Projeto: databricks-lakehouse-case

Evidências Técnicas — Databricks Lakehouse Case

Este documento reúne as principais evidências técnicas do projeto,
incluindo arquitetura implementada, governança, qualidade de dados,
camada analítica e validações executadas no ambiente Databricks.

---

00 — Observabilidade e Auditoria de Ingestão

Implementação de estrutura de auditoria operacional
para rastreamento completo das cargas da camada Bronze.

Tabela:

- control.ingestion_log

Objetivos da implementação:

- rastrear arquivos processados
- controlar tabelas de destino
- registrar volume carregado
- identificar falhas de ingestão
- armazenar timestamps operacionais
- permitir troubleshooting e replay
- registrar schema detectado
- garantir observabilidade do pipeline

Campos monitorados:

- source_file
- target_table
- records_loaded
- pipeline_execution_timestamp
- source_file_last_modified
- source_reference_date
- bronze_load_timestamp
- status
- validation_message
- detected_schema

A implementação da ingestion_log permite
rastreabilidade completa das execuções,
além de suportar análises operacionais
e troubleshooting de pipelines.

01 — Estruturas da Camada Gold

Validação da camada analítica final
com dimensões, fato e métricas agregadas.

Objetivo:

- disponibilizar dados analíticos organizados
- suportar consumo BI
- facilitar consultas executivas
- estruturar semantic layer

Tabelas implementadas:

- dim_cliente
- dim_produto
- dim_data
- dim_canal
- dim_regiao
- dim_vendedor
- fato_pedidos
- metricas_por_periodo
- metricas_por_categoria
- metricas_por_canal
- metricas_por_regiao
- metricas_operacionais
- vw_pedidos_analytics

---

02 — Quality Log

Implementação de estrutura de governança
e monitoramento da qualidade dos dados.

Objetivos:

- registrar problemas identificados
- armazenar regras executadas
- medir impacto operacional
- suportar troubleshooting

Principais verificações:

- status nulo
- campos obrigatórios ausentes
- regional inválida
- eventos inconsistentes
- itens órfãos
- ausência de correspondência dimensional

Tabela:

- silver.quality_log

---

03 — Resumo do Quality Log

Resumo consolidado das validações aplicadas
durante o processamento das tabelas Silver e Gold.

Indicadores:

- quantidade de checks executados
- total de registros afetados
- tabelas com maior incidência de inconsistências

Objetivo:

- evidenciar governança aplicada
- demonstrar rastreabilidade operacional
- suportar monitoramento contínuo

---

04 — Detecção de Replay / Duplicidade

Validação aplicada para garantir
ausência de duplicidade no cabeçalho dos pedidos.

Objetivo:

- evitar replay de ingestão
- impedir duplicidade operacional
- validar integridade do pipeline

Resultado:

- nenhuma duplicidade encontrada

---

05 — Granularidade da Fato

Validação da grain da tabela fato_pedidos.

Grão definido:

- 1 linha por item do pedido

Validação executada:

- comparação entre total de registros
  e chave composta:
  (order_id + item_seq)

Resultado:

- granularidade íntegra
- ausência de duplicidade

---

06 — Volume da Fato

Validação do volume final da tabela fato_pedidos.

Objetivos:

- confirmar carga completa
- validar consistência operacional
- evidenciar volume tratado

Resultado:

- 912 registros processados

---

07 — Integridade das Dimensões

Validação das surrogate keys
das tabelas dimensionais.

Objetivos:

- garantir unicidade
- evitar duplicidades
- validar integridade dimensional

Resultado:

- todas as SKs únicas

Dimensões validadas:

- dim_cliente
- dim_produto
- dim_data

---

08 — Receita por Período

Validação das métricas executivas
agregadas por período.

Indicadores:

- receita bruta
- receita líquida
- descontos
- ticket médio

Objetivos:

- suportar análises executivas
- evidenciar camada analítica
- demonstrar capacidade de agregação

Tabela:

- gold.metricas_por_periodo

---

09 — Métricas Operacionais

Camada analítica operacional
com indicadores de eficiência logística.

Indicadores:

- taxa de cancelamento
- taxa de atraso
- análises por canal
- análises por região

Tabela:

- gold.metricas_operacionais

Objetivo:

- demonstrar visão operacional
- suportar indicadores de SLA

---

10 — Semantic Layer

Implementação da camada semântica
para consumo analítico simplificado.

View criada:

- gold.vw_pedidos_analytics

Objetivos:

- simplificar consultas BI
- reduzir complexidade analítica
- centralizar joins de negócio
- facilitar integração com dashboards

A semantic layer consolida:

- clientes
- produtos
- regiões
- canais
- métricas financeiras
- status operacionais

---

11 — Integridade da Semantic Layer

Validação da consistência da view analítica final.

Verificações:

- clientes sem correspondência
- produtos sem correspondência
- canais sem correspondência
- regiões sem correspondência
- datas inválidas

Resultado:

- sem problemas de calendário
- inconsistências identificadas e rastreadas

Objetivo:

- evidenciar governança analítica
- garantir confiabilidade do consumo BI

---

12 — Dimensão Calendário

Validação da cobertura temporal
da dimensão de datas.

Objetivos:

- garantir cobertura completa

- suportar análises temporais
- validar calendário analítico

Resultado:

- período validado com sucesso

Tabela:

- gold.dim_data

---

13 — Top Produtos

Consulta analítica demonstrando
os produtos com maior geração de receita.

Objetivos:

- evidenciar uso da semantic layer
- demonstrar consumo analítico final
- validar métricas de negócio

Indicadores:

- produto
- categoria
- receita líquida

---

14 — Resumo Executivo do Ambiente

Consolidação técnica do ambiente Lakehouse.

Indicadores consolidados:

- total de registros da fato
- granularidade validada
- total de clientes
- total de produtos
- receita total
- total de validações executadas
- registros tratados
- registros disponíveis na semantic layer

Objetivo:

- fornecer visão executiva do projeto
- consolidar principais métricas técnicas
- evidenciar maturidade da solução

Considerações Técnicas Finais

O projeto foi desenvolvido seguindo arquitetura Lakehouse utilizando Databricks, Delta Lake e processamento distribuído com Apache Spark.

Principais capacidades implementadas:

- arquitetura Bronze / Silver / Gold
- controle de ingestão
- auditoria operacional
- quality checks automatizados
- quality log centralizado
- modelagem dimensional
- camada semântica analítica
- métricas executivas
- métricas operacionais
- troubleshooting
- replay prevention
- governança de dados
- observabilidade operacional

O projeto priorizou:

- rastreabilidade
- governança
- confiabilidade
- organização analítica
- reutilização
- clareza operacional
- separação de responsabilidades entre camadas