Realizado por Taiza Ferreira

ETL SUPER STORE

Modelagem Dimensional no BigQuery
com Pipeline de Atualização

Objetivo

Estruturar um processo completo de ETL e modelagem dimensional no BigQuery, organizando os dados da Super Store para análises estratégicas de vendas, clientes, produtos e mercados.

Descrição do caso

A Super Store **enfrentava desafios com grande volume de dados desestruturados**. O projeto **propôs um pipeline ETL robusto e um modelo estrela** no BigQuery, **visando organização, clareza e eficiência** na análise dos dados.

O que é ETL?

ETL é o **processo que prepara os dados** para análise de forma **automatizada e confiável**.

Extração: coleta de dados brutos

Transformação: limpeza e organização dos dados

Carga: envio para um sistema analítico (como o BigQuery)

A Base de Dados

51.290 registros

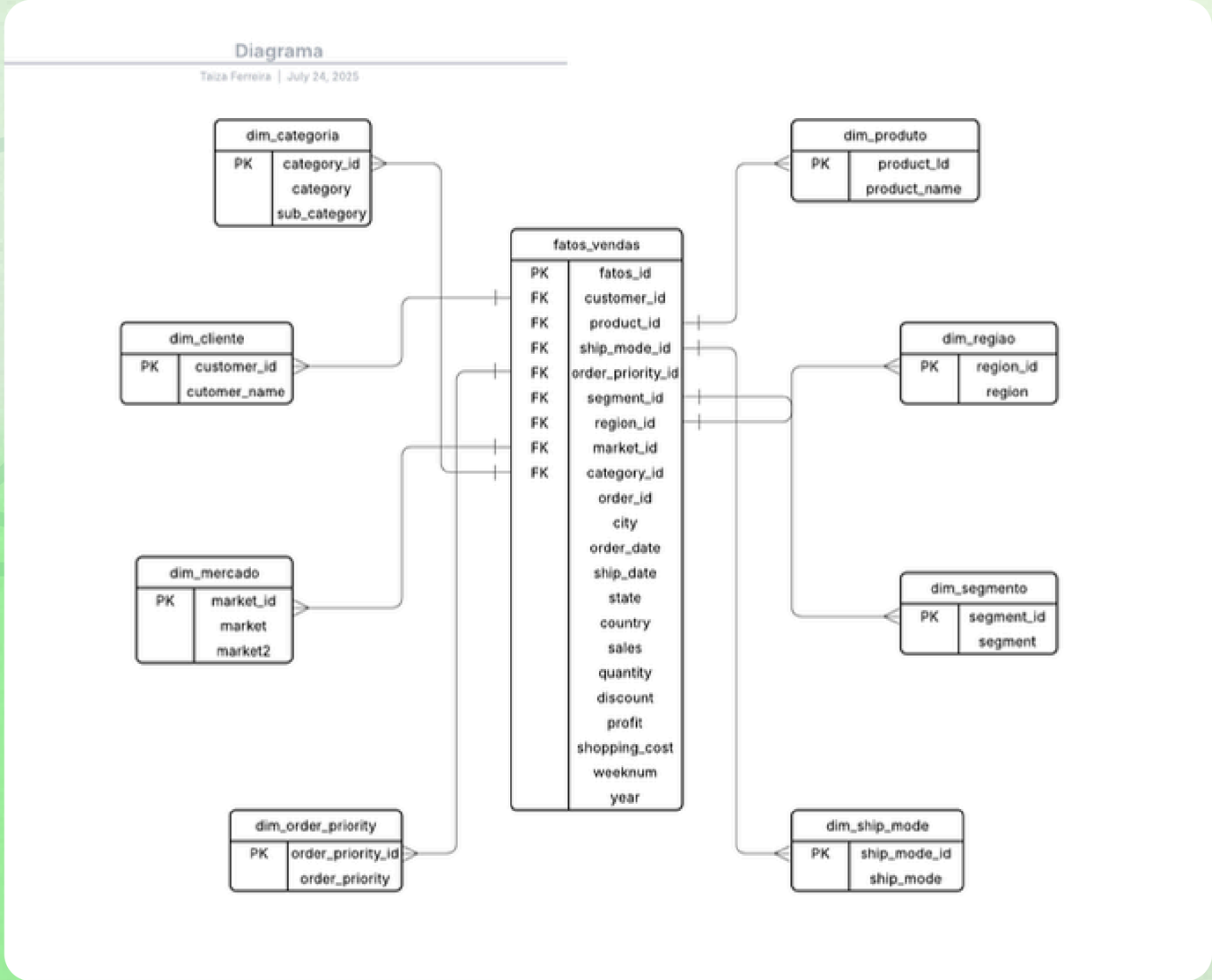
Fonte: Tabela superstore

Tabela: informações sobre pedidos, clientes, regiões, produtos...

Tipos de dados variados (textos, números, datas).

Dados consistentes, sem nulos ou duplicações

Esquema em Estrela



- Criado **1 tabela fato**: fato_vendas
- E **8 tabelas de dimensõe**s (cliente, produto, categoria, etc.)
- Utilizado **chaves primárias** e **estrangeiras** para garantir relacionamento e integridade dos dados.

Ao construir as tabelas de dimensão, levamos em conta que **alguns dados podem mudar ao longo do tempo**. Ex: um cliente pode mudar de segmento. Esse tipo de mudança pode afetar a análise se não for tratada corretamente.

O que são Slowly Changing Dimensions (SCD)?

Técnicas para **lidar com mudanças em dados** de dimensão **ao longo do tempo**.

Tipos principais:

- Tipo 1: **sobrescreve** (sem histórico)
- Tipo 2: **nova linha** (mantém histórico completo)
- Tipo 3: **guarda valor antigo e atual** (histórico parcial)

Usamos uma abordagem próxima ao Tipo 1, mas o modelo está pronto para evoluir para Tipo 2, caso seja necessário preservar histórico.

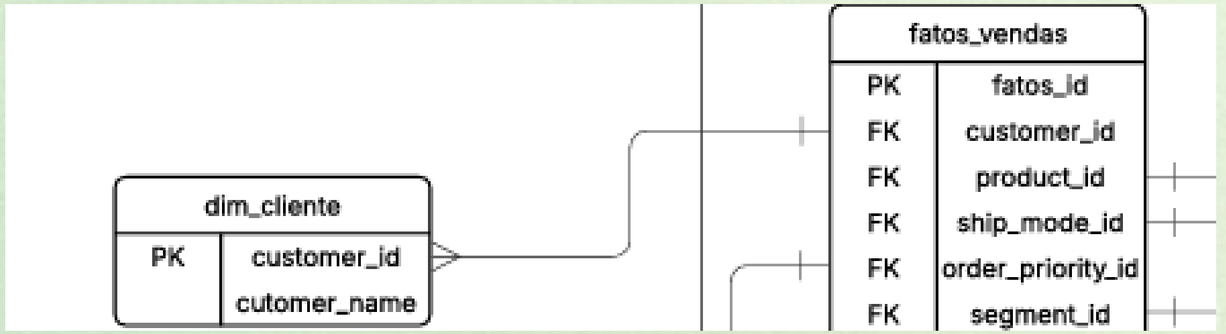
Verificação dos Relacionamentos

Para validar a integridade do modelo estrela, foi realizado JOINs entre a tabela **fato_vendas** e as **8 dimensões.**

O resultado confirmou que todos os identificadores estão **corretamente relacionados**, sem registros órfãos ou nulos.

Isso assegura que os dados podem ser analisados de forma integrada, confiável e **sem falhas de relacionamento.**

Exemplo: customer_id na fato encontrou correspondência exata na dim_cliente.



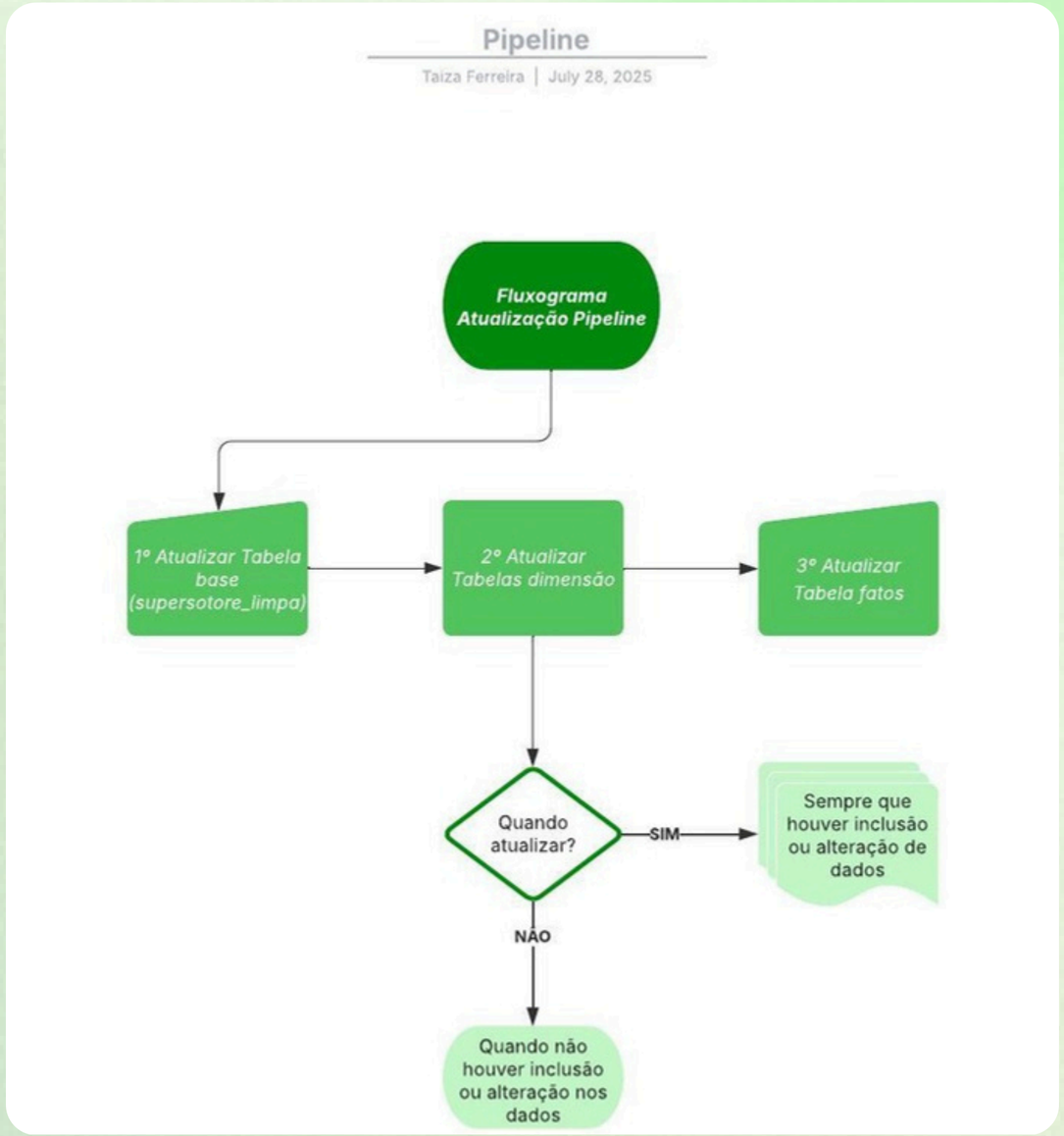
Pipeline de Atualização

Planejado o fluxo de atualização em 3 etapas:

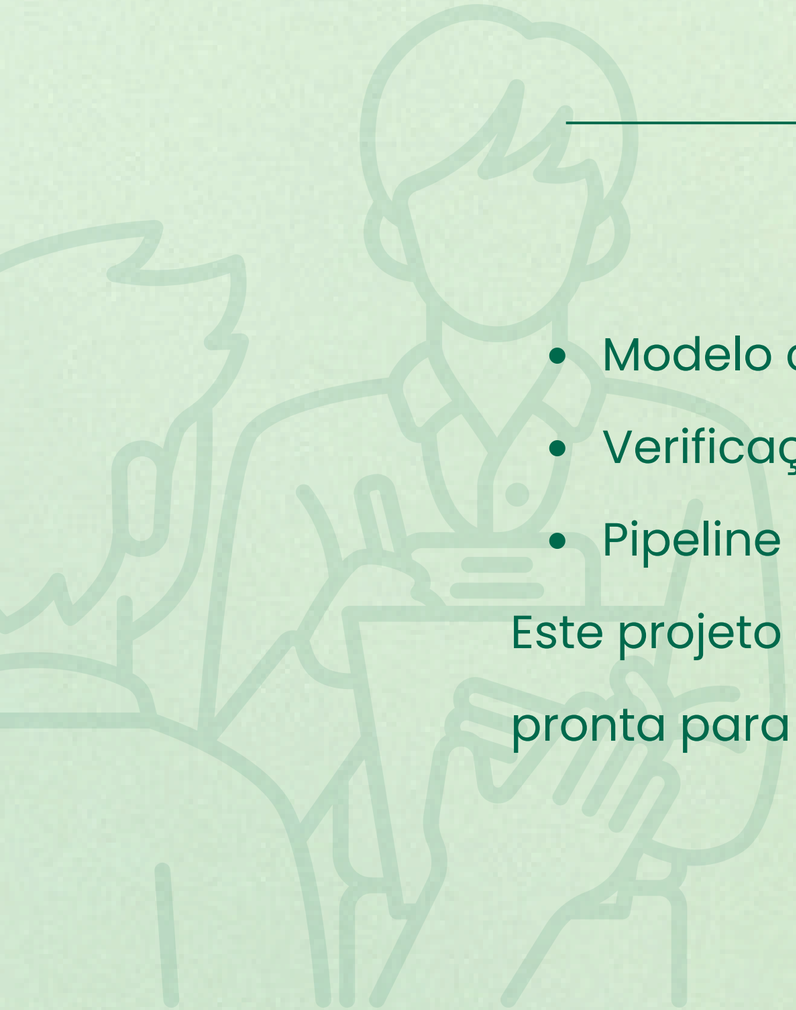
- Base limpa (superstore_limpa)
- Tabelas de dimensão
- Tabela fato

Ordem garante integridade e evita erros

Modelo pronto para futuras automações



Conclusão



- Modelo criado é confiável, escalável e pronto para análise.
- Verificações garantem integridade e consistência.
- Pipeline planejado garante manutenção segura dos dados.

Este projeto entrega uma base sólida, confiável e escalável, pronta para apoiar decisões estratégicas baseadas em dados

Recomendações

Automatizar o pipeline com ferramentas de agendamento e controle de fluxo de dados

Aplicar a estratégia de SCD mais adequada ao negócio

Implementar validações para garantir a qualidade dos dados

Documentar os processos e capacitar a equipe para sustentação

OBRIGADO!

linkedin.com/in/taiza-ferreira-dados/

TaizaFerreira - Overview

BigQuery | Power BI | Google Sheets | Google Colab | Looker Studio - TaizaFerreira

GitHub