Projeto de Exploração – Estrutura de Dados (ETL)

Organização e Estruturação de Dados para Análise

Thais Lira - Outubro 2025

Entendimento do Problema

Contexto, objetivo e visão geral do projeto

🧩 Contexto

A Super Store possui uma base extensa de transações, mas os dados estavam brutos e não estruturados, dificultando análises estratégicas e operacionais.

🎯 Objetivo

Construir uma base analítica confiável, utilizando o **modelo dimensional (star schema)**, com tabelas fato e dimensão, para possibilitar análises robustas de vendas, clientes, produtos, regiões e mais.

🔍 Visão Geral da Solução

- Estruturação completa dos dados via **processo ETL**
- Criação de **tabela intermediária limpa**
- Implementação do **modelo dimensional** no BigQuery
- Inclusão de **dados externos** para benchmarking internacional

Processo ETL

Etapas de preparação e carga dos dados

Extract (Extração)

Extração de dados brutos de múltiplas fontes

Transform (Transformação)

Limpeza, padronização e enriquecimento dos dados

Load (Carregamento)

Carregamento estruturado no modelo dimensional

📌 O processo foi conduzido manualmente e de forma sequencial, com foco total em garantir qualidade, integridade e rastreabilidade.

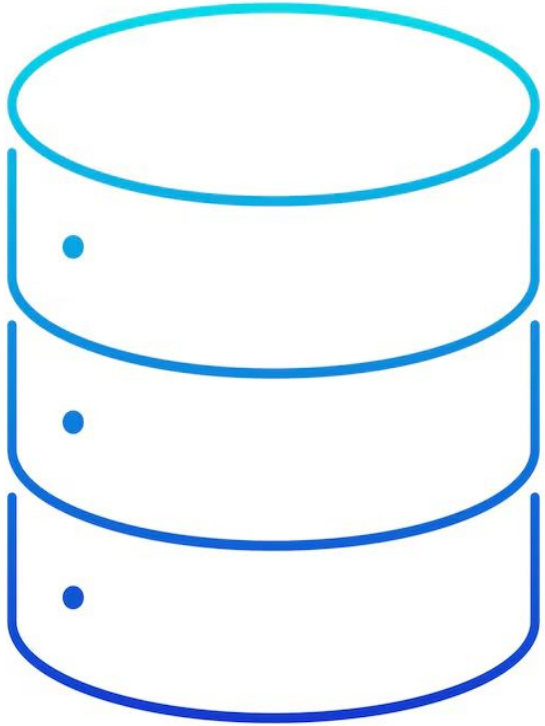
Preparação da Base de Dados

Importação, limpeza e padronização inicial

Base Inicial

Iniciamos com uma base de dados da **Super Store**, contendo **51.290 registros** de transações históricas.

Essa base serviu como ponto de partida para todas as etapas de preparação e modelagem de dados.



Etapas Realizadas

✔️ Verificação completa de valores nulos

✔️ Identificação e remoção de duplicados

✔️ Padronização de variáveis categóricas (`LOWER()`, `TRIM()`)

✔️ Identificação de discrepâncias e outliers

✔️ Criação da tabela intermediária `superstore_cleaned`

Identificação de Valores Nulos e Duplicados

Análise e tratamento de registros redundantes

51.290

Total de Linhas

Volume inicial da base de dados

51.255

Linhas Únicas

Linhas únicas restantes

35

Linhas Duplicadas

Linhas duplicadas removidas

Interpretação dos Resultados

- Nenhum valor nulo foi identificado nas colunas da base — os dados estavam completos.
- Foram confirmados **35 registros duplicados**, todos **redundantes**, ou seja, sem informações adicionais relevantes.
- Essas duplicatas poderiam distorcer métricas-chave, como vendas, lucro e quantidade.

Ação Realizada

✔️ Criação da tabela superstore_cleaned
→ Contendo apenas registros únicos e validados, utilizada nas próximas etapas do processo ETL.

💸 Quando vender mais significa **perder dinheiro**

Qual o custo real de aplicar um desconto agressivo? Os dados mostram a resposta

Faixa de Desconto	% Lucro Negativo	Interpretação
Sem desconto	0,04%	Situação ideal
1–10%	21%	Risco moderado
11–20%	24%	Risco elevado
21–30%	62%	Alto risco
> 30%	92%	Prejuízo quase certo

Conclusão crítica: Descontos elevados estão fortemente associados a prejuízo financeiro e devem ser revisados na política comercial..

Pesquisa de Outras Fontes

Benchmarking internacional com redes de supermercados

Dados de Concorrentes

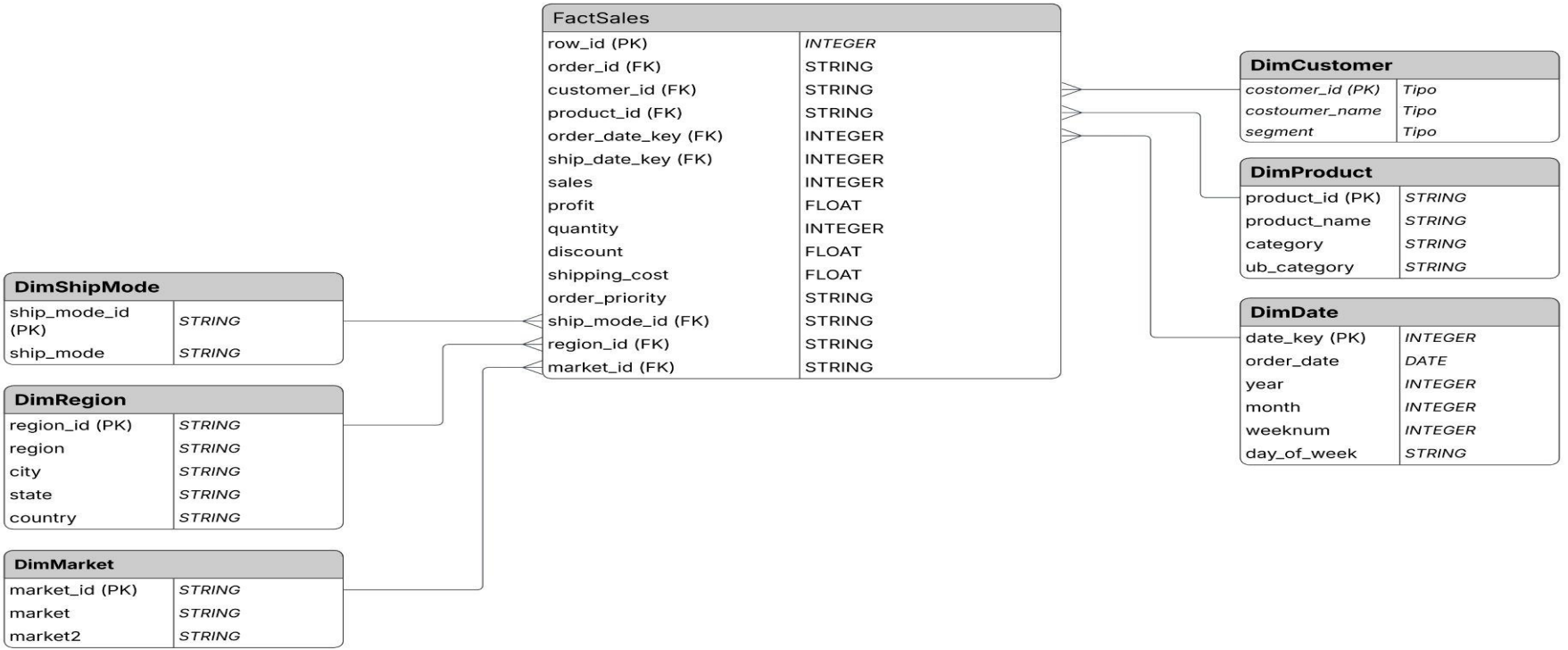
Para enriquecer a análise, foram integrados **dados de concorrentes internacionais** usando técnicas de **web scraping**.

Metodologia

- 🔎 Extração via IMPORTHTML do Google Sheets
- 🌐 Fonte: Wikipedia - lista de varejistas globais
- 🎯 Critério de seleção: foco em empresas com presença internacional significativa

Modelo Dimensional (Star Schema)

Estrutura da FactSales e tabelas de dimensão



Dá Origem à Decisão: Montagem da Base Analítica

Etapas de transformação até a criação da FactSales no modelo estrela

1. Superstore (Base Bruta)

Importação dos dados originais com validação inicial de estrutura e tipos de dados

2. Superstore_Cleaned (Intermediária)

Aplicação das regras de limpeza, deduplicação e padronização estabelecidas.

3. Tabelas Dimensão

População das dimensões Customer, Product, Date, Region, ShipMode e Market.

4. FactSales (Consolidação)

Criação da tabela fato com chaves estrangeiras e métricas calculadas.

Conclusões do Projeto

Principais entregas e resultados alcançados

1

✅ Qualidade dos Dados

- Eliminação de **35 registros duplicados**
  Padronização textual com uso de LOWER() e TRIM()
- Base de dados limpa e consistente

2

💡 Insights de Negócio

- Análise revelou que **descontos acima de 30%** resultam em **92% de prejuízo**
- Recomendação: **revisar política de descontos**

3

🧱 Estrutura Dimensional

- Implementação completa do modelo **Star Schema** no **BigQuery**
- Estrutura com **1 tabela fato** e **6 tabelas dimensão**

4

🌍 Integração Externa

- Inclusão de dados de **concorrentes internacionais**
- Benchmarking com **varejistas globais** para apoiar análises comparativas

Próximos Passos

Recomendações para evolução da solução

1	**Curto Prazo** • Automatizar o pipeline **ETL** (ex: com Airflow ou Cloud Composer) • Criar **dashboards executivos** em Power BI ou Looker Studio
2	**Médio Prazo** • Implementar **SCD Tipo 2** para rastreamento histórico de mudanças nas dimensões • Desenvolver **análises preditivas** com base em padrões de comportamento e vendas
3	**Longo Prazo** • Aplicar técnicas de **Machine Learning** para: • **Otimização de preços** • **Recomendações personalizadas** de produtos