ou outro mecanismo de consumo.

Cada produto de dados pode ser descrito por:

$$\mathbf{p_i = \{id, nome, tipo, finalidade, público_alvo, ativo_origem, formato, granularidade, periodicidade, versão, qualidade, restrição_acesso, mecanismo_consumo, limitações\}}$$

Além disso, o módulo registra o conjunto de consumidores:

$$\mathbf{C = \{c_1, c_2, ..., c_m\}}$$

em que cada consumidor $\mathbf{c_j}$ pode representar uma pessoa, equipe, sistema, dashboard, API, modelo, agente inteligente ou processo decisório.

Cada relação de consumo pode ser representada por:

$$\mathbf{r_{ij} = \{produto, consumidor, finalidade, frequência, nível_acesso, data_acesso, restrições, feedback\}}$$

O resultado do módulo é um conjunto de ativos disponibilizados, documentados, acessíveis, controlados e adequados ao uso.

6. Entradas do módulo

As entradas do Módulo 8 incluem os ativos analíticos persistidos e organizados no Módulo 7, bem como metadados, regras, restrições, métricas e requisitos de consumo.

As principais entradas são:

- datasets curados;
- data marts;
- views analíticas;
- tabelas agregadas;
- bases de modelagem;
- estruturas para grafos;
- indicadores calculados;
- datasets validados;
- datasets balanceados ou sinalizados;
- registros de qualidade;
- registros de sensibilidade;
- restrições de acesso;
- versões dos ativos;
- schemas analíticos;
- requisitos de usuários;
- requisitos de dashboards;
- requisitos de APIs;
- requisitos de relatórios;
- requisitos de modelos;
- requisitos de atualização;
- políticas de acesso;
- políticas de anonimização;
- políticas de retenção;
- documentação dos ativos;
- informações de lineage;
- critérios de aprovação para disponibilização.

7. Processamento

O processamento do Módulo 8 envolve atividades de seleção, preparação, publicação, controle, documentação, disponibilização e avaliação do consumo dos ativos analíticos.

7.1 Identificação dos consumidores e finalidades de uso

A primeira atividade consiste em identificar quem irá consumir os ativos produzidos pelo pipeline e para qual finalidade. Os consumidores podem ser gestores, pesquisadores, analistas, cientistas de dados, sistemas institucionais, dashboards, APIs, modelos computacionais, agentes inteligentes ou aplicações externas.

Cada consumidor pode exigir granularidade, formato, atualização, qualidade e restrição de acesso diferentes.

7.2 Seleção dos ativos aptos para disponibilização

A segunda atividade consiste em selecionar os ativos que podem ser disponibilizados. Essa seleção deve considerar status de qualidade, sensibilidade, restrição de acesso, completude documental, versão, finalidade de uso e eventuais riscos de viés ou equidade.

Ativos rejeitados, pendentes ou com risco crítico não devem ser disponibilizados sem revisão, sinalização ou restrição.

7.3 Definição do formato de disponibilização

A terceira atividade consiste em definir o formato mais adequado para cada ativo. O produto pode ser disponibilizado como dataset, tabela, view, arquivo, API, dashboard, relatório, indicador, grafo, modelo semântico, base de treinamento, feature table ou produto de dados institucional.

A escolha depende da finalidade de consumo, volume, frequência, público-alvo, sensibilidade e requisitos de desempenho.

7.4 Preparação para consumo

A quarta atividade consiste em preparar o ativo para consumo. Isso pode envolver agregação, filtragem, anonimização, pseudonimização, mascaramento, criação de views, materialização de consultas, geração de arquivos de exportação, criação de endpoints, configuração de dashboards ou publicação de relatórios.

Essa etapa deve preservar vínculo com o ativo de origem e registrar eventuais alterações realizadas especificamente para disponibilização.

7.5 Definição de regras de acesso

A quinta atividade consiste em definir quem pode acessar cada produto de dados, em qual nível de detalhe e sob quais condições. A regra de acesso deve considerar classificação de sensibilidade, finalidade de uso, perfil do consumidor, necessidade de anonimização e restrições legais ou institucionais.

Podem existir diferentes níveis de acesso, como público, interno, restrito, agregado, anonimizado, individualizado sob autorização ou bloqueado.

7.6 Publicação dos produtos de dados

A sexta atividade consiste em publicar o ativo no ambiente adequado. A publicação pode ocorrer em catálogo, portal interno, dashboard, API, repositório controlado, banco de dados analítico, ambiente de BI, ambiente de ciência de dados ou sistema institucional.

Cada publicação deve gerar registro de versão, data, responsável, status, consumidores autorizados e mecanismo de acesso.

7.7 Documentação mínima para consumo

A sétima atividade consiste em garantir que todo ativo disponibilizado possua documentação mínima. Essa documentação deve indicar finalidade, origem, versão, cobertura temporal, granularidade, atualização, qualidade, limitações, restrições de uso e responsável.

Essa documentação é fundamental para evitar interpretações equivocadas e uso indevido dos dados.

7.8 Registro de consumo

A oitava atividade consiste em registrar o uso dos produtos de dados. Isso pode incluir consumidores autorizados, frequência de acesso, aplicações dependentes, dashboards conectados, modelos consumidores, relatórios gerados e feedback recebido.

Esse registro permite avaliar utilidade, dependência, reuso e impacto dos ativos disponibilizados.

7.9 Avaliação da utilidade e adequação ao uso

A nona atividade consiste em avaliar se os produtos disponibilizados atendem às necessidades dos consumidores. Essa avaliação pode considerar desempenho, clareza, atualização, disponibilidade, facilidade de uso, aderência à finalidade e percepção de utilidade.

7.10 Encerramento do pipeline principal

A décima atividade consiste em registrar que o ativo percorreu o pipeline principal e foi entregue para consumo. A partir desse ponto, os módulos transversais passam a atuar sobre o ativo de forma contínua, mantendo governança, metadados, lineage, versionamento, monitoramento e apoio inteligente.

8. Técnicas, métodos e tecnologias aplicáveis

O Módulo 8 pode empregar diferentes técnicas e tecnologias para disponibilização e consumo.

Entre as técnicas de disponibilização, destacam-se:

- publicação de datasets curados;
- criação de data marts;
- criação de views analíticas;
- criação de views materializadas;
- publicação de APIs;
- geração de arquivos de exportação;
- construção de dashboards;
- publicação de relatórios analíticos;
- construção de indicadores;
- disponibilização de bases para modelagem;
- disponibilização de tabelas de features;
- disponibilização de estruturas de grafos;
- disponibilização de dados agregados ou anonimizados;
- criação de produtos de dados.

Entre as técnicas de controle e adequação ao consumo, podem ser utilizadas:

- controle de acesso baseado em papéis;
- anonimização;
- pseudonimização;

- mascaramento;
- agregação;
- filtragem por perfil de usuário;
- controle de granularidade;
- limitação de exportação;
- auditoria de acesso;
- versionamento de produtos;
- documentação de uso;
- termos de uso;
- sinalização de qualidade;
- alertas de limitação;
- registro de consumidores.

Entre as tecnologias possíveis, destacam-se:

- Power BI;
- Metabase;
- Apache Superset;
- Tableau;
- Looker;
- APIs REST;
- GraphQL;
- bancos relacionais;
- bancos colunares;
- data warehouses;
- lakehouses;
- catálogos de dados;
- portais de dados;
- repositórios Git;
- object storage;
- notebooks;
- ferramentas de BI;
- ferramentas de visualização;
- gateways de API;
- ambientes de ciência de dados;
- plataformas de MLOps;
- bancos de grafos.

9. Métricas do módulo

As métricas do Módulo 8 avaliam cobertura, acesso, uso, desempenho, atualização, qualidade percebida, aderência e reuso dos produtos disponibilizados.

Métrica	Sigla	Descrição
Taxa de ativos disponibilizados	TAD	Proporção de ativos aptos que foram disponibilizados
Taxa de produtos documentados	TPD	Proporção de produtos com documentação

		mínima
Taxa de produtos com qualidade informada	TPQI	Proporção de produtos com status/métricas de qualidade
Taxa de produtos com versão definida	TPVD	Proporção de produtos com versão registrada
Taxa de produtos com controle de acesso	TPCA	Proporção de produtos com regras de acesso
Tempo médio de disponibilização	TMD	Tempo entre ativo curado e produto publicado
Taxa de consumo efetivo	TCE	Proporção de produtos acessados por consumidores
Taxa de reutilização de produtos	TRP	Proporção de produtos usados em mais de uma finalidade
Frequência de acesso	FA	Quantidade de acessos por produto em determinado período
Tempo médio de resposta	TMR	Tempo de resposta de dashboards, APIs ou consultas
Taxa de atualização conforme periodicidade	TACP	Proporção de produtos atualizados conforme previsto
Taxa de produtos obsoletos	TPO	Proporção de produtos desatualizados ou substituídos
Satisfação do usuário	SU	Avaliação qualitativa ou quantitativa dos

		consumidores
Taxa de incidentes de acesso	TIA	Proporção de falhas, erros ou acessos negados indevidos
Taxa de adequação ao uso	TAU	Proporção de produtos avaliados como adequados à finalidade

9.1 Taxa de Ativos Disponibilizados — TAD

A **Taxa de Ativos Disponibilizados** mede a proporção de ativos considerados aptos para consumo que foram efetivamente disponibilizados.

TAD = número de ativos disponibilizados / número total de ativos aptos para disponibilização

Essa métrica indica a capacidade do pipeline de transformar ativos curados em produtos consumíveis.

9.2 Taxa de Produtos Documentados — TPD

A **Taxa de Produtos Documentados** mede a proporção de produtos de dados que possuem documentação mínima para uso.

TPD = número de produtos com documentação mínima / número total de produtos disponibilizados

Um produto pode ser considerado minimamente documentado quando possui finalidade, origem, versão, granularidade, qualidade, atualização, restrição de acesso e limitações.

9.3 Taxa de Consumo Efetivo — TCE

A **Taxa de Consumo Efetivo** mede a proporção de produtos disponibilizados que foram acessados ou utilizados por consumidores previstos.

TCE = número de produtos consumidos / número total de produtos disponibilizados

Essa métrica permite avaliar se a disponibilização gerou uso real.

9.4 Taxa de Reutilização de Produtos — TRP

A **Taxa de Reutilização de Produtos** mede a proporção de produtos utilizados em mais de uma análise, dashboard, modelo, relatório ou aplicação.

TRP = número de produtos reutilizados / número total de produtos disponibilizados

Essa métrica avalia o grau de reaproveitamento dos ativos produzidos.

9.5 Taxa de Adequação ao Uso — TAU

A **Taxa de Adequação ao Uso** mede a proporção de produtos avaliados como adequados à finalidade para a qual foram disponibilizados.

TAU = número de produtos adequados ao uso / número total de produtos avaliados

Essa métrica pode combinar avaliação técnica, feedback de usuários, desempenho, clareza, qualidade e atualidade.

10. Artefatos formalizados do módulo

Os artefatos do Módulo 8 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, regras de construção, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A8.1	Plano de Disponibilização Analítica	Documento técnico	TAD	Define produtos, públicos e formas de acesso
A8.2	Catálogo de Produtos de Dados	Catálogo estruturado	TPD	Organiza produtos consumíveis
A8.3	Matriz de Consumidores e Acessos	Matriz de acesso	TPCA	Controla quem pode consumir cada ativo
A8.4	Especificação de Produto de Dados	Ficha técnica	TPD, TPQI	Documenta finalidade, origem e limitações
A8.5	API, View ou Interface de Consumo	Endpoint, view ou interface	TMR, TCE	Materializa o acesso ao ativo
A8.6	Dashboard ou Painel Analítico	Interface visual	SU, TCE	Disponibiliza indicadores e visualizações

A8.7	Dataset Curado para Consumo	Tabela, arquivo ou data mart	TAD, TRP	Entrega dados preparados ao usuário
A8.8	Registro de Publicação e Versão	Manifesto/tabela	TPVD	Controla versões publicadas
A8.9	Registro de Uso e Feedback	Log ou relatório	TCE, SU	Avalia consumo e utilidade
A8.10	Relatório de Adequação ao Consumo	Relatório técnico	TAU	Avalia se o produto atende à finalidade

11. Artefato A8.1 — Plano de Disponibilização Analítica

11.1 Finalidade

O **Plano de Disponibilização Analítica** define quais ativos serão disponibilizados, para quais consumidores, em quais formatos, com quais restrições, em qual periodicidade e com quais critérios de qualidade.

11.2 Tipo de artefato na DSR

Na perspectiva da Design Science Research, esse artefato é um **método**, pois orienta o processo de disponibilização, e uma **instanciação**, quando documentado em planilha, relatório, arquivo estruturado ou ferramenta de gestão.

11.3 Estrutura mínima

Campo	Descrição
Produto previsto	Nome do produto de dados
Ativo de origem	Dataset, view, data mart ou indicador
Público-alvo	Gestor, pesquisador, sistema, dashboard, modelo etc.

Finalidade	Uso previsto
Formato	API, dashboard, dataset, relatório, view etc.
Granularidade	Nível de detalhe disponibilizado
Periodicidade	Frequência de atualização
Qualidade mínima	Status ou métricas exigidas
Restrição de acesso	Público, interno, restrito, anonimizado etc.
Responsável	Pessoa ou equipe responsável
Critério de publicação	Condição para disponibilizar

11.4 Métricas associadas

- taxa de ativos disponibilizados;
- percentual de produtos com público definido;
- percentual de produtos com periodicidade definida;
- percentual de produtos com critério de publicação;
- percentual de produtos com restrição de acesso definida.

11.5 Critérios de validação

O plano é considerado válido quando cada produto possui finalidade, público-alvo, formato, ativo de origem, qualidade mínima, restrição de acesso e responsável definidos.

11.6 Contribuição

Esse artefato evita que os dados sejam disponibilizados de forma improvisada, sem clareza sobre finalidade, público, formato e responsabilidade.

12. Artefato A8.2 — Catálogo de Produtos de Dados

12.1 Finalidade

O **Catálogo de Produtos de Dados** organiza os ativos disponibilizados para consumo, permitindo sua localização, compreensão e reutilização.

12.2 Estrutura mínima

Campo	Descrição
Identificador do produto	Código único
Nome do produto	Nome do dataset, dashboard, API, indicador etc.
Tipo	Dataset, API, dashboard, view, relatório, modelo etc.
Descrição	Explicação do produto
Finalidade	Uso previsto
Público-alvo	Consumidores esperados
Ativo de origem	Dados utilizados
Versão	Versão publicada
Qualidade	Status ou métricas associadas
Atualização	Periodicidade
Restrição de acesso	Regra de acesso
Responsável	Responsável pelo produto

Limitações	Cuidados de uso

12.3 Métricas associadas

- taxa de produtos documentados;
- taxa de produtos com qualidade informada;
- taxa de produtos com versão definida;
- taxa de produtos com responsável definido;
- taxa de reutilização de produtos.

12.4 Contribuição

Esse artefato organiza a camada de consumo do modelo de referência, permitindo que os produtos sejam encontrados, compreendidos, avaliados e reutilizados.

13. Artefato A8.3 — Matriz de Consumidores e Acessos

13.1 Finalidade

A **Matriz de Consumidores e Acessos** define quem pode consumir cada produto de dados e sob quais condições.

13.2 Estrutura mínima

Campo	Descrição
Produto	Produto de dados
Consumidor	Pessoa, equipe, sistema ou aplicação
Perfil	Gestor, analista, pesquisador, sistema etc.
Tipo de acesso	Leitura, exportação, API, dashboard, consulta
Granularidade permitida	Individual, agregado, anonimizado etc.
Restrição	Condição específica de uso

Validade	Período de autorização
Responsável pela autorização	Pessoa ou setor
Observações	Cuidados adicionais

13.3 Métricas associadas

- taxa de produtos com controle de acesso;
- número de consumidores autorizados;
- número de acessos expirados;
- taxa de produtos sensíveis com acesso restrito;
- taxa de incidentes de acesso.

13.4 Contribuição

Esse artefato garante que a disponibilização respeite segurança, privacidade, governança e finalidade de uso.

14. Artefato A8.4 — Especificação de Produto de Dados

14.1 Finalidade

A **Especificação de Produto de Dados** descreve um produto disponibilizado de forma padronizada, permitindo seu uso adequado.

14.2 Estrutura formal

Cada produto de dados pode ser descrito por:

p_i = {id, nome, tipo, finalidade, público, origem, formato, granularidade, qualidade, versão, acesso, limitações}

14.3 Estrutura mínima

Campo	Descrição
Nome do produto	Nome oficial
Tipo de produto	Dataset, API, dashboard, view, relatório

	etc.
Finalidade	Problema ou uso que atende
Público-alvo	Consumidores esperados
Fontes de origem	Bases utilizadas
Pipeline de origem	Processo que gerou o produto
Granularidade	Nível de detalhe
Cobertura temporal	Período contemplado
Indicadores ou atributos	Elementos principais
Métricas de qualidade	Status de qualidade
Atualização	Periodicidade
Versão	Versão publicada
Restrição de acesso	Condições de uso
Limitações	Cuidados de interpretação
Responsável	Responsável técnico ou institucional

14.4 Métricas associadas

- taxa de produtos documentados;
- completude da especificação;
- taxa de produtos com qualidade informada;
- taxa de produtos com limitações explícitas;
- taxa de produtos com versão definida.

14.5 Critérios de validação

A especificação é considerada adequada quando permite compreender finalidade, origem, qualidade, versão, restrições, limitações e forma de uso do produto.

14.6 Contribuição

Esse artefato transforma a disponibilização em um produto documentado, evitando uso inadequado ou interpretações equivocadas.

15. Artefato A8.5 — API, View ou Interface de Consumo

15.1 Finalidade

A **API, View ou Interface de Consumo** materializa o acesso controlado ao ativo analítico.

15.2 Possíveis formas

Forma	Descrição
API REST	Acesso programático por endpoints
GraphQL	Consulta flexível por aplicações
View SQL	Consulta controlada em banco analítico
View materializada	Consulta otimizada e pré-computada
Arquivo exportável	CSV, Parquet, JSON ou outro formato
Interface web	Acesso via portal ou aplicação
Endpoint para	Consulta por agentes inteligentes

agente

15.3 Estrutura mínima

Campo	Descrição
Produto associado	Produto exposto
Tipo de interface	API, view, arquivo, endpoint etc.
Local de acesso	URL, schema, tabela, caminho etc.
Parâmetros	Filtros ou argumentos disponíveis
Formato de saída	JSON, CSV, tabela, gráfico etc.
Autenticação	Necessária ou não
Restrição	Condições de acesso
Tempo de resposta esperado	Requisito de desempenho
Versão	Versão da interface

15.4 Métricas associadas

- tempo médio de resposta;
- taxa de consumo efetivo;
- taxa de consultas bem-sucedidas;
- taxa de incidentes de acesso;
- disponibilidade da interface.

15.5 Contribuição

Esse artefato operacionaliza o consumo dos dados, permitindo acesso controlado, programático ou interativo aos produtos analíticos.

16. Artefato A8.6 — Dashboard ou Painel Analítico

16.1 Finalidade

O **Dashboard ou Painel Analítico** disponibiliza indicadores, visualizações, filtros e informações analíticas para usuários finais.

16.2 Estrutura mínima

Campo	Descrição
Nome do painel	Identificação do dashboard
Objetivo	Finalidade decisória ou analítica
Público-alvo	Usuários previstos
Indicadores	Métricas apresentadas
Fontes de dados	Ativos utilizados
Granularidade	Nível de agregação
Filtros	Recortes disponíveis
Atualização	Periodicidade
Responsável	Equipe responsável

Restrições	Limitações de acesso ou interpretação
Versão	Versão publicada

16.3 Métricas associadas

- taxa de consumo efetivo;
- frequência de acesso;
- satisfação do usuário;
- tempo de carregamento;
- taxa de atualização conforme periodicidade.

16.4 Contribuição

Esse artefato transforma dados em visualizações e indicadores acessíveis, apoiando análise exploratória, acompanhamento e tomada de decisão.

17. Artefato A8.7 — Dataset Curado para Consumo

17.1 Finalidade

O **Dataset Curado para Consumo** é um ativo de dados preparado para uso direto por analistas, pesquisadores, modelos, dashboards ou sistemas.

17.2 Estrutura mínima

Campo	Descrição
Nome do dataset	Identificação
Finalidade	Uso previsto
Fontes de origem	Bases utilizadas
Estrutura	Atributos e tipos
Granularidade	Unidade de análise

Cobertura temporal	Período abrangido
Qualidade	Status e métricas
Versão	Versão publicada
Formato	CSV, Parquet, SQL, API etc.
Restrição de uso	Condições de acesso
Limitações	Cuidados de interpretação

17.3 Métricas associadas

- taxa de ativos disponibilizados;
- taxa de reutilização;
- taxa de produtos documentados;
- taxa de produtos com versão;
- taxa de adequação ao uso.

17.4 Contribuição

Esse artefato entrega dados prontos para uso, reduzindo retrabalho e promovendo reutilização de ativos confiáveis.

18. Artefato A8.8 — Registro de Publicação e Versão

18.1 Finalidade

O **Registro de Publicação e Versão** documenta quando, como e em qual versão um produto de dados foi disponibilizado.

18.2 Estrutura mínima

Campo	Descrição
Produto	Produto publicado

Versão	Identificador da versão
Data de publicação	Momento da publicação
Ativo de origem	Dataset ou data mart utilizado
Alterações	Mudanças em relação à versão anterior
Responsável	Pessoa ou sistema responsável
Status	Ativa, substituída, obsoleta, descontinuada
Consumidores impactados	Usuários ou sistemas afetados
Observações	Restrições ou alertas

18.3 Métricas associadas

- taxa de produtos com versão definida;
- número de versões por produto;
- taxa de produtos obsoletos;
- tempo médio entre versões;
- taxa de produtos descontinuados.

18.4 Contribuição

Esse artefato evita uso de produtos obsoletos e apoia rastreabilidade, auditoria e reprodutibilidade.

19. Artefato A8.9 — Registro de Uso e Feedback

19.1 Finalidade

O **Registro de Uso e Feedback** documenta o consumo dos produtos disponibilizados e a percepção dos usuários.

19.2 Estrutura mínima

Campo	Descrição
Produto	Produto consumido
Consumidor	Usuário, equipe, sistema ou aplicação
Tipo de uso	Consulta, dashboard, modelo, relatório, API etc.
Frequência	Periodicidade de uso
Data de acesso	Momento do acesso
Problemas relatados	Erros, dúvidas ou limitações
Feedback	Avaliação do consumidor
Solicitações	Melhorias ou novas demandas
Status	Aberto, analisado, resolvido

19.3 Métricas associadas

- taxa de consumo efetivo;
- frequência de acesso;
- satisfação do usuário;
- número de solicitações de melhoria;
- taxa de problemas resolvidos.

19.4 Contribuição

Esse artefato permite avaliar o valor real dos produtos disponibilizados e orientar melhoria contínua.

20. Artefato A8.10 — Relatório de Adequação ao Consumo

20.1 Finalidade

O **Relatório de Adequação ao Consumo** avalia se os produtos disponibilizados atendem às necessidades, restrições e finalidades dos consumidores.

20.2 Estrutura mínima

Campo	Descrição
Produto avaliado	Produto de dados
Consumidores	Usuários ou sistemas avaliados
Finalidade	Uso previsto
Desempenho	Tempo de resposta, disponibilidade
Documentação	Completude da documentação
Qualidade percebida	Avaliação do consumidor
Atualização	Aderência à periodicidade
Restrições	Problemas de acesso ou uso
Adequação	Adequado, adequado com ressalvas, inadequado
Recomendações	Melhorias sugeridas

20.3 Métricas associadas

- taxa de adequação ao uso;
- satisfação do usuário;
- tempo médio de resposta;
- taxa de produtos documentados;
- taxa de atualização conforme periodicidade.

20.4 Contribuição

Esse artefato fecha o ciclo do pipeline principal, verificando se o produto de dados realmente atende à finalidade para a qual foi construído.

21. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Datasets curados, data marts, views, indicadores, métricas de qualidade, versões, restrições de acesso, requisitos de consumo e documentação dos ativos.
Processamento	Identificação dos consumidores, seleção dos ativos, definição do formato de disponibilização, preparação para consumo, controle de acesso, publicação, documentação e avaliação do uso.
Artefatos	Plano de disponibilização, catálogo de produtos, matriz de consumidores, especificação de produto, APIs/views/interfaces, dashboards, datasets curados, registro de publicação, uso e relatório de adequação.
Métricas	TAD, TPD, TPQI, TPVD, TPCA, TMD, TCE, TRP, FA, TMR, TACP, TPO, SU, TIA e TAU.
Saídas	Produtos de dados disponibilizados, documentados, versionados, controlados, consumidos e avaliados quanto à adequação ao uso.

22. Critérios gerais de validação do módulo

O Módulo 8 pode ser considerado concluído quando:

1. os consumidores e finalidades de uso foram identificados;
2. os ativos aptos para consumo foram selecionados;

3. os produtos de dados foram definidos;
4. cada produto possui formato de disponibilização adequado;
5. os produtos possuem documentação mínima;
6. os produtos possuem versão registrada;
7. as regras de acesso foram definidas;
8. os produtos sensíveis foram agregados, anonimizados, filtrados ou restritos quando necessário;
9. os produtos foram publicados em ambiente adequado;
10. o uso ou consumo dos produtos foi registrado;
11. a adequação ao uso foi avaliada;
12. os produtos disponibilizados estão conectados aos módulos transversais de metadados, governança, lineage, versionamento e monitoramento.

23. Relação com os demais módulos do pipeline principal

O Módulo 8 é a etapa final do pipeline principal do UnespDataLens-RM.

Ele depende do **Módulo 1 — Fontes e Inventário de Dados**, pois utiliza informações sobre origem, finalidade, sensibilidade e restrições das fontes.

Depende do **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, pois os produtos disponibilizados precisam manter vínculo com os dados originalmente ingeridos.

Depende do **Módulo 3 — Integração e Harmonização**, pois muitos produtos de dados são derivados de bases integradas e harmonizadas.

Depende do **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, pois consome datasets tratados, padronizados e enriquecidos.

Depende do **Módulo 5 — Qualidade e Validação de Dados**, pois somente ativos com qualidade conhecida devem ser disponibilizados para consumo.

Depende do **Módulo 6 — Desbalanceamento, Vieses e Equidade**, pois produtos com riscos de viés, desbalanceamento ou limitações devem ser sinalizados, restringidos ou documentados.

Depende do **Módulo 7 — Armazenamento Analítico**, pois utiliza os ativos persistidos, data marts, views, tabelas e estruturas organizadas para consumo.

Com o Módulo 8, o pipeline principal se completa: as fontes foram identificadas, os dados foram extraídos, integrados, transformados, validados, avaliados quanto a vieses, armazenados analiticamente e finalmente disponibilizados para uso.

24. Relação com os módulos transversais

Embora encerre o pipeline principal, o Módulo 8 não encerra o ciclo de vida dos dados. A partir dele, tornam-se fundamentais os módulos transversais do UnespDataLens-RM.

O módulo de **Metadados, Catálogo e Semântica dos Dados** deve registrar e manter a documentação dos produtos disponibilizados, incluindo descrições, versões, qualidade, finalidade, responsáveis e limitações.

O módulo de **Governança, Segurança, Privacidade e Conformidade** deve controlar políticas de acesso, finalidade de uso, autorização, sensibilidade, anonimização, LGPD e responsabilidades.

O módulo de **Proveniência, Lineage e Rastreabilidade** deve permitir reconstruir a trajetória dos produtos desde as fontes originais até o consumo final.

O módulo de **Versionamento e Reprodutibilidade** deve preservar versões de produtos, datasets, dashboards, APIs, relatórios e configurações.

O módulo de **Representações Analíticas: Tabular, Temporal, Espacial, Textual, Semântica e Grafos** pode consumir produtos disponibilizados para gerar novas estruturas analíticas especializadas.

O módulo de **Feature Engineering e Feature Store** pode consumir datasets curados para gerar, armazenar e reutilizar features.

O módulo de **Monitoramento, Drift e Evolução dos Ativos** deve acompanhar atualização, uso, qualidade, obsolescência, alterações de schema, degradação de produtos e mudanças de distribuição.

O módulo de **Agentes Inteligentes de Apoio ao Pipeline** pode usar os produtos disponibilizados, seus metadados e registros de consumo para apoiar consultas, recomendações, explicações, descoberta de dados e apoio à tomada de decisão.

25. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 8 poderia disponibilizar produtos de dados derivados de matrículas, histórico acadêmico, notas, frequência, evasão, interações em ambientes virtuais, dados socioeconômicos e indicadores de permanência.

Exemplos de produtos disponibilizados incluem:

- dashboard de evasão por curso, turma e período;
- dataset curado de trajetória acadêmica;
- API de indicadores educacionais;
- data mart de permanência estudantil;
- relatório de desempenho por disciplina;
- painel de acompanhamento de estudantes em risco;
- base anonimizada para pesquisa;
- view agregada de indicadores por curso;
- estrutura de grafo para análise de relações estudante-disciplina-curso.

Cada produto deve possuir documentação mínima, versão, responsável, cobertura temporal, qualidade, restrição de acesso e limitações de interpretação.

Por exemplo, um painel de evasão não deve apenas apresentar percentuais. Ele deve indicar origem dos dados, período de atualização, definição de evasão adotada, nível de agregação, limitações conhecidas, qualidade da base e restrições de uso.

26. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, o Módulo 8 poderia disponibilizar produtos de dados voltados à análise territorial, acompanhamento de programas, monitoramento de serviços públicos e apoio à decisão governamental.

Exemplos de produtos disponibilizados incluem:

- painel de vulnerabilidade territorial;
- API de indicadores municipais;
- data mart de famílias atendidas;
- dashboard de cobertura de serviços;
- relatório de demandas por território;
- dataset agregado por bairro;
- mapa de equipamentos públicos;
- base anonimizada para estudos técnicos;
- view de indicadores de saúde, educação e assistência social.

Cada produto deve respeitar níveis de sensibilidade e privacidade. Dados individuais devem ser disponibilizados apenas mediante autorização e com controles adequados, enquanto dados públicos podem ser agregados e publicados em painéis ou relatórios.

27. Contribuição técnico-científica do módulo

A contribuição do Módulo 8 está em formalizar a disponibilização e o consumo analítico como a etapa final do pipeline principal de Engenharia Analítica de Dados.

Em vez de tratar o consumo como consequência natural do armazenamento, o UnespDataLens-RM propõe que a disponibilização seja planejada, documentada, controlada, avaliada e vinculada à finalidade de uso.

Essa formalização contribui para:

- transformar ativos analíticos em produtos de dados consumíveis;
- evitar disponibilização sem qualidade conhecida;
- reduzir riscos de uso indevido;
- controlar acesso conforme sensibilidade;
- documentar finalidade, versão, origem e limitações dos produtos;
- aumentar reuso e valor dos ativos produzidos;
- registrar consumidores e formas de uso;
- avaliar utilidade e adequação dos produtos;
- conectar o pipeline principal aos módulos transversais;
- apoiar decisões baseadas em evidências de forma mais confiável.

28. Síntese do módulo

O **Módulo 8 — Disponibilização e Consumo Analítico** encerra o pipeline principal do UnespDataLens-RM. Ele transforma ativos analíticos armazenados em produtos de dados acessíveis, documentados, versionados, controlados e adequados ao uso.

Sua importância está em materializar o valor produzido pelos módulos anteriores. O pipeline só gera impacto quando seus resultados são consumidos por pessoas, sistemas, modelos, dashboards, APIs, relatórios, grafos, indicadores ou processos decisórios.

Ao formalizar a disponibilização como uma etapa com artefatos, métricas, regras de acesso, documentação e avaliação de uso, o Módulo 8 fortalece a confiabilidade, a governança, a transparência e a utilidade dos ativos produzidos pelo modelo de referência.

A partir desse ponto, os módulos seguintes deixam de compor o fluxo linear principal e passam a atuar como **módulos transversais**, responsáveis por garantir metadados, governança, privacidade, lineage, versionamento, reprodutibilidade, monitoramento, representação especializada, feature store e apoio inteligente ao longo de todo o ciclo de vida dos dados.

Módulo Transversal 9 — Metadados, Catálogo e Semântica dos Dados

1. Finalidade do módulo

O **Módulo Transversal 9 — Metadados, Catálogo e Semântica dos Dados** tem como finalidade estruturar, organizar, documentar, padronizar, enriquecer semanticamente e disponibilizar informações descritivas sobre todos os ativos de dados produzidos, utilizados ou disponibilizados no modelo de referência **UnespDataLens-RM**.

Diferentemente dos módulos 1 a 8, que compõem o pipeline técnico-operacional principal — desde a identificação das fontes até a disponibilização e consumo analítico —, o Módulo 9 atua de forma **transversal**. Isso significa que ele não ocorre apenas em uma etapa isolada do pipeline, mas acompanha todo o ciclo de vida dos dados, registrando metadados, descrições, conceitos, relações, regras, versões, qualidade, sensibilidade, restrições, responsáveis e significados dos ativos produzidos em cada módulo.

Esse módulo é responsável por garantir que os dados não sejam tratados apenas como arquivos, tabelas, colunas, registros, datasets, dashboards ou APIs, mas como **ativos informacionais compreensíveis, documentados, localizáveis, reutilizáveis, governáveis, auditáveis e semanticamente interpretáveis**.

Em projetos de Engenharia Analítica de Dados, uma das fragilidades recorrentes é a produção de datasets, scripts, indicadores e produtos analíticos sem documentação adequada. Muitas vezes, os dados são extraídos, integrados, limpos, transformados, validados e disponibilizados, mas permanecem pouco compreensíveis para usuários que não participaram diretamente do pipeline. Essa ausência de documentação compromete a reutilização, dificulta auditorias, reduz a confiança, aumenta dependência de conhecimento tácito e favorece interpretações equivocadas.

No contexto do UnespDataLens-RM, o Módulo 9 tem papel central porque transforma informações técnicas, operacionais e semânticas geradas ao longo de todo o pipeline em **metadados estruturados e consultáveis**. Ele organiza um catálogo de dados, consolida dicionários, registra glossários, explicita relações semânticas, documenta produtos de dados, associa ativos a métricas de qualidade, registra restrições de acesso e prepara a base documental para governança, lineage, versionamento, monitoramento e agentes inteligentes.

Assim, o Módulo 9 não deve ser entendido como uma etapa final de documentação. Ele constitui uma infraestrutura transversal de conhecimento sobre os dados, responsável por tornar o ecossistema analítico compreensível, navegável e reutilizável.

2. Objetivo geral

O objetivo geral do Módulo 9 é **formalizar, estruturar e manter metadados técnicos, semânticos, operacionais, analíticos, de qualidade e de governança sobre os ativos do UnespDataLens-RM**, por meio de catálogos, dicionários, glossários, modelos semânticos e fichas de documentação que apoiem descoberta, compreensão, interoperabilidade, reuso, governança, rastreabilidade, explicabilidade e tomada de decisão baseada em evidências.

3. Objetivos específicos

São objetivos específicos do Módulo 9:

1. consolidar os metadados produzidos ao longo dos módulos do pipeline principal;
2. definir um modelo padronizado de metadados para fontes, datasets, atributos, entidades, indicadores, features, dashboards, APIs, data marts e produtos de dados;
3. organizar um catálogo de dados consultável, atualizado e integrado aos demais módulos do modelo de referência;
4. consolidar dicionários de dados a partir de fontes, datasets brutos, datasets integrados, datasets tratados, ativos armazenados e produtos disponibilizados;
5. registrar descrições semânticas de atributos, entidades, conceitos, indicadores, categorias e variáveis derivadas;
6. construir e manter glossários de termos do domínio, reduzindo ambiguidades terminológicas e conceituais;
7. mapear equivalências semânticas entre fontes, atributos, categorias, códigos e conceitos;
8. associar ativos de dados a informações de qualidade, sensibilidade, restrição de acesso, versão, origem, finalidade, responsável e limitação de uso;
9. apoiar a descoberta de dados por usuários humanos, sistemas computacionais e agentes inteligentes;
10. fornecer base documental para governança, lineage, reprodutibilidade, monitoramento, auditoria e uso responsável dos ativos analíticos;
11. definir métricas para avaliar cobertura, completude, atualidade, consistência, qualidade semântica, utilidade e reuso dos metadados.

4. Fundamentação conceitual do módulo

Metadados são informações que descrevem dados. No contexto do UnespDataLens-RM, os metadados assumem papel estrutural, pois permitem compreender o que são os dados, de onde vieram, como foram processados, qual é sua qualidade, quem é responsável por eles, quais restrições possuem, como podem ser utilizados e quais conceitos representam.

Os metadados podem ser classificados em diferentes categorias:

Categoria de metadado	Descrição	Exemplos
Metadados técnicos	Descrevem estrutura, formato e características computacionais dos dados.	Nome da tabela, tipo do atributo, schema, formato, encoding, chave, tamanho.
Metadados operacionais	Descrevem processos, execuções, atualizações e uso operacional.	Data de ingestão, pipeline, periodicidade, logs, status da execução.
Metadados	Descrevem significado dos	Definição de evasão, matrícula

semânticos	dados, conceitos e regras de negócio.	ativa, atendimento, vulnerabilidade, território.
Metadados de qualidade	Descrevem métricas, validações e status de qualidade dos ativos.	Completude, validade, consistência, unicidade, não conformidades, status de aprovação.
Metadados de governança	Descrevem responsabilidade, sensibilidade, acesso e conformidade.	Responsável, restrição de acesso, classificação de sensibilidade, finalidade de uso.
Metadados analíticos	Descrevem uso dos dados em análises, modelos, indicadores e dashboards.	Variável-alvo, features, indicadores, modelos consumidores, painéis associados.
Metadados de linhagem	Descrevem origem, dependências, transformações e trajetória dos dados.	Fonte original, regra aplicada, versão, dataset derivado, pipeline.
Metadados de consumo	Descrevem produtos disponibilizados e seus consumidores.	API, dashboard, relatório, público-alvo, frequência de acesso, feedback.

A dimensão semântica é especialmente importante. Em diferentes fontes ou instituições, o mesmo termo pode possuir significados distintos, enquanto termos diferentes podem representar conceitos equivalentes. Por exemplo, no domínio educacional, expressões como “evasão”, “abandono”, “cancelamento”, “trancamento”, “desistência” e “inatividade” podem indicar situações diferentes. Em políticas públicas, termos como “família”, “território”, “atendimento”, “benefício”, “vulnerabilidade” e “demanda” podem variar conforme o sistema, o órgão ou a regra institucional.

Sem documentação semântica, análises podem ser tecnicamente corretas e, ao mesmo tempo, conceitualmente equivocadas. Por essa razão, o Módulo 9 formaliza a semântica dos dados como elemento essencial da Engenharia Analítica de Dados.

À luz da **Design Science Research**, os artefatos do Módulo 9 podem ser compreendidos como constructos, modelos, métodos e instanciações. Eles formalizam o conhecimento sobre os dados, permitindo que o modelo de referência produza não apenas datasets e produtos analíticos, mas também uma estrutura de documentação, interpretação, descoberta e reuso.

5. Natureza transversal do módulo

O Módulo 9 é transversal porque acompanha e documenta todos os módulos do pipeline principal. Ele não ocorre apenas após a disponibilização dos dados. Ao contrário, seus artefatos são alimentados progressivamente desde o Módulo 1.

No **Módulo 1**, o Módulo 9 recebe informações sobre fontes, responsáveis, formatos, periodicidade, cobertura temporal, granularidade, restrições e sensibilidade.

No **Módulo 2**, recebe metadados de ingestão, logs, snapshots, manifestos, localizações e formatos dos dados brutos.

No **Módulo 3**, recebe regras de integração, chaves, tabelas de mapeamento, equivalências semânticas, conflitos e datasets integrados.

No **Módulo 4**, recebe regras de transformação, atributos derivados, parâmetros, fórmulas, registros de tratamento e datasets enriquecidos.

No **Módulo 5**, recebe métricas de qualidade, status dos ativos, não conformidades, critérios de aprovação e histórico de validação.

No **Módulo 6**, recebe informações sobre grupos, atributos sensíveis, desbalanceamento, vieses, métricas de equidade, riscos residuais e limitações de uso.

No **Módulo 7**, recebe informações sobre armazenamento analítico, zonas, schemas, localização, particionamento, versões e estruturas persistidas.

No **Módulo 8**, recebe informações sobre produtos de dados, consumidores, formas de acesso, dashboards, APIs, relatórios, versões publicadas e feedback de uso.

Dessa forma, o Módulo 9 atua como uma camada documental e semântica persistente, conectando o pipeline principal aos módulos transversais de governança, lineage, versionamento, monitoramento e agentes inteligentes.

6. Formalização sintética do módulo

O Módulo 9 pode ser compreendido como um processo de associação sistemática de metadados aos ativos produzidos pelo UnespDataLens-RM.

Considerando o conjunto de ativos de dados produzidos ou consumidos no modelo de referência:

$$\mathbf{A = \{a_1, a_2, ..., a_n\}}$$

em que cada $\mathbf{a_i}$ pode representar uma fonte, dataset bruto, dataset integrado, dataset tratado, data mart, atributo, feature, indicador, dashboard, API, relatório ou produto de dados, o módulo associa a cada ativo um conjunto de metadados:

$$\mathbf{M(a_i) = \{m_1, m_2, ..., m_k\}}$$

em que cada $\mathbf{m_j}$ representa um metadado técnico, operacional, semântico, analítico, de qualidade, de governança, de linhagem ou de consumo.

Cada ativo pode ser descrito formalmente por:

a_i = {id, nome, tipo, descrição, origem, responsável, domínio, formato, granularidade, periodicidade, versão, qualidade, sensibilidade, restrição_acesso, lineage, consumidores, limitações}

Além disso, o módulo define um conjunto de conceitos semânticos:

$C = \{c_1, c_2, ..., c_p\}$

em que cada conceito c_r pode ser descrito por:

c_r = {termo, definição, domínio, sinônimos, termos_relacionados, regra_negócio, exemplo, fonte_definição, status_validação}

Também pode ser definido um grafo semântico:

$G_s = (C, R)$

em que:

- **C** representa conceitos, entidades, atributos e indicadores;
- **R** representa relações semânticas, como equivalência, especialização, composição, associação, dependência, hierarquia ou temporalidade.

O resultado do módulo é um conjunto de artefatos documentais, semânticos e computacionais que tornam os ativos do UnespDataLens-RM compreensíveis, localizáveis, reutilizáveis, rastreáveis e governáveis.

7. Entradas do módulo

As entradas do Módulo 9 são produzidas ao longo de todo o pipeline principal e pelos demais módulos transversais.

As principais entradas são:

- inventário de fontes;
- mapa de bases;
- dicionário preliminar de dados;
- matriz de relacionamento entre fontes;
- registro de granularidade;
- classificação de sensibilidade;
- manifestos de ingestão;
- logs de execução;
- registros de snapshots;
- datasets brutos;
- planos de integração;
- tabelas de mapeamento de atributos;
- tabelas de equivalência semântica;
- registros de chaves e entidades;
- manifestos de integração;
- datasets integrados;
- registros de transformação;
- catálogos de atributos derivados;
- manifestos de transformação;
- datasets tratados e enriquecidos;

- planos de validação;
- relatórios de qualidade;
- registros de não conformidades;
- status de qualidade dos ativos;
- registros de grupos e atributos sensíveis;
- relatórios de vieses e equidade;
- registros de armazenamento analítico;
- schemas analíticos;
- data marts;
- views;
- APIs;
- dashboards;
- relatórios;
- produtos de dados;
- regras de acesso;
- informações de consumidores;
- glossários existentes;
- documentação institucional;
- regras de negócio;
- conhecimento de especialistas do domínio.

8. Processamento

O processamento do Módulo 9 envolve atividades contínuas de captura, consolidação, padronização, enriquecimento, curadoria, validação e disponibilização de metadados.

8.1 Captura de metadados ao longo do pipeline

A primeira atividade consiste em capturar metadados produzidos por cada módulo do pipeline. Essa captura pode ser manual, semiautomatizada ou automatizada, dependendo das ferramentas utilizadas.

O objetivo é evitar que informações importantes fiquem dispersas em scripts, notebooks, planilhas ou memória dos participantes do projeto.

8.2 Consolidação dos metadados

A segunda atividade consiste em consolidar os metadados capturados em uma estrutura padronizada. Informações sobre fonte, ingestão, integração, transformação, qualidade, armazenamento e consumo devem ser integradas em uma visão coerente do ativo de dados.

8.3 Padronização do modelo de metadados

A terceira atividade consiste em definir campos obrigatórios e opcionais para cada tipo de ativo. Uma fonte de dados, um atributo, um dataset curado, uma API e um dashboard não possuem exatamente os mesmos metadados, mas devem seguir um modelo compatível.

8.4 Construção e manutenção do catálogo de dados

A quarta atividade consiste em organizar os ativos em um catálogo consultável. O catálogo deve permitir localizar fontes, datasets, atributos, data marts, dashboards, APIs, indicadores, features, relatórios e produtos de dados.

8.5 Consolidação do dicionário de dados

A quinta atividade consiste em consolidar as descrições dos atributos produzidas ao longo do pipeline. O dicionário deve registrar nome técnico, nome semântico, descrição, tipo, domínio, formato, regra de preenchimento, origem, sensibilidade, exemplo de valor e limitações.

8.6 Construção do glossário de termos do domínio

A sexta atividade consiste em construir e manter um glossário de conceitos relevantes ao domínio. O glossário deve reduzir ambiguidades terminológicas e permitir que diferentes usuários compartilhem um vocabulário comum.

8.7 Enriquecimento semântico dos ativos

A sétima atividade consiste em associar atributos, entidades, indicadores, produtos e datasets a conceitos do glossário, taxonomias, regras de negócio ou modelos semânticos.

8.8 Mapeamento de equivalências semânticas

A oitava atividade consiste em registrar correspondências entre atributos, categorias, códigos e conceitos de diferentes fontes. Esse mapeamento apoia integração, harmonização, comparação entre bases e interpretação correta dos indicadores.

8.9 Curadoria e validação dos metadados

A nona atividade consiste em revisar e validar metadados críticos. A curadoria pode envolver especialistas do domínio, analistas de dados, gestores ou responsáveis pelas fontes.

Metadados incompletos, ambíguos, desatualizados ou contraditórios devem ser sinalizados e corrigidos.

8.10 Disponibilização dos metadados para uso humano e computacional

A décima atividade consiste em tornar os metadados disponíveis para consulta por usuários humanos, sistemas, ferramentas de BI, APIs, mecanismos de lineage, módulos de governança, componentes de monitoramento e agentes inteligentes.

9. Técnicas, métodos e tecnologias aplicáveis

O Módulo 9 pode empregar diferentes técnicas, métodos e tecnologias de documentação, catalogação e semântica.

Entre as técnicas de documentação e organização, destacam-se:

- catalogação de dados;
- modelagem de metadados;
- elaboração de dicionário de dados;
- construção de glossários;
- vocabulários controlados;
- taxonomias de domínio;
- modelagem conceitual;

- modelagem semântica;
- curadoria de metadados;
- documentação de regras de negócio;
- fichas de documentação de ativos;
- data cards;
- model cards;
- documentação de produtos de dados.

Entre as técnicas computacionais, podem ser utilizadas:

- extração automática de schema;
- inferência de tipos de dados;
- análise de cardinalidade;
- identificação de domínios de valores;
- indexação textual;
- busca semântica;
- classificação automática de metadados;
- sugestão automática de descrições;
- detecção de termos de domínio;
- vinculação entre atributos e conceitos;
- análise de similaridade entre nomes de atributos;
- agrupamento de atributos semelhantes;
- construção de grafos de conhecimento.

Entre os métodos e algoritmos aplicáveis, destacam-se:

- distância de Levenshtein;
- similaridade Jaro-Winkler;
- TF-IDF;
- embeddings textuais;
- similaridade por cosseno;
- clustering;
- classificação supervisionada;
- regras heurísticas;
- recuperação semântica de informações;
- busca vetorial;
- ontologias leves;
- grafos semânticos;
- modelos de linguagem para sugestão de documentação;
- validação assistida por especialista.

Ferramentas possíveis incluem Apache Atlas, DataHub, OpenMetadata, Amundsen, CKAN, dbt docs, Great Expectations Data Docs, repositórios Git, bancos relacionais, catálogos institucionais, bases vetoriais, APIs, planilhas estruturadas e interfaces de consulta.

10. Métricas do módulo

As métricas do Módulo 9 têm como objetivo avaliar cobertura, completude, consistência, atualidade, qualidade semântica, utilidade, descoberta e reuso dos metadados.

Métrica	Sigla	Descrição

Cobertura do catálogo	CC	Proporção de ativos registrados no catálogo
Índice de completude de metadados	ICM	Grau de preenchimento dos campos obrigatórios
Percentual de atributos documentados	PAD	Proporção de atributos com descrição, tipo e domínio
Percentual de atributos com definição semântica	PADS	Proporção de atributos com significado documentado
Percentual de ativos com responsável definido	PARD	Proporção de ativos com responsável identificado
Percentual de ativos com sensibilidade documentada	PASD	Proporção de ativos com classificação de sensibilidade
Percentual de ativos com qualidade documentada	PAQD	Proporção de ativos com métricas ou status de qualidade
Percentual de ativos com lineage associado	PALA	Proporção de ativos com origem e dependências registradas
Atualidade dos metadados	AM	Tempo desde a última atualização dos metadados
Consistência terminológica	CT	Grau de padronização dos termos usados
Cobertura do glossário	CG	Proporção de termos críticos definidos
Taxa de atributos ambíguos	TAA	Proporção de atributos com significado incerto

Taxa de equivalências semânticas validadas	TESV	Proporção de equivalências validadas
Índice de qualidade semântica	IQS	Grau de clareza, associação e validação semântica
Taxa de reutilização de ativos	TRA	Proporção de ativos reutilizados em análises ou pipelines
Tempo médio de descoberta	TMD	Tempo médio para localizar um ativo no catálogo
Utilidade percebida do catálogo	UPC	Avaliação dos usuários sobre utilidade do catálogo

10.1 Índice de Completude de Metadados — ICM

O **Índice de Completude de Metadados** avalia o grau de preenchimento dos campos obrigatórios de metadados de um ativo.

Para um ativo $\mathbf{a_i}$, o índice pode ser definido como:

$ICM(a_i) = CPO(a_i) / TCO$

em que:

- **$CPO(a_i)$** representa o número de campos obrigatórios preenchidos para o ativo $\mathbf{a_i}$;
- **TCO** representa o total de campos obrigatórios definidos para aquele tipo de ativo.

Como critério preliminar, um ativo pode ser considerado minimamente documentado quando apresentar:

$$\mathbf{ICM}(a_i) \geq 0{,}80$$

e possuir, obrigatoriamente, nome, descrição, origem, responsável, tipo, versão, sensibilidade, restrição de acesso e data de atualização.

10.2 Índice de Cobertura do Catálogo — ICC

O **Índice de Cobertura do Catálogo** mede a proporção de ativos registrados no catálogo em relação ao total de ativos identificados no modelo de referência.

ICC = número de ativos catalogados / número total de ativos identificados

Esse índice permite avaliar se o catálogo representa adequadamente o ecossistema analítico.

10.3 Índice de Qualidade Semântica — IQS

O **Índice de Qualidade Semântica** avalia o grau de clareza, padronização e validação das descrições semânticas.

Uma formulação simples pode considerar:

IQS(x) = itens semânticos atendidos / total de itens semânticos esperados

em que **x** pode representar um atributo, entidade, indicador, conceito ou produto de dados.

Os itens semânticos esperados podem incluir definição conceitual, domínio de valores, exemplo de uso, associação a termo do glossário, ausência de ambiguidade e validação por especialista.

10.4 Taxa de Ambiguidade Semântica — TAS

A **Taxa de Ambiguidade Semântica** mede a proporção de elementos catalogados que possuem significado incompleto, conflitante ou não validado.

TAS = número de elementos com ambiguidade semântica / número total de elementos avaliados

Valores elevados indicam necessidade de curadoria semântica.

10.5 Taxa de Ativos com Lineage Associado — TALA

A **Taxa de Ativos com Lineage Associado** mede a proporção de ativos catalogados que possuem origem, dependências ou transformações registradas.

TALA = número de ativos com lineage associado / número total de ativos catalogados

Essa métrica conecta o Módulo 9 ao módulo transversal de proveniência, lineage e rastreabilidade.

11. Artefatos formalizados do módulo

Os artefatos do Módulo 9 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, regras de construção, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A9.1	Modelo Transversal de Metadados	Estrutura formal	ICM	Padroniza a descrição dos ativos

A9.2	Catálogo Unificado de Dados	Catálogo consultável	ICC	Apoia descoberta, governança e reuso
A9.3	Dicionário de Dados Consolidado	Tabela de atributos	PAD, IQS	Reduz ambiguidades estruturais e semânticas
A9.4	Glossário de Termos do Domínio	Vocabulário controlado	CG	Padroniza conceitos críticos
A9.5	Modelo Semântico de Dados	Grafo conceitual	IQS, TAS	Representa conceitos e relações
A9.6	Mapa de Equivalência Semântica	Matriz ou tabela	TESV	Apoia harmonização entre fontes
A9.7	Ficha de Documentação de Ativo	Documento estruturado	ICM	Documenta ativos reutilizáveis
A9.8	Índice de Descoberta de Dados	Índice textual/semântico	TMD	Melhora localização e reuso
A9.9	Registro de Curadoria de Metadados	Tabela de revisão	AM, ICM	Controla atualização e revisão
A9.10	Relatório de Qualidade dos Metadados	Relatório técnico	ICM, IQS, TAS	Avalia a documentação produzida

12. Artefato A9.1 — Modelo Transversal de Metadados

12.1 Finalidade

O **Modelo Transversal de Metadados** define a estrutura mínima de informações que deve descrever os ativos produzidos ou utilizados no UnespDataLens-RM.

Esse modelo deve ser capaz de representar diferentes tipos de ativos, como fontes, datasets brutos, datasets integrados, datasets tratados, data marts, atributos, features, indicadores, dashboards, APIs, relatórios e produtos de dados.

12.2 Tipo de artefato na DSR

Esse artefato é um **modelo**, pois formaliza a estrutura de descrição dos ativos, e também um **constructo**, pois define conceitos centrais utilizados pelo modelo de referência.

12.3 Estrutura formal

O conjunto de ativos catalogáveis pode ser representado como:

$A = \{a_1, a_2, ..., a_n\}$

Cada ativo a_i pode ser descrito por:

a_i = {id, nome, tipo, descrição, origem, responsável, domínio, formato, granularidade, periodicidade, versão, sensibilidade, restrição, qualidade, lineage, consumidores, tags, limitações}

12.4 Campos obrigatórios

Campo	Descrição
Identificador	Código único do ativo
Nome	Nome oficial ou usual
Tipo	Fonte, dataset, atributo, feature, indicador, API, dashboard etc.
Descrição	Explicação sobre conteúdo ou finalidade
Origem	Fonte, módulo ou processo que gerou o ativo
Responsável	Pessoa, equipe, setor ou instituição responsável

Domínio	Área de aplicação
Formato	Estrutura técnica do ativo
Granularidade	Unidade de análise
Periodicidade	Frequência de atualização
Versão	Versão do ativo
Sensibilidade	Classificação de sensibilidade
Restrição de acesso	Política ou nível de acesso
Qualidade	Métricas ou status de qualidade
Lineage	Origem e dependências
Consumidores	Usuários, sistemas ou produtos que consomem o ativo
Tags	Palavras-chave
Limitações	Restrições ou cuidados de uso

12.5 Métricas associadas

- Índice de Completude de Metadados;
- percentual de ativos com responsável;
- percentual de ativos com sensibilidade;
- percentual de ativos com qualidade documentada;
- percentual de ativos com lineage;
- percentual de ativos com descrição semântica.

12.6 Critérios de validação

O modelo é considerado válido quando permite descrever diferentes tipos de ativos de forma padronizada e suficiente para apoiar descoberta, governança, qualidade, reuso, rastreabilidade e interpretação.

12.7 Contribuição

Esse artefato fornece a base estrutural para a documentação transversal dos ativos do UnespDataLens-RM.

13. Artefato A9.2 — Catálogo Unificado de Dados

13.1 Finalidade

O **Catálogo Unificado de Dados** organiza e disponibiliza os ativos do UnespDataLens-RM em uma estrutura consultável.

Ele permite que usuários localizem fontes, atributos, datasets, indicadores, features, modelos, dashboards, APIs, relatórios e produtos de dados, compreendendo sua origem, significado, qualidade, restrições, versões, responsáveis e condições de uso.

13.2 Tipo de artefato na DSR

O catálogo é uma **instanciação**, quando implementado em ferramenta computacional, planilha estruturada, banco de dados, API ou interface web. Também é um **modelo**, pois organiza os ativos e seus metadados de forma sistemática.

13.3 Estrutura mínima

Campo	Descrição
Identificador do ativo	Código único
Nome do ativo	Nome do dado, tabela, dataset, feature ou produto
Tipo de ativo	Fonte, atributo, dataset, indicador, modelo, dashboard etc.
Descrição	Explicação do ativo
Domínio	Área ou tema relacionado

Responsável	Responsável pelo ativo
Origem	Fonte, módulo ou pipeline de origem
Versão	Versão atual
Status	Ativo, depreciado, em validação, em revisão
Sensibilidade	Classificação de sensibilidade
Restrição de acesso	Política de acesso
Qualidade	Resultado ou status de qualidade
Tags	Palavras-chave
Lineage	Origem e dependências
Data de atualização	Última atualização
Limitações	Cuidados e restrições de uso

13.4 Regras de construção

Todo ativo relevante produzido ou consumido pelo modelo de referência deve ser registrado no catálogo. Ativos sem metadados mínimos devem receber status de incompleto ou em validação. Ativos depreciados devem permanecer registrados com indicação de obsolescência para preservar rastreabilidade histórica.

13.5 Métricas associadas

- Índice de Cobertura do Catálogo;
- Índice de Completude de Metadados;
- percentual de ativos com tags;
- percentual de ativos com lineage;

- percentual de ativos com qualidade documentada;
- percentual de ativos com restrição de acesso definida.

13.6 Critérios de validação

O catálogo é considerado operacional quando possui cobertura satisfatória dos ativos principais, metadados obrigatórios preenchidos, responsáveis definidos, status dos ativos e capacidade de consulta.

13.7 Contribuição

O catálogo transforma ativos dispersos em recursos localizáveis, compreensíveis, reutilizáveis e governáveis.

14. Artefato A9.3 — Dicionário de Dados Consolidado

14.1 Finalidade

O **Dicionário de Dados Consolidado** documenta os atributos utilizados nos ativos de dados do modelo de referência, incorporando descrições refinadas, domínios de valores, tipos, regras de preenchimento, sensibilidade, exemplos, fontes de origem e relações com conceitos do glossário.

14.2 Tipo de artefato na DSR

Esse artefato é um **modelo** e uma **instanciação**, pois formaliza a descrição dos atributos e pode ser implementado em tabela, banco, catálogo ou ferramenta específica.

14.3 Estrutura formal

Considerando um conjunto de atributos:

$$\mathbf{AT = \{at_1, at_2, ..., at_m\}}$$

cada atributo $\mathbf{at_j}$ deve ser descrito por:

$$\mathbf{at_j = \{id, nome_técnico, nome_semântico, descrição, tipo, domínio, formato, obrigatoriedade, regra, sensibilidade, fonte, exemplo, termo_associado, limitações\}}$$

14.4 Campos obrigatórios

Campo	Descrição
Identificador do atributo	Código único
Nome técnico	Nome físico no dataset

Nome semântico	Nome compreensível no domínio
Descrição	Significado do atributo
Tipo de dado	Texto, inteiro, decimal, data, booleano etc.
Domínio de valores	Valores permitidos ou categorias possíveis
Formato	Padrão esperado
Obrigatoriedade	Obrigatório ou opcional
Regra de preenchimento	Regra ou condição associada
Sensibilidade	Classificação do atributo
Fonte de origem	Fonte, tabela ou dataset de origem
Exemplo de valor	Exemplo ilustrativo
Termo do glossário	Conceito associado
Limitações	Cuidados de interpretação

14.5 Métricas associadas

- percentual de atributos documentados;
- percentual de atributos com domínio definido;
- percentual de atributos com termo associado ao glossário;
- percentual de atributos com sensibilidade classificada;

- taxa de atributos ambíguos;
- índice de qualidade semântica dos atributos.

14.6 Critérios de validação

Um atributo é considerado adequadamente documentado quando possui nome, descrição, tipo, domínio, fonte, sensibilidade e, quando aplicável, regra de preenchimento e termo associado ao glossário.

14.7 Contribuição

Esse artefato reduz ambiguidades, melhora a interoperabilidade semântica e apoia qualidade, transformação, integração, feature engineering e uso analítico dos dados.

15. Artefato A9.4 — Glossário de Termos do Domínio

15.1 Finalidade

O **Glossário de Termos do Domínio** define conceitos relevantes utilizados no projeto, estabelecendo vocabulário comum entre equipes técnicas, gestores, especialistas do domínio, pesquisadores e usuários dos dados.

15.2 Tipo de artefato na DSR

O glossário é um **constructo**, pois define conceitos fundamentais utilizados pelo modelo de referência e pelo domínio de aplicação.

15.3 Estrutura mínima

Campo	Descrição
Termo	Palavra ou expressão definida
Definição	Explicação conceitual
Domínio	Área à qual o termo pertence
Sinônimos	Termos equivalentes ou próximos
Termos relacionados	Conceitos associados

Exemplo de uso	Situação concreta de aplicação
Fonte da definição	Documento, especialista ou regra institucional
Status	Validado, em revisão ou preliminar
Observações	Cuidados ou ambiguidades

15.4 Métricas associadas

- cobertura do glossário;
- número de termos definidos;
- percentual de termos validados por especialista;
- número de termos ambíguos;
- número de sinônimos mapeados;
- taxa de conceitos sem definição.

15.5 Critérios de validação

Um termo é considerado validado quando possui definição clara, domínio associado, fonte da definição, exemplo de uso e, preferencialmente, validação por especialista.

15.6 Contribuição

O glossário reduz ambiguidades terminológicas e permite que atributos, indicadores e produtos de dados sejam interpretados de forma coerente.

16. Artefato A9.5 — Modelo Semântico de Dados

16.1 Finalidade

O **Modelo Semântico de Dados** representa conceitos, entidades, atributos, relações, regras de negócio e significados relevantes ao domínio do projeto.

Ele atua como uma ponte entre a estrutura técnica dos dados e o significado analítico dos conceitos representados.

16.2 Tipo de artefato na DSR

Esse artefato é um **modelo**, pois representa relações conceituais e semânticas entre elementos do domínio.

16.3 Estrutura formal

O modelo semântico pode ser representado como um grafo:

$G_S = (C, R)$

em que:

- **C** representa o conjunto de conceitos, entidades e atributos;
- **R** representa o conjunto de relações semânticas entre esses elementos.

Cada relação pode indicar equivalência, especialização, composição, dependência, associação, hierarquia ou temporalidade.

16.4 Exemplos de relações semânticas

Relação	Exemplo
Equivalência	“aluno” equivalente a “estudante” em determinado contexto
Especialização	“evasão definitiva” como tipo de “evasão”
Composição	“curso” composto por “disciplinas”
Associação	“estudante” associado a “matrícula”
Hierarquia	“campus” pertence a “instituição”
Temporalidade	“situação acadêmica” varia ao longo do tempo

16.5 Métricas associadas

- número de conceitos mapeados;
- número de relações semânticas;
- densidade do modelo semântico;
- percentual de atributos vinculados a conceitos;
- taxa de conceitos ambíguos;
- percentual de conceitos validados.

16.6 Critérios de validação

O modelo semântico é considerado válido quando representa conceitos críticos do domínio, explicita relações relevantes e reduz ambiguidades na interpretação dos dados.

16.7 Contribuição

Esse artefato apoia interoperabilidade semântica, integração, construção de indicadores, agentes inteligentes, explicabilidade e uso correto dos dados.

17. Artefato A9.6 — Mapa de Equivalência Semântica

17.1 Finalidade

O **Mapa de Equivalência Semântica** registra correspondências entre atributos, categorias, códigos ou conceitos equivalentes, parcialmente equivalentes ou conflitantes presentes em diferentes fontes.

17.2 Tipo de artefato na DSR

Esse artefato é um **modelo** e um **método**, pois representa equivalências e orienta harmonizações entre fontes.

17.3 Estrutura mínima

Campo	Descrição
Fonte A	Primeira fonte
Atributo ou conceito A	Elemento da primeira fonte
Fonte B	Segunda fonte
Atributo ou conceito B	Elemento da segunda fonte
Tipo de equivalência	Exata, parcial, aproximada, conflitante
Grau de confiança	Alto, médio ou baixo
Justificativa	Critério usado para indicar equivalência

Necessidade de transformação	Sim/não
Status	Validado, pendente, rejeitado
Observações	Limitações ou cuidados

17.4 Métricas associadas

- número de equivalências identificadas;
- percentual de equivalências validadas;
- taxa de equivalência exata;
- taxa de equivalência parcial;
- taxa de conflitos semânticos;
- grau médio de confiança das equivalências.

17.5 Critérios de validação

Uma equivalência é considerada validada quando possui justificativa documentada e, quando necessário, validação por especialista ou evidência técnica.

17.6 Contribuição

Esse artefato apoia integração, harmonização, transformação, interoperabilidade e redução de erros semânticos.

18. Artefato A9.7 — Ficha de Documentação de Ativo

18.1 Finalidade

A **Ficha de Documentação de Ativo** descreve, em formato sintético e padronizado, informações essenciais sobre um dataset, produto de dados, indicador, feature, dashboard, modelo analítico, API ou relatório.

18.2 Tipo de artefato na DSR

Esse artefato é uma **instanciação** e um **modelo documental**, pois padroniza a descrição de ativos reutilizáveis.

18.3 Estrutura mínima

Campo	Descrição

Nome do ativo	Nome do dataset, indicador, API, dashboard ou produto
Tipo	Dataset, feature, indicador, dashboard, modelo, API etc.
Finalidade	Uso previsto
Origem	Fontes e pipelines associados
Responsável	Responsável técnico ou institucional
Versão	Versão atual
Data de criação	Data de geração
Data de atualização	Última atualização
Granularidade	Unidade de análise
Cobertura temporal	Período abrangido
Variáveis principais	Atributos ou features centrais
Métricas de qualidade	Indicadores de qualidade associados
Sensibilidade	Classificação de sensibilidade
Restrições de uso	Condições de acesso ou uso

Limitações	Restrições interpretativas
Lineage	Origem e transformações principais
Consumidores	Usuários, sistemas ou produtos consumidores

18.4 Métricas associadas

- percentual de ativos com ficha;
- completude média das fichas;
- percentual de fichas atualizadas;
- percentual de fichas com lineage;
- percentual de fichas com qualidade documentada.

18.5 Critérios de validação

A ficha é considerada adequada quando permite que um usuário compreenda finalidade, origem, qualidade, restrições, limitações e condições de uso do ativo sem depender exclusivamente de conhecimento tácito.

18.6 Contribuição

Esse artefato apoia reutilização, transparência, documentação, governança, comunicação e transferência de conhecimento.

19. Artefato A9.8 — Índice de Descoberta de Dados

19.1 Finalidade

O **Índice de Descoberta de Dados** organiza mecanismos para localizar ativos no catálogo com base em palavras-chave, tags, conceitos, domínios, responsáveis, entidades, atributos, indicadores, restrições ou relações semânticas.

19.2 Tipo de artefato na DSR

Esse artefato é uma **instanciação** e um **método**, pois operacionaliza a descoberta de dados.

19.3 Componentes

- índice textual;
- tags;
- domínios;
- termos do glossário;
- entidades associadas;
- atributos relacionados;

- responsáveis;
- classificação de sensibilidade;
- status do ativo;
- data de atualização;
- relações com outros ativos;
- restrições de acesso;
- lineage associado;
- produtos consumidores.

19.4 Métricas associadas

- tempo médio para localizar um ativo;
- taxa de ativos encontrados por busca;
- precisão da busca;
- taxa de reutilização de ativos;
- número de consultas realizadas;
- satisfação dos usuários.

19.5 Contribuição

Esse artefato reduz sobrecarga informacional, melhora a descoberta de dados e apoia agentes inteligentes de consulta.

20. Artefato A9.9 — Registro de Curadoria de Metadados

20.1 Finalidade

O **Registro de Curadoria de Metadados** documenta revisões, atualizações, correções e validações realizadas sobre os metadados dos ativos.

20.2 Estrutura mínima

Campo	Descrição
Ativo	Ativo cujos metadados foram revisados
Campo revisado	Metadado alterado
Valor anterior	Conteúdo antes da revisão
Valor atualizado	Conteúdo após a revisão

Motivo da alteração	Justificativa
Responsável	Pessoa ou equipe responsável
Data da revisão	Momento da alteração
Status	Validado, pendente, em revisão
Evidência	Documento, fonte ou justificativa

20.3 Métricas associadas

- número de metadados revisados;
- percentual de ativos revisados;
- tempo médio desde a última atualização;
- percentual de revisões validadas;
- taxa de metadados desatualizados.

20.4 Contribuição

Esse artefato garante que o catálogo não se torne obsoleto, apoiando melhoria contínua, governança e confiabilidade da documentação.

21. Artefato A9.10 — Relatório de Qualidade dos Metadados

21.1 Finalidade

O **Relatório de Qualidade dos Metadados** avalia a condição documental e semântica dos metadados produzidos e mantidos pelo módulo.

21.2 Estrutura mínima

Campo	Descrição
Escopo avaliado	Fontes, atributos, datasets, produtos ou catálogo completo

Cobertura do catálogo	Proporção de ativos catalogados
Completude dos metadados	Grau de preenchimento
Atualidade	Tempo desde última atualização
Ambiguidades	Elementos com significado incerto
Termos sem definição	Conceitos ausentes no glossário
Ativos sem responsável	Itens sem dono definido
Ativos sem qualidade	Itens sem métricas associadas
Ativos sem lineage	Itens sem origem/dependência
Recomendações	Ações de melhoria

21.3 Métricas associadas

- ICM;
- ICC;
- IQS;
- TAS;
- TALA;
- cobertura do glossário;
- atualidade dos metadados.

21.4 Contribuição

Esse artefato permite avaliar a própria qualidade da documentação, evitando que o catálogo seja apenas uma coleção incompleta de descrições.

22. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Inventários, datasets, manifestos, logs, regras, relatórios de qualidade, registros de vieses, estruturas de armazenamento, produtos de dados, dashboards, APIs e documentação institucional.
Processamento	Captura, consolidação, padronização, catalogação, enriquecimento semântico, construção de glossário, mapeamento de equivalências, curadoria e disponibilização dos metadados.
Artefatos	Modelo transversal de metadados, catálogo unificado, dicionário consolidado, glossário, modelo semântico, equivalências semânticas, fichas de ativos, índice de descoberta, curadoria e relatório de qualidade.
Métricas	ICM, ICC, PAD, PADS, PASD, PAQD, PALA, AM, CT, CG, TAS, TESV, IQS, TRA, TMD e UPC.
Saídas	Ativos documentados, catalogados, semanticamente descritos, localizáveis, governáveis, reutilizáveis e preparados para lineage, versionamento, monitoramento, auditoria e agentes inteligentes.

23. Critérios gerais de validação do módulo

O Módulo 9 pode ser considerado adequadamente implementado quando:

1. os principais ativos produzidos pelo pipeline estão registrados no catálogo;
2. os metadados obrigatórios dos ativos críticos estão preenchidos;
3. fontes, datasets, atributos, indicadores, APIs, dashboards e produtos de dados possuem descrição mínima;
4. responsáveis, versões, restrições de acesso e classificações de sensibilidade estão documentados;
5. os principais termos do domínio estão definidos no glossário;
6. os atributos críticos possuem significado, domínio e regra de uso documentados;
7. equivalências semânticas relevantes entre fontes foram registradas;
8. o catálogo pode ser consultado por usuários humanos ou sistemas;
9. os ativos catalogados possuem relação com origem, versão, qualidade e lineage, quando aplicável;
10. os metadados foram avaliados quanto à completude, atualidade e qualidade semântica;
11. ambiguidades críticas foram identificadas e encaminhadas para curadoria;
12. os metadados apoiam efetivamente governança, rastreabilidade, versionamento, monitoramento e agentes inteligentes.

24. Relação com o pipeline principal

O Módulo 9 se relaciona com todos os módulos do pipeline principal do UnespDataLens-RM.

Com o **Módulo 1 — Fontes e Inventário de Dados**, recebe informações sobre fontes, responsáveis, formatos, periodicidade, granularidade, cobertura temporal, sensibilidade e restrições.

Com o **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, registra metadados sobre origem, data de ingestão, logs, snapshots, formatos e parâmetros de execução.

Com o **Módulo 3 — Integração e Harmonização**, incorpora regras de integração, chaves, entidades compartilhadas, equivalências semânticas, conflitos e datasets integrados.

Com o **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, documenta regras de transformação, atributos derivados, fórmulas, parâmetros, impactos e datasets tratados.

Com o **Módulo 5 — Qualidade e Validação de Dados**, incorpora métricas de qualidade, status de aprovação, não conformidades e histórico de validação.

Com o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, registra grupos, atributos sensíveis, métricas de fairness, riscos residuais e limitações de uso.

Com o **Módulo 7 — Armazenamento Analítico**, registra localização, estrutura, formato, schema, zona, versão e estado dos ativos armazenados.

Com o **Módulo 8 — Disponibilização e Consumo Analítico**, registra produtos de dados, dashboards, APIs, consumidores, versões publicadas, formas de acesso e feedback de uso.

25. Relação com os demais módulos transversais

O Módulo 9 fornece base documental para os demais módulos transversais do UnespDataLens-RM.

Para o módulo de **Governança, Segurança, Privacidade e Conformidade**, fornece informações sobre sensibilidade, restrições de acesso, responsáveis, finalidade de uso, qualidade e limitações dos ativos.

Para o módulo de **Proveniência, Lineage e Rastreabilidade**, fornece identificadores, descrições, versões, relações entre ativos, fontes de origem e dependências.

Para o módulo de **Versionamento e Reprodutibilidade**, fornece informações sobre versões, parâmetros, schemas, datasets, produtos e artefatos associados.

Para o módulo de **Representações Analíticas: Tabular, Temporal, Espacial, Textual, Semântica e Grafos**, fornece entidades, atributos, conceitos, relações e definições semânticas necessárias à modelagem.

Para o módulo de **Feature Engineering e Feature Store**, fornece dicionários, definições, regras de cálculo, granularidade, origem e contexto das features.

Para o módulo de **Monitoramento, Drift e Evolução dos Ativos**, fornece metadados de atualização, schemas, status, qualidade, versões e histórico dos ativos.

Para o módulo de **Agentes Inteligentes de Apoio ao Pipeline**, fornece a base documental e semântica para busca, explicação, recomendação, validação, navegação e geração de respostas contextualizadas.

26. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 9 pode organizar metadados sobre fontes como sistema acadêmico, ambiente virtual de aprendizagem, registros de matrícula, histórico escolar, notas, frequência, dados socioeconômicos, questionários, registros de evasão e atividades em plataformas digitais.

O catálogo pode registrar ativos como:

- base de matrículas por período letivo;
- histórico acadêmico;
- interações no AVA;
- indicadores de desempenho;
- features de permanência;
- base analítica de evasão;
- dashboard de evasão;
- API de indicadores educacionais;
- data mart de trajetória acadêmica.

O dicionário consolidado pode documentar atributos como RA, curso, disciplina, turma, período letivo, nota final, frequência, situação acadêmica, forma de ingresso, escola de origem, renda familiar e participação em atividades.

O glossário pode definir conceitos como estudante ativo, matrícula regular, trancamento, cancelamento, evasão, retenção, aprovação, reprovação, turma, curso, disciplina e período letivo.

O modelo semântico pode representar relações como:

- estudante realiza matrícula;
- estudante cursa disciplina;
- disciplina pertence a curso;
- estudante participa de turma;
- atividade pertence a ambiente virtual;
- evasão está associada à situação acadêmica;
- desempenho pode ser medido por nota, frequência e aprovação.

Essa estrutura reduz ambiguidades e apoia a construção da **UnespDataLens-RM** como especialização educacional do UnespDataLens-RM.

27. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, o Módulo 9 pode organizar metadados sobre fontes de saúde, educação, assistência social, território, habitação, trabalho, orçamento e atendimento ao cidadão.

O catálogo pode registrar ativos como:

- bases de famílias atendidas;
- indicadores territoriais;
- dados de equipamentos públicos;
- registros de atendimento;
- mapas de vulnerabilidade;
- painéis de gestão;
- datasets integrados por território;
- APIs de indicadores públicos;
- dashboards de cobertura de serviços.

O glossário pode definir conceitos como família, território, vulnerabilidade, atendimento, benefício, equipamento público, unidade de saúde, escola, ocorrência, demanda reprimida e indicador social.

O modelo semântico pode representar relações como:

- cidadão pertence a família;
- família reside em território;
- território possui equipamentos públicos;
- atendimento ocorre em unidade;
- política pública atende população-alvo;
- indicador social descreve território.

Esse modelo favorece a integração semântica entre secretarias, reduz duplicidades conceituais e apoia decisões baseadas em evidências.

28. Contribuição técnico-científica do módulo

A contribuição do Módulo 9 está em formalizar metadados, catálogo e semântica como componentes transversais da Engenharia Analítica de Dados.

Em vez de tratar documentação como atividade secundária ou posterior ao processamento, o UnespDataLens-RM incorpora os metadados como artefatos centrais para descoberta, governança, qualidade, interoperabilidade, rastreabilidade, reprodutibilidade, explicabilidade, reutilização e apoio a agentes inteligentes.

Essa formalização contribui para:

- transformar conhecimento tácito sobre dados em conhecimento explícito;
- reduzir ambiguidades semânticas;
- padronizar conceitos e descrições;
- apoiar descoberta e reutilização de ativos;
- documentar qualidade, sensibilidade e restrições de uso;
- conectar dados a lineage, versionamento e governança;
- viabilizar consulta por usuários humanos e agentes inteligentes;
- fortalecer transparência, auditoria e reprodutibilidade;
- melhorar comunicação entre equipes técnicas, gestores e especialistas de domínio;
- sustentar a evolução contínua do ecossistema analítico.

29. Síntese do módulo

O **Módulo Transversal 9 — Metadados, Catálogo e Semântica dos Dados** organiza o conhecimento sobre os dados no UnespDataLens-RM. Ele acompanha todo o pipeline principal, consolidando informações produzidas desde o inventário das fontes até a disponibilização dos produtos analíticos.

Sua importância é transversal porque permite que os dados sejam compreendidos, localizados, interpretados, governados, rastreados, reutilizados e explicados. Ao documentar não apenas a estrutura técnica dos dados, mas também seu significado, origem, restrições, qualidade, versões, relações, consumidores e limitações, o módulo fortalece a confiabilidade, a transparência, a governança e a reprodutibilidade dos processos analíticos.

Dessa forma, o Módulo 9 constitui uma contribuição central do UnespDataLens-RM, ao transformar metadados, catálogo e semântica em artefatos formais, avaliáveis e integrados ao ciclo completo de Engenharia Analítica de Dados.

Módulo Transversal 10 — Governança, Segurança, Privacidade e Conformidade

1. Finalidade do módulo

O **Módulo Transversal 10 — Governança, Segurança, Privacidade e Conformidade** tem como finalidade estabelecer princípios, políticas, regras, mecanismos, controles e artefatos para assegurar que os dados e ativos analíticos produzidos no modelo de referência **UnespDataLens-RM** sejam tratados de forma segura, responsável, controlada, auditável, ética e aderente às normas legais, institucionais e técnicas aplicáveis.

Diferentemente dos módulos 1 a 8, que compõem o pipeline técnico-operacional principal, este módulo atua de forma **transversal** sobre todo o ciclo de vida dos dados. Isso significa que governança, segurança, privacidade e conformidade não são tratadas apenas no final do processo, mas desde a identificação das fontes, passando pela ingestão, integração, transformação, validação, armazenamento, disponibilização, consumo, monitoramento e reutilização dos ativos analíticos.

No contexto do UnespDataLens-RM, os dados podem ter diferentes níveis de sensibilidade, finalidade, restrição de acesso e risco. Podem envolver dados públicos, dados institucionais, dados educacionais, dados administrativos, dados pessoais, dados sensíveis, dados anonimizados, dados agregados, dados territoriais, dados socioeconômicos, registros de interação, indicadores, modelos, dashboards, APIs e produtos de dados. Cada um desses ativos exige tratamento compatível com sua natureza, finalidade e risco.

Esse módulo é essencial porque pipelines analíticos não devem ser avaliados apenas pela capacidade técnica de integrar, transformar e disponibilizar dados. Eles também devem ser avaliados pela capacidade de garantir uso legítimo, responsável, seguro e controlado dos dados. Um pipeline tecnicamente eficiente pode ser inadequado se expuser dados sensíveis, permitir acessos indevidos, violar restrições institucionais, dificultar auditoria, não documentar finalidade de uso ou não respeitar princípios de minimização e necessidade.

Assim, o Módulo 10 formaliza a governança, a segurança, a privacidade e a conformidade como componentes estruturais da Engenharia Analítica de Dados. Ele define responsabilidades, papéis, políticas, permissões, classificações, controles, registros, evidências e critérios de validação que devem acompanhar os ativos ao longo de todo o modelo de referência.

2. Objetivo geral

O objetivo geral do Módulo 10 é **garantir que os dados e ativos analíticos do UnespDataLens-RM sejam gerenciados, protegidos, acessados, compartilhados, utilizados e descartados de forma governada, segura, responsável e conforme normas legais, institucionais e éticas**, por meio de políticas, controles, registros, matrizes de responsabilidade, classificação de sensibilidade, mecanismos de privacidade e critérios de conformidade.

3. Objetivos específicos

São objetivos específicos do Módulo 10:

1. definir políticas de governança de dados aplicáveis ao ciclo completo do UnespDataLens-RM;
2. estabelecer papéis, responsabilidades e responsáveis pelos dados, pipelines, produtos analíticos e decisões de acesso;
3. classificar fontes, atributos, datasets, produtos de dados e ativos analíticos quanto à sensibilidade, restrição e risco;
4. definir regras de acesso, uso, compartilhamento, publicação, retenção, anonimização e descarte de dados;
5. assegurar que dados pessoais e sensíveis sejam tratados de acordo com princípios de finalidade, necessidade, adequação, segurança, transparência e minimização;
6. estabelecer mecanismos de segurança, como controle de acesso, autenticação, autorização, criptografia, mascaramento, segregação de ambientes e auditoria;
7. registrar evidências de autorização, consentimento, base legal, finalidade de uso ou justificativa institucional, quando aplicável;
8. documentar riscos de privacidade, reidentificação, uso indevido, acesso indevido, exposição, vazamento ou compartilhamento inadequado;
9. definir critérios para disponibilização segura de produtos de dados, dashboards, APIs, datasets e relatórios;
10. fornecer insumos para os módulos transversais de metadados, lineage, versionamento, monitoramento e agentes inteligentes;
11. produzir artefatos formais que permitam auditar a conformidade do modelo de referência com políticas internas, normas externas e boas práticas de governança de dados.

4. Fundamentação conceitual do módulo

Governança de dados pode ser compreendida como o conjunto de princípios, políticas, papéis, processos e mecanismos utilizados para orientar a gestão dos dados ao longo de seu ciclo de vida. Ela define quem pode acessar, alterar, compartilhar, publicar, reutilizar, aprovar ou excluir dados, bem como sob quais condições essas ações podem ocorrer.

Segurança de dados refere-se aos mecanismos técnicos e organizacionais destinados a proteger dados contra acesso não autorizado, alteração indevida, perda, vazamento, exposição, indisponibilidade ou uso inadequado.

Privacidade refere-se à proteção dos direitos dos titulares dos dados e ao tratamento responsável de informações pessoais ou sensíveis, considerando princípios como finalidade, necessidade, minimização, transparência, segurança e limitação de acesso.

Conformidade refere-se à aderência a normas legais, políticas institucionais, contratos, regulamentos, diretrizes éticas, padrões técnicos e requisitos de auditoria aplicáveis ao uso dos dados.

No UnespDataLens-RM, essas quatro dimensões são tratadas de forma integrada:

Dimensão	Questão central	Exemplos
Governança	Quem é responsável pelos dados e quais regras orientam seu uso?	papéis, responsabilidades, políticas, aprovação de acesso

Segurança	Como os dados são protegidos contra acesso ou alteração indevida?	autenticação, autorização, criptografia, logs, segregação
Privacidade	Como dados pessoais e sensíveis são protegidos?	anonimização, minimização, pseudonimização, mascaramento
Conformidade	O uso dos dados está aderente a normas e políticas?	LGPD, regras institucionais, termos de uso, contratos, auditoria

Esse módulo parte do princípio de que governança e privacidade não são obstáculos ao uso de dados, mas condições para que o uso analítico seja confiável, legítimo e sustentável. A ausência de governança aumenta riscos de exposição, uso indevido, baixa qualidade, interpretações inadequadas, decisões não auditáveis e perda de confiança institucional.

À luz da **Design Science Research**, os artefatos do Módulo 10 são concebidos como modelos, métodos e instanciações que formalizam regras de governança, segurança, privacidade e conformidade aplicáveis ao modelo de referência. Esses artefatos tornam o tratamento dos dados auditável, justificável e avaliável.

5. Natureza transversal do módulo

O Módulo 10 é transversal porque suas regras, políticas e controles acompanham todos os módulos do pipeline principal e interagem com os demais módulos transversais.

No **Módulo 1 — Fontes e Inventário de Dados**, atua na classificação inicial de sensibilidade, identificação de responsáveis, restrições de acesso e finalidade original das fontes.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, atua no controle de permissões de extração, segurança das credenciais, registro de autorização, armazenamento seguro e proteção dos dados brutos.

No **Módulo 3 — Integração e Harmonização**, atua na avaliação de riscos gerados pela combinação de fontes, especialmente riscos de reidentificação, aumento de sensibilidade e uso indevido de dados integrados.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, atua na definição de regras para anonimização, pseudonimização, mascaramento, minimização, remoção de atributos sensíveis e controle de atributos derivados.

No **Módulo 5 — Qualidade e Validação de Dados**, atua na validação de conformidade dos dados com regras de acesso, formatos esperados, contratos, restrições institucionais e critérios mínimos de confiabilidade.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, atua na avaliação ética e responsável do uso de atributos sensíveis, proxies, grupos protegidos e métricas de equidade.

No **Módulo 7 — Armazenamento Analítico**, atua na definição de políticas de acesso, segregação de zonas, criptografia, retenção, controle de permissões e proteção de ativos sensíveis.

No **Módulo 8 — Disponibilização e Consumo Analítico**, atua na autorização de publicação, definição de granularidade, anonimização, agregação, restrição de acesso, termos de uso e auditoria de consumo.

Com o **Módulo 9 — Metadados, Catálogo e Semântica**, compartilha informações sobre sensibilidade, restrições, responsáveis, finalidade, qualidade, base legal, limitações e condições de uso dos ativos.

Dessa forma, o Módulo 10 opera como uma camada normativa e operacional que garante que o ciclo de dados do UnespDataLens-RM seja tecnicamente robusto e institucionalmente responsável.

6. Formalização sintética do módulo

O Módulo 10 pode ser compreendido como um processo transversal de associação de políticas, responsabilidades, permissões, classificações e controles aos ativos de dados do modelo de referência.

Considerando o conjunto de ativos de dados do UnespDataLens-RM:

$$\mathbf{A = \{a_1, a_2, ..., a_n\}}$$

em que cada $\mathbf{a_i}$ representa uma fonte, atributo, dataset, modelo, dashboard, API, relatório ou produto de dados, o módulo associa a cada ativo um conjunto de elementos de governança:

$$\mathbf{G(a_i) = \{papeis, responsáveis, permissões, sensibilidade, finalidade, restrições, riscos, controles, evidências\}}$$

Cada ativo pode ser descrito, do ponto de vista da governança, por:

$$\mathbf{a_i = \{id, nome, tipo, responsável, custodiante, proprietário, sensibilidade, finalidade, base_legal, restrição_acesso, risco_privacidade, controle_segurança, política_retenção, status_conformidade\}}$$

O módulo também define um conjunto de políticas:

$$\mathbf{P = \{p_1, p_2, ..., p_k\}}$$

em que cada política $\mathbf{p_j}$ pode ser descrita por:

$$\mathbf{p_j = \{id, nome, escopo, regra, ativos_afetados, responsável, critério_conformidade, evidência_requerida\}}$$

Além disso, define um conjunto de permissões:

$$\mathbf{R = \{r_1, r_2, ..., r_m\}}$$

em que cada permissão $\mathbf{r_l}$ representa uma relação entre consumidor, ativo e tipo de acesso:

r_l = {usuário_ou_sistema, ativo, tipo_acesso, finalidade, validade, autorização, restrição}

Dessa forma, o uso dos dados deixa de ser uma ação informal e passa a ser controlado por regras, papéis, permissões, evidências e critérios de conformidade.

7. Entradas do módulo

As entradas do Módulo 10 são produzidas ao longo do pipeline principal e pelos demais módulos transversais.

As principais entradas são:

- inventário de fontes;
- classificação de sensibilidade;
- registro de responsáveis;
- restrições de acesso;
- finalidade original das fontes;
- registros de permissões de ingestão;
- logs de execução;
- datasets brutos;
- datasets integrados;
- datasets tratados;
- datasets validados;
- datasets balanceados ou sinalizados;
- data marts;
- dashboards;
- APIs;
- produtos de dados;
- relatórios de qualidade;
- relatórios de vieses e equidade;
- metadados dos ativos;
- catálogo de dados;
- dicionários de dados;
- registros de lineage;
- versões e snapshots;
- registros de consumo;
- requisitos institucionais;
- políticas internas;
- termos de uso;
- contratos;
- normativas legais;
- regras de privacidade;
- decisões de comitês ou responsáveis;
- evidências de autorização.

8. Processamento

O processamento do Módulo 10 envolve atividades contínuas de classificação, definição de políticas, controle de acesso, gestão de riscos, aplicação de mecanismos de segurança, documentação de conformidade e auditoria.

8.1 Definição de políticas de governança

A primeira atividade consiste em estabelecer políticas que orientem o tratamento dos dados no UnespDataLens-RM. Essas políticas devem definir regras para acesso, uso, compartilhamento, publicação, retenção, descarte, anonimização, versionamento, responsabilização e auditoria.

As políticas devem ser explícitas, documentadas e associadas aos tipos de ativos, domínios de aplicação, níveis de sensibilidade e perfis de usuários.

8.2 Definição de papéis e responsabilidades

A segunda atividade consiste em definir papéis institucionais e técnicos relacionados aos dados. Entre os papéis possíveis estão proprietário do dado, custodiante, gestor, responsável técnico, analista, consumidor, aprovador de acesso, encarregado de privacidade e auditor.

A definição de papéis reduz ambiguidades sobre quem pode autorizar, alterar, publicar, corrigir ou restringir o uso de um ativo.

8.3 Classificação de sensibilidade e risco

A terceira atividade consiste em classificar os ativos quanto à sensibilidade, restrição de acesso e risco. Essa classificação deve considerar presença de dados pessoais, dados sensíveis, dados anonimizados, dados agregados, risco de reidentificação, impacto potencial de exposição e finalidade de uso.

Essa classificação deve ser atualizada quando os dados forem integrados, enriquecidos ou disponibilizados, pois a combinação de fontes pode aumentar o risco de identificação ou sensibilidade.

8.4 Definição de finalidade e adequação de uso

A quarta atividade consiste em registrar a finalidade de uso dos dados e avaliar se o tratamento realizado é compatível com essa finalidade. Dados coletados para uma finalidade específica não devem ser reutilizados de forma incompatível sem justificativa, autorização ou avaliação institucional.

8.5 Controle de acesso

A quinta atividade consiste em definir e aplicar regras de acesso aos ativos. O controle pode ocorrer por papéis, perfis, grupos, sistemas, tokens, credenciais, ambientes ou níveis de granularidade.

O acesso pode ser classificado como público, interno, restrito, sensível, anonimizado, agregado, sob autorização ou bloqueado.

8.6 Aplicação de mecanismos de privacidade

A sexta atividade consiste em aplicar medidas para reduzir riscos associados a dados pessoais ou sensíveis. Essas medidas podem incluir minimização, anonimização, pseudonimização, mascaramento, supressão, generalização, agregação, perturbação, criptografia ou limitação de granularidade.

A escolha da técnica deve considerar finalidade, risco, utilidade analítica e possibilidade de reidentificação.

8.7 Gestão de riscos de dados

A sétima atividade consiste em identificar, classificar e acompanhar riscos relacionados ao uso dos dados. Os riscos podem envolver vazamento, acesso indevido, uso incompatível, baixa qualidade, reidentificação, viés, decisão injusta, exposição de grupos, dependência excessiva de fonte externa ou uso de versão obsoleta.

Cada risco deve possuir severidade, probabilidade, impacto, ação mitigadora, responsável e status.

8.8 Registro de conformidade

A oitava atividade consiste em registrar evidências de conformidade. Essas evidências podem incluir autorização de acesso, justificativa de finalidade, classificação de sensibilidade, aprovação de publicação, termos de uso, logs de acesso, relatórios de anonimização, registros de consentimento ou base legal, quando aplicável.

8.9 Auditoria e rastreabilidade de acesso

A nona atividade consiste em registrar e auditar acessos, alterações, publicações e compartilhamentos. Logs de acesso, trilhas de auditoria e registros de consumo permitem identificar quem acessou o quê, quando, para qual finalidade e sob qual autorização.

8.10 Revisão periódica de políticas e permissões

A décima atividade consiste em revisar periodicamente políticas, permissões, classificações e riscos. A governança deve acompanhar mudanças nos dados, nos produtos, nas normas, nos usuários e nas finalidades de uso.

9. Técnicas, métodos e tecnologias aplicáveis

O Módulo 10 pode empregar diferentes técnicas, métodos e tecnologias para governança, segurança, privacidade e conformidade.

Entre as técnicas de governança, destacam-se:

- definição de políticas de dados;
- matriz RACI;
- data ownership;
- data stewardship;
- gestão de domínios;
- gestão de termos de uso;
- classificação de ativos;
- data contracts;
- workflow de aprovação;
- gestão de riscos;
- trilhas de auditoria;
- registro de finalidade;
- data lifecycle management.

Entre as técnicas de segurança, podem ser utilizadas:

- autenticação;

- autorização;
- controle de acesso baseado em papéis;
- controle de acesso baseado em atributos;
- criptografia em repouso;
- criptografia em trânsito;
- mascaramento de dados;
- segregação de ambientes;
- logs de acesso;
- auditoria de permissões;
- gestão de credenciais;
- rotação de chaves;
- controle de exportação;
- monitoramento de acessos.

Entre as técnicas de privacidade, destacam-se:

- minimização de dados;
- anonimização;
- pseudonimização;
- supressão;
- generalização;
- agregação;
- mascaramento;
- tokenização;
- hashing;
- k-anonymity;
- l-diversity;
- t-closeness;
- differential privacy, quando aplicável;
- avaliação de risco de reidentificação;
- controle de granularidade;
- separação de identificadores.

Entre as técnicas de conformidade, podem ser aplicadas:

- registro de base legal;
- registro de finalidade;
- registro de consentimento, quando aplicável;
- termos de uso;
- políticas de retenção;
- políticas de descarte;
- registros de autorização;
- relatórios de conformidade;
- auditoria interna;
- avaliação de impacto;
- revisão por comitê ou responsável institucional.

Ferramentas possíveis incluem catálogos de dados, sistemas de IAM, bancos de dados com controle de acesso, ferramentas de criptografia, logs centralizados, ferramentas de auditoria, Apache Ranger, Apache Atlas, DataHub, OpenMetadata, sistemas de gestão documental, repositórios Git, plataformas de data governance e ferramentas de observabilidade.

10. Métricas do módulo

As métricas do Módulo 10 têm como objetivo avaliar a maturidade, cobertura, efetividade e conformidade dos mecanismos de governança, segurança, privacidade e controle.

Métrica	Sigla	Descrição
Taxa de ativos com responsável definido	TARD	Proporção de ativos com proprietário ou responsável
Taxa de ativos com sensibilidade classificada	TASC	Proporção de ativos com classificação de sensibilidade
Taxa de ativos com finalidade documentada	TAFD	Proporção de ativos com finalidade registrada
Taxa de ativos com política de acesso	TAPA	Proporção de ativos com regra de acesso definida
Taxa de ativos sensíveis com controle explícito	TASCE	Proporção de ativos sensíveis com controle de acesso documentado
Taxa de acessos autorizados	TAA	Proporção de acessos compatíveis com regras definidas
Taxa de acessos não conformes	TANC	Proporção de acessos indevidos ou sem justificativa
Taxa de ativos com anonimização aplicada	TAAA	Proporção de ativos que exigem e receberam anonimização
Taxa de ativos com risco registrado	TARR	Proporção de ativos com avaliação de risco
Taxa de riscos mitigados	TRM	Proporção de riscos tratados ou reduzidos

Taxa de conformidade documental	TCD	Proporção de ativos com evidências mínimas de conformidade
Taxa de permissões vencidas	TPV	Proporção de permissões expiradas ainda ativas
Taxa de auditoria de acesso	TAuA	Proporção de acessos registrados em log
Índice de maturidade de governança	IMG	Grau de maturidade das práticas de governança
Taxa de produtos publicados com aprovação	TPPA	Proporção de produtos disponibilizados com aprovação formal

10.1 Taxa de Ativos com Sensibilidade Classificada — TASC

A **Taxa de Ativos com Sensibilidade Classificada** mede a proporção de ativos que possuem classificação formal de sensibilidade.

TASC = número de ativos com sensibilidade classificada / número total de ativos avaliados

Essa métrica indica se os ativos foram avaliados quanto ao risco de exposição e uso indevido.

10.2 Taxa de Ativos com Política de Acesso — TAPA

A **Taxa de Ativos com Política de Acesso** mede a proporção de ativos que possuem regra explícita de acesso.

TAPA = número de ativos com política de acesso definida / número total de ativos catalogados

Essa métrica é essencial para segurança e governança.

10.3 Taxa de Conformidade Documental — TCD

A **Taxa de Conformidade Documental** mede a proporção de ativos que possuem evidências mínimas de conformidade, como responsável, finalidade, sensibilidade, política de acesso e justificativa de uso.

TCD = número de ativos com documentação de conformidade / número total de ativos avaliados

10.4 Taxa de Riscos Mitigados — TRM

A **Taxa de Riscos Mitigados** mede a proporção de riscos registrados que receberam ação de mitigação.

TRM = número de riscos mitigados / número total de riscos identificados

10.5 Índice de Maturidade de Governança — IMG

O **Índice de Maturidade de Governança** pode ser estruturado em níveis:

Nível	Classificação	Caracterização
1	Inicial	Práticas informais, baixa documentação e controles pontuais
2	Documentado	Políticas e responsabilidades básicas registradas
3	Padronizado	Regras de acesso, sensibilidade e uso aplicadas de forma sistemática
4	Gerenciado	Controles, auditoria, riscos e conformidade monitorados
5	Otimizado	Governança automatizada, integrada e continuamente aprimorada

Esse índice permite avaliar a evolução institucional da governança de dados no contexto do modelo de referência.

11. Artefatos formalizados do módulo

Os artefatos do Módulo 10 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, regras de construção, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A10.1	Política de	Documento	IMG	Define princípios,

	Governança de Dados	normativo		regras e responsabilidades
A10.2	Matriz de Papéis e Responsabilidades	Matriz RACI	TARD	Define responsáveis por dados e decisões
A10.3	Classificação de Sensibilidade e Restrição	Matriz/tabela	TASC	Classifica ativos quanto ao risco
A10.4	Matriz de Acesso e Permissões	Matriz de controle	TAPA, TAA	Controla acesso aos ativos
A10.5	Registro de Finalidade e Base Legal	Tabela/ documento	TAFD, TCD	Documenta finalidade e justificativa de uso
A10.6	Plano de Privacidade e Proteção de Dados	Documento técnico	TAAA	Define técnicas de proteção
A10.7	Registro de Riscos de Dados	Matriz de risco	TARR, TRM	Identifica e acompanha riscos
A10.8	Registro de Auditoria e Acessos	Log estruturado	TAuA	Registra acessos e ações
A10.9	Relatório de Conformidade	Relatório técnico	TCD	Avalia aderência às regras
A10.10	Plano de Retenção e Descarte	Documento/ tabela	Conformidade de retenção	Define ciclo de vida dos dados

12. Artefato A10.1 — Política de Governança de Dados

12.1 Finalidade

A **Política de Governança de Dados** define os princípios, regras e diretrizes gerais para gestão dos dados no UnespDataLens-RM.

12.2 Tipo de artefato na DSR

Na perspectiva da Design Science Research, esse artefato é um **modelo normativo** e um **método**, pois orienta ações, responsabilidades e decisões ao longo do modelo de referência.

12.3 Estrutura mínima

Campo	Descrição
Escopo	Dados, ativos e processos cobertos pela política
Princípios	Segurança, finalidade, qualidade, transparência, minimização etc.
Papéis	Responsáveis e funções envolvidas
Regras de acesso	Condições para acesso aos ativos
Regras de uso	Finalidades permitidas e restrições
Regras de compartilhamento	Condições para disponibilização
Regras de retenção	Tempo de preservação
Regras de descarte	Condições para eliminação
Regras de auditoria	Evidências e logs exigidos
Critérios de conformidade	Condições de aderência

12.4 Métricas associadas

- índice de maturidade de governança;
- taxa de ativos com política associada;
- taxa de conformidade documental;
- taxa de ativos com responsável definido.

12.5 Critérios de validação

A política é considerada válida quando define escopo, princípios, papéis, regras de acesso, regras de uso, compartilhamento, retenção, descarte e critérios de auditoria.

12.6 Contribuição

Esse artefato estabelece a base normativa do uso de dados no modelo de referência, reduzindo arbitrariedade e aumentando responsabilidade institucional.

13. Artefato A10.2 — Matriz de Papéis e Responsabilidades

13.1 Finalidade

A **Matriz de Papéis e Responsabilidades** define quem é responsável por cada ativo, decisão, aprovação, correção, publicação ou monitoramento.

13.2 Estrutura mínima

Papel	Responsabilidade
Proprietário do dado	Define finalidade, autoriza uso e responde pelo ativo
Custodiante	Garante armazenamento, segurança e disponibilidade
Responsável técnico	Implementa pipelines, controles e documentação
Gestor de dados	Coordena políticas e qualidade dos dados

Analista	Consome dados conforme regras definidas
Aprovador de acesso	Autoriza ou nega solicitações de uso
Auditor	Verifica conformidade e rastreabilidade
Encarregado de privacidade	Avalia riscos de privacidade e conformidade

13.3 Métricas associadas

- taxa de ativos com responsável definido;
- percentual de papéis preenchidos;
- número de ativos sem proprietário;
- número de acessos sem aprovador definido.

13.4 Contribuição

Esse artefato evita indefinição de responsabilidade e apoia accountability no ciclo de dados.

14. Artefato A10.3 — Classificação de Sensibilidade e Restrição

14.1 Finalidade

A **Classificação de Sensibilidade e Restrição** registra o nível de sensibilidade dos ativos e define condições de acesso, uso e compartilhamento.

14.2 Estrutura mínima

Campo	Descrição
Ativo	Fonte, atributo, dataset ou produto
Tipo de dado	Público, interno, restrito, pessoal, sensível, anonimizado, agregado
Nível de sensibilidade	Baixo, moderado, alto, crítico

Risco de reidentificação	Baixo, médio, alto
Restrição de acesso	Público, interno, restrito, bloqueado
Técnica necessária	Anonimização, agregação, mascaramento etc.
Justificativa	Motivo da classificação
Responsável	Pessoa ou equipe responsável

14.3 Métricas associadas

- taxa de ativos com sensibilidade classificada;
- percentual de ativos críticos;
- percentual de ativos pessoais ou sensíveis;
- taxa de ativos com restrição compatível.

14.4 Contribuição

Esse artefato orienta decisões de segurança, privacidade, armazenamento, disponibilização e consumo.

15. Artefato A10.4 — Matriz de Acesso e Permissões

15.1 Finalidade

A **Matriz de Acesso e Permissões** define quem pode acessar cada ativo, em qual nível e com qual finalidade.

15.2 Estrutura mínima

Campo	Descrição
Ativo	Dataset, tabela, dashboard, API etc.

Usuário, perfil ou sistema	Consumidor autorizado
Tipo de acesso	Leitura, escrita, administração, exportação
Granularidade	Individual, agregado, anonimizado
Finalidade	Motivo do acesso
Validade	Prazo da permissão
Aprovador	Responsável pela autorização
Status	Ativo, vencido, revogado, pendente

15.3 Métricas associadas

- taxa de ativos com política de acesso;
- taxa de acessos autorizados;
- taxa de acessos não conformes;
- taxa de permissões vencidas;
- taxa de produtos sensíveis com controle explícito.

15.4 Contribuição

Esse artefato operacionaliza o controle de acesso e reduz riscos de uso indevido ou exposição não autorizada.

16. Artefato A10.5 — Registro de Finalidade e Base Legal

16.1 Finalidade

O **Registro de Finalidade e Base Legal** documenta por que um dado está sendo utilizado, para qual finalidade e com qual justificativa legal, institucional ou técnica.

16.2 Estrutura mínima

Campo	Descrição
Ativo	Dado ou produto analisado
Finalidade de uso	Objetivo do tratamento
Categoria de uso	Pesquisa, gestão, decisão, monitoramento, transparência etc.
Base legal ou justificativa	Fundamento legal, institucional ou autorização
Público-alvo	Quem utilizará o ativo
Compatibilidade	Compatível ou não com finalidade original
Restrições	Limitações de uso
Evidência	Documento, autorização ou registro

16.3 Métricas associadas

- taxa de ativos com finalidade documentada;
- taxa de conformidade documental;
- percentual de usos sem justificativa;
- percentual de ativos com finalidade incompatível.

16.4 Contribuição

Esse artefato reduz risco de uso inadequado dos dados e fortalece conformidade e transparência.

17. Artefato A10.6 — Plano de Privacidade e Proteção de Dados

17.1 Finalidade

O **Plano de Privacidade e Proteção de Dados** define técnicas e procedimentos para proteger dados pessoais, sensíveis ou de acesso restrito.

17.2 Estrutura mínima

Campo	Descrição
Ativo	Dataset, fonte ou produto
Risco identificado	Privacidade, reidentificação, exposição etc.
Técnica aplicada	Anonimização, mascaramento, agregação etc.
Atributos afetados	Campos protegidos
Justificativa	Motivo da técnica
Impacto na utilidade	Efeito sobre análise
Responsável	Pessoa ou equipe
Evidência	Relatório, log ou versão protegida

17.3 Métricas associadas

- taxa de ativos com anonimização aplicada;
- percentual de atributos sensíveis protegidos;
- risco residual de reidentificação;
- taxa de ativos protegidos antes da disponibilização.

17.4 Contribuição

Esse artefato permite equilibrar proteção de dados e utilidade analítica, tornando explícitas as decisões de privacidade.

18. Artefato A10.7 — Registro de Riscos de Dados

18.1 Finalidade

O **Registro de Riscos de Dados** documenta riscos associados ao uso, armazenamento, integração, publicação ou consumo dos ativos.

18.2 Tipos de risco

- acesso indevido;
- vazamento;
- reidentificação;
- uso incompatível;
- uso de dado obsoleto;
- baixa qualidade;
- viés ou impacto desproporcional;
- exposição de grupo vulnerável;
- erro de interpretação;
- dependência externa;
- descumprimento de política institucional.

18.3 Estrutura mínima

Campo	Descrição
Risco	Descrição do risco
Ativo afetado	Dado ou produto relacionado
Origem do risco	Fonte, integração, armazenamento, consumo etc.
Probabilidade	Baixa, média ou alta
Impacto	Baixo, médio, alto ou crítico
Severidade	Classificação combinada
Mitigação	Ação proposta
Responsável	Pessoa ou equipe

Status	Aberto, mitigado, aceito, encerrado

18.4 Métricas associadas

- taxa de ativos com risco registrado;
- número de riscos críticos;
- taxa de riscos mitigados;
- tempo médio de mitigação;
- risco residual.

18.5 Contribuição

Esse artefato transforma riscos difusos em elementos documentados, priorizáveis e monitoráveis.

19. Artefato A10.8 — Registro de Auditoria e Acessos

19.1 Finalidade

O **Registro de Auditoria e Acessos** documenta acessos, alterações, publicações, exportações e ações relevantes sobre os ativos.

19.2 Estrutura mínima

Campo	Descrição
Ativo	Dado, dataset, dashboard, API ou produto
Usuário ou sistema	Quem realizou a ação
Ação	Leitura, alteração, exportação, publicação etc.
Data e hora	Momento da ação
Finalidade	Motivo informado

Resultado	Sucesso, falha, negado
Ambiente	Produção, desenvolvimento, teste etc.
Evidência	Log, protocolo, registro técnico

19.3 Métricas associadas

- taxa de auditoria de acesso;
- número de acessos por ativo;
- taxa de acessos negados;
- taxa de acessos sem finalidade;
- taxa de acessos não conformes.

19.4 Contribuição

Esse artefato permite auditoria, investigação de incidentes, rastreabilidade de uso e responsabilização.

20. Artefato A10.9 — Relatório de Conformidade

20.1 Finalidade

O **Relatório de Conformidade** avalia se os ativos, processos e produtos do UnespDataLens-RM estão aderentes às políticas definidas.

20.2 Estrutura mínima

Campo	Descrição
Escopo avaliado	Ativos, módulo, produto ou pipeline
Políticas avaliadas	Regras aplicáveis
Evidências analisadas	Logs, permissões, metadados, registros

Conformidades	Itens aderentes
Não conformidades	Itens irregulares
Severidade	Grau dos problemas
Recomendações	Ações corretivas
Status final	Conforme, conforme com ressalvas, não conforme

20.3 Métricas associadas

- taxa de conformidade documental;
- taxa de ativos conformes;
- número de não conformidades;
- severidade média;
- taxa de correção de não conformidades.

20.4 Contribuição

Esse artefato permite avaliar institucionalmente a aderência do modelo de referência às regras de governança, segurança, privacidade e conformidade.

21. Artefato A10.10 — Plano de Retenção e Descarte

21.1 Finalidade

O **Plano de Retenção e Descarte** define por quanto tempo os dados serão mantidos, em quais condições e quando deverão ser arquivados, anonimizados, descartados ou descontinuados.

21.2 Estrutura mínima

Campo	Descrição
Ativo	Dado, dataset, produto ou versão
Tipo de dado	Público, interno, pessoal, sensível etc.

Período de retenção	Tempo de manutenção
Justificativa	Motivo da retenção
Ação final	Arquivar, anonimizar, excluir, descontinuar
Responsável	Pessoa ou equipe
Critério de descarte	Condição para eliminação
Evidência	Registro de execução

21.3 Métricas associadas

- percentual de ativos com política de retenção;
- número de ativos vencidos;
- taxa de descartes executados;
- taxa de ativos arquivados;
- taxa de conformidade de retenção.

21.4 Contribuição

Esse artefato garante que os dados não sejam mantidos indefinidamente sem justificativa, reduzindo riscos de privacidade, custo e obsolescência.

22. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Ativos de dados, classificações de sensibilidade, registros de acesso, metadados, produtos de dados, políticas institucionais, logs, permissões, riscos e requisitos legais.

Processamento	Definição de políticas, papéis, responsabilidades, sensibilidade, finalidade, permissões, privacidade, riscos, auditoria, conformidade, retenção e descarte.
Artefatos	Política de governança, matriz RACI, classificação de sensibilidade, matriz de acesso, registro de finalidade, plano de privacidade, registro de riscos, auditoria, relatório de conformidade e plano de retenção.
Métricas	TARD, TASC, TAFD, TAPA, TASCE, TAA, TANC, TAAA, TARR, TRM, TCD, TPV, TAuA, IMG e TPPA.
Saídas	Ativos governados, classificados, protegidos, controlados, auditáveis, conformes e adequados ao uso responsável no modelo de referência.

23. Critérios gerais de validação do módulo

O Módulo 10 pode ser considerado adequadamente implementado quando:

1. os principais ativos possuem responsável definido;
2. os ativos críticos possuem classificação de sensibilidade;
3. datasets e produtos disponibilizados possuem finalidade documentada;
4. dados pessoais e sensíveis possuem regras explícitas de acesso;
5. permissões de acesso estão registradas e possuem validade;
6. ativos sensíveis possuem controles de proteção compatíveis;
7. riscos de dados foram identificados, classificados e acompanhados;
8. publicações de produtos de dados possuem aprovação ou justificativa;
9. acessos relevantes são registrados em logs de auditoria;
10. existem evidências mínimas de conformidade documental;
11. há política de retenção e descarte para ativos relevantes;
12. não conformidades são registradas e encaminhadas para correção.

24. Relação com o pipeline principal

O Módulo 10 se relaciona com todos os módulos do pipeline principal do UnespDataLens-RM.

Com o **Módulo 1 — Fontes e Inventário de Dados**, orienta a classificação inicial de sensibilidade, finalidade, responsáveis e restrições.

Com o **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, controla permissões de extração, segurança de credenciais, armazenamento bruto e acesso aos dados originais.

Com o **Módulo 3 — Integração e Harmonização**, avalia riscos de reidentificação e aumento de sensibilidade decorrentes da combinação de bases.

Com o **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, orienta minimização, anonimização, mascaramento, supressão e proteção de atributos sensíveis.

Com o **Módulo 5 — Qualidade e Validação de Dados**, incorpora conformidade como dimensão de validação dos ativos.

Com o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, apoia uso responsável de atributos sensíveis, proxies e grupos protegidos.

Com o **Módulo 7 — Armazenamento Analítico**, define controles de acesso, criptografia, segregação de zonas, retenção e proteção de dados armazenados.

Com o **Módulo 8 — Disponibilização e Consumo Analítico**, controla publicação, acesso, granularidade, anonimização, termos de uso e auditoria do consumo.

25. Relação com os demais módulos transversais

O Módulo 10 depende fortemente do **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, pois utiliza metadados para identificar ativos, responsáveis, sensibilidade, qualidade, origem, finalidade, versões e restrições.

Com o módulo de **Proveniência, Lineage e Rastreabilidade**, compartilha registros de auditoria, origem, dependências, acesso e transformação dos dados.

Com o módulo de **Versionamento e Reprodutibilidade**, define quais versões podem ser utilizadas, preservadas, descartadas ou disponibilizadas.

Com o módulo de **Representações Analíticas: Tabular, Temporal, Espacial, Textual, Semântica e Grafos**, define restrições para exposição de relações sensíveis ou risco de reidentificação em grafos.

Com o módulo de **Feature Engineering e Feature Store**, orienta uso de atributos sensíveis, proxies e features que podem exigir restrição, anonimização ou monitoramento.

Com o módulo de **Monitoramento, Drift e Evolução dos Ativos**, acompanha acessos, riscos, permissões vencidas, mudanças de sensibilidade e conformidade ao longo do tempo.

Com o módulo de **Agentes Inteligentes de Apoio ao Pipeline**, define limites para consulta, resposta, explicação e recomendação sobre ativos sensíveis ou restritos.

26. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 10 poderia governar dados de estudantes, matrículas, histórico acadêmico, notas, frequência, dados socioeconômicos, interações em ambiente virtual, evasão e indicadores de permanência.

Esses dados podem conter informações pessoais, dados educacionais restritos, informações socioeconômicas e atributos sensíveis ou potencialmente sensíveis. Assim, o módulo definiria políticas de acesso diferenciadas para pesquisadores, gestores, analistas, orientadores, sistemas e dashboards.

Por exemplo, um dashboard público poderia apresentar apenas indicadores agregados por curso ou período, enquanto uma base individualizada de estudantes exigiria autorização específica, finalidade documentada, controle de acesso, anonimização ou pseudonimização, registro de uso e restrição de exportação.

O módulo também avaliaria riscos de reidentificação em bases anonimizadas, especialmente quando combinadas com curso, idade, turno, município, escola de origem ou condição socioeconômica.

27. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, o Módulo 10 poderia governar dados de saúde, educação, assistência social, território, habitação, trabalho, orçamento e atendimento ao cidadão.

Essas bases podem conter dados pessoais, registros familiares, informações de vulnerabilidade, atendimentos sociais, dados territoriais e indicadores sensíveis. O módulo definiria níveis de acesso para gestores, técnicos, pesquisadores, sistemas internos e painéis públicos.

Por exemplo, um painel público poderia apresentar indicadores agregados por município ou território, enquanto dados individualizados de famílias atendidas deveriam permanecer restritos, protegidos e acessíveis apenas mediante finalidade documentada e autorização.

A integração entre bases de diferentes secretarias exigiria atenção especial, pois a combinação de dados pode aumentar o risco de reidentificação mesmo quando cada base isoladamente parece pouco sensível.

28. Contribuição técnico-científica do módulo

A contribuição do Módulo 10 está em formalizar governança, segurança, privacidade e conformidade como dimensões transversais e avaliáveis da Engenharia Analítica de Dados.

Em vez de tratar essas dimensões como exigências externas ou controles administrativos isolados, o UnespDataLens-RM incorpora políticas, papéis, permissões, sensibilidade, finalidade, riscos, auditoria, retenção e conformidade como artefatos formais do modelo de referência.

Essa formalização contribui para:

- tornar explícitas responsabilidades sobre os dados;
- reduzir riscos de acesso indevido e uso incompatível;
- proteger dados pessoais e sensíveis;
- apoiar conformidade com normas legais e institucionais;
- controlar publicação e consumo de produtos analíticos;
- documentar finalidade, autorização e restrições de uso;
- permitir auditoria e responsabilização;
- apoiar uso ético e responsável de dados;
- fortalecer confiança institucional nos ativos produzidos;
- integrar governança ao ciclo completo de Engenharia Analítica de Dados.

29. Síntese do módulo

O **Módulo Transversal 10 — Governança, Segurança, Privacidade e Conformidade** estabelece as condições para que o UnespDataLens-RM produza, armazene, disponibilize e utilize dados de forma segura, responsável, controlada e aderente a normas institucionais e legais.

Sua atuação é transversal porque acompanha todos os módulos do pipeline principal, desde o inventário das fontes até o consumo dos produtos analíticos. Ao definir políticas, papéis, classificações, permissões, controles de privacidade, registros de risco, logs de auditoria, relatórios de conformidade e planos de retenção, o módulo transforma a governança em parte integrante da arquitetura analítica.

Dessa forma, o Módulo 10 fortalece a confiabilidade, a legitimidade, a segurança e a sustentabilidade do UnespDataLens-RM, garantindo que a produção de valor analítico ocorra com responsabilidade, transparência, proteção e conformidade.

Como o módulo de **Governança, Segurança, Privacidade e Conformidade** já foi estruturado anteriormente como transversal, este módulo de **Proveniência, Linhagem e Rastreabilidade** pode ser numerado como **Módulo Transversal 11** para evitar duplicidade. Mantive abaixo a redação como módulo transversal completo; se você preferir, basta ajustar a numeração no documento.

Módulo Transversal 11 — Proveniência, Linhagem e Rastreabilidade

1. Finalidade do módulo

O **Módulo Transversal 11 — Proveniência, Linhagem e Rastreabilidade** tem como finalidade registrar, representar, acompanhar e permitir a reconstrução da trajetória dos dados ao longo de todo o ciclo analítico do modelo de referência **UnespDataLens-RM**. Esse módulo atua de forma transversal sobre os módulos do pipeline principal, documentando a origem dos dados, os processos executados, as transformações aplicadas, as versões utilizadas, os artefatos gerados, as dependências entre ativos e os consumidores finais dos produtos de dados.

No contexto do UnespDataLens-RM, proveniência, linhagem e rastreabilidade são elementos fundamentais para garantir transparência, auditabilidade, reprodutibilidade, governança e confiança nos ativos analíticos produzidos. Sem esses elementos, torna-se difícil responder a perguntas essenciais, tais como: de onde veio determinado dado? Qual versão da fonte foi utilizada? Quais transformações foram aplicadas? Quais registros foram removidos, integrados ou enriquecidos? Qual regra gerou determinada variável? Qual dataset alimentou um dashboard, modelo ou indicador? Quem consumiu determinado produto de dados? Quais decisões foram tomadas com base nele?

A **proveniência** refere-se ao registro da origem dos dados, indicando fontes, sistemas, responsáveis, versões, datas de extração e condições iniciais de obtenção. A **linhagem dos dados** refere-se ao percurso dos dados ao longo do pipeline, incluindo ingestões, integrações, transformações, validações, armazenamentos, versões, produtos derivados e relações de dependência entre artefatos. A **rastreabilidade** corresponde à capacidade de acompanhar, auditar e reconstruir esse percurso, permitindo verificar como determinado ativo foi produzido e quais etapas, regras e decisões contribuíram para sua geração.

Por atuar transversalmente, o Módulo 11 não deve ser tratado como uma etapa final do pipeline. Ele acompanha todos os módulos do UnespDataLens-RM, desde o inventário das fontes até a disponibilização e consumo analítico. Em cada etapa, captura evidências, registros, manifestos, logs, versões, parâmetros, regras e relações que permitem reconstruir a história dos dados.

Assim, esse módulo transforma o pipeline em um processo transparente e auditável, no qual cada ativo analítico pode ser associado à sua origem, às suas transformações, às suas versões, aos seus responsáveis, às suas métricas de qualidade e aos seus usos.

2. Objetivo geral

O objetivo geral do Módulo 11 é **formalizar e operacionalizar a proveniência, a linhagem e a rastreabilidade dos dados no UnespDataLens-RM**, registrando origens, processos,

transformações, versões, dependências, artefatos e consumidores, de modo a permitir auditoria, reprodutibilidade, explicabilidade, governança e reconstrução dos ativos analíticos produzidos pelo modelo de referência.

3. Objetivos específicos

São objetivos específicos do Módulo 11:

1. registrar a origem das fontes de dados utilizadas no modelo de referência;
2. documentar as versões, snapshots, datas de extração e responsáveis associados a cada fonte;
3. capturar evidências dos processos de ingestão, integração, transformação, validação, armazenamento e disponibilização;
4. representar relações de dependência entre fontes, datasets, atributos, regras, features, indicadores, dashboards, APIs, modelos e produtos de dados;
5. documentar regras, parâmetros, scripts, consultas, workflows e decisões que produziram cada ativo analítico;
6. permitir reconstruir o percurso de um dado desde sua origem até seu uso final;
7. apoiar auditorias técnicas, institucionais e metodológicas sobre os ativos produzidos;
8. fornecer base para reprodutibilidade dos pipelines e comparação entre versões;
9. identificar impactos decorrentes de mudanças em fontes, schemas, regras ou datasets;
10. apoiar governança, segurança, privacidade, qualidade, versionamento, monitoramento e agentes inteligentes;
11. produzir artefatos formais de proveniência, linhagem e rastreabilidade, avaliáveis por métricas específicas.

4. Fundamentação conceitual do módulo

A proveniência, a linhagem e a rastreabilidade são conceitos relacionados, mas não equivalentes. No UnespDataLens-RM, esses conceitos devem ser tratados de forma articulada.

Conceito	Definição	Questão respondida
Proveniência	Registro da origem dos dados, incluindo fonte, responsável, versão, data, sistema e contexto de obtenção.	De onde o dado veio?
Linhagem dos dados	Registro do percurso dos dados ao longo do pipeline, incluindo transformações, integrações, dependências e produtos derivados.	Por onde o dado passou e como foi transformado?
Rastreabilidade	Capacidade de acompanhar, auditar e reconstruir o caminho dos dados, regras,	Como reconstruir, verificar ou auditar o

versões e artefatos. resultado?

A proveniência é o ponto inicial da rastreabilidade. Ela documenta a origem dos dados. A linhagem amplia essa visão, registrando o fluxo e as transformações ao longo do pipeline. A rastreabilidade, por sua vez, representa a capacidade prática de seguir esse caminho, verificar as evidências e reconstruir o processo.

No UnespDataLens-RM, esses conceitos são essenciais porque os dados passam por múltiplas etapas: fontes são inventariadas, extraídas, armazenadas em estado bruto, integradas, harmonizadas, transformadas, limpas, enriquecidas, avaliadas, armazenadas analiticamente, disponibilizadas e consumidas. Cada etapa pode alterar o conteúdo, a estrutura, a semântica, a qualidade, a sensibilidade ou a finalidade de uso dos dados.

Sem linhagem, um indicador pode ser visualizado em um dashboard sem que se saiba quais fontes foram usadas, quais regras foram aplicadas ou quais registros foram excluídos. Sem proveniência, um modelo pode ser treinado com dados cuja origem não é verificável. Sem rastreabilidade, um resultado analítico não pode ser auditado nem reproduzido.

À luz da **Design Science Research**, o Módulo 11 formaliza a proveniência, a linhagem e a rastreabilidade como artefatos técnico-metodológicos que tornam explícitas as relações entre dados, processos, decisões e produtos. Esses artefatos contribuem para transformar pipelines analíticos em processos auditáveis, explicáveis e cientificamente reprodutíveis.

5. Natureza transversal do módulo

O Módulo 11 é transversal porque acompanha todos os módulos do pipeline principal e se relaciona diretamente com os demais módulos transversais.

No **Módulo 1 — Fontes e Inventário de Dados**, registra a origem institucional, técnica e semântica das fontes, seus responsáveis, cobertura temporal, periodicidade, restrições e versões iniciais.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, registra a execução da ingestão, os logs, manifestos, snapshots, checksums, parâmetros, datas e destinos dos dados brutos.

No **Módulo 3 — Integração e Harmonização**, registra quais fontes foram integradas, quais chaves foram utilizadas, quais regras de pareamento foram aplicadas, quais conflitos foram encontrados e quais datasets integrados foram gerados.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, registra regras de transformação, atributos afetados, atributos derivados, registros removidos, imputações, parâmetros e impactos das alterações.

No **Módulo 5 — Qualidade e Validação de Dados**, registra regras de validação, métricas de qualidade, não conformidades, status de aprovação e ações corretivas.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, registra grupos avaliados, perdas por etapa, técnicas de mitigação, versões balanceadas, métricas de fairness e riscos residuais.

No **Módulo 7 — Armazenamento Analítico**, registra onde os ativos foram persistidos, em qual zona, formato, schema, versão, partição e condição de acesso.

No **Módulo 8 — Disponibilização e Consumo Analítico**, registra quais produtos foram publicados, quais ativos os alimentam, quais consumidores os utilizam, em qual versão e com qual finalidade.

Com o **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, compartilha descrições, identificadores, versões, relações semânticas e localização dos ativos.

Com o **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, compartilha trilhas de auditoria, registros de acesso, responsáveis, permissões, riscos e evidências de conformidade.

Dessa forma, o Módulo 11 cria uma visão encadeada do ciclo de vida dos dados, conectando fontes, processos, regras, artefatos, versões e consumidores.

6. Formalização sintética do módulo

O Módulo 11 pode ser formalizado como um grafo de linhagem dos dados.

Considerando o conjunto de ativos do UnespDataLens-RM:

$$\mathbf{A = \{a_1, a_2, ..., a_n\}}$$

em que cada $\mathbf{a_i}$ representa uma fonte, dataset bruto, dataset integrado, dataset tratado, data mart, feature, indicador, dashboard, API, modelo ou produto de dados, e considerando o conjunto de processos:

$$\mathbf{P = \{p_1, p_2, ..., p_k\}}$$

em que cada $\mathbf{p_j}$ representa uma operação de ingestão, integração, transformação, validação, armazenamento ou disponibilização, a linhagem pode ser representada por um grafo direcionado:

$$\mathbf{G^L = (V, E)}$$

em que:

- **V** representa o conjunto de vértices, incluindo ativos, processos, regras, versões e produtos;
- **E** representa o conjunto de arestas direcionadas, indicando dependências, transformações, consumo ou derivação.

Uma aresta pode ser representada por:

$$\mathbf{e_{ij} = \{origem, destino, tipo_relação, processo, regra, versão, data, responsável, evidência\}}$$

Cada ativo de dados pode ser descrito, sob a perspectiva da linhagem, por:

$$\mathbf{a_i = \{id, nome, tipo, fonte_origem, processos_aplicados, versões, predecessores, sucessores, regras, parâmetros, qualidade, responsável, consumidores\}}$$

Cada processo pode ser descrito por:

p_j = {id, tipo, entrada, saída, regra, parâmetro, executor, data_execução, versão, log, status}

O resultado do módulo é um grafo de linhagem capaz de responder a consultas como:

- quais fontes originaram determinado dashboard?
- quais transformações produziram determinada feature?
- quais datasets serão afetados se uma fonte mudar?
- qual versão de dados alimentou determinado relatório?
- quais produtos consumiram determinado dataset?
- quais regras foram aplicadas antes da disponibilização?

7. Entradas do módulo

As entradas do Módulo 11 são produzidas por todos os módulos do pipeline principal e pelos módulos transversais.

As principais entradas são:

- inventário de fontes;
- identificadores de fontes;
- responsáveis pelas fontes;
- versões e snapshots;
- manifestos de ingestão;
- logs de execução;
- registros de checksum;
- datasets brutos;
- planos de integração;
- matrizes de relacionamento;
- tabelas de mapeamento;
- regras de integração;
- registros de conflitos;
- manifestos de integração;
- datasets integrados;
- registros de transformação;
- regras de limpeza;
- regras de enriquecimento;
- catálogos de atributos derivados;
- manifestos de transformação;
- datasets tratados;
- regras de validação;
- relatórios de qualidade;
- registros de não conformidades;
- ações corretivas;
- registros de mitigação de viés;
- datasets balanceados ou sinalizados;
- mapas de armazenamento analítico;
- schemas analíticos;
- registros de versionamento;
- data marts;
- views;
- dashboards;
- APIs;
- relatórios;

- produtos de dados;
- registros de publicação;
- registros de consumo;
- metadados;
- políticas de governança;
- registros de acesso;
- registros de auditoria.

8. Processamento

O processamento do Módulo 11 envolve captura, padronização, associação, representação, validação e consulta da linhagem dos dados.

8.1 Captura de eventos de linhagem

A primeira atividade consiste em capturar eventos relevantes do pipeline. Cada execução de ingestão, integração, transformação, validação, armazenamento ou disponibilização deve gerar um evento de linhagem.

Um evento de linhagem registra o que foi executado, quando, por quem, sobre quais dados, com quais parâmetros, produzindo quais saídas.

8.2 Padronização dos registros de linhagem

A segunda atividade consiste em padronizar os registros de linhagem. Logs, manifestos, scripts, notebooks, workflows e relatórios devem ser traduzidos para uma estrutura comum de metadados de linhagem.

Essa padronização permite consultar e comparar eventos produzidos por diferentes ferramentas e módulos.

8.3 Associação entre entradas, processos e saídas

A terceira atividade consiste em associar cada processo às suas entradas e saídas. Um dataset tratado, por exemplo, deve apontar para o dataset integrado que o originou, para as regras aplicadas, para a versão do processo e para o artefato resultante.

Essa associação é a base para reconstruir o caminho dos dados.

8.4 Construção do grafo de linhagem

A quarta atividade consiste em representar os ativos, processos e dependências em um grafo direcionado. Nesse grafo, fontes e datasets alimentam processos; processos produzem novos datasets; datasets alimentam dashboards, APIs, modelos e relatórios.

O grafo permite navegação para frente e para trás: da fonte ao produto ou do produto à fonte.

8.5 Registro de linhagem em nível de ativo

A quinta atividade consiste em registrar a linhagem em nível de ativo. Esse nível indica relações entre fontes, datasets, tabelas, data marts, dashboards, APIs e modelos.

Por exemplo: “Dashboard de evasão” deriva do “Data Mart de permanência”, que deriva do “Dataset tratado de trajetória acadêmica”, que deriva das fontes “Matrícula”, “Histórico”, “AVA” e “Questionário socioeconômico”.

8.6 Registro de linhagem em nível de atributo

A sexta atividade consiste em registrar, quando necessário, a linhagem em nível de atributo. Esse nível é mais detalhado e indica quais atributos de origem contribuíram para gerar determinado atributo derivado, indicador ou feature.

Por exemplo: o atributo “taxa de aprovação acumulada” pode derivar de nota final, situação de matrícula, disciplina cursada e período letivo.

8.7 Registro de linhagem em nível de processo

A sétima atividade consiste em documentar os processos que transformam dados. Isso inclui scripts, consultas SQL, notebooks, workflows, parâmetros, bibliotecas, versões, ambiente de execução e status.

Esse nível é essencial para reprodutibilidade.

8.8 Validação da rastreabilidade

A oitava atividade consiste em verificar se os ativos críticos possuem linhagem suficiente para auditoria. Um produto de dados crítico deve permitir identificar suas fontes, versões, transformações, validações, armazenamento e publicação.

Ativos sem linhagem suficiente devem ser sinalizados como incompletos ou de uso restrito.

8.9 Consultas de impacto

A nona atividade consiste em permitir análise de impacto. Caso uma fonte, atributo, regra ou dataset seja alterado, o grafo de linhagem deve permitir identificar quais ativos derivados podem ser afetados.

8.10 Disponibilização da linhagem para auditoria e explicação

A décima atividade consiste em disponibilizar a linhagem para usuários autorizados, auditores, gestores, analistas, módulos de governança e agentes inteligentes, de modo que seja possível explicar como um resultado foi produzido.

9. Técnicas, métodos e tecnologias aplicáveis

O Módulo 11 pode empregar diferentes técnicas, métodos e tecnologias para capturar, representar e consultar proveniência e linhagem.

Entre as técnicas e métodos, destacam-se:

- captura de logs estruturados;
- manifestos de execução;
- registros de transformação;
- versionamento de datasets;
- rastreamento de dependências;

- grafos de linhagem;
- modelagem de proveniência;
- trilhas de auditoria;
- análise de impacto;
- rastreamento em nível de dataset;
- rastreamento em nível de atributo;
- rastreamento em nível de processo;
- data lineage;
- workflow lineage;
- provenance tracking;
- data observability;
- auditoria de pipelines;
- comparação de versões;
- documentação de dependências.

Entre os padrões e abordagens possíveis, destacam-se:

- OpenLineage;
- W3C PROV;
- modelos de grafos direcionados;
- registros baseados em eventos;
- metadados de execução;
- data contracts;
- manifestos de pipeline;
- versionamento por snapshots;
- identificadores persistentes;
- hashes e checksums;
- logs de auditoria.

Entre as tecnologias possíveis, destacam-se:

- Apache Atlas;
- DataHub;
- OpenMetadata;
- Marquez;
- OpenLineage;
- Airflow;
- dbt;
- Spark;
- NiFi;
- MLflow, quando houver modelos;
- DVC;
- LakeFS;
- Git;
- Neo4j;
- bancos de grafos;
- catálogos de dados;
- ferramentas de observabilidade;
- sistemas de logs centralizados.

10. Métricas do módulo

As métricas do Módulo 11 têm como objetivo avaliar cobertura, completude, profundidade, auditabilidade e utilidade da linhagem dos dados.

Métrica	Sigla	Descrição
Cobertura de linhagem de ativos	CLA	Proporção de ativos com linhagem registrada
Cobertura de proveniência	CP	Proporção de ativos com origem documentada
Cobertura de linhagem de atributos	CLAtr	Proporção de atributos críticos com linhagem registrada
Cobertura de processos rastreáveis	CPR	Proporção de processos com entrada, saída e parâmetros documentados
Taxa de ativos auditáveis	TAA	Proporção de ativos cuja trajetória pode ser reconstruída
Taxa de produtos com fontes identificáveis	TPFI	Proporção de produtos que permitem identificar fontes originais
Taxa de transformações rastreáveis	TTR	Proporção de transformações com regra, entrada e saída documentadas
Profundidade média da linhagem	PML	Número médio de etapas rastreáveis entre fonte e produto
Taxa de dependências documentadas	TDD	Proporção de relações entre ativos formalmente registradas
Taxa de impacto rastreável	TIR	Proporção de mudanças cuja propagação pode ser identificada
Taxa de linhagem incompleta	TLI	Proporção de ativos com lacunas de rastreabilidade

Tempo médio de reconstrução	TMR	Tempo necessário para reconstruir a trajetória de um ativo
Taxa de evidências associadas	TEA	Proporção de eventos com logs, manifestos ou registros associados
Taxa de consultas de linhagem respondidas	TCLR	Proporção de consultas respondidas pelo grafo de linhagem
Índice de auditabilidade do ativo	IAA	Grau em que um ativo pode ser auditado a partir de sua linhagem

10.1 Cobertura de Linhagem de Ativos — CLA

A **Cobertura de Linhagem de Ativos** mede a proporção de ativos que possuem relações de origem, transformação ou dependência registradas.

CLA = número de ativos com linhagem registrada / número total de ativos catalogados

Essa métrica indica se os ativos do modelo de referência possuem rastreabilidade mínima.

10.2 Cobertura de Proveniência — CP

A **Cobertura de Proveniência** mede a proporção de ativos com origem documentada.

CP = número de ativos com fonte de origem documentada / número total de ativos avaliados

Essa métrica avalia se é possível identificar de onde os dados vieram.

10.3 Taxa de Transformações Rastreáveis — TTR

A **Taxa de Transformações Rastreáveis** mede a proporção de transformações com entrada, saída, regra, parâmetro e evidência documentados.

TTR = número de transformações rastreáveis / número total de transformações aplicadas

Essa métrica é fundamental para reprodutibilidade.

10.4 Taxa de Ativos Auditáveis — TAA

A **Taxa de Ativos Auditáveis** mede a proporção de ativos cuja trajetória pode ser reconstruída de forma suficiente para auditoria.

TAA = número de ativos auditáveis / número total de ativos críticos

Um ativo é considerado auditável quando possui origem, versões, processos, regras, responsáveis, métricas e produtos derivados documentados.

10.5 Índice de Auditabilidade do Ativo — IAA

O **Índice de Auditabilidade do Ativo** pode ser definido como a proporção de elementos obrigatórios de linhagem preenchidos para um ativo.

$IAA(a_i) = ELP(a_i) / ELO$

em que:

- **$ELP(a_i)$** representa o número de elementos de linhagem preenchidos para o ativo **a_i**;
- **ELO** representa o total de elementos de linhagem obrigatórios.

Elementos obrigatórios podem incluir fonte, versão, processo, regra, parâmetro, entrada, saída, responsável, data e evidência.

11. Artefatos formalizados do módulo

Os artefatos do Módulo 11 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, regras de construção, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A11.1	Modelo de Proveniência dos Dados	Modelo conceitual	CP	Define como a origem dos dados será registrada
A11.2	Grafo de Linhagem dos Dados	Grafo direcionado	CLA, TAA	Representa dependências entre ativos e processos
A11.3	Registro de Eventos de Linhagem	Log estruturado	TEA	Captura execuções relevantes do pipeline
A11.4	Manifesto de Linhagem do Ativo	Documento estruturado	IAA	Resume trajetória de um ativo
A11.5	Matriz de	Matriz ativo ×	TDD	Mapeia relações e

	Dependência entre Ativos	ativo		impactos
A11.6	Registro de Linhagem de Atributos	Tabela atributo-origem	CLAtr	Rastreia atributos críticos e derivados
A11.7	Registro de Transformações Rastreáveis	Tabela técnica	TTR	Documenta regras e processos aplicados
A11.8	Relatório de Impacto de Mudanças	Relatório técnico	TIR	Avalia propagação de alterações
A11.9	Trilha de Auditoria Analítica	Log/relatório	TAA	Permite verificar produção e uso dos ativos
A11.10	Relatório de Completude da Linhagem	Relatório técnico	CLA, TLI	Avalia lacunas de rastreabilidade

12. Artefato A11.1 — Modelo de Proveniência dos Dados

12.1 Finalidade

O **Modelo de Proveniência dos Dados** define quais informações devem ser registradas para documentar a origem dos dados utilizados no UnespDataLens-RM.

12.2 Tipo de artefato na DSR

Esse artefato é um **modelo** e um **constructo**, pois formaliza os elementos essenciais para descrever a origem dos dados.

12.3 Estrutura mínima

Campo	Descrição
Fonte de origem	Sistema, base, arquivo, API ou repositório

Responsável pela fonte	Pessoa, setor ou instituição
Data de obtenção	Momento de extração ou acesso
Versão da fonte	Snapshot, data ou número de versão
Método de obtenção	API, SQL, arquivo, exportação etc.
Cobertura temporal	Período representado
Granularidade	Unidade de análise
Restrição de acesso	Condição de uso
Sensibilidade	Classificação dos dados
Evidência	Manifesto, log, checksum ou documentação

12.4 Métricas associadas

- cobertura de proveniência;
- percentual de fontes com responsável;
- percentual de fontes com versão;
- percentual de fontes com evidência de origem;
- percentual de fontes com restrição documentada.

12.5 Critérios de validação

A proveniência é considerada adequada quando a origem, responsável, data, versão, método de obtenção, cobertura, sensibilidade e evidência estão documentados.

12.6 Contribuição

Esse artefato garante que os dados possam ser vinculados à sua origem, evitando uso de informações sem procedência verificável.

13. Artefato A11.2 — Grafo de Linhagem dos Dados

13.1 Finalidade

O **Grafo de Linhagem dos Dados** representa as relações entre fontes, processos, datasets, atributos, produtos e consumidores ao longo do pipeline.

13.2 Tipo de artefato na DSR

Esse artefato é um **modelo**, pois representa relações e dependências, e uma **instanciação**, quando implementado em banco de grafos, catálogo, ferramenta de lineage ou estrutura computacional.

13.3 Estrutura formal

O grafo pode ser representado como:

$$\mathbf{G^L = (V, E)}$$

em que:

- **V** representa ativos, processos, versões, regras e produtos;
- **E** representa relações de origem, transformação, derivação, dependência ou consumo.

Cada vértice pode conter:

$$\mathbf{v_i = \{id, tipo, nome, versão, módulo, responsável, status\}}$$

Cada aresta pode conter:

$$\mathbf{e_{ij} = \{origem, destino, tipo_relação, processo, regra, data, evidência\}}$$

13.4 Tipos de vértices

- fonte;
- dataset bruto;
- dataset integrado;
- dataset tratado;
- dataset validado;
- data mart;
- atributo;
- feature;
- indicador;
- processo;
- regra;
- modelo;
- dashboard;
- API;
- relatório;

- produto de dados;
- consumidor.

13.5 Tipos de relações

- originou;
- foi extraído de;
- foi integrado com;
- foi transformado por;
- gerou;
- derivou de;
- validou;
- alimentou;
- consumiu;
- substituiu;
- depende de.

13.6 Métricas associadas

- cobertura de linhagem de ativos;
- profundidade média da linhagem;
- taxa de dependências documentadas;
- taxa de produtos com fontes identificáveis;
- taxa de consultas de linhagem respondidas.

13.7 Contribuição

Esse artefato permite visualizar e consultar a trajetória dos dados, apoiando auditoria, impacto de mudanças, governança, explicabilidade e reprodutibilidade.

14. Artefato A11.3 — Registro de Eventos de Linhagem

14.1 Finalidade

O **Registro de Eventos de Linhagem** captura eventos relevantes ocorridos durante a execução do pipeline.

14.2 Estrutura mínima

Campo	Descrição
Identificador do evento	Código único
Tipo de evento	Ingestão, integração, transformação, validação, publicação etc.

Ativos de entrada	Dados ou artefatos consumidos
Ativos de saída	Dados ou artefatos produzidos
Processo executado	Pipeline, script, notebook, consulta ou workflow
Parâmetros	Configurações utilizadas
Executor	Pessoa, sistema ou agente
Data e hora	Momento da execução
Status	Sucesso, falha, parcial
Evidência	Log, manifesto, checksum ou relatório

14.3 Métricas associadas

- taxa de evidências associadas;
- cobertura de processos rastreáveis;
- taxa de eventos com status;
- taxa de eventos com parâmetros documentados;
- taxa de eventos com entrada e saída registradas.

14.4 Contribuição

Esse artefato transforma execuções do pipeline em eventos auditáveis, preservando a história operacional dos dados.

15. Artefato A11.4 — Manifesto de Linhagem do Ativo

15.1 Finalidade

O **Manifesto de Linhagem do Ativo** resume a trajetória de um ativo específico, indicando suas fontes, processos, versões, transformações, validações, armazenamento e consumo.

15.2 Estrutura mínima

Campo	Descrição
Ativo	Dataset, indicador, dashboard, API ou produto
Fontes de origem	Fontes que contribuíram para o ativo
Versões utilizadas	Versões ou snapshots
Processos aplicados	Ingestão, integração, transformação etc.
Regras aplicadas	Regras de limpeza, integração, validação etc.
Datasets intermediários	Ativos gerados no percurso
Métricas de qualidade	Status ou indicadores associados
Responsáveis	Pessoas ou sistemas envolvidos
Produtos consumidores	Dashboards, APIs, modelos, relatórios
Evidências	Logs, manifestos, registros, checksums

15.3 Métricas associadas

- índice de auditabilidade do ativo;
- taxa de ativos auditáveis;
- cobertura de proveniência;
- taxa de transformações rastreáveis.

15.4 Contribuição

Esse artefato permite compreender a história completa de um ativo e apoiar auditoria, explicação e reprodutibilidade.

16. Artefato A11.5 — Matriz de Dependência entre Ativos

16.1 Finalidade

A **Matriz de Dependência entre Ativos** registra relações de dependência entre fontes, datasets, atributos, produtos e consumidores.

16.2 Estrutura mínima

Campo	Descrição
Ativo de origem	Ativo predecessor
Ativo dependente	Ativo sucessor
Tipo de dependência	Derivação, transformação, consumo, validação etc.
Processo associado	Processo que conecta os ativos
Nível	Dataset, atributo, produto ou processo
Criticidade	Baixa, média, alta, crítica
Evidência	Manifesto, log ou regra
Observações	Limitações ou cuidados

16.3 Métricas associadas

- taxa de dependências documentadas;
- número de dependências por ativo;
- número de ativos críticos dependentes;

- taxa de impacto rastreável;
- grau de centralidade dos ativos.

16.4 Contribuição

Esse artefato apoia análise de impacto, identificação de ativos críticos e gestão de mudanças no pipeline.

17. Artefato A11.6 — Registro de Linhagem de Atributos

17.1 Finalidade

O **Registro de Linhagem de Atributos** documenta a origem e as transformações de atributos críticos, derivados, sensíveis ou utilizados em indicadores, modelos e decisões.

17.2 Estrutura mínima

Campo	Descrição
Atributo final	Nome do atributo no ativo resultante
Ativo final	Dataset ou produto onde aparece
Atributos de origem	Campos que contribuíram para sua criação
Fontes de origem	Fontes ou datasets originais
Regra de derivação	Fórmula ou transformação
Transformações aplicadas	Limpeza, imputação, agregação etc.
Granularidade	Unidade de análise
Versão	Versão da regra ou atributo

Evidência	Script, manifesto ou registro
Limitações	Cuidados de interpretação

17.3 Métricas associadas

- cobertura de linhagem de atributos;
- percentual de atributos derivados rastreáveis;
- percentual de atributos sensíveis rastreáveis;
- taxa de atributos sem origem identificada.

17.4 Contribuição

Esse artefato é essencial para explicar indicadores, features, variáveis derivadas e atributos utilizados em modelos ou dashboards.

18. Artefato A11.7 — Registro de Transformações Rastreáveis

18.1 Finalidade

O **Registro de Transformações Rastreáveis** documenta regras, parâmetros, entradas, saídas e evidências associadas a transformações aplicadas aos dados.

18.2 Estrutura mínima

Campo	Descrição
Transformação	Código ou nome da transformação
Tipo	Limpeza, integração, enriquecimento, validação etc.
Entrada	Dataset ou atributo de entrada
Saída	Dataset ou atributo resultante
Regra	Descrição da regra aplicada

Parâmetros	Valores utilizados
Registros afetados	Quantidade ou percentual
Executor	Pessoa ou sistema
Data	Data da execução
Evidência	Script, notebook, log, manifesto

18.3 Métricas associadas

- taxa de transformações rastreáveis;
- taxa de evidências associadas;
- percentual de transformações com parâmetros;
- percentual de transformações com impacto documentado.

18.4 Contribuição

Esse artefato permite verificar como os dados foram modificados e sustenta reprodutibilidade das etapas de preparação.

19. Artefato A11.8 — Relatório de Impacto de Mudanças

19.1 Finalidade

O **Relatório de Impacto de Mudanças** avalia quais ativos podem ser afetados por alterações em fontes, schemas, regras, atributos ou processos.

19.2 Estrutura mínima

Campo	Descrição
Mudança analisada	Alteração em fonte, atributo, regra ou processo

Ativo alterado	Elemento de origem da mudança
Ativos impactados	Datasets, produtos, dashboards ou modelos afetados
Tipo de impacto	Estrutural, semântico, qualidade, consumo etc.
Severidade	Baixa, média, alta ou crítica
Ação recomendada	Reprocessar, revisar, bloquear, comunicar
Responsável	Pessoa ou equipe
Status	Aberto, tratado, encerrado

19.3 Métricas associadas

- taxa de impacto rastreável;
- número de ativos impactados;
- número de impactos críticos;
- tempo médio de avaliação de impacto;
- taxa de impactos tratados.

19.4 Contribuição

Esse artefato permite gerenciar mudanças de forma controlada, evitando efeitos não percebidos em produtos de dados.

20. Artefato A11.9 — Trilha de Auditoria Analítica

20.1 Finalidade

A **Trilha de Auditoria Analítica** consolida registros que permitem verificar a produção, alteração, publicação e consumo dos ativos analíticos.

20.2 Estrutura mínima

Campo	Descrição

Ativo	Dado, dataset, indicador ou produto
Ação	Criação, alteração, validação, publicação, consumo
Usuário ou sistema	Quem executou a ação
Data e hora	Momento da ação
Módulo	Etapa do modelo de referência
Evidência	Log, manifesto, relatório, versão
Resultado	Sucesso, falha, pendente
Observações	Informações adicionais

20.3 Métricas associadas

- taxa de ativos auditáveis;
- taxa de ações com evidência;
- taxa de auditoria por produto;
- tempo médio de reconstrução.

20.4 Contribuição

Esse artefato apoia responsabilização, governança, investigação de problemas e validação institucional dos resultados analíticos.

21. Artefato A11.10 — Relatório de Completude da Linhagem

21.1 Finalidade

O **Relatório de Completude da Linhagem** avalia lacunas de proveniência, linhagem e rastreabilidade dos ativos do modelo de referência.

21.2 Estrutura mínima

Campo	Descrição
Escopo avaliado	Ativos, módulo ou pipeline completo
Ativos avaliados	Quantidade e tipos
Ativos com proveniência	Percentual
Ativos com linhagem completa	Percentual
Ativos com lacunas	Quantidade e descrição
Processos sem evidência	Processos não rastreáveis
Atributos sem origem	Atributos críticos sem linhagem
Severidade	Classificação das lacunas
Recomendações	Ações de melhoria

21.3 Métricas associadas

- cobertura de linhagem de ativos;
- taxa de linhagem incompleta;
- cobertura de proveniência;
- taxa de processos rastreáveis;
- índice de auditabilidade.

21.4 Contribuição

Esse artefato permite avaliar a própria qualidade da rastreabilidade, identificando pontos frágeis do pipeline.

22. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Inventários, logs, manifestos, snapshots, regras, datasets, transformações, relatórios de qualidade, registros de armazenamento, produtos de dados, versões e registros de consumo.
Processamento	Captura de eventos, padronização de registros, associação entre entradas/processos/saídas, construção de grafo de linhagem, validação da rastreabilidade, análise de impacto e disponibilização para auditoria.
Artefatos	Modelo de proveniência, grafo de linhagem, eventos de linhagem, manifesto de ativo, matriz de dependência, linhagem de atributos, transformações rastreáveis, relatório de impacto, trilha de auditoria e relatório de completude.
Métricas	CLA, CP, CLAtr, CPR, TAA, TPFI, TTR, PML, TDD, TIR, TLI, TMR, TEA, TCLR e IAA.
Saídas	Ativos com origem, percurso, dependências, versões, transformações, consumidores e evidências documentadas, permitindo auditoria, reprodutibilidade e análise de impacto.

23. Critérios gerais de validação do módulo

O Módulo 11 pode ser considerado adequadamente implementado quando:

1. os ativos críticos possuem origem documentada;
2. os datasets brutos possuem registros de proveniência;
3. as ingestões possuem logs e manifestos associados;
4. as integrações possuem fontes, chaves, regras e saídas documentadas;
5. as transformações possuem entrada, saída, regra, parâmetro e evidência;
6. os datasets tratados possuem vínculo com datasets de origem;
7. os produtos de dados possuem relação com os ativos que os alimentam;
8. os atributos derivados críticos possuem linhagem documentada;
9. as versões e snapshots estão associados aos ativos correspondentes;
10. é possível navegar da fonte ao produto e do produto à fonte;
11. mudanças em fontes ou regras permitem análise de impacto;
12. os ativos críticos possuem índice de auditabilidade satisfatório.

24. Relação com o pipeline principal

O Módulo 11 se relaciona com todos os módulos do pipeline principal do UnespDataLens-RM.

Com o **Módulo 1 — Fontes e Inventário de Dados**, registra a proveniência inicial das fontes.

Com o **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, registra eventos de ingestão, snapshots, logs e datasets brutos.

Com o **Módulo 3 — Integração e Harmonização**, registra chaves, regras de pareamento, conflitos e datasets integrados.

Com o **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, registra regras, parâmetros, atributos afetados, atributos derivados e impactos.

Com o **Módulo 5 — Qualidade e Validação de Dados**, registra testes, métricas, não conformidades, status e ações corretivas.

Com o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, registra perdas por grupo, mitigação, versões balanceadas e riscos residuais.

Com o **Módulo 7 — Armazenamento Analítico**, registra zonas, localizações, versões, schemas e estruturas persistidas.

Com o **Módulo 8 — Disponibilização e Consumo Analítico**, registra produtos publicados, consumidores, dashboards, APIs, relatórios e versões de consumo.

25. Relação com os demais módulos transversais

O Módulo 11 depende do **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, pois utiliza identificadores, descrições, schemas, termos, relações semânticas e metadados dos ativos.

Relaciona-se com o **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, pois compartilha trilhas de auditoria, responsáveis, acessos, evidências, finalidades e registros de conformidade.

Relaciona-se com o módulo de **Versionamento e Reprodutibilidade**, pois a linhagem depende de versões, snapshots, parâmetros, ambientes e artefatos preservados.

Relaciona-se com o módulo de **Monitoramento, Drift e Evolução dos Ativos**, pois a linhagem permite identificar o impacto de mudanças, degradações, alterações de schema e evolução dos ativos.

Relaciona-se com o módulo de **Agentes Inteligentes de Apoio ao Pipeline**, pois agentes podem usar o grafo de linhagem para responder perguntas, explicar indicadores, rastrear erros, sugerir reprocessamentos e identificar impactos.

26. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 11 poderia registrar a linhagem de um painel de evasão.

O painel poderia ser derivado de um data mart de permanência, que por sua vez foi construído a partir de um dataset tratado de trajetória acadêmica. Esse dataset tratado poderia ter origem em bases de matrícula, histórico acadêmico, notas, frequência, ambiente virtual de aprendizagem e questionários socioeconômicos.

A linhagem permitiria responder:

- quais fontes alimentam o indicador de evasão?
- qual definição de evasão foi utilizada?
- quais períodos letivos foram considerados?
- quais registros foram removidos por inconsistência?
- quais atributos foram derivados?
- qual versão da base alimentou o painel?
- quais dashboards ou relatórios dependem desse dataset?
- quais produtos seriam afetados se a regra de evasão fosse alterada?

Esse tipo de rastreabilidade é essencial para garantir transparência e confiança em análises educacionais.

27. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, o Módulo 11 poderia registrar a linhagem de um indicador de vulnerabilidade territorial.

Esse indicador poderia ser derivado de bases de assistência social, saúde, educação, território, cadastro populacional e indicadores socioeconômicos externos. A linhagem permitiria identificar quais fontes contribuíram para o indicador, quais regras de agregação foram usadas, quais territórios foram considerados, quais versões das bases foram utilizadas e quais painéis consomem o resultado.

Se uma base de atendimento social fosse atualizada ou corrigida, o grafo de linhagem permitiria identificar automaticamente quais indicadores, dashboards, relatórios e decisões poderiam ser impactados.

28. Contribuição técnico-científica do módulo

A contribuição do Módulo 11 está em formalizar proveniência, linhagem e rastreabilidade como dimensões transversais da Engenharia Analítica de Dados.

Em vez de tratar a rastreabilidade como subproduto de logs ou documentação dispersa, o UnespDataLens-RM propõe que cada fonte, processo, transformação, versão, artefato e produto de dados seja conectado em uma estrutura formal de linhagem.

Essa formalização contribui para:

- tornar explícita a origem dos dados;
- permitir reconstrução dos pipelines;
- apoiar auditoria técnica e institucional;
- fortalecer reprodutibilidade científica;
- explicar indicadores, dashboards e modelos;
- identificar impactos de mudanças;
- conectar dados a regras, processos e versões;
- reduzir dependência de conhecimento tácito;

- apoiar governança, qualidade e conformidade;
- fornecer base para agentes inteligentes explicarem a trajetória dos dados.

29. Síntese do módulo

O **Módulo Transversal 11 — Proveniência, Linhagem e Rastreabilidade** registra e representa a trajetória dos dados ao longo de todo o UnespDataLens-RM. Ele conecta fontes, processos, transformações, validações, versões, armazenamentos, produtos e consumidores em uma estrutura auditável e consultável.

Sua importância é transversal porque permite compreender não apenas quais dados foram produzidos, mas como, quando, por quem, com quais regras, a partir de quais fontes e para quais usos. Ao formalizar a linhagem dos dados, o módulo fortalece transparência, governança, auditoria, explicabilidade, reprodutibilidade e confiança nos ativos analíticos.

Dessa forma, o Módulo 11 transforma o pipeline do UnespDataLens-RM em um processo rastreável, no qual cada produto analítico pode ser reconduzido às suas fontes, regras, versões, evidências e decisões de processamento.

Módulo Transversal 12 — Versionamento e Reprodutibilidade

1. Finalidade do módulo

O **Módulo Transversal 12 — Versionamento e Reprodutibilidade** tem como finalidade controlar, registrar, preservar e permitir a reconstrução das versões dos dados, scripts, configurações, parâmetros, ambientes, regras, modelos, artefatos e produtos analíticos produzidos ao longo do modelo de referência **UnespDataLens-RM**.

Esse módulo atua de forma transversal sobre todo o ciclo de vida dos dados, desde o inventário das fontes até a disponibilização e consumo dos produtos analíticos. Sua função é garantir que os resultados produzidos pelo modelo de referência possam ser compreendidos, auditados, repetidos, comparados, reprocessados e validados ao longo do tempo.

No contexto da Engenharia Analítica de Dados, versionamento e reprodutibilidade são dimensões essenciais para assegurar confiança, transparência e rigor científico-técnico. Um pipeline analítico pode produzir indicadores, dashboards, modelos, grafos, relatórios e produtos de dados; entretanto, se não for possível identificar quais versões dos dados, scripts, parâmetros, regras, bibliotecas, ambientes e artefatos foram utilizados, os resultados tornam-se frágeis, difíceis de auditar e impossíveis de reproduzir de forma confiável.

No **UnespDataLens-RM**, o versionamento não se limita ao controle de código-fonte. Ele abrange dados brutos, datasets integrados, datasets tratados, datasets validados, versões balanceadas ou sinalizadas, data marts, APIs, dashboards, indicadores, features, modelos, relatórios, parâmetros de execução, configurações, schemas, regras de transformação, regras de qualidade, políticas de acesso, snapshots e produtos disponibilizados para consumo.

A reprodutibilidade, por sua vez, refere-se à capacidade de reconstruir uma execução, um dataset, um indicador, um modelo, um painel ou um produto de dados a partir das versões, parâmetros, regras, dependências e ambientes registrados. Isso permite responder a questões como: qual versão da base foi usada? Qual regra gerou determinado indicador? Qual script produziu determinado dataset? Qual snapshot alimentou o dashboard? O resultado pode ser reproduzido hoje? A nova execução gera o mesmo resultado? Quais diferenças surgiram entre versões?

Assim, o Módulo 12 formaliza versionamento e reprodutibilidade como componentes transversais do modelo de referência, conectando dados, processos, linhagem, metadados, governança, qualidade, armazenamento e consumo analítico.

2. Objetivo geral

O objetivo geral do Módulo 12 é **garantir o controle de versões e a reprodutibilidade dos ativos analíticos produzidos pelo UnespDataLens-RM**, por meio do registro sistemático de versões de dados, códigos, parâmetros, configurações, ambientes, regras, artefatos e produtos, permitindo reconstrução, comparação, auditoria e validação dos resultados gerados.

3. Objetivos específicos

São objetivos específicos do Módulo 12:

1. controlar versões de fontes, datasets, snapshots, data marts, produtos de dados, dashboards, APIs, relatórios, features, modelos e demais ativos analíticos;
2. registrar versões de scripts, notebooks, consultas SQL, workflows, regras de transformação, regras de validação, configurações e parâmetros;
3. preservar ambientes computacionais, dependências, bibliotecas, containers, versões de ferramentas e configurações necessárias à execução dos pipelines;
4. permitir reconstruir datasets, indicadores, modelos, dashboards ou produtos de dados a partir de versões e parâmetros documentados;
5. comparar resultados produzidos por diferentes versões dos dados, regras, scripts ou ambientes;
6. identificar mudanças entre versões, incluindo alterações de schema, conteúdo, regras, parâmetros, qualidade, distribuição, sensibilidade ou produtos derivados;
7. apoiar auditoria, validação científica, governança, lineage, monitoramento e manutenção dos ativos analíticos;
8. reduzir dependência de conhecimento tácito, scripts não documentados e execuções manuais não rastreáveis;
9. definir critérios de aceitação para considerar um ativo, pipeline ou produto como reprodutível;
10. produzir artefatos formais de versionamento e reprodutibilidade, avaliáveis por métricas específicas.

4. Fundamentação conceitual do módulo

O versionamento corresponde ao controle sistemático de diferentes estados de um ativo ao longo do tempo. Em Engenharia Analítica de Dados, os ativos versionáveis incluem não apenas código, mas também dados, configurações, parâmetros, schemas, ambientes, modelos, resultados, dashboards e produtos de dados.

A reprodutibilidade corresponde à capacidade de executar novamente um processo analítico e obter o mesmo resultado, ou um resultado explicavelmente equivalente, a partir dos mesmos dados, versões, regras, parâmetros e ambiente. Em alguns contextos, é possível exigir reprodutibilidade exata; em outros, especialmente quando há aleatoriedade ou processamento distribuído, pode-se buscar reprodutibilidade controlada ou explicável.

No UnespDataLens-RM, é importante distinguir:

Conceito	Definição	Exemplo
Versionamento de dados	Controle de versões de datasets, snapshots, tabelas e arquivos.	Base de matrículas versão 2026.1.
Versionamento de código	Controle de scripts, notebooks, queries, pipelines e workflows.	Script de integração v2.3.
Versionamento de regras	Controle de regras de transformação, qualidade, cálculo e classificação.	Regra de evasão v1.1.
Versionamento de	Registro de bibliotecas, containers,	Python 3.11, pandas

ambiente	ferramentas e configurações.	2.x, imagem Docker.
Versionamento de produtos	Controle de versões de dashboards, APIs, relatórios e data marts.	Dashboard de evasão v4.0.
Reprodutibilidade técnica	Capacidade de reexecutar o pipeline com os mesmos insumos e parâmetros.	Reprocessar dataset tratado.
Reprodutibilidade analítica	Capacidade de reconstruir indicadores, métricas, modelos ou resultados.	Reproduzir taxa de evasão publicada.
Reprodutibilidade científica	Capacidade de permitir validação externa ou auditoria metodológica.	Reproduzir experimento de modelagem.

Esse módulo está diretamente associado à confiabilidade dos resultados. Sem versionamento, não é possível saber se uma diferença entre dois resultados ocorreu por mudança nos dados, alteração de regra, atualização de ambiente, correção de erro, mudança de filtro, nova versão de biblioteca ou modificação de parâmetro.

À luz da **Design Science Research**, o Módulo 12 formaliza versionamento e reprodutibilidade como artefatos técnico-metodológicos projetados para resolver problemas de perda de controle, baixa rastreabilidade, impossibilidade de auditoria, resultados não reproduzíveis e fragilidade científica dos pipelines analíticos.

5. Natureza transversal do módulo

O Módulo 12 é transversal porque todas as etapas do pipeline principal produzem artefatos que precisam ser versionados e, quando necessário, reproduzidos.

No **Módulo 1 — Fontes e Inventário de Dados**, versiona o inventário de fontes, dicionários preliminares, mapas de bases, registros de granularidade, classificações de sensibilidade e limitações conhecidas.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, versiona snapshots, dados brutos, manifestos de ingestão, configurações de extração, logs, checksums e parâmetros de conexão.

No **Módulo 3 — Integração e Harmonização**, versiona regras de integração, tabelas de mapeamento, chaves utilizadas, datasets integrados, manifestos de integração e registros de conflitos.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, versiona regras de transformação, scripts, atributos derivados, parâmetros de imputação, datasets tratados e relatórios de impacto.

No **Módulo 5 — Qualidade e Validação de Dados**, versiona regras de qualidade, planos de validação, limiares, relatórios de qualidade, não conformidades e status de aprovação.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, versiona métricas de distribuição, planos de mitigação, parâmetros de balanceamento, datasets balanceados ou sinalizados e relatórios de equidade.

No **Módulo 7 — Armazenamento Analítico**, versiona estruturas persistidas, schemas, data marts, views, tabelas, arquivos, partições e snapshots analíticos.

No **Módulo 8 — Disponibilização e Consumo Analítico**, versiona produtos de dados, dashboards, APIs, relatórios, datasets curados para consumo, registros de publicação e interfaces.

Com o **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, mantém versões de metadados, glossários, dicionários, fichas de ativos e modelos semânticos.

Com o **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, versiona políticas, permissões, classificações, registros de finalidade, planos de privacidade e regras de retenção.

Com o **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, utiliza a linhagem para reconstruir versões e fornece evidências para auditoria e reprocessamento.

6. Formalização sintética do módulo

O Módulo 12 pode ser formalizado como um sistema de controle de versões e reconstrução de execuções analíticas.

Considerando o conjunto de ativos do UnespDataLens-RM:

$$\mathbf{A = \{a_1, a_2, ..., a_n\}}$$

em que cada $\mathbf{a_i}$ representa uma fonte, dataset, atributo, regra, script, configuração, ambiente, produto ou relatório, cada ativo pode possuir um conjunto de versões:

$$\mathbf{V(a_i) = \{v_1, v_2, ..., v_k\}}$$

Cada versão $\mathbf{v_j}$ pode ser descrita por:

$$\mathbf{v_j = \{id_versão, ativo, data, autor, descrição_mudança, hash, estado, dependências, status, evidência\}}$$

Uma execução reprodutível pode ser representada por:

$$\mathbf{E = \{id, entradas, versões_entradas, processos, versões_processos, parâmetros, ambiente, saídas, versões_saídas, logs, evidências\}}$$

em que:

- **entradas** representam dados, fontes e artefatos consumidos;
- **processos** representam scripts, workflows, notebooks, consultas ou regras;
- **parâmetros** representam configurações usadas na execução;
- **ambiente** representa dependências computacionais;
- **saídas** representam datasets, indicadores, modelos, dashboards ou relatórios gerados;
- **evidências** representam logs, checksums, manifestos, hashes e registros de execução.

A condição mínima de reprodutibilidade pode ser expressa como:

R(E) = {dados_versionados, código_versionado, parâmetros_registrados, ambiente_documentado, saídas_identificadas, logs_preservados}

Uma execução é considerada reprodutível quando todos esses elementos estão disponíveis em grau suficiente para sua reconstrução.

7. Entradas do módulo

As entradas do Módulo 12 são produzidas por todos os módulos do modelo de referência.

As principais entradas são:

- inventários de fontes;
- dicionários de dados;
- mapas de bases;
- classificações de sensibilidade;
- snapshots de dados brutos;
- datasets brutos;
- configurações de ingestão;
- manifestos de ingestão;
- logs de execução;
- tabelas de integração;
- regras de harmonização;
- datasets integrados;
- registros de conflitos;
- regras de transformação;
- scripts;
- notebooks;
- consultas SQL;
- workflows;
- parâmetros de execução;
- datasets tratados;
- atributos derivados;
- regras de qualidade;
- relatórios de validação;
- datasets balanceados ou sinalizados;
- data marts;
- views;
- schemas;
- dashboards;
- APIs;
- relatórios;
- produtos de dados;
- metadados;
- políticas de governança;
- registros de lineage;
- ambientes computacionais;
- bibliotecas;
- containers;
- seeds de aleatoriedade;
- arquivos de configuração;
- hashes e checksums.

8. Processamento

O processamento do Módulo 12 envolve atividades contínuas de identificação, registro, preservação, comparação, reconstrução e validação de versões e execuções.

8.1 Identificação dos ativos versionáveis

A primeira atividade consiste em identificar quais ativos precisam ser versionados. Nem todos os artefatos possuem o mesmo nível de criticidade, mas todo ativo que impacta resultados analíticos, decisões, indicadores, modelos ou produtos deve possuir algum mecanismo de controle de versão.

Ativos críticos incluem fontes, datasets, regras, scripts, parâmetros, schemas, data marts, dashboards, APIs e modelos.

8.2 Definição da estratégia de versionamento

A segunda atividade consiste em definir como cada tipo de ativo será versionado. Dados podem ser versionados por snapshot, data, hash, tag ou mecanismo transacional. Código pode ser versionado em repositórios Git. Ambientes podem ser versionados por containers ou arquivos de dependências. Produtos podem ser versionados por publicação, data ou release.

8.3 Registro de versões dos dados

A terceira atividade consiste em registrar versões de datasets, tabelas, arquivos, snapshots e data marts. Cada versão deve indicar origem, data, schema, volume, checksum, responsável, status, relação com versões anteriores e motivo da criação.

8.4 Registro de versões de código, regras e parâmetros

A quarta atividade consiste em registrar scripts, notebooks, queries, workflows, regras de transformação, regras de qualidade, regras de cálculo e parâmetros utilizados nas execuções.

Essa etapa evita que resultados dependam de código local, scripts alterados sem histórico ou notebooks executados manualmente sem registro.

8.5 Registro do ambiente computacional

A quinta atividade consiste em documentar o ambiente utilizado na execução. Isso inclui versões de linguagem, bibliotecas, ferramentas, sistemas operacionais, containers, variáveis de ambiente, dependências e configurações relevantes.

Quando possível, o ambiente deve ser encapsulado em container, arquivo de dependências ou configuração declarativa.

8.6 Registro de execuções reprodutíveis

A sexta atividade consiste em gerar manifestos de execução que associem entradas, versões, processos, parâmetros, ambiente e saídas. Esse manifesto é o núcleo da reprodutibilidade.

Ele deve permitir identificar exatamente quais elementos foram utilizados para produzir determinado resultado.

8.7 Comparação entre versões

A sétima atividade consiste em comparar versões de dados, regras, scripts ou produtos para identificar mudanças. Essa comparação pode envolver diferenças de schema, volume, distribuição, qualidade, atributos, registros, métricas, desempenho ou resultados.

8.8 Reexecução controlada

A oitava atividade consiste em reexecutar pipelines, processos ou análises a partir de versões específicas. Essa reexecução pode ser usada para auditoria, validação, correção, atualização ou comparação científica.

8.9 Validação da reprodutibilidade

A nona atividade consiste em avaliar se a reexecução produziu resultados iguais, equivalentes ou justificadamente diferentes. Diferenças devem ser explicadas por mudanças em dados, ambiente, parâmetros, aleatoriedade, atualização de bibliotecas ou alterações de regras.

8.10 Preservação e arquivamento

A décima atividade consiste em definir quais versões devem ser preservadas, arquivadas ou descartadas. Versões associadas a publicações, decisões, relatórios oficiais, modelos implantados ou produtos disponibilizados devem ser preservadas de forma controlada.

9. Técnicas, métodos e tecnologias aplicáveis

O Módulo 12 pode empregar diferentes técnicas, métodos e tecnologias para versionamento e reprodutibilidade.

Entre as técnicas de versionamento, destacam-se:

- versionamento semântico;
- versionamento por data;
- versionamento por snapshot;
- versionamento por hash;
- versionamento por release;
- tags de versão;
- branches;
- controle de alterações;
- comparação entre versões;
- registro de changelog;
- versionamento de schema;
- versionamento de datasets;
- versionamento de modelos;
- versionamento de dashboards;
- versionamento de APIs.

Entre as técnicas de reprodutibilidade, podem ser utilizadas:

- manifestos de execução;
- arquivos de configuração;
- pipelines declarativos;
- parametrização;
- controle de seeds;
- notebooks versionados;
- containers;
- ambientes virtuais;
- gestão de dependências;
- workflows automatizados;
- logs estruturados;
- checksums;
- hashes;
- reconstrução de ambiente;
- data lineage;
- data contracts;
- testes de reexecução;
- comparação de resultados.

Entre as tecnologias possíveis, destacam-se:

- Git;
- GitLab;
- GitHub;
- DVC;
- LakeFS;
- Delta Lake;
- Apache Iceberg;
- Apache Hudi;
- MLflow;
- Airflow;
- Prefect;
- Dagster;
- Docker;
- Conda;
- Poetry;
- pip-tools;
- Makefiles;
- dbt;
- notebooks versionados;
- repositórios de artefatos;
- object storage;
- bancos com snapshots;
- sistemas de logs;
- catálogos de dados;
- ferramentas de lineage.

10. Métricas do módulo

As métricas do Módulo 12 avaliam cobertura, completude, estabilidade, reconstrução, comparação e confiabilidade do versionamento e da reprodutibilidade.

Métrica	Sigla	Descrição

Taxa de ativos versionados	TAV	Proporção de ativos críticos com versão registrada
Taxa de datasets versionados	TDV	Proporção de datasets com snapshot, hash ou versão
Taxa de código versionado	TCV	Proporção de scripts, notebooks e workflows sob controle de versão
Taxa de regras versionadas	TRV	Proporção de regras de transformação, qualidade ou cálculo versionadas
Taxa de parâmetros registrados	TPR	Proporção de execuções com parâmetros documentados
Taxa de ambientes documentados	TAD	Proporção de execuções com ambiente registrado
Taxa de execuções com manifesto	TEM	Proporção de execuções com manifesto completo
Taxa de execuções reprodutíveis	TER	Proporção de execuções que podem ser reconstruídas
Taxa de reexecuções bem-sucedidas	TRBS	Proporção de reexecuções concluídas sem erro
Divergência entre execuções	DEE	Diferença entre resultados original e reproduzido
Tempo médio de reprodução	TMR	Tempo necessário para reconstruir execução ou ativo
Taxa de versões com changelog	TVC	Proporção de versões com descrição de mudanças
Taxa de schemas versionados	TSV	Proporção de schemas com histórico de alterações
Taxa de produtos versionados	TPV	Proporção de dashboards, APIs, relatórios e produtos com versão
Índice de Reprodutibilidade do Ativo	IRA	Grau em que um ativo pode ser reconstruído

10.1 Taxa de Ativos Versionados — TAV

A **Taxa de Ativos Versionados** mede a proporção de ativos críticos que possuem versão registrada.

TAV = número de ativos críticos versionados / número total de ativos críticos

Essa métrica avalia a cobertura geral do versionamento.

10.2 Taxa de Execuções com Manifesto — TEM

A **Taxa de Execuções com Manifesto** mede a proporção de execuções que possuem manifesto contendo entradas, versões, processos, parâmetros, ambiente e saídas.

TEM = número de execuções com manifesto / número total de execuções avaliadas

Essa métrica é essencial para reprodutibilidade.

10.3 Taxa de Execuções Reprodutíveis — TER

A **Taxa de Execuções Reprodutíveis** mede a proporção de execuções que podem ser reconstruídas com base nas informações preservadas.

TER = número de execuções reprodutíveis / número total de execuções avaliadas

Uma execução é considerada reprodutível quando possui dados versionados, código versionado, parâmetros registrados, ambiente documentado, saídas identificadas e logs preservados.

10.4 Divergência entre Execuções — DEE

A **Divergência entre Execuções** mede a diferença entre o resultado original e o resultado reproduzido.

DEE = distância(resultado original, resultado reproduzido)

Essa distância pode ser calculada por diferença absoluta, diferença percentual, comparação de checksums, divergência de distribuições, diferença de métricas ou validação semântica do resultado.

10.5 Índice de Reprodutibilidade do Ativo — IRA

O **Índice de Reprodutibilidade do Ativo** pode ser definido como a proporção de elementos obrigatórios de reprodutibilidade disponíveis para um ativo.

$IRA(a_i) = EDP(a_i) / EDO$

em que:

- **$EDP(a_i)$** representa o número de elementos de dependência preenchidos para o ativo **a_i**;
- **EDO** representa o total de elementos obrigatórios.

Elementos obrigatórios podem incluir dados de entrada, versão dos dados, código, regra, parâmetros, ambiente, logs, manifesto, saídas e evidências.

11. Artefatos formalizados do módulo

Os artefatos do Módulo 12 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, regras de construção, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A12.1	Política de Versionamento	Documento normativo	TAV	Define critérios de versionamento
A12.2	Registro de Versões de Dados	Tabela/manifesto	TDV	Controla versões de datasets
A12.3	Registro de Versões de Código e Regras	Repositório/tabela	TCV, TRV	Controla scripts, regras e workflows
A12.4	Manifesto de Execução Reprodutível	Documento estruturado	TEM, TER	Associa entradas, processos, parâmetros e saídas
A12.5	Registro de Ambiente Computacional	Arquivo/ configuração	TAD	Documenta dependências e ambiente
A12.6	Registro de Parâmetros de Execução	Tabela/configuração	TPR	Preserva parâmetros usados
A12.7	Changelog de Ativos Analíticos	Documento/tabela	TVC	Registra alterações entre versões
A12.8	Relatório de Comparação entre Versões	Relatório técnico	DEE	Avalia diferenças entre versões
A12.9	Pacote de Reprodutibilidade	Repositório/pacote	IRA	Consolida dados, código, parâmetros e ambiente
A12.10	Relatório de Reprodutibilidade	Relatório técnico	TER, TRBS	Avalia capacidade de reprodução

12. Artefato A12.1 — Política de Versionamento

12.1 Finalidade

A **Política de Versionamento** define quais ativos devem ser versionados, como as versões serão identificadas, quando uma nova versão deve ser criada e quais informações mínimas devem ser preservadas.

12.2 Tipo de artefato na DSR

Esse artefato é um **modelo normativo** e um **método**, pois orienta práticas de versionamento ao longo do modelo de referência.

12.3 Estrutura mínima

Campo	Descrição
Escopo	Tipos de ativos cobertos
Critérios de versionamento	Condições para criar nova versão
Padrão de identificação	Semântico, data, hash, snapshot etc.
Responsável	Quem cria e aprova versões
Regras de preservação	Quais versões devem ser mantidas
Regras de descarte	Quando versões podem ser removidas
Evidências necessárias	Logs, manifestos, checksums, changelog
Integração com lineage	Relação com rastreabilidade
Critérios de conformidade	Condições mínimas de aderência

12.4 Métricas associadas

- taxa de ativos versionados;
- taxa de versões com changelog;
- taxa de produtos versionados;
- taxa de conformidade com política.

12.5 Critérios de validação

A política é considerada válida quando define escopo, padrão de versão, critérios de criação, preservação, descarte, responsáveis e evidências mínimas.

12.6 Contribuição

Esse artefato padroniza o versionamento e reduz inconsistências na gestão de versões dos ativos analíticos.

13. Artefato A12.2 — Registro de Versões de Dados

13.1 Finalidade

O **Registro de Versões de Dados** documenta versões de datasets, tabelas, arquivos, snapshots, data marts e produtos derivados.

13.2 Estrutura mínima

Campo	Descrição

Ativo de dados	Dataset, tabela, arquivo ou data mart
Versão	Identificador da versão
Data de criação	Momento da versão
Fonte de origem	Fonte ou dataset anterior
Snapshot	Referência ao estado preservado
Schema	Estrutura associada
Volume	Número de registros ou tamanho
Checksum/hash	Evidência de integridade
Responsável	Pessoa ou sistema
Status	Ativa, substituída, arquivada, descartada
Observações	Mudanças ou limitações

13.3 Métricas associadas

- taxa de datasets versionados;
- percentual de versões com checksum;
- percentual de versões com schema;
- número médio de versões por ativo;
- taxa de versões arquivadas.

13.4 Contribuição

Esse artefato garante controle sobre os estados dos dados utilizados ou produzidos pelo pipeline.

14. Artefato A12.3 — Registro de Versões de Código e Regras

14.1 Finalidade

O **Registro de Versões de Código e Regras** documenta scripts, notebooks, consultas SQL, workflows, regras de transformação, regras de qualidade, regras de cálculo e configurações associadas aos processos analíticos.

14.2 Estrutura mínima

Campo	Descrição
Artefato	Script, notebook, query, regra ou workflow

Versão	Identificador da versão
Repositório	Local onde está armazenado
Commit/hash	Identificador de controle
Data	Data da versão
Autor	Responsável pela alteração
Descrição da mudança	O que foi alterado
Ativos afetados	Datasets ou produtos impactados
Status	Ativo, depreciado, substituído
Observações	Dependências ou cuidados

14.3 Métricas associadas

- taxa de código versionado;
- taxa de regras versionadas;
- taxa de commits associados a execução;
- taxa de artefatos sem versão;
- taxa de regras depreciadas.

14.4 Contribuição

Esse artefato evita que resultados dependam de scripts não controlados ou regras alteradas sem histórico.

15. Artefato A12.4 — Manifesto de Execução Reprodutível

15.1 Finalidade

O **Manifesto de Execução Reprodutível** registra todos os elementos necessários para reconstruir uma execução do pipeline.

15.2 Estrutura formal

Uma execução pode ser descrita por:

E = {id, entradas, versões_entradas, processos, parâmetros, ambiente, saídas, logs, evidências}

15.3 Estrutura mínima

Campo	Descrição

Identificador da execução	Código único
Objetivo da execução	Finalidade
Dados de entrada	Datasets ou fontes utilizadas
Versões de entrada	Versões dos dados
Código ou workflow	Script, notebook ou pipeline
Versão do código	Commit, tag ou hash
Regras aplicadas	Transformações, validações ou cálculos
Parâmetros	Configurações utilizadas
Ambiente	Dependências e plataforma
Dados de saída	Ativos produzidos
Logs	Registros da execução
Evidências	Checksums, relatórios, manifestos
Responsável	Executor humano ou sistema
Data	Momento da execução

15.4 Métricas associadas

- taxa de execuções com manifesto;
- taxa de execuções reprodutíveis;
- índice de reprodutibilidade do ativo;
- taxa de parâmetros registrados.

15.5 Contribuição

Esse artefato é o núcleo da reprodutibilidade, pois associa dados, código, ambiente, parâmetros e saídas.

16. Artefato A12.5 — Registro de Ambiente Computacional

16.1 Finalidade

O **Registro de Ambiente Computacional** documenta as condições técnicas necessárias para executar ou reproduzir pipelines e análises.

16.2 Estrutura mínima

Campo	Descrição
Execução ou ativo	Processo associado
Linguagem	Python, R, SQL, Java etc.
Versão da linguagem	Versão utilizada
Bibliotecas	Dependências
Versões das bibliotecas	Versões específicas
Sistema operacional	Ambiente base
Container	Imagem Docker ou similar
Variáveis de ambiente	Configurações relevantes
Hardware ou cluster	Quando aplicável
Observações	Restrições ou dependências externas

16.3 Métricas associadas

- taxa de ambientes documentados;
- percentual de execuções com container;
- percentual de execuções com dependências versionadas;
- taxa de falhas por ambiente divergente.

16.4 Contribuição

Esse artefato reduz problemas de reexecução causados por diferenças de ambiente computacional.

17. Artefato A12.6 — Registro de Parâmetros de Execução

17.1 Finalidade

O **Registro de Parâmetros de Execução** preserva os parâmetros utilizados em processos de extração, integração, transformação, validação, balanceamento, modelagem, publicação ou monitoramento.

17.2 Estrutura mínima

Campo	Descrição
Processo	Pipeline, script ou regra

Execução	Identificador da execução
Parâmetro	Nome do parâmetro
Valor	Valor utilizado
Tipo	Texto, número, booleano, data etc.
Finalidade	Função do parâmetro
Valor padrão	Quando aplicável
Alterado por	Responsável
Data	Momento da execução

17.3 Métricas associadas

- taxa de parâmetros registrados;
- percentual de execuções parametrizadas;
- percentual de parâmetros com valor padrão;
- taxa de divergência por mudança de parâmetro.

17.4 Contribuição

Esse artefato evita que resultados dependam de configurações implícitas ou não documentadas.

18. Artefato A12.7 — Changelog de Ativos Analíticos

18.1 Finalidade

O **Changelog de Ativos Analíticos** registra alterações realizadas entre versões de dados, scripts, regras, schemas, dashboards, APIs, modelos ou produtos.

18.2 Estrutura mínima

Campo	Descrição
Ativo	Ativo alterado
Versão anterior	Identificador da versão anterior
Nova versão	Identificador da nova versão
Tipo de mudança	Schema, regra, conteúdo, parâmetro, ambiente, produto
Descrição da	Explicação da alteração

mudança	
Motivo	Correção, atualização, melhoria, ajuste metodológico
Impacto esperado	Ativos ou resultados afetados
Responsável	Autor da mudança
Data	Momento da alteração

18.3 Métricas associadas

- taxa de versões com changelog;
- número de mudanças por ativo;
- percentual de mudanças críticas documentadas;
- taxa de mudanças sem justificativa.

18.4 Contribuição

Esse artefato permite compreender por que versões diferem entre si e apoia auditoria de mudanças.

19. Artefato A12.8 — Relatório de Comparação entre Versões

19.1 Finalidade

O **Relatório de Comparação entre Versões** avalia diferenças entre duas ou mais versões de dados, regras, produtos ou resultados.

19.2 Estrutura mínima

Campo	Descrição
Ativo comparado	Dataset, regra, dashboard, API etc.
Versão A	Primeira versão
Versão B	Segunda versão
Diferença de schema	Alterações estruturais
Diferença de volume	Registros ou tamanho
Diferença de distribuição	Mudanças estatísticas
Diferença de qualidade	Alteração de métricas
Diferença de resultado	Indicadores ou saídas afetadas

Severidade	Baixa, média, alta ou crítica
Recomendações	Ações necessárias

19.3 Métricas associadas

- divergência entre execuções;
- variação de volume;
- variação de schema;
- variação de qualidade;
- variação de métricas analíticas.

19.4 Contribuição

Esse artefato permite avaliar o impacto de mudanças e justificar diferenças entre resultados.

20. Artefato A12.9 — Pacote de Reprodutibilidade

20.1 Finalidade

O **Pacote de Reprodutibilidade** consolida os elementos necessários para reproduzir um ativo, experimento, relatório, indicador, dashboard ou produto de dados.

20.2 Estrutura mínima

Elemento	Descrição
Dados de entrada	Datasets ou referências versionadas
Código	Scripts, notebooks, queries ou workflows
Parâmetros	Configurações utilizadas
Ambiente	Dependências, container ou arquivo de ambiente
Manifesto	Registro da execução
Saídas esperadas	Datasets, métricas, relatórios ou produtos
Instruções	Como reproduzir
Evidências	Logs, checksums, hashes ou relatórios

20.3 Métricas associadas

- índice de reprodutibilidade do ativo;
- taxa de execuções reprodutíveis;
- taxa de pacotes completos;

- tempo médio de reprodução.

20.4 Contribuição

Esse artefato viabiliza auditoria, validação científica, transferência tecnológica e reuso metodológico.

21. Artefato A12.10 — Relatório de Reprodutibilidade

21.1 Finalidade

O **Relatório de Reprodutibilidade** avalia se um ativo, execução ou produto pode ser reproduzido de forma confiável.

21.2 Estrutura mínima

Campo	Descrição
Ativo avaliado	Dataset, indicador, dashboard, modelo ou produto
Execução original	Referência da execução
Reexecução	Referência da reprodução
Dados utilizados	Versões de entrada
Código utilizado	Versão do código
Ambiente	Ambiente usado
Parâmetros	Configurações
Resultado original	Saída esperada
Resultado reproduzido	Saída obtida
Divergência	Diferença observada
Status	Reproduzido, reproduzido com divergência, não reproduzido
Justificativa	Explicação das diferenças

21.3 Métricas associadas

- taxa de reexecuções bem-sucedidas;
- divergência entre execuções;
- tempo médio de reprodução;
- taxa de execuções reprodutíveis;
- índice de reprodutibilidade.

21.4 Contribuição

Esse artefato permite avaliar empiricamente a capacidade de reprodução dos resultados produzidos pelo modelo de referência.

22. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Dados, scripts, notebooks, regras, parâmetros, ambientes, schemas, dashboards, APIs, produtos, logs, manifestos, versões, snapshots e registros de lineage.
Processamento	Identificação dos ativos versionáveis, definição da estratégia de versionamento, registro de versões, preservação de ambientes, criação de manifestos, comparação entre versões, reexecução e validação da reprodutibilidade.
Artefatos	Política de versionamento, registros de versões de dados, código e regras, manifesto de execução, ambiente computacional, parâmetros, changelog, relatório de comparação, pacote e relatório de reprodutibilidade.
Métricas	TAV, TDV, TCV, TRV, TPR, TAD, TEM, TER, TRBS, DEE, TMR, TVC, TSV, TPV e IRA.
Saídas	Ativos versionados, execuções documentadas, ambientes preservados, resultados reprodutíveis, pacotes de reprodutibilidade e evidências para auditoria.

23. Critérios gerais de validação do módulo

O Módulo 12 pode ser considerado adequadamente implementado quando:

1. os ativos críticos possuem versão registrada;
2. os datasets utilizados em produtos ou análises possuem snapshot ou referência versionada;
3. scripts, notebooks, workflows e queries críticos estão sob controle de versão;
4. regras de transformação, qualidade, cálculo e mitigação possuem versão;
5. parâmetros de execução estão registrados;
6. ambientes computacionais estão documentados ou encapsulados;
7. execuções críticas possuem manifesto reprodutível;
8. produtos publicados possuem versão e changelog;
9. reexecuções podem ser realizadas a partir dos registros preservados;
10. divergências entre execução original e reexecução podem ser identificadas e explicadas;
11. versões associadas a decisões, relatórios, dashboards ou modelos publicados são preservadas;
12. o índice de reprodutibilidade dos ativos críticos atinge patamar satisfatório.

24. Relação com o pipeline principal

O Módulo 12 se relaciona com todos os módulos do pipeline principal do UnespDataLens-RM.

Com o **Módulo 1 — Fontes e Inventário de Dados**, versiona o inventário, dicionários preliminares e classificações iniciais.

Com o **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, versiona snapshots, dados brutos, configurações de extração, logs e manifestos.

Com o **Módulo 3 — Integração e Harmonização**, versiona regras de integração, chaves, tabelas de mapeamento e datasets integrados.

Com o **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, versiona scripts, regras, parâmetros, atributos derivados e datasets tratados.

Com o **Módulo 5 — Qualidade e Validação de Dados**, versiona regras de qualidade, limiares, relatórios e status de validação.

Com o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, versiona planos de mitigação, técnicas aplicadas, parâmetros, métricas e datasets balanceados ou sinalizados.

Com o **Módulo 7 — Armazenamento Analítico**, versiona data marts, schemas, views, tabelas, arquivos e estruturas persistidas.

Com o **Módulo 8 — Disponibilização e Consumo Analítico**, versiona dashboards, APIs, relatórios, produtos de dados e registros de publicação.

25. Relação com os demais módulos transversais

O Módulo 12 depende do **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, pois o catálogo registra identificadores, descrições, versões, status e relações entre ativos.

Relaciona-se com o **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, pois define quais versões devem ser preservadas, protegidas, descartadas, arquivadas ou restringidas conforme políticas institucionais.

Relaciona-se com o **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, pois a reprodutibilidade depende da capacidade de reconstruir a trajetória dos ativos, identificar dependências e localizar versões corretas.

Relaciona-se com o módulo de **Monitoramento, Drift e Evolução dos Ativos**, pois versões sucessivas permitem comparar mudanças de schema, qualidade, distribuição, desempenho e consumo.

Relaciona-se com o módulo de **Agentes Inteligentes de Apoio ao Pipeline**, pois agentes podem consultar versões, explicar diferenças entre resultados, sugerir reprocessamentos e apoiar reconstrução de execuções.

26. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 12 poderia controlar versões de uma base analítica de evasão.

A base original de matrícula poderia possuir snapshots por semestre. As regras de definição de evasão poderiam ter versões distintas, por exemplo, uma regra baseada em ausência de matrícula no semestre seguinte e outra baseada em cancelamento formal. Os scripts de integração entre matrícula, histórico, notas, frequência e AVA seriam versionados em repositório. O ambiente computacional seria registrado com versões de bibliotecas, e os parâmetros de filtros por período, curso e situação acadêmica seriam preservados.

Se um dashboard de evasão publicado em determinado mês apresentasse uma taxa específica, o módulo permitiria reconstruir quais dados, regras, scripts, versões e parâmetros produziram aquele resultado.

Caso a regra de evasão fosse alterada posteriormente, o changelog explicaria a mudança, e o relatório de comparação entre versões indicaria como os percentuais foram afetados.

27. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, o Módulo 12 poderia versionar indicadores territoriais de vulnerabilidade.

Cada versão das bases de saúde, educação, assistência social e território seria preservada por data de extração. As regras de cálculo do indicador seriam versionadas, assim como os pesos, agregações, filtros e recortes territoriais. Os dashboards publicados teriam versões associadas aos datasets e parâmetros utilizados.

Se uma secretaria corrigisse dados de atendimento de meses anteriores, o módulo permitiria identificar quais indicadores precisam ser recalculados, comparar versões anteriores e novas, justificar diferenças e preservar o histórico de publicações.

28. Contribuição técnico-científica do módulo

A contribuição do Módulo 12 está em formalizar versionamento e reprodutibilidade como dimensões transversais da Engenharia Analítica de Dados.

Em vez de tratar reprodutibilidade como preocupação posterior ou restrita ao código-fonte, o UnespDataLens-RM propõe que dados, regras, parâmetros, ambientes, scripts, produtos e resultados sejam controlados de forma integrada.

Essa formalização contribui para:

- preservar histórico dos ativos analíticos;
- permitir reconstrução de resultados;
- reduzir dependência de conhecimento tácito;
- aumentar confiança em indicadores, dashboards e modelos;
- apoiar auditoria e validação científica;
- explicar diferenças entre versões;
- controlar mudanças em regras e datasets;
- fortalecer lineage e governança;
- facilitar transferência tecnológica;
- sustentar evolução controlada do ecossistema analítico.

29. Síntese do módulo

O **Módulo Transversal 12 — Versionamento e Reprodutibilidade** garante que os ativos e resultados produzidos pelo UnespDataLens-RM possam ser preservados, comparados, auditados e reconstruídos. Ele controla versões de dados, códigos, regras, parâmetros, ambientes, produtos e resultados, criando condições para reexecutar pipelines e validar resultados ao longo do tempo.

Sua atuação é transversal porque qualquer etapa do modelo de referência pode gerar artefatos versionáveis e resultados que precisam ser reproduzidos. Ao formalizar políticas, registros, manifestos, ambientes, changelogs, pacotes e relatórios de reprodutibilidade, o módulo fortalece a confiabilidade, a transparência, a auditoria e o rigor científico dos processos analíticos.

Dessa forma, o Módulo 12 transforma o UnespDataLens-RM em um modelo de referência não apenas reescrever o modulo transversal 12 - versionamento e reprodutibilidade capaz de produzir dados e produtos analíticos, mas também capaz de preservar as condições necessárias para explicar, repetir e validar aquilo que foi produzido.

Módulo Transversal 13 — Representações Analíticas: Tabular, Temporal, Espacial, Textual, Semântica e Grafos

1. Finalidade do módulo

O **Módulo Transversal 13 — Representações Analíticas: Tabular, Temporal, Espacial, Textual, Semântica e Grafos** tem como finalidade especificar, construir, documentar, validar e disponibilizar diferentes formas de representação dos dados produzidos pelo modelo de referência **UnespDataLens-RM**, de modo que esses dados possam ser utilizados de maneira adequada em análises exploratórias, indicadores, dashboards, modelos preditivos, análises temporais, grafos, grafos temporais, sistemas de recomendação, agentes inteligentes e processos de tomada de decisão baseada em evidências.

No contexto do UnespDataLens-RM, os dados não devem ser compreendidos apenas como tabelas estáticas. Um mesmo fenômeno pode ser representado sob diferentes perspectivas analíticas: como uma matriz tabular de atributos, como uma série temporal, como uma trajetória longitudinal, como uma sequência de eventos, como um grafo de relações entre entidades, como um grafo heterogêneo, como um grafo bipartido ou como um grafo temporal que evolui ao longo do tempo.

Essa multiplicidade de representações é fundamental porque diferentes problemas exigem diferentes formas de modelagem dos dados. Uma análise descritiva pode ser suficiente com uma representação tabular. Uma análise de evolução exige representação temporal. Uma análise de trajetória requer dados longitudinais. Uma análise de relações entre entidades exige grafos. Uma análise de interações dinâmicas exige grafos temporais ou redes evolutivas.

Assim, o Módulo 13 formaliza a etapa em que os ativos analíticos produzidos pelos módulos anteriores são reorganizados em estruturas computacionais apropriadas ao uso analítico pretendido. Ele atua de forma transversal porque depende dos dados extraídos, integrados, tratados, validados, governados, versionados e rastreáveis, mas também influencia diretamente a criação de features, a modelagem preditiva, os produtos de consumo, o monitoramento e a explicabilidade dos resultados.

2. Objetivo geral

O objetivo geral do Módulo 13 é **definir e operacionalizar representações analíticas tabulares, temporais, longitudinais, baseadas em eventos, relacionais e em grafos**, garantindo que cada representação seja construída de forma documentada, rastreável, versionada, semanticamente coerente, adequada ao problema analítico e compatível com os módulos de feature engineering, modelagem, monitoramento, governança e consumo analítico do UnespDataLens-RM.

3. Objetivos específicos

São objetivos específicos do Módulo 13:

1. definir a unidade de análise adequada para cada representação, como estudante, matrícula, curso, disciplina, território, atendimento, evento, instituição, período ou relação;
2. especificar representações tabulares para análises exploratórias, indicadores, dashboards e modelos tradicionais de aprendizagem de máquina;
3. especificar representações temporais para séries históricas, evolução de indicadores, dados longitudinais, trajetórias e janelas de observação;
4. especificar representações baseadas em eventos para logs, interações, registros administrativos, ações de usuários, sensores ou fluxos operacionais;
5. especificar representações em grafos para modelar relações entre entidades, atributos, eventos, pessoas, organizações, territórios, disciplinas, serviços ou processos;
6. especificar grafos temporais para representar relações que surgem, desaparecem, se fortalecem ou se modificam ao longo do tempo;
7. definir regras de construção das representações, incluindo seleção de atributos, granularidade, janelas temporais, relações, pesos, direções e tipos de arestas;
8. documentar metadados, origem, regras, versões, qualidade, limitações e finalidade de cada representação analítica;
9. avaliar a adequação da representação ao problema analítico, ao tipo de modelo, ao indicador ou ao produto de dados;
10. preservar a rastreabilidade entre dados originais, transformações, representações derivadas, features, modelos e produtos analíticos;
11. apoiar a construção de features tabulares, temporais, relacionais e baseadas em grafos;
12. permitir comparação entre diferentes paradigmas de representação de dados;
13. viabilizar análises de evolução, estabilidade, recorrência, transição, centralidade, comunidade, conectividade e dinâmica relacional;
14. fornecer estruturas adequadas para modelos preditivos, modelos explicáveis, algoritmos de grafos, modelos temporais, agentes inteligentes e visualizações analíticas.

4. Fundamentação conceitual do módulo

A escolha da representação analítica é uma decisão metodológica central em projetos de Engenharia Analítica de Dados. Dados provenientes de fontes heterogêneas podem ser organizados de formas distintas, e cada forma de representação evidencia aspectos diferentes do fenômeno analisado.

Uma representação **tabular** organiza os dados em linhas e colunas, geralmente com uma unidade de análise por linha e atributos descritivos por coluna. Essa estrutura é adequada para análises exploratórias, estatísticas descritivas, dashboards, modelos preditivos clássicos, regressão, classificação, agrupamento e construção de indicadores.

Uma representação **temporal** incorpora explicitamente a dimensão do tempo, permitindo analisar evolução, tendência, sazonalidade, recorrência, mudança, atraso, persistência, progressão ou ruptura. Ela pode assumir a forma de séries temporais, painéis longitudinais, sequências de eventos ou janelas temporais.

Uma representação **longitudinal** acompanha uma mesma entidade ao longo do tempo. No domínio educacional, por exemplo, pode representar a trajetória de um estudante por períodos letivos. Em políticas públicas, pode representar o histórico de atendimento de uma família ou território.

Uma representação **baseada em eventos** organiza registros como acontecimentos situados no tempo, associados a uma entidade, ação, contexto e timestamp. Essa representação é especialmente relevante para logs de sistemas, interações em ambientes virtuais, atendimentos, notificações, sensores, transações e fluxos operacionais.

Uma representação **em grafos** modela entidades como nós e relações como arestas. Essa estrutura é adequada quando o fenômeno depende de relações, conexões, dependências, interações, proximidades, fluxos ou estruturas de rede.

Uma representação **em grafos temporais** amplia o grafo tradicional ao incorporar o tempo, permitindo analisar como nós, arestas, pesos, comunidades, centralidades e relações evoluem. Essa representação é importante para estudar redes dinâmicas, interações educacionais, trajetórias institucionais, redes de atendimento, mobilidade, vínculos, fluxos e mudanças estruturais.

No UnespDataLens-RM, essas representações não competem entre si. Elas são complementares. Um mesmo conjunto de dados pode ser representado simultaneamente como:

- uma tabela de atributos;
- uma série temporal de indicadores;
- uma trajetória longitudinal de entidades;
- uma sequência de eventos;
- um grafo de relações;
- um grafo temporal;
- uma matriz de features para modelagem.

O papel do Módulo 13 é formalizar essa escolha, registrar sua justificativa e garantir que cada representação seja adequada, compreensível, rastreável, versionada e reutilizável.

5. Tipos de representações analíticas

O UnespDataLens-RM contempla diferentes tipos de representações analíticas.

Tipo de representação	Descrição	Exemplo
Representação tabular	Organiza dados em linhas e colunas, com unidade de análise e atributos.	Uma linha por estudante e colunas com idade, curso, frequência e desempenho.
Representação dimensional	Organiza fatos, dimensões e métricas para análise gerencial.	Fato matrícula, dimensão curso, dimensão tempo e métrica evasão.
Representação temporal	Organiza observações associadas a períodos ou instantes.	Taxa mensal de evasão por curso.
Representação longitudinal	Acompanha a mesma entidade em diferentes períodos.	Histórico acadêmico do estudante por semestre.
Representação baseada em eventos	Organiza ações ou ocorrências com timestamp.	Logs de acesso ao ambiente virtual.
Representação sequencial	Organiza eventos em ordem temporal.	Sequência de disciplinas cursadas por estudante.
Representação em janelas temporais	Agrega dados por intervalos de tempo definidos.	Features calculadas nos últimos 30, 60 ou 90 dias.
Representação em grafo homogêneo	Nós e arestas de um único tipo principal.	Rede de estudantes conectados por interação.
Representação em grafo heterogêneo	Nós e arestas de múltiplos tipos.	Estudantes, cursos, disciplinas, atividades e interações.
Representação bipartida	Relações entre dois tipos de entidades.	Estudantes conectados a disciplinas.
Representação ponderada	Arestas possuem pesos.	Número de interações entre estudante e atividade.
Representação direcionada	Relações possuem direção.	Fluxo de atendimento entre serviços.
Representação temporal em grafo	Relações variam ao longo do tempo.	Interações entre estudantes por semana.
Representação multimodal	Combina diferentes tipos de dados.	Dados tabulares, textos, logs e grafos no mesmo projeto.

6. Formalização sintética do módulo

O Módulo 13 pode ser formalizado a partir de um conjunto de ativos analíticos já produzidos pelos módulos anteriores.

Considerando o conjunto de ativos analíticos:

$$\mathbf{A = \{a_1, a_2, ..., a_n\}}$$

e o conjunto de representações possíveis:

$$\mathbf{R = \{r_1, r_2, ..., r_k\}}$$

cada representação $\mathbf{r_i}$ pode ser descrita por:

$$\mathbf{r_i = \{id, tipo, ativo_origem, unidade_análise, entidades, atributos, relações, tempo, granularidade, regras_construção, métricas, versão, finalidade, limitações, lineage\}}$$

Para uma representação tabular, pode-se definir:

$$\mathbf{T = (U, X, y)}$$

em que:

- **U** representa o conjunto de unidades de análise;
- **X** representa a matriz de atributos;
- **y** representa a variável-alvo, rótulo, indicador ou desfecho, quando aplicável.

Para uma representação temporal, pode-se definir:

$$\mathbf{S = \{(u, t, x_t)\}}$$

em que:

- **u** representa a unidade observada;
- **t** representa o instante, período ou janela temporal;
- $\mathbf{x_t}$ representa os atributos observados naquele tempo.

Para uma representação baseada em eventos, pode-se definir:

$$\mathbf{E = \{e_1, e_2, ..., e_m\}}$$

em que cada evento $\mathbf{e_i}$ pode ser descrito por:

$$\mathbf{e_i = \{id, entidade, tipo_evento, timestamp, contexto, atributos, fonte, versão\}}$$

Para uma representação em grafo, pode-se definir:

$$\mathbf{G = (V, E, X_v, X_e)}$$

em que:

- **V** representa o conjunto de nós;
- **E** representa o conjunto de arestas;
- $\mathbf{X_v}$ representa atributos dos nós;
- $\mathbf{X_e}$ representa atributos das arestas.

Para uma representação em grafo temporal, pode-se definir:

$G_t = (V_t, E_t, X_{vt}, X_{et}, T)$

em que:

- V_t representa os nós observados no tempo;
- E_t representa as arestas observadas no tempo;
- X_{vt} representa atributos temporais dos nós;
- X_{et} representa atributos temporais das arestas;
- **T** representa o conjunto de instantes, períodos ou janelas temporais.

Essa formalização permite documentar de forma clara como os dados foram reorganizados para análise, quais decisões foram tomadas e como as representações se relacionam com os ativos originais.

7. Entradas do módulo

As entradas do Módulo 13 são os ativos analíticos produzidos pelos módulos anteriores.

As principais entradas incluem:

- datasets brutos;
- datasets integrados;
- datasets tratados;
- datasets curados;
- data marts;
- views analíticas;
- logs de eventos;
- séries históricas;
- dados longitudinais;
- registros transacionais;
- atributos temporais;
- atributos categóricos;
- atributos numéricos;
- identificadores de entidades;
- chaves de relacionamento;
- registros de interação;
- dados de redes;
- dados de sensores;
- dados textuais estruturados ou semiestruturados;
- metadados;
- dicionários de dados;
- glossários;
- regras de negócio;
- registros de qualidade;
- registros de lineage;
- versões de datasets;
- contratos de dados;
- registros de sensibilidade;
- rótulos ou variáveis-alvo;
- features previamente construídas;
- objetivos analíticos;
- requisitos de consumo;
- critérios de governança.

8. Processamento

O processamento do Módulo 13 envolve a definição da finalidade analítica, escolha da representação, construção da estrutura, validação, documentação e disponibilização da representação.

8.1 Definição da finalidade analítica

A primeira atividade consiste em definir qual problema será analisado e qual tipo de representação é mais adequado. A finalidade pode ser descritiva, diagnóstica, preditiva, explicativa, relacional, temporal, prescritiva ou decisória.

Exemplos de finalidade incluem:

- calcular indicadores;
- criar dashboards;
- prever evasão;
- analisar trajetórias;
- identificar padrões temporais;
- detectar mudanças;
- modelar relações entre entidades;
- identificar comunidades;
- calcular centralidades;
- gerar features;
- alimentar agentes inteligentes;
- apoiar políticas públicas.

8.2 Definição da unidade de análise

A segunda atividade consiste em definir a unidade de análise da representação. Essa decisão é fundamental porque determina o nível de agregação e a interpretação dos resultados.

Exemplos de unidade de análise:

- estudante;
- matrícula;
- curso;
- disciplina;
- turma;
- atividade;
- atendimento;
- cidadão;
- família;
- território;
- município;
- instituição;
- evento;
- relação;
- período;
- trajetória.

8.3 Seleção de atributos, entidades e relações

A terceira atividade consiste em selecionar os atributos, entidades e relações relevantes. Para representações tabulares, selecionam-se colunas. Para representações temporais, selecionam-se atributos associados ao tempo. Para grafos, definem-se nós, arestas, tipos de relações, pesos, direções e atributos.

8.4 Definição da dimensão temporal

A quarta atividade consiste em definir como o tempo será tratado. Essa etapa é essencial para dados históricos, longitudinais, eventos, séries temporais e grafos temporais.

Devem ser definidos:

- atributo temporal de referência;
- granularidade temporal;
- janela de observação;
- janela de predição, quando aplicável;
- período de agregação;
- ordem dos eventos;
- periodicidade;
- tratamento de lacunas temporais;
- defasagens;
- alinhamento entre fontes com tempos diferentes.

8.5 Construção da representação tabular

A representação tabular organiza unidades de análise em linhas e atributos em colunas. Essa representação pode incluir atributos originais, derivados, agregados, temporais, categóricos, normalizados ou codificados.

Exemplo:

Estudante × atributos acadêmicos, socioeconômicos e comportamentais

8.6 Construção da representação temporal ou longitudinal

A representação temporal organiza observações ao longo do tempo. Pode ser estruturada como série temporal agregada, painel longitudinal, sequência de estados ou trajetória de eventos.

Exemplo:

Estudante × semestre × desempenho, frequência, situação acadêmica e interações

8.7 Construção da representação baseada em eventos

A representação baseada em eventos organiza ocorrências com timestamp, tipo de evento, entidade associada e contexto. Essa estrutura é adequada para logs e fluxos dinâmicos.

Exemplo:

Evento = {estudante, atividade, tipo_interação, timestamp, duração, recurso_acessado}

8.8 Construção da representação em grafo

A representação em grafo exige definir:

- tipos de nós;
- tipos de arestas;
- direção das relações;
- pesos das arestas;
- atributos dos nós;
- atributos das arestas;
- critérios de criação de relações;
- filtros;
- períodos;
- regras de agregação.

Exemplo:

Grafo educacional com estudantes, disciplinas, cursos, atividades e interações

8.9 Construção da representação em grafo temporal

A representação em grafo temporal incorpora a evolução das relações ao longo do tempo. Pode ser construída por snapshots temporais, arestas com timestamps, janelas deslizantes ou eventos relacionais.

Exemplo:

Grafo semanal de interações entre estudantes e atividades no ambiente virtual

8.10 Validação da representação

A representação deve ser validada quanto a completude, consistência, coerência semântica, cobertura temporal, granularidade, ausência de duplicidade, rastreabilidade, qualidade das relações e adequação à finalidade.

8.11 Documentação e versionamento

Cada representação deve ser documentada e versionada. Devem ser registrados origem, regras de construção, parâmetros, atributos utilizados, período, filtros, versão, limitações e lineage.

8.12 Disponibilização para consumo analítico

A representação validada pode ser disponibilizada para dashboards, modelos, algoritmos de grafos, feature store, relatórios, APIs, agentes inteligentes ou outros produtos analíticos.

9. Técnicas, métodos e tecnologias aplicáveis

O Módulo 13 pode empregar diferentes técnicas, métodos e tecnologias, dependendo da representação adotada.

Para representações tabulares, podem ser utilizadas:

- seleção de atributos;
- agregação;

- normalização;
- codificação de variáveis categóricas;
- transformação de variáveis;
- imputação;
- discretização;
- data profiling;
- matriz de features;
- modelagem dimensional;
- SQL;
- pandas;
- PySpark;
- DuckDB.

Para representações temporais e longitudinais, podem ser utilizadas:

- séries temporais;
- janelas temporais;
- janelas deslizantes;
- defasagens;
- variáveis lag;
- médias móveis;
- tendências;
- sazonalidade;
- detecção de mudança;
- análise de trajetória;
- painéis longitudinais;
- sequências de eventos;
- time-aware feature engineering.

Para representações em eventos, podem ser utilizadas:

- event logs;
- event sourcing;
- timestamps;
- sessões;
- ordenação temporal;
- agregação por janela;
- análise de sequência;
- mineração de processos;
- detecção de padrões de eventos;
- processamento de logs.

Para representações em grafos, podem ser utilizadas:

- modelagem de nós e arestas;
- grafos homogêneos;
- grafos heterogêneos;
- grafos bipartidos;
- grafos direcionados;
- grafos ponderados;
- métricas de centralidade;
- componentes conectados;
- comunidades;
- caminhos;
- similaridade estrutural;

- embeddings de grafos;
- Node2Vec;
- GraphSAGE;
- GCN;
- GAT.

Para grafos temporais, podem ser utilizadas:

- snapshots de grafos;
- arestas com timestamps;
- grafos dinâmicos;
- janelas temporais de rede;
- evolução de centralidade;
- persistência de arestas;
- detecção de comunidades dinâmicas;
- mudança estrutural;
- séries temporais de métricas de rede;
- modelos temporais em grafos.

Tecnologias aplicáveis incluem:

- pandas;
- PySpark;
- SQL;
- DuckDB;
- NetworkX;
- igraph;
- graph-tool;
- Neo4j;
- ArangoDB;
- Gephi;
- PyTorch Geometric;
- DGL;
- scikit-learn;
- MLflow;
- Featuretools;
- Feast;
- bancos relacionais;
- bancos de grafos;
- data warehouses;
- lakehouses;
- ferramentas de visualização.

10. Métricas do módulo

As métricas do Módulo 13 avaliam a qualidade, completude, coerência, cobertura temporal, adequação e utilidade das representações analíticas.

Métrica	Sigla	Descrição
Taxa de representações documentadas	TRD	Proporção de representações com documentação completa.

Taxa de ativos representados	TAR	Proporção de ativos analíticos com representação definida.
Taxa de unidades com representação válida	TURV	Proporção de unidades de análise representadas corretamente.
Completude da representação tabular	CRT	Proporção de valores esperados presentes na matriz tabular.
Densidade tabular	DT	Proporção de células preenchidas em relação ao total esperado.
Taxa de atributos derivados documentados	TADD	Proporção de atributos derivados com regra documentada.
Cobertura temporal	CT	Proporção de períodos esperados representados.
Regularidade temporal	RT	Grau de consistência na periodicidade das observações.
Taxa de lacunas temporais	TLT	Proporção de períodos ausentes ou incompletos.
Taxa de eventos com timestamp válido	TETV	Proporção de eventos com marca temporal válida.
Taxa de eventos ordenáveis	TEO	Proporção de eventos que podem ser ordenados temporalmente.
Densidade do grafo	DG	Relação entre arestas existentes e arestas possíveis.
Número de nós	NN	Quantidade de entidades representadas no grafo.
Número de arestas	NA	Quantidade de relações representadas.
Grau médio	GM	Média de conexões por nó.
Número de componentes	NC	Quantidade de subestruturas desconectadas.
Modularidade	MOD	Grau de estrutura comunitária do grafo.
Taxa de arestas persistentes	TAP	Proporção de relações que permanecem em múltiplos períodos.
Taxa de arestas temporais válidas	TATV	Proporção de arestas com informação temporal válida.
Índice de adequação da representação	IAR	Grau de adequação da representação à finalidade analítica.

10.1 Cobertura Temporal — CT

A **Cobertura Temporal** mede a proporção de períodos esperados que estão representados no ativo analítico.

CT = número de períodos representados / número de períodos esperados

Essa métrica é importante para avaliar séries históricas, dados longitudinais, janelas temporais e grafos temporais.

10.2 Taxa de Lacunas Temporais — TLT

A **Taxa de Lacunas Temporais** mede a proporção de períodos ausentes ou incompletos.

TLT = número de períodos ausentes ou incompletos / número total de períodos esperados

Valores elevados indicam problemas de continuidade temporal.

10.3 Densidade do Grafo — DG

A **Densidade do Grafo** mede a proporção de relações existentes em comparação com o número máximo de relações possíveis.

Para grafos simples não direcionados:

$$\mathbf{DG} = 2|\mathbf{E}| \,/\, |\mathbf{V}|(|\mathbf{V}| - 1)$$

em que:

- **|E|** representa o número de arestas;
- **|V|** representa o número de nós.

10.4 Taxa de Arestas Persistentes — TAP

A **Taxa de Arestas Persistentes** mede a proporção de relações que se mantêm ao longo de múltiplos períodos.

TAP = número de arestas presentes em mais de um período / número total de arestas observadas

Essa métrica é relevante para grafos temporais e redes evolutivas.

10.5 Índice de Adequação da Representação — IAR

O **Índice de Adequação da Representação** avalia se a forma escolhida é adequada à finalidade analítica.

Uma formulação possível é:

$$\mathbf{IAR} = w_1U + w_2A + w_3T + w_4R + w_5Q + w_6L$$

em que:

- **U** representa adequação da unidade de análise;
- **A** representa adequação dos atributos;
- **T** representa adequação temporal;
- **R** representa adequação relacional;
- **Q** representa qualidade da representação;
- **L** representa rastreabilidade e lineage;
- w_i representa o peso de cada dimensão.

11. Artefatos formalizados do módulo

Os artefatos do Módulo 13 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A13.1	Plano de Representação Analítica	Documento técnico	IAR	Define finalidade, unidade de análise e tipo de representação.
A13.2	Matriz Unidade × Atributo	Tabela ou matriz	CRT	Estrutura representação tabular.
A13.3	Especificação da Variável-Alvo e Unidade de Análise	Documento técnico	TRD	Define rótulo, unidade, período e escopo.
A13.4	Representação Temporal ou Longitudinal	Série, painel ou tabela temporal	CT, TLT	Representa evolução, trajetória ou histórico.
A13.5	Registro de Eventos e Estados	Log estruturado	TETV, TEO	Organiza eventos, timestamps e sequências.
A13.6	Modelo de Grafo Analítico	Documento ou grafo	DG, GM	Define nós, arestas, pesos, direção e atributos.
A13.7	Lista de Nós e Arestas	Tabelas node list e edge list	NN, NA	Materializa o grafo em estrutura computacional.
A13.8	Representação em Grafo Temporal	Grafo dinâmico ou snapshots	TAP, TATV	Representa relações evolutivas no tempo.
A13.9	Relatório de Métricas da Representação	Relatório técnico	IAR	Avalia qualidade, cobertura e adequação.
A13.1	Manifesto da	Documento	TRD	Registra regras,

0	Representação Analítica	reprodutível		parâmetros, versões, origem e lineage.
A13.11	Mapa de Transformação entre Representações	Diagrama ou tabela	IAR	Mostra como uma representação deriva de outra.
A13.12	Relatório de Limitações da Representação	Relatório técnico	TLT, IAR	Explicita restrições, lacunas, viés e incerteza.

12. Artefato A13.1 — Plano de Representação Analítica

12.1 Finalidade

O **Plano de Representação Analítica** define qual representação será construída, por que ela é adequada ao problema, quais dados serão utilizados, qual será a unidade de análise, qual dimensão temporal será considerada e quais produtos ou modelos serão alimentados.

12.2 Estrutura mínima

Campo	Descrição
Identificador da representação	Código único.
Finalidade analítica	Objetivo da representação.
Tipo de representação	Tabular, temporal, longitudinal, evento, grafo ou grafo temporal.
Unidade de análise	Entidade principal representada.
Fontes utilizadas	Datasets ou ativos de origem.
Atributos selecionados	Variáveis consideradas.
Dimensão temporal	Período, timestamp, janela ou sequência.
Relações	Nós, arestas, pesos ou dependências, quando aplicável.
Regras de construção	Procedimentos utilizados.
Limitações	Restrições, lacunas ou incertezas.
Versão	Identificador da representação.
Responsável	Pessoa ou equipe.

12.3 Métricas associadas

- índice de adequação da representação;
- taxa de representações documentadas;
- cobertura temporal;
- completude da representação;
- taxa de unidades com representação válida.

12.4 Critérios de validação

O plano é válido quando justifica a escolha da representação e documenta unidade de análise, fontes, atributos, tempo, regras, finalidade e limitações.

12.5 Contribuição

Esse artefato evita que representações sejam construídas de forma implícita ou improvisada, tornando explícitas as decisões metodológicas.

13. Artefato A13.2 — Matriz Unidade × Atributo

13.1 Finalidade

A **Matriz Unidade × Atributo** materializa a representação tabular dos dados, organizando unidades de análise em linhas e atributos em colunas.

13.2 Estrutura mínima

Campo	Descrição
Unidade de análise	Entidade representada na linha.
Identificador da unidade	Código único.
Atributos originais	Variáveis oriundas das fontes.
Atributos derivados	Variáveis construídas.
Período de referência	Quando aplicável.
Variável-alvo	Rótulo ou desfecho, quando aplicável.
Qualidade dos atributos	Métricas de completude, validade ou consistência.
Versão	Versão da matriz.
Lineage	Origem e transformações.

13.3 Métricas associadas

- completude da representação tabular;
- densidade tabular;
- taxa de atributos derivados documentados;
- taxa de unidades com representação válida.

13.4 Critérios de validação

A matriz é válida quando possui unidade de análise clara, atributos documentados, qualidade mínima e rastreabilidade.

13.5 Contribuição

Esse artefato fornece a base para análises estatísticas, dashboards e modelos tradicionais de aprendizagem de máquina.

14. Artefato A13.3 — Especificação da Variável-Alvo e Unidade de Análise

14.1 Finalidade

Este artefato define formalmente a unidade de análise e a variável-alvo ou rótulo, quando a representação será usada em modelos supervisionados, indicadores ou análises específicas.

14.2 Estrutura mínima

Campo	Descrição
Unidade de análise	Entidade analisada.
Variável-alvo	Rótulo, classe, indicador ou desfecho.
Definição semântica	Significado do rótulo.
Regra de construção	Como o rótulo é calculado ou atribuído.
Janela de observação	Período usado para construir atributos.
Janela de predição	Período associado ao desfecho.
Categorias	Classes ou valores possíveis.
Fonte	Origem do rótulo.
Versão	Versão da regra.
Limitações	Restrições interpretativas.

14.3 Métricas associadas

- taxa de rótulos documentados;
- taxa de regras versionadas;
- taxa de rótulos rastreáveis;
- índice de qualidade de rótulos.

14.4 Critérios de validação

O artefato é válido quando permite compreender, reproduzir e auditar a variável-alvo e sua relação com a unidade de análise.

14.5 Contribuição

Esse artefato reduz ambiguidade, data leakage e inconsistência na modelagem analítica.

15. Artefato A13.4 — Representação Temporal ou Longitudinal

15.1 Finalidade

A **Representação Temporal ou Longitudinal** organiza dados de forma a representar evolução, trajetória, sequência ou histórico de uma entidade ou indicador.

15.2 Estrutura mínima

Campo	Descrição
Unidade observada	Entidade acompanhada no tempo.
Atributo temporal	Data, período, semestre, mês, timestamp etc.
Granularidade	Dia, semana, mês, semestre, ano etc.
Atributos observados	Variáveis medidas no tempo.
Janela temporal	Período total coberto.
Frequência	Regularidade esperada.
Lacunas	Períodos ausentes.
Versão	Versão da representação.
Lineage	Origem e transformações.

15.3 Métricas associadas

- cobertura temporal;
- regularidade temporal;
- taxa de lacunas temporais;
- taxa de eventos com timestamp válido.

15.4 Critérios de validação

A representação é válida quando permite analisar evolução temporal com períodos, granularidade, lacunas e regras documentadas.

15.5 Contribuição

Esse artefato permite análises históricas, longitudinais, de tendência, trajetória, recorrência e mudança.

16. Artefato A13.5 — Registro de Eventos e Estados

16.1 Finalidade

O **Registro de Eventos e Estados** organiza acontecimentos, ações ou transições observadas ao longo do tempo.

16.2 Estrutura mínima

Campo	Descrição
Identificador do evento	Código único.
Entidade associada	Unidade vinculada ao evento.
Tipo de evento	Categoria da ocorrência.
Timestamp	Momento do evento.
Estado anterior	Quando aplicável.
Estado posterior	Quando aplicável.
Contexto	Curso, sistema, território, serviço, ambiente etc.
Fonte	Origem do evento.
Atributos do evento	Informações adicionais.
Versão	Versão do registro.

16.3 Métricas associadas

- taxa de eventos com timestamp válido;
- taxa de eventos ordenáveis;
- cobertura temporal;
- taxa de lacunas temporais.

16.4 Critérios de validação

O registro é válido quando eventos possuem entidade, tipo, timestamp, origem e contexto suficientes para análise.

16.5 Contribuição

Esse artefato permite análise de logs, trajetórias, sequências, transições e eventos em tempo real ou quase real.

17. Artefato A13.6 — Modelo de Grafo Analítico

17.1 Finalidade

O **Modelo de Grafo Analítico** define como entidades e relações serão representadas como nós e arestas.

17.2 Estrutura mínima

Campo	Descrição
Tipos de nós	Entidades representadas.
Tipos de arestas	Relações entre entidades.
Direção	Grafo direcionado ou não direcionado.
Peso	Critério de ponderação das relações.
Atributos dos nós	Características das entidades.
Atributos das arestas	Características das relações.
Critério de criação de aresta	Regra para estabelecer relação.
Período	Quando aplicável.
Fonte	Origem dos dados relacionais.
Limitações	Restrições e incertezas.

17.3 Métricas associadas

- número de nós;
- número de arestas;
- densidade do grafo;
- grau médio;
- número de componentes;
- modularidade.

17.4 Critérios de validação

O modelo é válido quando define claramente entidades, relações, pesos, direção, atributos, critérios e finalidade.

17.5 Contribuição

Esse artefato permite analisar relações, conexões, estruturas de rede e dependências entre entidades.

18. Artefato A13.7 — Lista de Nós e Arestas

18.1 Finalidade

A **Lista de Nós e Arestas** materializa o grafo em estruturas computacionais que podem ser carregadas em ferramentas de análise de redes ou bancos de grafos.

18.2 Estrutura mínima

Estrutura	Campos mínimos
Lista de nós	id_nó, tipo_nó, atributos, versão
Lista de arestas	id_origem, id_destino, tipo_aresta, peso, direção, período, atributos, versão

18.3 Métricas associadas

- número de nós;
- número de arestas;
- densidade;
- grau médio;
- taxa de arestas válidas.

18.4 Critérios de validação

A lista é válida quando nós e arestas possuem identificadores consistentes, tipos definidos e relações verificáveis.

18.5 Contribuição

Esse artefato viabiliza implementação prática de grafos em ferramentas computacionais.

19. Artefato A13.8 — Representação em Grafo Temporal

19.1 Finalidade

A **Representação em Grafo Temporal** modela relações entre entidades que evoluem ao longo do tempo.

19.2 Estrutura mínima

Campo	Descrição
Nós temporais	Entidades observadas por período.
Arestas temporais	Relações com timestamp ou janela temporal.

Peso temporal	Intensidade da relação no tempo.
Janela	Período de agregação.
Snapshot	Versão do grafo em determinado tempo.
Métricas temporais	Centralidade, densidade, persistência etc.
Evolução	Mudanças observadas.
Lineage	Origem das relações.

19.3 Métricas associadas

- taxa de arestas persistentes;
- taxa de arestas temporais válidas;
- cobertura temporal;
- densidade por período;
- variação de centralidade.

19.4 Critérios de validação

O grafo temporal é válido quando as relações possuem dimensão temporal clara e podem ser analisadas por período, evento ou janela.

19.5 Contribuição

Esse artefato permite estudar redes dinâmicas, evolução de relações, interações e mudanças estruturais.

20. Artefato A13.9 — Relatório de Métricas da Representação

20.1 Finalidade

O **Relatório de Métricas da Representação** avalia a qualidade, adequação e características estruturais da representação analítica.

20.2 Estrutura mínima

Campo	Descrição
Representação avaliada	Identificador.
Tipo	Tabular, temporal, evento, grafo etc.
Métricas calculadas	Completude, cobertura, densidade, lacunas etc.
Valores obtidos	Resultados das métricas.

Interpretação	Significado dos resultados.
Limitações	Restrições observadas.
Recomendações	Ajustes necessários.

20.3 Métricas associadas

- IAR;
- CRT;
- CT;
- TLT;
- DG;
- GM;
- MOD.

20.4 Critérios de validação

O relatório é válido quando permite avaliar se a representação é adequada à finalidade pretendida.

20.5 Contribuição

Esse artefato apoia comparação entre representações e seleção metodológica justificada.

21. Artefato A13.10 — Manifesto da Representação Analítica

21.1 Finalidade

O **Manifesto da Representação Analítica** registra as informações necessárias para reproduzir a construção de uma representação.

21.2 Estrutura mínima

Campo	Descrição
Representação	Identificador.
Ativos de origem	Datasets utilizados.
Versões	Versões dos ativos de origem.
Regras	Regras de construção.
Parâmetros	Filtros, janelas, agregações, pesos.
Código ou processo	Script, notebook ou pipeline.

Ambiente	Ambiente computacional.
Saída	Representação gerada.
Data	Data de geração.
Responsável	Pessoa ou equipe.
Limitações	Restrições conhecidas.

21.3 Métricas associadas

- taxa de representações documentadas;
- taxa de representações reprodutíveis;
- taxa de lineage associado;
- índice de adequação da representação.

21.4 Critérios de validação

O manifesto é válido quando permite reconstruir a representação a partir dos dados, regras, versões e parâmetros registrados.

21.5 Contribuição

Esse artefato conecta representação analítica, lineage, versionamento e reprodutibilidade.

22. Artefato A13.11 — Mapa de Transformação entre Representações

22.1 Finalidade

O **Mapa de Transformação entre Representações** documenta como uma representação deriva de outra, por exemplo, como uma tabela gera uma matriz de features, como eventos geram séries temporais ou como registros relacionais geram um grafo.

22.2 Estrutura mínima

Campo	Descrição
Representação de origem	Estrutura inicial.
Representação de destino	Estrutura derivada.
Regras de transformação	Procedimentos aplicados.
Parâmetros	Janelas, filtros, agregações, pesos.
Perdas ou agregações	Informações removidas ou

	consolidadas.
Lineage	Relação entre origem e destino.
Limitações	Restrições da conversão.

22.3 Métricas associadas

- taxa de transformações documentadas;
- índice de adequação da representação;
- taxa de perda de informação;
- taxa de lineage associado.

22.4 Critérios de validação

O mapa é válido quando permite compreender como uma representação foi derivada de outra.

22.5 Contribuição

Esse artefato favorece comparação entre paradigmas de representação e reutilização analítica.

23. Artefato A13.12 — Relatório de Limitações da Representação

23.1 Finalidade

O **Relatório de Limitações da Representação** documenta restrições, lacunas, incertezas, vieses e cuidados de interpretação associados a uma representação analítica.

23.2 Estrutura mínima

Campo	Descrição
Representação	Identificador.
Limitações de dados	Ausências, baixa cobertura, inconsistências.
Limitações temporais	Lacunas, defasagens, irregularidades.
Limitações relacionais	Relações inferidas, incompletas ou incertas.
Limitações semânticas	Ambiguidade ou mudanças de significado.
Riscos de viés	Sub-representação ou distorção.
Impacto no uso	Como afeta análise ou modelo.

Recomendações	Cuidados ou ações corretivas.

23.3 Métricas associadas

- taxa de lacunas temporais;
- índice de incerteza;
- índice de adequação da representação;
- taxa de relações incertas.

23.4 Critérios de validação

O relatório é válido quando explicita limitações relevantes para interpretação e uso responsável.

23.5 Contribuição

Esse artefato fortalece transparência, governança e comunicação responsável dos resultados.

24. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Datasets brutos, integrados, tratados e curados; logs; eventos; séries históricas; atributos temporais; chaves; entidades; relações; metadados; rótulos; regras; versões e objetivos analíticos.
Processamento	Definição da finalidade, unidade de análise, atributos, relações, dimensão temporal, construção de tabelas, séries, eventos, grafos, grafos temporais, validação, documentação e versionamento.
Artefatos	Plano de representação, matriz unidade × atributo, especificação de unidade e rótulo, representação temporal, registro de eventos, modelo de grafo, lista de nós e arestas, grafo temporal, relatório de métricas, manifesto e relatório de limitações.
Métricas	TRD, TAR, TURV, CRT, DT, CT, RT, TLT, TETV, TEO, DG, NN, NA, GM, NC, MOD, TAP, TATV e IAR.
Saídas	Representações analíticas documentadas, versionadas, rastreáveis e adequadas para dashboards, modelos, grafos, feature engineering, agentes inteligentes e tomada de decisão.

25. Critérios gerais de validação do módulo

O Módulo 13 pode ser considerado adequadamente implementado quando:

1. a finalidade analítica da representação está definida;
2. a unidade de análise está explicitada;
3. os ativos de origem estão identificados;

4. os atributos utilizados estão documentados;
5. a dimensão temporal está definida quando aplicável;
6. eventos possuem timestamp, tipo, entidade e contexto;
7. grafos possuem nós, arestas, pesos, direção e regras de construção documentados;
8. grafos temporais possuem janelas ou timestamps definidos;
9. representações possuem métricas de qualidade e adequação calculadas;
10. limitações, lacunas e incertezas estão documentadas;
11. a representação está versionada;
12. a representação possui lineage associado;
13. a representação pode ser reproduzida a partir de manifesto;
14. a representação é adequada ao uso pretendido;
15. os artefatos podem ser consumidos pelos módulos de feature engineering, modelagem, monitoramento, agentes inteligentes e disponibilização analítica.

26. Relação com os demais módulos do modelo de referência

O Módulo 13 depende diretamente dos módulos anteriores e fornece insumos para os módulos posteriores.

Com o **Módulo 1 — Fontes e Inventário de Dados**, utiliza informações sobre fontes, entidades, granularidade, cobertura temporal e restrições.

Com o **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, utiliza dados brutos, logs, eventos, janelas temporais e snapshots.

Com o **Módulo 3 — Integração e Harmonização**, utiliza dados integrados, chaves, entidades comuns e equivalências semânticas.

Com o **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, utiliza dados tratados, atributos derivados e regras de padronização.

Com o **Módulo 5 — Qualidade e Validação de Dados**, incorpora métricas de qualidade na avaliação das representações.

Com o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, considera representatividade, grupos, rótulos e riscos de viés.

Com o **Módulo 7 — Armazenamento Analítico**, utiliza datasets curados e data marts como base de representação.

Com o **Módulo 8 — Disponibilização e Consumo Analítico**, fornece representações para dashboards, APIs, relatórios e produtos de dados.

Com o **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, documenta significado, estrutura, relações e finalidade das representações.

Com o **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, respeita restrições de acesso, sensibilidade, finalidade e minimização.

Com o **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, registra origem, transformações e dependências das representações.

Com o **Módulo 12 — Versionamento e Reprodutibilidade**, versiona representações, regras, parâmetros e manifestos.

Com o **Módulo 14 — Feature Engineering e Feature Store**, fornece estruturas para criação de features tabulares, temporais, relacionais e de grafos.

Com o **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, permite acompanhar mudanças nas distribuições, relações, eventos e estruturas ao longo do tempo.

Com o **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, fornece representações e metadados que podem ser consultados, explicados e recomendados por agentes.

27. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 13 pode representar dados acadêmicos de diferentes formas.

Uma representação tabular pode conter uma linha por estudante, com atributos como curso, idade, desempenho médio, frequência, forma de ingresso, quantidade de reprovações e situação acadêmica.

Uma representação longitudinal pode acompanhar cada estudante por semestre, registrando evolução de desempenho, frequência, reprovações, trancamentos, interações no ambiente virtual e mudanças de situação acadêmica.

Uma representação baseada em eventos pode registrar interações no ambiente virtual, como acesso a materiais, envio de atividades, participação em fóruns, visualização de vídeos e realização de avaliações.

Uma representação em grafo pode modelar relações entre estudantes, disciplinas, cursos, atividades, docentes e interações.

Um grafo temporal pode representar como as interações entre estudantes e atividades evoluem semanalmente, permitindo analisar engajamento, isolamento, centralidade e mudança de comportamento ao longo do tempo.

Essas diferentes representações permitem estudar evasão, permanência, desempenho, engajamento, trajetórias, redes de aprendizagem e padrões temporais de forma mais rica do que uma única tabela estática.

28. Exemplo de aplicação em políticas públicas

Em políticas públicas, o Módulo 13 pode representar dados de atendimento, território, serviços, famílias, escolas, unidades de saúde, programas sociais e indicadores socioeconômicos.

Uma representação tabular pode conter indicadores agregados por território.

Uma representação temporal pode mostrar evolução mensal de atendimentos, vulnerabilidade, demanda por serviços ou cobertura de programas.

Uma representação longitudinal pode acompanhar famílias, territórios ou unidades ao longo do tempo.

Uma representação baseada em eventos pode registrar atendimentos, encaminhamentos, visitas, notificações ou ocorrências.

Uma representação em grafo pode modelar relações entre cidadãos, famílias, serviços, equipamentos públicos, territórios e programas.

Um grafo temporal pode representar a evolução dos vínculos entre famílias e serviços públicos, permitindo analisar continuidade de atendimento, encaminhamentos, sobreposição de políticas e redes de suporte.

29. Contribuição técnico-científica do módulo

A contribuição técnico-científica do Módulo 13 está em reconhecer que a forma de representar os dados é uma decisão metodológica essencial para a validade, a utilidade e a interpretabilidade dos resultados analíticos.

Em vez de restringir os dados a estruturas tabulares tradicionais, o UnespDataLens-RM propõe um módulo específico para formalizar representações tabulares, temporais, longitudinais, baseadas em eventos, relacionais e em grafos.

Essa formalização contribui para:

- ampliar a capacidade analítica do modelo de referência;
- representar fenômenos dinâmicos e relacionais;
- incorporar explicitamente a dimensão temporal;
- apoiar análise de trajetórias e eventos;
- permitir uso de grafos e grafos temporais;
- viabilizar features temporais e relacionais;
- melhorar modelagem preditiva e explicabilidade;
- apoiar visualizações mais adequadas;
- fortalecer rastreabilidade entre dados e modelos;
- reduzir ambiguidades na unidade de análise;
- melhorar comparação entre paradigmas de representação;
- apoiar pesquisas em diferentes domínios, como educação, políticas públicas, saúde e trabalho.

30. Síntese do módulo

O **Módulo Transversal 13 — Representações Analíticas: Tabular, Temporal, Espacial, Textual, Semântica e Grafos** define como os dados produzidos pelo UnespDataLens-RM devem ser organizados para análise computacional, modelagem, visualização e tomada de decisão.

Sua importância está em reconhecer que os dados não possuem uma única forma natural de representação. A escolha entre tabela, série temporal, trajetória longitudinal, evento, grafo ou grafo temporal depende da finalidade analítica, da unidade de análise, da temporalidade, das relações entre entidades, da qualidade dos dados e do tipo de produto esperado.

Ao formalizar representações analíticas documentadas, versionadas, rastreáveis, avaliáveis e reprodutíveis, o Módulo 13 fortalece a capacidade do UnespDataLens-RM de tratar dados estáticos, históricos, dinâmicos, temporais, relacionais e evolutivos, ampliando sua

relevância para análises avançadas, modelos preditivos, grafos, agentes inteligentes e tomada de decisão baseada em evidências.

Módulo Transversal 14 — Feature Engineering e Feature Store

1. Finalidade do módulo

O **Módulo Transversal 14 — Feature Engineering e Feature Store** tem como finalidade especificar, construir, documentar, validar, armazenar, versionar, reutilizar e monitorar atributos analíticos, também denominados **features**, produzidos a partir dos dados processados no modelo de referência **UnespDataLens-RM**.

Esse módulo atua de forma transversal porque a criação de features pode ocorrer em diferentes momentos do ciclo de dados. As features podem ser derivadas de fontes originais, dados integrados, dados tratados, dados validados, representações tabulares, séries temporais, grafos, indicadores, data marts, produtos de dados ou registros de consumo. Portanto, o módulo não pertence apenas a uma etapa linear do pipeline, mas se conecta a praticamente todos os módulos do UnespDataLens-RM.

No contexto da Engenharia Analítica de Dados, **feature engineering** corresponde ao processo de transformar dados disponíveis em atributos mais informativos, úteis e adequados para análises, modelos preditivos, indicadores, dashboards, sistemas de recomendação, análise de risco, agrupamentos, explicabilidade, tomada de decisão e agentes inteligentes. Esse processo envolve selecionar, construir, combinar, transformar, agregar, normalizar, codificar, temporalizar, contextualizar ou representar variáveis de modo que elas capturem melhor o fenômeno investigado.

A **feature store**, por sua vez, corresponde a uma estrutura organizada para armazenar, documentar, versionar, controlar, reutilizar e disponibilizar features de forma consistente. Ela permite que atributos criados em um projeto não fiquem dispersos em scripts, notebooks, planilhas ou pipelines isolados, mas sejam tratados como ativos analíticos reutilizáveis, governados, rastreáveis e monitoráveis.

No UnespDataLens-RM, esse módulo é essencial porque grande parte do valor analítico dos dados depende da qualidade das features construídas. Um modelo preditivo, um painel de indicadores ou uma análise exploratória podem ser fortemente influenciados por como as variáveis foram definidas, calculadas, agregadas, temporalizadas e disponibilizadas. Features mal definidas, sem documentação, sem versionamento ou sem controle de qualidade podem comprometer resultados, gerar vieses, dificultar reprodutibilidade e reduzir confiança nos produtos analíticos.

Assim, o Módulo 14 formaliza feature engineering e feature store como componentes transversais da Engenharia Analítica de Dados, garantindo que atributos analíticos sejam produzidos, avaliados e reutilizados de forma sistemática, documentada, rastreável e adequada ao uso.

2. Objetivo geral

O objetivo geral do Módulo 14 é **formalizar o processo de criação, documentação, validação, armazenamento, versionamento, reutilização e monitoramento de features**

no UnespDataLens-RM, garantindo que atributos analíticos sejam semanticamente coerentes, tecnicamente corretos, rastreáveis, governados, reprodutíveis e adequados para análises, modelos, dashboards, indicadores, grafos, APIs e agentes inteligentes.

3. Objetivos específicos

São objetivos específicos do Módulo 14:

1. identificar atributos candidatos à criação de features a partir de fontes, datasets tratados, data marts, representações analíticas e produtos de dados;
2. definir regras formais para criação de features derivadas, agregadas, temporais, relacionais, estatísticas, categóricas, textuais, semânticas ou baseadas em grafos;
3. documentar origem, fórmula, granularidade, janela temporal, atributos de entrada, regra de cálculo, versão, finalidade e limitações de cada feature;
4. avaliar a qualidade das features quanto à completude, validade, estabilidade, distribuição, correlação, redundância, vazamento de informação, sensibilidade e risco de viés;
5. armazenar features em uma estrutura organizada, reutilizável, versionada e controlada;
6. diferenciar features exploratórias, features aprovadas, features depreciadas, features sensíveis, features restritas e features prontas para produção;
7. evitar duplicação de atributos derivados e inconsistências entre cálculos realizados por diferentes equipes, scripts ou projetos;
8. garantir consistência entre features usadas em treinamento, validação, testes, dashboards, APIs e aplicações de consumo;
9. integrar features aos módulos de metadados, governança, linhagem, versionamento, representações analíticas, monitoramento e agentes inteligentes;
10. produzir artefatos formais que permitam auditoria, reprodutibilidade, reuso e avaliação da contribuição das features para análises e modelos.

4. Fundamentação conceitual do módulo

Features são atributos analíticos utilizados para representar aspectos relevantes de uma unidade de análise. Elas podem ser atributos originais, atributos derivados, indicadores, agregações, transformações, medidas temporais, métricas relacionais, embeddings, variáveis categorizadas, estatísticas por janela, scores, sinais de comportamento ou representações computacionais extraídas dos dados.

No UnespDataLens-RM, uma feature deve ser tratada como um **ativo analítico**, e não apenas como uma coluna criada em um script. Isso significa que cada feature deve possuir definição, finalidade, origem, regra de cálculo, granularidade, versão, qualidade, restrição, sensibilidade, linhagem e condição de uso.

As features podem assumir diferentes tipos:

Tipo de feature	Descrição	Exemplo
Original	Atributo preservado de uma fonte ou dataset.	idade, sexo, curso, município
Derivada	Atributo calculado a partir de outros atributos.	idade no ingresso, tempo desde matrícula

Agregada	Atributo calculado por grupo, período ou entidade.	média de notas por semestre
Temporal	Atributo associado a janelas, evolução ou sequência.	frequência acumulada nos últimos 3 meses
Categórica transformada	Codificação de categorias.	curso codificado, faixa de renda
Estatística	Medida resumida de distribuição ou comportamento.	média, desvio padrão, tendência
Relacional	Atributo derivado de relações entre entidades.	número de disciplinas conectadas ao estudante
Baseada em grafos	Métrica ou embedding extraído de grafo.	grau, centralidade, PageRank, Node2Vec
Semântica ou textual	Atributo extraído de texto, categoria ou significado.	tópico dominante, embedding textual
Indicador composto	Combinação de múltiplas variáveis.	índice de vulnerabilidade, score de risco

A feature store tem como função organizar essas features em uma estrutura reutilizável. Ela pode ser simples, como um catálogo estruturado de features com tabelas versionadas, ou mais sofisticada, envolvendo repositórios online e offline, APIs, pipelines automatizados, validação, monitoramento e integração com modelos.

No contexto do UnespDataLens-RM, a feature store pode cumprir funções como:

- armazenar features validadas;
- documentar regras de cálculo;
- controlar versões;
- registrar linhagem;
- evitar duplicidade de features;
- garantir consistência entre treinamento e consumo;
- apoiar reuso em diferentes projetos;
- controlar acesso a features sensíveis;
- monitorar estabilidade e drift das features;
- fornecer insumos para modelos, dashboards, indicadores e agentes inteligentes.

À luz da **Design Science Research**, o Módulo 14 formaliza feature engineering e feature store como artefatos técnico-metodológicos voltados a resolver problemas recorrentes em pipelines analíticos: criação dispersa de variáveis, ausência de documentação, baixa reprodutibilidade, duplicação de cálculos, inconsistência entre ambientes, vazamento de informação, falta de controle de qualidade e dificuldade de reuso.

5. Natureza transversal do módulo

O Módulo 14 é transversal porque features podem ser produzidas, consumidas, avaliadas e monitoradas em diferentes pontos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, identifica atributos candidatos e unidades de análise relevantes.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, utiliza dados brutos e snapshots como base para features reprocessáveis.

No **Módulo 3 — Integração e Harmonização**, utiliza entidades integradas, chaves e relações entre fontes para criar features combinadas.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, recebe atributos tratados, padronizados e enriquecidos, além de produzir atributos derivados.

No **Módulo 5 — Qualidade e Validação de Dados**, utiliza métricas de qualidade para aprovar, rejeitar ou sinalizar features.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, avalia se features afetam grupos de forma desigual, se funcionam como proxies sensíveis ou se introduzem riscos de viés.

No **Módulo 7 — Armazenamento Analítico**, utiliza data marts, views e datasets persistidos para armazenar ou materializar features.

No **Módulo 8 — Disponibilização e Consumo Analítico**, disponibiliza features para dashboards, APIs, modelos, produtos de dados e agentes inteligentes.

No **Módulo 9 — Metadados, Catálogo e Semântica**, registra definições, descrições, termos, domínios, significado e uso das features.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, controla acesso, sensibilidade, restrição de uso e finalidade das features.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, registra origem, atributos de entrada, regras e dependências das features.

No **Módulo 12 — Versionamento e Reprodutibilidade**, controla versões, parâmetros, regras e ambientes usados para gerar features.

No **Módulo 13 — Representações Analíticas**, consome representações tabulares, temporais, espaciais, textuais, semânticas e em grafos para criar features específicas dessas estruturas.

6. Formalização sintética do módulo

O Módulo 14 pode ser formalizado como um processo de geração e gestão de features a partir de ativos analíticos.

Considerando o conjunto de ativos disponíveis:

$$\mathbf{A = \{a_1, a_2, ..., a_n\}}$$

e o conjunto de unidades de análise:

$$\mathbf{U = \{u_1, u_2, ..., u_m\}}$$

o módulo produz um conjunto de features:

$F = \{f_1, f_2, ..., f_k\}$

Cada feature **f_i** pode ser descrita por:

f_i = {id, nome, descrição, tipo, unidade_análise, atributos_origem, regra_cálculo, parâmetros, janela_temporal, granularidade, versão, qualidade, sensibilidade, finalidade, lineage}

A matriz de features pode ser representada por:

$X \in R^{n \times p}$

em que:

- **n** representa o número de unidades de análise;
- **p** representa o número de features;
- **X_{ij}** representa o valor da feature **j** para a unidade **i**.

A feature store pode ser representada como um repositório estruturado:

FS = {F, M, V, Q, G, L}

em que:

- **F** representa o conjunto de features;
- **M** representa metadados das features;
- **V** representa versões;
- **Q** representa métricas de qualidade;
- **G** representa regras de governança;
- **L** representa linhagem.

Dessa forma, uma feature não é apenas uma variável calculada, mas um ativo com definição, origem, regra, qualidade, versão, governança e rastreabilidade.

7. Entradas do módulo

As entradas do Módulo 14 são compostas por ativos produzidos pelo pipeline principal e por informações dos módulos transversais.

As principais entradas são:

- datasets brutos;
- datasets integrados;
- datasets tratados;
- datasets validados;
- datasets balanceados ou sinalizados;
- data marts;
- views analíticas;
- produtos de dados;
- representações tabulares;
- representações temporais;
- grafos;
- grafos temporais;
- atributos originais;

- atributos derivados;
- indicadores;
- logs;
- eventos;
- registros de interação;
- tabelas de entidades;
- tabelas de relações;
- variáveis-alvo;
- unidades de análise;
- janelas temporais;
- metadados;
- dicionários de dados;
- glossários;
- regras de negócio;
- métricas de qualidade;
- registros de viés;
- registros de linhagem;
- versões;
- requisitos de modelos;
- requisitos de dashboards;
- requisitos de APIs;
- requisitos de análise.

8. Processamento

O processamento do Módulo 14 envolve identificação, criação, validação, documentação, armazenamento, versionamento, disponibilização e monitoramento das features.

8.1 Identificação de necessidades analíticas

A primeira atividade consiste em identificar quais features são necessárias para responder a uma pergunta analítica, construir um modelo, alimentar um dashboard, gerar um indicador, compor um grafo, apoiar uma decisão ou disponibilizar um produto de dados.

Essa identificação deve considerar a finalidade, a unidade de análise, a variável-alvo, a temporalidade, as relações, os grupos relevantes e as restrições de uso.

8.2 Seleção de atributos e fontes de origem

A segunda atividade consiste em selecionar os atributos, datasets, eventos, relações ou indicadores que servirão de base para criação das features. Essa seleção deve considerar qualidade, sensibilidade, completude, rastreabilidade e adequação semântica dos atributos.

8.3 Definição das regras de criação de features

A terceira atividade consiste em definir regras formais para gerar features. Essas regras podem envolver agregações, transformações, codificações, janelas temporais, cálculos estatísticos, métricas de grafos, combinações de atributos, normalizações, discretizações ou embeddings.

Cada regra deve ser documentada com fórmula, parâmetros, atributos de origem, granularidade, unidade de análise, janela temporal e justificativa.

8.4 Construção de features tabulares

A quarta atividade consiste em construir features em formato tabular. Essas features podem incluir atributos originais, atributos derivados, indicadores, codificações categóricas, agregações por unidade de análise, normalizações e variáveis de contexto.

8.5 Construção de features temporais

A quinta atividade consiste em criar features que capturam evolução temporal, como médias móveis, tendências, variações, contagens por janela, eventos recentes, tempo desde o último evento, sequência de estados, persistência, recorrência ou sazonalidade.

8.6 Construção de features relacionais e baseadas em grafos

A sexta atividade consiste em criar features a partir de relações entre entidades. Isso pode incluir grau, grau ponderado, centralidade, PageRank, número de vizinhos, pertencimento a comunidades, distância média, componentes, densidade local, embeddings de nós ou métricas de grafos temporais.

8.7 Avaliação de qualidade das features

A sétima atividade consiste em avaliar as features quanto à completude, validade, distribuição, estabilidade, redundância, multicolinearidade, correlação, sensibilidade, risco de vazamento de informação e adequação ao uso.

Features com baixa qualidade, instabilidade ou risco elevado devem ser sinalizadas, revisadas ou excluídas.

8.8 Avaliação de vazamento de informação

A oitava atividade consiste em verificar se uma feature contém informação que não estaria disponível no momento da predição ou decisão. Esse problema, conhecido como **data leakage**, pode gerar resultados artificialmente elevados e comprometer a validade de modelos.

A verificação deve considerar janela de observação, janela de predição, data de disponibilidade do atributo e sequência temporal dos eventos.

8.9 Registro e documentação das features

A nona atividade consiste em registrar cada feature em um catálogo ou feature store, documentando nome, descrição, tipo, origem, regra, unidade de análise, granularidade, janela temporal, versão, qualidade, sensibilidade, limitações e consumidores.

8.10 Armazenamento e disponibilização na feature store

A décima atividade consiste em armazenar features aprovadas em uma estrutura organizada, permitindo reuso por modelos, dashboards, APIs, análises e agentes inteligentes.

A feature store deve preservar versões, metadados, lineage, status de qualidade e regras de acesso.

9. Técnicas, métodos e algoritmos aplicáveis

O Módulo 14 pode empregar diferentes técnicas de feature engineering.

Entre as técnicas tabulares, destacam-se:

- seleção de atributos;
- criação de atributos derivados;
- normalização;
- padronização;
- discretização;
- binarização;
- one-hot encoding;
- ordinal encoding;
- target encoding, quando adequado;
- agregações por grupo;
- interações entre variáveis;
- transformação logarítmica;
- cálculo de proporções;
- criação de faixas;
- tratamento de categorias raras;
- redução de dimensionalidade;
- análise de correlação;
- análise de multicolinearidade.

Entre as técnicas temporais, destacam-se:

- janelas móveis;
- médias móveis;
- somatórios por janela;
- contagens por período;
- defasagens temporais;
- tempo desde o último evento;
- frequência recente;
- tendência;
- variação percentual;
- volatilidade;
- sequência de estados;
- persistência;
- recorrência;
- sazonalidade;
- extração de eventos;
- agregações por período.

Entre as técnicas relacionais e baseadas em grafos, destacam-se:

- grau;
- grau ponderado;
- centralidade de intermediação;
- centralidade de proximidade;
- centralidade de autovetor;
- PageRank;
- coeficiente de agrupamento;
- número de vizinhos;

- densidade local;
- componentes conectados;
- modularidade;
- pertencimento a comunidade;
- Louvain;
- Label Propagation;
- distância média;
- métricas bipartidas;
- projeções de grafos;
- persistência de vínculos;
- embeddings como Node2Vec, DeepWalk, GraphSAGE, GCN ou GAT, quando aplicável.

Entre as técnicas de avaliação de features, destacam-se:

- análise de completude;
- análise de distribuição;
- estabilidade temporal;
- Population Stability Index;
- correlação;
- VIF;
- mutual information;
- feature importance;
- permutation importance;
- SHAP, quando houver modelo;
- testes de vazamento;
- análise por grupos;
- análise de fairness;
- avaliação de redundância;
- avaliação de drift.

Ferramentas possíveis incluem Python, pandas, scikit-learn, PySpark, SQL, DuckDB, dbt, Featuretools, Feast, Tecton, Hopsworks, MLflow, DVC, NetworkX, igraph, Neo4j, PyTorch Geometric, DGL e ambientes de lakehouse.

10. Métricas do módulo

As métricas do Módulo 14 avaliam cobertura, qualidade, estabilidade, reuso, rastreabilidade, contribuição analítica e governança das features.

Métrica	Sigla	Descrição
Número de features criadas	NFC	Quantidade total de features produzidas
Taxa de features documentadas	TFD	Proporção de features com documentação mínima
Taxa de features com lineage	TFL	Proporção de features com origem e regra rastreáveis
Taxa de features versionadas	TFV	Proporção de features com versão registrada

Taxa de features aprovadas	TFA	Proporção de features aprovadas para uso
Taxa de features reutilizadas	TFR	Proporção de features usadas em mais de um produto ou modelo
Completude média das features	CMF	Grau médio de preenchimento das features
Estabilidade temporal das features	ETF	Grau de estabilidade ao longo do tempo
Taxa de features com drift	TFDri	Proporção de features com mudança significativa de distribuição
Taxa de features sensíveis	TFS	Proporção de features sensíveis ou restritas
Taxa de features com risco de proxy	TFRP	Proporção de features que podem representar atributo sensível indiretamente
Taxa de features com vazamento	TFVaz	Proporção de features com risco de data leakage
Redundância entre features	RF	Grau de correlação ou sobreposição entre features
Contribuição preditiva média	CPM	Importância média das features em modelos
Índice de qualidade da feature	IQF	Indicador composto de qualidade de uma feature
Taxa de features monitoradas	TFM	Proporção de features acompanhadas ao longo do tempo

10.1 Taxa de Features Documentadas — TFD

A **Taxa de Features Documentadas** mede a proporção de features com documentação mínima.

TFD = número de features documentadas / número total de features criadas

Uma feature pode ser considerada documentada quando possui nome, descrição, origem, regra de cálculo, unidade de análise, granularidade, versão e finalidade.

10.2 Taxa de Features com Lineage — TFL

A **Taxa de Features com Lineage** mede a proporção de features cuja origem e regra de derivação estão registradas.

TFL = número de features com lineage / número total de features criadas

Essa métrica é essencial para auditoria, explicabilidade e reprodutibilidade.

10.3 Completude Média das Features — CMF

A **Completude Média das Features** avalia o grau médio de preenchimento das features.

CMF = soma das taxas de completude das features / número total de features avaliadas

Features com baixa completude podem ser inadequadas para modelos ou dashboards.

10.4 Taxa de Features com Vazamento — TFVaz

A **Taxa de Features com Vazamento** mede a proporção de features que apresentam risco de utilizar informação futura, indevida ou indisponível no momento da decisão.

TFVaz = número de features com risco de vazamento / número total de features avaliadas

Essa métrica é especialmente importante em problemas preditivos temporais.

10.5 Índice de Qualidade da Feature — IQF

O **Índice de Qualidade da Feature** pode ser definido como um indicador composto que considera completude, validade, estabilidade, rastreabilidade, documentação e ausência de vazamento.

Uma formulação possível é:

$$IQF(f_i) = w_1C + w_2V + w_3E + w_4L + w_5D + w_6(1 - Z)$$

em que:

- **C** representa completude;
- **V** representa validade;
- **E** representa estabilidade;
- **L** representa lineage;
- **D** representa documentação;
- **Z** representa risco de vazamento;
- w_i representa o peso de cada dimensão.

11. Artefatos formalizados do módulo

Os artefatos do Módulo 14 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, regras de construção, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A14.1	Plano de Feature Engineering	Documento técnico	IQF	Define features, finalidade e regras

A14.2	Catálogo de Features	Catálogo estruturado	TFD	Documenta features disponíveis
A14.3	Registro de Regras de Cálculo de Features	Tabela técnica	TFL	Formaliza origem e fórmulas
A14.4	Matriz de Features	Tabela unidade × features	CMF	Materializa features por unidade
A14.5	Registro de Features Temporais	Tabela temporal	ETF, TFDri	Documenta janelas e evolução
A14.6	Registro de Features Relacionais e de Grafos	Tabela/grafo	TFL	Documenta métricas relacionais
A14.7	Relatório de Qualidade das Features	Relatório técnico	IQF	Avalia qualidade e adequação
A14.8	Registro de Vazamento e Riscos de Feature	Relatório/tabela	TFVaz	Identifica leakage e riscos
A14.9	Feature Store	Repositório estruturado	TFR, TFV	Armazena e reutiliza features
A14.10	Manifesto de Feature Set	Documento estruturado	TFD, TFL	Registra conjunto de features usado

12. Artefato A14.1 — Plano de Feature Engineering

12.1 Finalidade

O **Plano de Feature Engineering** define quais features serão criadas, para qual finalidade, com quais dados de origem, unidade de análise, regras de cálculo, janelas temporais, granularidade e critérios de validação.

12.2 Tipo de artefato na DSR

Esse artefato é um **método**, pois orienta a construção das features, e uma **instanciação**, quando documentado em relatório, planilha, notebook, arquivo YAML, JSON ou ferramenta específica.

12.3 Estrutura mínima

Campo	Descrição
Objetivo analítico	Finalidade da feature ou conjunto de features
Unidade de análise	Estudante, matrícula, município, família, atendimento etc.

Variável-alvo	Quando aplicável
Features previstas	Lista de atributos a criar
Dados de origem	Datasets e atributos utilizados
Regra de cálculo	Fórmula ou procedimento
Janela temporal	Quando aplicável
Granularidade	Nível de detalhe
Critério de validação	Condição para aceitar a feature
Riscos	Vazamento, viés, sensibilidade ou instabilidade
Responsável	Pessoa ou equipe

12.4 Métricas associadas

- número de features planejadas;
- taxa de features documentadas;
- taxa de features com regra definida;
- taxa de features com risco avaliado;
- índice de qualidade da feature.

12.5 Critérios de validação

O plano é considerado válido quando cada feature prevista possui finalidade, origem, regra de cálculo, unidade de análise, granularidade, versão e critério de validação.

12.6 Contribuição

Esse artefato evita criação improvisada de atributos e torna o processo de feature engineering explícito, auditável e reprodutível.

13. Artefato A14.2 — Catálogo de Features

13.1 Finalidade

O **Catálogo de Features** registra as features disponíveis, suas definições, regras, versões, status, qualidade, origem, restrições e consumidores.

13.2 Estrutura mínima

Campo	Descrição
Identificador da feature	Código único
Nome da feature	Nome técnico ou semântico
Descrição	Significado analítico

Tipo	Original, derivada, temporal, grafo, textual etc.
Unidade de análise	Entidade à qual se aplica
Dados de origem	Ativos utilizados
Regra de cálculo	Fórmula ou procedimento
Versão	Versão da feature
Qualidade	Status ou métricas
Sensibilidade	Classificação
Status	Experimental, aprovada, depreciada, restrita
Consumidores	Modelos, dashboards, APIs, análises
Limitações	Cuidados de uso

13.3 Métricas associadas

- taxa de features documentadas;
- taxa de features aprovadas;
- taxa de features reutilizadas;
- taxa de features versionadas;
- taxa de features com lineage.

13.4 Contribuição

Esse artefato transforma features em ativos reutilizáveis, reduzindo duplicidade e aumentando consistência entre projetos.

14. Artefato A14.3 — Registro de Regras de Cálculo de Features

14.1 Finalidade

O **Registro de Regras de Cálculo de Features** documenta como cada feature é produzida.

14.2 Estrutura formal

Cada regra pode ser descrita por:

r_i = {feature, atributos_origem, fórmula, parâmetros, janela, granularidade, condição, versão, responsável}

14.3 Estrutura mínima

Campo	Descrição
Feature	Nome ou identificador
Atributos de origem	Campos utilizados
Datasets de origem	Ativos de entrada
Fórmula ou regra	Como a feature é calculada
Parâmetros	Valores utilizados
Janela temporal	Quando aplicável
Condição de aplicação	Filtros ou critérios
Granularidade	Unidade de cálculo
Versão da regra	Controle de versão
Responsável	Pessoa ou sistema

14.4 Métricas associadas

- taxa de features com regra documentada;
- taxa de features com lineage;
- taxa de regras versionadas;
- taxa de regras com parâmetros documentados.

14.5 Contribuição

Esse artefato permite reproduzir features e auditar sua construção.

15. Artefato A14.4 — Matriz de Features

15.1 Finalidade

A **Matriz de Features** materializa as features em formato unidade × atributo, permitindo uso em modelos, análises, dashboards ou produtos de dados.

15.2 Estrutura formal

A matriz pode ser representada por:

$$X \in R^{n \times p}$$

em que:

- **n** representa o número de unidades de análise;
- **p** representa o número de features;

- X_{ij} representa o valor da feature **j** para a unidade **i**.

15.3 Estrutura mínima

Campo	Descrição
Identificador da unidade	Chave da unidade de análise
Features	Atributos calculados
Versão das features	Versões utilizadas
Data de geração	Momento de criação
Janela temporal	Quando aplicável
Qualidade	Status da matriz
Sensibilidade	Classificação
Variável-alvo	Quando aplicável

15.4 Métricas associadas

- completude média das features;
- dimensionalidade;
- taxa de features derivadas;
- taxa de features sensíveis;
- taxa de features com drift.

15.5 Contribuição

Esse artefato permite utilizar features de forma estruturada e consistente em análises e modelos.

16. Artefato A14.5 — Registro de Features Temporais

16.1 Finalidade

O **Registro de Features Temporais** documenta atributos criados a partir de janelas, sequências, trajetórias, eventos e evolução temporal.

16.2 Estrutura mínima

Campo	Descrição
Feature temporal	Nome da feature
Unidade de análise	Entidade acompanhada
Janela temporal	Período considerado

Função de agregação	Média, soma, contagem, tendência etc.
Evento base	Evento que gera a feature
Data de referência	Ponto temporal usado
Atributos de origem	Campos utilizados
Regra de cálculo	Fórmula
Risco de vazamento	Avaliação
Versão	Versão da regra

16.3 Métricas associadas

- estabilidade temporal das features;
- taxa de features com drift;
- taxa de features com vazamento;
- cobertura temporal;
- taxa de lacunas temporais.

16.4 Contribuição

Esse artefato permite construir atributos que capturam trajetória, mudança, recorrência e comportamento ao longo do tempo.

17. Artefato A14.6 — Registro de Features Relacionais e de Grafos

17.1 Finalidade

O **Registro de Features Relacionais e de Grafos** documenta features construídas a partir de relações entre entidades e métricas de grafos.

17.2 Estrutura mínima

Campo	Descrição
Feature relacional	Nome da feature
Grafo de origem	Representação utilizada
Tipo de nó	Entidade associada
Tipo de aresta	Relação utilizada
Métrica	Grau, centralidade, PageRank etc.

Peso	Critério de ponderação
Janela temporal	Quando aplicável
Regra de cálculo	Procedimento
Versão	Versão da representação
Limitações	Cuidados de interpretação

17.3 Métricas associadas

- taxa de features relacionais documentadas;
- taxa de features com grafo de origem identificado;
- taxa de features com versão;
- estabilidade das métricas de grafo;
- contribuição preditiva das features relacionais.

17.4 Contribuição

Esse artefato permite incorporar estrutura relacional e dinâmica de redes em análises e modelos.

18. Artefato A14.7 — Relatório de Qualidade das Features

18.1 Finalidade

O **Relatório de Qualidade das Features** avalia se as features são adequadas ao uso analítico.

18.2 Estrutura mínima

Campo	Descrição
Feature avaliada	Nome da feature
Completude	Percentual de valores preenchidos
Validade	Conformidade com domínio
Distribuição	Perfil estatístico
Estabilidade	Variação temporal
Correlação	Relação com outras features
Redundância	Sobreposição com outras features

Sensibilidade	Classificação
Risco de viés	Avaliação por grupos
Status	Aprovada, ressalva, rejeitada

18.3 Métricas associadas

- índice de qualidade da feature;
- completude média;
- taxa de features aprovadas;
- taxa de features sensíveis;
- taxa de features com drift.

18.4 Contribuição

Esse artefato evita uso de features sem avaliação de qualidade, estabilidade ou risco.

19. Artefato A14.8 — Registro de Vazamento e Riscos de Feature

19.1 Finalidade

O **Registro de Vazamento e Riscos de Feature** identifica features com risco de vazamento de informação, uso inadequado, proxy sensível, instabilidade ou interpretação equivocada.

19.2 Estrutura mínima

Campo	Descrição
Feature	Nome da feature
Tipo de risco	Leakage, proxy, sensibilidade, instabilidade etc.
Evidência	Motivo da identificação
Severidade	Baixa, média, alta ou crítica
Impacto	Consequência possível
Ação recomendada	Remover, restringir, revisar, monitorar
Status	Aberto, mitigado, aceito, rejeitado
Responsável	Pessoa ou equipe

19.3 Métricas associadas

- taxa de features com vazamento;
- taxa de features com risco de proxy;
- taxa de riscos mitigados;
- número de features rejeitadas por risco;
- severidade média dos riscos.

19.4 Contribuição

Esse artefato reduz risco de modelos artificialmente bons, decisões enviesadas ou uso inadequado de atributos sensíveis.

20. Artefato A14.9 — Feature Store

20.1 Finalidade

A **Feature Store** é o repositório estruturado de features aprovadas, documentadas, versionadas, governadas e reutilizáveis.

20.2 Estrutura mínima

Componente	Descrição
Repositório offline	Features para treinamento, análise e histórico
Repositório online	Features para consumo em tempo real ou APIs, quando aplicável
Catálogo de features	Definições e metadados
Registro de versões	Histórico das features
Registro de qualidade	Métricas e status
Registro de lineage	Origem e regras
Controle de acesso	Permissões e restrições
Monitoramento	Estabilidade e drift

20.3 Métricas associadas

- taxa de features reutilizadas;
- taxa de features versionadas;
- taxa de features monitoradas;
- número de consumidores por feature;
- taxa de features depreciadas.

20.4 Contribuição

Esse artefato transforma features em ativos governados e reutilizáveis, reduzindo inconsistências e retrabalho.

21. Artefato A14.10 — Manifesto de Feature Set

21.1 Finalidade

O **Manifesto de Feature Set** documenta o conjunto de features utilizado em uma análise, modelo, dashboard, API ou produto de dados.

21.2 Estrutura mínima

Campo	Descrição
Feature set	Nome do conjunto
Finalidade	Uso previsto
Features incluídas	Lista de features
Versões	Versões de cada feature
Unidade de análise	Entidade representada
Janela temporal	Quando aplicável
Variável-alvo	Quando aplicável
Qualidade	Status das features
Restrições	Sensibilidade ou acesso
Consumidor	Modelo, dashboard, relatório ou API
Data de criação	Momento da geração

21.3 Métricas associadas

- taxa de feature sets documentados;
- taxa de features com lineage;
- taxa de features versionadas;
- completude do manifesto;
- índice de qualidade do feature set.

21.4 Contribuição

Esse artefato garante que análises, modelos e produtos possam ser auditados e reproduzidos a partir das features utilizadas.

22. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Datasets tratados, integrados, validados e armazenados;

	representações tabulares, temporais, espaciais, textuais, semânticas e em grafos; atributos, eventos, relações, metadados, qualidade, linhagem, versões e requisitos analíticos.
Processamento	Identificação de necessidades, seleção de atributos, criação de features tabulares, temporais e relacionais, avaliação de qualidade, verificação de vazamento, documentação, armazenamento e disponibilização na feature store.
Artefatos	Plano de feature engineering, catálogo de features, regras de cálculo, matriz de features, features temporais, features relacionais/grafos, relatório de qualidade, registro de riscos, feature store e manifesto de feature set.
Métricas	NFC, TFD, TFL, TFV, TFA, TFR, CMF, ETF, TFDri, TFS, TFRP, TFVaz, RF, CPM, IQF e TFM.
Saídas	Features documentadas, validadas, versionadas, rastreáveis, governadas, armazenadas e reutilizáveis para modelos, dashboards, APIs, análises, grafos, indicadores e agentes inteligentes.

23. Critérios gerais de validação do módulo

O Módulo 14 pode ser considerado adequadamente implementado quando:

1. as necessidades analíticas foram identificadas;
2. a unidade de análise das features foi definida;
3. os atributos de origem foram documentados;
4. as regras de cálculo foram registradas;
5. as features possuem versionamento;
6. a qualidade das features foi avaliada;
7. riscos de vazamento foram verificados;
8. features sensíveis ou proxies foram identificadas;
9. features aprovadas foram armazenadas na feature store;
10. features possuem lineage suficiente para auditoria;
11. features usadas em modelos, dashboards ou APIs possuem manifesto;
12. features críticas são monitoradas quanto a drift, estabilidade e uso.

24. Relação com o pipeline principal

O Módulo 14 se relaciona com todos os módulos do pipeline principal.

Com o **Módulo 1 — Fontes e Inventário de Dados**, utiliza atributos, entidades, granularidade e fontes candidatas para criação de features.

Com o **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, utiliza snapshots e dados brutos para garantir reprocessabilidade.

Com o **Módulo 3 — Integração e Harmonização**, utiliza relações entre fontes e entidades integradas para criar features combinadas.

Com o **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, utiliza dados tratados e também formaliza atributos derivados mais complexos.

Com o **Módulo 5 — Qualidade e Validação de Dados**, utiliza métricas de qualidade para aprovar ou rejeitar features.

Com o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, avalia risco de viés, proxies sensíveis e impacto por grupo.

Com o **Módulo 7 — Armazenamento Analítico**, utiliza data marts e estruturas persistidas como base para features.

Com o **Módulo 8 — Disponibilização e Consumo Analítico**, fornece features para produtos, dashboards, APIs, modelos e agentes inteligentes.

25. Relação com os demais módulos transversais

O Módulo 14 depende do **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, pois cada feature deve possuir definição, descrição, conceito associado, domínio, finalidade e documentação semântica.

Relaciona-se com o **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, pois features podem ser sensíveis, restritas ou funcionar como proxies de atributos protegidos.

Relaciona-se com o **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, pois cada feature deve apontar para seus atributos de origem, regras, versões e processos.

Relaciona-se com o **Módulo 12 — Versionamento e Reprodutibilidade**, pois features devem possuir versão, parâmetros, ambiente e manifestos de execução.

Relaciona-se com o **Módulo 13 — Representações Analíticas**, pois features podem ser derivadas de tabelas, séries temporais, eventos, grafos e grafos temporais.

Relaciona-se com o módulo de **Monitoramento, Drift e Evolução dos Ativos**, pois features devem ser monitoradas quanto a estabilidade, drift, degradação, obsolescência e alteração de distribuição.

Relaciona-se com o módulo de **Agentes Inteligentes de Apoio ao Pipeline**, pois agentes podem consultar o catálogo de features, sugerir novas features, explicar regras de cálculo, detectar redundâncias e recomendar reuso.

26. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 14 pode criar features para análise de evasão, desempenho, permanência ou engajamento estudantil.

Exemplos de features tabulares incluem idade no ingresso, escola de origem, curso, turno, forma de ingresso, renda familiar, média geral, número de reprovações e frequência média.

Exemplos de features temporais incluem evolução da média por semestre, frequência acumulada nos últimos períodos, número de reprovações recentes, tempo desde a última matrícula ativa, variação de desempenho e sequência de situações acadêmicas.

Exemplos de features relacionais ou de grafos incluem número de disciplinas conectadas ao estudante, centralidade em rede de interações no AVA, grau em grafo estudante-disciplina, proximidade com estudantes de mesmo curso, participação em comunidades acadêmicas ou embeddings de nós.

Essas features poderiam ser armazenadas em uma feature store educacional, documentadas com regras, versões, qualidade, sensibilidade e riscos de uso.

27. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, o Módulo 14 pode criar features para análise de vulnerabilidade, cobertura de serviços, risco territorial ou priorização de atendimento.

Exemplos de features tabulares incluem renda média territorial, número de equipamentos públicos, taxa de atendimento, índice de escolaridade, cobertura de saúde e indicadores socioeconômicos.

Exemplos de features temporais incluem evolução mensal de atendimentos, crescimento da demanda, frequência de ocorrência por território, tendência de vulnerabilidade e sazonalidade de serviços.

Exemplos de features relacionais ou de grafos incluem centralidade de equipamentos públicos, distância relacional entre famílias e serviços, número de vínculos com políticas públicas, comunidades territoriais e persistência de atendimentos ao longo do tempo.

Essas features poderiam alimentar painéis, modelos de priorização, indicadores territoriais ou APIs de apoio à gestão pública.

28. Contribuição técnico-científica do módulo

A contribuição do Módulo 14 está em formalizar feature engineering e feature store como componentes transversais da Engenharia Analítica de Dados.

Em vez de tratar features como colunas criadas informalmente em scripts isolados, o UnespDataLens-RM propõe que elas sejam tratadas como ativos analíticos documentados, avaliados, versionados, rastreáveis, governados e reutilizáveis.

Essa formalização contribui para:

- aumentar o valor analítico dos dados;
- reduzir duplicidade e inconsistência de cálculos;
- melhorar reprodutibilidade de modelos e dashboards;
- documentar origem e regra de cada feature;
- avaliar qualidade, estabilidade e risco de vazamento;
- apoiar análise tabular, temporal e relacional;
- permitir reutilização de features em diferentes produtos;
- controlar features sensíveis ou com risco de proxy;
- fortalecer a integração entre dados, modelos, dashboards e agentes inteligentes;
- estruturar uma base reutilizável de atributos analíticos no ecossistema UnespDataLens-RM.

29. Síntese do módulo

O **Módulo Transversal 14 — Feature Engineering e Feature Store** organiza o processo de criação, documentação, validação, armazenamento e reutilização de atributos analíticos no UnespDataLens-RM. Ele transforma variáveis derivadas, indicadores, métricas temporais, métricas relacionais e embeddings em ativos formais, governados e rastreáveis.

Sua importância é transversal porque features podem ser produzidas a partir de diferentes representações e consumidas por diversos produtos, como modelos, dashboards, APIs, relatórios, grafos e agentes inteligentes.

Ao formalizar feature engineering e feature store por meio de artefatos, métricas, regras, manifestos e critérios de validação, o módulo fortalece a qualidade, a reprodutibilidade, a governança e o reuso dos atributos analíticos produzidos pelo UnespDataLens-RM.

Módulo Transversal 15 — Monitoramento, Drift e Evolução dos Ativos

1. Finalidade do módulo

O **Módulo Transversal 15 — Monitoramento, Drift e Evolução dos Ativos** tem como finalidade acompanhar, avaliar e registrar continuamente o comportamento, a estabilidade, a qualidade, a atualização, a disponibilidade, o uso, a evolução e as mudanças dos ativos analíticos produzidos pelo modelo de referência **UnespDataLens-RM**.

No contexto do UnespDataLens-RM, os ativos analíticos não são elementos estáticos. Fontes de dados podem mudar, schemas podem ser alterados, atributos podem deixar de ser preenchidos, categorias podem ser renomeadas, indicadores podem mudar de comportamento, features podem perder estabilidade, rótulos podem sofrer alteração de distribuição, modelos podem degradar, dashboards podem ficar desatualizados, APIs podem falhar, produtos podem deixar de ser usados e relações em grafos podem evoluir ao longo do tempo.

Por isso, o Módulo 15 atua como um componente transversal de acompanhamento contínuo da saúde do ecossistema analítico. Seu papel é identificar mudanças relevantes, detectar problemas emergentes, sinalizar riscos, apoiar ações corretivas, documentar a evolução dos ativos e fornecer evidências para governança, qualidade, reprodutibilidade, observabilidade, DataOps, MLOps e tomada de decisão baseada em dados.

Esse módulo também explicita que o UnespDataLens-RM deve contemplar dados históricos, temporais, dinâmicos, eventuais, incrementais e, quando necessário, dados em tempo quase real ou streaming. O monitoramento deve considerar não apenas a existência dos dados, mas sua evolução ao longo do tempo, sua latência de atualização, sua estabilidade estrutural, sua distribuição estatística, sua validade semântica e seu impacto sobre produtos analíticos.

Assim, o Módulo 15 transforma o acompanhamento dos dados em um processo sistemático, documentado, mensurável e acionável, evitando que falhas, degradações ou mudanças silenciosas comprometam a confiabilidade dos resultados produzidos pelo UnespDataLens-RM.

2. Objetivo geral

O objetivo geral do Módulo 15 é **monitorar continuamente os ativos analíticos do UnespDataLens-RM quanto à qualidade, atualização, disponibilidade, estabilidade, distribuição, schema, features, rótulos, relações, uso, desempenho e evolução**, detectando drift, falhas, obsolescência, degradações e mudanças relevantes que possam impactar análises, modelos, dashboards, APIs, relatórios, grafos, agentes inteligentes e decisões baseadas em evidências.

3. Objetivos específicos

São objetivos específicos do Módulo 15:

1. definir quais ativos analíticos devem ser monitorados;
2. estabelecer métricas, baselines, limiares e critérios de alerta para cada tipo de ativo;
3. acompanhar a atualização, disponibilidade e latência dos dados;
4. detectar falhas de ingestão, atrasos, lacunas temporais e interrupções de fluxo;
5. monitorar mudanças de volume, completude, consistência, validade e atualidade;
6. detectar alterações estruturais, como schema drift, mudança de tipos, remoção ou inclusão de atributos;
7. acompanhar mudanças nas distribuições estatísticas dos dados, caracterizando data drift;
8. monitorar estabilidade, completude e comportamento de features, caracterizando feature drift;
9. acompanhar mudanças em rótulos, classes, desfechos e variáveis-alvo, caracterizando label drift;
10. identificar mudanças na relação entre variáveis de entrada e desfechos, caracterizando concept drift;
11. monitorar grafos, relações, arestas, centralidades, comunidades e estruturas temporais;
12. acompanhar uso, acesso, reutilização e obsolescência de datasets, dashboards, APIs, features e produtos;
13. registrar incidentes, alertas, causas prováveis, impactos e ações corretivas;
14. apoiar processos de DataOps e MLOps;
15. fornecer evidências para agentes inteligentes explicarem alertas, riscos, falhas e recomendações;
16. apoiar a evolução controlada dos ativos analíticos ao longo do tempo.

4. Fundamentação conceitual do módulo

O monitoramento de ativos analíticos parte do reconhecimento de que dados, pipelines, produtos, features, modelos e representações evoluem continuamente. Essa evolução pode ser esperada, como o crescimento natural do volume de registros, ou inesperada, como a alteração silenciosa de um schema. Pode ser benéfica, como a inclusão de novos atributos relevantes, ou problemática, como a perda de completude em um campo crítico.

O conceito de **drift** refere-se a mudanças relevantes no comportamento dos dados ou dos modelos ao longo do tempo. No UnespDataLens-RM, o drift deve ser interpretado de forma ampla, abrangendo mudanças estruturais, estatísticas, temporais, semânticas, relacionais e preditivas.

Os principais tipos de drift contemplados são:

Tipo de drift	Descrição	Exemplo
Data Drift	Mudança na distribuição dos dados de entrada.	A idade média dos estudantes ingressantes muda ao longo dos anos.
Schema Drift	Mudança na estrutura dos dados.	Uma coluna é removida, renomeada ou muda de tipo.

Feature Drift	Mudança no comportamento, completude ou distribuição de features.	Uma feature de engajamento deixa de ser preenchida adequadamente.
Label Drift	Mudança na distribuição ou definição da variável-alvo.	A proporção de estudantes evadidos muda abruptamente após alteração de regra.
Concept Drift	Mudança na relação entre atributos de entrada e desfecho.	O padrão que explicava evasão deixa de ser válido após nova política institucional.
Semantic Drift	Mudança no significado de um atributo, categoria ou indicador.	“Trancamento” passa a ser tratado administrativamente de forma diferente.
Relational Drift	Mudança na estrutura de relações entre entidades.	A rede de interações entre estudantes e atividades muda ao longo do semestre.
Operational Drift	Mudança no comportamento operacional de pipelines e produtos.	Um pipeline passa a atrasar ou falhar com frequência.

O Módulo 15 também se relaciona com **Data Observability**, mas não se confunde totalmente com ela. O monitoramento acompanha métricas, estados, alertas e mudanças. A observabilidade amplia esse acompanhamento ao permitir compreender causas, impactos, dependências e contexto dos problemas observados.

No UnespDataLens-RM, monitoramento e observabilidade devem atuar de forma integrada: o monitoramento detecta que algo mudou; a observabilidade ajuda a explicar por que mudou, o que foi afetado e quais ações devem ser tomadas.

5. Escopo do monitoramento

O Módulo 15 pode monitorar diferentes tipos de ativos do UnespDataLens-RM.

Tipo de ativo	Aspectos monitorados

Fontes de dados	Disponibilidade, atualização, periodicidade, acesso, cobertura e mudanças de schema.
Processos de ingestão	Execuções, falhas, tempo, volume, latência, registros rejeitados e logs.
Datasets brutos	Integridade, volume, schema, versão, atualização e retenção.
Datasets integrados	Conflitos, chaves, taxa de pareamento, completude e consistência.
Datasets tratados	Qualidade, perdas, transformações, ausências, outliers e estabilidade.
Datasets curados	Atualização, disponibilidade, qualidade, consumo e adequação ao uso.
Features	Completude, distribuição, drift, estabilidade, vazamento e uso.
Rótulos	Completude, distribuição, mudança de regra, balanceamento e label drift.
Representações temporais	Cobertura temporal, lacunas, regularidade, atrasos e evolução.
Grafos	Nós, arestas, densidade, centralidade, comunidades e relações incertas.
Grafos temporais	Persistência de arestas, surgimento de relações, desaparecimento e evolução estrutural.
Dashboards	Atualização, acesso, uso, obsolescência e inconsistências.

APIs	Disponibilidade, latência, erros, versão e volume de consumo.
Modelos	Desempenho, drift, fairness, explicabilidade, uso e necessidade de retreinamento.
Agentes inteligentes	Utilidade, erros, fontes consultadas, violações, recomendações e feedback.

6. Formalização sintética do módulo

Considerando o conjunto de ativos analíticos monitoráveis:

$$\mathbf{A = \{a_1, a_2, ..., a_n\}}$$

e o conjunto de métricas associadas a cada ativo:

$$\mathbf{M(a_i) = \{m_1, m_2, ..., m_k\}}$$

cada métrica pode ser observada ao longo do tempo:

$$\mathbf{m_j(a_i, t)}$$

em que:

- $\mathbf{a_i}$ representa o ativo monitorado;
- $\mathbf{m_j}$ representa a métrica observada;
- **t** representa o instante, período ou janela temporal de monitoramento.

Para cada ativo, pode-se definir um baseline:

$$\mathbf{B(a_i) = \{b_1, b_2, ..., b_k\}}$$

em que $\mathbf{b_j}$ representa o valor esperado, histórico ou de referência da métrica $\mathbf{m_j}$.

A variação observada pode ser representada por:

$$\mathbf{\Delta(a_i, m_j, t) = m_j(a_i, t) - b_j}$$

Quando a variação ultrapassa um limiar definido, é gerado um alerta:

alerta = {ativo, métrica, valor_observado, baseline, limiar, severidade, data, causa_provável, impacto, ação_recomendada}

Para detecção de drift, pode-se representar a comparação entre uma distribuição de referência e uma distribuição atual:

$$\mathbf{Drift(X_r, X_t) = distância(P(X_r), P(X_t))}$$

em que:

- $\mathbf{X_r}$ representa os dados de referência;
- $\mathbf{X_t}$ representa os dados observados no tempo atual;
- **P(X)** representa a distribuição dos dados;
- a distância pode ser medida por métodos estatísticos, divergências, testes ou métricas específicas.

Para grafos temporais, pode-se monitorar a evolução de uma métrica estrutural:

$$\mathbf{g(G_t) = \{g_1, g_2, ..., g_n\}}$$

em que $\mathbf{G_t}$ representa o grafo em determinado tempo e **g** representa métricas como densidade, centralidade, modularidade, componentes ou persistência de arestas.

7. Entradas do módulo

As entradas do Módulo 15 são compostas por ativos, metadados, logs, métricas, versões, baselines, contratos, regras e registros produzidos pelos demais módulos do UnespDataLens-RM.

As principais entradas são:

- inventário de fontes;
- periodicidade esperada das fontes;
- contratos de dados;
- logs de ingestão;
- logs de execução;
- manifestos de ingestão;
- datasets brutos;
- datasets integrados;
- datasets tratados;
- datasets curados;
- data marts;
- schemas esperados;
- schemas observados;
- métricas de qualidade;
- registros de validação;
- registros de falhas;
- registros de incidentes;
- metadados;
- catálogo de dados;
- glossários;
- regras de transformação;
- registros de lineage;
- versões de ativos;
- snapshots;
- feature store;
- catálogo de features;
- rótulos e variáveis-alvo;
- representações temporais;
- grafos;
- grafos temporais;
- dashboards;
- APIs;
- modelos;

- produtos analíticos;
- registros de uso;
- feedback de usuários;
- políticas de governança;
- permissões de acesso;
- baselines e limiares de alerta.

8. Processamento

O processamento do Módulo 15 envolve definição de ativos críticos, coleta de métricas, comparação com baselines, detecção de mudanças, geração de alertas, análise de impacto, registro de incidentes e recomendação de ações.

8.1 Definição dos ativos monitorados

A primeira atividade consiste em identificar quais ativos devem ser monitorados. Nem todos os ativos exigem o mesmo nível de acompanhamento. Ativos críticos, sensíveis, amplamente consumidos, utilizados por modelos ou dependentes de atualização frequente devem receber prioridade.

8.2 Definição de métricas, baselines e limiares

A segunda atividade consiste em definir, para cada ativo, quais métricas serão monitoradas, quais valores de referência serão adotados e quais limiares indicam alerta, atenção ou falha crítica.

Os baselines podem ser definidos com base em:

- valores históricos;
- médias móveis;
- janelas temporais;
- contratos de dados;
- SLAs;
- regras de qualidade;
- versões de referência;
- comportamento esperado do domínio.

8.3 Coleta contínua ou periódica de métricas

A terceira atividade consiste em coletar métricas em intervalos definidos. Essa coleta pode ocorrer em batch, incremental, near real-time ou streaming, dependendo da natureza do ativo.

Exemplos:

- volume de registros por ingestão;
- completude de atributos críticos;
- número de falhas por pipeline;
- latência de atualização;
- distribuição de uma feature;
- proporção de classes de um rótulo;
- número de nós e arestas de um grafo;
- tempo de resposta de uma API;

- frequência de acesso a um dashboard.

8.4 Detecção de mudanças e drift

A quarta atividade consiste em comparar métricas atuais com baselines ou versões anteriores, detectando variações relevantes.

Podem ser aplicados métodos como:

- comparação de médias;
- comparação de proporções;
- testes estatísticos;
- divergência entre distribuições;
- PSI;
- KS Statistic;
- análise de tendência;
- detecção de anomalias;
- comparação de schemas;
- comparação de categorias;
- métricas de estabilidade;
- métricas de grafos temporais.

8.5 Classificação da severidade

A quinta atividade consiste em classificar a severidade do problema identificado. Uma mudança pode ser classificada como:

- informativa;
- baixa;
- média;
- alta;
- crítica.

A classificação deve considerar impacto, ativo afetado, consumidores, sensibilidade, uso em modelos, uso em dashboards, abrangência temporal e risco decisório.

8.6 Análise de impacto

A sexta atividade consiste em identificar quais ativos, produtos, features, modelos, dashboards, APIs ou usuários podem ser impactados por uma mudança.

Essa análise depende diretamente do Módulo 11 — Proveniência, Linhagem e Rastreabilidade.

8.7 Geração de alertas

A sétima atividade consiste em gerar alertas estruturados, contendo ativo afetado, métrica, valor observado, baseline, severidade, data, evidência, causa provável, impacto e ação recomendada.

8.8 Registro de incidentes

A oitava atividade consiste em registrar incidentes quando uma alteração compromete qualidade, atualização, disponibilidade, estabilidade, governança ou confiabilidade de um ativo.

8.9 Recomendação de ações corretivas

A nona atividade consiste em recomendar ações como:

- reprocessar pipeline;
- revisar schema;
- atualizar contrato de dados;
- recalcular feature;
- revisar rótulo;
- bloquear publicação;
- atualizar dashboard;
- retreinar modelo;
- revisar regra de transformação;
- acionar responsável;
- classificar produto como depreciado;
- descontinuar ativo obsoleto.

8.10 Acompanhamento da evolução dos ativos

A décima atividade consiste em acompanhar a evolução histórica dos ativos, registrando crescimento, mudança, obsolescência, substituição, melhoria, degradação e uso ao longo do tempo.

9. Técnicas, métodos e tecnologias aplicáveis

O Módulo 15 pode empregar diferentes técnicas e tecnologias para monitoramento, detecção de drift, observabilidade e evolução dos ativos.

Entre as técnicas de monitoramento, destacam-se:

- monitoramento de volume;
- monitoramento de schema;
- monitoramento de completude;
- monitoramento de atualidade;
- monitoramento de disponibilidade;
- monitoramento de latência;
- monitoramento de uso;
- monitoramento de qualidade;
- monitoramento de drift;
- monitoramento de modelos;
- monitoramento de grafos temporais.

Entre os métodos de detecção de drift, podem ser utilizados:

- Population Stability Index;
- Kolmogorov-Smirnov Test;
- Jensen-Shannon Divergence;
- Kullback-Leibler Divergence;
- Wasserstein Distance;
- Chi-square Test;

- comparação de proporções;
- análise de tendência;
- detecção de anomalias;
- controle estatístico de processo;
- comparação de embeddings;
- métricas de estabilidade temporal;
- comparação de grafos por snapshots.

Entre as tecnologias aplicáveis, podem ser consideradas:

- Evidently AI;
- WhyLabs;
- Great Expectations;
- Soda Core;
- Deequ;
- Pandera;
- Prometheus;
- Grafana;
- OpenLineage;
- Marquez;
- MLflow;
- Feast;
- Airflow;
- Prefect;
- Dagster;
- dbt tests;
- DataHub;
- OpenMetadata;
- logs estruturados;
- bancos de séries temporais;
- bancos de grafos;
- ferramentas de observabilidade;
- painéis operacionais;
- sistemas de alerta.

10. Métricas do módulo

As métricas do Módulo 15 avaliam cobertura de monitoramento, qualidade, atualização, drift, estabilidade, uso, incidentes e evolução dos ativos.

Métrica	Sigla	Descrição
Taxa de ativos monitorados	TAM	Proporção de ativos críticos com monitoramento ativo.
Taxa de baselines definidos	TBD	Proporção de ativos monitorados com baseline documentado.

Taxa de métricas ativas	TMA	Proporção de métricas configuradas e coletadas periodicamente.
Taxa de atualização conforme periodicidade	TACP	Proporção de ativos atualizados no prazo esperado.
Latência média de atualização	LMA	Tempo médio entre atualização na fonte e disponibilidade no produto.
Taxa de falhas detectadas	TFD	Proporção de falhas identificadas por monitoramento.
Taxa de falhas silenciosas	TFS	Proporção de problemas identificados tardiamente ou sem alerta automático.
Taxa de data drift	TDD	Proporção de atributos ou ativos com mudança estatística relevante.
Taxa de schema drift	TSD	Proporção de ativos com alteração estrutural detectada.
Taxa de feature drift	TFDri	Proporção de features com mudança de distribuição ou estabilidade.
Taxa de label drift	TLD	Proporção de rótulos com mudança relevante de distribuição.
Taxa de concept drift	TCD	Proporção de modelos ou problemas com mudança na relação entrada-saída.
Taxa de relational drift	TRD	Proporção de grafos ou relações com mudança estrutural relevante.
Taxa de produtos obsoletos	TPO	Proporção de produtos sem atualização, uso ou

		manutenção.
Taxa de incidentes resolvidos	TIR	Proporção de incidentes tratados em determinado período.
Tempo médio de detecção	TMD	Tempo entre ocorrência do problema e sua detecção.
Tempo médio de resolução	TMR	Tempo entre abertura e resolução do incidente.
Índice de evolução controlada dos ativos	IECA	Indicador composto de estabilidade, atualização, monitoramento e resolução.

10.1 Taxa de Ativos Monitorados — TAM

A **Taxa de Ativos Monitorados** mede a proporção de ativos críticos acompanhados por métricas, baselines e alertas.

TAM = número de ativos críticos monitorados / número total de ativos críticos

Essa métrica indica a cobertura do monitoramento no ecossistema.

10.2 Taxa de Atualização Conforme Periodicidade — TACP

A **Taxa de Atualização Conforme Periodicidade** mede a proporção de ativos atualizados dentro do prazo esperado.

TACP = número de atualizações realizadas no prazo / número total de atualizações esperadas

Essa métrica é especialmente importante para produtos que dependem de atualização recorrente.

10.3 Taxa de Schema Drift — TSD

A **Taxa de Schema Drift** mede a proporção de ativos que apresentaram alteração estrutural relevante.

TSD = número de ativos com schema drift / número total de ativos monitorados quanto a schema

Essa métrica permite acompanhar estabilidade estrutural dos dados.

10.4 Taxa de Feature Drift — TFDri

A **Taxa de Feature Drift** mede a proporção de features que apresentaram mudança relevante de distribuição, completude ou estabilidade.

TFDri = número de features com drift / número total de features monitoradas

Essa métrica é relevante para modelos e feature stores.

10.5 Taxa de Label Drift — TLD

A **Taxa de Label Drift** mede a proporção de rótulos ou variáveis-alvo cuja distribuição mudou de forma relevante.

TLD = número de rótulos com drift / número total de rótulos monitorados

Essa métrica é essencial para avaliar estabilidade de modelos supervisionados e indicadores.

10.6 Índice de Evolução Controlada dos Ativos — IECA

O **Índice de Evolução Controlada dos Ativos** sintetiza a capacidade do ecossistema de acompanhar e controlar mudanças.

Uma formulação possível é:

$$\mathbf{IECA} = \mathbf{w_1TAM} + \mathbf{w_2TBD} + \mathbf{w_3TMA} + \mathbf{w_4TACP} + \mathbf{w_5TIR} - \mathbf{w_6TFS} - \mathbf{w_7TPO} - \mathbf{w_8TSD}$$

em que:

- **TAM** representa a taxa de ativos monitorados;
- **TBD** representa a taxa de baselines definidos;
- **TMA** representa a taxa de métricas ativas;
- **TACP** representa a atualização conforme periodicidade;
- **TIR** representa a taxa de incidentes resolvidos;
- **TFS** representa a taxa de falhas silenciosas;
- **TPO** representa a taxa de produtos obsoletos;
- **TSD** representa a taxa de schema drift;
- $\mathbf{w_i}$ representa o peso atribuído a cada dimensão.

Quanto maior o IECA, maior a capacidade do UnespDataLens-RM de acompanhar a evolução dos seus ativos de forma controlada.

11. Artefatos formalizados do módulo

Os artefatos do Módulo 15 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição

A15.1	Plano de Monitoramento dos Ativos	Documento técnico	TAM	Define ativos, métricas, baselines, limiares e responsáveis.
A15.2	Registro de Baselines	Tabela ou repositório de referência	TBD	Preserva valores esperados para comparação temporal.
A15.3	Painel de Monitoramento	Dashboard operacional	TMA	Acompanha saúde, atualização, drift, falhas e uso.
A15.4	Registro de Alertas	Tabela ou sistema de alertas	TFD	Documenta variações, severidade, evidências e responsáveis.
A15.5	Relatório de Data Drift	Relatório técnico	TDD	Avalia mudanças estatísticas nos dados.
A15.6	Relatório de Schema Drift	Relatório técnico	TSD	Avalia mudanças estruturais nos ativos.
A15.7	Relatório de Feature Drift	Relatório técnico	TFDri	Avalia estabilidade e distribuição das features.
A15.8	Relatório de Label e Concept Drift	Relatório técnico	TLD, TCD	Avalia mudanças em rótulos e relação entrada-saída.
A15.9	Registro de Incidentes e Ações Corretivas	Tabela ou sistema operacional	TIR	Documenta falhas, causas, impactos e tratamento.
A15.10	Relatório de Evolução dos Ativos	Relatório técnico-gerencial	IECA	Avalia mudanças, obsolescência, uso e maturidade dos ativos.

A15.11	Mapa de Impacto de Mudanças	Grafo ou matriz de dependências	TAIR	Relaciona mudanças a produtos, modelos, features e usuários afetados.
A15.12	Registro de Obsolescência e Descontinuação	Tabela de curadoria	TPO	Documenta ativos depreciados, substituídos ou retirados de uso.

12. Artefato A15.1 — Plano de Monitoramento dos Ativos

12.1 Finalidade

O **Plano de Monitoramento dos Ativos** define quais ativos serão monitorados, quais métricas serão acompanhadas, quais baselines serão utilizados, quais limiares acionarão alertas, quem será responsável e quais ações devem ser tomadas em caso de desvio.

12.2 Estrutura mínima

Campo	Descrição
Ativo monitorado	Dataset, feature, dashboard, API, modelo, grafo ou fonte.
Tipo de ativo	Categoria do ativo.
Métricas monitoradas	Métricas associadas ao ativo.
Baseline	Valor de referência.
Limiar	Valor que aciona alerta.
Periodicidade	Frequência de monitoramento.

Responsável	Pessoa, equipe ou sistema.
Severidade	Critério para classificação de alerta.
Ação recomendada	Procedimento em caso de desvio.
Relação com produtos	Produtos dependentes do ativo.

12.3 Métricas associadas

- taxa de ativos monitorados;
- taxa de baselines definidos;
- taxa de métricas ativas;
- taxa de falhas detectadas.

12.4 Critérios de validação

O plano é válido quando ativos críticos possuem métricas, baselines, limiares, responsáveis e ações associadas.

12.5 Contribuição

Esse artefato torna o monitoramento planejado, mensurável e acionável.

13. Artefato A15.2 — Registro de Baselines

13.1 Finalidade

O **Registro de Baselines** preserva valores de referência para comparação com observações futuras.

13.2 Estrutura mínima

Campo	Descrição
Ativo	Ativo associado.
Métrica	Métrica de referência.

Valor baseline	Valor esperado ou histórico.
Período de referência	Janela usada para calcular baseline.
Método de cálculo	Média, mediana, distribuição, regra, contrato etc.
Versão	Versão do baseline.
Data de criação	Momento de definição.
Responsável	Pessoa ou processo responsável.
Limite inferior	Quando aplicável.
Limite superior	Quando aplicável.

13.3 Métricas associadas

- taxa de baselines definidos;
- taxa de baselines atualizados;
- taxa de alertas com baseline associado.

13.4 Critérios de validação

O baseline é válido quando possui métrica, período de referência, método de cálculo, versão e responsável.

13.5 Contribuição

Esse artefato permite detectar mudanças e drift de forma objetiva.

14. Artefato A15.3 — Painel de Monitoramento

14.1 Finalidade

O **Painel de Monitoramento** consolida visualmente o estado dos ativos analíticos, exibindo métricas, alertas, falhas, atualizações, drift, incidentes e uso.

14.2 Estrutura mínima

Componente	Descrição
Visão geral	Estado dos ativos críticos.
Atualização	Atrasos, periodicidade e latência.
Qualidade	Completude, validade, consistência e atualidade.
Drift	Data drift, schema drift, feature drift e label drift.
Incidentes	Falhas abertas, críticas e resolvidas.
Uso	Acessos, reutilização e produtos sem uso.
Responsáveis	Equipes associadas a ativos e incidentes.
Severidade	Classificação dos alertas.

14.3 Métricas associadas

- taxa de métricas ativas;
- taxa de falhas detectadas;
- taxa de incidentes resolvidos;
- tempo médio de detecção.

14.4 Critérios de validação

O painel é válido quando permite acompanhar ativos críticos, identificar alertas e apoiar ações de resposta.

14.5 Contribuição

Esse artefato fortalece observabilidade, governança operacional e tomada de decisão técnica.

15. Artefato A15.4 — Registro de Alertas

15.1 Finalidade

O **Registro de Alertas** documenta eventos de monitoramento que indicam desvio, falha, mudança, drift ou risco.

15.2 Estrutura mínima

Campo	Descrição
ID do alerta	Código único.
Ativo afetado	Ativo relacionado.
Métrica	Métrica que gerou o alerta.
Valor observado	Valor atual.
Baseline	Valor esperado.
Limiar	Regra de disparo.
Severidade	Baixa, média, alta, crítica.
Evidência	Log, métrica, gráfico ou relatório.
Causa provável	Hipótese inicial.

Impacto	Produtos ou usuários afetados.
Responsável	Pessoa ou equipe.
Status	Aberto, em análise, resolvido, aceito.

15.3 Métricas associadas

- taxa de alertas com evidência;
- taxa de falhas detectadas;
- tempo médio de resposta;
- taxa de alertas resolvidos.

15.4 Critérios de validação

O alerta é válido quando possui métrica, evidência, severidade, ativo afetado, responsável e status.

15.5 Contribuição

Esse artefato transforma variações detectadas em itens rastreáveis de gestão operacional.

16. Artefato A15.5 — Relatório de Data Drift

16.1 Finalidade

O **Relatório de Data Drift** avalia mudanças estatísticas nos dados de entrada em relação a uma referência.

16.2 Estrutura mínima

Campo	Descrição
Ativo avaliado	Dataset ou feature set.
Atributos avaliados	Variáveis monitoradas.
Distribuição de referência	Período ou versão base.

Distribuição atual	Período observado.
Métrica de drift	PSI, KS, divergência etc.
Atributos com drift	Lista de variáveis afetadas.
Severidade	Baixa, média, alta, crítica.
Interpretação	Explicação do resultado.
Recomendação	Ação sugerida.

16.3 Métricas associadas

- taxa de data drift;
- número de atributos com drift;
- índice de estabilidade dos atributos;
- taxa de ativos com drift crítico.

16.4 Critérios de validação

O relatório é válido quando compara distribuições de referência e atual, identifica variáveis afetadas e sugere ações.

16.5 Contribuição

Esse artefato apoia estabilidade analítica, modelos preditivos e interpretação temporal dos dados.

17. Artefato A15.6 — Relatório de Schema Drift

17.1 Finalidade

O **Relatório de Schema Drift** documenta mudanças estruturais em atributos, tipos, domínios, chaves, categorias ou formatos.

17.2 Estrutura mínima

Campo	Descrição

Ativo avaliado	Fonte, dataset, tabela, API ou produto.
Schema esperado	Estrutura de referência.
Schema observado	Estrutura atual.
Mudanças detectadas	Adição, remoção, renomeação, tipo, domínio etc.
Severidade	Grau de impacto.
Produtos afetados	Dashboards, modelos, APIs, features etc.
Ação recomendada	Ajuste, bloqueio, versionamento, comunicação.

17.3 Métricas associadas

- taxa de schema drift;
- taxa de mudanças críticas;
- tempo médio de resolução;
- taxa de produtos impactados.

17.4 Critérios de validação

O relatório é válido quando identifica alterações estruturais e seus impactos.

17.5 Contribuição

Esse artefato reduz falhas silenciosas causadas por mudanças de estrutura.

18. Artefato A15.7 — Relatório de Feature Drift

18.1 Finalidade

O **Relatório de Feature Drift** avalia mudanças em features utilizadas por modelos, dashboards, indicadores ou análises.

18.2 Estrutura mínima

Campo	Descrição
Feature avaliada	Nome e versão.
Feature set	Conjunto associado.
Distribuição de referência	Período ou versão base.
Distribuição atual	Período observado.
Completude	Valores ausentes ou inválidos.
Métrica de drift	Método utilizado.
Impacto potencial	Modelo ou produto afetado.
Recomendação	Revisar, recalcular, substituir ou monitorar.

18.3 Métricas associadas

- taxa de feature drift;
- taxa de features instáveis;
- taxa de features com completude degradada;
- taxa de modelos impactados.

18.4 Critérios de validação

O relatório é válido quando avalia estabilidade, completude, distribuição e impacto das features.

18.5 Contribuição

Esse artefato apoia feature store, MLOps, qualidade de modelos e reprodutibilidade.

19. Artefato A15.8 — Relatório de Label e Concept Drift

19.1 Finalidade

O **Relatório de Label e Concept Drift** avalia mudanças nos rótulos, classes, desfechos e na relação entre atributos de entrada e variável-alvo.

19.2 Estrutura mínima

Campo	Descrição
Rótulo avaliado	Variável-alvo ou desfecho.
Distribuição de referência	Distribuição histórica.
Distribuição atual	Distribuição observada.
Mudança de regra	Se houve alteração semântica ou operacional.
Métrica de drift	Método utilizado.
Evidência de concept drift	Mudança na relação entrada-saída.
Modelos afetados	Modelos dependentes do rótulo.
Recomendação	Revisar rótulo, retreinar modelo, ajustar regra.

19.3 Métricas associadas

- taxa de label drift;
- taxa de concept drift;
- índice de qualidade de rótulos;
- taxa de modelos com degradação.

19.4 Critérios de validação

O relatório é válido quando identifica mudanças relevantes nos rótulos ou no conceito modelado.

19.5 Contribuição

Esse artefato fortalece Label Quality, MLOps e avaliação contínua de modelos.

20. Artefato A15.9 — Registro de Incidentes e Ações Corretivas

20.1 Finalidade

O **Registro de Incidentes e Ações Corretivas** documenta problemas detectados, impactos, causas prováveis, responsáveis, decisões e soluções aplicadas.

20.2 Estrutura mínima

Campo	Descrição
ID do incidente	Código único.
Alerta associado	Referência ao alerta.
Ativo afetado	Dataset, pipeline, feature, modelo, dashboard etc.
Tipo de incidente	Falha, drift, atraso, obsolescência, qualidade etc.
Causa provável	Hipótese ou diagnóstico.
Impacto	Produtos, usuários ou decisões afetadas.
Ação corretiva	Procedimento realizado.
Responsável	Pessoa ou equipe.

Prazo	Data esperada de resolução.
Status	Aberto, resolvido, aceito, recorrente.

20.3 Métricas associadas

- taxa de incidentes resolvidos;
- tempo médio de resolução;
- taxa de recorrência de incidentes;
- taxa de ações corretivas concluídas.

20.4 Critérios de validação

O registro é válido quando permite acompanhar o ciclo completo do incidente, da detecção à resolução.

20.5 Contribuição

Esse artefato permite gestão operacional, melhoria contínua e redução de dívida de dados.

21. Artefato A15.10 — Relatório de Evolução dos Ativos

21.1 Finalidade

O **Relatório de Evolução dos Ativos** documenta como fontes, datasets, features, modelos, dashboards, APIs, grafos e produtos evoluíram ao longo do tempo.

21.2 Estrutura mínima

Campo	Descrição
Ativo	Identificador.
Período analisado	Janela temporal.
Mudanças observadas	Volume, schema, qualidade, uso, drift etc.
Tendências	Crescimento, queda, estabilidade ou obsolescência.

Incidentes relacionados	Falhas ou alertas associados.
Uso	Acessos, reutilização, abandono.
Recomendações	Manter, revisar, atualizar, depreciar ou descontinuar.

21.3 Métricas associadas

- índice de evolução controlada dos ativos;
- taxa de produtos obsoletos;
- taxa de ativos atualizados;
- taxa de reutilização.

21.4 Critérios de validação

O relatório é válido quando apresenta histórico, interpretação e recomendações sobre a evolução dos ativos.

21.5 Contribuição

Esse artefato apoia governança, curadoria, planejamento e sustentabilidade do ecossistema.

22. Artefato A15.11 — Mapa de Impacto de Mudanças

22.1 Finalidade

O **Mapa de Impacto de Mudanças** identifica quais ativos, produtos, modelos, features, dashboards, APIs e usuários são afetados por determinada alteração.

22.2 Estrutura mínima

Campo	Descrição
Mudança detectada	Drift, falha, schema, atraso etc.
Ativo de origem	Fonte ou ativo alterado.

Ativos dependentes	Datasets, features, modelos ou produtos.
Tipo de dependência	Direta ou indireta.
Severidade	Grau de impacto.
Usuários afetados	Consumidores ou áreas impactadas.
Ação recomendada	Revisão, bloqueio, comunicação, reprocessamento.

22.3 Métricas associadas

- taxa de ativos com impacto rastreável;
- número de produtos impactados;
- tempo médio de análise de impacto;
- taxa de mudanças comunicadas.

22.4 Critérios de validação

O mapa é válido quando permite rastrear dependências e estimar impacto de mudanças.

22.5 Contribuição

Esse artefato conecta monitoramento com lineage e governança de mudanças.

23. Artefato A15.12 — Registro de Obsolescência e Descontinuação

23.1 Finalidade

O **Registro de Obsolescência e Descontinuação** documenta ativos que se tornaram desatualizados, redundantes, pouco utilizados, substituídos ou inadequados.

23.2 Estrutura mínima

Campo	Descrição
Ativo	Identificador.

Tipo	Dataset, dashboard, API, feature, modelo etc.
Motivo	Baixo uso, substituição, desatualização, risco, duplicidade.
Produto substituto	Quando aplicável.
Data de marcação	Data de obsolescência.
Responsável	Pessoa ou equipe.
Status	Ativo, depreciado, obsoleto, descontinuado.
Comunicação	Usuários notificados.
Preservação histórica	Se deve ser mantido para auditoria.

23.3 Métricas associadas

- taxa de produtos obsoletos;
- taxa de ativos depreciados;
- taxa de ativos substituídos;
- taxa de ativos sem uso.

23.4 Critérios de validação

O registro é válido quando ativos obsoletos são identificados, justificados e tratados de forma controlada.

23.5 Contribuição

Esse artefato reduz data overload, data debt e risco de uso de produtos desatualizados.

24. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Ativos analíticos, logs, metadados, baselines, contratos, versões, schemas, métricas de qualidade, features, rótulos, modelos, grafos, produtos, APIs, dashboards e registros de uso.
Processamento	Definição de ativos críticos, coleta de métricas, comparação com baselines, detecção de drift, geração de alertas, análise de impacto, registro de incidentes e recomendação de ações.
Artefatos	Plano de monitoramento, baselines, painel, alertas, relatórios de drift, incidentes, mapa de impacto, relatório de evolução e registro de obsolescência.
Métricas	TAM, TBD, TMA, TACP, LMA, TFD, TFS, TDD, TSD, TFDri, TLD, TCD, TRD, TPO, TIR, TMD, TMR e IECA.
Saídas	Alertas, relatórios de drift, incidentes, recomendações, ações corretivas, registros de evolução, produtos depreciados e evidências para governança, DataOps, MLOps e agentes inteligentes.

25. Critérios gerais de validação do módulo

O Módulo 15 pode ser considerado adequadamente implementado quando:

1. ativos críticos estão identificados;
2. ativos críticos possuem métricas de monitoramento definidas;
3. baselines e limiares estão documentados;
4. métricas são coletadas periodicamente ou continuamente;
5. falhas de atualização são detectadas;
6. atrasos e latências são monitorados;
7. schema drift é identificado e registrado;
8. data drift é avaliado para atributos relevantes;
9. feature drift é monitorado para features críticas;
10. label drift é monitorado para variáveis-alvo relevantes;
11. concept drift é avaliado em modelos em uso;
12. grafos e grafos temporais possuem métricas estruturais acompanhadas quando aplicável;
13. alertas possuem evidências e severidade;
14. incidentes são registrados e acompanhados;
15. lineage permite identificar produtos impactados;
16. ações corretivas são documentadas;
17. ativos obsoletos ou sem uso são sinalizados;

18. agentes inteligentes podem consultar registros para explicar alertas e recomendações;
19. a evolução dos ativos pode ser analisada historicamente.

26. Relação com os demais módulos do modelo de referência

O Módulo 15 se relaciona com todos os módulos do UnespDataLens-RM.

Com o **Módulo 1 — Fontes e Inventário de Dados**, acompanha disponibilidade, periodicidade, cobertura e mudanças das fontes.

Com o **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, monitora falhas de ingestão, logs, volume, latência, janelas temporais, snapshots e eventos.

Com o **Módulo 3 — Integração e Harmonização**, acompanha taxa de pareamento, conflitos, chaves, integração e estabilidade entre fontes.

Com o **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, monitora perdas, transformações, atributos derivados, mudanças de distribuição e efeitos de regras.

Com o **Módulo 5 — Qualidade e Validação de Dados**, utiliza métricas de completude, validade, consistência, unicidade, integridade e atualidade.

Com o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, acompanha mudanças em grupos, representatividade, fairness e distribuição de classes.

Com o **Módulo 7 — Armazenamento Analítico**, monitora disponibilidade, atualização, crescimento, particionamento, desempenho e integridade dos ativos persistidos.

Com o **Módulo 8 — Disponibilização e Consumo Analítico**, acompanha uso, atualização, obsolescência e disponibilidade de dashboards, APIs e produtos.

Com o **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, utiliza metadados para contextualizar alertas e atualizar status dos ativos.

Com o **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, fornece evidências para auditoria, conformidade, acesso e riscos.

Com o **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, utiliza lineage para análise de impacto.

Com o **Módulo 12 — Versionamento e Reprodutibilidade**, compara versões, snapshots, schemas e execuções.

Com o **Módulo 13 — Representações Analíticas**, monitora representações tabulares, temporais, eventos, grafos e grafos temporais.

Com o **Módulo 14 — Feature Engineering e Feature Store**, acompanha estabilidade, drift e uso de features.

Com o **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, fornece alertas, incidentes, métricas e explicações que podem ser consultados por agentes.

27. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, o Módulo 15 pode monitorar bases de matrícula, histórico acadêmico, notas, frequência, interações em ambiente virtual, rótulos de evasão, features educacionais, dashboards institucionais e modelos preditivos.

Por exemplo, se a base de interações do ambiente virtual apresentar queda abrupta no volume de eventos em determinado período, o monitoramento pode gerar um alerta, comparar o volume atual com o baseline histórico, verificar a janela temporal da ingestão, consultar logs do Módulo 2, identificar produtos afetados por lineage e recomendar reprocessamento ou revisão da fonte.

Outro exemplo ocorre quando a distribuição do rótulo “evasão” muda após alteração administrativa na regra de classificação. Nesse caso, o módulo pode detectar label drift, registrar a mudança, indicar impacto em modelos preditivos e recomendar revisão da regra de rotulagem ou retreinamento do modelo.

Também é possível monitorar grafos temporais de interação entre estudantes e atividades, acompanhando mudanças de centralidade, isolamento, densidade e persistência de relações ao longo do semestre.

28. Exemplo de aplicação em políticas públicas

Em políticas públicas, o Módulo 15 pode monitorar bases de atendimento, saúde, educação, assistência social, território, indicadores socioeconômicos, dashboards territoriais, APIs e modelos de priorização.

Por exemplo, se uma secretaria deixar de enviar dados mensais de atendimento, o monitoramento pode identificar ausência de atualização, gerar alerta, registrar incidente, consultar contratos de dados, identificar indicadores e dashboards afetados e recomendar acionamento do responsável pela fonte.

Outro exemplo ocorre quando uma categoria de atendimento muda de nome ou deixa de existir. O módulo pode detectar schema drift ou semantic drift, indicar produtos impactados, acionar curadoria semântica e recomendar atualização das regras de harmonização.

29. Contribuição técnico-científica do módulo

A contribuição técnico-científica do Módulo 15 está em formalizar o monitoramento contínuo dos ativos analíticos como parte essencial da Engenharia Analítica de Dados.

Em vez de considerar que o pipeline termina na disponibilização dos produtos, o UnespDataLens-RM reconhece que dados, features, modelos, grafos, dashboards e APIs continuam evoluindo após sua publicação. Portanto, precisam ser acompanhados, avaliados, explicados, corrigidos, atualizados ou descontinuados ao longo do tempo.

Essa formalização contribui para:

- detectar falhas silenciosas;
- monitorar dados temporais, eventos e fluxos dinâmicos;
- acompanhar data drift, schema drift, feature drift, label drift e concept drift;
- avaliar evolução de grafos e grafos temporais;
- reduzir obsolescência de produtos;

- apoiar DataOps e MLOps;
- fortalecer observabilidade e governança;
- aumentar confiança nos produtos analíticos;
- reduzir data debt e data overload;
- fornecer evidências para auditoria e tomada de decisão;
- permitir resposta rápida a incidentes;
- sustentar evolução controlada do ecossistema analítico.

30. Síntese do módulo

O **Módulo Transversal 15 — Monitoramento, Drift e Evolução dos Ativos** acompanha a saúde, estabilidade, atualização, uso, qualidade e evolução dos ativos analíticos do UnespDataLens-RM.

Sua importância está em reconhecer que dados, pipelines, features, rótulos, grafos, modelos, dashboards e produtos analíticos mudam ao longo do tempo. Essas mudanças podem comprometer análises, indicadores, modelos e decisões se não forem detectadas, documentadas e tratadas.

Ao incorporar métricas, baselines, alertas, relatórios de drift, registros de incidentes, análise de impacto, monitoramento de uso, controle de obsolescência e integração com DataOps, MLOps e agentes inteligentes, o Módulo 15 fortalece a capacidade do UnespDataLens-RM de manter ativos analíticos confiáveis, governados, rastreáveis, evolutivos e adequados à tomada de decisão baseada em evidências.

Módulo Transversal 16 — Agentes Inteligentes de Apoio ao Pipeline

1. Finalidade do módulo

O **Módulo Transversal 16 — Agentes Inteligentes de Apoio ao Pipeline** tem como finalidade especificar, integrar, controlar, avaliar e monitorar agentes inteligentes capazes de apoiar usuários, analistas, pesquisadores, gestores e equipes técnicas na navegação, compreensão, documentação, validação, explicação, recomendação e manutenção dos processos do modelo de referência **UnespDataLens-RM**.

Esse módulo atua de forma transversal porque os agentes inteligentes podem interagir com diferentes etapas do pipeline principal e com diferentes artefatos produzidos pelos módulos transversais. Eles podem apoiar o inventário de fontes, a consulta ao catálogo, a geração de documentação, a explicação da linhagem dos dados, a interpretação de métricas de qualidade, a identificação de problemas de governança, a recomendação de transformações, a análise de drift, a busca de produtos de dados, a explicação de features, a interpretação de dashboards, a identificação de riscos e a orientação sobre restrições de acesso.

No contexto do UnespDataLens-RM, os agentes inteligentes não devem ser compreendidos como substitutos dos módulos técnicos, nem como mecanismos autônomos de decisão sem controle humano. Eles são concebidos como componentes de apoio, operando sobre metadados, catálogos, glossários, logs, registros de qualidade, regras, versões, grafos de linhagem, features, relatórios, produtos de dados, políticas de governança e registros de monitoramento. Seu papel é ampliar a capacidade de consulta, explicação, recomendação e documentação do ecossistema analítico, mantendo rastreabilidade, validação, segurança e conformidade.

Os agentes podem assumir diferentes funções. Um agente pode auxiliar na descoberta de fontes de dados. Outro pode explicar a origem de um indicador. Outro pode identificar datasets com queda de qualidade. Outro pode sugerir revisão de features com drift. Outro pode explicar quais dashboards dependem de uma fonte alterada. Outro pode apoiar a geração de fichas de documentação. Outro pode orientar usuários sobre restrições de acesso, finalidade de uso e limitações de um produto de dados.

Entretanto, o uso de agentes inteligentes em pipelines de dados exige cuidados metodológicos, técnicos e éticos. As respostas geradas devem ser fundamentadas em evidências registradas nos módulos do modelo de referência, como metadados, lineage, versões, logs, catálogos, relatórios de qualidade, políticas de governança e registros de monitoramento. O agente deve indicar, sempre que possível, a origem da informação utilizada, sinalizar incertezas, respeitar permissões de acesso e evitar inferências não fundamentadas sobre dados sensíveis, incompletos ou não autorizados.

Assim, o Módulo 16 formaliza os agentes inteligentes como componentes transversais de apoio ao pipeline, orientados por evidências, governança, rastreabilidade, controle humano, versionamento, segurança e critérios de confiabilidade.

2. Objetivo geral

O objetivo geral do Módulo 16 é **integrar agentes inteligentes ao UnespDataLens-RM como mecanismos de apoio à consulta, explicação, documentação, recomendação, validação e manutenção dos ativos analíticos**, garantindo que suas respostas, sugestões e ações sejam baseadas em metadados, evidências, regras, lineage, versões, políticas de governança, registros de monitoramento e validação humana quando necessário.

3. Objetivos específicos

São objetivos específicos do Módulo 16:

1. definir os tipos de agentes inteligentes que poderão atuar no apoio ao pipeline analítico;
2. especificar funções, escopos, permissões, limites e responsabilidades de cada agente;
3. integrar agentes aos catálogos de dados, metadados, glossários, dicionários, modelos semânticos, registros de qualidade, lineage, versionamento, feature store e monitoramento;
4. permitir consultas em linguagem natural sobre fontes, datasets, atributos, produtos de dados, dashboards, APIs, features, grafos, indicadores, regras e versões;
5. apoiar a explicação de indicadores, features, datasets e produtos analíticos a partir de sua origem, regras, versões, métricas de qualidade e limitações;
6. recomendar ações de melhoria, revisão, reprocessamento, documentação, validação, monitoramento ou governança com base em evidências registradas;
7. auxiliar na geração, revisão e atualização de documentação técnica, dicionários, fichas de ativos, relatórios, descrições de produtos e registros de qualidade;
8. apoiar usuários na descoberta de dados e produtos disponíveis, respeitando permissões, restrições de acesso e classificações de sensibilidade;
9. interpretar alertas de drift, evolução dos ativos, obsolescência, falhas de qualidade e incidentes de monitoramento;
10. registrar interações, recomendações, respostas, fontes utilizadas, limitações, validações humanas e feedback dos usuários;
11. avaliar qualidade, utilidade, confiabilidade, rastreabilidade, segurança e efetividade das respostas e recomendações dos agentes;
12. assegurar que agentes inteligentes operem sob controle humano e em conformidade com regras de segurança, privacidade, governança e uso responsável.

4. Fundamentação conceitual do módulo

Agentes inteligentes podem ser compreendidos como componentes computacionais capazes de interpretar solicitações, acessar informações, executar tarefas, gerar respostas, propor recomendações ou apoiar decisões com algum grau de autonomia operacional. No UnespDataLens-RM, sua atuação deve ser orientada por dados, metadados, regras, documentos, catálogos, registros e evidências produzidas pelo próprio modelo de referência.

Esses agentes podem empregar diferentes abordagens técnicas, como modelos de linguagem, recuperação de informação, busca semântica, embeddings, regras, workflows, APIs, grafos de conhecimento, sistemas de recomendação, análise de logs, bases vetoriais e mecanismos de validação. Contudo, sua atuação deve ser limitada pelo escopo do pipeline, pelas permissões do usuário e pelas políticas de governança.

Os agentes inteligentes do UnespDataLens-RM podem ser classificados conforme sua função principal:

Tipo de agente	Função principal	Exemplo de atuação
Agente de descoberta de dados	Auxilia usuários a encontrar fontes, datasets, features e produtos.	“Quais bases possuem dados de evasão?”
Agente de documentação	Gera, revisa ou complementa descrições, dicionários e fichas.	Sugerir descrição para atributos sem documentação.
Agente de qualidade	Consulta métricas e identifica problemas de qualidade.	Apontar datasets com queda de completude.
Agente de lineage	Explica origem, dependências e trajetória dos dados.	“De onde vem este indicador?”
Agente de governança	Orienta sobre acesso, sensibilidade, finalidade e restrições.	Informar se um dataset pode ser compartilhado.
Agente de versionamento	Localiza versões, changelogs e execuções reprodutíveis.	Explicar diferenças entre duas versões de um dashboard.
Agente de representações analíticas	Apoia escolha e interpretação de tabelas, séries e grafos.	Sugerir representação temporal ou grafo para um problema.
Agente de feature engineering	Sugere, explica, reutiliza ou avalia features.	Sugerir features temporais para evasão.
Agente de monitoramento	Interpreta alertas, drift e evolução dos ativos.	Explicar alerta de mudança de distribuição.
Agente de consumo analítico	Auxilia usuários na interpretação de produtos disponibilizados.	Explicar indicador em painel institucional.
Agente de manutenção do pipeline	Recomenda ações corretivas ou reprocessamentos.	Sugerir reexecução após drift de schema.

A atuação desses agentes deve obedecer aos seguintes princípios:

1. **evidência antes de inferência**: respostas devem se apoiar em registros, metadados, logs, documentos, métricas ou relatórios disponíveis;
2. **rastreabilidade da resposta**: sempre que possível, o agente deve registrar quais fontes, artefatos ou evidências sustentam sua resposta;
3. **controle de acesso**: o agente só deve consultar ou expor informações compatíveis com as permissões do usuário;
4. **validação humana**: recomendações críticas devem depender de revisão ou aprovação humana;
5. **transparência**: limitações, incertezas e ausência de dados devem ser explicitadas;

6. **não substituição da governança**: agentes apoiam o processo, mas não substituem políticas, responsáveis e decisões institucionais;
7. **não autonomização indevida**: agentes não devem executar alterações críticas sem autorização;
8. **responsabilidade e auditabilidade**: interações, recomendações e respostas devem ser registradas para auditoria e melhoria contínua.

À luz da **Design Science Research**, o Módulo 16 formaliza agentes inteligentes como artefatos computacionais e metodológicos voltados a resolver problemas de sobrecarga informacional, baixa descoberta de dados, dificuldade de interpretação, documentação insuficiente, complexidade de lineage, monitoramento contínuo e apoio à tomada de decisão baseada em evidências.

5. Natureza transversal do módulo

O Módulo 16 é transversal porque os agentes inteligentes podem atuar sobre todos os módulos do pipeline principal e sobre todos os módulos transversais.

No **Módulo 1 — Fontes e Inventário de Dados**, os agentes podem auxiliar na consulta ao inventário, identificação de fontes relevantes, sugestão de campos de documentação e comparação entre fontes.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, podem explicar falhas de ingestão, consultar logs, identificar fontes atrasadas e sugerir verificações iniciais.

No **Módulo 3 — Integração e Harmonização**, podem apoiar a identificação de chaves, entidades compartilhadas, conflitos, equivalências semânticas e riscos de integração.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, podem sugerir regras de limpeza, explicar atributos derivados, identificar transformações aplicadas e apontar impactos.

No **Módulo 5 — Qualidade e Validação de Dados**, podem interpretar relatórios de qualidade, indicar não conformidades, explicar métricas e sugerir ações corretivas.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, podem consultar métricas de distribuição, fairness, grupos afetados, riscos residuais e limitações de uso.

No **Módulo 7 — Armazenamento Analítico**, podem localizar ativos armazenados, explicar zonas, schemas, versões, formatos e estruturas persistidas.

No **Módulo 8 — Disponibilização e Consumo Analítico**, podem orientar usuários sobre produtos disponíveis, dashboards, APIs, datasets, relatórios, versões e restrições de acesso.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, utilizam catálogo, glossário, dicionários e modelos semânticos como base principal de conhecimento.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, devem respeitar políticas de acesso, sensibilidade, finalidade, retenção, privacidade e conformidade.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, podem explicar origem, percurso e dependências dos dados.

No **Módulo 12 — Versionamento e Reprodutibilidade**, podem identificar versões, explicar diferenças entre execuções e auxiliar na reconstrução de resultados.

No **Módulo 13 — Representações Analíticas**, podem explicar representações tabulares, temporais, espaciais, textuais, semânticas e em grafos, além de sugerir métricas e visualizações.

No **Módulo 14 — Feature Engineering e Feature Store**, podem consultar features, explicar regras de cálculo, identificar riscos de vazamento, sugerir reuso e interpretar drift de features.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, podem interpretar alertas, explicar mudanças, identificar impactos e recomendar ações de manutenção.

6. Formalização sintética do módulo

O Módulo 16 pode ser formalizado como um conjunto de agentes inteligentes que operam sobre ativos, metadados, evidências e regras do UnespDataLens-RM.

Considerando o conjunto de agentes:

$$\mathbf{AG = \{ag_1, ag_2, ..., ag_n\}}$$

cada agente $\mathbf{ag_i}$ pode ser descrito por:

$$\mathbf{ag_i = \{id, nome, função, escopo, fontes_conhecimento, permissões, ferramentas, entradas, saídas, limites, métricas, validação_humana\}}$$

Considerando o conjunto de ativos do modelo de referência:

$$\mathbf{A = \{a_1, a_2, ..., a_m\}}$$

e o conjunto de bases de conhecimento e evidências:

$$\mathbf{K = \{metadados, catálogo, glossário, lineage, versões, logs, relatórios, políticas, métricas, regras, alertas\}}$$

um agente pode ser representado por uma função:

$$\mathbf{ag_i(q, K, C) \rightarrow r}$$

em que:

- **q** representa uma consulta, tarefa ou solicitação;
- **K** representa o conjunto de conhecimentos e evidências disponíveis;
- **C** representa o contexto do usuário, suas permissões, finalidade e restrições;
- **r** representa uma resposta, explicação, recomendação, relatório, alerta interpretado ou ação proposta.

Cada resposta gerada por um agente deve ser acompanhada, sempre que possível, por um registro:

$$\mathbf{resposta = \{consulta, resposta, fontes_utilizadas, nível_confiança, limitações, usuário, data, validação\}}$$

Dessa forma, a atuação do agente não é uma caixa-preta isolada, mas um processo documentado, controlado, rastreável e avaliável.

7. Entradas do módulo

As entradas do Módulo 16 são compostas por ativos, documentos, registros, métricas, políticas e artefatos produzidos por todo o UnespDataLens-RM.

As principais entradas são:

- inventário de fontes;
- dicionários de dados;
- glossários;
- modelos semânticos;
- catálogos de dados;
- catálogos de produtos;
- registros de qualidade;
- relatórios de não conformidade;
- registros de viés e equidade;
- registros de sensibilidade;
- políticas de governança;
- matriz de acesso;
- registros de lineage;
- grafos de linhagem;
- registros de versionamento;
- changelogs;
- manifestos de execução;
- registros de monitoramento;
- alertas de drift;
- logs de ingestão;
- logs de acesso;
- dashboards;
- APIs;
- data marts;
- feature store;
- catálogo de features;
- representações tabulares;
- séries temporais;
- grafos;
- produtos de dados;
- documentação institucional;
- regras de negócio;
- perguntas dos usuários;
- permissões e perfis de acesso;
- feedback dos usuários.

8. Processamento

O processamento do Módulo 16 envolve recepção da solicitação, interpretação, verificação de permissões, recuperação de evidências, geração de resposta ou recomendação, validação e registro da interação.

8.1 Recepção e interpretação da solicitação

A primeira atividade consiste em receber uma pergunta, comando, tarefa ou solicitação do usuário ou de outro sistema. O agente deve interpretar a intenção da solicitação, identificar o escopo, reconhecer entidades mencionadas, verificar o tipo de resposta esperada e determinar quais bases de conhecimento precisam ser consultadas.

8.2 Verificação de contexto, perfil e permissão

A segunda atividade consiste em verificar quem está solicitando, qual é a finalidade, qual é o perfil de acesso e quais restrições se aplicam. O agente não deve acessar ou expor informações além do permitido ao usuário ou sistema solicitante.

8.3 Recuperação de informações e evidências

A terceira atividade consiste em consultar os artefatos relevantes do UnespDataLens-RM, como catálogo, metadados, lineage, registros de qualidade, logs, versões, relatórios, políticas, feature store, dashboards, produtos de dados ou registros de monitoramento.

A recuperação pode ser baseada em busca textual, busca semântica, consultas estruturadas, grafos, APIs, regras ou combinação dessas abordagens.

8.4 Raciocínio assistido por regras e contexto

A quarta atividade consiste em interpretar as evidências recuperadas e gerar uma resposta coerente com as regras do modelo de referência. O agente pode resumir, comparar, explicar, sugerir, classificar, priorizar ou alertar, mas deve manter vínculo com as evidências disponíveis.

8.5 Geração de respostas, explicações ou recomendações

A quinta atividade consiste em produzir uma resposta compreensível ao usuário. Essa resposta pode assumir a forma de explicação, lista de ativos, recomendação de ação, resumo de qualidade, interpretação de indicador, sugestão de feature, alerta de governança, análise de impacto ou documentação preliminar.

8.6 Registro das fontes utilizadas

A sexta atividade consiste em registrar quais artefatos, documentos, metadados, logs ou métricas foram utilizados para gerar a resposta. Esse registro aumenta a transparência e permite auditoria.

8.7 Sinalização de incertezas e limitações

A sétima atividade consiste em indicar quando a resposta possui incerteza, quando dados estão incompletos, quando o agente não possui permissão, quando não há evidência suficiente ou quando a solicitação exige validação humana.

8.8 Validação humana quando necessário

A oitava atividade consiste em encaminhar recomendações críticas para validação humana. Isso é especialmente importante para ações que envolvem exclusão de dados, alteração de regras, publicação de produtos, acesso a dados sensíveis, mitigação de viés, reprocessamento crítico ou mudanças em pipelines.

8.9 Registro da interação e feedback

A nona atividade consiste em registrar a interação, incluindo pergunta, resposta, usuário, data, fontes utilizadas, nível de confiança, limitações e feedback do usuário.

8.10 Aprimoramento contínuo do agente

A décima atividade consiste em utilizar feedback, avaliações, erros, lacunas documentais e interações anteriores para melhorar prompts, regras, bases de conhecimento, fluxos de recuperação, documentação e critérios de resposta.

9. Técnicas, métodos e tecnologias aplicáveis

O Módulo 16 pode empregar diferentes técnicas e tecnologias de inteligência artificial, recuperação de informação, automação e interação.

Entre as técnicas de recuperação e consulta, destacam-se:

- busca textual;
- busca semântica;
- embeddings;
- bases vetoriais;
- recuperação aumentada por contexto;
- consulta a catálogos;
- consulta a grafos de linhagem;
- consulta a APIs;
- consulta a logs;
- consulta SQL;
- consulta em bancos de grafos;
- indexação de documentos;
- rankeamento de resultados;
- filtragem por permissão.

Entre as técnicas de agentes inteligentes, podem ser utilizadas:

- modelos de linguagem;
- agentes baseados em ferramentas;
- workflows orientados por tarefas;
- sistemas de recomendação;
- regras especialistas;
- classificadores;
- sumarização automática;
- geração assistida de documentação;
- análise de intenção;
- extração de entidades;
- planejamento de tarefas;
- validação por regras;
- human-in-the-loop;
- recuperação aumentada por geração sobre catálogo e documentação;
- geração controlada por evidências.

Entre as técnicas de governança dos agentes, destacam-se:

- controle de acesso;

- filtros por sensibilidade;
- logs de interação;
- validação humana;
- auditoria de respostas;
- avaliação de alucinação;
- avaliação de utilidade;
- versionamento de prompts;
- testes de regressão de respostas;
- políticas de uso;
- limites de ação;
- rastreamento de fontes;
- monitoramento de segurança.

Entre as tecnologias possíveis, destacam-se:

- modelos de linguagem;
- bancos vetoriais;
- catálogos de dados;
- APIs internas;
- orquestradores de agentes;
- ferramentas de busca;
- bancos de grafos;
- dashboards;
- sistemas de logs;
- repositórios de documentação;
- ferramentas de observabilidade;
- mecanismos de autenticação;
- notebooks;
- serviços de geração de texto;
- ferramentas de workflow;
- interfaces conversacionais.

10. Métricas do módulo

As métricas do Módulo 16 avaliam cobertura, utilidade, confiabilidade, rastreabilidade, segurança, qualidade e eficiência dos agentes inteligentes.

Métrica	Sigla	Descrição
Taxa de respostas baseadas em evidências	TRBE	Proporção de respostas com fontes ou artefatos associados
Taxa de respostas úteis	TRU	Proporção de respostas avaliadas como úteis pelos usuários
Taxa de respostas corretas	TRC	Proporção de respostas validadas como corretas
Taxa de respostas com limitação explícita	TRLE	Proporção de respostas que sinalizam incerteza quando necessário
Taxa de recomendações aceitas	TRA	Proporção de recomendações aceitas por usuários ou responsáveis

Taxa de recomendações rejeitadas	TRR	Proporção de recomendações recusadas
Taxa de interações auditáveis	TIA	Proporção de interações com registro completo
Taxa de violações de acesso	TVA	Proporção de tentativas ou respostas que violam restrições
Taxa de consultas resolvidas	TCR	Proporção de consultas respondidas sem encaminhamento
Tempo médio de resposta	TMR	Tempo médio para gerar resposta
Taxa de encaminhamento humano	TEH	Proporção de solicitações encaminhadas para validação humana
Taxa de erro do agente	TEA	Proporção de respostas incorretas, incompletas ou inadequadas
Cobertura da base de conhecimento	CBK	Proporção de artefatos indexados ou acessíveis ao agente
Índice de confiabilidade da resposta	ICR	Grau de confiança atribuído à resposta
Índice de utilidade do agente	IUA	Grau agregado de utilidade percebida e efetividade

10.1 Taxa de Respostas Baseadas em Evidências — TRBE

A **Taxa de Respostas Baseadas em Evidências** mede a proporção de respostas que indicam ou registram os artefatos utilizados como base.

TRBE = número de respostas com evidências associadas / número total de respostas geradas

Essa métrica é essencial para evitar respostas sem fundamento documental.

10.2 Taxa de Respostas Úteis — TRU

A **Taxa de Respostas Úteis** mede a proporção de respostas consideradas úteis pelos usuários.

TRU = número de respostas avaliadas como úteis / número total de respostas avaliadas

Essa métrica avalia a contribuição prática dos agentes.

10.3 Taxa de Violações de Acesso — TVA

A **Taxa de Violações de Acesso** mede a proporção de interações em que o agente tentou acessar, expor ou recomendar uso de informação fora das permissões definidas.

TVA = número de violações de acesso / número total de interações avaliadas

Essa métrica deve tender a zero.

10.4 Taxa de Encaminhamento Humano — TEH

A **Taxa de Encaminhamento Humano** mede a proporção de solicitações que exigiram validação humana.

TEH = número de solicitações encaminhadas para validação humana / número total de solicitações

Essa métrica ajuda a avaliar o grau de autonomia controlada dos agentes.

10.5 Índice de Confiabilidade da Resposta — ICR

O **Índice de Confiabilidade da Resposta** pode ser definido como uma composição de evidência, consistência, validação e conformidade.

Uma formulação possível é:

$$ICR(r_i) = w_1E + w_2C + w_3V + w_4S$$

em que:

- **E** representa presença de evidências;
- **C** representa consistência com os artefatos consultados;
- **V** representa validação humana ou automática;
- **S** representa conformidade com segurança e acesso;
- w_i representa o peso de cada dimensão.

11. Artefatos formalizados do módulo

Os artefatos do Módulo 16 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A16.1	Arquitetura de Agentes Inteligentes	Documento técnico/diagrama	CBK	Define agentes, fontes e integrações
A16.2	Catálogo de Agentes	Tabela/catálogo	TRU	Documenta função, escopo e limites
A16.3	Base de Conhecimento dos Agentes	Índice textual/vetorial/grafo	CBK	Organiza evidências consultáveis
A16.4	Política de Uso dos	Documento	TVA	Define limites,

	Agentes	normativo		permissões e restrições
A16.5	Registro de Interações dos Agentes	Log estruturado	TIA	Documenta perguntas, respostas e evidências
A16.6	Registro de Recomendações	Tabela de recomendações	TRA	Acompanha sugestões e decisões
A16.7	Relatório de Avaliação dos Agentes	Relatório técnico	TRC, TRU	Avalia qualidade e utilidade
A16.8	Registro de Validação Humana	Tabela/workflow	TEH	Controla revisão de ações críticas
A16.9	Registro de Erros e Limitações	Tabela/relatório	TEA	Documenta falhas, incertezas e lacunas
A16.10	Manifesto de Resposta do Agente	Documento estruturado	TRBE, ICR	Registra evidências da resposta

12. Artefato A16.1 — Arquitetura de Agentes Inteligentes

12.1 Finalidade

A **Arquitetura de Agentes Inteligentes** define como os agentes serão organizados, quais bases de conhecimento poderão acessar, quais ferramentas poderão utilizar, quais permissões terão e como se integrarão aos módulos do UnespDataLens-RM.

12.2 Tipo de artefato na DSR

Esse artefato é um **modelo** e uma **instanciação**, pois representa a estrutura conceitual e computacional dos agentes no modelo de referência.

12.3 Estrutura mínima

Elemento	Descrição
Tipos de agentes	Descoberta, documentação, qualidade, lineage, governança, monitoramento etc.
Fontes de conhecimento	Catálogo, metadados, logs, lineage, versões, relatórios
Ferramentas	APIs, busca, bancos, grafos, dashboards

disponíveis	
Controle de acesso	Permissões e restrições por usuário
Mecanismo de recuperação	Busca textual, semântica, grafo, SQL
Mecanismo de resposta	Modelo de linguagem, regras, templates
Validação humana	Quando necessária
Registro de interação	Logs, evidências e feedback
Monitoramento	Métricas de qualidade e segurança

12.4 Métricas associadas

- cobertura da base de conhecimento;
- taxa de interações auditáveis;
- taxa de respostas baseadas em evidências;
- taxa de violações de acesso;
- taxa de respostas úteis.

12.5 Critérios de validação

A arquitetura é considerada válida quando define agentes, escopos, bases consultáveis, permissões, ferramentas, mecanismos de validação e registros de auditoria.

12.6 Contribuição

Esse artefato organiza os agentes como componentes controlados do modelo de referência, evitando uso improvisado, inseguro ou desconectado da governança.

13. Artefato A16.2 — Catálogo de Agentes

13.1 Finalidade

O **Catálogo de Agentes** documenta os agentes disponíveis, suas funções, escopo, limites, permissões, fontes de conhecimento e responsáveis.

13.2 Estrutura mínima

Campo	Descrição
Identificador do agente	Código único
Nome	Nome do agente
Função	Descoberta, documentação, qualidade, lineage etc.

Escopo	Módulos ou ativos que pode consultar
Fontes de conhecimento	Artefatos acessíveis
Ferramentas	APIs, busca, catálogo, grafo etc.
Permissões	Níveis de acesso
Limites	Ações que não pode executar
Responsável	Pessoa ou equipe
Status	Experimental, ativo, restrito, depreciado

13.3 Métricas associadas

- número de agentes catalogados;
- taxa de agentes com escopo definido;
- taxa de agentes com limites documentados;
- taxa de agentes com responsável;
- taxa de agentes avaliados.

13.4 Contribuição

Esse artefato torna explícito o papel de cada agente e evita sobreposição, ambiguidade ou uso indevido.

14. Artefato A16.3 — Base de Conhecimento dos Agentes

14.1 Finalidade

A **Base de Conhecimento dos Agentes** organiza os documentos, metadados, catálogos, registros, grafos, relatórios e artefatos que podem ser consultados pelos agentes.

14.2 Estrutura mínima

Componente	Descrição
Catálogo de dados	Ativos, descrições, responsáveis e restrições
Dicionário de dados	Atributos, tipos, domínios e significados
Glossário	Termos e conceitos do domínio
Lineage	Origem, dependências e transformações
Qualidade	Métricas, relatórios e não conformidades

Versionamento	Versões, changelogs e manifestos
Governança	Políticas, acessos e restrições
Monitoramento	Alertas, drift e evolução
Feature store	Features, regras e versões
Produtos	Dashboards, APIs, relatórios e datasets

14.3 Métricas associadas

- cobertura da base de conhecimento;
- taxa de artefatos indexados;
- taxa de artefatos atualizados;
- taxa de consultas com retorno;
- tempo médio de recuperação.

14.4 Contribuição

Esse artefato garante que as respostas dos agentes sejam fundamentadas em conhecimento estruturado do modelo de referência.

15. Artefato A16.4 — Política de Uso dos Agentes

15.1 Finalidade

A **Política de Uso dos Agentes** define limites, responsabilidades, permissões, restrições, validações e regras para uso seguro dos agentes inteligentes.

15.2 Estrutura mínima

Campo	Descrição
Escopo	Agentes e usos cobertos
Usuários autorizados	Perfis permitidos
Dados acessíveis	Tipos de ativos permitidos
Dados proibidos	Informações restritas ou sensíveis
Ações permitidas	Consultar, explicar, recomendar, documentar
Ações restritas	Alterar, excluir, publicar, liberar acesso
Validação humana	Casos obrigatórios
Registro obrigatório	Logs e evidências
Critérios de	Controle de acesso e privacidade

segurança	
Tratamento de erro	Como reportar falhas

15.3 Métricas associadas

- taxa de violações de acesso;
- taxa de interações auditáveis;
- taxa de encaminhamento humano;
- taxa de erros críticos;
- taxa de conformidade das interações.

15.4 Contribuição

Esse artefato garante que agentes operem como apoio governado e não como mecanismos autônomos sem controle.

16. Artefato A16.5 — Registro de Interações dos Agentes

16.1 Finalidade

O **Registro de Interações dos Agentes** documenta perguntas, respostas, usuários, evidências, fontes utilizadas, limitações e feedback.

16.2 Estrutura mínima

Campo	Descrição
Identificador da interação	Código único
Usuário ou sistema	Quem interagiu
Agente	Agente utilizado
Consulta	Pergunta ou tarefa
Resposta	Resposta gerada
Fontes utilizadas	Artefatos consultados
Nível de confiança	Estimativa de confiabilidade
Limitações	Incertezas ou restrições
Data e hora	Momento da interação
Feedback	Avaliação do usuário

16.3 Métricas associadas

- taxa de interações auditáveis;
- taxa de respostas baseadas em evidências;
- taxa de respostas úteis;
- taxa de respostas com limitação explícita;
- tempo médio de resposta.

16.4 Contribuição

Esse artefato permite auditar e melhorar continuamente a atuação dos agentes.

17. Artefato A16.6 — Registro de Recomendações

17.1 Finalidade

O **Registro de Recomendações** documenta sugestões geradas por agentes, como reprocessamento, revisão de qualidade, atualização de metadados, criação de features, bloqueio de produto ou investigação de drift.

17.2 Estrutura mínima

Campo	Descrição
Recomendação	Ação sugerida
Agente	Agente que recomendou
Evidência	Métrica, alerta, relatório ou fonte usada
Ativo afetado	Dataset, produto, feature, dashboard etc.
Severidade	Baixa, média, alta ou crítica
Responsável sugerido	Pessoa ou equipe
Status	Aceita, rejeitada, em análise, executada
Justificativa humana	Motivo da decisão
Data	Momento da recomendação

17.3 Métricas associadas

- taxa de recomendações aceitas;
- taxa de recomendações rejeitadas;
- taxa de recomendações executadas;
- tempo médio de resposta à recomendação;
- taxa de recomendações com evidência.

17.4 Contribuição

Esse artefato transforma sugestões de agentes em itens rastreáveis e avaliáveis.

18. Artefato A16.7 — Relatório de Avaliação dos Agentes

18.1 Finalidade

O **Relatório de Avaliação dos Agentes** avalia qualidade, utilidade, segurança, rastreabilidade e confiabilidade das respostas e recomendações produzidas.

18.2 Estrutura mínima

Campo	Descrição
Agente avaliado	Nome ou código
Período	Intervalo analisado
Número de interações	Quantidade
Respostas úteis	Avaliação de utilidade
Respostas corretas	Validação técnica
Erros identificados	Falhas ou inconsistências
Violações de acesso	Problemas de segurança
Recomendações aceitas	Aderência prática
Limitações	Lacunas observadas
Melhorias propostas	Ajustes necessários

18.3 Métricas associadas

- taxa de respostas corretas;
- taxa de respostas úteis;
- taxa de erro do agente;
- taxa de violações de acesso;
- índice de utilidade do agente.

18.4 Contribuição

Esse artefato permite avaliar empiricamente se os agentes agregam valor ao modelo de referência.

19. Artefato A16.8 — Registro de Validação Humana

19.1 Finalidade

O **Registro de Validação Humana** documenta revisões feitas por especialistas, gestores ou responsáveis sobre respostas, recomendações ou ações sugeridas pelos agentes.

19.2 Estrutura mínima

Campo	Descrição
Item validado	Resposta, recomendação ou ação
Agente	Agente responsável
Validador	Pessoa que revisou
Decisão	Aprovado, corrigido, rejeitado
Justificativa	Motivo
Alteração solicitada	Quando aplicável
Data	Momento da validação
Status	Concluído ou pendente

19.3 Métricas associadas

- taxa de encaminhamento humano;
- taxa de recomendações aprovadas;
- taxa de correções humanas;
- taxa de rejeições;
- tempo médio de validação.

19.4 Contribuição

Esse artefato garante supervisão humana em decisões críticas e reduz riscos de automação indevida.

20. Artefato A16.9 — Registro de Erros e Limitações

20.1 Finalidade

O **Registro de Erros e Limitações** documenta falhas, respostas incorretas, lacunas de conhecimento, incertezas, limitações de acesso, inconsistências e problemas dos agentes.

20.2 Estrutura mínima

Campo	Descrição
Agente	Agente envolvido

Interação	Consulta associada
Tipo de erro	Informação incorreta, incompleta, acesso, interpretação
Causa provável	Falta de metadado, ambiguidade, erro de recuperação etc.
Severidade	Baixa, média, alta ou crítica
Correção proposta	Ação de melhoria
Responsável	Pessoa ou equipe
Status	Aberto, corrigido, aceito
Evidência	Registro da interação

20.3 Métricas associadas

- taxa de erro do agente;
- número de erros críticos;
- taxa de erros corrigidos;
- taxa de lacunas de conhecimento;
- taxa de recorrência de erro.

20.4 Contribuição

Esse artefato permite melhoria contínua dos agentes e da própria base documental do modelo de referência.

21. Artefato A16.10 — Manifesto de Resposta do Agente

21.1 Finalidade

O **Manifesto de Resposta do Agente** registra de forma estruturada como uma resposta foi produzida.

21.2 Estrutura mínima

Campo	Descrição
Consulta	Pergunta ou tarefa original
Resposta gerada	Conteúdo produzido
Agente	Agente responsável
Fontes utilizadas	Artefatos consultados

Evidências principais	Trechos, métricas ou registros usados
Regras aplicadas	Filtros, permissões, políticas
Limitações	Incertezas ou lacunas
Nível de confiança	Baixo, médio ou alto
Necessita validação humana	Sim/não
Data	Momento da resposta

21.3 Métricas associadas

- taxa de respostas baseadas em evidências;
- índice de confiabilidade da resposta;
- taxa de respostas com limitação explícita;
- taxa de interações auditáveis.

21.4 Contribuição

Esse artefato fortalece transparência, rastreabilidade e confiabilidade das respostas dos agentes.

22. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Catálogos, metadados, glossários, lineage, versões, logs, relatórios de qualidade, políticas, registros de monitoramento, feature store, produtos de dados, dashboards, APIs e consultas dos usuários.
Processamento	Interpretação da solicitação, verificação de permissões, recuperação de evidências, geração de resposta, recomendação, explicação, validação humana quando necessária e registro da interação.
Artefatos	Arquitetura de agentes, catálogo de agentes, base de conhecimento, política de uso, registros de interação, recomendações, avaliação, validação humana, erros e manifesto de resposta.
Métricas	TRBE, TRU, TRC, TRLE, TRA, TRR, TIA, TVA, TCR, TMR, TEH, TEA, CBK, ICR e IUA.
Saídas	Respostas, explicações, recomendações, documentação assistida, alertas interpretados, consultas ao catálogo, apoio à governança e registros auditáveis das interações.

23. Critérios gerais de validação do módulo

O Módulo 16 pode ser considerado adequadamente implementado quando:

1. os agentes possuem funções e escopos definidos;
2. as bases de conhecimento utilizadas pelos agentes estão identificadas;
3. os agentes respeitam permissões e restrições de acesso;
4. as respostas são baseadas em evidências sempre que possível;
5. interações são registradas de forma auditável;
6. recomendações críticas passam por validação humana;
7. erros e limitações são documentados;
8. a utilidade das respostas é avaliada pelos usuários;
9. violações de acesso são monitoradas e tratadas;
10. os agentes indicam incertezas ou ausência de evidência quando necessário;
11. as recomendações dos agentes podem ser aceitas, rejeitadas ou revisadas por responsáveis;
12. os agentes contribuem efetivamente para descoberta, explicação, documentação, qualidade, governança, lineage, monitoramento ou reuso dos ativos.

24. Relação com o pipeline principal

O Módulo 16 apoia todos os módulos do pipeline principal.

Com o **Módulo 1 — Fontes e Inventário de Dados**, auxilia na descoberta, documentação e consulta das fontes.

Com o **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, apoia interpretação de logs, falhas, manifestos e registros de ingestão.

Com o **Módulo 3 — Integração e Harmonização**, auxilia na compreensão de chaves, conflitos, relações e equivalências.

Com o **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, explica transformações, regras, atributos derivados e impactos.

Com o **Módulo 5 — Qualidade e Validação de Dados**, interpreta métricas, não conformidades, relatórios e recomendações de correção.

Com o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, consulta métricas de fairness, grupos afetados, riscos e limitações.

Com o **Módulo 7 — Armazenamento Analítico**, localiza ativos, zonas, schemas, versões e estruturas persistidas.

Com o **Módulo 8 — Disponibilização e Consumo Analítico**, orienta usuários sobre produtos, dashboards, APIs, relatórios, versões, acesso e limitações.

25. Relação com os demais módulos transversais

O Módulo 16 depende do **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, pois os agentes precisam de metadados, glossários, dicionários, modelos semânticos e catálogos para responder de forma contextualizada.

Relaciona-se com o **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, pois os agentes devem respeitar políticas, permissões, finalidades, sensibilidade e restrições de acesso.

Relaciona-se com o **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, pois agentes podem explicar a trajetória dos dados, dependências e impacto de mudanças.

Relaciona-se com o **Módulo 12 — Versionamento e Reprodutibilidade**, pois agentes podem localizar versões, explicar diferenças e auxiliar na reconstrução de execuções.

Relaciona-se com o **Módulo 13 — Representações Analíticas**, pois agentes podem explicar tabelas, séries temporais, grafos e grafos temporais.

Relaciona-se com o **Módulo 14 — Feature Engineering e Feature Store**, pois agentes podem consultar features, explicar regras de cálculo, sugerir reuso e identificar riscos.

Relaciona-se com o **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, pois agentes podem interpretar alertas, explicar drift, sugerir ações corretivas e apoiar manutenção contínua.

26. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, agentes inteligentes poderiam apoiar usuários e equipes técnicas na consulta e interpretação de dados sobre matrícula, evasão, desempenho, frequência, permanência, interações no ambiente virtual e indicadores institucionais.

Um gestor poderia perguntar: “quais cursos apresentam maior evasão no último período?”. O agente poderia consultar o catálogo de produtos, identificar o dashboard ou dataset correspondente, verificar a definição de evasão, indicar a versão utilizada, apresentar limitações e apontar a origem dos dados.

Um pesquisador poderia perguntar: “quais features estão disponíveis para modelar evasão?”. O agente poderia consultar a feature store, listar features relacionadas à trajetória acadêmica, desempenho, frequência, interação no AVA e perfil socioeconômico, indicando qualidade, versão, sensibilidade e risco de vazamento.

Um analista poderia perguntar: “por que o indicador de evasão mudou nesta versão?”. O agente poderia consultar o changelog, o lineage e o relatório de comparação entre versões para explicar se a mudança decorreu de nova regra, atualização de fonte, correção de dados ou alteração de período.

27. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, agentes inteligentes poderiam apoiar gestores, técnicos e pesquisadores na consulta a indicadores territoriais, bases de atendimento, dados de saúde, educação, assistência social e produtos analíticos integrados.

Um gestor poderia perguntar: “quais territórios apresentam aumento recente de demanda por atendimento?”. O agente poderia consultar produtos de dados, dashboards e relatórios de monitoramento, indicar os territórios com variação relevante e apontar a origem dos dados.

Um técnico poderia perguntar: “quais bases alimentam o indicador de vulnerabilidade?”. O agente poderia consultar o grafo de linhagem e explicar quais fontes, regras, versões e atributos foram usados.

Um responsável pela governança poderia perguntar: “este dataset pode ser compartilhado externamente?”. O agente poderia consultar classificação de sensibilidade, política de acesso, finalidade, base legal e nível de agregação, recomendando validação humana antes de qualquer compartilhamento.

28. Contribuição técnico-científica do módulo

A contribuição do Módulo 16 está em formalizar agentes inteligentes como componentes transversais de apoio à Engenharia Analítica de Dados, integrados aos artefatos, métricas, metadados, governança, lineage, versionamento, monitoramento, feature store, representações analíticas e produtos do UnespDataLens-RM.

Em vez de utilizar agentes apenas como interfaces genéricas de pergunta e resposta, o UnespDataLens-RM propõe agentes orientados por evidências, com escopo definido, controle de acesso, registro de interações, validação humana, avaliação de utilidade e rastreabilidade das respostas.

Essa formalização contribui para:

- melhorar descoberta e reutilização de dados;
- reduzir sobrecarga informacional;
- apoiar documentação técnica;
- explicar indicadores, datasets, features e produtos;
- facilitar consulta ao catálogo e ao lineage;
- apoiar detecção e interpretação de problemas de qualidade;
- orientar usuários sobre governança e restrições;
- recomendar ações corretivas com base em evidências;
- aumentar transparência e auditabilidade das interações;
- integrar IA generativa e agentes ao pipeline de dados com controle, segurança e responsabilidade.

29. Síntese do módulo

O **Módulo Transversal 16 — Agentes Inteligentes de Apoio ao Pipeline** introduz mecanismos inteligentes para apoiar a navegação, interpretação, documentação, explicação, recomendação e manutenção do ecossistema UnespDataLens-RM.

Sua importância é transversal porque os agentes podem atuar sobre todos os módulos do modelo de referência, desde a identificação das fontes até o consumo dos produtos analíticos, conectando informações dispersas em catálogos, metadados, lineage, logs, versões, relatórios, dashboards, features, produtos de dados, registros de monitoramento e políticas de governança.

Ao formalizar agentes inteligentes por meio de arquitetura, catálogo, base de conhecimento, política de uso, registros de interação, validação humana, avaliação de qualidade e manifestos de resposta, o módulo fortalece a confiabilidade, a transparência, a governança, a rastreabilidade e a utilidade do UnespDataLens-RM como modelo de referência para Engenharia Analítica de Dados.

Módulo Transversal 16 — Agentes Inteligentes de Apoio ao Pipeline

1. Finalidade do módulo

O **Módulo Transversal 16 — Agentes Inteligentes de Apoio ao Pipeline** tem como finalidade especificar, integrar, controlar, avaliar e monitorar agentes inteligentes capazes de apoiar usuários, analistas, pesquisadores, gestores e equipes técnicas na navegação, compreensão, documentação, validação, explicação, recomendação e manutenção dos processos do modelo de referência **UnespDataLens-RM**.

Esse módulo atua de forma transversal porque os agentes inteligentes podem interagir com diferentes etapas do pipeline principal e com diferentes artefatos produzidos pelos módulos transversais. Eles podem apoiar o inventário de fontes, a consulta ao catálogo, a geração de documentação, a explicação da linhagem dos dados, a interpretação de métricas de qualidade, a identificação de problemas de governança, a recomendação de transformações, a análise de drift, a busca de produtos de dados, a explicação de features, a interpretação de dashboards, a identificação de riscos e a orientação sobre restrições de acesso.

No contexto do UnespDataLens-RM, os agentes inteligentes não devem ser compreendidos como substitutos dos módulos técnicos, nem como mecanismos autônomos de decisão sem controle humano. Eles são concebidos como componentes de apoio, operando sobre metadados, catálogos, glossários, logs, registros de qualidade, regras, versões, grafos de linhagem, features, relatórios, produtos de dados, políticas de governança e registros de monitoramento. Seu papel é ampliar a capacidade de consulta, explicação, recomendação e documentação do ecossistema analítico, mantendo rastreabilidade, validação, segurança e conformidade.

Os agentes podem assumir diferentes funções. Um agente pode auxiliar na descoberta de fontes de dados. Outro pode explicar a origem de um indicador. Outro pode identificar datasets com queda de qualidade. Outro pode sugerir revisão de features com drift. Outro pode explicar quais dashboards dependem de uma fonte alterada. Outro pode apoiar a geração de fichas de documentação. Outro pode orientar usuários sobre restrições de acesso, finalidade de uso e limitações de um produto de dados.

Entretanto, o uso de agentes inteligentes em pipelines de dados exige cuidados metodológicos, técnicos e éticos. As respostas geradas devem ser fundamentadas em evidências registradas nos módulos do modelo de referência, como metadados, lineage, versões, logs, catálogos, relatórios de qualidade, políticas de governança e registros de monitoramento. O agente deve indicar, sempre que possível, a origem da informação utilizada, sinalizar incertezas, respeitar permissões de acesso e evitar inferências não fundamentadas sobre dados sensíveis, incompletos ou não autorizados.

Assim, o Módulo 16 formaliza os agentes inteligentes como componentes transversais de apoio ao pipeline, orientados por evidências, governança, rastreabilidade, controle humano, versionamento, segurança e critérios de confiabilidade.

2. Objetivo geral

O objetivo geral do Módulo 16 é **integrar agentes inteligentes ao UnespDataLens-RM como mecanismos de apoio à consulta, explicação, documentação, recomendação, validação e manutenção dos ativos analíticos**, garantindo que suas respostas, sugestões e ações sejam baseadas em metadados, evidências, regras, lineage, versões, políticas de governança, registros de monitoramento e validação humana quando necessário.

3. Objetivos específicos

São objetivos específicos do Módulo 16:

1. definir os tipos de agentes inteligentes que poderão atuar no apoio ao pipeline analítico;
2. especificar funções, escopos, permissões, limites e responsabilidades de cada agente;
3. integrar agentes aos catálogos de dados, metadados, glossários, dicionários, modelos semânticos, registros de qualidade, lineage, versionamento, feature store e monitoramento;
4. permitir consultas em linguagem natural sobre fontes, datasets, atributos, produtos de dados, dashboards, APIs, features, grafos, indicadores, regras e versões;
5. apoiar a explicação de indicadores, features, datasets e produtos analíticos a partir de sua origem, regras, versões, métricas de qualidade e limitações;
6. recomendar ações de melhoria, revisão, reprocessamento, documentação, validação, monitoramento ou governança com base em evidências registradas;
7. auxiliar na geração, revisão e atualização de documentação técnica, dicionários, fichas de ativos, relatórios, descrições de produtos e registros de qualidade;
8. apoiar usuários na descoberta de dados e produtos disponíveis, respeitando permissões, restrições de acesso e classificações de sensibilidade;
9. interpretar alertas de drift, evolução dos ativos, obsolescência, falhas de qualidade e incidentes de monitoramento;
10. registrar interações, recomendações, respostas, fontes utilizadas, limitações, validações humanas e feedback dos usuários;
11. avaliar qualidade, utilidade, confiabilidade, rastreabilidade, segurança e efetividade das respostas e recomendações dos agentes;
12. assegurar que agentes inteligentes operem sob controle humano e em conformidade com regras de segurança, privacidade, governança e uso responsável.

4. Fundamentação conceitual do módulo

Agentes inteligentes podem ser compreendidos como componentes computacionais capazes de interpretar solicitações, acessar informações, executar tarefas, gerar respostas, propor recomendações ou apoiar decisões com algum grau de autonomia operacional. No UnespDataLens-RM, sua atuação deve ser orientada por dados, metadados, regras, documentos, catálogos, registros e evidências produzidas pelo próprio modelo de referência.

Esses agentes podem empregar diferentes abordagens técnicas, como modelos de linguagem, recuperação de informação, busca semântica, embeddings, regras, workflows, APIs, grafos de conhecimento, sistemas de recomendação, análise de logs, bases vetoriais e mecanismos de validação. Contudo, sua atuação deve ser limitada pelo escopo do pipeline, pelas permissões do usuário e pelas políticas de governança.

Os agentes inteligentes do UnespDataLens-RM podem ser classificados conforme sua função principal:

Tipo de agente	Função principal	Exemplo de atuação
Agente de descoberta de dados	Auxilia usuários a encontrar fontes, datasets, features e produtos.	“Quais bases possuem dados de evasão?”
Agente de documentação	Gera, revisa ou complementa descrições, dicionários e fichas.	Sugerir descrição para atributos sem documentação.
Agente de qualidade	Consulta métricas e identifica problemas de qualidade.	Apontar datasets com queda de completude.
Agente de lineage	Explica origem, dependências e trajetória dos dados.	“De onde vem este indicador?”
Agente de governança	Orienta sobre acesso, sensibilidade, finalidade e restrições.	Informar se um dataset pode ser compartilhado.
Agente de versionamento	Localiza versões, changelogs e execuções reprodutíveis.	Explicar diferenças entre duas versões de um dashboard.
Agente de representações analíticas	Apoia escolha e interpretação de tabelas, séries e grafos.	Sugerir representação temporal ou grafo para um problema.
Agente de feature engineering	Sugere, explica, reutiliza ou avalia features.	Sugerir features temporais para evasão.
Agente de monitoramento	Interpreta alertas, drift e evolução dos ativos.	Explicar alerta de mudança de distribuição.
Agente de consumo analítico	Auxilia usuários na interpretação de produtos disponibilizados.	Explicar indicador em painel institucional.
Agente de manutenção do pipeline	Recomenda ações corretivas ou reprocessamentos.	Sugerir reexecução após drift de schema.

A atuação desses agentes deve obedecer aos seguintes princípios:

1. **evidência antes de inferência**: respostas devem se apoiar em registros, metadados, logs, documentos, métricas ou relatórios disponíveis;
2. **rastreabilidade da resposta**: sempre que possível, o agente deve registrar quais fontes, artefatos ou evidências sustentam sua resposta;
3. **controle de acesso**: o agente só deve consultar ou expor informações compatíveis com as permissões do usuário;
4. **validação humana**: recomendações críticas devem depender de revisão ou aprovação humana;
5. **transparência**: limitações, incertezas e ausência de dados devem ser explicitadas;

6. **não substituição da governança**: agentes apoiam o processo, mas não substituem políticas, responsáveis e decisões institucionais;
7. **não autonomização indevida**: agentes não devem executar alterações críticas sem autorização;
8. **responsabilidade e auditabilidade**: interações, recomendações e respostas devem ser registradas para auditoria e melhoria contínua.

À luz da **Design Science Research**, o Módulo 16 formaliza agentes inteligentes como artefatos computacionais e metodológicos voltados a resolver problemas de sobrecarga informacional, baixa descoberta de dados, dificuldade de interpretação, documentação insuficiente, complexidade de lineage, monitoramento contínuo e apoio à tomada de decisão baseada em evidências.

5. Natureza transversal do módulo

O Módulo 16 é transversal porque os agentes inteligentes podem atuar sobre todos os módulos do pipeline principal e sobre todos os módulos transversais.

No **Módulo 1 — Fontes e Inventário de Dados**, os agentes podem auxiliar na consulta ao inventário, identificação de fontes relevantes, sugestão de campos de documentação e comparação entre fontes.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, podem explicar falhas de ingestão, consultar logs, identificar fontes atrasadas e sugerir verificações iniciais.

No **Módulo 3 — Integração e Harmonização**, podem apoiar a identificação de chaves, entidades compartilhadas, conflitos, equivalências semânticas e riscos de integração.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, podem sugerir regras de limpeza, explicar atributos derivados, identificar transformações aplicadas e apontar impactos.

No **Módulo 5 — Qualidade e Validação de Dados**, podem interpretar relatórios de qualidade, indicar não conformidades, explicar métricas e sugerir ações corretivas.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, podem consultar métricas de distribuição, fairness, grupos afetados, riscos residuais e limitações de uso.

No **Módulo 7 — Armazenamento Analítico**, podem localizar ativos armazenados, explicar zonas, schemas, versões, formatos e estruturas persistidas.

No **Módulo 8 — Disponibilização e Consumo Analítico**, podem orientar usuários sobre produtos disponíveis, dashboards, APIs, datasets, relatórios, versões e restrições de acesso.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, utilizam catálogo, glossário, dicionários e modelos semânticos como base principal de conhecimento.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, devem respeitar políticas de acesso, sensibilidade, finalidade, retenção, privacidade e conformidade.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, podem explicar origem, percurso e dependências dos dados.

No **Módulo 12 — Versionamento e Reprodutibilidade**, podem identificar versões, explicar diferenças entre execuções e auxiliar na reconstrução de resultados.

No **Módulo 13 — Representações Analíticas**, podem explicar representações tabulares, temporais, espaciais, textuais, semânticas e em grafos, além de sugerir métricas e visualizações.

No **Módulo 14 — Feature Engineering e Feature Store**, podem consultar features, explicar regras de cálculo, identificar riscos de vazamento, sugerir reuso e interpretar drift de features.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, podem interpretar alertas, explicar mudanças, identificar impactos e recomendar ações de manutenção.

6. Formalização sintética do módulo

O Módulo 16 pode ser formalizado como um conjunto de agentes inteligentes que operam sobre ativos, metadados, evidências e regras do UnespDataLens-RM.

Considerando o conjunto de agentes:

$$\mathbf{AG = \{ag_1, ag_2, ..., ag_n\}}$$

cada agente $\mathbf{ag_i}$ pode ser descrito por:

$$\mathbf{ag_i = \{id, nome, função, escopo, fontes_conhecimento, permissões, ferramentas, entradas, saídas, limites, métricas, validação_humana\}}$$

Considerando o conjunto de ativos do modelo de referência:

$$\mathbf{A = \{a_1, a_2, ..., a_m\}}$$

e o conjunto de bases de conhecimento e evidências:

$$\mathbf{K = \{metadados, catálogo, glossário, lineage, versões, logs, relatórios, políticas, métricas, regras, alertas\}}$$

um agente pode ser representado por uma função:

$$\mathbf{ag_i(q, K, C) \rightarrow r}$$

em que:

- **q** representa uma consulta, tarefa ou solicitação;
- **K** representa o conjunto de conhecimentos e evidências disponíveis;
- **C** representa o contexto do usuário, suas permissões, finalidade e restrições;
- **r** representa uma resposta, explicação, recomendação, relatório, alerta interpretado ou ação proposta.

Cada resposta gerada por um agente deve ser acompanhada, sempre que possível, por um registro:

$$\mathbf{resposta = \{consulta, resposta, fontes_utilizadas, nível_confiança, limitações, usuário, data, validação\}}$$

Dessa forma, a atuação do agente não é uma caixa-preta isolada, mas um processo documentado, controlado, rastreável e avaliável.

7. Entradas do módulo

As entradas do Módulo 16 são compostas por ativos, documentos, registros, métricas, políticas e artefatos produzidos por todo o UnespDataLens-RM.

As principais entradas são:

- inventário de fontes;
- dicionários de dados;
- glossários;
- modelos semânticos;
- catálogos de dados;
- catálogos de produtos;
- registros de qualidade;
- relatórios de não conformidade;
- registros de viés e equidade;
- registros de sensibilidade;
- políticas de governança;
- matriz de acesso;
- registros de lineage;
- grafos de linhagem;
- registros de versionamento;
- changelogs;
- manifestos de execução;
- registros de monitoramento;
- alertas de drift;
- logs de ingestão;
- logs de acesso;
- dashboards;
- APIs;
- data marts;
- feature store;
- catálogo de features;
- representações tabulares;
- séries temporais;
- grafos;
- produtos de dados;
- documentação institucional;
- regras de negócio;
- perguntas dos usuários;
- permissões e perfis de acesso;
- feedback dos usuários.

8. Processamento

O processamento do Módulo 16 envolve recepção da solicitação, interpretação, verificação de permissões, recuperação de evidências, geração de resposta ou recomendação, validação e registro da interação.

8.1 Recepção e interpretação da solicitação

A primeira atividade consiste em receber uma pergunta, comando, tarefa ou solicitação do usuário ou de outro sistema. O agente deve interpretar a intenção da solicitação, identificar o escopo, reconhecer entidades mencionadas, verificar o tipo de resposta esperada e determinar quais bases de conhecimento precisam ser consultadas.

8.2 Verificação de contexto, perfil e permissão

A segunda atividade consiste em verificar quem está solicitando, qual é a finalidade, qual é o perfil de acesso e quais restrições se aplicam. O agente não deve acessar ou expor informações além do permitido ao usuário ou sistema solicitante.

8.3 Recuperação de informações e evidências

A terceira atividade consiste em consultar os artefatos relevantes do UnespDataLens-RM, como catálogo, metadados, lineage, registros de qualidade, logs, versões, relatórios, políticas, feature store, dashboards, produtos de dados ou registros de monitoramento.

A recuperação pode ser baseada em busca textual, busca semântica, consultas estruturadas, grafos, APIs, regras ou combinação dessas abordagens.

8.4 Raciocínio assistido por regras e contexto

A quarta atividade consiste em interpretar as evidências recuperadas e gerar uma resposta coerente com as regras do modelo de referência. O agente pode resumir, comparar, explicar, sugerir, classificar, priorizar ou alertar, mas deve manter vínculo com as evidências disponíveis.

8.5 Geração de respostas, explicações ou recomendações

A quinta atividade consiste em produzir uma resposta compreensível ao usuário. Essa resposta pode assumir a forma de explicação, lista de ativos, recomendação de ação, resumo de qualidade, interpretação de indicador, sugestão de feature, alerta de governança, análise de impacto ou documentação preliminar.

8.6 Registro das fontes utilizadas

A sexta atividade consiste em registrar quais artefatos, documentos, metadados, logs ou métricas foram utilizados para gerar a resposta. Esse registro aumenta a transparência e permite auditoria.

8.7 Sinalização de incertezas e limitações

A sétima atividade consiste em indicar quando a resposta possui incerteza, quando dados estão incompletos, quando o agente não possui permissão, quando não há evidência suficiente ou quando a solicitação exige validação humana.

8.8 Validação humana quando necessário

A oitava atividade consiste em encaminhar recomendações críticas para validação humana. Isso é especialmente importante para ações que envolvem exclusão de dados, alteração de regras, publicação de produtos, acesso a dados sensíveis, mitigação de viés, reprocessamento crítico ou mudanças em pipelines.

8.9 Registro da interação e feedback

A nona atividade consiste em registrar a interação, incluindo pergunta, resposta, usuário, data, fontes utilizadas, nível de confiança, limitações e feedback do usuário.

8.10 Aprimoramento contínuo do agente

A décima atividade consiste em utilizar feedback, avaliações, erros, lacunas documentais e interações anteriores para melhorar prompts, regras, bases de conhecimento, fluxos de recuperação, documentação e critérios de resposta.

9. Técnicas, métodos e tecnologias aplicáveis

O Módulo 16 pode empregar diferentes técnicas e tecnologias de inteligência artificial, recuperação de informação, automação e interação.

Entre as técnicas de recuperação e consulta, destacam-se:

- busca textual;
- busca semântica;
- embeddings;
- bases vetoriais;
- recuperação aumentada por contexto;
- consulta a catálogos;
- consulta a grafos de linhagem;
- consulta a APIs;
- consulta a logs;
- consulta SQL;
- consulta em bancos de grafos;
- indexação de documentos;
- rankeamento de resultados;
- filtragem por permissão.

Entre as técnicas de agentes inteligentes, podem ser utilizadas:

- modelos de linguagem;
- agentes baseados em ferramentas;
- workflows orientados por tarefas;
- sistemas de recomendação;
- regras especialistas;
- classificadores;
- sumarização automática;
- geração assistida de documentação;
- análise de intenção;
- extração de entidades;
- planejamento de tarefas;
- validação por regras;
- human-in-the-loop;
- recuperação aumentada por geração sobre catálogo e documentação;
- geração controlada por evidências.

Entre as técnicas de governança dos agentes, destacam-se:

- controle de acesso;

- filtros por sensibilidade;
- logs de interação;
- validação humana;
- auditoria de respostas;
- avaliação de alucinação;
- avaliação de utilidade;
- versionamento de prompts;
- testes de regressão de respostas;
- políticas de uso;
- limites de ação;
- rastreamento de fontes;
- monitoramento de segurança.

Entre as tecnologias possíveis, destacam-se:

- modelos de linguagem;
- bancos vetoriais;
- catálogos de dados;
- APIs internas;
- orquestradores de agentes;
- ferramentas de busca;
- bancos de grafos;
- dashboards;
- sistemas de logs;
- repositórios de documentação;
- ferramentas de observabilidade;
- mecanismos de autenticação;
- notebooks;
- serviços de geração de texto;
- ferramentas de workflow;
- interfaces conversacionais.

10. Métricas do módulo

As métricas do Módulo 16 avaliam cobertura, utilidade, confiabilidade, rastreabilidade, segurança, qualidade e eficiência dos agentes inteligentes.

Métrica	Sigla	Descrição
Taxa de respostas baseadas em evidências	TRBE	Proporção de respostas com fontes ou artefatos associados
Taxa de respostas úteis	TRU	Proporção de respostas avaliadas como úteis pelos usuários
Taxa de respostas corretas	TRC	Proporção de respostas validadas como corretas
Taxa de respostas com limitação explícita	TRLE	Proporção de respostas que sinalizam incerteza quando necessário
Taxa de recomendações aceitas	TRA	Proporção de recomendações aceitas por usuários ou responsáveis

Taxa de recomendações rejeitadas	TRR	Proporção de recomendações recusadas
Taxa de interações auditáveis	TIA	Proporção de interações com registro completo
Taxa de violações de acesso	TVA	Proporção de tentativas ou respostas que violam restrições
Taxa de consultas resolvidas	TCR	Proporção de consultas respondidas sem encaminhamento
Tempo médio de resposta	TMR	Tempo médio para gerar resposta
Taxa de encaminhamento humano	TEH	Proporção de solicitações encaminhadas para validação humana
Taxa de erro do agente	TEA	Proporção de respostas incorretas, incompletas ou inadequadas
Cobertura da base de conhecimento	CBK	Proporção de artefatos indexados ou acessíveis ao agente
Índice de confiabilidade da resposta	ICR	Grau de confiança atribuído à resposta
Índice de utilidade do agente	IUA	Grau agregado de utilidade percebida e efetividade

10.1 Taxa de Respostas Baseadas em Evidências — TRBE

A **Taxa de Respostas Baseadas em Evidências** mede a proporção de respostas que indicam ou registram os artefatos utilizados como base.

TRBE = número de respostas com evidências associadas / número total de respostas geradas

Essa métrica é essencial para evitar respostas sem fundamento documental.

10.2 Taxa de Respostas Úteis — TRU

A **Taxa de Respostas Úteis** mede a proporção de respostas consideradas úteis pelos usuários.

TRU = número de respostas avaliadas como úteis / número total de respostas avaliadas

Essa métrica avalia a contribuição prática dos agentes.

10.3 Taxa de Violações de Acesso — TVA

A **Taxa de Violações de Acesso** mede a proporção de interações em que o agente tentou acessar, expor ou recomendar uso de informação fora das permissões definidas.

TVA = número de violações de acesso / número total de interações avaliadas

Essa métrica deve tender a zero.

10.4 Taxa de Encaminhamento Humano — TEH

A **Taxa de Encaminhamento Humano** mede a proporção de solicitações que exigiram validação humana.

TEH = número de solicitações encaminhadas para validação humana / número total de solicitações

Essa métrica ajuda a avaliar o grau de autonomia controlada dos agentes.

10.5 Índice de Confiabilidade da Resposta — ICR

O **Índice de Confiabilidade da Resposta** pode ser definido como uma composição de evidência, consistência, validação e conformidade.

Uma formulação possível é:

$$ICR(r_i) = w_1E + w_2C + w_3V + w_4S$$

em que:

- **E** representa presença de evidências;
- **C** representa consistência com os artefatos consultados;
- **V** representa validação humana ou automática;
- **S** representa conformidade com segurança e acesso;
- w_i representa o peso de cada dimensão.

11. Artefatos formalizados do módulo

Os artefatos do Módulo 16 são especificados como unidades técnico-metodológicas formais. Cada artefato possui finalidade, estrutura mínima, representação, métricas associadas, critérios de validação e relação com os demais módulos.

Código	Artefato	Representação	Métrica principal	Contribuição
A16.1	Arquitetura de Agentes Inteligentes	Documento técnico/diagrama	CBK	Define agentes, fontes e integrações
A16.2	Catálogo de Agentes	Tabela/catálogo	TRU	Documenta função, escopo e limites
A16.3	Base de Conhecimento dos Agentes	Índice textual/vetorial/grafo	CBK	Organiza evidências consultáveis
A16.4	Política de Uso dos	Documento	TVA	Define limites,

	Agentes	normativo		permissões e restrições
A16.5	Registro de Interações dos Agentes	Log estruturado	TIA	Documenta perguntas, respostas e evidências
A16.6	Registro de Recomendações	Tabela de recomendações	TRA	Acompanha sugestões e decisões
A16.7	Relatório de Avaliação dos Agentes	Relatório técnico	TRC, TRU	Avalia qualidade e utilidade
A16.8	Registro de Validação Humana	Tabela/workflow	TEH	Controla revisão de ações críticas
A16.9	Registro de Erros e Limitações	Tabela/relatório	TEA	Documenta falhas, incertezas e lacunas
A16.10	Manifesto de Resposta do Agente	Documento estruturado	TRBE, ICR	Registra evidências da resposta

12. Artefato A16.1 — Arquitetura de Agentes Inteligentes

12.1 Finalidade

A **Arquitetura de Agentes Inteligentes** define como os agentes serão organizados, quais bases de conhecimento poderão acessar, quais ferramentas poderão utilizar, quais permissões terão e como se integrarão aos módulos do UnespDataLens-RM.

12.2 Tipo de artefato na DSR

Esse artefato é um **modelo** e uma **instanciação**, pois representa a estrutura conceitual e computacional dos agentes no modelo de referência.

12.3 Estrutura mínima

Elemento	Descrição
Tipos de agentes	Descoberta, documentação, qualidade, lineage, governança, monitoramento etc.
Fontes de conhecimento	Catálogo, metadados, logs, lineage, versões, relatórios
Ferramentas	APIs, busca, bancos, grafos, dashboards

disponíveis	
Controle de acesso	Permissões e restrições por usuário
Mecanismo de recuperação	Busca textual, semântica, grafo, SQL
Mecanismo de resposta	Modelo de linguagem, regras, templates
Validação humana	Quando necessária
Registro de interação	Logs, evidências e feedback
Monitoramento	Métricas de qualidade e segurança

12.4 Métricas associadas

- cobertura da base de conhecimento;
- taxa de interações auditáveis;
- taxa de respostas baseadas em evidências;
- taxa de violações de acesso;
- taxa de respostas úteis.

12.5 Critérios de validação

A arquitetura é considerada válida quando define agentes, escopos, bases consultáveis, permissões, ferramentas, mecanismos de validação e registros de auditoria.

12.6 Contribuição

Esse artefato organiza os agentes como componentes controlados do modelo de referência, evitando uso improvisado, inseguro ou desconectado da governança.

13. Artefato A16.2 — Catálogo de Agentes

13.1 Finalidade

O **Catálogo de Agentes** documenta os agentes disponíveis, suas funções, escopo, limites, permissões, fontes de conhecimento e responsáveis.

13.2 Estrutura mínima

Campo	Descrição
Identificador do agente	Código único
Nome	Nome do agente
Função	Descoberta, documentação, qualidade, lineage etc.

Escopo	Módulos ou ativos que pode consultar
Fontes de conhecimento	Artefatos acessíveis
Ferramentas	APIs, busca, catálogo, grafo etc.
Permissões	Níveis de acesso
Limites	Ações que não pode executar
Responsável	Pessoa ou equipe
Status	Experimental, ativo, restrito, depreciado

13.3 Métricas associadas

- número de agentes catalogados;
- taxa de agentes com escopo definido;
- taxa de agentes com limites documentados;
- taxa de agentes com responsável;
- taxa de agentes avaliados.

13.4 Contribuição

Esse artefato torna explícito o papel de cada agente e evita sobreposição, ambiguidade ou uso indevido.

14. Artefato A16.3 — Base de Conhecimento dos Agentes

14.1 Finalidade

A **Base de Conhecimento dos Agentes** organiza os documentos, metadados, catálogos, registros, grafos, relatórios e artefatos que podem ser consultados pelos agentes.

14.2 Estrutura mínima

Componente	Descrição
Catálogo de dados	Ativos, descrições, responsáveis e restrições
Dicionário de dados	Atributos, tipos, domínios e significados
Glossário	Termos e conceitos do domínio
Lineage	Origem, dependências e transformações
Qualidade	Métricas, relatórios e não conformidades

Versionamento	Versões, changelogs e manifestos
Governança	Políticas, acessos e restrições
Monitoramento	Alertas, drift e evolução
Feature store	Features, regras e versões
Produtos	Dashboards, APIs, relatórios e datasets

14.3 Métricas associadas

- cobertura da base de conhecimento;
- taxa de artefatos indexados;
- taxa de artefatos atualizados;
- taxa de consultas com retorno;
- tempo médio de recuperação.

14.4 Contribuição

Esse artefato garante que as respostas dos agentes sejam fundamentadas em conhecimento estruturado do modelo de referência.

15. Artefato A16.4 — Política de Uso dos Agentes

15.1 Finalidade

A **Política de Uso dos Agentes** define limites, responsabilidades, permissões, restrições, validações e regras para uso seguro dos agentes inteligentes.

15.2 Estrutura mínima

Campo	Descrição
Escopo	Agentes e usos cobertos
Usuários autorizados	Perfis permitidos
Dados acessíveis	Tipos de ativos permitidos
Dados proibidos	Informações restritas ou sensíveis
Ações permitidas	Consultar, explicar, recomendar, documentar
Ações restritas	Alterar, excluir, publicar, liberar acesso
Validação humana	Casos obrigatórios
Registro obrigatório	Logs e evidências
Critérios de	Controle de acesso e privacidade

segurança	
Tratamento de erro	Como reportar falhas

15.3 Métricas associadas

- taxa de violações de acesso;
- taxa de interações auditáveis;
- taxa de encaminhamento humano;
- taxa de erros críticos;
- taxa de conformidade das interações.

15.4 Contribuição

Esse artefato garante que agentes operem como apoio governado e não como mecanismos autônomos sem controle.

16. Artefato A16.5 — Registro de Interações dos Agentes

16.1 Finalidade

O **Registro de Interações dos Agentes** documenta perguntas, respostas, usuários, evidências, fontes utilizadas, limitações e feedback.

16.2 Estrutura mínima

Campo	Descrição
Identificador da interação	Código único
Usuário ou sistema	Quem interagiu
Agente	Agente utilizado
Consulta	Pergunta ou tarefa
Resposta	Resposta gerada
Fontes utilizadas	Artefatos consultados
Nível de confiança	Estimativa de confiabilidade
Limitações	Incertezas ou restrições
Data e hora	Momento da interação
Feedback	Avaliação do usuário

16.3 Métricas associadas

- taxa de interações auditáveis;
- taxa de respostas baseadas em evidências;
- taxa de respostas úteis;
- taxa de respostas com limitação explícita;
- tempo médio de resposta.

16.4 Contribuição

Esse artefato permite auditar e melhorar continuamente a atuação dos agentes.

17. Artefato A16.6 — Registro de Recomendações

17.1 Finalidade

O **Registro de Recomendações** documenta sugestões geradas por agentes, como reprocessamento, revisão de qualidade, atualização de metadados, criação de features, bloqueio de produto ou investigação de drift.

17.2 Estrutura mínima

Campo	Descrição
Recomendação	Ação sugerida
Agente	Agente que recomendou
Evidência	Métrica, alerta, relatório ou fonte usada
Ativo afetado	Dataset, produto, feature, dashboard etc.
Severidade	Baixa, média, alta ou crítica
Responsável sugerido	Pessoa ou equipe
Status	Aceita, rejeitada, em análise, executada
Justificativa humana	Motivo da decisão
Data	Momento da recomendação

17.3 Métricas associadas

- taxa de recomendações aceitas;
- taxa de recomendações rejeitadas;
- taxa de recomendações executadas;
- tempo médio de resposta à recomendação;
- taxa de recomendações com evidência.

17.4 Contribuição

Esse artefato transforma sugestões de agentes em itens rastreáveis e avaliáveis.

18. Artefato A16.7 — Relatório de Avaliação dos Agentes

18.1 Finalidade

O **Relatório de Avaliação dos Agentes** avalia qualidade, utilidade, segurança, rastreabilidade e confiabilidade das respostas e recomendações produzidas.

18.2 Estrutura mínima

Campo	Descrição
Agente avaliado	Nome ou código
Período	Intervalo analisado
Número de interações	Quantidade
Respostas úteis	Avaliação de utilidade
Respostas corretas	Validação técnica
Erros identificados	Falhas ou inconsistências
Violações de acesso	Problemas de segurança
Recomendações aceitas	Aderência prática
Limitações	Lacunas observadas
Melhorias propostas	Ajustes necessários

18.3 Métricas associadas

- taxa de respostas corretas;
- taxa de respostas úteis;
- taxa de erro do agente;
- taxa de violações de acesso;
- índice de utilidade do agente.

18.4 Contribuição

Esse artefato permite avaliar empiricamente se os agentes agregam valor ao modelo de referência.

19. Artefato A16.8 — Registro de Validação Humana

19.1 Finalidade

O **Registro de Validação Humana** documenta revisões feitas por especialistas, gestores ou responsáveis sobre respostas, recomendações ou ações sugeridas pelos agentes.

19.2 Estrutura mínima

Campo	Descrição
Item validado	Resposta, recomendação ou ação
Agente	Agente responsável
Validador	Pessoa que revisou
Decisão	Aprovado, corrigido, rejeitado
Justificativa	Motivo
Alteração solicitada	Quando aplicável
Data	Momento da validação
Status	Concluído ou pendente

19.3 Métricas associadas

- taxa de encaminhamento humano;
- taxa de recomendações aprovadas;
- taxa de correções humanas;
- taxa de rejeições;
- tempo médio de validação.

19.4 Contribuição

Esse artefato garante supervisão humana em decisões críticas e reduz riscos de automação indevida.

20. Artefato A16.9 — Registro de Erros e Limitações

20.1 Finalidade

O **Registro de Erros e Limitações** documenta falhas, respostas incorretas, lacunas de conhecimento, incertezas, limitações de acesso, inconsistências e problemas dos agentes.

20.2 Estrutura mínima

Campo	Descrição
Agente	Agente envolvido

Interação	Consulta associada
Tipo de erro	Informação incorreta, incompleta, acesso, interpretação
Causa provável	Falta de metadado, ambiguidade, erro de recuperação etc.
Severidade	Baixa, média, alta ou crítica
Correção proposta	Ação de melhoria
Responsável	Pessoa ou equipe
Status	Aberto, corrigido, aceito
Evidência	Registro da interação

20.3 Métricas associadas

- taxa de erro do agente;
- número de erros críticos;
- taxa de erros corrigidos;
- taxa de lacunas de conhecimento;
- taxa de recorrência de erro.

20.4 Contribuição

Esse artefato permite melhoria contínua dos agentes e da própria base documental do modelo de referência.

21. Artefato A16.10 — Manifesto de Resposta do Agente

21.1 Finalidade

O **Manifesto de Resposta do Agente** registra de forma estruturada como uma resposta foi produzida.

21.2 Estrutura mínima

Campo	Descrição
Consulta	Pergunta ou tarefa original
Resposta gerada	Conteúdo produzido
Agente	Agente responsável
Fontes utilizadas	Artefatos consultados

Evidências principais	Trechos, métricas ou registros usados
Regras aplicadas	Filtros, permissões, políticas
Limitações	Incertezas ou lacunas
Nível de confiança	Baixo, médio ou alto
Necessita validação humana	Sim/não
Data	Momento da resposta

21.3 Métricas associadas

- taxa de respostas baseadas em evidências;
- índice de confiabilidade da resposta;
- taxa de respostas com limitação explícita;
- taxa de interações auditáveis.

21.4 Contribuição

Esse artefato fortalece transparência, rastreabilidade e confiabilidade das respostas dos agentes.

22. Estrutura operacional do módulo

Elemento	Descrição
Entradas	Catálogos, metadados, glossários, lineage, versões, logs, relatórios de qualidade, políticas, registros de monitoramento, feature store, produtos de dados, dashboards, APIs e consultas dos usuários.
Processamento	Interpretação da solicitação, verificação de permissões, recuperação de evidências, geração de resposta, recomendação, explicação, validação humana quando necessária e registro da interação.
Artefatos	Arquitetura de agentes, catálogo de agentes, base de conhecimento, política de uso, registros de interação, recomendações, avaliação, validação humana, erros e manifesto de resposta.
Métricas	TRBE, TRU, TRC, TRLE, TRA, TRR, TIA, TVA, TCR, TMR, TEH, TEA, CBK, ICR e IUA.
Saídas	Respostas, explicações, recomendações, documentação assistida, alertas interpretados, consultas ao catálogo, apoio à governança e registros auditáveis das interações.

23. Critérios gerais de validação do módulo

O Módulo 16 pode ser considerado adequadamente implementado quando:

1. os agentes possuem funções e escopos definidos;
2. as bases de conhecimento utilizadas pelos agentes estão identificadas;
3. os agentes respeitam permissões e restrições de acesso;
4. as respostas são baseadas em evidências sempre que possível;
5. interações são registradas de forma auditável;
6. recomendações críticas passam por validação humana;
7. erros e limitações são documentados;
8. a utilidade das respostas é avaliada pelos usuários;
9. violações de acesso são monitoradas e tratadas;
10. os agentes indicam incertezas ou ausência de evidência quando necessário;
11. as recomendações dos agentes podem ser aceitas, rejeitadas ou revisadas por responsáveis;
12. os agentes contribuem efetivamente para descoberta, explicação, documentação, qualidade, governança, lineage, monitoramento ou reuso dos ativos.

24. Relação com o pipeline principal

O Módulo 16 apoia todos os módulos do pipeline principal.

Com o **Módulo 1 — Fontes e Inventário de Dados**, auxilia na descoberta, documentação e consulta das fontes.

Com o **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, apoia interpretação de logs, falhas, manifestos e registros de ingestão.

Com o **Módulo 3 — Integração e Harmonização**, auxilia na compreensão de chaves, conflitos, relações e equivalências.

Com o **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, explica transformações, regras, atributos derivados e impactos.

Com o **Módulo 5 — Qualidade e Validação de Dados**, interpreta métricas, não conformidades, relatórios e recomendações de correção.

Com o **Módulo 6 — Desbalanceamento, Vieses e Equidade**, consulta métricas de fairness, grupos afetados, riscos e limitações.

Com o **Módulo 7 — Armazenamento Analítico**, localiza ativos, zonas, schemas, versões e estruturas persistidas.

Com o **Módulo 8 — Disponibilização e Consumo Analítico**, orienta usuários sobre produtos, dashboards, APIs, relatórios, versões, acesso e limitações.

25. Relação com os demais módulos transversais

O Módulo 16 depende do **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, pois os agentes precisam de metadados, glossários, dicionários, modelos semânticos e catálogos para responder de forma contextualizada.

Relaciona-se com o **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, pois os agentes devem respeitar políticas, permissões, finalidades, sensibilidade e restrições de acesso.

Relaciona-se com o **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, pois agentes podem explicar a trajetória dos dados, dependências e impacto de mudanças.

Relaciona-se com o **Módulo 12 — Versionamento e Reprodutibilidade**, pois agentes podem localizar versões, explicar diferenças e auxiliar na reconstrução de execuções.

Relaciona-se com o **Módulo 13 — Representações Analíticas**, pois agentes podem explicar tabelas, séries temporais, grafos e grafos temporais.

Relaciona-se com o **Módulo 14 — Feature Engineering e Feature Store**, pois agentes podem consultar features, explicar regras de cálculo, sugerir reuso e identificar riscos.

Relaciona-se com o **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, pois agentes podem interpretar alertas, explicar drift, sugerir ações corretivas e apoiar manutenção contínua.

26. Exemplo de aplicação no domínio educacional

Em uma aplicação educacional, agentes inteligentes poderiam apoiar usuários e equipes técnicas na consulta e interpretação de dados sobre matrícula, evasão, desempenho, frequência, permanência, interações no ambiente virtual e indicadores institucionais.

Um gestor poderia perguntar: “quais cursos apresentam maior evasão no último período?”. O agente poderia consultar o catálogo de produtos, identificar o dashboard ou dataset correspondente, verificar a definição de evasão, indicar a versão utilizada, apresentar limitações e apontar a origem dos dados.

Um pesquisador poderia perguntar: “quais features estão disponíveis para modelar evasão?”. O agente poderia consultar a feature store, listar features relacionadas à trajetória acadêmica, desempenho, frequência, interação no AVA e perfil socioeconômico, indicando qualidade, versão, sensibilidade e risco de vazamento.

Um analista poderia perguntar: “por que o indicador de evasão mudou nesta versão?”. O agente poderia consultar o changelog, o lineage e o relatório de comparação entre versões para explicar se a mudança decorreu de nova regra, atualização de fonte, correção de dados ou alteração de período.

27. Exemplo de aplicação em políticas públicas

Em uma aplicação em políticas públicas, agentes inteligentes poderiam apoiar gestores, técnicos e pesquisadores na consulta a indicadores territoriais, bases de atendimento, dados de saúde, educação, assistência social e produtos analíticos integrados.

Um gestor poderia perguntar: “quais territórios apresentam aumento recente de demanda por atendimento?”. O agente poderia consultar produtos de dados, dashboards e relatórios de monitoramento, indicar os territórios com variação relevante e apontar a origem dos dados.

Um técnico poderia perguntar: “quais bases alimentam o indicador de vulnerabilidade?”. O agente poderia consultar o grafo de linhagem e explicar quais fontes, regras, versões e atributos foram usados.

Um responsável pela governança poderia perguntar: “este dataset pode ser compartilhado externamente?”. O agente poderia consultar classificação de sensibilidade, política de acesso, finalidade, base legal e nível de agregação, recomendando validação humana antes de qualquer compartilhamento.

28. Contribuição técnico-científica do módulo

A contribuição do Módulo 16 está em formalizar agentes inteligentes como componentes transversais de apoio à Engenharia Analítica de Dados, integrados aos artefatos, métricas, metadados, governança, lineage, versionamento, monitoramento, feature store, representações analíticas e produtos do UnespDataLens-RM.

Em vez de utilizar agentes apenas como interfaces genéricas de pergunta e resposta, o UnespDataLens-RM propõe agentes orientados por evidências, com escopo definido, controle de acesso, registro de interações, validação humana, avaliação de utilidade e rastreabilidade das respostas.

Essa formalização contribui para:

- melhorar descoberta e reutilização de dados;
- reduzir sobrecarga informacional;
- apoiar documentação técnica;
- explicar indicadores, datasets, features e produtos;
- facilitar consulta ao catálogo e ao lineage;
- apoiar detecção e interpretação de problemas de qualidade;
- orientar usuários sobre governança e restrições;
- recomendar ações corretivas com base em evidências;
- aumentar transparência e auditabilidade das interações;
- integrar IA generativa e agentes ao pipeline de dados com controle, segurança e responsabilidade.

29. Síntese do módulo

O **Módulo Transversal 16 — Agentes Inteligentes de Apoio ao Pipeline** introduz mecanismos inteligentes para apoiar a navegação, interpretação, documentação, explicação, recomendação e manutenção do ecossistema UnespDataLens-RM.

Sua importância é transversal porque os agentes podem atuar sobre todos os módulos do modelo de referência, desde a identificação das fontes até o consumo dos produtos analíticos, conectando informações dispersas em catálogos, metadados, lineage, logs, versões, relatórios, dashboards, features, produtos de dados, registros de monitoramento e políticas de governança.

Ao formalizar agentes inteligentes por meio de arquitetura, catálogo, base de conhecimento, política de uso, registros de interação, validação humana, avaliação de qualidade e manifestos de resposta, o módulo fortalece a confiabilidade, a transparência, a governança, a rastreabilidade e a utilidade do UnespDataLens-RM como modelo de referência para Engenharia Analítica de Dados.

Dimensão Complementar — Preparação Especializada por Natureza do Dado

1. Definição da dimensão

A dimensão Preparação Especializada por Natureza do Dado refere-se à capacidade do UnespDataLens-RM de orientar a preparação dos dados de acordo com suas características estruturais, temporais, espaciais, textuais, semânticas, relacionais, multimodais ou dinâmicas. Essa dimensão reconhece que a preparação de dados não pode ser tratada como uma etapa homogênea, composta apenas por operações genéricas de limpeza e padronização.

No contexto do modelo de referência, preparar dados significa transformar registros brutos ou parcialmente integrados em ativos analíticos adequados ao uso, preservando significado, qualidade, rastreabilidade, governança, versionamento e reprodutibilidade. Entretanto, cada natureza de dado demanda operações próprias. Dados tabulares exigem validação de tipos, domínios e chaves; dados temporais exigem ordenação, janelas e tratamento de frequência; dados espaciais exigem geocodificação e validação topológica; dados textuais exigem normalização linguística e representação semântica; dados em grafos exigem construção de nós e arestas; dados multimodais exigem alinhamento entre modalidades; e dados em streaming exigem tratamento de eventos, latência e janelas de processamento.

Assim, essa dimensão complementa o Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados, fornecendo uma orientação metodológica para que a preparação seja especializada conforme o tipo de dado, sem fragmentar o modelo em módulos específicos para cada técnica analítica.

2. Problema tratado

Modelos tradicionais de ETL frequentemente descrevem a preparação dos dados de forma genérica, assumindo implicitamente dados estruturados, tabulares e relativamente estáveis. Essa abordagem é insuficiente em ecossistemas analíticos contemporâneos, nos quais os dados podem assumir diferentes naturezas, formatos, granularidades e dinâmicas de produção.

Quando dados de naturezas distintas são preparados por procedimentos genéricos, podem ocorrer perdas de significado, redução da qualidade analítica, distorções nos resultados, baixa reprodutibilidade, fragilidade na documentação, dificuldade de integração e ausência de métricas específicas. Um texto livre não pode ser preparado da mesma forma que uma tabela relacional; uma série temporal não pode ser tratada como um conjunto de registros independentes; um grafo não pode ser reduzido apenas a uma tabela de atributos; um dado espacial não pode ignorar projeção, escala e topologia; e um fluxo contínuo não pode ser tratado apenas como uma carga batch convencional.

Essa dimensão trata, portanto, da lacuna metodológica existente entre a preparação genérica dos dados e a preparação especializada exigida por diferentes formas de representação, análise e consumo. Seu papel é explicitar que a natureza do dado condiciona as operações de preparação, as métricas de qualidade, os artefatos produzidos, os registros de lineage e os critérios de validação.

3. Objetivo da dimensão

O objetivo da dimensão Preparação Especializada por Natureza do Dado é orientar a especificação, execução, documentação, avaliação e evolução de subpipelines de preparação adequados aos diferentes tipos de dados presentes no ecossistema analítico do UnespDataLens-RM.

Essa dimensão busca assegurar que cada subpipeline explicite suas entradas, operações, técnicas, parâmetros, regras, métricas, artefatos, critérios de validação, relações de dependência, mecanismos de versionamento e registros de lineage, permitindo que os ativos produzidos sejam confiáveis, reutilizáveis, governados, auditáveis e reprodutíveis.

4. Escopo da dimensão

O escopo da dimensão inclui a preparação de dados estruturados, temporais, espaciais, textuais, semânticos, em grafos, multimodais e em fluxo contínuo. A dimensão não define técnicas analíticas finais, como análise de sentimentos, classificação, regressão, clustering, previsão, sumarização, recomendação ou detecção de anomalias. Essas tarefas pertencem a etapas posteriores de modelagem, análise ou consumo analítico.

A dimensão atua sobre os processos de preparação que antecedem ou sustentam essas tarefas. No caso de análise de sentimentos, por exemplo, o foco da dimensão não é classificar textos como positivos, negativos ou neutros, mas preparar o corpus textual por meio de coleta, limpeza, normalização, anonimização, enriquecimento, representação, validação, documentação, versionamento e rastreabilidade.

5. Princípios orientadores

A dimensão é orientada pelos seguintes princípios: adequação metodológica, preservação de significado, especialização por natureza do dado, rastreabilidade, reprodutibilidade, governança, qualidade específica, interoperabilidade e utilidade analítica.

A adequação metodológica estabelece que cada tipo de dado deve ser preparado por operações compatíveis com sua estrutura e finalidade. A preservação de significado determina que as transformações não devem eliminar informações relevantes para a análise. A especialização por natureza do dado reconhece que dados tabulares, temporais, espaciais, textuais, semânticos, grafos, multimodais e streaming exigem métodos distintos. A rastreabilidade exige registro de todas as operações aplicadas. A reprodutibilidade requer versionamento de dados, scripts, parâmetros, ambientes e artefatos. A governança assegura controle de acesso, classificação de sensibilidade e finalidade de uso. A qualidade específica exige métricas próprias para cada tipo de dado. A interoperabilidade garante que as saídas possam ser consumidas por módulos posteriores. A utilidade analítica assegura que a preparação esteja alinhada ao problema de análise.

6. Entradas da dimensão

As entradas da dimensão são dados brutos, semibrutos ou parcialmente integrados provenientes dos módulos anteriores, especialmente do Módulo 1, do Módulo 2 e do Módulo 3. Essas entradas podem incluir tabelas relacionais, arquivos estruturados, planilhas, APIs, logs, eventos, séries temporais, camadas geoespaciais, documentos, textos livres, mensagens, imagens, áudio, vídeo, grafos, fluxos contínuos e registros multimodais.

Também são entradas relevantes os metadados preliminares, dicionários de dados, contratos de dados, glossários, regras de negócio, classificações de sensibilidade, restrições de acesso, finalidade analítica, requisitos de qualidade, demandas dos usuários, registros de origem, informações de periodicidade e limitações conhecidas das fontes.

7. Processos gerais da dimensão

Independentemente da natureza do dado, a dimensão envolve um conjunto de processos gerais: identificação da natureza predominante do dado; seleção do subpipeline de preparação adequado; definição das regras de transformação, limpeza e enriquecimento; registro dos parâmetros e das decisões metodológicas; execução das operações específicas; avaliação da qualidade antes e depois da preparação; geração dos artefatos intermediários e finais; registro de lineage; versionamento dos dados e das regras; e validação das saídas para uso em representações analíticas, feature engineering, modelagem, dashboards, APIs ou relatórios.
Esses processos devem ser documentados em artefatos próprios, permitindo que equipes técnicas, pesquisadores, gestores e usuários compreendam o que foi feito, por que foi feito, quais dados foram afetados, quais versões foram utilizadas e quais limitações permanecem após a preparação.

8. Subpipeline de Preparação de Dados Estruturados

A Preparação de Dados Estruturados contempla dados organizados em tabelas, atributos, registros, chaves, domínios e restrições. Inclui operações como padronização de tipos, tratamento de valores ausentes, remoção ou consolidação de duplicidades, validação de domínios, normalização de categorias, verificação de integridade referencial, tratamento de outliers, recodificação de atributos, agregações e criação de atributos derivados.
As saídas incluem datasets estruturados tratados, tabelas padronizadas, dicionários atualizados, regras de transformação, relatórios de qualidade, registros de impacto das operações e versões documentadas dos conjuntos de dados.

9. Subpipeline de Preparação de Dados Temporais

A Preparação de Dados Temporais contempla registros com dimensão temporal explícita, como séries históricas, logs, eventos, trajetórias, sequências e dados longitudinais. Inclui operações como ordenação temporal, validação de timestamps, tratamento de lacunas temporais, interpolação quando justificada, agregação por janelas, criação de defasagens, cálculo de tendências, identificação de periodicidade, alinhamento temporal entre fontes e definição de janelas de observação.
As saídas incluem séries temporais tratadas, janelas analíticas, atributos temporais derivados, datasets longitudinais, registros de cobertura temporal, documentação de periodicidade e relatórios de qualidade temporal.

10. Subpipeline de Preparação de Dados Espaciais

A Preparação de Dados Espaciais contempla dados associados a localização, coordenadas, territórios, regiões, unidades espaciais ou objetos geográficos. Inclui operações como geocodificação, validação de coordenadas, padronização de sistemas de

referência, transformação de projeções, associação a unidades territoriais, tratamento de geometrias inválidas, cálculo de distâncias, agregações espaciais, junções espaciais e validação topológica.
As saídas incluem camadas geográficas tratadas, bases georreferenciadas, atributos territoriais derivados, matrizes espaciais, mapas de cobertura, registros de projeção, relatórios de validade topológica e documentação das operações espaciais.

11. Subpipeline de Preparação de Dados Textuais e Semânticos

A Preparação de Dados Textuais e Semânticos contempla os processos necessários para transformar textos livres, respostas abertas, comentários, avaliações, fóruns, redes sociais, e-mails, chats, documentos, notícias, transcrições e registros qualitativos em ativos analíticos adequados para mineração de textos, PLN, análise semântica, LLMs e tarefas analíticas posteriores.
Esse subpipeline inclui organização e unificação de fontes textuais, preservação do corpus bruto, remoção ou marcação de duplicidades, anonimização, remoção de HTML, URLs, ruídos, caracteres indesejados e problemas de codificação, tratamento de emojis quando necessário, normalização de caixa e acentuação, tokenização, segmentação de sentenças, stemming, lematização, tratamento de stopwords, preservação de negações e intensificadores, detecção de idioma, correção ortográfica, expansão de abreviações, uso de dicionários de domínio, extração de entidades, construção de vocabulário, vetorização, geração de matrizes termo-documento, geração de matrizes TF-IDF, geração de embeddings e registro de metadados textuais.
As saídas incluem corpus bruto preservado, corpus limpo, corpus anonimizado, corpus normalizado, vocabulário, dicionários de domínio, matriz termo-documento, matriz TF-IDF, embeddings persistidos, metadados textuais, relatório de qualidade textual, relatório de anonimização, pipeline NLP documentado, registros de pré-processamento, manifesto de execução, versões do corpus e lineage textual.
Tarefas como análise de sentimentos, classificação de emoções, detecção de tópicos, sumarização, detecção de toxicidade e extração de entidades não são definidas como preparação. Elas são tarefas analíticas posteriores que dependem da qualidade da preparação textual. Essa separação é metodologicamente importante porque preserva a distinção entre Engenharia de Dados e mineração, modelagem ou análise baseada em IA.

12. Subpipeline de Preparação de Dados em Grafos

A Preparação de Dados em Grafos contempla dados representados por entidades e relações. Inclui definição de nós, arestas, tipos de relação, direção, pesos, atributos de nós, atributos de arestas, remoção de relações inválidas, resolução de entidades, deduplicação de nós, construção de matrizes de adjacência, normalização de pesos, validação de conectividade e geração de atributos estruturais preliminares.
As saídas incluem grafos tratados, tabelas de nós e arestas, matrizes de adjacência, métricas estruturais preliminares, atributos relacionais derivados, registros de construção do grafo e documentação das regras de relacionamento.

13. Subpipeline de Preparação de Dados Multimodais

A Preparação de Dados Multimodais contempla dados compostos por diferentes modalidades, como texto, imagem, áudio, vídeo, sensores, tabelas e eventos. Inclui operações de sincronização, alinhamento temporal, padronização de formatos, extração de metadados, associação entre modalidades, validação de completude multimodal e geração de representações intermediárias.
As saídas incluem datasets multimodais alinhados, metadados por modalidade, registros de sincronização, embeddings por modalidade, representações intermediárias e documentação das relações entre as modalidades.

14. Subpipeline de Preparação de Dados em Fluxo ou Streaming

A Preparação de Dados em Fluxo ou Streaming contempla dados produzidos continuamente ou com baixa latência. Inclui ingestão incremental, ordenação de eventos, tratamento de eventos atrasados ou fora de ordem, definição de janelas, agregações contínuas, detecção de anomalias em fluxo, controle de estado, validação em tempo de execução, monitoramento de latência e geração de snapshots.
As saídas incluem fluxos tratados, janelas de eventos, agregações incrementais, logs de processamento, métricas de latência, registros de estado, alertas, snapshots versionados e relatórios de estabilidade operacional.

15. Artefatos relacionados

A dimensão pode produzir ou orientar diferentes artefatos, de acordo com o subpipeline utilizado: especificação do subpipeline de preparação; registro de regras de transformação por natureza de dado; dataset tratado por tipo de dado; corpus textual preparado; vocabulário e dicionários de domínio; matriz termo-documento; matriz TF-IDF; embeddings persistidos; séries temporais tratadas; janelas analíticas; camadas geoespaciais tratadas; grafos tratados; tabelas de nós e arestas; dataset multimodal alinhado; fluxos processados; snapshots de streaming; relatório de qualidade específico por natureza de dado; relatório de impacto das transformações; manifesto de execução; registro de versionamento; e registro de lineage dos artefatos produzidos.

16. Métricas associadas

As métricas devem ser definidas conforme a natureza do dado. Para dados estruturados, podem ser avaliadas completude, consistência, validade de domínio, unicidade, integridade e conformidade. Para dados temporais, podem ser avaliadas cobertura temporal, regularidade, lacunas, latência, eventos fora de ordem e consistência de janelas. Para dados espaciais, podem ser avaliadas taxa de geocodificação válida, validade topológica, cobertura territorial, precisão espacial e consistência de projeção.
Para dados textuais e semânticos, podem ser avaliadas taxa de textos válidos, taxa de textos vazios, duplicidade textual, taxa de ruído, cobertura lexical, diversidade textual, tamanho médio dos documentos, distribuição de idiomas, taxa de anonimização, consistência dos metadados textuais e qualidade dos rótulos quando existirem. Para grafos, podem ser avaliadas conectividade, densidade, taxa de nós duplicados, taxa de arestas

válidas, coerência de tipos de relação e estabilidade estrutural. Para dados multimodais, podem ser avaliadas completude por modalidade, taxa de alinhamento, sincronização temporal e consistência entre modalidades. Para streaming, podem ser avaliadas latência, throughput, taxa de eventos fora de ordem, taxa de perda, estabilidade de processamento e grau de reprodutibilidade do fluxo.

17. Critérios de validação

A dimensão é considerada adequadamente aplicada quando o subpipeline de preparação está explicitamente associado à natureza do dado; quando as operações executadas são compatíveis com essa natureza; quando os parâmetros e regras estão documentados; quando há métricas de qualidade específicas; quando os artefatos intermediários e finais estão versionados; quando existe registro de lineage; quando as restrições de governança são respeitadas; e quando as saídas são adequadas ao uso nos módulos posteriores do UnespDataLens-RM.

18. Relação com os módulos do UnespDataLens-RM

A dimensão relaciona-se diretamente com o Módulo 1, pois depende da identificação da natureza das fontes; com o Módulo 2, pois utiliza dados brutos preservados; com o Módulo 3, pois recebe dados integrados ou harmonizados; com o Módulo 4, pois orienta a execução das operações de preparação; com o Módulo 5, pois define métricas de qualidade específicas; com o Módulo 6, pois pode revelar desbalanceamentos, vieses ou limitações de representação; com o Módulo 9, pois depende de metadados e semântica; com o Módulo 10, pois envolve governança, segurança e privacidade; com o Módulo 11, pois exige lineage; com o Módulo 12, pois exige versionamento e reprodutibilidade; com o Módulo 13, pois prepara dados para representações analíticas; com o Módulo 14, pois subsidia feature engineering; com o Módulo 15, pois produz ativos monitoráveis; e com o Módulo 16, pois pode ser consultada e explicada por agentes inteligentes.

19. Exemplo de aplicação: preparação textual para análise de sentimentos

Em um cenário de análise de sentimentos, o UnespDataLens-RM não trata a análise de sentimentos como etapa de preparação. A preparação ocorre no subpipeline de Dados Textuais e Semânticos, onde comentários, respostas abertas ou mensagens são coletados, preservados em corpus bruto, limpos, anonimizados, normalizados, enriquecidos, vetorizados, documentados, versionados e avaliados quanto à qualidade.

Após essa preparação, a análise de sentimentos pode ser executada como técnica analítica posterior, utilizando léxicos, classificadores supervisionados, modelos de linguagem, embeddings ou abordagens híbridas. O produto analítico resultante, como escore de sentimento, classe de polaridade, indicador agregado ou dashboard, deve manter vínculo de lineage com o corpus, as regras de pré-processamento, a representação textual, o modelo utilizado e a versão dos artefatos.

Essa separação fortalece o modelo de referência porque impede que uma técnica específica seja confundida com a etapa de Engenharia de Dados. A contribuição do UnespDataLens-RM está em preparar os dados textuais de forma governada, rastreável,

versionada e reprodutível, permitindo que diferentes tarefas de PLN e IA sejam realizadas posteriormente com maior confiabilidade.

20. Contribuição para o modelo de referência

A dimensão Preparação Especializada por Natureza do Dado contribui para o UnespDataLens-RM ao explicitar que a diversidade dos dados exige pipelines de preparação especializados, e não apenas operações genéricas de ETL. Essa contribuição amplia a capacidade do modelo de referência para lidar com cenários contemporâneos de Inteligência Artificial, PLN, mineração de textos, análise semântica, grafos, dados espaciais, séries temporais, dados multimodais e streaming.
A dimensão também fortalece a generalidade do modelo sem torná-lo excessivamente específico. Em vez de criar módulos para tarefas particulares, como análise de sentimentos, classificação de emoções ou detecção de tópicos, o modelo formaliza a preparação textual e semântica como uma capacidade de Engenharia Analítica de Dados. Desse modo, diferentes tarefas analíticas podem ser instanciadas posteriormente sobre ativos preparados com rigor metodológico.

21. Síntese da dimensão

A dimensão Preparação Especializada por Natureza do Dado organiza os processos de preparação conforme o tipo de dado utilizado no ecossistema analítico. Ela atua de forma articulada ao Módulo 4 e orienta a definição de subpipelines, operações, métricas, artefatos, critérios de validação, versionamento e lineage para dados estruturados, temporais, espaciais, textuais, semânticos, em grafos, multimodais e em fluxo contínuo.
Com essa dimensão, o UnespDataLens-RM mantém sua natureza de modelo de referência genérico e reutilizável, ao mesmo tempo em que reconhece a complexidade técnica e metodológica dos dados contemporâneos. A preparação textual para análise de sentimentos torna-se, assim, um exemplo de aplicação da dimensão, e não um módulo específico do modelo.

Dimensão Complementar — Data Overload

1. Definição da dimensão

A dimensão **Data Overload**, ou **sobrecarga de dados**, refere-se à condição em que a quantidade, diversidade, dispersão, redundância, velocidade de produção ou complexidade dos dados disponíveis ultrapassa a capacidade dos usuários, sistemas, equipes técnicas e processos analíticos de localizar, compreender, selecionar, interpretar e utilizar esses dados de maneira eficiente, confiável e orientada à decisão.

No contexto do **UnespDataLens-RM**, a sobrecarga de dados não deve ser compreendida apenas como excesso de volume. Embora o crescimento do volume de registros seja um fator relevante, a sobrecarga também pode decorrer da multiplicidade de fontes, da existência de muitos atributos, da produção de versões sucessivas de datasets, da criação redundante de dashboards, da proliferação de indicadores semelhantes, da geração

excessiva de features, da baixa documentação dos ativos, da duplicação de relatórios e da dificuldade de distinguir quais produtos analíticos são oficiais, atualizados, confiáveis ou adequados a determinado uso.

Assim, a sobrecarga de dados ocorre quando há abundância de dados, mas baixa capacidade de transformá-los em conhecimento útil. Nessa situação, o problema não é a inexistência de informação, mas o excesso de ativos pouco organizados, pouco priorizados, pouco documentados ou pouco governados. Esse cenário dificulta a descoberta de dados relevantes, aumenta o esforço de análise, reduz a confiança nos produtos analíticos e pode comprometer a tomada de decisão baseada em evidências.

No UnespDataLens-RM, a dimensão Data Overload é tratada como uma capacidade complementar e transversal, pois se manifesta em diferentes partes do modelo de referência: no inventário de fontes, na integração de bases, na criação de atributos derivados, no armazenamento analítico, na disponibilização de produtos, na feature store, no catálogo de dados, no monitoramento de uso e na atuação dos agentes inteligentes.

2. Problema tratado

Em ecossistemas analíticos complexos, diferentes sistemas, equipes, projetos e áreas institucionais tendem a produzir continuamente novas bases, relatórios, planilhas, dashboards, APIs, indicadores, features, versões e produtos de dados. Quando esses ativos crescem sem critérios claros de curadoria, catalogação, priorização, governança e descontinuação, o ambiente analítico passa a apresentar excesso de informação disponível e baixa capacidade de uso efetivo.

Essa situação pode gerar diversos problemas. Usuários podem não saber qual base utilizar, quais dados estão atualizados, qual indicador é oficial, qual dashboard substitui versões anteriores, quais atributos são confiáveis, quais produtos possuem qualidade suficiente, quais dados possuem restrição de acesso e quais ativos já estão obsoletos. Como consequência, diferentes usuários podem tomar decisões com base em versões distintas ou interpretações divergentes dos mesmos fenômenos.

A sobrecarga de dados pode produzir efeitos como:

- dificuldade de localizar ativos analíticos relevantes;
- aumento do tempo necessário para encontrar dados úteis;
- duplicação de datasets, dashboards, relatórios e indicadores;
- criação de múltiplas versões de uma mesma informação;
- uso de ativos obsoletos, incompletos ou não oficiais;
- baixa reutilização de produtos de dados existentes;
- proliferação de features redundantes ou pouco relevantes;
- aumento da dependência de conhecimento tácito;
- dificuldade de distinguir dados confiáveis de dados experimentais;
- maior risco de interpretações conflitantes;
- aumento do custo de manutenção do ecossistema analítico;
- acúmulo de produtos sem responsável, sem uso ou sem atualização;
- redução da confiança institucional nos dados;
- dificuldade de priorizar ações analíticas e decisórias.

No UnespDataLens-RM, essa dimensão reconhece que a maturidade de um modelo de referência para Engenharia Analítica de Dados não depende apenas da capacidade de produzir, integrar e disponibilizar dados, mas também da capacidade de **organizar, priorizar, recomendar, reutilizar, consolidar e descontinuar ativos de dados**.

3. Objetivo da dimensão

O objetivo da dimensão **Data Overload** é **reduzir a sobrecarga informacional gerada pelo excesso de fontes, atributos, datasets, versões, dashboards, APIs, indicadores, features e produtos analíticos**, por meio de mecanismos de catalogação, curadoria, classificação, priorização, recomendação, monitoramento de uso, redução de redundância e descontinuação controlada de ativos obsoletos ou pouco relevantes.

Essa dimensão busca assegurar que o ecossistema analítico do UnespDataLens-RM permaneça utilizável, navegável, compreensível e orientado à decisão, mesmo diante do crescimento contínuo do número de dados e produtos.

4. Relação com o UnespDataLens-RM

A dimensão Data Overload se articula com diversos módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, contribui para evitar a inclusão indiscriminada de fontes sem finalidade clara, priorizando fontes relevantes, documentadas, com responsável definido e alinhadas aos objetivos analíticos.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, orienta a criação de critérios para ingestão controlada, evitando a acumulação desnecessária de cópias, arquivos, versões e snapshots sem valor analítico ou sem política de retenção.

No **Módulo 3 — Integração e Harmonização**, contribui para identificar redundâncias entre bases, atributos equivalentes, entidades duplicadas e informações semanticamente sobrepostas.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, orienta a seleção criteriosa de atributos e evita a produção excessiva de variáveis derivadas sem justificativa analítica, sem uso previsto ou sem documentação.

No **Módulo 7 — Armazenamento Analítico**, apoia a organização de zonas, data marts, views e datasets curados, evitando o acúmulo descontrolado de estruturas analíticas semelhantes ou obsoletas.

No **Módulo 8 — Disponibilização e Consumo Analítico**, contribui para priorizar produtos úteis, reduzir dashboards redundantes, sinalizar produtos depreciados e orientar usuários para os ativos mais adequados.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, encontra seu principal mecanismo de mitigação, pois o catálogo, os metadados e o glossário tornam os ativos localizáveis, compreensíveis e comparáveis.

No **Módulo 14 — Feature Engineering e Feature Store**, ajuda a evitar a proliferação de features duplicadas, redundantes, instáveis, pouco utilizadas ou semanticamente equivalentes.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, permite identificar ativos pouco acessados, desatualizados, redundantes, sem responsável ou com baixo valor de uso.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, pode ser tratada por mecanismos de recomendação, busca em linguagem natural, sumarização, orientação ao usuário e priorização de ativos relevantes.

5. Estratégias de mitigação

A mitigação da sobrecarga de dados no UnespDataLens-RM pode envolver estratégias técnicas, metodológicas e de governança.

Estratégia	Descrição
Catalogação estruturada	Registrar fontes, datasets, dashboards, APIs, features e produtos com descrição, finalidade, versão, qualidade, responsável e restrições.
Curadoria contínua	Revisar periodicamente ativos existentes para manter, consolidar, atualizar, arquivar ou descontinuar produtos.
Classificação por relevância	Priorizar ativos conforme utilidade, qualidade, criticidade, frequência de uso, domínio e finalidade.
Sinalização de status	Marcar ativos como oficial, experimental, em validação, ativo, depreciado, obsoleto, restrito ou recomendado.
Redução de redundância	Identificar datasets, dashboards, indicadores e features equivalentes ou sobrepostos.
Governança de publicação	Definir critérios mínimos para que um novo produto seja disponibilizado aos usuários.
Monitoramento de uso	Acompanhar acessos, reutilização, frequência de consumo e produtos sem uso.
Recomendação de ativos	Utilizar agentes inteligentes ou mecanismos de busca para sugerir dados adequados à finalidade do usuário.
Documentação mínima obrigatória	Exigir metadados essenciais antes da publicação ou reutilização de um ativo.
Descontinuação controlada	Arquivar, substituir ou retirar ativos obsoletos mantendo histórico e rastreabilidade.
Consolidação de produtos semelhantes	Unificar dashboards, indicadores ou datasets redundantes em produtos oficiais.
Gestão de versões	Diferenciar versões vigentes, históricas, experimentais e depreciadas.

6. Métricas associadas

A dimensão Data Overload pode ser avaliada por métricas relacionadas à quantidade, redundância, uso, descoberta, obsolescência e reutilização dos ativos analíticos.

Métrica	Sigla	Descrição
Número total de ativos catalogados	NTAC	Quantidade de fontes, datasets, produtos, dashboards, APIs e features registrados.
Taxa de ativos sem uso	TASU	Proporção de ativos sem acesso ou consumo em determinado período.
Taxa de ativos redundantes	TAR	Proporção de ativos com conteúdo, finalidade ou indicadores equivalentes.
Taxa de ativos obsoletos	TAO	Proporção de ativos desatualizados, depreciados ou sem manutenção.
Tempo médio de descoberta	TMD	Tempo necessário para localizar um ativo adequado.
Taxa de reutilização de ativos	TRA	Proporção de ativos reutilizados em mais de um produto, análise ou pipeline.
Taxa de ativos com finalidade documentada	TAFD	Proporção de ativos que possuem finalidade explícita.
Taxa de ativos com responsável definido	TARD	Proporção de ativos com responsável técnico ou institucional.
Taxa de ativos recomendados	TAREC	Proporção de ativos sinalizados como prioritários, oficiais ou recomendados.
Taxa de produtos depreciados	TPD	Proporção de produtos marcados como substituídos ou em descontinuação.
Índice de sobrecarga de dados	ISD	Indicador composto de redundância, obsolescência, baixo uso e dificuldade de descoberta.

Uma formulação possível para o **Índice de Sobrecarga de Dados** é:

$$\mathbf{ISD} = w_1 TASU + w_2 TAR + w_3 TAO + w_4(1 - TRA) + w_5 TMD_n$$

em que:

- **TASU** representa a taxa de ativos sem uso;
- **TAR** representa a taxa de ativos redundantes;
- **TAO** representa a taxa de ativos obsoletos;
- **TRA** representa a taxa de reutilização de ativos;
- $\mathbf{TMD_n}$ representa o tempo médio de descoberta normalizado;
- w_i representa o peso atribuído a cada dimensão.

Quanto maior o valor do ISD, maior a evidência de sobrecarga de dados no ecossistema analítico.

7. Artefatos relacionados

A dimensão Data Overload pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de catálogo, disponibilização, monitoramento, feature store e agentes inteligentes.

Código	Artefato	Finalidade
DO1	Mapa de Ativos Analíticos	Identificar fontes, datasets, dashboards, APIs, features, indicadores e produtos existentes.
DO2	Relatório de Redundância de Ativos	Detectar ativos semelhantes, duplicados, sobrepostos ou semanticamente equivalentes.
DO3	Relatório de Uso dos Ativos	Avaliar acesso, consumo, reutilização, frequência de uso e abandono.
DO4	Registro de Ativos Obsoletos	Identificar ativos desatualizados, sem responsável, sem uso ou substituídos.
DO5	Matriz de Relevância dos Ativos	Classificar ativos por utilidade, qualidade, criticidade, uso e adequação ao domínio.
DO6	Plano de Curadoria e Descontinuação	Definir ações para manter, revisar, consolidar, arquivar ou excluir ativos.
DO7	Índice de Descoberta de Dados	Avaliar facilidade de localização de ativos relevantes.
DO8	Registro de Recomendações de Uso	Sugerir ativos oficiais, prioritários ou mais adequados a determinada finalidade.
DO9	Lista de Produtos Oficiais	Consolidar dashboards, indicadores, APIs e datasets considerados referência.
DO10	Registro de Substituição de Ativos	Documentar quando um ativo foi substituído por outro mais atualizado ou confiável.

8. Critérios de validação

A dimensão Data Overload pode ser considerada adequadamente tratada quando:

1. os principais ativos analíticos estão catalogados;
2. cada ativo possui finalidade, responsável, versão, qualidade e status documentados;
3. ativos redundantes ou semanticamente equivalentes são identificados;
4. ativos obsoletos, depreciados ou sem uso são sinalizados;
5. existe processo de curadoria periódica;

6. produtos oficiais ou recomendados são distinguíveis de produtos experimentais;
7. usuários conseguem localizar ativos relevantes com menor esforço;
8. dashboards, APIs, datasets e features possuem critérios de publicação;
9. produtos pouco utilizados são revisados, consolidados ou descontinuados;
10. versões antigas são preservadas apenas quando possuem justificativa histórica, científica, legal ou operacional;
11. mecanismos de recomendação, busca ou agentes inteligentes apoiam a descoberta de ativos;
12. a taxa de reutilização de ativos aumenta ao longo do tempo;
13. o tempo médio de descoberta de dados diminui;
14. a criação de novos produtos passa por critérios mínimos de governança e documentação.

9. Exemplo de aplicação no domínio educacional

Em um projeto educacional, a sobrecarga de dados pode ocorrer quando existem múltiplas bases sobre estudantes, matrículas, notas, frequência, evasão, interações no ambiente virtual, desempenho, questionários e indicadores institucionais, mas sem clareza sobre qual base é oficial, qual está atualizada e qual deve ser usada para cada finalidade.

Por exemplo, podem existir vários dashboards de evasão produzidos em momentos diferentes, cada um com definição distinta de evasão, período de referência diferente, regra de cálculo específica e bases de origem distintas. Sem curadoria, usuários podem utilizar painéis divergentes e tomar decisões com base em informações inconsistentes.

A dimensão Data Overload permite organizar esses ativos, identificar redundâncias, definir produtos oficiais, documentar diferenças entre versões, sinalizar dashboards depreciados e recomendar o produto mais adequado para cada finalidade. Também permite identificar features educacionais duplicadas, bases pouco utilizadas e indicadores que precisam ser consolidados.

10. Exemplo de aplicação em políticas públicas

Em políticas públicas, a sobrecarga pode ocorrer quando diferentes secretarias produzem múltiplos relatórios, planilhas, indicadores e painéis sobre territórios, famílias, atendimentos, vulnerabilidade, escolas, unidades de saúde e equipamentos públicos.

Sem organização, gestores podem ter dificuldade para localizar o indicador correto, compreender sua origem, saber se está atualizado ou identificar se há outro produto mais confiável.

A dimensão Data Overload permite catalogar os produtos, classificar ativos por relevância, identificar indicadores duplicados, monitorar uso, arquivar produtos obsoletos e orientar gestores para os dados mais adequados à decisão. Também pode apoiar a construção de uma lista de produtos oficiais, reduzindo inconsistências entre áreas e melhorando a confiança no uso de dados integrados.

11. Contribuição para o modelo de referência

A dimensão **Data Overload** contribui para o UnespDataLens-RM ao reconhecer que a produção contínua de dados e produtos analíticos, quando não acompanhada de curadoria e governança, pode gerar complexidade excessiva, redundância e perda de efetividade.

Essa dimensão fortalece o modelo de referência ao:

- reduzir excesso de ativos sem uso;
- melhorar a descoberta de dados;
- aumentar a reutilização de ativos;
- evitar duplicidade de produtos analíticos;
- reduzir confusão entre versões;
- apoiar curadoria contínua;
- orientar a publicação responsável de novos produtos;
- melhorar governança dos ativos analíticos;
- reduzir dívida de dados;
- tornar a tomada de decisão mais objetiva;
- apoiar agentes inteligentes na recomendação de dados relevantes;
- favorecer a consolidação de produtos oficiais e confiáveis.

12. Síntese da dimensão

A dimensão **Data Overload** trata da sobrecarga gerada pelo excesso de dados, fontes, atributos, versões, dashboards, APIs, indicadores, features e produtos analíticos em ecossistemas complexos. No UnespDataLens-RM, essa dimensão atua como uma capacidade complementar para garantir que os dados produzidos sejam não apenas abundantes, mas organizados, relevantes, localizáveis, reutilizáveis, governados e adequados ao uso.

Ao incorporar estratégias de catalogação, curadoria, priorização, monitoramento de uso, redução de redundância, recomendação e descontinuação controlada, o modelo de referência reduz a complexidade informacional e aumenta a efetividade dos dados na produção de conhecimento e na tomada de decisão baseada em evidências.

Dimensão Complementar — Data Silos

1. Definição da dimensão

A dimensão **Data Silos**, ou **silos de dados**, refere-se à condição em que dados relevantes para análise, gestão, pesquisa ou tomada de decisão permanecem isolados em sistemas, setores, equipes, bases, planilhas, plataformas ou repositórios distintos, sem integração adequada, sem interoperabilidade semântica, sem padronização técnica e sem mecanismos consistentes de compartilhamento, governança e reutilização.

No contexto do **UnespDataLens-RM**, os silos de dados representam uma das principais barreiras para a construção de ativos analíticos integrados, confiáveis e úteis à tomada de decisão baseada em evidências. Dados podem existir em diferentes sistemas institucionais, ambientes acadêmicos, plataformas administrativas, bases externas, arquivos locais, planilhas departamentais, dashboards isolados, repositórios de pesquisa ou bases públicas, mas sem articulação entre si.

Essa fragmentação dificulta a compreensão ampla dos fenômenos analisados. Em muitos casos, cada fonte contém apenas uma parte da realidade: uma base registra matrícula, outra registra desempenho, outra registra frequência, outra registra interações, outra registra dados socioeconômicos e outra registra indicadores institucionais. Quando essas bases permanecem isoladas, torna-se difícil analisar trajetórias, padrões, relações, causas, riscos, desigualdades, impactos e oportunidades de intervenção.

Assim, a dimensão Data Silos trata da necessidade de superar o isolamento dos dados por meio de inventário, integração, harmonização, interoperabilidade, metadados, semântica, governança, lineage, armazenamento analítico e disponibilização controlada de ativos integrados.

2. Problema tratado

Em ecossistemas analíticos complexos, os dados frequentemente são produzidos por diferentes sistemas e áreas organizacionais, cada uma com seus próprios formatos, regras, identificadores, periodicidades, definições, responsáveis, restrições e finalidades. Essa organização descentralizada pode ser necessária do ponto de vista operacional, mas tende a produzir fragmentação quando não há mecanismos adequados de integração e governança.

Os silos de dados podem gerar problemas como:

- dificuldade de localizar fontes relevantes;
- impossibilidade de cruzar dados de diferentes sistemas;
- duplicação de informações em bases distintas;
- divergência entre indicadores produzidos por diferentes áreas;
- ausência de visão integrada sobre estudantes, cidadãos, serviços, processos ou territórios;
- baixa interoperabilidade entre sistemas;
- uso de identificadores incompatíveis;
- ausência de chaves comuns para integração;
- diferença de granularidade entre bases;
- dificuldade de harmonizar conceitos e categorias;
- dependência de extrações manuais e planilhas paralelas;

- baixa rastreabilidade da origem dos dados;
- aumento de inconsistências e retrabalho;
- perda de qualidade analítica;
- enfraquecimento da governança institucional dos dados.

No UnespDataLens-RM, os silos de dados são tratados como uma dimensão complementar porque o modelo de referência busca exatamente transformar dados dispersos e heterogêneos em ativos analíticos integrados, documentados, governados e reutilizáveis. A superação dos silos não significa centralizar tudo de forma indiscriminada, mas criar condições técnicas, semânticas e institucionais para que os dados possam ser conectados, compreendidos, governados e utilizados de forma adequada.

3. Objetivo da dimensão

O objetivo da dimensão **Data Silos** é **reduzir o isolamento técnico, semântico, organizacional e operacional dos dados**, promovendo integração, interoperabilidade, compartilhamento controlado, harmonização conceitual, rastreabilidade e governança entre fontes, sistemas, áreas e produtos analíticos.

Essa dimensão busca garantir que os dados do ecossistema UnespDataLens-RM possam ser articulados em uma visão analítica mais ampla, permitindo análises integradas, indicadores consistentes, produtos de dados reutilizáveis e decisões baseadas em evidências mais completas.

4. Relação com o UnespDataLens-RM

A dimensão Data Silos se articula diretamente com diversos módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, os silos são inicialmente identificados por meio do mapeamento das fontes, sistemas, responsáveis, formatos, periodicidades, granularidades, restrições e limitações conhecidas.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, os dados isolados são trazidos para um ambiente controlado, preservando sua forma original e sua proveniência.

No **Módulo 3 — Integração e Harmonização**, ocorre o enfrentamento mais direto dos silos, por meio da identificação de chaves, entidades comuns, regras de relacionamento, compatibilidade de granularidade, equivalências semânticas e resolução de conflitos.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, os dados oriundos de silos distintos são padronizados, limpos e enriquecidos para permitir uso conjunto.

No **Módulo 5 — Qualidade e Validação de Dados**, são avaliadas inconsistências produzidas pela integração de fontes isoladas, como duplicidades, divergências, valores ausentes e conflitos de domínio.

No **Módulo 7 — Armazenamento Analítico**, os dados integrados são organizados em estruturas persistentes, como datasets curados, data marts, views, lakehouses ou repositórios analíticos.

No **Módulo 8 — Disponibilização e Consumo Analítico**, os ativos integrados são disponibilizados como produtos de dados, dashboards, APIs, relatórios ou bases analíticas para consumo controlado.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, a redução de silos é apoiada pela documentação de fontes, atributos, conceitos, equivalências semânticas e relações entre ativos.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, são definidos critérios para compartilhamento controlado entre áreas, respeitando sensibilidade, finalidade, permissão e base institucional.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, registra-se como dados provenientes de silos distintos foram combinados, transformados e utilizados.

No **Módulo 13 — Representações Analíticas**, dados antes isolados podem ser reorganizados em representações tabulares, temporais, espaciais, textuais, semânticas e em grafos, permitindo análises integradas de entidades e relações.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, agentes podem auxiliar usuários a descobrir fontes relacionadas, compreender equivalências, identificar produtos integrados e sugerir conexões entre ativos.

5. Estratégias de mitigação

A mitigação de silos de dados no UnespDataLens-RM pode envolver estratégias técnicas, semânticas, organizacionais e de governança.

Estratégia	Descrição
Inventário de fontes	Mapear sistemas, bases, planilhas, APIs, repositórios e responsáveis pelos dados.
Mapeamento de entidades comuns	Identificar entidades compartilhadas entre fontes, como estudante, curso, município, território, cidadão, atendimento ou serviço.
Definição de chaves de integração	Estabelecer identificadores, chaves compostas ou regras de correspondência entre bases.
Harmonização semântica	Resolver diferenças conceituais entre termos, categorias, códigos e definições.
Padronização técnica	Compatibilizar formatos, tipos, domínios, codificações, datas e unidades de medida.
Integração controlada	Combinar fontes com registro de regras, conflitos, limitações e qualidade da integração.
Catálogo unificado	Registrar fontes e ativos em catálogo comum, permitindo descoberta e reutilização.

Governança de compartilhamento	Definir regras de acesso, finalidade, responsabilidade e restrições para uso entre áreas.
Modelos semânticos comuns	Criar vocabulários, glossários, ontologias leves ou modelos conceituais compartilhados.
APIs e produtos de dados integrados	Disponibilizar ativos integrados de forma controlada e reutilizável.
Data marts por domínio	Organizar dados integrados em estruturas analíticas orientadas a temas ou áreas.
Lineage entre fontes e produtos	Registrar como fontes isoladas foram combinadas para produzir ativos integrados.

6. Métricas associadas

A dimensão Data Silos pode ser avaliada por métricas relacionadas à integração, interoperabilidade, compartilhamento, redundância e cobertura das fontes.

Métrica	Sigla	Descrição
Número de fontes isoladas identificadas	NFII	Quantidade de fontes relevantes ainda não integradas ou catalogadas.
Taxa de fontes catalogadas	TFC	Proporção de fontes registradas em catálogo comum.
Taxa de fontes integradas	TFI	Proporção de fontes integradas ao ambiente analítico.
Taxa de entidades com chave comum	TECC	Proporção de entidades que possuem identificador ou regra de correspondência.
Taxa de atributos harmonizados	TAH	Proporção de atributos equivalentes harmonizados entre fontes.
Taxa de conflitos semânticos resolvidos	TCSR	Proporção de conflitos conceituais tratados.
Taxa de ativos integrados reutilizados	TAIR	Proporção de ativos integrados usados em mais de um produto ou análise.
Taxa de redundância entre silos	TRS	Proporção de informações duplicadas em fontes distintas.
Índice de interoperabilidade analítica	IIA	Grau em que fontes podem ser combinadas para análise.
Índice de redução de silos	IRS	Indicador composto de catalogação, integração,

harmonização e reuso.

Uma formulação possível para o **Índice de Redução de Silos** é:

$$\mathbf{IRS} = \mathbf{w_1TFC} + \mathbf{w_2TFI} + \mathbf{w_3TECC} + \mathbf{w_4TAH} + \mathbf{w_5TAIR} - \mathbf{w_6TRS}$$

em que:

- **TFC** representa a taxa de fontes catalogadas;
- **TFI** representa a taxa de fontes integradas;
- **TECC** representa a taxa de entidades com chave comum;
- **TAH** representa a taxa de atributos harmonizados;
- **TAIR** representa a taxa de ativos integrados reutilizados;
- **TRS** representa a taxa de redundância entre silos;
- w_i representa o peso atribuído a cada dimensão.

Quanto maior o IRS, maior o grau de superação dos silos no ecossistema analítico.

7. Artefatos relacionados

A dimensão Data Silos pode ser operacionalizada por meio de artefatos complementares, articulados aos módulos de inventário, integração, catálogo, governança e linhagem.

Código	Artefato	Finalidade
DS1	Mapa de Silos de Dados	Identificar fontes, sistemas, áreas e bases isoladas.
DS2	Matriz de Fontes por Domínio	Relacionar fontes a áreas, temas, entidades e finalidades.
DS3	Mapa de Entidades Compartilhadas	Identificar entidades comuns entre diferentes fontes.
DS4	Matriz de Chaves de Integração	Registrar chaves, identificadores e regras de correspondência.
DS5	Mapa de Equivalência Semântica	Harmonizar termos, atributos, categorias e conceitos.
DS6	Registro de Conflitos entre Fontes	Documentar divergências técnicas, semânticas e operacionais.
DS7	Plano de Integração entre Silos	Definir prioridade, regras, riscos e estratégia de integração.
DS8	Catálogo Unificado de Fontes	Disponibilizar visão integrada das fontes existentes.
DS9	Relatório de	Avaliar capacidade de combinação entre fontes.

	Interoperabilidade	
DS10	Registro de Ativos Integrados	Documentar datasets, data marts ou produtos resultantes da integração.

8. Critérios de validação

A dimensão Data Silos pode ser considerada adequadamente tratada quando:

1. as principais fontes isoladas foram identificadas;
2. os responsáveis por cada fonte estão documentados;
3. as fontes relevantes estão catalogadas;
4. entidades comuns entre fontes foram mapeadas;
5. chaves de integração foram definidas ou justificadas como inexistentes;
6. conflitos técnicos e semânticos foram registrados;
7. atributos equivalentes foram harmonizados;
8. restrições de acesso e compartilhamento foram documentadas;
9. ativos integrados foram produzidos com lineage registrado;
10. produtos analíticos integrados são reutilizados por diferentes análises ou usuários;
11. redundâncias entre bases foram identificadas e reduzidas;
12. usuários conseguem localizar fontes relacionadas a um mesmo domínio ou fenômeno;
13. a integração de fontes respeita regras de governança, segurança e privacidade;
14. a visão analítica integrada melhora a compreensão do fenômeno estudado.

9. Exemplo de aplicação no domínio educacional

Em um projeto educacional, os dados de interesse podem estar distribuídos em diferentes silos: sistema acadêmico, ambiente virtual de aprendizagem, questionários socioeconômicos, registros de bolsas, histórico escolar, frequência, biblioteca, sistemas de apoio pedagógico e bases institucionais externas.

Cada sistema pode utilizar identificadores, formatos, períodos e conceitos próprios. A situação acadêmica de um estudante pode estar em uma base, seu desempenho em outra, suas interações no AVA em outra e seu perfil socioeconômico em outra. Se essas fontes permanecem isoladas, análises sobre evasão, permanência, desempenho ou engajamento ficam incompletas.

A dimensão Data Silos permite mapear essas fontes, identificar entidades comuns, definir chaves de integração, harmonizar conceitos como matrícula, evasão, trancamento, aprovação e reprovação, produzir datasets integrados e disponibilizar produtos analíticos que representem a trajetória educacional de forma mais completa.

10. Exemplo de aplicação em políticas públicas

Em políticas públicas, os silos de dados são recorrentes porque diferentes secretarias e órgãos mantêm bases próprias sobre saúde, educação, assistência social, habitação, trabalho, território, orçamento, atendimento e infraestrutura.

Uma família em situação de vulnerabilidade pode aparecer em bases distintas, como cadastro social, atendimento de saúde, matrícula escolar, programas habitacionais e

benefícios públicos. Se essas bases não forem integradas ou, ao menos, interoperáveis, o gestor terá uma visão fragmentada da realidade.

A dimensão Data Silos permite mapear fontes por secretaria, identificar entidades comuns, harmonizar conceitos territoriais, compatibilizar períodos, documentar restrições de acesso e criar produtos integrados que apoiem análise intersetorial, priorização de ações e avaliação de políticas públicas.

11. Contribuição para o modelo de referência

A dimensão **Data Silos** contribui para o UnespDataLens-RM ao reconhecer que a fragmentação dos dados é uma barreira central para análises robustas, integradas e confiáveis.

Essa dimensão fortalece o modelo de referência ao:

- tornar visíveis fontes antes isoladas;
- reduzir fragmentação informacional;
- apoiar integração entre sistemas e áreas;
- melhorar interoperabilidade técnica e semântica;
- reduzir duplicidades e divergências entre bases;
- permitir visão analítica mais completa dos fenômenos;
- apoiar governança do compartilhamento de dados;
- preservar rastreabilidade entre fontes e produtos integrados;
- fortalecer o catálogo como ponto comum de descoberta;
- permitir análises interdisciplinares, intersetoriais e longitudinais;
- ampliar a reutilização de ativos integrados;
- reduzir dependência de planilhas e extrações manuais.

12. Síntese da dimensão

A dimensão **Data Silos** trata do isolamento de dados em sistemas, áreas, bases, planilhas, plataformas ou repositórios distintos, sem integração técnica, semântica e governada. No UnespDataLens-RM, essa dimensão atua como capacidade complementar para transformar fontes dispersas em ativos analíticos integrados, documentados, rastreáveis, interoperáveis e reutilizáveis.

Ao incorporar estratégias de inventário, integração, harmonização, catálogo unificado, governança de compartilhamento, lineage e produtos de dados integrados, o modelo de referência reduz a fragmentação informacional e amplia a capacidade de análise e tomada de decisão baseada em evidências.

Dimensão Complementar — Data Debt

1. Definição da dimensão

A dimensão **Data Debt**, ou **dívida de dados**, refere-se ao acúmulo de problemas, pendências, improvisações, inconsistências, lacunas documentais, decisões técnicas provisórias e fragilidades estruturais nos dados, nos pipelines e nos produtos analíticos, que passam a comprometer gradualmente a qualidade, a confiabilidade, a rastreabilidade, a manutenção e a capacidade de evolução de um ecossistema de dados.

No contexto do **UnespDataLens-RM**, a dívida de dados representa o conjunto de passivos técnicos, metodológicos, semânticos e organizacionais associados ao ciclo de vida dos dados. Ela pode surgir quando fontes são incorporadas sem inventário adequado, datasets são criados sem documentação, transformações são realizadas manualmente, regras de negócio não são versionadas, atributos derivados são gerados sem rastreabilidade, dashboards são publicados sem validação, dados são armazenados sem política de retenção, ou produtos analíticos permanecem ativos mesmo após se tornarem obsoletos.

A dívida de dados não se manifesta necessariamente como erro imediato. Muitas vezes, ela aparece como uma solução provisória que permite avançar rapidamente em um projeto, mas que cria dificuldades futuras para manutenção, reprocessamento, auditoria, integração, governança e reuso. Assim como ocorre com a dívida técnica em desenvolvimento de software, a dívida de dados pode crescer silenciosamente até tornar o ecossistema analítico difícil de compreender, corrigir, reproduzir ou escalar.

No UnespDataLens-RM, essa dimensão é tratada como uma capacidade complementar e transversal, pois a dívida de dados pode se acumular em qualquer módulo do modelo de referência: no inventário de fontes, na ingestão, na integração, na transformação, na qualidade, na governança, na linhagem, no versionamento, nas representações analíticas, na feature store, no monitoramento e nos produtos de consumo.

2. Problema tratado

Em muitos ambientes institucionais, projetos analíticos são desenvolvidos sob pressão de tempo, com dados incompletos, documentação limitada, sistemas legados, integrações manuais, planilhas paralelas, regras implícitas e conhecimento concentrado em poucas pessoas. Essas condições favorecem a criação de soluções pontuais, funcionais no curto prazo, mas frágeis no médio e longo prazo.

A dívida de dados pode se manifestar de diversas formas:

- fontes utilizadas sem documentação adequada;
- ausência de responsável institucional ou técnico pelo dado;
- regras de transformação registradas apenas em scripts ou planilhas;
- datasets sem versão, sem lineage ou sem descrição;
- indicadores calculados de forma diferente por equipes distintas;
- atributos derivados sem fórmula documentada;
- dashboards publicados sem indicação de origem, data ou regra de cálculo;
- dados duplicados em diferentes repositórios;
- schemas alterados sem comunicação ou registro;
- pipelines dependentes de execução manual;
- ausência de testes de qualidade automatizados;

- metadados incompletos ou desatualizados;
- features criadas sem avaliação de vazamento, viés ou estabilidade;
- produtos de dados obsoletos mantidos em uso;
- decisões analíticas baseadas em conhecimento tácito.

Esses problemas aumentam o esforço necessário para compreender, corrigir ou reutilizar os dados. Com o tempo, a dívida de dados reduz a velocidade de novos projetos, aumenta o risco de erro, dificulta auditoria, reduz confiança nos produtos analíticos e compromete a sustentabilidade do ecossistema de dados.

No UnespDataLens-RM, a dimensão Data Debt busca tornar essa dívida visível, mensurável e gerenciável, evitando que fragilidades acumuladas comprometam a confiabilidade dos ativos analíticos.

3. Objetivo da dimensão

O objetivo da dimensão **Data Debt** é **identificar, classificar, mensurar, priorizar e reduzir passivos acumulados nos dados, pipelines, metadados, regras, versões, produtos e processos analíticos**, promovendo melhoria contínua, governança, rastreabilidade, documentação, reprodutibilidade e sustentabilidade do ecossistema UnespDataLens-RM.

Essa dimensão busca transformar problemas ocultos ou dispersos em registros explícitos, passíveis de acompanhamento, priorização e correção.

4. Relação com o UnespDataLens-RM

A dimensão Data Debt se articula com todos os módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, a dívida pode surgir quando fontes são utilizadas sem responsável, descrição, periodicidade, cobertura temporal, restrição de acesso ou limitação documentada.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, a dívida pode ocorrer quando ingestões são manuais, logs são incompletos, snapshots não são preservados ou dados brutos são armazenados sem manifesto.

No **Módulo 3 — Integração e Harmonização**, a dívida aparece quando regras de integração não são documentadas, chaves são escolhidas informalmente, conflitos são ignorados ou equivalências semânticas não são registradas.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, a dívida se acumula quando transformações são executadas sem registro, atributos derivados não possuem fórmula ou tratamentos de ausentes, duplicidades e outliers não são justificados.

No **Módulo 5 — Qualidade e Validação de Dados**, a dívida ocorre quando não existem testes de qualidade, limiares, relatórios de validação ou histórico de não conformidades.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, a dívida pode surgir quando distribuições de grupos não são avaliadas, atributos sensíveis são ignorados ou riscos de viés não são documentados.

No **Módulo 7 — Armazenamento Analítico**, a dívida aparece em estruturas sem organização por zonas, datasets duplicados, schemas não versionados, data marts sem responsável ou ativos sem política de retenção.

No **Módulo 8 — Disponibilização e Consumo Analítico**, a dívida se manifesta em dashboards sem documentação, APIs sem versão, produtos sem uso, datasets disponibilizados sem restrição clara ou relatórios sem data de referência.

No **Módulo 9 — Metadados, Catálogo e Semântica**, a dívida é reduzida por meio de documentação, glossários, dicionários, modelos semânticos e fichas de ativos.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, a dívida é mitigada por políticas, papéis, permissões, classificação de sensibilidade e registros de conformidade.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, a dívida é enfrentada pela capacidade de reconstruir origem, trajetória e dependências dos dados.

No **Módulo 12 — Versionamento e Reprodutibilidade**, a dívida é reduzida por versões, manifestos, ambientes preservados e pacotes de reprodutibilidade.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, a dívida é acompanhada por meio da identificação de ativos obsoletos, inconsistentes, desatualizados ou não monitorados.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, agentes podem auxiliar na identificação de lacunas documentais, ativos sem responsável, produtos obsoletos, regras ausentes e prioridades de correção.

5. Estratégias de mitigação

A mitigação da dívida de dados no UnespDataLens-RM pode envolver estratégias de documentação, governança, automação, revisão, padronização e melhoria contínua.

Estratégia	Descrição
Inventário de passivos de dados	Identificar fontes, datasets, regras, produtos e processos com lacunas ou fragilidades.
Classificação da dívida	Organizar problemas por tipo: documentação, qualidade, lineage, versão, governança, segurança, automação ou uso.
Priorização por impacto	Classificar dívidas conforme risco, criticidade, frequência de uso e impacto decisório.
Plano de redução de dívida	Definir ações para corrigir, documentar, versionar, reprocessar, consolidar ou descontinuar ativos.
Documentação mínima obrigatória	Exigir metadados essenciais para fontes, datasets, features, dashboards e produtos.
Automação de validações	Implantar testes de qualidade, checks de schema, logs e monitoramento recorrente.

Versionamento e changelog	Registrar mudanças em dados, regras, produtos e pipelines.
Curadoria periódica	Revisar ativos obsoletos, redundantes, pouco utilizados ou sem responsável.
Lineage obrigatório para ativos críticos	Garantir rastreabilidade mínima para datasets, indicadores e produtos relevantes.
Governança de publicação	Impedir disponibilização de produtos sem documentação, qualidade e responsável definidos.
Descontinuação controlada	Arquivar ou substituir ativos que aumentam complexidade sem gerar valor.
Revisão pós-incidente	Registrar causas e correções após falhas, inconsistências ou erros analíticos.

6. Métricas associadas

A dimensão Data Debt pode ser avaliada por métricas relacionadas à documentação, qualidade, rastreabilidade, versionamento, automação, obsolescência e governança dos ativos.

Métrica	Sigla	Descrição
Taxa de ativos com dívida identificada	TADI	Proporção de ativos com algum passivo registrado.
Taxa de ativos sem responsável	TASR	Proporção de ativos sem responsável técnico ou institucional.
Taxa de ativos sem documentação mínima	TADM	Proporção de ativos sem metadados essenciais.
Taxa de ativos sem lineage	TASL	Proporção de ativos sem rastreabilidade suficiente.
Taxa de ativos sem versão	TASV	Proporção de ativos críticos sem versionamento.
Taxa de produtos obsoletos	TPO	Proporção de produtos desatualizados ou sem manutenção.
Taxa de regras não documentadas	TRND	Proporção de regras de transformação, integração ou cálculo sem registro.
Taxa de pipelines manuais	TPM	Proporção de processos dependentes de execução manual.
Taxa de dívidas críticas	TDC	Proporção de dívidas classificadas como críticas.

Taxa de resolução da dívida	TRD	Proporção de dívidas tratadas em determinado período.
Tempo médio de resolução da dívida	TMRD	Tempo médio para corrigir ou mitigar uma dívida registrada.
Índice de dívida de dados	IDD	Indicador composto do nível de dívida acumulada no ecossistema.

Uma formulação possível para o **Índice de Dívida de Dados** é:

$$\mathbf{IDD} = \mathbf{w_1TADM} + \mathbf{w_2TASL} + \mathbf{w_3TASV} + \mathbf{w_4TRND} + \mathbf{w_5TPM} + \mathbf{w_6TPO} + \mathbf{w_7TDC} - \mathbf{w_8TRD}$$

em que:

- **TADM** representa ativos sem documentação mínima;
- **TASL** representa ativos sem lineage;
- **TASV** representa ativos sem versão;
- **TRND** representa regras não documentadas;
- **TPM** representa pipelines manuais;
- **TPO** representa produtos obsoletos;
- **TDC** representa dívidas críticas;
- **TRD** representa resolução da dívida;
- $\mathbf{w_i}$ representa o peso atribuído a cada dimensão.

Quanto maior o IDD, maior o nível de dívida de dados acumulada no ecossistema analítico.

7. Artefatos relacionados

A dimensão Data Debt pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de catálogo, qualidade, lineage, versionamento, governança e monitoramento.

Código	Artefato	Finalidade
DD1	Inventário de Dívida de Dados	Registrar passivos identificados em fontes, datasets, produtos e pipelines.
DD2	Matriz de Classificação da Dívida	Classificar dívidas por tipo, severidade, origem e impacto.
DD3	Registro de Regras Não Documentadas	Identificar regras de negócio, transformação ou cálculo sem documentação.
DD4	Mapa de Ativos sem Responsável	Identificar ativos sem ownership definido.
DD5	Relatório de Ativos sem Lineage	Listar datasets, dashboards, features e produtos sem rastreabilidade.

DD6	Relatório de Produtos Obsoletos	Identificar produtos desatualizados, depreciados ou sem uso.
DD7	Plano de Redução da Dívida de Dados	Definir ações, responsáveis, prazos e prioridades de correção.
DD8	Registro de Ações de Mitigação	Acompanhar correções, reprocessamentos, documentação e descontinuação.
DD9	Painel de Dívida de Dados	Monitorar evolução da dívida por tipo, módulo e criticidade.
DD10	Relatório de Maturidade da Dívida de Dados	Avaliar evolução da gestão da dívida no ecossistema.

8. Critérios de validação

A dimensão Data Debt pode ser considerada adequadamente tratada quando:

1. os principais tipos de dívida de dados foram identificados;
2. ativos críticos foram avaliados quanto à documentação, lineage, versão, responsável e qualidade;
3. dívidas foram classificadas por severidade, origem e impacto;
4. dívidas críticas possuem plano de mitigação;
5. regras de transformação, integração e cálculo foram documentadas;
6. ativos críticos possuem responsável definido;
7. datasets e produtos relevantes possuem versionamento e lineage mínimos;
8. produtos obsoletos ou sem uso foram revisados;
9. pipelines manuais críticos foram documentados ou automatizados;
10. ações de mitigação são registradas e acompanhadas;
11. há redução mensurável do índice de dívida de dados ao longo do tempo;
12. novas dívidas são prevenidas por critérios de publicação, documentação e governança.

9. Exemplo de aplicação no domínio educacional

Em um projeto educacional, a dívida de dados pode ocorrer quando diferentes bases sobre matrícula, frequência, notas, evasão, desempenho e interações em ambiente virtual são utilizadas sem documentação padronizada, sem versões preservadas e sem definição clara das regras de cálculo.

Por exemplo, uma taxa de evasão pode ter sido calculada em diferentes dashboards com regras distintas, sem registro da versão da base, sem definição formal de evasão e sem indicação do período de referência. Esse cenário cria dívida de dados porque dificulta comparar resultados, reproduzir indicadores e explicar divergências.

A dimensão Data Debt permite identificar esse passivo, registrar as diferentes regras existentes, definir uma regra oficial, versionar os datasets utilizados, documentar o indicador, descontinuar dashboards obsoletos e reduzir a dependência de conhecimento tácito.

10. Exemplo de aplicação em políticas públicas

Em políticas públicas, a dívida de dados pode ocorrer quando diferentes secretarias mantêm planilhas históricas, bases locais, relatórios manuais e dashboards sem documentação, sem governança e sem integração.

Por exemplo, um indicador de vulnerabilidade territorial pode ser produzido a partir de múltiplas planilhas, com regras de agregação não documentadas e sem registro das fontes originais. Se o responsável técnico deixar a equipe, a instituição pode perder a capacidade de atualizar ou reproduzir o indicador.

A dimensão Data Debt permite mapear esses passivos, documentar fontes e regras, criar lineage, definir responsáveis, migrar processos manuais para pipelines controlados e instituir critérios mínimos para novos produtos de dados.

11. Contribuição para o modelo de referência

A dimensão **Data Debt** contribui para o UnespDataLens-RM ao reconhecer que problemas acumulados nos dados, pipelines e produtos analíticos podem comprometer a sustentabilidade do ecossistema de dados.

Essa dimensão fortalece o modelo de referência ao:

- tornar visíveis passivos ocultos nos dados;
- reduzir dependência de conhecimento tácito;
- melhorar documentação e rastreabilidade;
- aumentar capacidade de manutenção dos pipelines;
- apoiar reprodutibilidade de resultados;
- reduzir risco de uso de produtos obsoletos;
- orientar priorização de correções;
- prevenir acúmulo de novas fragilidades;
- melhorar governança e responsabilidade sobre os dados;
- apoiar monitoramento da maturidade analítica;
- aumentar confiança nos ativos produzidos;
- sustentar evolução controlada do UnespDataLens-RM.

12. Síntese da dimensão

A dimensão **Data Debt** trata do acúmulo de passivos técnicos, metodológicos, semânticos e organizacionais nos dados, pipelines, regras, metadados, produtos e processos analíticos. No UnespDataLens-RM, essa dimensão atua como capacidade complementar para identificar, mensurar, priorizar e reduzir fragilidades que comprometem qualidade, governança, rastreabilidade, versionamento, reprodutibilidade e manutenção dos ativos analíticos.

Ao incorporar estratégias de inventário da dívida, classificação de severidade, documentação mínima, lineage, versionamento, curadoria, automação, monitoramento e planos de mitigação, o modelo de referência reduz riscos acumulados e fortalece a sustentabilidade do ecossistema de Engenharia Analítica de Dados.

Perfeito. O problema estava na tabela da seção **7. Artefatos relacionados**, em que a estrutura da tabela ficou inconsistente. Abaixo está o módulo completo refeito, com a tabela corrigida e a redação revisada.

Dimensão Complementar — Data Observability

1. Definição da dimensão

A dimensão **Data Observability**, ou **observabilidade de dados**, refere-se à capacidade de acompanhar, diagnosticar, explicar e compreender o estado, o comportamento, a qualidade, a disponibilidade, a atualização, a integridade e a evolução dos dados, pipelines e produtos analíticos ao longo do tempo.

No contexto do **UnespDataLens-RM**, a observabilidade de dados representa a capacidade de saber, de forma sistemática, rastreável e acionável, se os dados estão sendo extraídos corretamente, se os pipelines foram executados como esperado, se os datasets estão completos, se houve mudança de schema, se a qualidade se degradou, se os produtos analíticos estão atualizados, se há falhas recorrentes, se ocorreu drift, se determinados ativos deixaram de ser consumidos ou se algum problema compromete a confiança nos resultados disponibilizados.

A observabilidade de dados vai além do monitoramento tradicional. Enquanto o monitoramento identifica estados, métricas e alertas, a observabilidade busca oferecer condições para **compreender as causas, os impactos e o contexto dos problemas observados**. Assim, não basta saber que uma base apresentou queda de registros; é necessário compreender qual fonte falhou, qual pipeline foi afetado, quais produtos dependem dessa base, quais usuários podem ser impactados, qual versão está comprometida e qual ação corretiva deve ser tomada.

No UnespDataLens-RM, essa dimensão é complementar e transversal porque depende da integração entre logs, métricas, metadados, qualidade, linhagem, versionamento, governança, armazenamento, disponibilização, consumo e monitoramento. A observabilidade transforma eventos técnicos e operacionais em conhecimento acionável sobre a saúde do ecossistema analítico.

2. Problema tratado

Em ecossistemas analíticos complexos, falhas nos dados podem ocorrer de forma silenciosa. Uma fonte pode deixar de ser atualizada, um campo pode mudar de tipo, uma categoria pode desaparecer, uma regra de transformação pode falhar, uma API pode retornar dados incompletos, uma tabela pode ser carregada parcialmente, uma feature pode sofrer drift, um dashboard pode continuar publicado com dados obsoletos ou um produto de dados pode apresentar resultados inconsistentes sem que os usuários percebam imediatamente.

Esses problemas são especialmente críticos porque decisões institucionais, científicas ou de gestão podem ser tomadas com base em dados aparentemente disponíveis, mas tecnicamente comprometidos.

A ausência de observabilidade pode gerar problemas como:

- falhas de ingestão não detectadas;
- atrasos de atualização sem alerta;
- degradação gradual da qualidade dos dados;
- mudanças de schema que quebram pipelines;
- dashboards publicados com dados desatualizados;
- indicadores calculados com dados incompletos;
- dificuldade de identificar a causa de inconsistências;
- ausência de visibilidade sobre dependências entre ativos;
- aumento do tempo para diagnosticar problemas;
- falta de evidências para auditoria;
- baixa confiança dos usuários nos dados;
- dificuldade de priorizar ações corretivas;
- problemas recorrentes não registrados;
- ausência de histórico operacional do pipeline;
- baixa capacidade de explicar incidentes a usuários e gestores.

No UnespDataLens-RM, a dimensão Data Observability busca tornar esses problemas visíveis, mensuráveis, explicáveis e acionáveis, reduzindo o risco de que falhas nos dados permaneçam ocultas ou sejam percebidas apenas após impactarem produtos analíticos e decisões.

3. Objetivo da dimensão

O objetivo da dimensão **Data Observability** é **fornecer visibilidade contínua, explicável e acionável sobre o estado dos dados, pipelines e produtos analíticos**, permitindo detectar, diagnosticar, priorizar e corrigir problemas de qualidade, atualização, disponibilidade, schema, volume, drift, desempenho, linhagem e consumo.

Essa dimensão busca garantir que o UnespDataLens-RM não apenas produza dados e produtos analíticos, mas também acompanhe sua saúde operacional, técnica e analítica ao longo do tempo, permitindo atuação preventiva e corretiva sobre falhas, degradações e mudanças relevantes.

4. Relação com o UnespDataLens-RM

A dimensão Data Observability se articula com diversos módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, apoia a observação da disponibilidade das fontes, periodicidade de atualização, responsáveis, acessibilidade, cobertura temporal e mudanças na estrutura ou no escopo dos dados fornecidos.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, acompanha logs de ingestão, falhas, volumes extraídos, checksums, atrasos, dados ausentes, integridade dos snapshots e sucesso ou falha das execuções.

No **Módulo 3 — Integração e Harmonização**, permite observar taxas de pareamento, registros não integrados, conflitos entre fontes, falhas de chaves, divergências de granularidade e queda na capacidade de relacionamento entre bases.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, acompanha transformações executadas, registros afetados, atributos criados, perdas, erros, mudanças de distribuição e impactos das regras aplicadas.

No **Módulo 5 — Qualidade e Validação de Dados**, fornece base para acompanhar completude, consistência, validade, unicidade, atualidade, conformidade e evolução da qualidade dos dados.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, observa mudanças na distribuição de grupos, perdas por etapa, drift por grupo e alterações em métricas de equidade.

No **Módulo 7 — Armazenamento Analítico**, monitora disponibilidade, desempenho de consultas, atualização de data marts, integridade de tabelas, crescimento de volume, organização das zonas de dados e acessibilidade dos ativos persistidos.

No **Módulo 8 — Disponibilização e Consumo Analítico**, acompanha atualização, uso, disponibilidade, tempo de resposta, falhas e obsolescência de dashboards, APIs, relatórios e produtos de dados.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, utiliza metadados para contextualizar alertas, identificar responsáveis, localizar ativos afetados e compreender o significado dos dados monitorados.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, apoia a identificação de acessos indevidos, permissões vencidas, ativos sem responsável, políticas não cumpridas e riscos operacionais associados ao uso de dados sensíveis.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, utiliza lineage para identificar causas e impactos de falhas, respondendo quais produtos são afetados por determinado problema.

No **Módulo 12 — Versionamento e Reprodutibilidade**, compara versões e execuções para verificar quando um problema foi introduzido e se uma execução pode ser reproduzida.

No **Módulo 13 — Representações Analíticas**, observa a estabilidade de representações tabulares, temporais, espaciais, textuais, semânticas, relacionais, grafos e grafos temporais.

No **Módulo 14 — Feature Engineering e Feature Store**, acompanha estabilidade, completude, drift, uso e qualidade das features.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, atua como dimensão complementar que amplia o monitoramento, incorporando diagnóstico, explicação, contexto e análise de causa.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, fornece insumos para que agentes expliquem alertas, recomendem ações corretivas, indiquem produtos impactados e apoiem usuários na interpretação da saúde dos dados.

5. Estratégias de implementação e mitigação

A implementação da observabilidade de dados no UnespDataLens-RM pode envolver estratégias técnicas, metodológicas e organizacionais.

Estratégia	Descrição
Logs estruturados	Registrar eventos de execução, erros, entradas, saídas, parâmetros, status, duração e evidências dos pipelines.
Métricas operacionais	Acompanhar tempo de execução, volume processado, falhas, atrasos, disponibilidade e desempenho.
Métricas de qualidade	Monitorar completude, validade, consistência, duplicidade, atualidade, conformidade e integridade.
Monitoramento de schema	Detectar inclusão, remoção, renomeação ou alteração de tipos de atributos.
Monitoramento de volume	Identificar variações inesperadas no número de registros, arquivos, eventos ou entidades.
Monitoramento de distribuição	Detectar mudanças estatísticas em atributos, features, variáveis-alvo ou indicadores.
Alertas configuráveis	Gerar alertas com severidade, responsável, evidência, impacto e ação recomendada.
Análise de causa raiz	Relacionar falhas a fontes, processos, versões, regras, dependências e produtos impactados.
Lineage operacional	Identificar dependências entre fontes, pipelines, datasets, features, dashboards, APIs e produtos.
Painéis de saúde dos dados	Disponibilizar visão consolidada sobre o estado dos ativos analíticos.
Registro de incidentes	Documentar problemas, causas, impactos, responsáveis, ações corretivas e recorrência.
Acompanhamento de SLAs de dados	Definir expectativas de atualização, disponibilidade, qualidade e tempo de resposta para ativos críticos.
Integração com agentes inteligentes	Permitir explicação automática de alertas, impactos, causas prováveis e recomendações de ação.
Revisão pós-incidente	Registrar lições aprendidas, causas recorrentes e ajustes necessários nos pipelines.

6. Métricas associadas

A dimensão Data Observability pode ser avaliada por métricas relacionadas à visibilidade, diagnóstico, alertas, qualidade operacional, resolução de incidentes e cobertura de observabilidade.

Métrica	Sigla	Descrição

Taxa de ativos observáveis	TAO	Proporção de ativos críticos com métricas e logs disponíveis.
Taxa de pipelines observáveis	TPO	Proporção de pipelines com registros de execução, entradas, saídas e status.
Taxa de logs estruturados	TLE	Proporção de execuções com logs padronizados e consultáveis.
Taxa de métricas coletadas	TMC	Proporção de métricas esperadas efetivamente coletadas.
Taxa de alertas com evidência	TAE	Proporção de alertas acompanhados de métricas, logs ou registros que justificam sua geração.
Taxa de incidentes diagnosticados	TID	Proporção de incidentes com causa provável identificada.
Tempo médio de detecção	TMD	Tempo entre a ocorrência do problema e sua detecção.
Tempo médio de diagnóstico	TMDi	Tempo entre a detecção do problema e a identificação da causa provável.
Tempo médio de resolução	TMR	Tempo entre a abertura e a resolução do incidente.
Taxa de recorrência de incidentes	TRI	Proporção de incidentes que se repetem em determinado período.
Taxa de ativos com SLA definido	TASLA	Proporção de ativos críticos com expectativas formais de atualização, qualidade ou disponibilidade.
Taxa de falhas silenciosas	TFS	Proporção de problemas identificados tardiamente ou sem alerta automático.
Taxa de ativos com impacto rastreável	TAIR	Proporção de ativos cujos impactos podem ser identificados por lineage.
Índice de observabilidade de dados	IOD	Indicador composto de logs, métricas, alertas, diagnóstico e rastreabilidade.

Uma formulação possível para o **Índice de Observabilidade de Dados** é:

$$\mathbf{IOD} = w_1\mathbf{TAO} + w_2\mathbf{TPO} + w_3\mathbf{TLE} + w_4\mathbf{TMC} + w_5\mathbf{TAE} + w_6\mathbf{TID} + w_7\mathbf{TAIR} - w_8\mathbf{TRI} - w_9\mathbf{TFS}$$

em que:

- **TAO** representa a taxa de ativos observáveis;
- **TPO** representa a taxa de pipelines observáveis;
- **TLE** representa a taxa de logs estruturados;
- **TMC** representa a taxa de métricas coletadas;
- **TAE** representa a taxa de alertas com evidência;
- **TID** representa a taxa de incidentes diagnosticados;
- **TAIR** representa a taxa de ativos com impacto rastreável;
- **TRI** representa a taxa de recorrência de incidentes;
- **TFS** representa a taxa de falhas silenciosas;
- w_i representa o peso atribuído a cada dimensão.

Quanto maior o IOD, maior a capacidade do ecossistema de observar, diagnosticar e explicar o comportamento dos dados, pipelines e produtos analíticos.

7. Artefatos relacionados

A dimensão Data Observability pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de qualidade, monitoramento, lineage, versionamento, governança, armazenamento, consumo e agentes inteligentes.

Código	Artefato	Representação	Finalidade
DOB1	Plano de Observabilidade de Dados	Documento técnico	Definir ativos observáveis, métricas, logs, alertas, responsáveis, periodicidade e critérios de severidade.
DOB2	Registro de Métricas Operacionais	Tabela ou base de métricas	Acompanhar execução, volume, tempo, erros, disponibilidade, atualização e desempenho dos pipelines.
DOB3	Registro de Logs Estruturados	Log padronizado	Preservar eventos de pipeline, entradas, saídas, parâmetros, erros, status e evidências de execução.
DOB4	Painel de Saúde dos Dados	Dashboard operacional	Consolidar visão sobre qualidade, atualização, alertas, falhas, disponibilidade, drift e consumo dos ativos.
DOB5	Registro de Alertas com Evidência	Tabela de alertas	Documentar alertas, métricas associadas, severidade, impacto, responsável e ação sugerida.
DOB6	Registro de Incidentes de Dados	Tabela ou relatório de incidentes	Documentar problemas, causas prováveis, impactos, responsáveis, ações corretivas e recorrência.
DOB7	Relatório de	Relatório técnico	Explicar a origem provável de

	Análise de Causa Raiz		falhas, degradações, atrasos, alterações de schema ou inconsistências.
DOB8	Mapa de Impacto Operacional	Grafo, matriz ou tabela de dependências	Relacionar falhas a datasets, features, dashboards, APIs, relatórios e produtos afetados.
DOB9	Registro de SLAs de Dados	Documento ou tabela de compromissos	Definir expectativas de atualização, disponibilidade, qualidade, tempo de resposta e critérios de violação.
DOB10	Relatório de Observabilidade dos Ativos	Relatório gerencial e técnico	Avaliar cobertura, lacunas, recorrências, evolução e maturidade da observabilidade no ecossistema.
DOB11	Registro de Falhas Silenciosas	Tabela de ocorrências	Identificar problemas detectados tardiamente ou sem alerta automático, indicando lacunas de observabilidade.
DOB12	Registro de Ações Pós-Incidente	Tabela de acompanhamento	Documentar correções, reprocessamentos, ajustes de regra, atualização de baseline e prevenção de recorrência.

8. Critérios de validação

A dimensão Data Observability pode ser considerada adequadamente tratada quando:

1. os ativos críticos possuem métricas observáveis;
2. pipelines críticos possuem logs estruturados;
3. execuções registram entradas, saídas, parâmetros, tempo, status e erros;
4. métricas de qualidade são coletadas de forma recorrente;
5. mudanças de schema são detectadas e registradas;
6. variações anômalas de volume são identificadas;
7. alertas possuem evidência, severidade, responsável e ação recomendada;
8. incidentes são registrados com causa provável, impacto e ação corretiva;
9. lineage permite identificar produtos afetados por falhas;
10. responsáveis por ativos e incidentes estão definidos;
11. painéis de saúde dos dados estão disponíveis para acompanhamento;
12. tempos de detecção, diagnóstico e resolução são monitorados;
13. problemas recorrentes são identificados e tratados;
14. falhas silenciosas são reduzidas ao longo do tempo;
15. agentes inteligentes podem consultar registros de observabilidade para explicar alertas e recomendar ações.

9. Exemplo de aplicação no domínio educacional

Em um projeto educacional, a observabilidade de dados pode acompanhar a saúde de bases de matrícula, histórico acadêmico, notas, frequência, interações em ambiente virtual, evasão, desempenho e dashboards institucionais.

Por exemplo, se a base de interações do ambiente virtual apresentar queda abrupta de registros em determinado período, o sistema de observabilidade pode gerar um alerta, comparar o volume atual com o histórico, verificar logs de ingestão, identificar se houve falha de extração, consultar a linhagem para descobrir quais features e dashboards dependem dessa fonte e recomendar reprocessamento ou revisão da conexão.

Da mesma forma, se um dashboard de evasão estiver disponível, mas a base de matrícula utilizada possuir atraso de carga, o painel de saúde dos dados pode sinalizar que o produto está operacionalmente ativo, porém analiticamente desatualizado.

Outro exemplo ocorre quando a coluna que representa situação acadêmica muda de domínio, passando a incluir uma nova categoria. A observabilidade pode detectar a alteração, registrar o drift de schema ou domínio, indicar quais regras de cálculo de evasão podem ser impactadas e sugerir revisão da harmonização semântica.

10. Exemplo de aplicação em políticas públicas

Em políticas públicas, a observabilidade pode acompanhar bases de saúde, educação, assistência social, território, atendimento, benefícios, equipamentos públicos e indicadores de vulnerabilidade.

Por exemplo, se uma secretaria deixar de enviar dados mensais de atendimento, a observabilidade pode detectar ausência de atualização, gerar alerta, identificar os indicadores afetados, apontar dashboards dependentes e registrar o incidente. Se a falha for recorrente, o relatório de observabilidade pode indicar necessidade de revisão do processo de integração ou de definição de SLA com a fonte fornecedora.

Outro exemplo ocorre quando uma categoria de atendimento muda de nome em uma fonte. A observabilidade de schema e domínio pode detectar a alteração, alertar sobre risco de inconsistência histórica e recomendar atualização das regras de harmonização.

Também pode ocorrer de um painel territorial continuar sendo acessado, mesmo com dados desatualizados. Nesse caso, a observabilidade permite sinalizar ao usuário que o produto está disponível, mas possui restrição de atualidade, evitando interpretações equivocadas.

11. Contribuição para o modelo de referência

A dimensão **Data Observability** contribui para o UnespDataLens-RM ao transformar a saúde dos dados e pipelines em uma capacidade explícita, mensurável e acionável.

Essa dimensão fortalece o modelo de referência ao:

- aumentar a visibilidade sobre o estado dos dados;
- reduzir falhas silenciosas;
- antecipar problemas antes que afetem decisões;
- acelerar o diagnóstico de incidentes;
- relacionar falhas a fontes, processos e produtos impactados;
- apoiar melhoria contínua dos pipelines;
- fortalecer a confiança nos produtos analíticos;

- reduzir dívida de dados;
- complementar monitoramento, lineage e versionamento;
- apoiar governança operacional;
- fornecer evidências para auditoria;
- melhorar comunicação entre equipes técnicas, gestores e usuários;
- permitir que agentes inteligentes expliquem alertas e recomendem ações.

12. Síntese da dimensão

A dimensão **Data Observability** trata da capacidade de observar, diagnosticar e explicar o comportamento dos dados, pipelines e produtos analíticos ao longo do tempo. No UnespDataLens-RM, essa dimensão atua como capacidade complementar para garantir que problemas de qualidade, atualização, schema, volume, distribuição, disponibilidade, desempenho e consumo sejam detectados, contextualizados e tratados de forma sistemática.

Ao incorporar logs estruturados, métricas, alertas, painéis de saúde, análise de causa raiz, SLAs, mapas de impacto, registros de incidentes e ações pós-incidente, o modelo de referência fortalece sua capacidade de manter ativos analíticos confiáveis, auditáveis, monitoráveis e adequados à tomada de decisão baseada em evidências.

Dimensão Complementar — Schema Drift

1. Definição da dimensão

A dimensão **Schema Drift**, ou **deriva de esquema**, refere-se às mudanças estruturais que ocorrem nos dados ao longo do tempo, afetando atributos, tipos de dados, nomes de campos, formatos, domínios de valores, chaves, cardinalidades, granularidades ou estruturas de tabelas, arquivos, APIs, eventos, documentos ou produtos analíticos.

No contexto do **UnespDataLens-RM**, o schema drift ocorre quando a estrutura esperada de uma fonte, dataset, tabela, arquivo, API, data mart, feature set, grafo ou produto de dados se altera em relação a uma versão anterior ou a um contrato esperado. Essas mudanças podem ser planejadas, como a inclusão de novos campos em uma base institucional, ou não planejadas, como a alteração silenciosa do tipo de um atributo, a remoção de uma coluna, a mudança de nome de uma categoria, a substituição de códigos ou a modificação de uma chave de integração.

A deriva de esquema é uma dimensão crítica da Engenharia Analítica de Dados porque os pipelines dependem de estruturas previamente conhecidas. Quando um schema muda sem registro, validação ou adaptação, processos de ingestão, integração, transformação, validação, armazenamento, dashboards, APIs, modelos, features, grafos e indicadores podem falhar ou, ainda pior, continuar funcionando com resultados incorretos.

No UnespDataLens-RM, a dimensão Schema Drift é tratada como uma capacidade complementar e transversal, pois mudanças estruturais podem ocorrer em qualquer etapa do ciclo de dados e podem impactar múltiplos módulos, produtos e consumidores. Seu tratamento exige monitoramento, versionamento, validação de schema, lineage, metadados, governança, alertas, contratos de dados e análise de impacto.

2. Problema tratado

Em ecossistemas analíticos reais, as fontes de dados evoluem continuamente. Sistemas institucionais são atualizados, APIs mudam suas respostas, planilhas recebem novas colunas, categorias são renomeadas, campos deixam de ser preenchidos, identificadores são substituídos, tipos de dados são alterados e regras de negócio passam por revisões.

Essas mudanças nem sempre são comunicadas às equipes responsáveis pelos pipelines analíticos. Como resultado, processos que dependem de uma estrutura esperada podem falhar ou produzir resultados inconsistentes.

O schema drift pode gerar problemas como:

- falhas de ingestão devido à remoção ou renomeação de campos;
- erros de transformação causados por mudança de tipo de dado;
- quebra de integrações por alteração de chaves;
- perda de compatibilidade entre versões de datasets;
- dashboards com indicadores incorretos devido à mudança de significado de atributos;
- features calculadas com base em campos alterados;

- modelos preditivos degradados por mudança de estrutura ou distribuição dos atributos;
- APIs incompatíveis com consumidores existentes;
- dificuldade de comparar séries históricas;
- perda de rastreabilidade entre versões;
- aumento de retrabalho técnico;
- inconsistência entre produtos analíticos;
- risco de decisões baseadas em dados estruturalmente alterados sem validação.

No UnespDataLens-RM, o tratamento do schema drift é essencial para garantir que mudanças estruturais sejam detectadas, registradas, avaliadas e tratadas antes que comprometam a confiabilidade dos ativos analíticos.

3. Objetivo da dimensão

O objetivo da dimensão **Schema Drift** é **detectar, registrar, classificar, avaliar e tratar mudanças estruturais nos dados e ativos analíticos**, garantindo que alterações em schemas, atributos, tipos, nomes, domínios, chaves, categorias e granularidades sejam controladas, documentadas, versionadas e avaliadas quanto ao seu impacto nos pipelines, produtos, features, modelos, dashboards, APIs e decisões.

Essa dimensão busca assegurar que a evolução estrutural dos dados ocorra de forma controlada, transparente, rastreável e compatível com a manutenção da qualidade analítica.

4. Relação com o UnespDataLens-RM

A dimensão Schema Drift se articula com diversos módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, contribui para registrar a estrutura inicial das fontes, seus atributos, formatos, tipos, periodicidade, granularidade e limitações conhecidas, criando uma referência para comparação futura.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, atua na detecção de alterações estruturais no momento da ingestão, verificando se os campos esperados estão presentes, se os tipos permanecem compatíveis e se o formato dos dados foi alterado.

No **Módulo 3 — Integração e Harmonização**, é fundamental para identificar mudanças em chaves, identificadores, atributos equivalentes, granularidades e entidades compartilhadas entre fontes.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, permite verificar se regras de transformação continuam válidas após alterações no schema de entrada.

No **Módulo 5 — Qualidade e Validação de Dados**, relaciona-se aos testes estruturais, validação de domínio, conformidade de tipos, presença de atributos obrigatórios e detecção de mudanças em categorias.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, pode indicar mudanças estruturais que afetam atributos sensíveis, grupos de análise, variáveis-alvo ou dimensões de equidade.

No **Módulo 7 — Armazenamento Analítico**, apoia a gestão de schemas analíticos, data marts, views, tabelas, partições e estruturas persistidas.

No **Módulo 8 — Disponibilização e Consumo Analítico**, ajuda a evitar que dashboards, APIs, relatórios e produtos de dados sejam impactados por mudanças estruturais não comunicadas.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, registra versões de schemas, descrições de atributos, domínios, categorias, conceitos e equivalências semânticas.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, permite identificar quais produtos, features, dashboards e modelos dependem de um campo alterado.

No **Módulo 12 — Versionamento e Reprodutibilidade**, registra versões de schemas e possibilita comparar alterações entre versões.

No **Módulo 14 — Feature Engineering e Feature Store**, verifica se features dependem de atributos removidos, renomeados ou semanticamente alterados.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, atua como uma das formas específicas de drift a serem monitoradas.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, agentes podem explicar mudanças de schema, identificar impactos e recomendar ações corretivas.

5. Tipos de Schema Drift

O schema drift pode assumir diferentes formas, cada uma com impacto específico no pipeline analítico.

Tipo de drift	Descrição	Exemplo
Adição de atributo	Novo campo passa a existir na fonte ou dataset.	Inclusão do campo `modalidade_ingresso`.
Remoção de atributo	Campo esperado deixa de existir.	Remoção da coluna `frequencia`.
Renomeação de atributo	Campo mantém significado semelhante, mas muda de nome.	`cod_curso` passa a ser `codigo_curso`.
Mudança de tipo	Tipo de dado é alterado.	Campo de data passa de `date`

		para `string`.
Mudança de domínio	Conjunto de valores permitidos muda.	Nova categoria em situação acadêmica.
Mudança de codificação	Códigos ou padrões de representação são alterados.	Sexo: `M/F` passa para `1/2`.
Mudança de granularidade	Nível de detalhe dos dados muda.	Frequência diária passa a ser mensal.
Mudança de chave	Identificador usado para integração muda.	RA substituído por identificador institucional.
Mudança de cardinalidade	Relação entre entidades muda.	Um estudante passa a ter múltiplos vínculos ativos.
Mudança semântica	O nome permanece, mas o significado muda.	“Evasão” passa a incluir cancelamento e abandono.
Mudança de formato	Estrutura do arquivo ou resposta da API muda.	CSV passa a JSON; API altera estrutura aninhada.
Mudança de obrigatoriedade	Campo antes obrigatório passa a ser opcional, ou o inverso.	Campo de renda deixa de ser preenchido obrigatoriamente.

Nem todo schema drift é necessariamente negativo. Algumas mudanças representam evolução legítima das fontes. O problema ocorre quando a mudança não é detectada, documentada, avaliada ou incorporada corretamente ao pipeline.

6. Estratégias de mitigação

A mitigação de schema drift no UnespDataLens-RM pode envolver mecanismos técnicos, metodológicos e de governança.

Estratégia	Descrição
Registro de schema de referência	Documentar a estrutura esperada dos dados, incluindo atributos, tipos, domínios e chaves.
Validação automática de schema	Comparar dados ingeridos com schema esperado a cada execução.
Versionamento de schemas	Registrar alterações estruturais ao longo do tempo.
Data contracts	Definir acordos formais sobre estrutura, periodicidade, qualidade e mudanças esperadas.
Alertas de mudança estrutural	Gerar alertas quando atributos forem adicionados, removidos, renomeados ou alterados.
Classificação de severidade	Avaliar se a mudança é baixa, média, alta ou crítica.
Análise de impacto via lineage	Identificar quais datasets, features, dashboards, APIs e modelos são afetados.
Compatibilidade retroativa	Manter suporte a versões anteriores quando necessário.
Mapeamento de renomeações	Registrar equivalências entre nomes antigos e novos de atributos.
Validação semântica	Verificar se mudanças estruturais também implicam mudança de significado.

Reprocessamento controlado	Reexecutar pipelines afetados após ajuste de schema.
Governança de mudanças	Exigir comunicação, aprovação ou registro para alterações estruturais em fontes críticas.

7. Métricas associadas

A dimensão Schema Drift pode ser avaliada por métricas relacionadas à frequência, severidade, detecção, impacto e resolução de mudanças estruturais.

Métrica	Sigla	Descrição
Taxa de ativos com schema documentado	TASD	Proporção de ativos que possuem schema registrado.
Taxa de ativos com schema versionado	TASV	Proporção de ativos que possuem histórico de versões de schema.
Taxa de schema drift detectado	TSDD	Proporção de ativos com mudanças estruturais identificadas.
Taxa de mudanças críticas de schema	TMCS	Proporção de mudanças classificadas como críticas.
Taxa de mudanças não comunicadas	TMNC	Proporção de mudanças detectadas sem aviso ou registro prévio.
Tempo médio de detecção de schema drift	TMDS	Tempo entre a mudança estrutural e sua detecção.
Tempo médio de resolução de schema drift	TMRS	Tempo entre a detecção e a correção ou adaptação.

Taxa de pipelines impactados	TPI	Proporção de pipelines afetados por mudanças de schema.
Taxa de produtos impactados	TPrI	Proporção de dashboards, APIs, relatórios ou produtos afetados.
Taxa de atributos renomeados mapeados	TARM	Proporção de renomeações com equivalência registrada.
Taxa de compatibilidade preservada	TCP	Proporção de mudanças que preservam compatibilidade com versões anteriores.
Índice de estabilidade de schema	IES	Indicador composto da estabilidade estrutural dos ativos.

Uma formulação possível para o **Índice de Estabilidade de Schema** é:

$$\mathbf{IES} = \mathbf{w_1 TASD} + \mathbf{w_2 TASV} + \mathbf{w_3 TCP} + \mathbf{w_4 TARM} - \mathbf{w_5 TSDD} - \mathbf{w_6 TMCS} - \mathbf{w_7 TMNC}$$

em que:

- **TASD** representa a taxa de ativos com schema documentado;
- **TASV** representa a taxa de ativos com schema versionado;
- **TCP** representa a taxa de compatibilidade preservada;
- **TARM** representa a taxa de atributos renomeados mapeados;
- **TSDD** representa a taxa de schema drift detectado;
- **TMCS** representa a taxa de mudanças críticas de schema;
- **TMNC** representa a taxa de mudanças não comunicadas;
- $\mathbf{w_i}$ representa o peso atribuído a cada dimensão.

Quanto maior o IES, maior a estabilidade estrutural e a capacidade de controle sobre mudanças de schema.

8. Artefatos relacionados

A dimensão Schema Drift pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de inventário, ingestão, qualidade, metadados, lineage, versionamento, monitoramento e governança.

Código	Artefato	Representação	Finalidade

SD1	Registro de Schema de Referência	Tabela técnica ou contrato de dados	Documentar atributos esperados, tipos, domínios, chaves, obrigatoriedade e granularidade.
SD2	Histórico de Versões de Schema	Changelog ou tabela versionada	Registrar alterações estruturais ao longo do tempo.
SD3	Relatório de Detecção de Schema Drift	Relatório técnico	Comparar schema observado com schema esperado e registrar divergências.
SD4	Registro de Mudanças Estruturais	Tabela de mudanças	Documentar adições, remoções, renomeações, mudanças de tipo, domínio ou chave.
SD5	Matriz de Impacto de Schema	Matriz ou grafo de dependências	Identificar pipelines, features, dashboards, APIs e produtos afetados por mudanças.
SD6	Mapa de Equivalência de Atributos	Tabela de correspondência	Registrar equivalências entre atributos antigos, novos, renomeados ou substituídos.
SD7	Plano de Adaptação de Schema	Documento técnico	Definir ações para ajustar pipelines, validações, produtos e consumidores afetados.
SD8	Registro de Compatibilidade Retroativa	Documento ou tabela técnica	Indicar se mudanças preservam compatibilidade com versões anteriores.
SD9	Alerta de Schema Drift	Registro de alerta	Notificar responsáveis sobre mudanças estruturais detectadas.
SD10	Relatório de Estabilidade de Schema	Relatório técnico-gerencial	Avaliar frequência, severidade, impacto e resolução das mudanças de schema.

9. Critérios de validação

A dimensão Schema Drift pode ser considerada adequadamente tratada quando:

1. os ativos críticos possuem schema de referência documentado;
2. schemas de fontes, datasets, data marts, APIs, features e produtos são versionados;
3. mudanças estruturais são detectadas automaticamente ou registradas formalmente;
4. alterações de atributos, tipos, domínios, chaves e granularidades são documentadas;
5. mudanças críticas geram alerta para responsáveis;
6. impactos sobre pipelines, dashboards, APIs, features e modelos são identificados por lineage;
7. renomeações e substituições de atributos possuem mapeamento de equivalência;
8. mudanças planejadas são comunicadas antes de afetar consumidores;
9. mudanças não planejadas são registradas como incidentes ou alertas;
10. pipelines afetados são revisados, adaptados ou reprocessados;
11. produtos publicados indicam quando houve alteração estrutural relevante;
12. usuários são informados quando mudanças de schema afetam interpretação de indicadores ou resultados.

10. Exemplo de aplicação no domínio educacional

Em um projeto educacional, o schema drift pode ocorrer quando a base de matrícula passa a incluir uma nova categoria de situação acadêmica, quando o campo de frequência muda de percentual para quantidade de presenças, quando o identificador do estudante é alterado, ou quando a base de interações do ambiente virtual passa a registrar eventos em novo formato.

Por exemplo, se a variável `situacao_academica` originalmente continha as categorias “ativo”, “trancado”, “cancelado” e “concluído”, mas passa a incluir “abandono”, “transferido” ou “mobilidade”, as regras de cálculo de evasão podem ser afetadas. Sem detecção e documentação, dashboards e modelos podem interpretar incorretamente os novos valores.

A dimensão Schema Drift permite detectar a nova categoria, registrar a mudança, avaliar impacto nos indicadores de evasão, atualizar regras de harmonização, versionar o schema e informar os consumidores dos produtos educacionais.

11. Exemplo de aplicação em políticas públicas

Em políticas públicas, o schema drift pode ocorrer quando uma secretaria altera o formato de envio de dados, muda códigos de atendimento, substitui identificadores territoriais, altera categorias de serviços ou modifica a periodicidade de registros.

Por exemplo, uma base de atendimentos sociais pode mudar o campo `tipo_atendimento` para uma nova codificação, agrupando categorias antes separadas. Essa alteração pode impactar indicadores de demanda, dashboards territoriais e análises de vulnerabilidade.

A dimensão Schema Drift permite identificar a mudança, comparar o schema atual com o anterior, mapear categorias antigas e novas, avaliar quais produtos foram impactados e orientar o reprocessamento dos indicadores.

12. Contribuição para o modelo de referência

A dimensão **Schema Drift** contribui para o UnespDataLens-RM ao reconhecer que a estrutura dos dados é dinâmica e que mudanças estruturais precisam ser controladas para preservar qualidade, integração, rastreabilidade, reprodutibilidade e confiança.

Essa dimensão fortalece o modelo de referência ao:

- detectar alterações estruturais em fontes e datasets;
- reduzir falhas inesperadas em pipelines;
- preservar compatibilidade entre versões;
- proteger dashboards, APIs, features e modelos contra mudanças silenciosas;
- apoiar análise de impacto por meio de lineage;
- fortalecer versionamento de schemas;
- melhorar comunicação entre produtores e consumidores de dados;
- reduzir dívida de dados;
- apoiar governança de mudanças;
- garantir que a evolução estrutural dos dados ocorra de forma controlada.

13. Síntese da dimensão

A dimensão **Schema Drift** trata das mudanças estruturais que afetam schemas, atributos, tipos, domínios, chaves, categorias, formatos e granularidades dos dados ao longo do tempo. No UnespDataLens-RM, essa dimensão atua como capacidade complementar para detectar, registrar, avaliar e tratar alterações estruturais que possam impactar pipelines, produtos analíticos, dashboards, APIs, features, modelos e decisões.

Ao incorporar validação de schema, versionamento, contratos de dados, alertas, análise de impacto, lineage, governança de mudanças e registros de compatibilidade, o modelo de referência fortalece sua capacidade de lidar com a evolução dos dados sem comprometer a confiabilidade e a continuidade dos processos analíticos.

Dimensão Complementar — Data Contracts

1. Definição da dimensão

A dimensão **Data Contracts**, ou **contratos de dados**, refere-se ao conjunto de acordos formais, explícitos e verificáveis que definem as responsabilidades, expectativas, regras, estruturas, padrões de qualidade, periodicidade, segurança, governança, versionamento e condições de uso associadas ao fornecimento, processamento, compartilhamento e consumo de dados entre produtores e consumidores.

No contexto do **UnespDataLens-RM**, os contratos de dados funcionam como instrumentos técnico-metodológicos para reduzir ambiguidades, prevenir falhas, controlar mudanças, melhorar a comunicação entre áreas, proteger pipelines contra alterações inesperadas e assegurar que os dados utilizados em análises, dashboards, APIs, features, modelos e produtos analíticos atendam a requisitos mínimos previamente definidos.

Um contrato de dados pode estabelecer, por exemplo, quais campos uma fonte deve fornecer, quais tipos de dados são esperados, quais atributos são obrigatórios, qual a periodicidade de atualização, quais domínios de valores são aceitos, quais regras de qualidade devem ser respeitadas, quais mudanças precisam ser comunicadas previamente, quais restrições de acesso se aplicam e quais produtos dependem daquele ativo.

Assim, contratos de dados não são apenas documentos administrativos. Eles representam uma forma de governança técnica dos dados, permitindo que fontes, pipelines e produtos analíticos sejam tratados com maior previsibilidade, confiabilidade e responsabilidade.

No UnespDataLens-RM, essa dimensão é complementar e transversal porque contratos de dados podem ser aplicados a fontes, datasets, APIs, tabelas, data marts, features, produtos de dados, dashboards, modelos e processos de integração. Eles se relacionam diretamente com inventário, ingestão, integração, qualidade, governança, lineage, versionamento, schema drift, observabilidade e monitoramento.

2. Problema tratado

Em ecossistemas analíticos complexos, muitos problemas decorrem da ausência de acordos explícitos entre quem produz os dados e quem os consome. Fontes podem mudar de estrutura sem aviso, atributos podem ser removidos, categorias podem ser alteradas, periodicidades podem deixar de ser cumpridas, campos obrigatórios podem passar a vir vazios, APIs podem mudar sua resposta, regras de negócio podem ser modificadas e produtos analíticos podem ser impactados sem que os responsáveis sejam informados.

A ausência de contratos de dados pode gerar problemas como:

- pipelines quebrados por mudanças inesperadas em fontes;
- alterações de schema não comunicadas;
- dados entregues fora da periodicidade esperada;
- queda de qualidade sem responsabilização;
- ausência de definição sobre atributos obrigatórios;
- divergência entre produtores e consumidores sobre o significado dos dados;
- dificuldade de estabelecer expectativas de atualização;

- uso de dados sem clareza sobre finalidade e restrições;
- ausência de critérios mínimos para publicação de produtos;
- conflitos entre áreas sobre responsabilidade pelos dados;
- maior ocorrência de schema drift não planejado;
- aumento de falhas silenciosas;
- dificuldade de auditoria e rastreabilidade;
- perda de confiança nos produtos analíticos.

No UnespDataLens-RM, a dimensão Data Contracts busca prevenir esses problemas por meio da formalização de expectativas técnicas, semânticas, operacionais e de governança entre os diversos componentes do ecossistema analítico.

3. Objetivo da dimensão

O objetivo da dimensão **Data Contracts** é **formalizar acordos explícitos entre produtores e consumidores de dados**, definindo estrutura, qualidade, periodicidade, responsabilidade, versionamento, governança, segurança, mudanças permitidas e condições de uso dos ativos analíticos do UnespDataLens-RM.

Essa dimensão busca aumentar a previsibilidade e a confiabilidade dos pipelines, reduzir falhas decorrentes de mudanças não comunicadas, melhorar a governança dos dados e fortalecer a relação entre fontes, processos e produtos analíticos.

4. Relação com o UnespDataLens-RM

A dimensão Data Contracts se articula com diversos módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, os contratos ajudam a definir o que se espera de cada fonte: responsável, periodicidade, formato, cobertura temporal, granularidade, atributos obrigatórios e restrições de acesso.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, os contratos orientam a validação inicial dos dados recebidos, verificando se estrutura, volume, formato, campos obrigatórios e periodicidade estão conforme o esperado.

No **Módulo 3 — Integração e Harmonização**, os contratos contribuem para estabilizar chaves de integração, entidades compartilhadas, domínios de valores e regras de correspondência entre fontes.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, os contratos indicam quais atributos devem existir, quais regras de transformação dependem deles e quais alterações exigem revisão do pipeline.

No **Módulo 5 — Qualidade e Validação de Dados**, os contratos fornecem critérios explícitos para testes de qualidade, limiares, domínios válidos, obrigatoriedade, completude mínima e consistência.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, podem estabelecer requisitos mínimos para preservação de atributos necessários à avaliação de grupos, representatividade e equidade.

No **Módulo 7 — Armazenamento Analítico**, orientam a persistência de ativos em estruturas compatíveis com schemas contratados, versões e requisitos de consumo.

No **Módulo 8 — Disponibilização e Consumo Analítico**, definem condições para consumo de datasets, dashboards, APIs, relatórios e produtos de dados, incluindo atualização, disponibilidade, granularidade e restrições.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, os contratos são registrados como metadados formais dos ativos, associados a definições, domínios, responsáveis, regras e limitações.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, estabelecem responsabilidades, permissões, bases de uso, finalidades, restrições e obrigações de segurança.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, permitem identificar quais produtos dependem de um contrato e quais impactos ocorrem quando ele é violado.

No **Módulo 12 — Versionamento e Reprodutibilidade**, os contratos são versionados, permitindo comparar versões e reconstruir quais regras estavam vigentes em determinado momento.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, os contratos fornecem baselines e critérios para geração de alertas quando há violação de estrutura, qualidade, periodicidade ou disponibilidade.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, agentes podem consultar contratos para explicar violações, orientar usuários e recomendar ações corretivas.

5. Componentes de um contrato de dados

Um contrato de dados no UnespDataLens-RM pode conter componentes técnicos, semânticos, operacionais e de governança.

Componente	Descrição	Exemplo
Identificação do ativo	Nome, código, descrição e tipo do ativo contratado.	Base de matrículas semestrais.
Produtor dos dados	Área, sistema ou responsável pela geração dos dados.	Sistema acadêmico institucional.
Consumidores	Pipelines, produtos, equipes ou sistemas que dependem do ativo.	Dashboard de evasão, feature store.
Schema esperado	Atributos, tipos, obrigatoriedade, chaves e domínios.	`id_estudante`, `curso`, `situacao`, `periodo`.

Periodicidade	Frequência esperada de atualização ou entrega.	Mensal, semestral, diária.
Critérios de qualidade	Requisitos mínimos de completude, validade, consistência e unicidade.	Completude mínima de 95% no campo curso.
Regras semânticas	Definições de conceitos, categorias e códigos.	Definição de estudante ativo.
Política de mudança	Regras para alteração de schema, periodicidade ou domínio.	Mudanças críticas devem ser avisadas previamente.
SLA de dados	Expectativas de disponibilidade, atualização e resposta.	Atualização até o quinto dia útil.
Restrições de uso	Condições de acesso, privacidade e finalidade.	Uso apenas para análise institucional agregada.
Versionamento	Histórico de versões do contrato e do schema.	Contrato v1.2.
Ações em caso de violação	Procedimentos quando o contrato não for cumprido.	Bloquear ingestão, gerar alerta, notificar responsável.

6. Estratégias de implementação

A implementação de Data Contracts no UnespDataLens-RM pode envolver estratégias técnicas, metodológicas e institucionais.

Estratégia	Descrição
Definição de contratos por fonte crítica	Criar contratos para fontes relevantes que alimentam produtos, indicadores ou modelos importantes.

Contratos legíveis por humanos e máquinas	Manter documentação textual e representação estruturada em YAML, JSON, tabelas ou catálogo.
Validação automática contra contratos	Verificar se dados recebidos atendem a schema, tipos, domínios, obrigatoriedade e qualidade mínima.
Versionamento de contratos	Registrar mudanças nos contratos e permitir comparação entre versões.
Governança de mudanças	Exigir comunicação ou aprovação para alterações críticas em fontes contratadas.
Alertas de violação contratual	Gerar notificações quando periodicidade, schema, qualidade ou disponibilidade não forem cumpridos.
Integração com catálogo de dados	Registrar contratos junto aos metadados dos ativos.
Integração com lineage	Identificar produtos impactados por violação de contrato.
Definição de responsáveis	Associar produtores, consumidores e responsáveis técnicos a cada contrato.
Critérios de severidade	Classificar violações como baixa, média, alta ou crítica.
Revisão periódica	Atualizar contratos conforme evolução das fontes, regras e necessidades analíticas.
Registro de exceções	Documentar casos em que uma violação é aceita temporariamente.

7. Métricas associadas

A dimensão Data Contracts pode ser avaliada por métricas relacionadas à cobertura, conformidade, estabilidade, qualidade, violação e resolução de contratos.

Métrica	Sigla	Descrição
Taxa de ativos críticos com contrato	TACC	Proporção de ativos críticos que possuem contrato formal.
Taxa de contratos versionados	TCV	Proporção de contratos com histórico de versões.
Taxa de contratos com schema definido	TCSD	Proporção de contratos que especificam atributos, tipos, chaves e domínios.
Taxa de contratos com qualidade mínima definida	TCQM	Proporção de contratos com critérios mínimos de qualidade.
Taxa de contratos com SLA definido	TCSLA	Proporção de contratos com periodicidade, atualização ou disponibilidade definida.
Taxa de conformidade contratual	TCC	Proporção de execuções ou entregas que cumprem o contrato.
Taxa de violações contratuais	TVC	Proporção de execuções ou entregas com descumprimento do contrato.
Taxa de violações críticas	TVCr	Proporção de violações classificadas como críticas.
Tempo médio de resolução de violação	TMRV	Tempo médio entre detecção e correção da violação.
Taxa de mudanças comunicadas	TMC	Proporção de mudanças estruturais comunicadas antes de afetar consumidores.

Taxa de contratos revisados	TCR	Proporção de contratos revisados em determinado período.
Índice de maturidade de contratos de dados	IMCD	Indicador composto de cobertura, conformidade, versionamento e revisão.

Uma formulação possível para o **Índice de Maturidade de Contratos de Dados** é:

$$\mathbf{IMCD} = \mathbf{w_1TACC} + \mathbf{w_2TCV} + \mathbf{w_3TCSD} + \mathbf{w_4TCQM} + \mathbf{w_5TCSLA} + \mathbf{w_6TCC} + \mathbf{w_7TMC} - \mathbf{w_8TVC} - \mathbf{w_9TVCr}$$

em que:

- **TACC** representa a taxa de ativos críticos com contrato;
- **TCV** representa a taxa de contratos versionados;
- **TCSD** representa a taxa de contratos com schema definido;
- **TCQM** representa a taxa de contratos com qualidade mínima definida;
- **TCSLA** representa a taxa de contratos com SLA definido;
- **TCC** representa a taxa de conformidade contratual;
- **TMC** representa a taxa de mudanças comunicadas;
- **TVC** representa a taxa de violações contratuais;
- **TVCr** representa a taxa de violações críticas;
- $\mathbf{w_i}$ representa o peso atribuído a cada dimensão.

Quanto maior o IMCD, maior a maturidade do ecossistema na formalização e gestão de contratos de dados.

8. Artefatos relacionados

A dimensão Data Contracts pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de inventário, ingestão, qualidade, governança, metadados, lineage, versionamento, observabilidade e monitoramento.

Código	Artefato	Representação	Finalidade
DC1	Modelo de Contrato de Dados	Documento técnico ou template estruturado	Padronizar os elementos mínimos de um contrato de dados.
DC2	Registro de Contratos de Dados	Catálogo ou tabela de contratos	Manter inventário dos contratos existentes, seus responsáveis, versões e status.

DC3	Especificação de Schema Contratado	Tabela, YAML, JSON ou data contract	Definir atributos, tipos, domínios, chaves, obrigatoriedade e granularidade esperados.
DC4	Regras de Qualidade Contratadas	Tabela de regras ou arquivo validável	Definir completude mínima, validade, unicidade, consistência e limites aceitáveis.
DC5	Registro de SLA de Dados	Tabela ou documento de compromisso	Definir periodicidade, atualização, disponibilidade e tempo de resposta esperado.
DC6	Registro de Violações Contratuais	Log ou tabela de incidentes	Documentar descumprimentos de schema, qualidade, periodicidade, disponibilidade ou domínio.
DC7	Relatório de Conformidade Contratual	Relatório técnico-gerencial	Avaliar o grau de cumprimento dos contratos por fonte, ativo ou período.
DC8	Registro de Mudanças Contratuais	Changelog de contrato	Documentar alterações de schema, regras, periodicidade, qualidade ou responsáveis.
DC9	Matriz Produtor–Consumidor	Matriz de dependências	Identificar produtores, consumidores e produtos impactados por cada contrato.
DC10	Plano de Resposta a Violações	Documento operacional	Definir ações diante de violações, como alerta, bloqueio, reprocessamento ou exceção.
DC11	Registro de Exceções Contratuais	Tabela de exceções aprovadas	Documentar violações temporariamente aceitas, justificativas, prazos e responsáveis.
DC12	Painel de	Dashboard de	Monitorar cobertura,

Contratos de Dados	acompanhamento	conformidade, violações, revisões e maturidade dos contratos.

9. Critérios de validação

A dimensão Data Contracts pode ser considerada adequadamente tratada quando:

1. ativos críticos possuem contratos de dados formalizados;
2. contratos especificam produtor, consumidor, finalidade e responsável;
3. schemas esperados estão documentados;
4. atributos obrigatórios, tipos, domínios e chaves estão definidos;
5. critérios mínimos de qualidade estão explicitados;
6. periodicidade, atualização ou disponibilidade esperada estão registradas;
7. regras de mudança de schema ou domínio estão estabelecidas;
8. contratos possuem versionamento;
9. violações contratuais geram alertas e registros;
10. impactos sobre consumidores são identificados por lineage;
11. mudanças críticas são comunicadas previamente;
12. exceções são registradas, justificadas e acompanhadas;
13. contratos são revisados periodicamente;
14. consumidores conseguem verificar se um ativo está em conformidade com seu contrato.

10. Exemplo de aplicação no domínio educacional

Em um projeto educacional, um contrato de dados pode ser estabelecido para a base de matrículas utilizada em análises de evasão, desempenho e permanência.

Esse contrato pode definir que a base deve ser atualizada semestralmente, conter os campos `id_estudante`, `curso`, `periodo_letivo`, `situacao_academica`, `data_matricula` e `forma_ingresso`, além de especificar quais campos são obrigatórios, quais categorias são aceitas para situação acadêmica, qual é a regra de identificação do estudante e quais alterações precisam ser comunicadas previamente.

Se a fonte passar a incluir uma nova categoria de situação acadêmica, como “mobilidade” ou “abandono”, o contrato orienta que a mudança seja registrada, comunicada, versionada e avaliada quanto ao impacto nos indicadores de evasão e nos modelos preditivos.

Dessa forma, o contrato reduz o risco de que dashboards e features educacionais sejam afetados por mudanças estruturais ou semânticas não tratadas.

11. Exemplo de aplicação em políticas públicas

Em políticas públicas, um contrato de dados pode ser definido entre uma secretaria fornecedora de dados de atendimento social e o ecossistema analítico responsável por produzir indicadores territoriais.

O contrato pode estabelecer periodicidade mensal, campos obrigatórios, códigos válidos de tipo de atendimento, identificadores territoriais, critérios mínimos de completude, restrições de acesso e regras para alteração de categorias.

Se uma secretaria deixar de enviar dados no prazo ou modificar códigos de atendimento sem comunicação, o contrato permite gerar alerta, registrar violação, identificar dashboards e indicadores afetados e acionar o responsável pela fonte.

Assim, contratos de dados contribuem para tornar a integração intersetorial mais previsível, auditável e confiável.

12. Contribuição para o modelo de referência

A dimensão **Data Contracts** contribui para o UnespDataLens-RM ao formalizar expectativas entre produtores e consumidores de dados, reduzindo falhas causadas por mudanças inesperadas, ambiguidade semântica, baixa qualidade, ausência de periodicidade e falta de responsabilização.

Essa dimensão fortalece o modelo de referência ao:

- aumentar previsibilidade no uso de fontes e ativos;
- reduzir impactos de schema drift;
- formalizar requisitos mínimos de qualidade;
- melhorar comunicação entre produtores e consumidores;
- apoiar governança de mudanças;
- permitir validação automática de dados recebidos;
- fortalecer rastreabilidade entre contratos, fontes e produtos;
- reduzir falhas silenciosas;
- apoiar observabilidade e monitoramento;
- melhorar confiabilidade dos produtos analíticos;
- reduzir dívida de dados;
- sustentar evolução controlada do ecossistema.

13. Síntese da dimensão

A dimensão **Data Contracts** trata da formalização de acordos explícitos entre produtores e consumidores de dados, especificando estrutura, qualidade, periodicidade, responsabilidade, versionamento, governança, segurança e regras de mudança dos ativos analíticos.

No UnespDataLens-RM, essa dimensão atua como capacidade complementar para tornar o ciclo de dados mais previsível, confiável, rastreável e governado. Ao incorporar contratos de dados, validação automática, versionamento, SLAs, registros de violação, análise de impacto e planos de resposta, o modelo de referência reduz riscos operacionais e fortalece a confiança nos dados utilizados para análise, modelagem, dashboards, APIs, produtos e tomada de decisão baseada em evidências.

Dimensão Complementar — Qualidade de Metadados

1. Definição da dimensão

A dimensão **Qualidade de Metadados** refere-se ao grau em que os metadados associados aos ativos de dados são completos, corretos, atualizados, consistentes, compreensíveis, padronizados, rastreáveis, úteis e adequados ao uso analítico, técnico, gerencial e científico.

No contexto do **UnespDataLens-RM**, os metadados são informações que descrevem, contextualizam e qualificam os ativos de dados. Eles indicam, por exemplo, a origem de uma fonte, o significado de um atributo, o responsável por um dataset, a periodicidade de atualização, a granularidade dos registros, a versão de um produto, as regras de transformação aplicadas, a sensibilidade dos dados, as restrições de acesso, os indicadores de qualidade, a linhagem, as limitações e a finalidade de uso.

Assim, a qualidade dos metadados é uma condição fundamental para que os dados possam ser encontrados, compreendidos, integrados, governados, reutilizados, auditados e interpretados corretamente. Dados tecnicamente corretos podem perder valor analítico se seus metadados forem incompletos, ambíguos, desatualizados ou inconsistentes.

No UnespDataLens-RM, a dimensão Qualidade de Metadados é tratada como uma capacidade complementar e transversal, pois os metadados são produzidos e consumidos por todos os módulos do modelo de referência. Eles apoiam inventário de fontes, ingestão, integração, transformação, validação, governança, linhagem, versionamento, armazenamento, disponibilização, feature engineering, monitoramento, agentes inteligentes e consumo analítico.

2. Problema tratado

Em ecossistemas analíticos complexos, é comum que os dados estejam disponíveis, mas que faltem informações suficientes para compreendê-los e utilizá-los corretamente. Um dataset pode existir em um repositório, mas sem descrição clara. Uma coluna pode ter nome técnico, mas sem significado semântico. Um indicador pode estar publicado em um dashboard, mas sem definição formal. Uma feature pode estar disponível em uma feature store, mas sem regra de cálculo documentada. Um produto de dados pode ser consumido por gestores, mas sem indicação de versão, período de referência ou limitações.

A baixa qualidade dos metadados pode gerar problemas como:

- dificuldade de localizar ativos relevantes;
- interpretação incorreta de atributos e indicadores;
- uso de dados fora de contexto;
- duplicação de datasets e produtos analíticos;
- baixa reutilização de ativos existentes;
- dificuldade de integração entre fontes;
- ambiguidade semântica entre termos semelhantes;
- ausência de responsável pelo dado;
- desconhecimento da qualidade ou confiabilidade de um ativo;
- dificuldade de rastrear origem e transformações;

- impossibilidade de reproduzir resultados;
- uso indevido de dados sensíveis;
- falhas de governança e conformidade;
- aumento da dependência de conhecimento tácito;
- baixa confiança dos usuários nos produtos analíticos.

No UnespDataLens-RM, essa dimensão busca garantir que os metadados não sejam tratados apenas como documentação auxiliar, mas como ativos estruturantes da Engenharia Analítica de Dados. A qualidade dos metadados determina a capacidade do modelo de referência de sustentar descoberta, governança, rastreabilidade, reprodutibilidade e tomada de decisão baseada em evidências.

3. Objetivo da dimensão

O objetivo da dimensão **Qualidade de Metadados** é **avaliar, aprimorar e manter a completude, consistência, atualidade, precisão, padronização, rastreabilidade e utilidade dos metadados associados aos ativos analíticos do UnespDataLens-RM**.

Essa dimensão busca assegurar que fontes, datasets, atributos, regras, features, indicadores, data marts, dashboards, APIs, relatórios, modelos, grafos e produtos de dados sejam acompanhados por metadados suficientes para sua compreensão, uso, auditoria, governança e reutilização.

4. Relação com o UnespDataLens-RM

A dimensão Qualidade de Metadados se articula com diversos módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, garante que fontes sejam registradas com nome, descrição, origem, responsável, domínio, formato, periodicidade, cobertura temporal, granularidade, restrições, sensibilidade e limitações conhecidas.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, assegura que datasets brutos, snapshots, logs, checksums, data de ingestão, método de extração, parâmetros e destino estejam documentados.

No **Módulo 3 — Integração e Harmonização**, apoia a documentação de chaves, entidades, equivalências semânticas, conflitos, regras de correspondência, granularidade e fontes integradas.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, garante que regras de transformação, atributos derivados, parâmetros, tratamentos de ausentes, duplicidades, outliers e impactos estejam registrados.

No **Módulo 5 — Qualidade e Validação de Dados**, conecta metadados a regras de validação, métricas de qualidade, limiares, não conformidades, status de aprovação e ações corretivas.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, documenta atributos sensíveis, grupos avaliados, distribuições, métricas de fairness, riscos residuais e limitações de uso.

No **Módulo 7 — Armazenamento Analítico**, assegura que zonas, schemas, localização, formato, particionamento, versão, data marts e estruturas persistidas estejam descritos.

No **Módulo 8 — Disponibilização e Consumo Analítico**, garante que produtos de dados, dashboards, APIs e relatórios possuam finalidade, público-alvo, versão, restrição de acesso, periodicidade, qualidade, limitações e responsável.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, a qualidade de metadados constitui o foco principal de avaliação e manutenção, pois esse módulo organiza o catálogo, dicionários, glossários, modelos semânticos e fichas de ativos.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, metadados de qualidade sustentam classificação de sensibilidade, finalidade, base de uso, permissões, responsáveis e conformidade.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, metadados qualificados permitem rastrear origem, dependências, regras, versões, processos e consumidores.

No **Módulo 12 — Versionamento e Reprodutibilidade**, metadados registram versões, parâmetros, ambientes, changelogs, manifestos e condições de reexecução.

No **Módulo 13 — Representações Analíticas**, metadados documentam unidade de análise, granularidade, atributos, relações, eventos, janelas temporais, grafos e limitações.

No **Módulo 14 — Feature Engineering e Feature Store**, metadados qualificam features por origem, regra, versão, qualidade, sensibilidade, finalidade e consumidores.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, metadados permitem identificar responsáveis, periodicidade, baselines, versões e produtos impactados.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, metadados de qualidade são fundamentais para respostas confiáveis, contextualizadas, rastreáveis e úteis.

5. Dimensões da qualidade de metadados

A qualidade dos metadados pode ser avaliada por diferentes dimensões complementares.

Dimensão	Descrição	Exemplo
Completude	Grau em que os campos obrigatórios de metadados estão preenchidos.	Dataset possui descrição, origem, responsável, versão e periodicidade.
Precisão	Grau em que os metadados representam corretamente o ativo descrito.	Descrição do atributo corresponde ao seu significado real.

Consistência	Ausência de contradições entre metadados de diferentes ativos ou módulos.	O mesmo indicador possui a mesma definição no catálogo e no dashboard.
Atualidade	Grau em que os metadados refletem a versão mais recente do ativo.	Metadados atualizados após mudança de schema.
Padronização	Uso de vocabulários, formatos, campos e classificações comuns.	Todos os ativos usam padrão único de classificação de sensibilidade.
Clareza	Facilidade de compreensão dos metadados por usuários técnicos e não técnicos.	Descrição de indicador em linguagem compreensível.
Rastreabilidade	Capacidade de conectar metadados à origem, versão, regras e consumidores do ativo.	Feature aponta para datasets e regras de cálculo utilizados.
Unicidade	Ausência de registros duplicados ou conflitantes para o mesmo ativo.	Um dashboard possui um único registro oficial no catálogo.
Utilidade	Grau em que os metadados ajudam usuários a descobrir, compreender e usar o ativo.	Usuário consegue decidir se um dataset serve ao seu problema.
Conformidade	Aderência a padrões, políticas, contratos e requisitos de governança.	Ativos sensíveis possuem finalidade e restrição documentadas.

Essas dimensões permitem avaliar se os metadados são suficientes para apoiar o uso responsável e confiável dos dados.

6. Estratégias de melhoria da qualidade de metadados

A melhoria da qualidade de metadados no UnespDataLens-RM pode envolver estratégias técnicas, metodológicas e de governança.

Estratégia	Descrição
Modelo padronizado de metadados	Definir campos obrigatórios e opcionais para cada tipo de ativo.
Catálogo unificado	Centralizar registros de fontes, datasets, features, produtos, dashboards e APIs.
Dicionário de dados consolidado	Documentar atributos, tipos, domínios, significados e regras de uso.
Glossário controlado	Padronizar termos, conceitos, sinônimos e definições do domínio.
Validação de metadados	Verificar completude, consistência, atualidade e conformidade dos registros.
Curadoria periódica	Revisar metadados antigos, incompletos, ambíguos ou desatualizados.
Responsabilidade definida	Associar cada ativo e seus metadados a um responsável técnico ou institucional.
Automação de captura	Extrair automaticamente schemas, tipos, volumes, datas, versões e estatísticas.
Enriquecimento semântico	Associar atributos a conceitos, entidades, domínios e regras de negócio.
Integração com lineage	Vincular metadados à origem, transformações, versões e consumidores.
Integração com governança	Registrar sensibilidade, restrições, finalidade e políticas

	aplicáveis.
Uso de agentes inteligentes	Apoiar geração, revisão, sugestão e identificação de lacunas em metadados.
Controle de versões de metadados	Registrar mudanças nos metadados e preservar histórico.
Critérios de publicação	Impedir disponibilização de ativos críticos sem metadados mínimos.

7. Métricas associadas

A dimensão Qualidade de Metadados pode ser avaliada por métricas relacionadas à completude, consistência, atualização, rastreabilidade, padronização e utilidade dos registros.

Métrica	Sigla	Descrição
Taxa de ativos com metadados mínimos	TAMM	Proporção de ativos com campos obrigatórios preenchidos.
Índice de completude de metadados	ICM	Grau de preenchimento dos campos esperados.
Taxa de atributos documentados	TAD	Proporção de atributos com descrição, tipo, domínio e significado.
Taxa de ativos com responsável definido	TARD	Proporção de ativos com responsável técnico ou institucional.
Taxa de metadados atualizados	TMA	Proporção de metadados revisados ou atualizados no período esperado.

Taxa de inconsistências de metadados	TIM	Proporção de registros com contradições, duplicidades ou divergências.
Taxa de ativos com lineage associado	TALA	Proporção de ativos com metadados conectados à linhagem.
Taxa de ativos com sensibilidade documentada	TASD	Proporção de ativos com classificação de sensibilidade registrada.
Taxa de termos com definição no glossário	TTDG	Proporção de termos críticos definidos no glossário.
Taxa de metadados versionados	TMV	Proporção de metadados com histórico de alterações.
Tempo médio de descoberta	TMD	Tempo necessário para localizar e compreender um ativo relevante.
Índice de qualidade de metadados	IQM	Indicador composto de completude, precisão, consistência, atualidade e utilidade.

Uma formulação possível para o **Índice de Qualidade de Metadados** é:

$$\mathbf{IQM} = \mathbf{w_1ICM} + \mathbf{w_2TAD} + \mathbf{w_3TARD} + \mathbf{w_4TMA} + \mathbf{w_5TALA} + \mathbf{w_6TASD} + \mathbf{w_7TTDG} - \mathbf{w_8TIM}$$

em que:

- **ICM** representa o índice de completude de metadados;
- **TAD** representa a taxa de atributos documentados;
- **TARD** representa a taxa de ativos com responsável definido;
- **TMA** representa a taxa de metadados atualizados;
- **TALA** representa a taxa de ativos com lineage associado;
- **TASD** representa a taxa de ativos com sensibilidade documentada;
- **TTDG** representa a taxa de termos definidos no glossário;
- **TIM** representa a taxa de inconsistências de metadados;
- w_i representa o peso atribuído a cada dimensão.

Quanto maior o IQM, maior a qualidade dos metadados associados ao ecossistema analítico.

8. Artefatos relacionados

A dimensão Qualidade de Metadados pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de catálogo, semântica, governança, lineage, versionamento, qualidade, monitoramento e agentes inteligentes.

Código	Artefato	Representação	Finalidade
QM1	Modelo de Qualidade de Metadados	Documento técnico	Definir dimensões, critérios e campos obrigatórios para avaliação dos metadados.
QM2	Perfil de Metadados por Tipo de Ativo	Tabela ou template	Especificar metadados mínimos para fontes, datasets, atributos, features, dashboards, APIs e produtos.
QM3	Relatório de Completude de Metadados	Relatório técnico	Avaliar campos preenchidos, ausentes e obrigatórios por ativo.
QM4	Relatório de Consistência de Metadados	Relatório técnico	Identificar contradições, duplicidades, divergências e registros conflitantes.
QM5	Registro de Atualização de Metadados	Changelog ou tabela versionada	Documentar revisões, alterações, responsáveis e datas de atualização.
QM6	Registro de Lacunas de Metadados	Tabela de pendências	Listar ativos sem descrição, responsável, versão, sensibilidade, finalidade ou lineage.
QM7	Matriz de Responsabilidade de Metadados	Matriz ativo × responsável	Definir ownership sobre atualização e curadoria dos metadados.
QM8	Relatório de Qualidade Semântica	Relatório técnico	Avaliar clareza, definição, ambiguidade, equivalência e padronização dos conceitos.

QM9	Painel de Qualidade de Metadados	Dashboard	Monitorar completude, atualização, inconsistências, lacunas e evolução da qualidade.
QM10	Plano de Curadoria de Metadados	Documento operacional	Definir ações, prioridades, responsáveis e prazos para melhoria dos metadados.
QM11	Registro de Validação de Metadados	Tabela de revisão	Documentar validação humana ou automática dos metadados críticos.
QM12	Índice de Qualidade de Metadados	Indicador composto	Sintetizar a qualidade dos metadados por ativo, domínio, módulo ou produto.

9. Critérios de validação

A dimensão Qualidade de Metadados pode ser considerada adequadamente tratada quando:

1. os ativos críticos possuem metadados mínimos definidos;
2. fontes, datasets, atributos, features, dashboards, APIs e produtos possuem descrição clara;
3. os responsáveis pelos ativos e metadados estão identificados;
4. atributos críticos possuem tipo, domínio, significado e regra de uso documentados;
5. termos relevantes estão definidos em glossário;
6. metadados estão atualizados em relação às versões vigentes dos ativos;
7. mudanças nos metadados são registradas;
8. inconsistências e duplicidades são identificadas e tratadas;
9. metadados de sensibilidade, finalidade e restrição de acesso estão preenchidos quando aplicável;
10. ativos críticos possuem associação com lineage;
11. usuários conseguem localizar e compreender ativos com menor esforço;
12. agentes inteligentes conseguem consultar metadados confiáveis para responder perguntas e gerar recomendações;
13. existe processo de curadoria periódica;
14. o índice de qualidade de metadados melhora ao longo do tempo.

10. Exemplo de aplicação no domínio educacional

Em um projeto educacional, a qualidade de metadados é essencial para compreender bases de matrícula, histórico escolar, notas, frequência, evasão, desempenho, interações em ambientes virtuais e dados socioeconômicos.

Por exemplo, um atributo chamado `situacao` pode ser ambíguo se não houver metadados indicando se ele representa situação da matrícula, situação acadêmica do estudante, situação no curso ou situação em uma disciplina. Sem essa definição, análises de evasão ou permanência podem ser interpretadas incorretamente.

A dimensão Qualidade de Metadados permite documentar o significado do atributo, seus valores possíveis, sua fonte, sua regra de atualização, seu período de validade, sua relação com indicadores de evasão e suas limitações. Isso reduz ambiguidades e melhora a confiabilidade dos produtos educacionais.

11. Exemplo de aplicação em políticas públicas

Em políticas públicas, diferentes órgãos podem usar termos semelhantes com significados distintos, como atendimento, ocorrência, família, beneficiário, território, vulnerabilidade ou serviço. Sem metadados qualificados, a integração dessas bases pode gerar interpretações incorretas.

Por exemplo, uma secretaria pode registrar “atendimento” como qualquer contato com o cidadão, enquanto outra pode registrar apenas atendimentos formalizados em sistema. Se essa diferença não estiver documentada, indicadores integrados podem superestimar ou subestimar a cobertura de serviços públicos.

A dimensão Qualidade de Metadados permite registrar definições, domínios, regras de negócio, responsáveis, periodicidade, limitações e equivalências semânticas, fortalecendo a análise intersetorial e a tomada de decisão baseada em evidências.

12. Contribuição para o modelo de referência

A dimensão **Qualidade de Metadados** contribui para o UnespDataLens-RM ao reconhecer que a confiabilidade dos dados depende também da qualidade das informações que descrevem, contextualizam e qualificam esses dados.

Essa dimensão fortalece o modelo de referência ao:

- melhorar descoberta e compreensão dos ativos;
- reduzir ambiguidades semânticas;
- apoiar integração entre fontes;
- fortalecer governança e conformidade;
- aumentar rastreabilidade e reprodutibilidade;
- reduzir dependência de conhecimento tácito;
- melhorar qualidade dos catálogos e glossários;
- apoiar agentes inteligentes com informações confiáveis;
- reduzir data overload, data silos e data debt;
- aumentar confiança dos usuários nos produtos analíticos;
- favorecer reuso responsável dos dados;
- sustentar tomada de decisão baseada em evidências.

13. Síntese da dimensão

A dimensão **Qualidade de Metadados** trata da completude, precisão, consistência, atualidade, padronização, clareza, rastreabilidade, utilidade e conformidade dos metadados associados aos ativos analíticos. No UnespDataLens-RM, essa dimensão atua como

capacidade complementar para garantir que dados, features, indicadores, dashboards, APIs, grafos, modelos e produtos sejam compreensíveis, localizáveis, governáveis, auditáveis e reutilizáveis.

Ao incorporar modelos de metadados, catálogos, dicionários, glossários, validações, relatórios de completude, curadoria, versionamento e indicadores de qualidade, o modelo de referência fortalece sua capacidade de transformar dados em ativos analíticos confiáveis e semanticamente compreensíveis.

Dimensão Complementar — Data Discoverability

1. Definição da dimensão

A dimensão **Data Discoverability**, ou **descoberta de dados**, refere-se à capacidade de localizar, identificar, compreender, avaliar e selecionar ativos de dados relevantes dentro de um ecossistema analítico. Essa dimensão envolve os mecanismos, metadados, catálogos, descrições, classificações, buscas, recomendações, filtros, glossários, relações semânticas e informações de qualidade que permitem aos usuários encontrar os dados mais adequados para uma finalidade específica.

No contexto do **UnespDataLens-RM**, a descoberta de dados não significa apenas “procurar uma base” em um repositório. Ela envolve a capacidade de responder a perguntas como: quais dados existem? onde estão? quem é o responsável? qual é a versão vigente? qual é a origem? quais atributos estão disponíveis? qual é a qualidade? quais restrições de acesso existem? quais produtos utilizam esses dados? quais dados são oficiais, experimentais, obsoletos ou recomendados? quais ativos podem ser reutilizados em uma nova análise?

A Data Discoverability é, portanto, uma condição fundamental para que dados, datasets, dashboards, APIs, relatórios, features, indicadores, grafos, modelos e produtos analíticos sejam efetivamente utilizados. Um ativo de dados pode existir e possuir qualidade técnica, mas, se não puder ser encontrado, compreendido ou avaliado por seus potenciais usuários, seu valor analítico e institucional é reduzido.

No UnespDataLens-RM, essa dimensão é complementar e transversal porque depende da articulação entre inventário de fontes, metadados, catálogo, semântica, governança, qualidade, lineage, versionamento, disponibilização, consumo analítico, feature store, monitoramento e agentes inteligentes.

2. Problema tratado

Em ecossistemas analíticos complexos, é comum que diferentes equipes, sistemas e projetos produzam grande quantidade de fontes, datasets, relatórios, dashboards, indicadores, APIs e features. Entretanto, esses ativos muitas vezes permanecem dispersos, pouco documentados, sem padronização de nomes, sem descrição clara, sem responsável identificado, sem indicação de qualidade, sem versão registrada e sem mecanismo eficiente de busca.

Essa situação gera um paradoxo: os dados existem, mas não são encontrados; os produtos foram produzidos, mas não são reutilizados; os indicadores estão disponíveis, mas não são compreendidos; as features foram calculadas, mas são recriadas por outras equipes; as bases estão armazenadas, mas os usuários não sabem se podem utilizá-las.

A baixa capacidade de descoberta de dados pode gerar problemas como:

- dificuldade de localizar fontes, datasets, dashboards, APIs e features relevantes;
- duplicação de esforços analíticos;
- criação repetida de produtos semelhantes;
- baixa reutilização de ativos existentes;
- dependência de conhecimento tácito de pessoas específicas;

- uso de dados inadequados por falta de alternativas visíveis;
- dificuldade de distinguir ativos oficiais, experimentais, obsoletos ou depreciados;
- demora na preparação de novas análises;
- pouca transparência sobre origem, qualidade e restrições dos dados;
- aumento da sobrecarga informacional;
- perda de confiança no ecossistema analítico;
- redução do valor dos ativos produzidos.

No UnespDataLens-RM, a dimensão Data Discoverability busca resolver esse problema ao estruturar mecanismos que tornem os ativos de dados localizáveis, compreensíveis, avaliáveis e reutilizáveis.

3. Objetivo da dimensão

O objetivo da dimensão **Data Discoverability** é **ampliar a capacidade de descoberta, compreensão, avaliação e reutilização dos ativos analíticos do UnespDataLens-RM**, por meio de catálogos, metadados qualificados, glossários, mecanismos de busca, filtros, classificações, recomendações, indicadores de qualidade, lineage, versionamento e agentes inteligentes.

Essa dimensão busca reduzir o tempo e o esforço necessários para encontrar dados relevantes, ao mesmo tempo em que aumenta a confiança do usuário na escolha do ativo mais adequado à sua finalidade analítica.

4. Relação com o UnespDataLens-RM

A dimensão Data Discoverability se articula com diversos módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, a descoberta começa com o mapeamento das fontes existentes, seus responsáveis, domínios, formatos, periodicidades, coberturas temporais, granularidades e restrições.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, os registros de ingestão, snapshots e manifestos ajudam a tornar visíveis os dados incorporados ao ambiente analítico.

No **Módulo 3 — Integração e Harmonização**, a descoberta é ampliada pela documentação de entidades comuns, chaves, equivalências semânticas e relações entre fontes.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, atributos tratados, variáveis derivadas e datasets enriquecidos precisam ser documentados para serem localizados e reutilizados.

No **Módulo 5 — Qualidade e Validação de Dados**, métricas de qualidade auxiliam o usuário a avaliar se um ativo descoberto é adequado ao uso pretendido.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, registros de representatividade, grupos avaliados, limitações e riscos de viés contribuem para uma descoberta responsável.

No **Módulo 7 — Armazenamento Analítico**, a organização de zonas, data marts, views e datasets curados facilita a localização de ativos prontos para consumo.

No **Módulo 8 — Disponibilização e Consumo Analítico**, produtos como dashboards, APIs, relatórios e datasets curados devem ser encontráveis, descritos e classificados por finalidade, público-alvo e restrição de acesso.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, ocorre a principal sustentação da descoberta de dados, por meio de catálogo, dicionário, glossário, modelos semânticos, fichas de ativos e índices de busca.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, a descoberta deve respeitar permissões, classificação de sensibilidade, finalidade de uso e políticas de acesso.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, a descoberta é enriquecida pela possibilidade de localizar ativos relacionados por origem, transformação, dependência e consumo.

No **Módulo 12 — Versionamento e Reprodutibilidade**, a descoberta deve permitir identificar versões vigentes, históricas, experimentais, depreciadas ou reproduzíveis.

No **Módulo 13 — Representações Analíticas**, os usuários podem descobrir representações tabulares, temporais, relacionais e em grafos adequadas ao problema analítico.

No **Módulo 14 — Feature Engineering e Feature Store**, a descoberta de features permite reuso, evita duplicação e apoia construção de modelos e indicadores.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, informações de uso, atualização, obsolescência e estabilidade contribuem para recomendar ou depreciar ativos.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, agentes podem apoiar consultas em linguagem natural, recomendação de ativos, explicação de produtos, busca semântica e orientação ao usuário.

5. Componentes da descoberta de dados

A descoberta de dados envolve diferentes componentes que, combinados, aumentam a capacidade de localização e compreensão dos ativos.

Componente	Descrição	Exemplo
Catálogo de dados	Registro centralizado dos ativos disponíveis.	Catálogo de datasets, dashboards e APIs.
Metadados descritivos	Informações sobre nome, descrição, domínio, responsável e finalidade.	Dataset de matrículas semestrais.
Metadados	Informações sobre formato, schema,	Tabela em Parquet, versão

técnicos	tipo, localização e versão.	2.1.
Metadados de qualidade	Indicadores sobre completude, validade, consistência e atualidade.	Completude de 97% no campo curso.
Metadados de governança	Sensibilidade, acesso, finalidade e restrições.	Uso restrito a análise agregada.
Glossário e semântica	Definição de termos, conceitos e equivalências.	Definição institucional de evasão.
Lineage	Origem, transformações e dependências.	Dashboard depende da base de matrícula.
Busca textual e semântica	Localização por palavras-chave, conceitos ou similaridade.	Buscar “evasão”, “abandono”, “permanência”.
Filtros e facetas	Refinamento por domínio, tipo, qualidade, versão ou responsável.	Filtrar por dados educacionais e ativos oficiais.
Recomendação de ativos	Sugestão de dados relevantes para determinada finalidade.	Recomendar feature set para evasão.
Sinalização de status	Indicação de ativo oficial, experimental, obsoleto ou depreciado.	Dashboard oficial de evasão.
Feedback de usuários	Avaliação de utilidade, clareza e adequação dos ativos.	Usuários avaliam dataset como útil.

6. Estratégias de implementação

A implementação da Data Discoverability no UnespDataLens-RM pode envolver estratégias técnicas, semânticas, metodológicas e de governança.

Estratégia	Descrição
Catálogo unificado de ativos	Centralizar fontes, datasets, data marts, dashboards, APIs, relatórios, features e produtos.
Indexação de metadados	Tornar nomes, descrições, atributos, tags e glossários pesquisáveis.
Busca semântica	Permitir localização por conceitos relacionados, sinônimos e termos do domínio.
Tags e classificações	Classificar ativos por domínio, finalidade, status, sensibilidade, tipo e qualidade.
Ficha de ativo	Apresentar informações essenciais para compreensão e uso de cada ativo.
Sinalização de qualidade	Exibir métricas ou status de qualidade para apoiar seleção de ativos.
Sinalização de status	Indicar se o ativo é oficial, experimental, em validação, depreciado ou obsoleto.
Relacionamento entre ativos	Mostrar fontes, derivados, consumidores, versões e produtos relacionados.
Glossário integrado	Associar termos do domínio a datasets, atributos, indicadores e produtos.
Recomendação baseada em uso	Sugerir ativos mais utilizados, mais recentes ou mais adequados ao contexto.

Controle de acesso na busca	Exibir apenas ativos permitidos ou metadados compatíveis com o perfil do usuário.
Feedback e avaliação de utilidade	Coletar percepções dos usuários sobre clareza, relevância e adequação dos ativos.
Agentes inteligentes de descoberta	Permitir consultas em linguagem natural sobre dados, indicadores, features e produtos.
Curadoria contínua do catálogo	Revisar ativos duplicados, obsoletos, mal descritos ou sem responsável.

7. Métricas associadas

A dimensão Data Discoverability pode ser avaliada por métricas relacionadas à cobertura, busca, localização, uso, relevância, reutilização e satisfação dos usuários.

Métrica	Sigla	Descrição
Taxa de ativos catalogados	TAC	Proporção de ativos existentes registrados no catálogo.
Taxa de ativos pesquisáveis	TAP	Proporção de ativos indexados e recuperáveis por busca.
Taxa de ativos com descrição adequada	TADA	Proporção de ativos com descrição clara, finalidade e contexto de uso.
Taxa de ativos com tags ou classificações	TATC	Proporção de ativos classificados por domínio, tipo, status ou finalidade.
Taxa de ativos com qualidade visível	TAQV	Proporção de ativos com métricas ou status de qualidade acessíveis ao usuário.

Taxa de ativos com lineage visível	TALV	Proporção de ativos com origem, dependências ou consumidores disponíveis para consulta.
Tempo médio de descoberta	TMD	Tempo necessário para localizar um ativo adequado.
Taxa de sucesso na busca	TSB	Proporção de consultas que retornam ativos úteis.
Taxa de reutilização de ativos descobertos	TRAD	Proporção de ativos localizados que são reutilizados em análises ou produtos.
Taxa de consultas sem resultado útil	TCSU	Proporção de buscas que não retornam ativos relevantes.
Índice de satisfação com descoberta	ISD	Grau de satisfação dos usuários com a localização e compreensão dos ativos.
Índice de descobribilidade de dados	IDD	Indicador composto da capacidade de localização, avaliação e reutilização dos ativos.

Uma formulação possível para o **Índice de Descobribilidade de Dados** é:

$$\mathbf{IDD} = w_1\mathbf{TAC} + w_2\mathbf{TAP} + w_3\mathbf{TADA} + w_4\mathbf{TATC} + w_5\mathbf{TAQV} + w_6\mathbf{TALV} + w_7\mathbf{TSB} + w_8\mathbf{TRAD} - w_9\mathbf{TCSU} - w_{10}\mathbf{TMD_n}$$

em que:

- **TAC** representa a taxa de ativos catalogados;
- **TAP** representa a taxa de ativos pesquisáveis;
- **TADA** representa a taxa de ativos com descrição adequada;
- **TATC** representa a taxa de ativos com tags ou classificações;
- **TAQV** representa a taxa de ativos com qualidade visível;
- **TALV** representa a taxa de ativos com lineage visível;
- **TSB** representa a taxa de sucesso na busca;
- **TRAD** representa a taxa de reutilização de ativos descobertos;
- **TCSU** representa a taxa de consultas sem resultado útil;
- $\mathbf{TMD_n}$ representa o tempo médio de descoberta normalizado;
- w_i representa o peso atribuído a cada dimensão.

Quanto maior o IDD, maior a capacidade do ecossistema de tornar seus ativos localizáveis, compreensíveis e reutilizáveis.

8. Artefatos relacionados

A dimensão Data Discoverability pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de inventário, catálogo, metadados, semântica, governança, lineage, consumo analítico, feature store, monitoramento e agentes inteligentes.

Código	Artefato	Representação	Finalidade
DDI1	Catálogo de Ativos Descobertos	Catálogo ou repositório pesquisável	Registrar fontes, datasets, data marts, dashboards, APIs, features, grafos, modelos e produtos disponíveis.
DDI2	Ficha de Descoberta do Ativo	Documento ou página de catálogo	Apresentar descrição, finalidade, responsável, versão, qualidade, restrições, origem e consumidores.
DDI3	Índice de Busca de Dados	Índice textual ou semântico	Permitir busca por nomes, atributos, tags, glossário, descrições e conceitos relacionados.
DDI4	Taxonomia de Classificação de Ativos	Tabela ou modelo classificatório	Organizar ativos por domínio, tipo, finalidade, status, sensibilidade e qualidade.
DDI5	Mapa de Relações entre Ativos	Grafo ou matriz de dependências	Mostrar relações entre fontes, datasets, features, dashboards, APIs, modelos e produtos.
DDI6	Registro de Termos de Busca e Lacunas	Tabela de consultas	Identificar buscas frequentes, termos sem resultado, ambiguidade e necessidades não atendidas.
DDI7	Relatório de Uso e Reutilização	Relatório técnico-gerencial	Avaliar quais ativos descobertos são acessados, reutilizados ou abandonados.

DDI8	Registro de Recomendações de Ativos	Tabela ou serviço de recomendação	Sugerir ativos relevantes com base em finalidade, contexto, qualidade, uso e permissões.
DDI9	Painel de Discoverability	Dashboard	Monitorar cobertura do catálogo, sucesso de busca, tempo de descoberta e reutilização.
DDI10	Plano de Curadoria da Descoberta	Documento operacional	Definir ações para melhorar descrição, tags, indexação, classificação e organização dos ativos.
DDI11	Registro de Ativos Oficiais e Recomendados	Lista controlada	Diferenciar ativos oficiais, experimentais, depreciados, obsoletos e recomendados.
DDI12	Relatório de Experiência de Descoberta	Relatório qualitativo e quantitativo	Avaliar satisfação, dificuldades, termos ambíguos e efetividade da busca para os usuários.

9. Critérios de validação

A dimensão Data Discoverability pode ser considerada adequadamente tratada quando:

1. os ativos relevantes estão registrados em catálogo;
2. os ativos possuem descrições claras e compreensíveis;
3. fontes, datasets, dashboards, APIs, features e produtos podem ser encontrados por busca;
4. usuários conseguem filtrar ativos por domínio, tipo, finalidade, status, qualidade e restrição;
5. ativos oficiais, experimentais, depreciados e obsoletos são distinguíveis;
6. métricas de qualidade estão visíveis ou acessíveis durante a descoberta;
7. lineage, versões e dependências podem ser consultados;
8. termos relevantes estão definidos em glossário e associados aos ativos;
9. buscas por termos comuns retornam resultados úteis;
10. consultas sem resultado útil são registradas e analisadas;
11. ativos descobertos são efetivamente reutilizados;
12. usuários conseguem identificar se um ativo é adequado à sua finalidade;
13. agentes inteligentes conseguem apoiar a descoberta por linguagem natural;
14. o tempo médio de descoberta diminui ao longo do tempo.

10. Exemplo de aplicação no domínio educacional

Em um projeto educacional, a Data Discoverability permite que pesquisadores, gestores e equipes técnicas encontrem rapidamente bases relacionadas a matrícula, histórico acadêmico, evasão, frequência, desempenho, interações em ambiente virtual, perfil socioeconômico, bolsas, permanência e indicadores institucionais.

Por exemplo, ao buscar por “evasão”, o usuário poderia encontrar datasets oficiais, dashboards relacionados, features disponíveis, indicadores validados, definições no glossário, modelos que utilizam essa variável, versões históricas e restrições de acesso. Também poderia distinguir entre produtos oficiais, experimentais e depreciados.

Essa capacidade reduz duplicação de análises, melhora a escolha dos dados, aumenta a reutilização de ativos e fortalece a confiabilidade dos estudos educacionais.

11. Exemplo de aplicação em políticas públicas

Em políticas públicas, a descoberta de dados pode apoiar gestores e analistas na localização de bases sobre território, vulnerabilidade, saúde, educação, assistência social, atendimento, infraestrutura, orçamento e indicadores socioeconômicos.

Por exemplo, ao buscar por “vulnerabilidade territorial”, o usuário poderia encontrar indicadores oficiais, bases de origem, dashboards territoriais, mapas, relatórios, regras de cálculo, periodicidade de atualização, nível de agregação e restrições de uso.

Com isso, a Data Discoverability reduz a dependência de contatos informais entre áreas, melhora a transparência dos dados disponíveis e apoia decisões intersetoriais baseadas em evidências.

12. Contribuição para o modelo de referência

A dimensão **Data Discoverability** contribui para o UnespDataLens-RM ao transformar a localização e compreensão dos ativos de dados em uma capacidade explícita do modelo de referência.

Essa dimensão fortalece o modelo de referência ao:

- tornar ativos de dados localizáveis e compreensíveis;
- reduzir dependência de conhecimento tácito;
- aumentar reutilização de datasets, dashboards, APIs e features;
- reduzir duplicação de esforços;
- melhorar governança do catálogo;
- apoiar seleção de dados adequados à finalidade;
- diminuir sobrecarga informacional;
- reduzir silos de dados;
- melhorar qualidade dos metadados;
- apoiar agentes inteligentes na recomendação de ativos;
- ampliar confiança dos usuários nos produtos analíticos;
- acelerar análises e processos decisórios.

13. Síntese da dimensão

A dimensão **Data Discoverability** trata da capacidade de tornar ativos de dados localizáveis, compreensíveis, avaliáveis e reutilizáveis. No UnespDataLens-RM, essa

dimensão atua como capacidade complementar para assegurar que fontes, datasets, features, dashboards, APIs, indicadores, grafos, modelos e produtos analíticos possam ser encontrados e selecionados com base em finalidade, qualidade, versão, governança, lineage e adequação ao uso.

Ao incorporar catálogos, metadados qualificados, glossários, busca textual e semântica, filtros, recomendações, lineage visível, sinalização de status, métricas de qualidade e agentes inteligentes, o modelo de referência amplia a capacidade de transformar dados disponíveis em ativos efetivamente úteis para análise, pesquisa, gestão e tomada de decisão baseada em evidências.

Dimensão Complementar — Data Minimization

1. Definição da dimensão

A dimensão **Data Minimization**, ou **minimização de dados**, refere-se ao princípio técnico, metodológico e de governança segundo o qual apenas os dados estritamente necessários, adequados, proporcionais e justificados para uma determinada finalidade analítica devem ser coletados, ingeridos, integrados, transformados, armazenados, disponibilizados ou utilizados.

No contexto do **UnespDataLens-RM**, a minimização de dados não significa reduzir indiscriminadamente a quantidade de dados disponíveis, nem eliminar atributos relevantes para análise, modelagem ou tomada de decisão. Trata-se de estabelecer critérios formais para avaliar quais dados são necessários, por que são necessários, para qual finalidade serão utilizados, por quanto tempo devem ser mantidos, quem poderá acessá-los e quais riscos estão associados ao seu uso.

A minimização de dados é especialmente relevante em ecossistemas analíticos que lidam com dados pessoais, sensíveis, institucionais, educacionais, territoriais, administrativos ou sociais. A coleta excessiva de atributos, o armazenamento indefinido de dados brutos, a integração de bases sem finalidade clara, a criação de features com variáveis desnecessárias e a disponibilização ampla de dados detalhados aumentam riscos de exposição, uso indevido, viés, reidentificação, complexidade operacional e dívida de dados.

No UnespDataLens-RM, essa dimensão é complementar e transversal porque se relaciona com todas as etapas do ciclo de vida dos dados: inventário de fontes, ingestão, integração, transformação, qualidade, governança, armazenamento, disponibilização, feature engineering, monitoramento e agentes inteligentes. Ela contribui para assegurar que o modelo de referência produza ativos analíticos úteis e robustos, mas sem ampliar desnecessariamente riscos técnicos, éticos, legais e institucionais.

2. Problema tratado

Em projetos de Engenharia Analítica de Dados, é comum a tendência de coletar e armazenar o maior número possível de atributos, sob a justificativa de que “podem ser úteis no futuro”. Embora essa estratégia pareça vantajosa em curto prazo, ela pode gerar riscos e custos significativos.

A ausência de minimização de dados pode produzir problemas como:

- coleta de atributos sem finalidade analítica definida;
- ingestão de dados pessoais ou sensíveis sem necessidade clara;
- armazenamento excessivo de dados brutos;
- aumento do risco de exposição ou uso indevido;
- maior possibilidade de reidentificação de indivíduos;
- aumento da complexidade de governança;
- dificuldade de controlar permissões de acesso;
- maior volume de dados a monitorar, proteger e auditar;
- criação de features irrelevantes ou redundantes;
- aumento do risco de viés e discriminação algorítmica;

- maior custo de armazenamento e processamento;
- ampliação da dívida de dados;
- dificuldade de justificar o uso de determinados atributos;
- redução da confiança dos usuários e gestores no ecossistema analítico.

No UnespDataLens-RM, a dimensão Data Minimization busca responder a uma questão central: **quais dados são realmente necessários para atender à finalidade analítica, científica, institucional ou decisória proposta?**

Essa questão deve ser tratada antes da ingestão, durante a integração, na criação de atributos derivados, na publicação de produtos e no consumo por usuários e sistemas.

3. Objetivo da dimensão

O objetivo da dimensão **Data Minimization** é **estabelecer critérios para limitar a coleta, ingestão, armazenamento, transformação, disponibilização e uso de dados ao mínimo necessário para uma finalidade analítica explícita, legítima, documentada e proporcional**, reduzindo riscos técnicos, éticos, legais, institucionais e operacionais.

Essa dimensão busca garantir que os ativos analíticos do UnespDataLens-RM sejam suficientes para produzir conhecimento e apoiar decisões, mas sem incorporar dados excessivos, irrelevantes, sensíveis ou desnecessários.

4. Relação com o UnespDataLens-RM

A dimensão Data Minimization se articula com diversos módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, orienta a avaliação da necessidade de cada fonte, atributo, período, granularidade e nível de detalhamento antes de sua incorporação ao modelo de referência.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, contribui para definir quais dados devem ser extraídos, quais devem permanecer fora do ambiente analítico, quais devem ser anonimizados, pseudonimizados ou agregados e quais devem ter retenção limitada.

No **Módulo 3 — Integração e Harmonização**, orienta a integração apenas de entidades, chaves, atributos e relações necessários à finalidade analítica, evitando cruzamentos excessivos ou desproporcionais.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, contribui para remover, mascarar, generalizar ou agregar atributos desnecessários, sensíveis ou de alto risco.

No **Módulo 5 — Qualidade e Validação de Dados**, apoia a validação da adequação dos dados ao uso, considerando não apenas qualidade técnica, mas também pertinência, necessidade e proporcionalidade.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, permite avaliar se determinados atributos aumentam risco de discriminação, inferência indevida ou uso injustificado de características sensíveis.

No **Módulo 7 — Armazenamento Analítico**, orienta políticas de retenção, segregação, agregação, anonimização, granularidade e exclusão de dados não necessários.

No **Módulo 8 — Disponibilização e Consumo Analítico**, define que produtos, dashboards, APIs, relatórios e datasets devem expor apenas o nível de detalhe necessário ao público e à finalidade.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, exige que finalidade, sensibilidade, restrição de uso, justificativa de coleta e necessidade analítica sejam documentadas como metadados.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, a minimização atua como princípio central para reduzir riscos de privacidade, exposição, uso indevido e descumprimento normativo.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, permite rastrear por que determinados dados foram coletados, transformados, agregados, removidos ou disponibilizados.

No **Módulo 12 — Versionamento e Reprodutibilidade**, registra versões de datasets minimizados, regras de exclusão, agregação, anonimização e justificativas de uso.

No **Módulo 14 — Feature Engineering e Feature Store**, orienta a criação de features necessárias, justificadas, não redundantes e compatíveis com a finalidade do modelo ou análise.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, permite identificar ativos com excesso de atributos, dados sem uso, campos sensíveis não utilizados e produtos que precisam ser revisados.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, agentes podem alertar sobre dados excessivos, atributos sensíveis, ausência de justificativa de uso ou necessidade de agregação.

5. Princípios da minimização de dados

A minimização de dados pode ser estruturada por princípios que orientam sua aplicação no modelo de referência.

Princípio	Descrição	Exemplo
Necessidade	Coletar e utilizar apenas dados necessários para a finalidade definida.	Usar faixa etária em vez de data de nascimento quando a idade exata não for necessária.
Adequação	Garantir que os dados sejam compatíveis com o objetivo analítico.	Usar dados acadêmicos para análise de evasão, evitando atributos sem relação clara.

Proporcionalidade	Evitar uso de dados excessivos em relação ao benefício analítico esperado.	Não integrar dados sensíveis se indicadores agregados forem suficientes.
Finalidade explícita	Documentar por que o dado será usado.	Registrar que determinado atributo será usado para avaliar desigualdade de acesso.
Granularidade mínima	Utilizar o menor nível de detalhe necessário.	Trabalhar com município em vez de endereço completo.
Retenção limitada	Manter dados apenas pelo período necessário.	Arquivar ou eliminar dados brutos após produção de dataset curado.
Acesso mínimo	Restringir acesso aos dados conforme função e finalidade.	Usuário de dashboard acessa agregados, não registros individuais.
Redução de sensibilidade	Aplicar anonimização, pseudonimização, agregação ou mascaramento quando possível.	Substituir identificadores diretos por códigos internos.
Revisão contínua	Reavaliar periodicamente a necessidade dos dados armazenados.	Remover atributos que deixaram de ser usados em produtos analíticos.

6. Estratégias de implementação

A implementação da Data Minimization no UnespDataLens-RM pode envolver estratégias técnicas, metodológicas e de governança.

Estratégia	Descrição
Justificativa de coleta por atributo	Registrar a finalidade de cada atributo incorporado ao pipeline.

Classificação de necessidade	Classificar atributos como essencial, útil, opcional, redundante ou desnecessário.
Avaliação de sensibilidade	Identificar dados pessoais, sensíveis, identificadores diretos e atributos com risco de inferência.
Redução de granularidade	Substituir dados detalhados por agregações quando suficiente para a finalidade.
Pseudonimização	Substituir identificadores diretos por identificadores técnicos controlados.
Anonimização ou agregação	Reduzir possibilidade de identificação individual em produtos de consumo.
Política de retenção	Definir prazos de manutenção, arquivamento e descarte de dados.
Controle de acesso por finalidade	Associar permissões ao objetivo de uso, e não apenas ao perfil do usuário.
Revisão de atributos não utilizados	Identificar campos que não são usados em análises, dashboards, features ou modelos.
Minimização em produtos de consumo	Expor apenas atributos e níveis de detalhe necessários ao público-alvo.
Minimização em feature engineering	Evitar features redundantes, irrelevantes, sensíveis ou com risco de vazamento.
Registro de decisão de minimização	Documentar atributos removidos, agregados, mascarados ou mantidos com justificativa.
Monitoramento de excesso	Detectar ativos com muitos atributos não utilizados ou

de dados	sem finalidade documentada.
Validação humana de dados sensíveis	Exigir revisão antes de integrar ou disponibilizar dados sensíveis.

7. Métricas associadas

A dimensão Data Minimization pode ser avaliada por métricas relacionadas à necessidade, sensibilidade, retenção, granularidade, uso efetivo e redução de exposição.

Métrica	Sigla	Descrição
Taxa de atributos com finalidade documentada	TAFD	Proporção de atributos com justificativa de uso registrada.
Taxa de atributos essenciais	TAE	Proporção de atributos classificados como essenciais para a finalidade.
Taxa de atributos sem uso	TASU	Proporção de atributos não utilizados em análises, produtos, features ou modelos.
Taxa de atributos sensíveis	TAS	Proporção de atributos classificados como sensíveis ou de alto risco.
Taxa de atributos sensíveis justificados	TASJ	Proporção de atributos sensíveis com justificativa formal de uso.
Taxa de atributos minimizados	TAM	Proporção de atributos removidos, agregados, mascarados ou pseudonimizados.
Taxa de produtos com granularidade adequada	TPGA	Proporção de produtos que expõem apenas o nível de detalhe necessário.
Taxa de ativos com política	TAPR	Proporção de ativos com prazo de retenção

de retenção		definido.
Taxa de ativos com excesso de atributos	TAEA	Proporção de ativos com atributos sem finalidade, sem uso ou redundantes.
Taxa de acesso por finalidade	TAF	Proporção de acessos associados a uma finalidade explícita.
Índice de minimização de dados	IMD	Indicador composto de finalidade, necessidade, redução de exposição e retenção.

Uma formulação possível para o **Índice de Minimização de Dados** é:

$$\mathbf{IMD} = \mathbf{w_1TAFD} + \mathbf{w_2TAE} + \mathbf{w_3TASJ} + \mathbf{w_4TAM} + \mathbf{w_5TPGA} + \mathbf{w_6TAPR} + \mathbf{w_7TAF} - \mathbf{w_8TASU} - \mathbf{w_9TAEA}$$

em que:

- **TAFD** representa a taxa de atributos com finalidade documentada;
- **TAE** representa a taxa de atributos essenciais;
- **TASJ** representa a taxa de atributos sensíveis justificados;
- **TAM** representa a taxa de atributos minimizados;
- **TPGA** representa a taxa de produtos com granularidade adequada;
- **TAPR** representa a taxa de ativos com política de retenção;
- **TAF** representa a taxa de acesso por finalidade;
- **TASU** representa a taxa de atributos sem uso;
- **TAEA** representa a taxa de ativos com excesso de atributos;
- $\mathbf{w_i}$ representa o peso atribuído a cada dimensão.

Quanto maior o IMD, maior a aderência do ecossistema analítico ao princípio de minimização de dados.

8. Artefatos relacionados

A dimensão Data Minimization pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de inventário, governança, qualidade, armazenamento, feature engineering, disponibilização, monitoramento e agentes inteligentes.

Código	Artefato	Representação	Finalidade
DM1	Matriz de Necessidade	Tabela atributo ×	Justificar a necessidade de cada

	dos Dados	finalidade	atributo em relação à finalidade analítica.
DM2	Registro de Classificação de Sensibilidade	Tabela de classificação	Identificar atributos pessoais, sensíveis, identificadores diretos e atributos de risco.
DM3	Relatório de Atributos sem Uso	Relatório técnico	Listar atributos não utilizados em análises, produtos, features ou modelos.
DM4	Plano de Minimização de Dados	Documento operacional	Definir ações de remoção, agregação, mascaramento, pseudonimização ou retenção limitada.
DM5	Registro de Decisões de Minimização	Tabela de decisões	Documentar atributos mantidos, removidos, agregados ou mascarados e suas justificativas.
DM6	Mapa de Granularidade dos Dados	Tabela ou matriz	Avaliar se o nível de detalhe é adequado à finalidade de uso.
DM7	Política de Retenção e Descarte	Documento normativo	Definir prazos de manutenção, arquivamento, descarte e exceções.
DM8	Relatório de Exposição de Dados	Relatório de risco	Avaliar quantidade de dados sensíveis, identificadores e atributos de alto risco expostos.
DM9	Registro de Pseudonimização e Anonimização	Tabela técnica	Documentar técnicas aplicadas para reduzir identificação direta ou indireta.
DM10	Painel de Minimização	Dashboard	Monitorar atributos sem uso, sensíveis, minimizados, justificados

	de Dados		e ativos com retenção definida.
DM11	Registro de Acessos por Finalidade	Log governado	Relacionar consumo de dados à finalidade declarada e autorizada.
DM12	Relatório de Aderência à Minimização	Relatório técnico-gerencial	Avaliar conformidade do ecossistema com critérios de minimização.

9. Critérios de validação

A dimensão Data Minimization pode ser considerada adequadamente tratada quando:

1. atributos críticos possuem finalidade de uso documentada;
2. dados pessoais e sensíveis são identificados e classificados;
3. atributos sensíveis possuem justificativa explícita de uso;
4. atributos sem uso são identificados e revisados;
5. produtos analíticos expõem apenas o nível de granularidade necessário;
6. datasets de consumo evitam identificadores diretos quando não essenciais;
7. políticas de retenção e descarte estão definidas;
8. decisões de remoção, agregação, mascaramento ou pseudonimização são documentadas;
9. acessos a dados são vinculados a finalidades autorizadas;
10. features sensíveis, redundantes ou sem justificativa são revisadas;
11. produtos de dados de uso amplo priorizam agregação e restrição de granularidade;
12. ativos com excesso de atributos são monitorados;
13. agentes inteligentes ou rotinas de monitoramento identificam dados excessivos ou não justificados;
14. o índice de minimização de dados melhora ao longo do tempo.

10. Exemplo de aplicação no domínio educacional

Em um projeto educacional, a minimização de dados é essencial para evitar o uso excessivo de informações pessoais dos estudantes. Para uma análise de evasão, por exemplo, pode ser necessário utilizar dados de trajetória acadêmica, matrícula, desempenho, frequência e algumas variáveis socioeconômicas. Entretanto, nem sempre é necessário manter identificadores diretos, endereço completo, documentos pessoais ou atributos sensíveis em datasets de consumo analítico.

Nesse contexto, o UnespDataLens-RM pode substituir identificadores diretos por códigos pseudonimizados, utilizar faixa etária em vez de data de nascimento, trabalhar com município ou região em vez de endereço completo, disponibilizar dados agregados em dashboards e documentar a finalidade de cada atributo sensível utilizado.

Dessa forma, a análise educacional preserva sua capacidade explicativa e preditiva, mas reduz riscos de exposição, reidentificação e uso indevido.

11. Exemplo de aplicação em políticas públicas

Em políticas públicas, a minimização de dados é relevante quando diferentes bases de atendimento, saúde, assistência social, educação e território são integradas. Embora a integração possa apoiar decisões intersetoriais, ela também pode aumentar riscos de exposição de indivíduos, famílias ou grupos vulneráveis.

Por exemplo, para construir um indicador territorial de vulnerabilidade, pode não ser necessário disponibilizar registros individuais de cidadãos. Em muitos casos, indicadores agregados por território, faixa etária, tipo de atendimento ou período são suficientes para apoiar a decisão.

A dimensão Data Minimization permite avaliar quais atributos devem ser mantidos, agregados, mascarados ou removidos, documentando a finalidade de uso e reduzindo riscos associados à exposição de dados sensíveis.

12. Contribuição para o modelo de referência

A dimensão **Data Minimization** contribui para o UnespDataLens-RM ao incorporar o princípio de que dados devem ser coletados, tratados e disponibilizados na medida necessária para uma finalidade legítima, explícita e proporcional.

Essa dimensão fortalece o modelo de referência ao:

- reduzir exposição desnecessária de dados;
- diminuir riscos de privacidade e reidentificação;
- apoiar conformidade e governança;
- reduzir complexidade de armazenamento e controle de acesso;
- evitar criação excessiva de atributos e features;
- reduzir data overload e data debt;
- melhorar clareza sobre finalidade de uso;
- apoiar uso responsável de dados sensíveis;
- fortalecer confiança institucional no ecossistema analítico;
- melhorar segurança dos produtos de consumo;
- orientar decisões de retenção, descarte, agregação e anonimização;
- tornar o pipeline mais proporcional, ético e sustentável.

13. Síntese da dimensão

A dimensão **Data Minimization** trata da necessidade de limitar a coleta, integração, transformação, armazenamento, disponibilização e uso de dados ao mínimo necessário para uma finalidade explícita, legítima e documentada. No UnespDataLens-RM, essa dimensão atua como capacidade complementar para reduzir riscos, controlar exposição, evitar excesso de atributos, orientar governança e fortalecer o uso responsável dos dados.

Ao incorporar justificativa de uso por atributo, classificação de sensibilidade, redução de granularidade, anonimização, pseudonimização, retenção limitada, controle de acesso por finalidade e monitoramento de atributos sem uso, o modelo de referência amplia sua capacidade de produzir ativos analíticos úteis, confiáveis, proporcionais e compatíveis com princípios de governança, segurança e responsabilidade.

Dimensão Complementar — Label Quality

1. Definição da dimensão

A dimensão **Label Quality**, ou **qualidade dos rótulos**, refere-se ao grau em que as variáveis-alvo, classes, categorias, desfechos, eventos ou indicadores utilizados em análises computacionais, modelos preditivos, classificações, agrupamentos supervisionados, painéis e processos decisórios representam corretamente o fenômeno que se pretende analisar.

No contexto do **UnespDataLens-RM**, o termo “rótulo” não se limita a classes formais de modelos de aprendizagem de máquina. Ele pode representar qualquer variável que assuma papel de referência analítica, como evasão, permanência, aprovação, reprovação, risco, vulnerabilidade, desempenho, abandono, inadimplência, ocorrência, atendimento, prioridade, sucesso, falha, categoria de evento ou situação institucional.

A qualidade dos rótulos é uma dimensão crítica porque modelos, indicadores e decisões dependem diretamente da forma como esses rótulos são definidos, construídos, registrados, validados e interpretados. Um modelo preditivo pode apresentar boa métrica estatística, mas produzir resultados pouco confiáveis se a variável-alvo estiver mal definida, desatualizada, inconsistente, enviesada, ruidosa ou construída com regras frágeis.

No UnespDataLens-RM, essa dimensão é complementar e transversal porque a qualidade dos rótulos depende de múltiplos módulos do modelo de referência: inventário de fontes, integração, transformação, qualidade dos dados, equidade, metadados, governança, linhagem, versionamento, feature engineering, monitoramento e agentes inteligentes. A Label Quality garante que os desfechos analisados sejam semanticamente claros, tecnicamente consistentes, metodologicamente defensáveis e adequados ao uso.

2. Problema tratado

Em muitos projetos analíticos, grande parte da atenção é dedicada aos atributos preditores, às features, aos modelos e às métricas de desempenho. Entretanto, a variável-alvo ou o rótulo utilizado para treinar, avaliar ou interpretar modelos pode conter problemas relevantes.

Esses problemas podem ocorrer quando a definição do rótulo é ambígua, quando diferentes bases utilizam categorias distintas, quando eventos são registrados tardiamente, quando há erro de classificação, quando regras de negócio mudam ao longo do tempo, quando há inconsistência entre sistemas, quando os rótulos são derivados de registros administrativos incompletos ou quando o desfecho observado não representa adequadamente o fenômeno real.

A baixa qualidade dos rótulos pode gerar problemas como:

- definição ambígua da variável-alvo;
- inconsistência entre categorias ou classes;
- rótulos ausentes, incorretos ou desatualizados;
- rótulos construídos por regras não documentadas;
- divergência entre diferentes fontes sobre o mesmo desfecho;

- mudanças históricas na regra de classificação;
- atraso entre ocorrência do evento e registro no sistema;
- ruído na classe utilizada para treinamento de modelos;
- viés nos rótulos devido a processos institucionais desiguais;
- baixa comparabilidade entre períodos;
- dificuldade de reproduzir resultados;
- modelos preditivos treinados sobre desfechos mal definidos;
- indicadores baseados em categorias inconsistentes;
- decisões apoiadas em classificações pouco confiáveis.

No UnespDataLens-RM, a dimensão Label Quality busca tornar explícito que a qualidade do rótulo é tão importante quanto a qualidade dos atributos explicativos. Se o rótulo estiver incorreto ou mal especificado, todo o processo analítico subsequente pode ser comprometido.

3. Objetivo da dimensão

O objetivo da dimensão **Label Quality** é **avaliar, documentar, validar, monitorar e aprimorar a qualidade dos rótulos, variáveis-alvo, classes, categorias e desfechos utilizados em análises, modelos, indicadores e produtos analíticos**.

Essa dimensão busca garantir que os rótulos sejam semanticamente bem definidos, metodologicamente justificados, tecnicamente consistentes, temporalmente válidos, rastreáveis, versionados e adequados à finalidade analítica.

4. Relação com o UnespDataLens-RM

A dimensão Label Quality se articula com diversos módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, contribui para identificar quais fontes registram ou permitem derivar rótulos, quais categorias são usadas, quais responsáveis mantêm esses registros e quais limitações existem.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, permite preservar o estado original dos rótulos, suas categorias, códigos, datas de registro e possíveis inconsistências.

No **Módulo 3 — Integração e Harmonização**, é essencial para resolver divergências entre fontes que classificam o mesmo fenômeno de formas distintas.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, orienta a construção de rótulos derivados, recodificação de categorias, tratamento de classes ausentes e padronização de desfechos.

No **Módulo 5 — Qualidade e Validação de Dados**, estabelece regras específicas para avaliar completude, validade, consistência, atualidade e unicidade dos rótulos.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, permite avaliar se os rótulos possuem distribuição desigual, ruído diferencial entre grupos ou viés institucional.

No **Módulo 8 — Disponibilização e Consumo Analítico**, garante que indicadores, dashboards e produtos que utilizam rótulos explicitem definição, regra de cálculo, período de referência e limitações.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, documenta o significado dos rótulos, suas categorias, regras de derivação, termos equivalentes, mudanças históricas e relação com conceitos do domínio.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, avalia implicações éticas, institucionais e de sensibilidade no uso de rótulos que classificam indivíduos, grupos ou territórios.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, registra como o rótulo foi construído, quais fontes foram usadas, quais transformações foram aplicadas e quais produtos dependem dele.

No **Módulo 12 — Versionamento e Reprodutibilidade**, preserva versões de regras de rotulagem, categorias, janelas temporais e critérios de classificação.

No **Módulo 14 — Feature Engineering e Feature Store**, evita vazamento de informação entre features e rótulos, além de documentar a relação entre atributos preditores e variável-alvo.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, acompanha mudanças na distribuição dos rótulos, surgimento de novas categorias, alterações de regras e degradação da estabilidade temporal.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, agentes podem explicar definições de rótulos, identificar inconsistências, alertar sobre mudanças e apoiar documentação das variáveis-alvo.

5. Dimensões da qualidade dos rótulos

A qualidade dos rótulos pode ser avaliada por diferentes dimensões complementares.

Dimensão	Descrição	Exemplo
Definição semântica	Clareza conceitual sobre o que o rótulo representa.	Definir formalmente o que é evasão.
Completude	Grau em que os registros possuem rótulo preenchido.	Todos os estudantes possuem situação final registrada.
Correção	Grau em que o rótulo corresponde ao estado real do fenômeno.	Estudante marcado como evadido de fato não realizou rematrícula.
Consistência	Ausência de contradições entre fontes ou registros.	A mesma matrícula não aparece como ativa e cancelada no

		mesmo período.
Atualidade	Grau em que o rótulo reflete a situação mais recente ou o período correto.	Situação acadêmica atualizada após encerramento do semestre.
Estabilidade temporal	Capacidade de manter significado comparável ao longo do tempo.	Regra de evasão preservada entre anos ou versionada.
Granularidade adequada	Nível de detalhe compatível com a análise.	Diferenciar abandono, cancelamento e trancamento quando necessário.
Rastreabilidade	Capacidade de reconstruir como o rótulo foi definido ou derivado.	Rótulo calculado a partir de matrícula, rematrícula e status final.
Balanceamento	Distribuição adequada entre classes ou categorias.	Classes evadido e não evadido avaliadas quanto à proporção.
Equidade	Ausência de ruído ou erro sistemático em grupos específicos.	Rótulo de risco não super-representa determinado grupo por falha de registro.
Reprodutibilidade	Capacidade de reconstruir o rótulo usando as mesmas regras e versões.	Rótulo de evasão reproduzido a partir do manifesto de execução.

Essas dimensões permitem avaliar se um rótulo é adequado para sustentar análises, indicadores, modelos ou decisões.

6. Estratégias de melhoria da qualidade dos rótulos

A melhoria da Label Quality no UnespDataLens-RM pode envolver estratégias técnicas, metodológicas, semânticas e de governança.

Estratégia	Descrição

Definição formal do rótulo	Documentar o conceito, escopo, regra e finalidade da variável-alvo.
Dicionário de categorias	Registrar classes, códigos, significados, equivalências e limitações.
Validação cruzada entre fontes	Comparar rótulos registrados em diferentes bases ou sistemas.
Regras de derivação documentadas	Explicitar como rótulos são calculados a partir de outros atributos.
Versionamento das regras de rotulagem	Registrar mudanças em critérios, categorias e janelas temporais.
Análise de ruído de rótulo	Identificar inconsistências, classes improváveis ou registros suspeitos.
Validação temporal	Verificar se o rótulo respeita o período de observação e evita uso de informação futura.
Avaliação de balanceamento	Medir proporção entre classes e identificar classes raras ou sub-representadas.
Avaliação de equidade dos rótulos	Verificar se erros ou ausências de rótulos afetam grupos de forma desigual.
Separação entre rótulo e features	Evitar vazamento de informação entre variável-alvo e atributos preditores.
Curadoria humana	Revisar rótulos críticos ou ambíguos com apoio de

	especialistas do domínio.
Monitoramento de label drift	Acompanhar mudanças na distribuição dos rótulos ao longo do tempo.
Documentação de limitações	Explicitar incertezas, atrasos, classes agregadas e restrições de interpretação.
Auditoria de rótulos críticos	Avaliar rótulos utilizados em modelos ou decisões de alto impacto.

7. Métricas associadas

A dimensão Label Quality pode ser avaliada por métricas relacionadas à completude, consistência, ruído, estabilidade, balanceamento, rastreabilidade e equidade dos rótulos.

Métrica	Sigla	Descrição
Taxa de rótulos documentados	TRD	Proporção de rótulos com definição, categorias e regra de construção documentadas.
Taxa de completude de rótulos	TCR	Proporção de registros com rótulo preenchido.
Taxa de rótulos inconsistentes	TRI	Proporção de registros com rótulos contraditórios ou incompatíveis.
Taxa de rótulos ausentes	TRA	Proporção de registros sem rótulo quando ele deveria existir.
Taxa de rótulos corrigidos	TRC	Proporção de rótulos revisados ou corrigidos após validação.
Taxa de regras de rotulagem	TRRV	Proporção de regras de construção de rótulos com

versionadas		controle de versão.
Taxa de rótulos rastreáveis	TRR	Proporção de rótulos cuja origem e regra de derivação podem ser reconstruídas.
Índice de ruído de rótulos	IRR	Estimativa de registros com rótulo potencialmente incorreto.
Índice de balanceamento de classes	IBC	Grau de equilíbrio entre categorias do rótulo.
Taxa de label drift	TLD	Proporção de variação relevante na distribuição dos rótulos ao longo do tempo.
Taxa de rótulos com avaliação de equidade	TRAE	Proporção de rótulos avaliados quanto a efeitos diferenciais entre grupos.
Índice de qualidade de rótulos	IQR	Indicador composto da qualidade global dos rótulos.

Uma formulação possível para o **Índice de Qualidade de Rótulos** é:

$$\mathbf{IQR} = \mathbf{w_1TRD} + \mathbf{w_2TCR} + \mathbf{w_3TRRV} + \mathbf{w_4TRR} + \mathbf{w_5IBC} + \mathbf{w_6TRAE} - \mathbf{w_7TRI} - \mathbf{w_8TRA} - \mathbf{w_9IRR} - \mathbf{w_{10}TLD}$$

em que:

- **TRD** representa a taxa de rótulos documentados;
- **TCR** representa a taxa de completude de rótulos;
- **TRRV** representa a taxa de regras de rotulagem versionadas;
- **TRR** representa a taxa de rótulos rastreáveis;
- **IBC** representa o índice de balanceamento de classes;
- **TRAE** representa a taxa de rótulos com avaliação de equidade;
- **TRI** representa a taxa de rótulos inconsistentes;
- **TRA** representa a taxa de rótulos ausentes;
- **IRR** representa o índice de ruído de rótulos;
- **TLD** representa a taxa de label drift;
- $\mathbf{w_i}$ representa o peso atribuído a cada dimensão.

Quanto maior o IQR, maior a qualidade dos rótulos utilizados no ecossistema analítico.

8. Artefatos relacionados

A dimensão Label Quality pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de qualidade, metadados, governança, lineage, versionamento, feature engineering, monitoramento e agentes inteligentes.

Código	Artefato	Representação	Finalidade
LQ1	Catálogo de Rótulos	Catálogo ou tabela de variáveis-alvo	Registrar rótulos, desfechos, classes, categorias, responsáveis e finalidades.
LQ2	Dicionário de Categorias do Rótulo	Tabela semântica	Definir categorias, códigos, significados, equivalências e limitações.
LQ3	Especificação da Regra de Rotulagem	Documento técnico ou regra computacional	Formalizar como o rótulo é atribuído, calculado ou derivado.
LQ4	Registro de Versionamento de Rótulos	Changelog ou tabela versionada	Registrar mudanças em regras, categorias, janelas temporais e critérios de classificação.
LQ5	Relatório de Completude de Rótulos	Relatório técnico	Avaliar registros com rótulos preenchidos, ausentes ou indefinidos.
LQ6	Relatório de Consistência de Rótulos	Relatório técnico	Identificar contradições entre fontes, períodos, categorias ou regras.
LQ7	Relatório de Ruído de Rótulos	Relatório analítico	Identificar registros potencialmente mal rotulados, ambíguos ou suspeitos.
LQ8	Relatório de Balanceamento de	Relatório estatístico	Avaliar distribuição das categorias do rótulo e classes

	Classes		raras.
LQ9	Registro de Label Drift	Série temporal ou relatório	Monitorar mudanças na distribuição dos rótulos ao longo do tempo.
LQ10	Relatório de Equidade dos Rótulos	Relatório técnico	Avaliar se ausências, erros ou distribuições de rótulos afetam grupos de forma desigual.
LQ11	Manifesto de Construção do Rótulo	Documento reprodutível	Registrar fontes, regras, versões, parâmetros e evidências usados na construção do rótulo.
LQ12	Plano de Curadoria de Rótulos	Documento operacional	Definir ações de revisão, correção, validação e manutenção dos rótulos.

9. Critérios de validação

A dimensão Label Quality pode ser considerada adequadamente tratada quando:

1. os rótulos críticos possuem definição formal;
2. categorias e códigos dos rótulos estão documentados;
3. regras de construção, derivação ou classificação são explícitas;
4. rótulos possuem responsável técnico ou institucional;
5. rótulos possuem completude mínima aceitável;
6. inconsistências entre fontes ou períodos são identificadas;
7. regras de rotulagem são versionadas;
8. a origem dos rótulos pode ser rastreada;
9. mudanças na definição dos rótulos são registradas e comunicadas;
10. rótulos utilizados em modelos são avaliados quanto a ruído, balanceamento e estabilidade;
11. rótulos sensíveis ou de alto impacto passam por avaliação de equidade;
12. janelas temporais de observação e predição são documentadas;
13. riscos de vazamento de informação entre rótulo e features são avaliados;
14. agentes ou rotinas de monitoramento conseguem identificar label drift ou inconsistências.

10. Exemplo de aplicação no domínio educacional

Em um projeto educacional, um rótulo frequentemente utilizado é “evasão”. Entretanto, evasão pode ser definida de diferentes formas: cancelamento formal, ausência de

rematrícula, abandono, trancamento prolongado, desligamento, transferência ou não conclusão em determinado prazo.

Se a definição de evasão não for formalizada, dois dashboards ou modelos podem apresentar resultados diferentes mesmo utilizando bases semelhantes. Além disso, mudanças administrativas na classificação da situação acadêmica podem alterar o significado do rótulo ao longo do tempo.

A dimensão Label Quality permite documentar a definição adotada, registrar categorias incluídas ou excluídas, definir a janela temporal de observação, versionar a regra de rotulagem, avaliar completude e consistência, monitorar mudanças na distribuição do rótulo e indicar limitações interpretativas.

11. Exemplo de aplicação em políticas públicas

Em políticas públicas, um rótulo pode representar vulnerabilidade, risco social, atendimento concluído, reincidência, prioridade, situação de família, ocorrência, adesão a programa ou sucesso de uma intervenção.

Por exemplo, uma classificação de “família em situação de vulnerabilidade” pode ser construída a partir de critérios socioeconômicos, territoriais, composição familiar, acesso a serviços e registros administrativos. Se esses critérios não forem claros ou se variarem entre secretarias, o rótulo pode apresentar inconsistência e produzir decisões injustas ou pouco confiáveis.

A dimensão Label Quality permite formalizar a regra de classificação, documentar fontes, avaliar completude dos registros, verificar estabilidade temporal, analisar possíveis vieses e garantir que a classificação seja interpretada corretamente nos produtos analíticos.

12. Contribuição para o modelo de referência

A dimensão **Label Quality** contribui para o UnespDataLens-RM ao reconhecer que a qualidade de análises, modelos e decisões depende diretamente da qualidade dos rótulos utilizados como referência.

Essa dimensão fortalece o modelo de referência ao:

- melhorar definição e documentação de variáveis-alvo;
- reduzir inconsistências entre fontes e produtos;
- apoiar construção confiável de modelos preditivos;
- reduzir ruído e ambiguidade nos desfechos;
- evitar vazamento de informação entre features e rótulos;
- permitir avaliação de balanceamento e label drift;
- fortalecer reprodutibilidade de indicadores e modelos;
- apoiar análise de equidade e justiça algorítmica;
- reduzir dependência de regras tácitas;
- melhorar interpretação de dashboards e produtos analíticos;
- aumentar confiança na tomada de decisão baseada em evidências.

13. Síntese da dimensão

A dimensão **Label Quality** trata da qualidade dos rótulos, variáveis-alvo, classes, categorias e desfechos utilizados em análises, modelos, indicadores e produtos analíticos. No UnespDataLens-RM, essa dimensão atua como capacidade complementar para garantir que os rótulos sejam semanticamente definidos, tecnicamente consistentes, metodologicamente justificados, rastreáveis, versionados, monitoráveis e adequados ao uso.

Ao incorporar catálogos de rótulos, dicionários de categorias, regras de rotulagem, versionamento, análise de completude, consistência, ruído, balanceamento, label drift e equidade, o modelo de referência fortalece a confiabilidade dos processos analíticos e reduz o risco de decisões baseadas em desfechos mal definidos ou inconsistentes.

Dimensão Complementar — Data Uncertainty

1. Definição da dimensão

A dimensão **Data Uncertainty**, ou **incerteza nos dados**, refere-se ao grau de dúvida, imprecisão, incompletude, ambiguidade, variabilidade, instabilidade ou limitação associado aos dados, atributos, rótulos, indicadores, features, modelos, representações e produtos analíticos. Essa incerteza pode estar relacionada à origem dos dados, à forma de coleta, ao processo de integração, à qualidade dos registros, à temporalidade, à definição semântica dos conceitos, às regras de transformação, à construção de variáveis, à representatividade da amostra ou à interpretação dos resultados.

No contexto do **UnespDataLens-RM**, a incerteza não deve ser tratada apenas como erro. Ela representa uma característica inerente a muitos ecossistemas de dados, especialmente quando os dados são heterogêneos, administrativos, históricos, integrados de múltiplas fontes, produzidos por processos humanos, sujeitos a mudanças de schema, coletados em diferentes granularidades ou utilizados para apoiar decisões complexas.

A incerteza pode estar presente quando um dado está ausente, quando um registro é inconsistente, quando uma categoria é ambígua, quando a origem de um atributo não é clara, quando uma regra de transformação possui exceções, quando uma variável-alvo é construída por aproximação, quando uma fonte não cobre todo o fenômeno estudado, quando um indicador é calculado com dados defasados, ou quando um modelo produz predições com diferentes níveis de confiança.

No UnespDataLens-RM, essa dimensão é complementar e transversal porque a incerteza pode surgir em todos os módulos do modelo de referência: no inventário de fontes, na ingestão, na integração, na transformação, na validação, na equidade, no armazenamento, no consumo, nos metadados, na governança, na linhagem, no versionamento, nas representações analíticas, na feature store, no monitoramento e nos agentes inteligentes.

2. Problema tratado

Em muitos projetos analíticos, os dados são tratados como se fossem exatos, completos, estáveis e semanticamente inequívocos. Entretanto, na prática, dados institucionais, educacionais, administrativos e de políticas públicas frequentemente carregam diferentes formas de incerteza.

Essa incerteza pode surgir por múltiplas razões: campos não preenchidos, registros duplicados, diferenças entre sistemas, atrasos de atualização, erros de digitação, mudanças de categorias, integração parcial entre bases, ausência de metadados, baixa qualidade dos rótulos, viés de cobertura, dados agregados, amostras incompletas, regras de negócio implícitas, transformações manuais ou limitações de acesso.

A ausência de tratamento explícito da incerteza pode gerar problemas como:

- interpretação excessivamente confiante de dados incompletos;
- uso de indicadores sem explicitação de limitações;
- modelos preditivos apresentados sem estimativa de confiança;
- decisões baseadas em dados desatualizados ou parcialmente observados;
- integração de fontes com correspondências incertas;
- rótulos construídos por aproximação sem sinalização;
- dashboards que ocultam margens de erro, defasagens ou lacunas;
- dificuldade de comparar resultados entre períodos ou fontes;
- ausência de registro sobre qualidade e confiabilidade dos atributos;
- subestimação de riscos analíticos;
- conclusões indevidas a partir de dados frágeis;
- baixa transparência sobre limitações metodológicas;
- perda de confiança quando inconsistências são identificadas posteriormente.

No UnespDataLens-RM, a dimensão Data Uncertainty busca tornar a incerteza visível, documentada, mensurável e comunicável, permitindo que usuários, gestores, pesquisadores e sistemas compreendam não apenas “o que os dados dizem”, mas também **com que grau de confiança, limitação e contexto essa informação pode ser utilizada**.

3. Objetivo da dimensão

O objetivo da dimensão **Data Uncertainty** é **identificar, classificar, mensurar, documentar, comunicar e monitorar incertezas associadas aos dados, atributos, rótulos, features, indicadores, modelos e produtos analíticos**, garantindo que limitações, ambiguidades, lacunas, estimativas e níveis de confiança sejam explicitados no ciclo de vida dos dados.

Essa dimensão busca fortalecer a interpretação responsável dos resultados produzidos pelo UnespDataLens-RM, evitando que produtos analíticos sejam apresentados como verdades absolutas quando dependem de dados incompletos, incertos, aproximados ou sujeitos a restrições metodológicas.

4. Relação com o UnespDataLens-RM

A dimensão Data Uncertainty se articula com diversos módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, contribui para registrar limitações conhecidas das fontes, cobertura temporal, cobertura populacional, granularidade, periodicidade, confiabilidade, responsável e riscos de incompletude.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, permite registrar falhas de extração, perdas de registros, atrasos, dados parciais, inconsistências no arquivo original e incertezas associadas à origem dos dados.

No **Módulo 3 — Integração e Harmonização**, atua na identificação de incertezas em chaves de integração, correspondências entre entidades, conflitos entre fontes, diferenças de granularidade e equivalências semânticas.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, permite documentar incertezas geradas por imputação, recodificação, agregação, normalização, remoção de outliers, criação de variáveis derivadas e tratamento de valores ausentes.

No **Módulo 5 — Qualidade e Validação de Dados**, complementa as métricas de qualidade ao registrar não apenas se há erro ou ausência, mas o grau de confiabilidade, severidade e impacto das não conformidades.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, apoia a identificação de incertezas relacionadas à representatividade dos grupos, dados ausentes diferenciais, rótulos enviesados e estimativas instáveis em grupos minoritários.

No **Módulo 7 — Armazenamento Analítico**, orienta a preservação de metadados de incerteza associados aos datasets, versões, zonas, data marts e produtos derivados.

No **Módulo 8 — Disponibilização e Consumo Analítico**, garante que dashboards, APIs, relatórios e produtos de dados comuniquem limitações, defasagens, níveis de confiança, qualidade e restrições interpretativas.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, a incerteza deve ser documentada como metadado, incluindo limitações de fonte, confiabilidade, nível de evidência, completude, ambiguidade semântica e adequação ao uso.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, a incerteza apoia decisões sobre uso adequado, restrições, riscos, finalidade e responsabilidade na interpretação dos dados.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, permite identificar onde a incerteza foi introduzida, propagada ou reduzida ao longo do pipeline.

No **Módulo 12 — Versionamento e Reprodutibilidade**, registra versões de dados, regras, parâmetros e decisões que influenciam níveis de incerteza e possibilidade de reconstrução dos resultados.

No **Módulo 13 — Representações Analíticas**, a incerteza pode ser associada a tabelas, séries temporais, grafos, grafos temporais, relações inferidas, eventos e janelas de observação.

No **Módulo 14 — Feature Engineering e Feature Store**, permite qualificar features por estabilidade, confiabilidade, ausência, imputação, variabilidade e risco de erro.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, acompanha mudanças na incerteza ao longo do tempo, como aumento de valores ausentes, queda de confiabilidade ou instabilidade de indicadores.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, agentes podem explicar níveis de confiança, sinalizar limitações, identificar lacunas e evitar respostas assertivas quando há incerteza relevante.

5. Tipos de incerteza nos dados

A incerteza nos dados pode assumir diferentes formas, que devem ser identificadas e documentadas no UnespDataLens-RM.

Tipo de incerteza	Descrição	Exemplo
Incerteza de origem	Dúvida sobre procedência, responsável, processo de coleta ou confiabilidade da fonte.	Base recebida sem documentação da origem.
Incerteza de completude	Ausência parcial de registros, atributos ou períodos.	Dados de frequência incompletos em determinado semestre.
Incerteza de medição	Imprecisão no valor registrado ou na forma de mensuração.	Tempo de acesso registrado com baixa precisão.
Incerteza semântica	Ambiguidade no significado de termos, categorias ou atributos.	O termo “abandono” possui significados diferentes entre sistemas.
Incerteza temporal	Dúvida sobre data de ocorrência, atualização ou validade do dado.	Situação acadêmica atualizada com atraso.
Incerteza de integração	Dúvida sobre correspondência entre registros de diferentes fontes.	Estudantes pareados por nome e data de nascimento.
Incerteza de transformação	Incerteza introduzida por limpeza, imputação, agregação ou recodificação.	Valor ausente substituído por média do grupo.
Incerteza de rótulo	Dúvida sobre a correção ou definição da variável-alvo.	Rótulo de evasão derivado por ausência de rematrícula.
Incerteza	Limitação decorrente de amostra	Dados disponíveis apenas de cursos que usam determinado

amostral	parcial ou não representativa.	sistema.
Incerteza preditiva	Grau de confiança associado a modelos, scores ou classificações.	Predição de risco com baixa probabilidade de confiança.
Incerteza relacional	Dúvida sobre a existência, força ou direção de relações em grafos.	Relação entre estudante e atividade inferida por logs incompletos.
Incerteza de uso	Dúvida sobre adequação do dado a determinada finalidade.	Dataset criado para gestão usado em modelagem preditiva sem validação.

Nem toda incerteza invalida o uso dos dados. O aspecto central é que a incerteza seja identificada, documentada, mensurada quando possível e comunicada adequadamente aos consumidores.

6. Estratégias de tratamento da incerteza

O tratamento da Data Uncertainty no UnespDataLens-RM pode envolver estratégias técnicas, metodológicas, estatísticas, semânticas e de governança.

Estratégia	Descrição
Registro explícito de limitações	Documentar lacunas, restrições, cobertura e condições de uso dos dados.
Metadados de confiabilidade	Associar ativos a níveis de confiança, qualidade, cobertura e completude.
Sinalização de dados estimados	Marcar valores imputados, inferidos, agregados ou aproximados.
Validação cruzada entre fontes	Comparar registros ou indicadores provenientes de fontes diferentes.

Controle de janelas temporais	Documentar períodos de observação, atualização e validade dos dados.
Análise de sensibilidade	Avaliar como mudanças nos dados, parâmetros ou regras afetam resultados.
Intervalos de confiança	Comunicar incerteza estatística em estimativas, indicadores e modelos.
Classificação de severidade da incerteza	Avaliar se a incerteza é baixa, média, alta ou crítica para o uso.
Propagação de incerteza	Registrar como incertezas de entrada afetam datasets, features e produtos derivados.
Curadoria humana	Revisar casos ambíguos, rótulos críticos e registros de alto impacto.
Monitoramento de incerteza	Acompanhar evolução de ausências, instabilidade, ruído e confiabilidade.
Comunicação visual da incerteza	Exibir alertas, faixas, notas, selos de qualidade ou indicadores de confiança em dashboards.
Uso de agentes inteligentes com cautela	Orientar agentes a sinalizar ausência de evidência, baixa confiança ou limitações.

7. Métricas associadas

A dimensão Data Uncertainty pode ser avaliada por métricas relacionadas à completude, confiabilidade, ambiguidade, estabilidade, estimativa, propagação e comunicação da incerteza.

Métrica	Sigla	Descrição

Taxa de ativos com incerteza documentada	TAID	Proporção de ativos que registram limitações, incertezas ou restrições interpretativas.
Taxa de atributos com nível de confiança	TANC	Proporção de atributos com avaliação de confiabilidade ou qualidade associada.
Taxa de valores ausentes	TVA	Proporção de valores não preenchidos em atributos críticos.
Taxa de valores imputados	TVI	Proporção de valores estimados ou imputados.
Taxa de registros com origem incerta	TROI	Proporção de registros sem origem suficientemente documentada.
Taxa de correspondências incertas	TCI	Proporção de integrações realizadas com baixa confiança de pareamento.
Taxa de rótulos incertos	TRI	Proporção de rótulos derivados, ambíguos ou com baixa confiabilidade.
Taxa de indicadores com intervalo de confiança	TIIC	Proporção de indicadores acompanhados de medida de incerteza.
Taxa de produtos com limitação explícita	TPLE	Proporção de dashboards, APIs, relatórios ou datasets com limitações visíveis.
Índice de estabilidade dos atributos	IEA	Grau de estabilidade dos atributos ao longo do tempo.
Índice de incerteza dos dados	IID	Indicador composto do nível de incerteza associado aos ativos analíticos.

Uma formulação possível para o **Índice de Incerteza dos Dados** é:

$$\mathbf{IID} = w_1\mathbf{TVA} + w_2\mathbf{TVI} + w_3\mathbf{TROI} + w_4\mathbf{TCI} + w_5\mathbf{TRI} + w_6(1 - \mathbf{IEA}) - w_7\mathbf{TAID} - w_8\mathbf{TPLE}$$

em que:

- **TVA** representa a taxa de valores ausentes;
- **TVI** representa a taxa de valores imputados;
- **TROI** representa a taxa de registros com origem incerta;
- **TCI** representa a taxa de correspondências incertas;
- **TRI** representa a taxa de rótulos incertos;
- **IEA** representa o índice de estabilidade dos atributos;
- **TAID** representa a taxa de ativos com incerteza documentada;
- **TPLE** representa a taxa de produtos com limitação explícita;
- w_i representa o peso atribuído a cada dimensão.

Quanto maior o IID, maior o nível de incerteza associado ao ecossistema analítico. Entretanto, a interpretação do índice deve considerar o domínio, a finalidade e a criticidade do uso.

8. Artefatos relacionados

A dimensão Data Uncertainty pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de qualidade, metadados, lineage, versionamento, feature engineering, monitoramento, governança e agentes inteligentes.

Código	Artefato	Representação	Finalidade
DU1	Registro de Incertezas dos Dados	Tabela de incertezas	Documentar fontes, atributos, registros, rótulos ou produtos com incertezas relevantes.
DU2	Matriz de Tipos de Incerteza	Matriz classificatória	Classificar incertezas por origem, completude, semântica, temporalidade, integração, rótulo ou predição.
DU3	Relatório de Completude e Lacunas	Relatório técnico	Identificar ausências, períodos não cobertos, atributos incompletos e registros parciais.
DU4	Registro de Valores Imputados ou Estimados	Tabela técnica	Marcar valores inferidos, imputados, aproximados ou agregados.
DU5	Relatório de Confiabilidade de	Relatório técnico-gerencial	Avaliar fontes quanto a origem, atualização, estabilidade,

	Fontes		documentação e qualidade.
DU6	Registro de Correspondências Incertas	Tabela de pareamento	Documentar integrações entre entidades com grau de confiança reduzido.
DU7	Relatório de Propagação de Incerteza	Relatório analítico	Avaliar como incertezas de entrada afetam datasets, features, indicadores e produtos derivados.
DU8	Registro de Limitações Interpretativas	Documento ou ficha de ativo	Explicitar restrições de uso, escopo, validade e cuidados na interpretação.
DU9	Painel de Incerteza dos Dados	Dashboard	Monitorar ausências, imputações, origem incerta, rótulos incertos e estabilidade dos ativos.
DU10	Manifesto de Incerteza do Produto Analítico	Documento associado ao produto	Registrar incertezas, limitações, níveis de confiança e restrições de uso de dashboards, APIs ou relatórios.
DU11	Relatório de Sensibilidade Analítica	Relatório estatístico	Avaliar como variações em dados, regras ou parâmetros alteram resultados.
DU12	Plano de Mitigação da Incerteza	Documento operacional	Definir ações para reduzir, comunicar ou controlar incertezas críticas.

9. Critérios de validação

A dimensão Data Uncertainty pode ser considerada adequadamente tratada quando:

1. ativos críticos possuem limitações e incertezas documentadas;
2. fontes possuem avaliação de confiabilidade, cobertura e atualização;
3. atributos críticos possuem avaliação de completude e estabilidade;
4. valores imputados, estimados ou inferidos são sinalizados;

5. integrações incertas possuem grau de confiança registrado;
6. rótulos ambíguos ou derivados possuem regra e limitação documentadas;
7. indicadores relevantes apresentam informação sobre qualidade, cobertura ou confiança;
8. produtos analíticos explicitam limitações interpretativas;
9. incertezas introduzidas em etapas do pipeline podem ser rastreadas por lineage;
10. incertezas críticas geram alertas, revisão humana ou restrição de uso;
11. agentes inteligentes sinalizam baixa confiança ou ausência de evidência;
12. análises sensíveis consideram variação de parâmetros, regras ou dados;
13. incertezas são monitoradas ao longo do tempo;
14. usuários conseguem distinguir dados observados, imputados, inferidos e estimados.

10. Exemplo de aplicação no domínio educacional

Em um projeto educacional, a incerteza pode ocorrer quando dados de frequência são incompletos, quando registros de interação no ambiente virtual dependem da configuração da plataforma, quando a situação acadêmica é atualizada com atraso, quando o rótulo de evasão é derivado da ausência de rematrícula ou quando diferentes sistemas registram o mesmo evento de forma distinta.

Por exemplo, se um estudante não aparece rematriculado no período seguinte, ele pode ser classificado como evadido. No entanto, essa classificação pode ser incerta se houver atraso de atualização, transferência não registrada, trancamento ainda não processado ou mudança administrativa. A dimensão Data Uncertainty permite registrar essa incerteza, diferenciar rótulos confirmados de rótulos prováveis e comunicar a limitação nos dashboards e modelos.

Outro exemplo ocorre quando interações em ambiente virtual são usadas para medir engajamento. Se parte das disciplinas não utiliza o ambiente virtual de forma consistente, a ausência de interação não significa necessariamente ausência de engajamento. Essa limitação deve ser documentada e considerada na interpretação dos resultados.

11. Exemplo de aplicação em políticas públicas

Em políticas públicas, a incerteza pode ocorrer quando bases de diferentes secretarias possuem cobertura desigual, quando registros são atualizados em prazos distintos, quando indicadores são calculados com dados parciais, quando territórios possuem subnotificação ou quando categorias de atendimento não são padronizadas.

Por exemplo, um indicador de vulnerabilidade territorial pode ser calculado a partir de dados de atendimento social, saúde, educação e renda. Se uma das fontes apresenta defasagem ou subregistro, o indicador final deve refletir essa limitação, evitando que o território seja interpretado como menos vulnerável apenas por ausência de dados.

A dimensão Data Uncertainty permite registrar fontes com menor confiabilidade, sinalizar indicadores com dados parciais, comunicar limitações e apoiar decisões com maior transparência sobre o grau de confiança das evidências.

12. Contribuição para o modelo de referência

A dimensão **Data Uncertainty** contribui para o UnespDataLens-RM ao reconhecer que dados e produtos analíticos não são igualmente confiáveis em todas as situações e que incertezas precisam ser explicitadas para apoiar decisões responsáveis.

Essa dimensão fortalece o modelo de referência ao:

- tornar limitações dos dados visíveis e documentadas;
- evitar interpretações excessivamente assertivas;
- apoiar comunicação responsável de indicadores;
- melhorar avaliação de qualidade e confiabilidade;
- fortalecer rastreabilidade de decisões analíticas;
- reduzir riscos de conclusões indevidas;
- apoiar análise de sensibilidade;
- qualificar rótulos, features e modelos;
- melhorar governança do uso de dados incertos;
- orientar agentes inteligentes a responder com cautela;
- aumentar transparência sobre lacunas e restrições;
- fortalecer confiança na tomada de decisão baseada em evidências.

13. Síntese da dimensão

A dimensão **Data Uncertainty** trata da identificação, documentação, mensuração, comunicação e monitoramento das incertezas associadas aos dados, atributos, rótulos, features, indicadores, modelos e produtos analíticos. No UnespDataLens-RM, essa dimensão atua como capacidade complementar para garantir que limitações, lacunas, ambiguidades, estimativas, imputações, instabilidades e níveis de confiança sejam considerados ao longo do ciclo de vida dos dados.

Ao incorporar registros de incerteza, metadados de confiabilidade, sinalização de valores imputados, análise de completude, propagação de incerteza, limitações interpretativas, análise de sensibilidade e comunicação explícita nos produtos analíticos, o modelo de referência fortalece sua capacidade de produzir evidências úteis, transparentes e metodologicamente responsáveis.

Dimensão Complementar — Interoperabilidade Semântica

1. Definição da dimensão

A dimensão **Interoperabilidade Semântica** refere-se à capacidade de diferentes fontes, sistemas, bases, atributos, indicadores, categorias, produtos analíticos e áreas institucionais compartilharem, integrarem e interpretarem dados de forma coerente, preservando o significado dos conceitos utilizados.

No contexto do **UnespDataLens-RM**, a interoperabilidade semântica não se limita à compatibilidade técnica entre formatos, tabelas, arquivos ou APIs. Ela trata da compatibilidade de significado entre dados produzidos por diferentes sistemas, áreas, projetos ou domínios. Dois sistemas podem utilizar o mesmo termo com significados distintos, ou termos diferentes para representar o mesmo conceito. Também podem usar categorias, códigos, regras, granularidades e períodos diferentes para descrever fenômenos semelhantes.

Assim, a interoperabilidade semântica busca responder a perguntas como: o que este atributo significa? este conceito tem o mesmo significado em outra fonte? duas categorias podem ser comparadas? este indicador é calculado com a mesma regra em diferentes produtos? esta variável representa o mesmo fenômeno ao longo do tempo? este termo possui sinônimos, equivalências ou especializações? quais conceitos precisam ser harmonizados antes da integração?

No UnespDataLens-RM, essa dimensão é complementar e transversal porque a integração analítica depende não apenas de juntar dados, mas de garantir que os dados integrados tenham significado compatível. Sem interoperabilidade semântica, a integração pode produzir datasets tecnicamente corretos, porém conceitualmente inconsistentes.

2. Problema tratado

Em ecossistemas analíticos complexos, os dados são frequentemente produzidos por diferentes sistemas, setores, equipes, projetos e domínios. Cada um pode adotar terminologias próprias, códigos específicos, classificações particulares, regras de negócio internas e definições distintas para conceitos aparentemente semelhantes.

A ausência de interoperabilidade semântica pode gerar problemas como:

- uso do mesmo termo com significados diferentes;
- uso de termos diferentes para o mesmo conceito;
- categorias incompatíveis entre fontes;
- indicadores com nomes semelhantes, mas regras de cálculo distintas;
- dificuldade de integrar bases de domínios diferentes;
- inconsistência entre dashboards e relatórios;
- ambiguidade na interpretação de atributos;
- conflitos entre regras de negócio;
- baixa comparabilidade entre períodos, cursos, unidades, territórios ou sistemas;

- perda de significado durante processos de integração;
- criação de variáveis derivadas semanticamente frágeis;
- dificuldade de reutilização de dados por outros projetos;
- aumento de retrabalho analítico;
- risco de decisões baseadas em interpretações equivocadas.

No UnespDataLens-RM, essa dimensão busca enfrentar a fragmentação semântica dos dados, permitindo que fontes heterogêneas sejam integradas e utilizadas com significado explícito, harmonizado, documentado e rastreável.

3. Objetivo da dimensão

O objetivo da dimensão **Interoperabilidade Semântica** é **garantir que dados, atributos, categorias, indicadores, rótulos, features e produtos analíticos possam ser compreendidos, comparados, integrados e reutilizados de forma semanticamente coerente**, por meio de glossários, dicionários, modelos conceituais, vocabulários controlados, mapeamentos de equivalência, regras de harmonização e registros de conflitos semânticos.

Essa dimensão busca assegurar que o UnespDataLens-RM produza ativos analíticos não apenas tecnicamente integrados, mas também conceitualmente consistentes e interpretáveis.

4. Relação com o UnespDataLens-RM

A dimensão Interoperabilidade Semântica se articula com diversos módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, contribui para identificar conceitos, termos, entidades, atributos, domínios, classificações e regras de negócio presentes em cada fonte.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, preserva o significado original dos dados, mantendo registros dos nomes, categorias, códigos e definições conforme aparecem nas fontes de origem.

No **Módulo 3 — Integração e Harmonização**, atua de forma central, permitindo mapear equivalências, resolver conflitos, alinhar categorias, compatibilizar domínios e documentar decisões semânticas.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, orienta recodificações, agregações, padronizações, criação de atributos derivados e enriquecimento semântico.

No **Módulo 5 — Qualidade e Validação de Dados**, complementa a validação técnica ao avaliar se os valores, categorias, domínios e conceitos são semanticamente válidos.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, é importante para garantir que grupos, categorias sensíveis, classes sociais, perfis ou atributos demográficos sejam definidos e comparados de forma consistente.

No **Módulo 8 — Disponibilização e Consumo Analítico**, assegura que dashboards, APIs, relatórios, indicadores e produtos de dados apresentem conceitos compreensíveis e definições explícitas.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, constitui uma das dimensões centrais, pois depende de glossários, dicionários, taxonomias, modelos semânticos e mapas de equivalência.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, apoia a definição correta de finalidades, classificações de sensibilidade, restrições e responsabilidades associadas aos dados.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, permite rastrear quando e como conceitos foram transformados, harmonizados, agregados ou recodificados.

No **Módulo 12 — Versionamento e Reprodutibilidade**, registra versões de definições, categorias, vocabulários e regras semânticas, permitindo reconstruir interpretações utilizadas em determinado período.

No **Módulo 13 — Representações Analíticas**, apoia a construção coerente de representações tabulares, temporais, espaciais, textuais, semânticas e em grafos, garantindo que entidades e relações tenham significado consistente.

No **Módulo 14 — Feature Engineering e Feature Store**, permite que features sejam interpretadas, reutilizadas e comparadas com base em regras semânticas documentadas.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, permite detectar mudanças semânticas ao longo do tempo, como alteração de categorias, redefinição de indicadores ou mudança no significado de um atributo.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, fornece base para que agentes respondam consultas conceituais, expliquem indicadores, identifiquem equivalências e alertem sobre ambiguidades semânticas.

5. Tipos de problemas semânticos

A interoperabilidade semântica busca identificar e tratar diferentes tipos de problemas de significado nos dados.

Tipo de problema	Descrição	Exemplo
Homônimos semânticos	Mesmo termo usado com significados diferentes.	“Situação” pode significar situação da matrícula, do estudante ou da disciplina.
Sinônimos conceituais	Termos diferentes usados para o mesmo conceito.	“Aluno”, “estudante” e “discente”.
Categorias incompatíveis	Fontes usam classificações distintas para o mesmo	Uma base usa “cancelado”; outra usa “desligado”.

	fenômeno.	
Granularidade semântica diferente	Conceitos são registrados em níveis de detalhe distintos.	Curso, habilitação, turma e disciplina.
Mudança de significado no tempo	Um termo ou categoria muda de sentido ao longo dos anos.	“Evasão” passa a incluir abandono e cancelamento.
Indicadores com regras distintas	Indicadores semelhantes possuem cálculos diferentes.	Taxa de permanência calculada por período ou por coorte.
Entidades parcialmente equivalentes	Entidades parecem iguais, mas representam escopos distintos.	Matrícula em disciplina versus matrícula no curso.
Códigos sem documentação	Valores codificados não possuem significado explícito.	Status `1`, `2`, `3` sem dicionário associado.
Ambiguidade de atributo	Nome do atributo não permite compreender seu significado.	Campo `tipo` sem descrição.
Equivalência contextual	Termos são equivalentes apenas em determinados contextos.	“Unidade” pode ser campus, secretaria ou unidade de atendimento.

Esses problemas não são apenas técnicos; eles afetam diretamente a validade das análises e a confiança nos produtos produzidos.

6. Estratégias de implementação

A implementação da Interoperabilidade Semântica no UnespDataLens-RM pode envolver estratégias metodológicas, documentais, computacionais e de governança.

Estratégia	Descrição
Glossário controlado	Definir termos, conceitos, sinônimos, escopo, exemplos e restrições de uso.
Dicionário de dados semântico	Documentar atributos, tipos, domínios, categorias, significados e regras de interpretação.
Mapeamento de equivalências	Relacionar termos, atributos ou categorias equivalentes entre fontes.
Registro de conflitos semânticos	Documentar divergências conceituais entre sistemas ou áreas.
Vocabulário controlado	Padronizar termos autorizados para conceitos críticos.
Taxonomias de domínio	Organizar conceitos em níveis, categorias e relações hierárquicas.
Modelos conceituais compartilhados	Representar entidades, atributos, relações e regras de negócio comuns.
Ontologias leves	Formalizar conceitos e relações com maior rigor semântico quando necessário.
Camada semântica analítica	Disponibilizar definições padronizadas para indicadores, métricas, dimensões e atributos.
Validação semântica na integração	Verificar se atributos combinados possuem significado compatível.
Versionamento de definições	Registrar mudanças em termos, categorias e regras conceituais ao longo do tempo.

Curadoria semântica	Revisar periodicamente definições, equivalências, conflitos e termos ambíguos.
Agentes inteligentes semânticos	Apoiar busca, explicação, identificação de sinônimos e resolução de ambiguidades.
Governança de conceitos críticos	Definir responsáveis por conceitos, indicadores e categorias institucionais.

7. Métricas associadas

A dimensão Interoperabilidade Semântica pode ser avaliada por métricas relacionadas à definição, equivalência, padronização, conflito, reuso e qualidade semântica.

Métrica	Sigla	Descrição
Taxa de termos críticos definidos	TTCD	Proporção de termos críticos com definição formal no glossário.
Taxa de atributos com significado documentado	TASD	Proporção de atributos com definição semântica registrada.
Taxa de categorias com domínio definido	TCDD	Proporção de categorias ou códigos com significado documentado.
Taxa de equivalências semânticas mapeadas	TESM	Proporção de equivalências identificadas e documentadas entre fontes.
Taxa de conflitos semânticos registrados	TCSR	Proporção de divergências conceituais formalmente registradas.
Taxa de conflitos semânticos resolvidos	TCSRes	Proporção de conflitos tratados por harmonização, documentação ou decisão de governança.

Taxa de indicadores com regra semântica definida	TIRSD	Proporção de indicadores com definição, escopo e regra de cálculo documentados.
Taxa de ativos aderentes ao vocabulário controlado	TAAV	Proporção de ativos que utilizam termos padronizados.
Taxa de conceitos versionados	TCV	Proporção de conceitos críticos com histórico de versões.
Índice de interoperabilidade semântica	IIS	Indicador composto da capacidade de compreensão e integração conceitual.

Uma formulação possível para o **Índice de Interoperabilidade Semântica** é:

$$\mathbf{IIS} = \mathbf{w_1TTCD} + \mathbf{w_2TASD} + \mathbf{w_3TCDD} + \mathbf{w_4TESM} + \mathbf{w_5TCSRes} + \mathbf{w_6TIRSD} + \mathbf{w_7TAAV} + \mathbf{w_8TCV} - \mathbf{w_9TCSR}$$

em que:

- **TTCD** representa a taxa de termos críticos definidos;
- **TASD** representa a taxa de atributos com significado documentado;
- **TCDD** representa a taxa de categorias com domínio definido;
- **TESM** representa a taxa de equivalências semânticas mapeadas;
- **TCSRes** representa a taxa de conflitos semânticos resolvidos;
- **TIRSD** representa a taxa de indicadores com regra semântica definida;
- **TAAV** representa a taxa de ativos aderentes ao vocabulário controlado;
- **TCV** representa a taxa de conceitos versionados;
- **TCSR** representa a taxa de conflitos semânticos registrados;
- $\mathbf{w_i}$ representa o peso atribuído a cada dimensão.

Quanto maior o IIS, maior a capacidade do ecossistema de integrar e interpretar dados de forma semanticamente coerente.

8. Artefatos relacionados

A dimensão Interoperabilidade Semântica pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de inventário, integração, metadados, catálogo, governança, lineage, versionamento, representações analíticas e agentes inteligentes.

Código	Artefato	Representação	Finalidade
IS1	Glossário Semântico	Glossário ou base	Definir termos críticos,

	Controlado	terminológica	sinônimos, escopo, exemplos e restrições de uso.
IS2	Dicionário Semântico de Atributos	Tabela ou catálogo	Documentar significado, domínio, unidade, granularidade, fonte e regra de interpretação dos atributos.
IS3	Mapa de Equivalência Semântica	Matriz ou grafo de equivalências	Relacionar termos, atributos, categorias e conceitos equivalentes entre fontes.
IS4	Registro de Conflitos Semânticos	Tabela de conflitos	Documentar divergências de significado, regras, categorias, códigos ou escopos entre fontes.
IS5	Vocabulário Controlado do Domínio	Lista controlada	Definir termos padronizados autorizados para uso em datasets, dashboards, indicadores e produtos.
IS6	Taxonomia de Conceitos	Estrutura hierárquica	Organizar conceitos em categorias, níveis, subtipos e relações de especialização.
IS7	Modelo Conceitual Compartilhado	Diagrama, grafo ou modelo entidade-relação conceitual	Representar entidades, atributos, relações e conceitos comuns entre fontes e produtos.
IS8	Registro de Versionamento Semântico	Changelog conceitual	Documentar mudanças em definições, categorias, equivalências e regras semânticas ao longo do tempo.
IS9	Relatório de Aderência Semântica	Relatório técnico	Avaliar conformidade dos ativos com glossário, vocabulário, taxonomia e regras de interpretação.

IS10	Camada Semântica Analítica	Modelo lógico ou camada de métricas	Padronizar definições de indicadores, métricas, dimensões e atributos para consumo analítico.
IS11	Plano de Curadoria Semântica	Documento operacional	Definir responsáveis, prazos, prioridades e ações para resolver ambiguidades e conflitos.
IS12	Relatório de Interoperabilidade Semântica	Relatório técnico-gerencial	Avaliar cobertura, conflitos, equivalências, lacunas e evolução da interoperabilidade semântica.

9. Critérios de validação

A dimensão Interoperabilidade Semântica pode ser considerada adequadamente tratada quando:

1. termos críticos do domínio possuem definição formal;
2. atributos relevantes possuem significado documentado;
3. categorias, códigos e domínios possuem interpretação explícita;
4. sinônimos e termos equivalentes são mapeados;
5. conflitos semânticos entre fontes são registrados;
6. conflitos críticos possuem decisão de harmonização ou restrição de uso;
7. indicadores possuem definição, escopo e regra de cálculo documentados;
8. conceitos críticos possuem responsável ou curador definido;
9. mudanças em definições ou categorias são versionadas;
10. produtos analíticos utilizam termos aderentes ao vocabulário controlado;
11. usuários conseguem compreender diferenças entre conceitos semelhantes;
12. agentes inteligentes conseguem explicar termos, equivalências e ambiguidades;
13. integrações entre fontes registram decisões semânticas;
14. a comparação entre fontes, períodos ou produtos é semanticamente justificável.

10. Exemplo de aplicação no domínio educacional

Em um projeto educacional, a interoperabilidade semântica é fundamental para integrar dados sobre matrícula, evasão, desempenho, frequência, permanência, curso, turma, disciplina, estudante e interações em ambiente virtual.

Por exemplo, o termo “evasão” pode possuir diferentes interpretações: cancelamento formal, abandono, ausência de rematrícula, desligamento, transferência ou trancamento prolongado. Se essas definições não forem explicitadas, dashboards, modelos e relatórios podem apresentar resultados divergentes mesmo quando analisam o mesmo fenômeno.

A dimensão Interoperabilidade Semântica permite definir formalmente o conceito de evasão, registrar categorias relacionadas, mapear termos equivalentes, diferenciar evasão de trancamento ou transferência, versionar mudanças de definição e indicar qual regra foi utilizada em cada produto analítico.

Outro exemplo envolve o termo “matrícula”, que pode representar matrícula no curso, matrícula em disciplina, vínculo institucional ou inscrição em atividade. A interoperabilidade semântica permite distinguir esses conceitos e evitar integrações incorretas.

11. Exemplo de aplicação em políticas públicas

Em políticas públicas, a interoperabilidade semântica é essencial porque diferentes secretarias podem usar termos semelhantes com significados distintos. Conceitos como atendimento, família, beneficiário, território, vulnerabilidade, ocorrência, serviço, equipamento público e risco podem variar conforme o órgão responsável.

Por exemplo, uma secretaria pode registrar “atendimento” como qualquer contato com o cidadão, enquanto outra considera apenas atendimentos formalizados em sistema. Se essas diferenças não forem documentadas, indicadores integrados podem gerar interpretações equivocadas sobre cobertura de serviços.

A dimensão Interoperabilidade Semântica permite criar glossários intersetoriais, mapear equivalências, registrar conflitos conceituais, definir taxonomias de serviços e estabelecer uma camada semântica comum para análise territorial, social e institucional.

12. Contribuição para o modelo de referência

A dimensão **Interoperabilidade Semântica** contribui para o UnespDataLens-RM ao reconhecer que a integração de dados não depende apenas de compatibilidade técnica, mas também de compatibilidade conceitual.

Essa dimensão fortalece o modelo de referência ao:

- reduzir ambiguidades conceituais;
- melhorar integração entre fontes heterogêneas;
- apoiar construção de indicadores consistentes;
- favorecer reuso de dados e produtos;
- aumentar qualidade dos metadados;
- reduzir conflitos entre áreas e sistemas;
- fortalecer catálogos, glossários e modelos semânticos;
- apoiar agentes inteligentes com conhecimento conceitual confiável;
- melhorar rastreabilidade de decisões semânticas;
- reduzir riscos de interpretações equivocadas;
- apoiar comparabilidade histórica e institucional;
- fortalecer a tomada de decisão baseada em evidências.

13. Síntese da dimensão

A dimensão **Interoperabilidade Semântica** trata da capacidade de integrar, comparar, interpretar e reutilizar dados preservando o significado dos conceitos envolvidos. No UnespDataLens-RM, essa dimensão atua como capacidade complementar para garantir

que atributos, categorias, indicadores, rótulos, features, grafos, dashboards, APIs e produtos analíticos sejam semanticamente compreensíveis, compatíveis e rastreáveis.

Ao incorporar glossários controlados, dicionários semânticos, mapas de equivalência, registros de conflitos, vocabulários controlados, taxonomias, modelos conceituais, versionamento semântico, camada semântica analítica e curadoria de conceitos, o modelo de referência fortalece sua capacidade de produzir ativos analíticos integrados não apenas tecnicamente, mas também conceitualmente.

Dimensão Complementar — DataOps

1. Definição da dimensão

A dimensão **DataOps** refere-se ao conjunto de princípios, práticas, processos, métodos e mecanismos técnicos voltados à automação, orquestração, versionamento, monitoramento, validação, colaboração, governança e melhoria contínua dos pipelines de dados. Seu objetivo é tornar o ciclo de vida dos dados mais confiável, reprodutível, rastreável, auditável, eficiente e adaptável a mudanças.

No contexto do **UnespDataLens-RM**, DataOps representa a capacidade de operacionalizar o modelo de referência como um ecossistema analítico contínuo, no qual fontes, ingestões, transformações, validações, armazenamentos, disponibilizações, metadados, linhagem, versionamentos, monitoramentos e produtos de dados são tratados como processos controlados, automatizados e avaliáveis.

A dimensão DataOps não se limita à automação técnica de scripts ou à execução de pipelines. Ela envolve uma abordagem metodológica para garantir que os dados sejam produzidos e entregues com qualidade, previsibilidade e rastreabilidade. Isso inclui padronização de processos, controle de versões, testes de qualidade, orquestração de tarefas, gestão de dependências, monitoramento de falhas, auditoria de execuções, documentação automática, gestão de ambientes, colaboração entre equipes e resposta rápida a incidentes.

No UnespDataLens-RM, DataOps é uma dimensão complementar e transversal, pois apoia todos os módulos do modelo de referência. Ela funciona como uma infraestrutura operacional e metodológica que permite transformar o desenho conceitual do modelo de referência em processos executáveis, monitoráveis, reprodutíveis e sustentáveis.

2. Problema tratado

Em muitos projetos de dados, pipelines são construídos de forma ad hoc, com scripts manuais, pouca documentação, baixa rastreabilidade, ausência de testes, dependência de pessoas específicas, dificuldade de reexecução e pouca visibilidade sobre falhas. Embora esses processos possam funcionar em curto prazo, tendem a se tornar frágeis, difíceis de manter e pouco confiáveis à medida que o ecossistema analítico cresce.

A ausência de práticas de DataOps pode gerar problemas como:

- pipelines executados manualmente;
- ausência de orquestração entre etapas;
- scripts sem versionamento;
- regras de transformação não documentadas;
- falhas de execução não detectadas;
- dificuldade de reprocessar dados;
- inconsistência entre ambientes de desenvolvimento e produção;
- baixa rastreabilidade de execuções;
- ausência de testes automatizados de qualidade;
- dificuldade de reproduzir datasets e produtos;
- dependência de conhecimento tácito;
- demora para identificar e corrigir incidentes;
- retrabalho técnico;

- baixa confiabilidade dos produtos analíticos;
- dificuldade de escalar o ecossistema de dados.

No UnespDataLens-RM, a dimensão DataOps busca reduzir esses problemas ao estabelecer práticas operacionais padronizadas para construção, execução, validação, monitoramento, versionamento e manutenção dos pipelines de Engenharia Analítica de Dados.

3. Objetivo da dimensão

O objetivo da dimensão **DataOps** é **estruturar práticas operacionais, automatizadas, colaborativas, versionadas, testáveis, monitoráveis e reprodutíveis para o desenvolvimento, execução e manutenção dos pipelines de dados do UnespDataLens-RM**.

Essa dimensão busca garantir que os processos de dados não sejam apenas conceitualmente definidos, mas também operacionalmente confiáveis, controlados, auditáveis e sustentáveis ao longo do tempo.

4. Relação com o UnespDataLens-RM

A dimensão DataOps se articula com todos os módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, DataOps contribui para padronizar a coleta de informações sobre fontes, responsáveis, formatos, periodicidades, restrições e requisitos operacionais de ingestão.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, DataOps atua de forma central por meio da automação de ingestões, controle de execução, logs, retries, checkpoints, manifestos, validação inicial e armazenamento controlado dos dados brutos.

No **Módulo 3 — Integração e Harmonização**, apoia a orquestração de dependências entre fontes, execução controlada de integrações, validação de chaves, registro de conflitos e rastreabilidade das regras de harmonização.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, DataOps permite executar transformações de forma padronizada, versionada, testada e reprodutível.

No **Módulo 5 — Qualidade e Validação de Dados**, fornece mecanismos para testes automatizados, validações recorrentes, bloqueio de pipelines em caso de falhas críticas e geração de relatórios de qualidade.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, apoia a execução periódica de métricas de representatividade, fairness, distribuição de grupos e verificação de impactos de transformações.

No **Módulo 7 — Armazenamento Analítico**, contribui para processos controlados de carga, atualização, particionamento, versionamento, publicação e organização de datasets analíticos.

No **Módulo 8 — Disponibilização e Consumo Analítico**, apoia a atualização automatizada de dashboards, APIs, views, relatórios e produtos de dados, com controle de versão, status e disponibilidade.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, permite captura automática de metadados, atualização do catálogo, extração de schemas, registro de execuções e sincronização com produtos de dados.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, DataOps garante que controles de acesso, políticas, logs, classificações e restrições sejam incorporados aos fluxos operacionais.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, DataOps fornece registros de execução, entradas, saídas, regras, versões, parâmetros e dependências necessários à construção da linhagem.

No **Módulo 12 — Versionamento e Reprodutibilidade**, DataOps atua no versionamento de dados, scripts, ambientes, parâmetros, manifestos e produtos gerados.

No **Módulo 13 — Representações Analíticas**, permite que representações tabulares, temporais, espaciais, textuais, semânticas e em grafos sejam geradas de forma reprodutível e controlada.

No **Módulo 14 — Feature Engineering e Feature Store**, apoia pipelines automatizados de criação, validação, publicação, versionamento e monitoramento de features.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, DataOps fornece a infraestrutura operacional para monitorar pipelines, detectar falhas, gerar alertas e acionar ações corretivas.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, agentes podem consultar logs, status, falhas, versões, dependências e relatórios operacionais produzidos por práticas de DataOps.

5. Princípios de DataOps

A dimensão DataOps pode ser estruturada a partir de princípios operacionais que orientam a construção e manutenção dos pipelines.

Princípio	Descrição	Exemplo
Automação	Reduzir execução manual de tarefas recorrentes.	Agendar ingestão automática de fontes periódicas.
Orquestração	Controlar dependências e sequência de etapas do pipeline.	Executar integração apenas após ingestão bem-sucedida.

Versionamento	Controlar versões de dados, scripts, regras, parâmetros e produtos.	Versionar transformação que gera dataset curado.
Testabilidade	Incorporar testes automatizados de schema, qualidade e consistência.	Bloquear carga se completude mínima não for atendida.
Observabilidade	Registrar logs, métricas, status, falhas e alertas.	Painel de execução dos pipelines.
Reprodutibilidade	Permitir reconstruir execuções e resultados.	Manifesto com entradas, código, parâmetros e saída.
Colaboração	Apoiar trabalho coordenado entre equipes técnicas, analistas e gestores.	Repositório compartilhado com revisão de mudanças.
Governança operacional	Incorporar políticas, permissões e responsabilidades ao fluxo.	Pipeline respeita restrição de acesso ao dado sensível.
Melhoria contínua	Revisar falhas, incidentes, gargalos e oportunidades de otimização.	Relatório pós-incidente orienta ajuste do pipeline.
Entrega confiável	Garantir que produtos de dados sejam atualizados, validados e disponibilizados com segurança.	Dashboard só é atualizado após validação do dataset.

6. Estratégias de implementação

A implementação de DataOps no UnespDataLens-RM pode envolver estratégias técnicas, metodológicas e organizacionais.

Estratégia	Descrição

Orquestração de pipelines	Definir workflows, dependências, agendamentos, retries, checkpoints e tratamento de falhas.
Controle de versão de código	Versionar scripts, notebooks, consultas, configurações e regras de transformação.
Versionamento de dados e artefatos	Controlar snapshots, datasets, schemas, features, produtos e relatórios.
Testes automatizados de dados	Aplicar testes de schema, completude, validade, unicidade, domínio e consistência.
Integração contínua de pipelines	Validar mudanças em código, regras e configurações antes de publicação.
Entrega contínua de produtos de dados	Publicar datasets, dashboards e APIs após validação e aprovação.
Logs estruturados de execução	Registrar entrada, saída, parâmetros, tempo, status, erros e responsável.
Manifestos de execução	Documentar versões, dependências, ambiente, parâmetros e produtos gerados.
Ambientes controlados	Separar desenvolvimento, teste, homologação e produção.
Monitoramento de execução	Acompanhar falhas, atrasos, tempo de execução, volume processado e disponibilidade.
Gestão de incidentes	Registrar falhas, causa provável, impacto, responsável e ação corretiva.
Padronização de	Organizar scripts, configurações, documentação e

repositórios	artefatos de forma consistente.
Revisão e aprovação de mudanças	Controlar alterações críticas em pipelines, schemas e produtos.
Documentação automática	Atualizar catálogo, lineage e metadados a partir das execuções.

7. Métricas associadas

A dimensão DataOps pode ser avaliada por métricas relacionadas à automação, confiabilidade, reprodutibilidade, testabilidade, monitoramento, incidentes e entrega de produtos.

Métrica	Sigla	Descrição
Taxa de pipelines automatizados	TPA	Proporção de pipelines executados automaticamente.
Taxa de pipelines orquestrados	TPO	Proporção de pipelines com dependências, agendamento e controle de execução.
Taxa de execuções bem-sucedidas	TEBS	Proporção de execuções concluídas sem falhas.
Taxa de execuções com logs estruturados	TELE	Proporção de execuções com logs completos e consultáveis.
Taxa de pipelines com testes automatizados	TPTA	Proporção de pipelines com validações automáticas.
Taxa de falhas detectadas automaticamente	TFDA	Proporção de falhas identificadas por monitoramento ou testes.

Tempo médio de recuperação	TMR	Tempo médio para corrigir falha e restaurar execução.
Taxa de pipelines versionados	TPV	Proporção de pipelines com código, regras e configurações versionados.
Taxa de execuções reprodutíveis	TER	Proporção de execuções que podem ser reconstruídas.
Taxa de produtos entregues com validação	TPEV	Proporção de produtos publicados após validação automática ou humana.
Taxa de incidentes recorrentes	TIR	Proporção de falhas repetidas em determinado período.
Índice de maturidade DataOps	IMDOPS	Indicador composto da maturidade operacional dos pipelines.

Uma formulação possível para o **Índice de Maturidade DataOps** é:

$$\mathbf{IMDOPS} = w_1\mathbf{TPA} + w_2\mathbf{TPO} + w_3\mathbf{TEBS} + w_4\mathbf{TELE} + w_5\mathbf{TPTA} + w_6\mathbf{TFDA} + w_7\mathbf{TPV} + w_8\mathbf{TER} + w_9\mathbf{TPEV} - w_{10}\mathbf{TIR}$$

em que:

- **TPA** representa a taxa de pipelines automatizados;
- **TPO** representa a taxa de pipelines orquestrados;
- **TEBS** representa a taxa de execuções bem-sucedidas;
- **TELE** representa a taxa de execuções com logs estruturados;
- **TPTA** representa a taxa de pipelines com testes automatizados;
- **TFDA** representa a taxa de falhas detectadas automaticamente;
- **TPV** representa a taxa de pipelines versionados;
- **TER** representa a taxa de execuções reprodutíveis;
- **TPEV** representa a taxa de produtos entregues com validação;
- **TIR** representa a taxa de incidentes recorrentes;
- w_i representa o peso atribuído a cada dimensão.

Quanto maior o IMDOPS, maior a maturidade operacional do ecossistema de dados.

8. Artefatos relacionados

A dimensão DataOps pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de ingestão, transformação, qualidade, armazenamento,

disponibilização, metadados, lineage, versionamento, observabilidade, monitoramento e agentes inteligentes.

Código	Artefato	Representação	Finalidade
DOPS1	Arquitetura Operacional DataOps	Documento técnico ou diagrama	Definir componentes, fluxos, ambientes, dependências e responsabilidades operacionais.
DOPS2	Catálogo de Pipelines	Catálogo ou tabela operacional	Registrar pipelines, finalidade, responsável, frequência, entradas, saídas e status.
DOPS3	Workflow Orquestrado	DAG, fluxo ou especificação de execução	Representar dependências, ordem de execução, retries, checkpoints e critérios de falha.
DOPS4	Repositório Versionado de Pipelines	Repositório de código	Armazenar scripts, notebooks, configurações, regras e documentação técnica.
DOPS5	Plano de Testes de Dados	Documento técnico ou suíte de testes	Definir validações de schema, qualidade, domínio, consistência e integridade.
DOPS6	Registro de Execuções	Log estruturado	Documentar data, status, duração, entradas, saídas, parâmetros, erros e responsável.
DOPS7	Manifesto de Execução DataOps	Documento reprodutível	Registrar versões de dados, código, ambiente, parâmetros, dependências e artefatos gerados.
DOPS8	Painel Operacional de Pipelines	Dashboard	Monitorar execuções, falhas, atrasos, duração, volumes e disponibilidade.
DOPS9	Registro de	Tabela ou sistema	Documentar falhas, causas,

	Incidentes Operacionais	de incidentes	impactos, responsáveis, ações corretivas e recorrência.
DOPS10	Plano de Recuperação de Pipeline	Documento operacional	Definir procedimentos para reexecução, rollback, correção e reprocessamento.
DOPS11	Relatório de Maturidade DataOps	Relatório técnico-gerencial	Avaliar automação, orquestração, testes, versionamento, reprodutibilidade e monitoramento.
DOPS12	Plano de Melhoria Contínua DataOps	Documento de evolução	Definir ações, prioridades e prazos para aumento da maturidade operacional.

9. Critérios de validação

A dimensão DataOps pode ser considerada adequadamente tratada quando:

1. pipelines críticos estão catalogados;
2. pipelines recorrentes são automatizados;
3. dependências entre etapas estão formalizadas;
4. execuções geram logs estruturados;
5. scripts, regras e configurações são versionados;
6. datasets e produtos críticos possuem manifestos de execução;
7. testes de dados são executados automaticamente;
8. falhas críticas geram alertas;
9. incidentes operacionais são registrados e tratados;
10. produtos de dados são publicados apenas após validação;
11. execuções podem ser reproduzidas;
12. ambientes de desenvolvimento, teste e produção são controlados;
13. responsáveis por pipelines estão definidos;
14. métricas de maturidade DataOps são acompanhadas ao longo do tempo.

10. Exemplo de aplicação no domínio educacional

Em um projeto educacional, DataOps pode ser aplicado à construção de pipelines que extraem dados de matrícula, histórico acadêmico, frequência, interações em ambiente virtual, notas e dados socioeconômicos.

Por exemplo, um pipeline de evasão pode ser orquestrado para executar ingestão das bases acadêmicas, validação de schema, verificação de completude, integração por estudante e período, construção do rótulo de evasão, geração de features temporais, validação de qualidade, atualização do dataset curado e publicação de um dashboard institucional.

Cada execução registra logs, versões de entrada, regras aplicadas, parâmetros, métricas de qualidade, tempo de execução e produtos gerados. Se a base de matrícula não for atualizada no prazo ou apresentar mudança de schema, o pipeline gera alerta, bloqueia a publicação e registra o incidente.

11. Exemplo de aplicação em políticas públicas

Em políticas públicas, DataOps pode apoiar pipelines que integram bases de saúde, educação, assistência social, território, atendimento e indicadores socioeconômicos.

Por exemplo, um pipeline territorial pode executar mensalmente a ingestão de dados de diferentes secretarias, validar schemas, harmonizar códigos territoriais, calcular indicadores agregados, atualizar data marts, publicar dashboards e registrar lineage. Se uma secretaria enviar dados incompletos, o pipeline pode gerar alerta, registrar violação, acionar responsável e impedir atualização de indicadores críticos.

Assim, DataOps fortalece a confiabilidade operacional dos produtos de dados utilizados por gestores públicos.

12. Contribuição para o modelo de referência

A dimensão **DataOps** contribui para o UnespDataLens-RM ao transformar os módulos do modelo de referência em processos operacionais controlados, automatizados, monitoráveis e reprodutíveis.

Essa dimensão fortalece o modelo de referência ao:

- reduzir execução manual de pipelines;
- aumentar confiabilidade das entregas de dados;
- padronizar processos operacionais;
- automatizar testes de qualidade;
- melhorar rastreabilidade de execuções;
- apoiar reprodutibilidade dos resultados;
- reduzir falhas silenciosas;
- acelerar diagnóstico e recuperação de incidentes;
- integrar versionamento, qualidade, lineage e monitoramento;
- reduzir dívida de dados;
- apoiar escalabilidade do ecossistema;
- melhorar comunicação entre equipes técnicas e usuários.

13. Síntese da dimensão

A dimensão **DataOps** trata da operacionalização confiável, automatizada, versionada, testável, observável e reprodutível dos pipelines de dados. No UnespDataLens-RM, essa dimensão atua como capacidade complementar para garantir que os processos definidos pelo modelo de referência possam ser executados, monitorados, auditados, corrigidos e evoluídos de forma sistemática.

Ao incorporar automação, orquestração, versionamento, testes de dados, logs estruturados, manifestos de execução, painéis operacionais, gestão de incidentes e melhoria contínua, o modelo de referência fortalece sua capacidade de produzir ativos analíticos confiáveis, atualizados, governados e adequados à tomada de decisão baseada em evidências.

Dimensão Complementar — MLOps

1. Definição da dimensão

A dimensão **MLOps** refere-se ao conjunto de princípios, práticas, processos e mecanismos técnicos voltados à gestão do ciclo de vida de modelos de aprendizagem de máquina, desde a preparação dos dados e construção de features até o treinamento, avaliação, versionamento, implantação, monitoramento, explicação, governança, atualização e descontinuação de modelos.

No contexto do **UnespDataLens-RM**, MLOps representa a capacidade de operacionalizar modelos analíticos, preditivos, classificatórios, explicativos ou prescritivos de forma controlada, reprodutível, rastreável, monitorável e governada. Essa dimensão não se restringe ao treinamento de algoritmos. Ela envolve todo o ecossistema que permite transformar modelos em artefatos confiáveis, avaliáveis e integrados aos pipelines de Engenharia Analítica de Dados.

MLOps inclui práticas como versionamento de datasets de treinamento, controle de features, registro de experimentos, comparação de modelos, validação de desempenho, análise de fairness, explicabilidade, empacotamento de modelos, implantação controlada, monitoramento de drift, avaliação contínua, auditoria de decisões, atualização periódica e descontinuação de modelos obsoletos.

No UnespDataLens-RM, essa dimensão é complementar e transversal porque modelos dependem dos dados, metadados, qualidade, rótulos, features, linhagem, versionamento, governança, monitoramento e agentes inteligentes produzidos pelos demais módulos do modelo de referência. Um modelo não deve ser visto como um componente isolado, mas como um ativo analítico derivado de dados, regras, escolhas metodológicas, parâmetros, versões e contextos de uso.

2. Problema tratado

Em muitos projetos de Ciência de Dados e Inteligência Artificial, modelos são desenvolvidos em notebooks, scripts ou experimentos pontuais, mas sem controle adequado sobre os dados utilizados, as versões das features, os parâmetros, o ambiente de execução, as métricas, os critérios de validação, os riscos de viés, as limitações ou as condições de implantação.

Essa situação pode gerar modelos difíceis de reproduzir, comparar, auditar, atualizar ou explicar. Um modelo pode apresentar bom desempenho experimental, mas não estar preparado para uso contínuo, pois depende de dados instáveis, features sem versionamento, rótulos frágeis, métricas incompletas, ausência de monitoramento ou falta de governança.

A ausência de MLOps pode gerar problemas como:

- experimentos não reprodutíveis;
- modelos treinados com datasets sem versão;
- uso de features não documentadas;
- parâmetros e hiperparâmetros não registrados;
- ausência de rastreabilidade entre modelo, dados e código;
- comparação inadequada entre modelos;

- implantação de modelos sem validação suficiente;
- degradação de desempenho não detectada;
- data drift, feature drift, label drift ou concept drift sem monitoramento;
- ausência de explicabilidade ou interpretabilidade;
- baixa avaliação de fairness e vieses;
- dificuldade de auditar decisões apoiadas por modelos;
- modelos obsoletos mantidos em uso;
- falta de critérios para retreinamento;
- risco de uso indevido de predições em processos decisórios.

No UnespDataLens-RM, a dimensão MLOps busca evitar que modelos sejam tratados apenas como resultados de experimentos, formalizando-os como ativos analíticos versionados, monitorados, avaliados, governados e integrados ao ciclo de vida dos dados.

3. Objetivo da dimensão

O objetivo da dimensão **MLOps** é **estruturar práticas para construção, avaliação, versionamento, implantação, monitoramento, governança e manutenção de modelos de aprendizagem de máquina no ecossistema UnespDataLens-RM**, garantindo que esses modelos sejam reprodutíveis, rastreáveis, explicáveis, avaliáveis, atualizáveis e adequados ao uso analítico ou decisório.

Essa dimensão busca assegurar que modelos derivados dos dados do UnespDataLens-RM sejam tecnicamente robustos, metodologicamente documentados, operacionalmente monitorados e institucionalmente responsáveis.

4. Relação com o UnespDataLens-RM

A dimensão MLOps se articula com diversos módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, MLOps depende da identificação das fontes que alimentam os modelos, seus responsáveis, limitações, periodicidade, granularidade e restrições de uso.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, garante que os dados utilizados no treinamento e avaliação dos modelos possam ser rastreados até sua origem e versão inicial.

No **Módulo 3 — Integração e Harmonização**, apoia a construção de bases integradas para modelagem, preservando regras de relacionamento, chaves, conflitos e limitações.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, registra transformações aplicadas aos dados antes da modelagem, incluindo tratamento de ausentes, normalizações, recodificações e enriquecimentos.

No **Módulo 5 — Qualidade e Validação de Dados**, utiliza métricas de qualidade para avaliar se os dados são adequados ao treinamento, validação e teste dos modelos.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, incorpora avaliação de distribuição de classes, representatividade, métricas de fairness, mitigação de viés e análise de impacto entre grupos.

No **Módulo 7 — Armazenamento Analítico**, utiliza datasets curados, data marts e bases analíticas como insumos controlados para treinamento e avaliação.

No **Módulo 8 — Disponibilização e Consumo Analítico**, modelos podem gerar scores, classificações, recomendações, indicadores ou serviços analíticos consumidos por dashboards, APIs ou relatórios.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, modelos devem ser descritos por metadados, incluindo finalidade, dados de entrada, variável-alvo, features, métricas, limitações e responsáveis.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, MLOps deve respeitar políticas de acesso, finalidade de uso, dados sensíveis, privacidade, auditoria e responsabilidade institucional.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, registra a relação entre modelo, datasets, features, rótulos, código, parâmetros, versões e produtos derivados.

No **Módulo 12 — Versionamento e Reprodutibilidade**, controla versões de dados, features, código, experimentos, modelos, ambientes, parâmetros e resultados.

No **Módulo 13 — Representações Analíticas**, modelos podem utilizar representações tabulares, temporais, relacionais, grafos ou grafos temporais como entrada.

No **Módulo 14 — Feature Engineering e Feature Store**, MLOps depende do controle de features, regras de cálculo, versionamento, qualidade, drift e risco de vazamento.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, acompanha degradação de desempenho, drift dos dados, drift das features, drift dos rótulos, mudanças no comportamento do modelo e necessidade de retreinamento.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, agentes podem apoiar explicação de modelos, consulta a experimentos, interpretação de métricas, identificação de drift e recomendação de retreinamento.

5. Componentes de MLOps

A dimensão MLOps pode ser estruturada em componentes que cobrem o ciclo completo de modelos no UnespDataLens-RM.

Componente	Descrição	Exemplo
Dataset de treinamento	Base utilizada para treinar o modelo, com versão e qualidade registradas.	Dataset curado de evasão v2.1.
Dataset de validação e teste	Bases separadas para avaliação do desempenho e generalização.	Conjunto de validação temporal.

Feature set	Conjunto de atributos utilizados pelo modelo.	Features acadêmicas, temporais e socioeconômicas.
Variável-alvo ou rótulo	Desfecho que o modelo busca prever, classificar ou estimar.	Evasão no próximo período.
Experimento	Execução controlada de treinamento e avaliação.	XGBoost com parâmetros específicos.
Modelo treinado	Artefato resultante do processo de aprendizagem.	Modelo preditivo salvo e versionado.
Métricas de avaliação	Indicadores de desempenho e qualidade do modelo.	Acurácia, F1, AUC, recall, precisão.
Métricas de fairness	Indicadores de equidade entre grupos.	SPD, DI, EOD, AOD, FPR e FNR por grupo.
Explicabilidade	Métodos para interpretação do comportamento do modelo.	SHAP, LIME, importância de variáveis.
Registro de modelo	Catálogo ou repositório de modelos e versões.	Model registry.
Implantação	Disponibilização do modelo para uso analítico ou operacional.	API de score de risco.
Monitoramento	Acompanhamento de desempenho, drift, uso e estabilidade.	Alerta de queda de recall.
Retreinamento	Atualização do modelo com novos dados ou novas condições.	Retreinamento semestral.

Descontinuação	Retirada controlada de modelos obsoletos ou inadequados.	Modelo substituído por nova versão.

6. Estratégias de implementação

A implementação de MLOps no UnespDataLens-RM pode envolver estratégias técnicas, metodológicas e de governança.

Estratégia	Descrição
Registro de experimentos	Documentar dados, features, rótulos, algoritmos, parâmetros, métricas e resultados.
Versionamento de modelos	Controlar versões de modelos treinados, seus artefatos e configurações.
Versionamento de datasets	Registrar versões dos dados usados em treinamento, validação, teste e produção.
Versionamento de features	Controlar regras, janelas temporais, parâmetros e versões das features.
Separação treino-validação-teste	Definir protocolos de avaliação adequados ao problema e ao tempo.
Validação temporal	Evitar vazamento temporal e avaliar generalização em períodos futuros.
Avaliação de desempenho	Medir métricas adequadas ao problema, como acurácia, F1, AUC, RMSE ou MAE.
Avaliação de fairness	Avaliar desempenho e erros por grupos relevantes ou sensíveis.

Explicabilidade do modelo	Utilizar métodos para explicar variáveis, decisões, predições e limitações.
Registro de limitações	Documentar escopo, contexto, restrições e riscos do modelo.
Governança de implantação	Estabelecer critérios para publicação, uso, revisão e aprovação de modelos.
Monitoramento em produção	Acompanhar drift, desempenho, estabilidade, uso, latência e falhas.
Critérios de retreinamento	Definir quando o modelo deve ser atualizado ou substituído.
Auditoria de decisões	Registrar predições, versões, entradas, saídas e explicações quando aplicável.
Descontinuação controlada	Retirar modelos obsoletos preservando histórico, justificativa e versões anteriores.

7. Métricas associadas

A dimensão MLOps pode ser avaliada por métricas relacionadas à reprodutibilidade, desempenho, fairness, estabilidade, implantação, monitoramento e governança dos modelos.

Métrica	Sigla	Descrição
Taxa de modelos registrados	TMR	Proporção de modelos com registro formal no catálogo ou model registry.
Taxa de modelos versionados	TMV	Proporção de modelos com controle de versão.

Taxa de experimentos reprodutíveis	TER	Proporção de experimentos com dados, código, parâmetros e ambiente registrados.
Taxa de modelos com dataset versionado	TMDV	Proporção de modelos associados a versões dos datasets utilizados.
Taxa de modelos com feature set documentado	TMFD	Proporção de modelos com features, regras e versões documentadas.
Taxa de modelos com rótulo documentado	TMRD	Proporção de modelos com variável-alvo formalmente definida.
Taxa de modelos com avaliação de desempenho	TMAD	Proporção de modelos com métricas de desempenho registradas.
Taxa de modelos com avaliação de fairness	TMAF	Proporção de modelos avaliados quanto a equidade entre grupos.
Taxa de modelos explicáveis	TME	Proporção de modelos com relatório de explicabilidade ou interpretabilidade.
Taxa de modelos monitorados	TMM	Proporção de modelos em uso com monitoramento ativo.
Taxa de modelos com drift detectado	TMDD	Proporção de modelos com evidência de drift nos dados, features, rótulos ou desempenho.
Taxa de modelos retreinados conforme critério	TMRC	Proporção de modelos atualizados conforme regra de retreinamento.
Taxa de modelos obsoletos	TMO	Proporção de modelos sem uso, sem monitoramento ou superados por nova versão.

Índice de maturidade MLOps	IMLOps	Indicador composto da maturidade operacional e governança dos modelos.

Uma formulação possível para o **Índice de Maturidade MLOps** é:

$$\mathbf{IMLOps} = \mathbf{w_1TMR} + \mathbf{w_2TMV} + \mathbf{w_3TER} + \mathbf{w_4TMDV} + \mathbf{w_5TMFD} + \mathbf{w_6TMRD} + \mathbf{w_7TMAD} + \mathbf{w_8TMAF} + \mathbf{w_9TME} + \mathbf{w_{10}TMM} + \mathbf{w_{11}TMRC} - \mathbf{w_{12}TMO}$$

em que:

- **TMR** representa a taxa de modelos registrados;
- **TMV** representa a taxa de modelos versionados;
- **TER** representa a taxa de experimentos reprodutíveis;
- **TMDV** representa a taxa de modelos com dataset versionado;
- **TMFD** representa a taxa de modelos com feature set documentado;
- **TMRD** representa a taxa de modelos com rótulo documentado;
- **TMAD** representa a taxa de modelos com avaliação de desempenho;
- **TMAF** representa a taxa de modelos com avaliação de fairness;
- **TME** representa a taxa de modelos explicáveis;
- **TMM** representa a taxa de modelos monitorados;
- **TMRC** representa a taxa de modelos retreinados conforme critério;
- **TMO** representa a taxa de modelos obsoletos;
- $\mathbf{w_i}$ representa o peso atribuído a cada dimensão.

Quanto maior o IMLOps, maior a maturidade do ecossistema na gestão do ciclo de vida de modelos.

8. Artefatos relacionados

A dimensão MLOps pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de qualidade, feature store, versionamento, linhagem, governança, monitoramento, explicabilidade e agentes inteligentes.

Código	Artefato	Representação	Finalidade
MLO1	Arquitetura MLOps	Documento técnico ou diagrama	Definir componentes, fluxos, ambientes, registros, integrações e responsabilidades do ciclo de modelos.
MLO2	Catálogo de Modelos	Model registry ou tabela de modelos	Registrar modelos, versões, finalidade, responsável, status, métricas e restrições.

MLO3	Registro de Experimentos	Tabela ou ferramenta de tracking	Documentar datasets, features, rótulos, algoritmos, parâmetros, métricas e resultados.
MLO4	Manifesto de Treinamento	Documento reprodutível	Registrar dados, versões, código, ambiente, parâmetros, bibliotecas, seeds e artefatos gerados.
MLO5	Relatório de Avaliação de Desempenho	Relatório técnico	Avaliar métricas de desempenho, erro, estabilidade, validação temporal e comparação entre modelos.
MLO6	Relatório de Fairness do Modelo	Relatório técnico	Avaliar desempenho, erros e impactos diferenciais entre grupos.
MLO7	Relatório de Explicabilidade	Relatório técnico	Explicar variáveis relevantes, predições, comportamento global e limitações do modelo.
MLO8	Registro de Implantação do Modelo	Documento ou tabela operacional	Documentar versão implantada, ambiente, API, data, responsável, status e critérios de uso.
MLO9	Painel de Monitoramento de Modelos	Dashboard	Monitorar drift, desempenho, latência, uso, falhas e estabilidade do modelo.
MLO10	Registro de Retreinamento	Changelog ou tabela de atualização	Documentar critérios, datas, versões, motivos e resultados de retreinamento.
MLO11	Registro de Predições e Auditoria	Log governado	Registrar entradas, saídas, versão do modelo, explicações e contexto quando aplicável.

MLO12	Plano de Descontinuação de Modelo	Documento operacional	Definir critérios, justificativa, impacto e substituição de modelos obsoletos.
MLO13	Model Card	Documento padronizado	Sintetizar finalidade, dados, desempenho, limitações, riscos, fairness, uso recomendado e uso não recomendado.
MLO14	Plano de Governança de Modelos	Documento normativo	Definir papéis, aprovações, responsabilidades, restrições, auditoria e revisão dos modelos.

9. Critérios de validação

A dimensão MLOps pode ser considerada adequadamente tratada quando:

1. modelos críticos estão catalogados;
2. modelos possuem finalidade, responsável, versão e status definidos;
3. datasets de treinamento, validação e teste estão versionados;
4. features utilizadas pelos modelos estão documentadas e rastreáveis;
5. variáveis-alvo ou rótulos possuem definição e regra de construção;
6. experimentos registram algoritmos, parâmetros, ambiente e resultados;
7. métricas de desempenho são calculadas e comparadas;
8. modelos de alto impacto possuem avaliação de fairness;
9. modelos utilizados em decisões possuem explicabilidade ou justificativa de interpretabilidade;
10. modelos implantados possuem registro de versão e ambiente;
11. modelos em uso são monitorados quanto a drift, desempenho e estabilidade;
12. critérios de retreinamento estão definidos;
13. predições relevantes podem ser auditadas;
14. modelos obsoletos são descontinuados de forma controlada;
15. limitações, riscos e usos não recomendados são documentados.

10. Exemplo de aplicação no domínio educacional

Em um projeto educacional, MLOps pode ser aplicado ao desenvolvimento e manutenção de modelos preditivos de evasão, desempenho, risco acadêmico, engajamento ou recomendação de apoio pedagógico.

Por exemplo, um modelo de evasão pode ser treinado a partir de dados de matrícula, histórico acadêmico, desempenho, frequência, interações no ambiente virtual e perfil socioeconômico. A dimensão MLOps exige que o dataset utilizado seja versionado, que o rótulo de evasão seja formalmente definido, que as features sejam documentadas, que o experimento registre parâmetros e métricas, que o modelo seja avaliado por período e por grupos, que suas principais variáveis sejam explicadas e que sua implantação seja monitorada.

Se a distribuição dos estudantes mudar, se novas regras acadêmicas forem adotadas ou se o desempenho do modelo cair em determinado grupo, o monitoramento MLOps deve gerar alertas e orientar revisão ou retreinamento.

11. Exemplo de aplicação em políticas públicas

Em políticas públicas, MLOps pode apoiar modelos de priorização de atendimento, previsão de demanda, classificação de vulnerabilidade, detecção de risco, alocação de recursos ou avaliação de programas.

Por exemplo, um modelo de priorização territorial pode usar dados de saúde, educação, assistência social, renda e infraestrutura. Nesse caso, MLOps exige rastreabilidade entre dados, features, rótulo e predição, além de avaliação de desempenho, fairness entre territórios ou grupos populacionais, explicabilidade dos fatores associados ao score e monitoramento de drift.

Como esses modelos podem influenciar decisões públicas, a governança deve definir limites de uso, revisão humana, auditoria das predições e documentação clara sobre riscos e limitações.

12. Contribuição para o modelo de referência

A dimensão **MLOps** contribui para o UnespDataLens-RM ao formalizar modelos de aprendizagem de máquina como ativos analíticos que precisam ser versionados, documentados, monitorados, auditados e governados.

Essa dimensão fortalece o modelo de referência ao:

- aumentar reprodutibilidade de experimentos;
- melhorar controle sobre datasets, features e rótulos;
- apoiar comparação sistemática de modelos;
- incorporar fairness e explicabilidade ao ciclo de modelagem;
- reduzir risco de modelos obsoletos ou degradados;
- permitir monitoramento de drift e desempenho;
- apoiar decisões sobre retreinamento;
- integrar modelos a pipelines de dados governados;
- fortalecer auditoria de predições;
- reduzir dívida de modelos e dados;
- aumentar confiança no uso de inteligência artificial;
- sustentar implantação responsável de modelos analíticos.

13. Síntese da dimensão

A dimensão **MLOps** trata da gestão operacional, metodológica e governada do ciclo de vida de modelos de aprendizagem de máquina. No UnespDataLens-RM, essa dimensão atua como capacidade complementar para garantir que modelos sejam construídos, avaliados, versionados, implantados, monitorados, explicados, atualizados e descontinuados de forma controlada.

Ao incorporar registro de experimentos, versionamento de dados, features e modelos, avaliação de desempenho, fairness, explicabilidade, model registry, monitoramento de drift, auditoria de predições, critérios de retreinamento e governança de modelos, o modelo de

referência fortalece sua capacidade de transformar modelos computacionais em ativos analíticos confiáveis, rastreáveis e adequados à tomada de decisão baseada em evidências.

Dimensão Complementar — Data Valuation

1. Definição

Data Valuation refere-se ao processo de identificação, mensuração, interpretação e gestão do valor dos dados ao longo de seu ciclo de vida. Essa dimensão parte do pressuposto de que nem todos os dados possuem o mesmo valor analítico, operacional, estratégico, científico ou institucional. Em pipelines de dados, especialmente em contextos educacionais, o valor de um ativo de dados não depende apenas de sua existência ou volume, mas de sua capacidade de gerar conhecimento confiável, apoiar decisões, reduzir incertezas, melhorar modelos analíticos, orientar intervenções e produzir impacto mensurável para os atores envolvidos.

No contexto do UnespDataLens-RM, a dimensão **Data Valuation** tem como objetivo estabelecer critérios para avaliar a relevância e a utilidade dos dados utilizados nos processos de extração, integração, transformação, modelagem, explicabilidade, governança e tomada de decisão. Essa avaliação permite distinguir dados críticos de dados redundantes, dados de alto valor de dados de baixo impacto, dados estratégicos de dados meramente operacionais e dados necessários de dados excessivos ou pouco justificáveis.

Dessa forma, Data Valuation complementa dimensões como qualidade de dados, observabilidade, minimização, governança, proveniência e reprodutibilidade. Enquanto a qualidade de dados busca verificar se os dados são corretos, completos, consistentes e atualizados, a valoração de dados busca responder a uma questão adicional: **qual é a contribuição efetiva desse dado para os objetivos analíticos, científicos, educacionais ou institucionais do pipeline?**

---

2. Objetivo da dimensão

O objetivo da dimensão **Data Valuation** é fornecer mecanismos conceituais, metodológicos e operacionais para estimar o valor dos dados em função de sua utilidade, criticidade, impacto, custo, risco, reutilização e contribuição para a geração de conhecimento.

Essa dimensão busca apoiar decisões como:

- quais dados devem ser priorizados na coleta;
- quais fontes possuem maior valor analítico;
- quais atributos contribuem efetivamente para modelos preditivos ou explicativos;
- quais dados são essenciais para decisões educacionais;
- quais dados apresentam baixo valor e alto custo de manutenção;
- quais dados devem ser preservados, versionados ou descartados;
- quais ativos de dados justificam maior investimento em governança, qualidade e documentação.

Assim, Data Valuation contribui para tornar o pipeline mais racional, sustentável, auditável e orientado a valor, evitando tanto a subutilização de dados relevantes quanto a manutenção indiscriminada de dados sem finalidade clara.

---

3. Justificativa

Em projetos de Ciência de Dados, Mineração de Dados Educacionais e Learning Analytics, é comum que grandes volumes de dados sejam coletados, armazenados e processados sem uma avaliação explícita de seu valor. Essa prática pode gerar aumento de complexidade, custos computacionais, riscos de privacidade, redundância, baixa governança e dificuldade de interpretação dos resultados.

A ausência de uma dimensão de Data Valuation pode levar a problemas como:

- coleta excessiva de dados sem justificativa analítica;
- manutenção de atributos irrelevantes para o objetivo do estudo;
- dependência de fontes com baixo impacto decisório;
- aumento desnecessário de custos de armazenamento e processamento;
- maior exposição a riscos éticos e legais;
- dificuldade de explicar por que determinados dados foram utilizados;
- baixa rastreabilidade entre dados, modelos, métricas e decisões.

No domínio educacional, essa questão é ainda mais sensível, pois os dados podem envolver informações acadêmicas, socioeconômicas, comportamentais e institucionais dos estudantes. Portanto, a valoração dos dados deve considerar não apenas sua contribuição estatística ou preditiva, mas também sua pertinência pedagógica, sua legitimidade institucional, seu potencial de apoio à decisão e seus riscos associados.

Nesse sentido, Data Valuation atua como uma dimensão complementar essencial para evitar que o pipeline seja orientado apenas pela disponibilidade de dados. Em vez disso, propõe que os dados sejam avaliados por sua contribuição efetiva para a compreensão de fenômenos educacionais, para a melhoria de modelos analíticos e para o suporte a decisões responsáveis.

---

4. Escopo da dimensão

A dimensão Data Valuation abrange a avaliação do valor dos dados em diferentes níveis do pipeline:

4.1 Valor da fonte de dados

Avalia a relevância de cada fonte utilizada no pipeline, considerando sua cobertura, confiabilidade, atualidade, granularidade, custo de obtenção, facilidade de integração e contribuição para os objetivos analíticos.

Exemplos de fontes:

- sistemas acadêmicos;
- ambientes virtuais de aprendizagem;
- questionários socioeconômicos;
- registros de matrícula;
- dados de desempenho;
- dados de frequência;
- dados administrativos;
- bases públicas;
- dados longitudinais.

4.2 Valor do atributo ou variável

Avalia a contribuição de cada variável para análises, modelos, explicações ou decisões. Uma variável pode ter alto valor por ser fortemente associada ao fenômeno estudado, por melhorar o desempenho de um modelo, por favorecer a interpretação dos resultados ou por representar um fator educacionalmente relevante.

Exemplos:

- idade;
- renda familiar;
- desempenho em disciplinas iniciais;
- frequência;
- número de reprovações;
- tempo de acesso ao AVA;
- curso;
- modalidade;
- situação de matrícula.

4.3 Valor do conjunto de dados

Avalia o valor do dataset como ativo analítico, considerando completude, representatividade, potencial de reutilização, documentação, histórico, granularidade, qualidade e aplicabilidade a diferentes problemas.

4.4 Valor do dado para o modelo

Analisa a contribuição dos dados para o desempenho, estabilidade, generalização, explicabilidade e robustez dos modelos analíticos.

4.5 Valor do dado para a decisão

Avalia se o dado contribui para apoiar decisões educacionais concretas, como identificação de estudantes em risco, planejamento de intervenções, revisão curricular, políticas de permanência, ações pedagógicas ou estratégias institucionais.

4.6 Valor do dado em relação ao risco

Considera se o valor analítico de um dado justifica seu custo, sensibilidade, risco de privacidade, risco de viés, risco de uso indevido ou esforço de governança.

---

5. Componentes da dimensão Data Valuation

A dimensão pode ser organizada nos seguintes componentes:

5.1 Criticidade

Indica o grau de importância do dado para o funcionamento do pipeline ou para a tomada de decisão. Dados críticos são aqueles cuja ausência, erro ou baixa qualidade compromete significativamente os resultados analíticos.

5.2 Utilidade analítica

Representa a contribuição do dado para análises descritivas, diagnósticas, preditivas, prescritivas ou explicativas.

5.3 Impacto decisório

Avalia se o dado contribui para decisões institucionais ou pedagógicas relevantes.

5.4 Potencial explicativo

Indica se o dado ajuda a compreender os fatores associados a determinado fenômeno, como evasão, desempenho, retenção, engajamento ou permanência.

5.5 Contribuição preditiva

Mensura a contribuição do dado para a melhoria do desempenho de modelos analíticos.

5.6 Reutilização

Avalia se o dado pode ser reaproveitado em diferentes estudos, modelos, dashboards, relatórios ou processos institucionais.

5.7 Custo de obtenção e manutenção

Considera o esforço técnico, operacional, computacional e institucional necessário para coletar, tratar, armazenar, atualizar e governar o dado.

5.8 Risco associado

Avalia riscos de privacidade, sensibilidade, discriminação, viés, exposição indevida, uso inadequado ou não conformidade legal.

5.9 Temporalidade do valor

Considera que o valor dos dados pode mudar ao longo do tempo. Dados altamente relevantes em determinado período podem perder valor analítico em outro contexto, especialmente quando ocorrem mudanças curriculares, institucionais, sociais ou tecnológicas.

---

6. Relação com outras dimensões do modelo

A dimensão Data Valuation possui forte relação com outras dimensões complementares do UnespDataLens-RM.

6.1 Relação com Data Quality

Dados de alto valor precisam ter qualidade suficiente para serem utilizados. Entretanto, um dado pode ter alta qualidade técnica e baixo valor analítico. Por exemplo, um atributo pode estar completo, correto e padronizado, mas não contribuir para os objetivos da análise.

6.2 Relação com Data Minimization

A valoração dos dados apoia a minimização, pois permite identificar quais dados são realmente necessários. Dados de baixo valor e alto risco devem ser questionados quanto à sua permanência no pipeline.

6.3 Relação com Data Governance

A governança deve priorizar ativos de dados com maior valor estratégico, maior criticidade ou maior risco. Data Valuation ajuda a definir níveis diferenciados de controle, documentação, acesso e monitoramento.

6.4 Relação com Proveniência e Linhagem

A valoração depende da capacidade de rastrear a origem, transformação e uso dos dados. Sem proveniência, torna-se difícil justificar o valor atribuído a uma fonte, variável ou dataset.

6.5 Relação com Feature Engineering e Feature Store

A avaliação do valor das variáveis contribui para selecionar, documentar e reutilizar atributos relevantes em modelos analíticos.

6.6 Relação com Data Observability

A observabilidade permite monitorar se ativos de dados de alto valor continuam disponíveis, atualizados e confiáveis ao longo do tempo.

6.7 Relação com Fairness e Bias

Dados de alto valor preditivo podem apresentar riscos de viés ou discriminação. Portanto, a valoração não deve considerar apenas desempenho, mas também implicações éticas e sociais.

---

7. Critérios de avaliação do valor dos dados

A avaliação do valor dos dados pode considerar os seguintes critérios:

Critério	Descrição	Pergunta orientadora
Relevância analítica	Grau de contribuição para os objetivos da análise	Este dado ajuda a responder ao problema investigado?
Contribuição preditiva	Impacto no desempenho dos modelos	A presença desse dado melhora a predição?
Potencial explicativo	Capacidade de apoiar interpretação	Este dado ajuda a explicar o fenômeno?
Utilidade decisória	Contribuição para decisões práticas	Este dado apoia alguma decisão concreta?

Criticidade	Impacto da ausência ou erro do dado	O pipeline é prejudicado sem esse dado?
Reutilização	Potencial de uso em outros processos	Este dado pode ser reaproveitado?
Custo	Esforço de coleta, tratamento e manutenção	O custo de manter esse dado é justificável?
Risco	Sensibilidade, privacidade ou viés	O uso desse dado pode gerar riscos?
Temporalidade	Manutenção do valor ao longo do tempo	Este dado continua relevante com o passar do tempo?
Governabilidade	Facilidade de controle e documentação	É possível governar esse dado adequadamente?

8. Métricas associadas

A dimensão Data Valuation pode ser operacionalizada por meio de métricas quantitativas, qualitativas ou híbridas.

8.1 Índice de Valor Analítico do Dado — IVAD

Representa o grau de contribuição de um dado para o objetivo analítico do pipeline.

$$
IVAD(d) = \frac{R_a(d) + U_d(d) + P_e(d) + C_p(d)}{4}
$$

Em que:

- $R_a(d)$ = relevância analítica do dado;
- $U_d(d)$ = utilidade decisória;
- $P_e(d)$ = potencial explicativo;
- $C_p(d)$ = contribuição preditiva.

Cada componente pode ser avaliado em escala de 0 a 1 ou de 1 a 5.

8.2 Índice de Custo-Risco do Dado — ICRD

Representa o custo e o risco associados ao uso de um dado.

$$
ICRD(d) = \frac{C_o(d) + R_p(d) + R_v(d) + E_g(d)}{4}
$$

Em que:

- $C_o(d)$ = custo operacional;
- $R_p(d)$ = risco de privacidade;
- $R_v(d)$ = risco de viés;
- $E_g(d)$ = esforço de governança.

8.3 Índice de Valor Líquido do Dado — IVLD

Representa a relação entre valor analítico e custo-risco.

$$
IVLD(d) = IVAD(d) - ICRD(d)
$$

Quanto maior o IVLD, maior a justificativa para uso e manutenção do dado no pipeline.

8.4 Taxa de Reutilização do Dado — TRD

Mede a frequência com que um dado é reutilizado em diferentes análises, modelos ou produtos analíticos.

$$
TRD(d) = \frac{N_{usos}(d)}{N_{processos}}
$$

Em que:

- $N_{usos}(d)$ = número de processos em que o dado é utilizado;
- $N_{processos}$ = número total de processos analíticos avaliados.

8.5 Índice de Criticidade do Dado — ICD

Mede o impacto da ausência ou degradação do dado no pipeline.

$$
ICD(d) = \frac{I_m(d) + I_d(d) + I_o(d)}{3}
$$

Em que:

- $I_m(d)$ = impacto no modelo;
- $I_d(d)$ = impacto na decisão;
- $I_o(d)$ = impacto operacional.

---

9. Classificação dos dados por valor

A partir dos critérios e métricas, os dados podem ser classificados em quatro categorias.

9.1 Dados de alto valor estratégico

São dados com alta relevância analítica, alto impacto decisório, forte potencial explicativo ou elevada contribuição para modelos. Devem ser priorizados em processos de qualidade, governança, documentação, versionamento e monitoramento.

Exemplo: desempenho acadêmico inicial em estudos de evasão, quando comprovadamente associado ao abandono.

9.2 Dados de valor operacional

São dados necessários para funcionamento do pipeline, integração de bases, identificação de registros ou controle institucional, ainda que não tenham alta contribuição analítica direta.

Exemplo: identificadores internos, códigos de curso, códigos de turma e datas de matrícula.

9.3 Dados de valor contextual

São dados que ajudam a interpretar resultados, segmentar análises ou contextualizar fenômenos, mas que isoladamente podem ter baixo poder preditivo.

Exemplo: modalidade de ensino, turno, polo, ano de ingresso ou tipo de escola de origem.

9.4 Dados de baixo valor ou valor questionável

São dados com pouca contribuição analítica, baixo uso, alto custo, alto risco ou ausência de justificativa clara. Devem ser revisados, descartados ou mantidos apenas mediante justificativa explícita.

Exemplo: atributos coletados por disponibilidade, mas sem relação com os objetivos do estudo ou com decisões institucionais.

---

10. Processo de aplicação da dimensão

A aplicação da dimensão Data Valuation pode seguir o seguinte processo:

Etapa 1 — Inventariar os ativos de dados

Identificar fontes, datasets, tabelas, atributos, metadados, variáveis derivadas e produtos analíticos utilizados no pipeline.

Etapa 2 — Associar dados aos objetivos do pipeline

Relacionar cada dado aos objetivos analíticos, modelos, dashboards, decisões ou indicadores que ele apoia.

Etapa 3 — Avaliar valor analítico e decisório

Mensurar ou qualificar a contribuição do dado para análise, explicação, predição, decisão e reutilização.

Etapa 4 — Avaliar custo, risco e governabilidade

Identificar esforço de coleta e manutenção, riscos de privacidade, riscos de viés, sensibilidade e exigências de governança.

Etapa 5 — Classificar os dados por valor

Categorizar os dados em alto valor estratégico, valor operacional, valor contextual ou baixo valor.

Etapa 6 — Definir ações de gestão

Decidir se o dado deve ser priorizado, documentado, monitorado, enriquecido, mantido, descartado, anonimizado, restringido ou revisado.

Etapa 7 — Monitorar a evolução do valor

Reavaliar periodicamente o valor dos dados, considerando mudanças no contexto, nos modelos, nos objetivos institucionais e nas políticas educacionais.

---

11. Artefatos relacionados

Artefato	Finalidade
Inventário de ativos de dados	Lista fontes, tabelas, atributos e datasets disponíveis
Matriz de valor dos dados	Relaciona cada dado a critérios de valor, custo e risco
Catálogo de dados valorados	Registra a classificação dos dados segundo seu valor
Relatório de criticidade	Identifica dados essenciais para o pipeline
Matriz dado-decisão	Relaciona dados às decisões educacionais apoiadas
Matriz dado-modelo	Relaciona dados aos modelos analíticos que os utilizam
Relatório de custo-risco	Documenta riscos, custos e exigências de governança

Plano de priorização de dados	Define quais ativos devem receber maior atenção
Histórico de reavaliação	Registra mudanças no valor dos dados ao longo do tempo

12. Exemplo aplicado ao contexto educacional

Em um estudo sobre evasão estudantil, diferentes dados podem apresentar valores distintos.

O atributo “desempenho no primeiro semestre” pode ter alto valor analítico e preditivo, pois pode estar fortemente associado à evasão e contribuir para identificar estudantes em risco. Além disso, possui valor decisório, pois pode orientar ações de acompanhamento acadêmico.

O atributo “renda familiar” pode ter valor contextual e explicativo, pois ajuda a compreender desigualdades e vulnerabilidades, mas também exige cuidado ético e governança adequada, por envolver informação sensível.

O atributo “cor favorita do estudante”, caso eventualmente coletado, tende a ter baixo valor para o problema de evasão, pois dificilmente contribui para análise, predição, explicação ou decisão. Nesse caso, sua manutenção no pipeline seria questionável.

Já o atributo “identificador do estudante” possui valor operacional, pois permite integração e rastreamento dos registros, mas deve ser protegido, anonimizado ou pseudonimizado em etapas analíticas.

Esse exemplo mostra que a valoração dos dados não se limita à contribuição estatística. Ela envolve a relação entre utilidade, custo, risco, explicabilidade, governança e finalidade educacional.

---

13. Riscos e desafios

A implementação da dimensão Data Valuation apresenta alguns desafios:

- dificuldade de mensurar valor de forma objetiva;
- dependência do contexto institucional;
- variação do valor dos dados ao longo do tempo;
- risco de supervalorizar dados com alto poder preditivo, mas baixo significado educacional;
- risco de subvalorizar dados contextuais importantes;
- conflito entre valor analítico e risco ético;
- necessidade de participação de especialistas de domínio;
- dificuldade de comparar dados de naturezas diferentes;
- ausência de metadados suficientes para avaliar valor;

- risco de decisões automatizadas baseadas em dados mal interpretados.

Por isso, Data Valuation deve combinar critérios quantitativos e qualitativos, envolvendo tanto métricas computacionais quanto julgamento especializado.

---

14. Boas práticas

A adoção da dimensão Data Valuation deve considerar as seguintes boas práticas:

1. atribuir valor aos dados sempre em relação a um objetivo explícito;
2. evitar considerar volume de dados como sinônimo de valor;
3. documentar por que cada dado é utilizado;
4. revisar periodicamente dados de baixo uso ou baixo impacto;
5. avaliar conjuntamente valor, custo e risco;
6. priorizar governança para dados críticos e sensíveis;
7. envolver especialistas educacionais na interpretação do valor;
8. distinguir valor preditivo de valor pedagógico;
9. considerar a possibilidade de descarte ou minimização de dados sem finalidade clara;
10. manter registro histórico das decisões de valoração.

---

15. Contribuição para o UnespDataLens-RM

A dimensão Data Valuation contribui para o UnespDataLens-RM ao introduzir uma perspectiva orientada a valor na gestão dos dados educacionais. Ela permite que o pipeline não seja avaliado apenas pela quantidade de dados processados ou pelo desempenho dos modelos, mas também pela relevância, utilidade, criticidade, explicabilidade, sustentabilidade e responsabilidade no uso dos dados.

Ao incorporar Data Valuation, a arquitetura passa a oferecer suporte mais robusto para decisões como seleção de fontes, priorização de variáveis, descarte de dados redundantes, investimento em qualidade, definição de políticas de governança e justificativa do uso de dados sensíveis.

Essa dimensão reforça o caráter técnico-metodológico do UnespDataLens-RM, pois transforma a gestão dos dados em um processo avaliável, documentado, rastreável e orientado à geração de valor educacional. Com isso, contribui para que os ativos de dados sejam tratados não apenas como insumos computacionais, mas como recursos estratégicos para a produção de conhecimento, a melhoria dos processos educacionais e o apoio à tomada de decisão baseada em evidências.

---

16. Síntese da dimensão

A dimensão **Data Valuation** estabelece critérios para determinar o valor dos dados em um pipeline analítico educacional. Seu foco está em avaliar quanto cada fonte, atributo, dataset ou produto de dados contribui para análise, predição, explicação, decisão, reutilização e impacto institucional, considerando simultaneamente custos, riscos e exigências de governança.

Em síntese, essa dimensão responde às seguintes perguntas:

- Este dado é necessário?
- Este dado contribui para o objetivo analítico?
- Este dado melhora a explicação ou a predição?
- Este dado apoia alguma decisão educacional?
- O valor desse dado justifica seu custo e seu risco?
- Este dado deve ser priorizado, mantido, monitorado, restringido ou descartado?

Com isso, Data Valuation fortalece a capacidade do UnespDataLens-RM de produzir pipelines mais eficientes, responsáveis, sustentáveis e orientados à geração de valor.

Essa dimensão ficou alinhada à lógica de artefatos metodológicos descrita no guia de DSR, especialmente na exigência de explicitar componentes, relações, procedimentos, critérios, entradas, saídas e mecanismos internos do artefato.

Dimensão Complementar — MLOps

1. Definição da dimensão

A dimensão **MLOps** refere-se ao conjunto de princípios, práticas, processos e mecanismos técnicos voltados à gestão do ciclo de vida de modelos de aprendizagem de máquina, desde a preparação dos dados e construção de features até o treinamento, avaliação, versionamento, implantação, monitoramento, explicação, governança, atualização e descontinuação de modelos.

No contexto do **UnespDataLens-RM**, MLOps representa a capacidade de operacionalizar modelos analíticos, preditivos, classificatórios, explicativos ou prescritivos de forma controlada, reprodutível, rastreável, monitorável e governada. Essa dimensão não se restringe ao treinamento de algoritmos. Ela envolve todo o ecossistema que permite transformar modelos em artefatos confiáveis, avaliáveis e integrados aos pipelines de Engenharia Analítica de Dados.

MLOps inclui práticas como versionamento de datasets de treinamento, controle de features, registro de experimentos, comparação de modelos, validação de desempenho, análise de fairness, explicabilidade, empacotamento de modelos, implantação controlada, monitoramento de drift, avaliação contínua, auditoria de decisões, atualização periódica e descontinuação de modelos obsoletos.

No UnespDataLens-RM, essa dimensão é complementar e transversal porque modelos dependem dos dados, metadados, qualidade, rótulos, features, linhagem, versionamento, governança, monitoramento e agentes inteligentes produzidos pelos demais módulos do modelo de referência. Um modelo não deve ser visto como um componente isolado, mas como um ativo analítico derivado de dados, regras, escolhas metodológicas, parâmetros, versões e contextos de uso.

2. Problema tratado

Em muitos projetos de Ciência de Dados e Inteligência Artificial, modelos são desenvolvidos em notebooks, scripts ou experimentos pontuais, mas sem controle adequado sobre os dados utilizados, as versões das features, os parâmetros, o ambiente de execução, as métricas, os critérios de validação, os riscos de viés, as limitações ou as condições de implantação.

Essa situação pode gerar modelos difíceis de reproduzir, comparar, auditar, atualizar ou explicar. Um modelo pode apresentar bom desempenho experimental, mas não estar preparado para uso contínuo, pois depende de dados instáveis, features sem versionamento, rótulos frágeis, métricas incompletas, ausência de monitoramento ou falta de governança.

A ausência de MLOps pode gerar problemas como:

- experimentos não reprodutíveis;
- modelos treinados com datasets sem versão;
- uso de features não documentadas;
- parâmetros e hiperparâmetros não registrados;
- ausência de rastreabilidade entre modelo, dados e código;
- comparação inadequada entre modelos;

- implantação de modelos sem validação suficiente;
- degradação de desempenho não detectada;
- data drift, feature drift, label drift ou concept drift sem monitoramento;
- ausência de explicabilidade ou interpretabilidade;
- baixa avaliação de fairness e vieses;
- dificuldade de auditar decisões apoiadas por modelos;
- modelos obsoletos mantidos em uso;
- falta de critérios para retreinamento;
- risco de uso indevido de predições em processos decisórios.

No UnespDataLens-RM, a dimensão MLOps busca evitar que modelos sejam tratados apenas como resultados de experimentos, formalizando-os como ativos analíticos versionados, monitorados, avaliados, governados e integrados ao ciclo de vida dos dados.

3. Objetivo da dimensão

O objetivo da dimensão **MLOps** é **estruturar práticas para construção, avaliação, versionamento, implantação, monitoramento, governança e manutenção de modelos de aprendizagem de máquina no ecossistema UnespDataLens-RM**, garantindo que esses modelos sejam reprodutíveis, rastreáveis, explicáveis, avaliáveis, atualizáveis e adequados ao uso analítico ou decisório.

Essa dimensão busca assegurar que modelos derivados dos dados do UnespDataLens-RM sejam tecnicamente robustos, metodologicamente documentados, operacionalmente monitorados e institucionalmente responsáveis.

4. Relação com o UnespDataLens-RM

A dimensão MLOps se articula com diversos módulos do UnespDataLens-RM.

No **Módulo 1 — Fontes e Inventário de Dados**, MLOps depende da identificação das fontes que alimentam os modelos, seus responsáveis, limitações, periodicidade, granularidade e restrições de uso.

No **Módulo 2 — Extração, Ingestão e Armazenamento Bruto**, garante que os dados utilizados no treinamento e avaliação dos modelos possam ser rastreados até sua origem e versão inicial.

No **Módulo 3 — Integração e Harmonização**, apoia a construção de bases integradas para modelagem, preservando regras de relacionamento, chaves, conflitos e limitações.

No **Módulo 4 — Transformação, Limpeza, Enriquecimento e Preparação dos Dados**, registra transformações aplicadas aos dados antes da modelagem, incluindo tratamento de ausentes, normalizações, recodificações e enriquecimentos.

No **Módulo 5 — Qualidade e Validação de Dados**, utiliza métricas de qualidade para avaliar se os dados são adequados ao treinamento, validação e teste dos modelos.

No **Módulo 6 — Desbalanceamento, Vieses e Equidade**, incorpora avaliação de distribuição de classes, representatividade, métricas de fairness, mitigação de viés e análise de impacto entre grupos.

No **Módulo 7 — Armazenamento Analítico**, utiliza datasets curados, data marts e bases analíticas como insumos controlados para treinamento e avaliação.

No **Módulo 8 — Disponibilização e Consumo Analítico**, modelos podem gerar scores, classificações, recomendações, indicadores ou serviços analíticos consumidos por dashboards, APIs ou relatórios.

No **Módulo 9 — Metadados, Catálogo e Semântica dos Dados**, modelos devem ser descritos por metadados, incluindo finalidade, dados de entrada, variável-alvo, features, métricas, limitações e responsáveis.

No **Módulo 10 — Governança, Segurança, Privacidade e Conformidade**, MLOps deve respeitar políticas de acesso, finalidade de uso, dados sensíveis, privacidade, auditoria e responsabilidade institucional.

No **Módulo 11 — Proveniência, Linhagem e Rastreabilidade**, registra a relação entre modelo, datasets, features, rótulos, código, parâmetros, versões e produtos derivados.

No **Módulo 12 — Versionamento e Reprodutibilidade**, controla versões de dados, features, código, experimentos, modelos, ambientes, parâmetros e resultados.

No **Módulo 13 — Representações Analíticas**, modelos podem utilizar representações tabulares, temporais, relacionais, grafos ou grafos temporais como entrada.

No **Módulo 14 — Feature Engineering e Feature Store**, MLOps depende do controle de features, regras de cálculo, versionamento, qualidade, drift e risco de vazamento.

No **Módulo 15 — Monitoramento, Drift e Evolução dos Ativos**, acompanha degradação de desempenho, drift dos dados, drift das features, drift dos rótulos, mudanças no comportamento do modelo e necessidade de retreinamento.

No **Módulo 16 — Agentes Inteligentes de Apoio ao Pipeline**, agentes podem apoiar explicação de modelos, consulta a experimentos, interpretação de métricas, identificação de drift e recomendação de retreinamento.

5. Componentes de MLOps

A dimensão MLOps pode ser estruturada em componentes que cobrem o ciclo completo de modelos no UnespDataLens-RM.

Componente	Descrição	Exemplo
Dataset de treinamento	Base utilizada para treinar o modelo, com versão e qualidade registradas.	Dataset curado de evasão v2.1.
Dataset de validação e teste	Bases separadas para avaliação do desempenho e generalização.	Conjunto de validação temporal.
Feature set	Conjunto de atributos utilizados pelo modelo.	Features acadêmicas, temporais e socioeconômicas.
Variável-alvo ou	Desfecho que o modelo busca	Evasão no próximo período.

rótulo	prever, classificar ou estimar.	
Experimento	Execução controlada de treinamento e avaliação.	XGBoost com parâmetros específicos.
Modelo treinado	Artefato resultante do processo de aprendizagem.	Modelo preditivo salvo e versionado.
Métricas de avaliação	Indicadores de desempenho e qualidade do modelo.	Acurácia, F1, AUC, recall, precisão.
Métricas de fairness	Indicadores de equidade entre grupos.	SPD, DI, EOD, AOD, FPR e FNR por grupo.
Explicabilidade	Métodos para interpretação do comportamento do modelo.	SHAP, LIME, importância de variáveis.
Registro de modelo	Catálogo ou repositório de modelos e versões.	Model registry.
Implantação	Disponibilização do modelo para uso analítico ou operacional.	API de score de risco.
Monitoramento	Acompanhamento de desempenho, drift, uso e estabilidade.	Alerta de queda de recall.
Retreinamento	Atualização do modelo com novos dados ou novas condições.	Retreinamento semestral.
Descontinuação	Retirada controlada de modelos obsoletos ou inadequados.	Modelo substituído por nova versão.

6. Estratégias de implementação

A implementação de MLOps no UnespDataLens-RM pode envolver estratégias técnicas, metodológicas e de governança.

Estratégia	Descrição
Registro de experimentos	Documentar dados, features, rótulos, algoritmos, parâmetros, métricas e resultados.
Versionamento de modelos	Controlar versões de modelos treinados, seus artefatos e configurações.
Versionamento de datasets	Registrar versões dos dados usados em treinamento, validação, teste e produção.
Versionamento de features	Controlar regras, janelas temporais, parâmetros e versões das features.

Separação treino-validação-teste	Definir protocolos de avaliação adequados ao problema e ao tempo.
Validação temporal	Evitar vazamento temporal e avaliar generalização em períodos futuros.
Avaliação de desempenho	Medir métricas adequadas ao problema, como acurácia, F1, AUC, RMSE ou MAE.
Avaliação de fairness	Avaliar desempenho e erros por grupos relevantes ou sensíveis.
Explicabilidade do modelo	Utilizar métodos para explicar variáveis, decisões, predições e limitações.
Registro de limitações	Documentar escopo, contexto, restrições e riscos do modelo.
Governança de implantação	Estabelecer critérios para publicação, uso, revisão e aprovação de modelos.
Monitoramento em produção	Acompanhar drift, desempenho, estabilidade, uso, latência e falhas.
Critérios de retreinamento	Definir quando o modelo deve ser atualizado ou substituído.
Auditoria de decisões	Registrar predições, versões, entradas, saídas e explicações quando aplicável.
Descontinuação controlada	Retirar modelos obsoletos preservando histórico, justificativa e versões anteriores.

7. Métricas associadas

A dimensão MLOps pode ser avaliada por métricas relacionadas à reprodutibilidade, desempenho, fairness, estabilidade, implantação, monitoramento e governança dos modelos.

Métrica	Sigla	Descrição
Taxa de modelos registrados	TMR	Proporção de modelos com registro formal no catálogo ou model registry.
Taxa de modelos versionados	TMV	Proporção de modelos com controle de versão.
Taxa de experimentos reprodutíveis	TER	Proporção de experimentos com dados, código, parâmetros e ambiente registrados.
Taxa de modelos com dataset versionado	TMDV	Proporção de modelos associados a versões dos datasets utilizados.

Taxa de modelos com feature set documentado	TMFD	Proporção de modelos com features, regras e versões documentadas.
Taxa de modelos com rótulo documentado	TMRD	Proporção de modelos com variável-alvo formalmente definida.
Taxa de modelos com avaliação de desempenho	TMAD	Proporção de modelos com métricas de desempenho registradas.
Taxa de modelos com avaliação de fairness	TMAF	Proporção de modelos avaliados quanto a equidade entre grupos.
Taxa de modelos explicáveis	TME	Proporção de modelos com relatório de explicabilidade ou interpretabilidade.
Taxa de modelos monitorados	TMM	Proporção de modelos em uso com monitoramento ativo.
Taxa de modelos com drift detectado	TMDD	Proporção de modelos com evidência de drift nos dados, features, rótulos ou desempenho.
Taxa de modelos retreinados conforme critério	TMRC	Proporção de modelos atualizados conforme regra de retreinamento.
Taxa de modelos obsoletos	TMO	Proporção de modelos sem uso, sem monitoramento ou superados por nova versão.
Índice de maturidade MLOps	IMLOps	Indicador composto da maturidade operacional e governança dos modelos.

Uma formulação possível para o **Índice de Maturidade MLOps** é:

$$\mathbf{IMLOps} = \mathbf{w_1TMR} + \mathbf{w_2TMV} + \mathbf{w_3TER} + \mathbf{w_4TMDV} + \mathbf{w_5TMFD} + \mathbf{w_6TMRD} + \mathbf{w_7TMAD} + \mathbf{w_8TMAF} + \mathbf{w_9TME} + \mathbf{w_{10}TMM} + \mathbf{w_{11}TMRC} - \mathbf{w_{12}TMO}$$

em que:

- **TMR** representa a taxa de modelos registrados;
- **TMV** representa a taxa de modelos versionados;
- **TER** representa a taxa de experimentos reprodutíveis;
- **TMDV** representa a taxa de modelos com dataset versionado;
- **TMFD** representa a taxa de modelos com feature set documentado;
- **TMRD** representa a taxa de modelos com rótulo documentado;
- **TMAD** representa a taxa de modelos com avaliação de desempenho;
- **TMAF** representa a taxa de modelos com avaliação de fairness;
- **TME** representa a taxa de modelos explicáveis;
- **TMM** representa a taxa de modelos monitorados;
- **TMRC** representa a taxa de modelos retreinados conforme critério;
- **TMO** representa a taxa de modelos obsoletos;
- $\mathbf{w_i}$ representa o peso atribuído a cada dimensão.

Quanto maior o IMLOps, maior a maturidade do ecossistema na gestão do ciclo de vida de modelos.

8. Artefatos relacionados

A dimensão MLOps pode ser operacionalizada por meio de artefatos complementares, integrados aos módulos de qualidade, feature store, versionamento, linhagem, governança, monitoramento, explicabilidade e agentes inteligentes.

Código	Artefato	Representação	Finalidade
MLO1	Arquitetura MLOps	Documento técnico ou diagrama	Definir componentes, fluxos, ambientes, registros, integrações e responsabilidades do ciclo de modelos.
MLO2	Catálogo de Modelos	Model registry ou tabela de modelos	Registrar modelos, versões, finalidade, responsável, status, métricas e restrições.
MLO3	Registro de Experimentos	Tabela ou ferramenta de tracking	Documentar datasets, features, rótulos, algoritmos, parâmetros, métricas e resultados.
MLO4	Manifesto de Treinamento	Documento reprodutível	Registrar dados, versões, código, ambiente, parâmetros, bibliotecas, seeds e artefatos gerados.
MLO5	Relatório de Avaliação de Desempenho	Relatório técnico	Avaliar métricas de desempenho, erro, estabilidade, validação temporal e comparação entre modelos.
MLO6	Relatório de Fairness do Modelo	Relatório técnico	Avaliar desempenho, erros e impactos diferenciais entre grupos.
MLO7	Relatório de Explicabilidade	Relatório técnico	Explicar variáveis relevantes, predições, comportamento global e limitações do modelo.
MLO8	Registro de Implantação do Modelo	Documento ou tabela operacional	Documentar versão implantada, ambiente, API, data, responsável, status e critérios de uso.
MLO9	Painel de Monitoramento de Modelos	Dashboard	Monitorar drift, desempenho, latência, uso, falhas e estabilidade do modelo.
MLO10	Registro de Retreinamento	Changelog ou tabela de atualização	Documentar critérios, datas, versões, motivos e resultados de retreinamento.

MLO11	Registro de Predições e Auditoria	Log governado	Registrar entradas, saídas, versão do modelo, explicações e contexto quando aplicável.
MLO12	Plano de Descontinuação de Modelo	Documento operacional	Definir critérios, justificativa, impacto e substituição de modelos obsoletos.
MLO13	Model Card	Documento padronizado	Sintetizar finalidade, dados, desempenho, limitações, riscos, fairness, uso recomendado e uso não recomendado.
MLO14	Plano de Governança de Modelos	Documento normativo	Definir papéis, aprovações, responsabilidades, restrições, auditoria e revisão dos modelos.

9. Critérios de validação

A dimensão MLOps pode ser considerada adequadamente tratada quando:

1. modelos críticos estão catalogados;
2. modelos possuem finalidade, responsável, versão e status definidos;
3. datasets de treinamento, validação e teste estão versionados;
4. features utilizadas pelos modelos estão documentadas e rastreáveis;
5. variáveis-alvo ou rótulos possuem definição e regra de construção;
6. experimentos registram algoritmos, parâmetros, ambiente e resultados;
7. métricas de desempenho são calculadas e comparadas;
8. modelos de alto impacto possuem avaliação de fairness;
9. modelos utilizados em decisões possuem explicabilidade ou justificativa de interpretabilidade;
10. modelos implantados possuem registro de versão e ambiente;
11. modelos em uso são monitorados quanto a drift, desempenho e estabilidade;
12. critérios de retreinamento estão definidos;
13. predições relevantes podem ser auditadas;
14. modelos obsoletos são descontinuados de forma controlada;
15. limitações, riscos e usos não recomendados são documentados.

10. Exemplo de aplicação no domínio educacional

Em um projeto educacional, MLOps pode ser aplicado ao desenvolvimento e manutenção de modelos preditivos de evasão, desempenho, risco acadêmico, engajamento ou recomendação de apoio pedagógico.

Por exemplo, um modelo de evasão pode ser treinado a partir de dados de matrícula, histórico acadêmico, desempenho, frequência, interações no ambiente virtual e perfil socioeconômico. A dimensão MLOps exige que o dataset utilizado seja versionado, que o rótulo de evasão seja formalmente definido, que as features sejam documentadas, que o experimento registre parâmetros e métricas, que o modelo seja avaliado por período e por grupos, que suas principais variáveis sejam explicadas e que sua implantação seja monitorada.

Se a distribuição dos estudantes mudar, se novas regras acadêmicas forem adotadas ou se o desempenho do modelo cair em determinado grupo, o monitoramento MLOps deve gerar alertas e orientar revisão ou retreinamento.

11. Exemplo de aplicação em políticas públicas

Em políticas públicas, MLOps pode apoiar modelos de priorização de atendimento, previsão de demanda, classificação de vulnerabilidade, detecção de risco, alocação de recursos ou avaliação de programas.

Por exemplo, um modelo de priorização territorial pode usar dados de saúde, educação, assistência social, renda e infraestrutura. Nesse caso, MLOps exige rastreabilidade entre dados, features, rótulo e predição, além de avaliação de desempenho, fairness entre territórios ou grupos populacionais, explicabilidade dos fatores associados ao score e monitoramento de drift.

Como esses modelos podem influenciar decisões públicas, a governança deve definir limites de uso, revisão humana, auditoria das predições e documentação clara sobre riscos e limitações.

12. Contribuição para o modelo de referência

A dimensão **MLOps** contribui para o UnespDataLens-RM ao formalizar modelos de aprendizagem de máquina como ativos analíticos que precisam ser versionados, documentados, monitorados, auditados e governados.

Essa dimensão fortalece o modelo de referência ao:

- aumentar reprodutibilidade de experimentos;
- melhorar controle sobre datasets, features e rótulos;
- apoiar comparação sistemática de modelos;
- incorporar fairness e explicabilidade ao ciclo de modelagem;
- reduzir risco de modelos obsoletos ou degradados;
- permitir monitoramento de drift e desempenho;
- apoiar decisões sobre retreinamento;
- integrar modelos a pipelines de dados governados;
- fortalecer auditoria de predições;
- reduzir dívida de modelos e dados;
- aumentar confiança no uso de inteligência artificial;
- sustentar implantação responsável de modelos analíticos.

13. Síntese da dimensão

A dimensão **MLOps** trata da gestão operacional, metodológica e governada do ciclo de vida de modelos de aprendizagem de máquina. No UnespDataLens-RM, essa dimensão atua como capacidade complementar para garantir que modelos sejam construídos, avaliados, versionados, implantados, monitorados, explicados, atualizados e descontinuados de forma controlada.

Ao incorporar registro de experimentos, versionamento de dados, features e modelos, avaliação de desempenho, fairness, explicabilidade, model registry, monitoramento de drift, auditoria de predições, critérios de retreinamento e governança de modelos, o modelo de

referência fortalece sua capacidade de transformar modelos computacionais em ativos analíticos confiáveis, rastreáveis e adequados à tomada de decisão baseada em evidências.