

Projeto de Análise e Tratamento de Dados – Olist (MySQL)

Este documento apresenta, de forma detalhada, um projeto completo de engenharia e análise de dados desenvolvido em MySQL, utilizando o dataset público da Olist, um dos maiores e-commerces do Brasil. O projeto foi estruturado com base na metodologia STAR (Situation, Task, Action, Result), permitindo uma descrição clara do contexto, das responsabilidades assumidas, das ações executadas e dos resultados obtidos ao longo do processo.

S — Situation (Situação)

O cenário inicial do projeto envolvia um grande volume de dados transacionais armazenados em múltiplos arquivos CSV, representando diferentes áreas do negócio, como pedidos, clientes, vendedores, produtos, pagamentos, avaliações e informações de geolocalização. Apesar da riqueza dos dados, o material encontrava-se em estado bruto, sem tratamento adequado para análise.

Durante a análise inicial, foram identificados diversos problemas que comprometiam a qualidade e a confiabilidade das informações, tais como campos vazios ou nulos, tipos de dados inconsistentes, registros duplicados e ausência de índices para otimização de consultas. Além disso, não havia uma estrutura analítica definida que permitisse responder perguntas de negócio de forma eficiente.

Diante desse contexto, surgiu a necessidade de transformar os dados brutos em uma base relacional estruturada, confiável e preparada para suportar análises exploratórias, relatórios gerenciais e futuras soluções de Business Intelligence.

T — Task (Tarefa)

O objetivo principal do projeto foi construir uma solução de dados robusta utilizando MySQL, garantindo qualidade, integridade e desempenho. Para isso, foram definidas as seguintes responsabilidades:

- 1 Modelar e criar tabelas relacionais representando as principais entidades do negócio.
- 2 Realizar a carga de dados a partir de arquivos CSV, respeitando padrões de formatação e consistência.
- 3 Tratar dados ausentes, inconsistentes ou incorretos.
- 4 Identificar e eliminar registros duplicados, garantindo unicidade das informações.
- 5 Otimizar o banco de dados por meio da criação de índices.
- 6 Desenvolver consultas analíticas capazes de gerar insights relevantes para o negócio.

Todas as tarefas foram executadas com foco em boas práticas de engenharia de dados, priorizando desempenho, escalabilidade e clareza.

A — Action (Ação)

As ações realizadas ao longo do projeto foram organizadas em etapas bem definidas, garantindo controle e rastreabilidade das transformações aplicadas aos dados.

1. Modelagem e Criação das Tabelas

Inicialmente, foram criadas tabelas relacionais no MySQL para representar as principais entidades do ecossistema Olist, incluindo clientes, pedidos, itens de pedidos, pagamentos, produtos, vendedores, avaliações e geolocalização. Cada tabela foi definida com tipos de dados adequados, visando consistência e suporte a análises futuras.

2. Carga de Dados (ETL – Extract & Load)

Após a modelagem, foi realizada a carga dos dados por meio do comando `LOAD DATA LOCAL INFILE`, permitindo a importação eficiente dos arquivos CSV para o banco de dados. Nessa etapa, foram padronizados delimitadores, aspas e encoding, garantindo a correta interpretação das informações.

3. Limpeza e Tratamento de Dados (Transform)

- 1 Conversão de campos vazios em valores NULL, facilitando o tratamento analítico.
- 2 Identificação e exclusão de registros completamente vazios.
- 3 Avaliação de tamanhos e tipos de variáveis para evitar inconsistências.
- 4 Detecção de registros duplicados com base em chaves de negócio, como `review_id`.
- 5 Remoção segura de duplicidades utilizando colunas auxiliares.

4. Otimização de Performance

Com os dados tratados, foram criados índices em colunas estratégicas, melhorando significativamente o desempenho das consultas. Também foram realizadas validações de unicidade e remoção de índices temporários após seu uso.

5. Consultas Analíticas

Por fim, foram desenvolvidas consultas analíticas para responder questões relevantes do negócio, como tempo de aprovação de pedidos, análise de compras por dia da semana, tempo de entrega ao cliente, classificação de clientes por recorrência e cálculo do valor total dos pedidos.

R — Result (Resultado)

Ao final do projeto, foi entregue uma base de dados totalmente estruturada, limpa e confiável. As principais conquistas incluem:

- 1 Redução significativa de inconsistências e registros duplicados.
- 2 Melhoria expressiva no desempenho das consultas SQL.
- 3 Dados prontos para análises exploratórias, dashboards e relatórios gerenciais.
- 4 Estrutura escalável para futuras integrações, como Data Warehouses e ferramentas de BI.

O projeto demonstra domínio de modelagem relacional, SQL avançado e boas práticas de engenharia de dados, sendo altamente relevante para portfólios nas áreas de Análise e Engenharia de Dados.

Tecnologias Utilizadas

- 1 MySQL
- 2 SQL (DDL, DML e consultas analíticas)
- 3 Dataset público da Olist

Considerações Finais

Este projeto pode ser facilmente expandido para a construção de um Data Warehouse, integração com ferramentas de Business Intelligence como Power BI, Tableau ou Looker, além da automação de pipelines ETL. Trata-se de um projeto completo, bem documentado e ideal para demonstrar competências técnicas em ambientes acadêmicos e profissionais.