Projeto: databricks-lakehouse-case

Orquestração do Pipeline

Visão Geral

O projeto implementa uma orquestração ponta a ponta utilizando Databricks Jobs com execução serverless e DAG orientado a dependências entre camadas Bronze, Silver, Gold e frameworks operacionais.

A execução foi estruturada para demonstrar:

- Arquitetura Medallion (Bronze → Silver → Gold)
- Pipeline distribuído e desacoplado
- Processamento incremental
- Data Quality Validation
- Observabilidade operacional
- Replay handling
- Governança e auditoria
- Semantic Layer analítica
- Logs operacionais centralizados

DAG Orquestrado - Jobs & Pipelines

A pipeline foi implementada utilizando Databricks Jobs com dependências explícitas entre tarefas.

Fluxo executado:

1. Bronze Ingestion
2. Silver Processing
3. Gold Analytics
4. Data Quality Validation
5. Operational Evidence

Evidência da execução orquestrada

Estrutura da Pipeline

1. Bronze Ingestion

Notebook responsável pela ingestão raw multi-formato na camada Bronze.

Funcionalidades implementadas

- Ingestão CSV, JSON, XLSX e NDJSON
- Controle de schema
- Controle de replay
- Controle de duplicidade
- Logging operacional
- Metadata tracking
- Controle de execução incremental

Evidência

---

2. Silver Processing

Camada responsável pela padronização e tratamento operacional dos dados.

Funcionalidades implementadas

- Deduplicação
- Tratamento de tipos
- Normalização
- Regras de negócio
- Correção de inconsistências
- Enriquecimento operacional
- Data quality derivado

Evidência

---

3. Gold Analytics

Camada semântica analítica para consumo BI e analytics.

Funcionalidades implementadas

- Star Schema
- Dimensional Modeling
- Tabela fato
- Dimensões analíticas
- Semantic Layer
- Métricas operacionais
- Views analíticas

Evidência

Framework de Qualidade

Data Quality Validation

O projeto implementa validações operacionais automatizadas para garantir integridade dos dados processados.

Validações implementadas

- Catálogos disponíveis
- Schemas válidos
- Volumes configurados
- Tabelas Bronze
- Tabelas Silver
- Tabelas Gold
- Tabelas Control
- Integridade estrutural

Evidência

Observabilidade Operacional

Ingestion Log

Tabela operacional utilizada para rastreamento completo das ingestões executadas.

Informações rastreadas

- Arquivo de origem
- Tabela de destino
- Quantidade de registros
- Timestamp da execução
- Data de referência
- Status operacional
- Schema detectado
- Metadata do arquivo

Evidência

Quality Log

Tabela centralizada responsável pelo registro das validações e correções realizadas na camada Silver.

Validações monitoradas

- Status inválidos
- Datas inconsistentes
- Valores derivados
- Duplicidades removidas
- Registros inválidos
- Campos nulos
- Normalizações operacionais

Evidência

Semantic Layer Analítica

A camada Gold disponibiliza uma semantic layer pronta para consumo analítico e integração com ferramentas BI.

Objetos analíticos criados

Dimensões

- dim_cliente
- dim_produto
- dim_data
- dim_regiao
- dim_vendedor
- dim_canal

Fato

- fato_pedidos

Views Analíticas

- vw_pedidos_analytics

Métricas

- metricas_operacionais
- metricas_por_regiao
- metricas_por_periodo
- metricas_por_categoria
- metricas_por_canal

Evidência da camada Gold

Características Técnicas Implementadas

Arquitetura

- Medallion Architecture
- Lakehouse Architecture
- Delta Lake
- PySpark
- Databricks Serverless
- DAG Orchestration

Governança

- Ingestion Log
- Quality Log
- Metadata Tracking
- Replay Handling
- Data Lineage operacional

Qualidade

- Data Quality Framework
- Schema Validation
- Regras de negócio
- Deduplicação
- Normalização
- Tratamento de inconsistências

Analytics

- Star Schema
- Semantic Layer
- Fato e Dimensões
- Views analíticas
- Métricas operacionais

Resultado Final

A solução entregue representa uma plataforma Lakehouse corporativa funcional com:

- Pipeline orquestrado ponta a ponta
- Observabilidade operacional
- Governança de dados

- Governança de dados
- Framework de qualidade
- Camada analítica Gold
- Evidências operacionais reais
- Execução distribuída serverless
- Logs centralizados
- Controle de replay
- Data lineage operacional