

Relatório Executivo: Pipeline ETL - Gastos Diretos Brasil.IO

1. Objetivo do Projeto

O principal objetivo deste projeto foi automatizar a Extração, Transformação e Carga (ETL) dos dados de "Gastos Diretos" da API pública do Brasil.IO, implementando uma arquitetura Data Lakehouse robusta (RAW, BRONZE, SILVER, GOLD).

O processo foi desenhado para ser resiliente contra falhas de rede e limites de requisição (erro 429), garantindo que os dados finais sejam de alta qualidade e estejam prontos para consumo em ferramentas de Business Intelligence (BI) ou modelos de Machine Learning (ML).

2. Fluxo de Trabalho (Data Lakehouse)

O pipeline segue uma arquitetura Data Lakehouse, onde os dados evoluem em quatro camadas para garantir qualidade e otimização para consumo:

1. RAW (Bruta): É o ponto de coleta. Armazena os dados brutos, exatamente como vieram da API (formato JSON), e utiliza Checkpointing para garantir a continuidade do download em caso de falha.
2. BRONZE: Nesta etapa, o dado é convertido para o formato colunar otimizado (Parquet) e recebe seu primeiro particionamento por ano/mês, servindo como fonte de dados semi-estruturados.
3. SILVER: É a camada de limpeza e qualidade (Data Wrangling). Aqui os dados são padronizados, tipificados (conversão para float, caixa alta) e têm seu *schema* corrigido (Aliasing), mantendo o particionamento por ano/mês.
4. GOLD (Serviço): É o destino final. Contém os dados agregados e modelados (Métricas de BI), otimizados para velocidade de leitura, e é a camada que será consumida por relatórios e análises.

3. Principais Regras de Transformação Implementadas

As regras de Data Wrangling (Camada SILVER) e de Modelagem (Camada GOLD) garantiram a utilidade do conjunto de dados:

3.1. Camada SILVER (Qualidade e Refinamento)

- Consistência de Schema (Aliasing): Colunas com nomes inconsistentes da API (ex: valor, nome_orgao) foram renomeadas para um padrão interno (valor_pagamento, orgao_nome).
- Tipagem de Dados: O campo de fato principal (valor_pagamento) foi convertido de texto para o tipo numérico (float), removendo registros onde a conversão falhou (garantia de integridade).
- Padronização de Texto: Colunas de dimensão (ex: orgao_nome, municipio_nome) foram convertidas para CAIXA ALTA para evitar inconsistências no agrupamento.
- Limpeza de Pastas: Implementação de lógica para apagar e recriar as pastas BRONZE e SILVER antes do processamento, garantindo que não haja resíduos de execuções anteriores com *schema* incompleto.

3.2. Camada GOLD (Agregação e Serviço)

- Artefato de Dados: O foco foi criar uma tabela de fatos agregada.
- Agregação: Os dados foram agrupados por ano, mês, orgao_nome e municipio_nome.
- Métricas de Fato: Calculadas as métricas de BI chave: total_gasto_mensal (soma) e total_trasacoes_mensal (contagem).

4. Conclusão e Status

O pipeline está totalmente operacional e robusto, capaz de baixar e processar mais de 1.3 milhão de registros da API [Brasil.IO](#). O resultado final na camada GOLD (dataset/gold) é um conjunto de dados pronto para ser lido e visualizado por analistas e cientistas de dados, com garantia de qualidade e consistência.