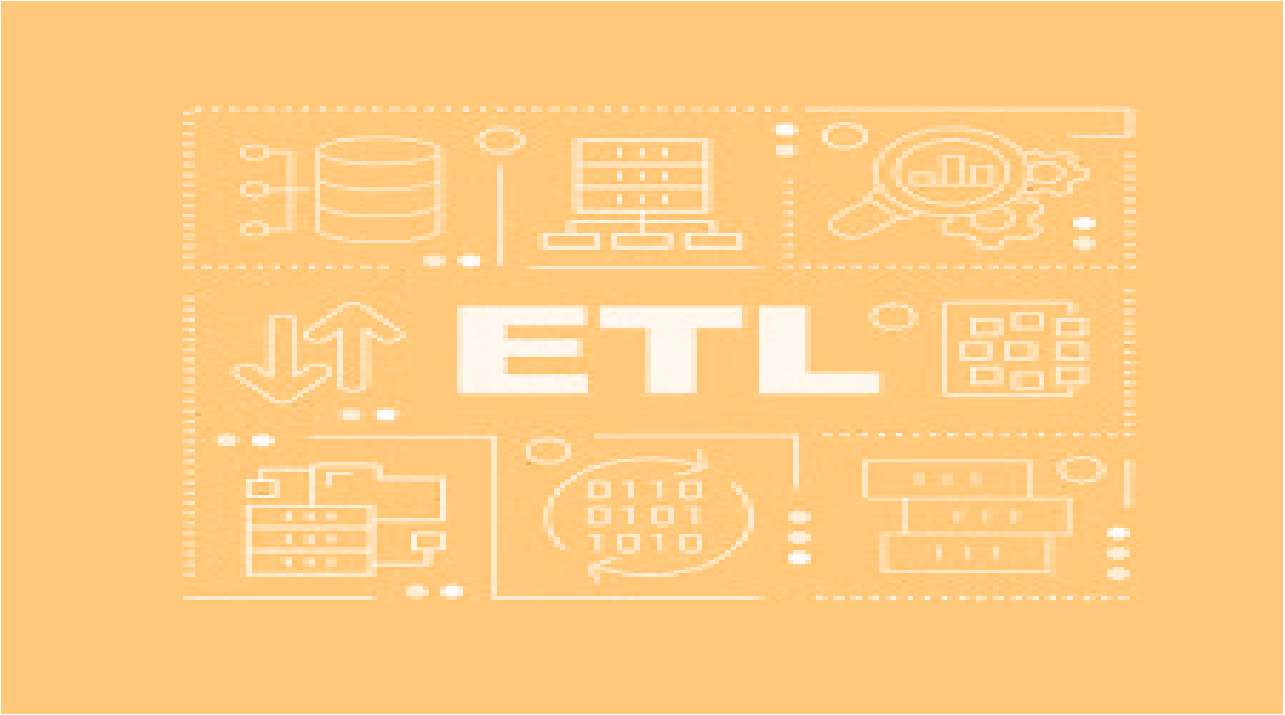

RELATÓRIO SUPER STORE

PROJETO 5 ETL (extração, transformação e carga)

RESPONSÁVEL

Taiza de Souza Santos Ferreira

25.07.2025

1. OBJETIVO

Este projeto teve como objetivo estruturar um processo completo de ETL (Extração, Transformação e Carga) e modelagem dimensional no BigQuery, com base nos dados da Super Store. O foco foi organizar a base de dados de forma eficiente para análises de vendas, clientes, produtos e mercados, viabilizando uma visão estratégica dos dados por meio de um modelo estrela (star schema).

2. DESCRIÇÃO DO CASO

A Super Store enfrentava dificuldades com grande volume de dados desestruturados. A proposta foi construir um pipeline ETL robusto, modelar os dados utilizando boas práticas de Data Warehouse e implementar tudo no BigQuery, com suporte visual pelo LucidApp.

3. METODOLOGIA

O projeto seguiu uma abordagem ágil, com entregas iterativas e foco na clareza, consistência e integridade dos dados. Utilizamos a modelagem dimensional de Kimball (modelo estrela), estruturamos um pipeline de atualização e aplicamos boas práticas de engenharia e governança de dados.

4. BASE DE DADOS

Utilizamos a base `superstore`, com 51.290 registros. A tabela contém dados de vendas detalhadas, incluindo:

- Identificadores únicos (cliente, produto, pedido)
- Atributos categóricos (região, mercado, categoria, segmento)
- Métricas (quantidade, lucro, vendas)
- Informações temporais (datas, semanas, ano)

A estrutura variada da base permitiu a extração de insights estratégicos com qualidade.

5. PROCESSAMENTO E PREPARAÇÃO DOS DADOS

Foram realizadas as seguintes análises:

- **Valores nulos:** Nenhuma ausência foi identificada.
- **Duplicatas:** Nenhuma duplicação em chaves relevantes.
- **Categóricas:** Padronização de letras minúsculas para consistência.
- **Numéricas:** Validação de tipos e faixas.
- **Tipos de dados:** Todos os campos foram ajustados corretamente entre texto, números e datas.

6. INTEGRAÇÃO COM FONTES EXTERNAS

Realizou-se uma integração via web scraping com dados da Wikipédia (lista das maiores empresas por receita), cruzando o setor e país com os dados da Super Store. Isso permitiu identificar potenciais concorrentes e contextos de atuação global.

7. ESTRUTURA DE DADOS (MODELO ESTRELA)

Foi desenvolvido um modelo estrela com:

- **1 tabela fato:** `fato_vendas`, contendo dados quantitativos (vendas, lucro, quantidade)
- **8 tabelas de dimensões:** contendo atributos descritivos como:
    - `dim_regiao`, `dim_mercado`, `dim_categoria`, `dim_order_priority`, `dim_cliente`, `dim_produto`, `dim_ship_mode`, `dim_segmento`

Cada dimensão possui um identificador único (ID), utilizado na tabela fato como chave estrangeira. Isso garante a integridade referencial e otimiza as consultas analíticas.

8. VERIFICAÇÃO DE RELACIONAMENTOS

Foram realizados `JOINs` no BigQuery entre a tabela fato e todas as tabelas de dimensão. O resultado confirmou que todos os relacionamentos estavam funcionando corretamente, sem retornos nulos, validando a estrutura e a integridade do modelo de dados.

9. PIPELINE DE ATUALIZAÇÃO DE DADOS

O pipeline foi planejado em três etapas:

1. **Atualização da base limpa (`superstore_limpa`)**
2. **Atualização das dimensões (`dim_*`)**
3. **Atualização da tabela fato (`fato_vendas`)**

A ordem respeita as dependências entre tabelas, garantindo que as referências estejam disponíveis antes da carga na fato. Foi considerada a possibilidade de atualizações incrementais para otimizar desempenho e evitar recarregamentos completos.

10. CONCLUSÃO

O projeto resultou em uma base de dados analítica robusta, bem estruturada e pronta para suportar a geração de dashboards, relatórios e análises estratégicas. A aplicação de boas práticas de modelagem dimensional, somada à validação dos relacionamentos e à organização do pipeline, garante qualidade, escalabilidade e confiabilidade dos dados.

11. RECOMENDAÇÕES FINAIS

- **Automatizar o pipeline ETL** com ferramentas como Airflow ou Cloud Composer.
- **Monitorar a qualidade dos dados** com validações contínuas.
- **Documentar** todo o modelo e os fluxos de atualização.
- **Expandir indicadores** com novas variáveis de negócio.
- **Otimizar performance** com particionamento e clustering no BigQuery.
- **Capacitar a equipe** com treinamentos em BI e engenharia de dados.