Documentação do Pipeline de Dados – Databricks - "Yellow Taxi Trip"

1. Visão Geral
Este documento descreve as etapas, decisões técnicas e resultados obtidos na construção do pipeline "Yellow Taxi Trip" de ingestão, tratamento, análise e automação de dados no Databricks. O projeto segue o conceito de camadas Bronze, Prata e Ouro para organização dos dados, utilizando Delta Lake como formato de armazenamento.

2. Arquitetura e Camadas

Estrutura no Data Lake:

/tlc_trip/
bronze/
yellow_taxi_trip

prata/
tlc_trip.prata.yellow_taxi_trip
tlc_trip.prata.yellow_taxi_trip_invalida
tlc_trip.prata.yellow_taxi_trip_removido
tlc_trip.prata.yellow_taxi_trip_tratada

ouro/
tlc_trip.ouro.taxi_metrics_by_pulocation
tlc_trip.ouro.taxi_metrics_daily

Justificativa das camadas:

- Bronze: armazenamento fiel à origem, sem alterações estruturais, mas com schema definido.
- Prata: dados limpos, normalizados e prontos para análise.
- Ouro: agregações e indicadores que respondem a perguntas de negócio.

Organização dos Notebooks:

/tlc_trip/orchestra => notebook responsável por executar os notebooks das camadas Bronze, Prata e Ouro
/tlc_trip/bronze/brz_yellow_taxi_trip => notebook responsável pela carga dos dados "yellow_taxi_trip" na camada Bronze
/tlc_trip/prata/pra_yellow_taxi_trip => notebook responsável pela carga dos dados "yellow_taxi_trip" na camada Prata
/tlc_trip/ouro/our_yellow_taxi_trip => notebook responsável pela carga dos dados "yellow_taxi_trip" na camada Ouro

3. Camada Bronze – Ingestão de Dados

Fonte original: API de trip records do Yellow Taxi disponível em NYC Open Data (Socrata), que fornece os dados em formatos como JSON ou CSV, paginados mensalmente.

Periodicidade: ingestão mensal, contemplando múltiplos arquivos por mês com paginação ou múltiplas chamadas (limites de linhas por requisição).

Processo técnico: chamadas automatizadas à API para cada mês, com tratamento de paginação, garantindo ingestão incremental e rastreabilidade.

4. Camada Prata – Tratamento e Qualidade dos Dados

Objetivos:
- Remover registros inconsistentes (ex.: campos obrigatórios nulos, valores inválidos).

- Normalizar colunas (tipos de dados, formatação de texto, padronização de datas).
- Aplicar validações de negócio (ex.: valores dentro de intervalos esperados).

Exemplos de regras aplicadas:
- Descartar registros com `tpep_pickup_datetime` ou `VendorID` nulos.
- Remover viagens com distância negativa ou valor de tarifa zero.

5. Camada Ouro – Análises e Métricas

Objetivo: disponibilizar visões prontas para consumo analítico e tomada de decisão.

Processo: agregação e cálculo de métricas, criação de tabelas de indicadores e disponibilização via tabelas Delta.

Exemplos de métricas criadas:
- Top 5 dias com maior faturamento.
- Locais de partida com maior número de corridas.
- Relação entre distância média e gorjeta média por local de partida.