Engenheiro de Dados – OLIST

Candidato: Felipe Azevedo Pinagé

Proposta

Foi proposta uma nova modelagem relacional dimensional (dimensão e fatos) para os datasets disponíveis do Olist. Um dos principais motivos da proposta é a melhoria da relação entre clientes e produtos, pedidos e localidades, e até clientes e vendedores. Dessa forma, podemos diminuir a quantidade de *joins,* facilitando o acesso aos dados mais relevantes, reduzindo o tempo de *query* e tornando as análises mais rápidas.

O novo modelo pode ser observado na Figura 1. Para isso foram definidas três tabelas fato: 1) vendas; 2) pagamentos; e 3) e comentários (*reviews*). Além das quatro tabelas de dimensões: 1) data; 2) produto; 3) localidade; e 4) clientes.

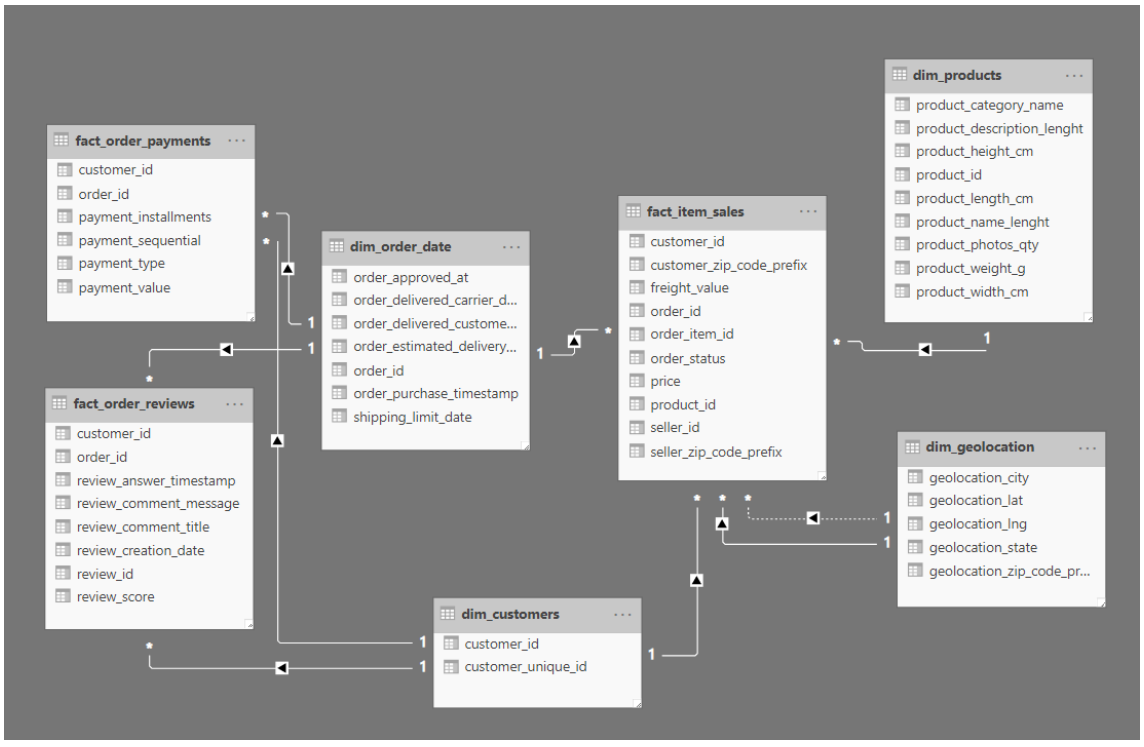

Figura 1. Modelagem Fato e Dimensões para e-commerce Olist (gerado pelo Power BI).

Portanto, através do modelo acima, será construído *data warehouse* (DW) onde as *queries* poderão ser executadas, a fim de comparar a performance da solução proposta com o modelo original fornecido pelo Olist.

Etapas e Ferramentas

Este projeto foi dividido em quatro etapas principais:

1. Planejamento do novo modelo dimensional;
2. Criação do *data lake* baseado na solução Amazon S3 como plataforma de armazenamento primária;
3. Processamento de ETL (*Extract-Transform-Load*) para prepara os dados para a integração. Todo o processamento foi feito em PySpark através da plataforma Databricks;
4. Construção do DW utilizando a solução Amazon Redshift para a gravação de tabelas e esquemas.

Além das etapas mencionadas, algumas visualizações dos dados foram criadas para explorar as possíveis análises a serem feitas, utilizando a ferramenta Power BI.

Visualizações dos Dados

Segue abaixo algumas visualizações de análises (construídas em Power BI) que podem ter seu tempo reduzido através da modelagem proposta.

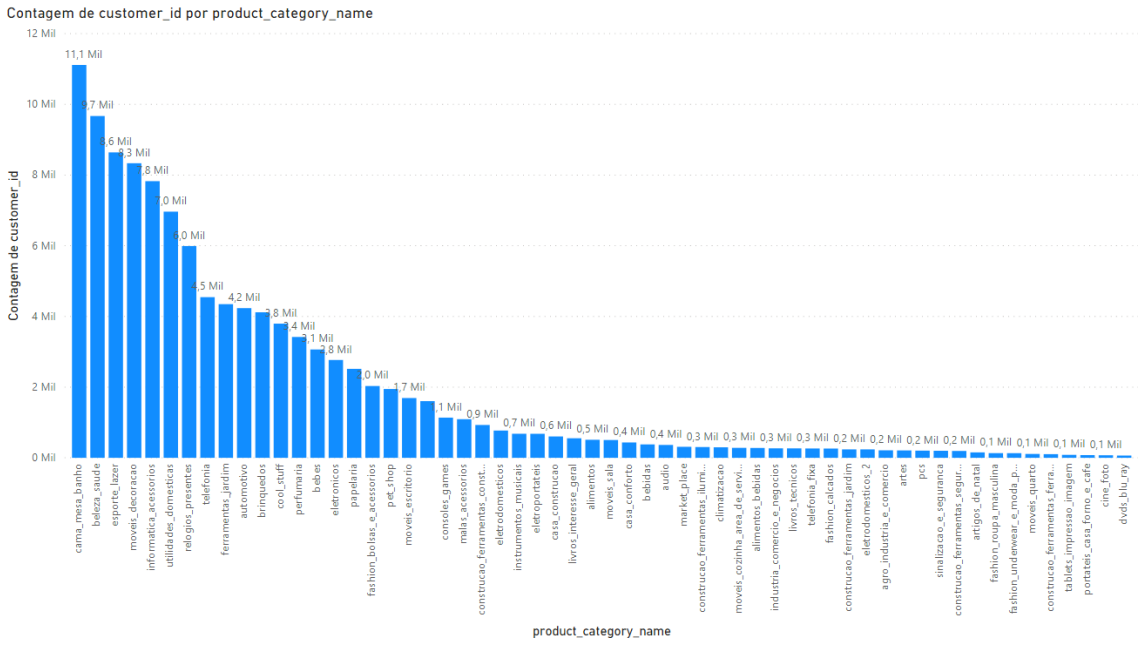

Figura 2. Número de clientes por categoria de produto.

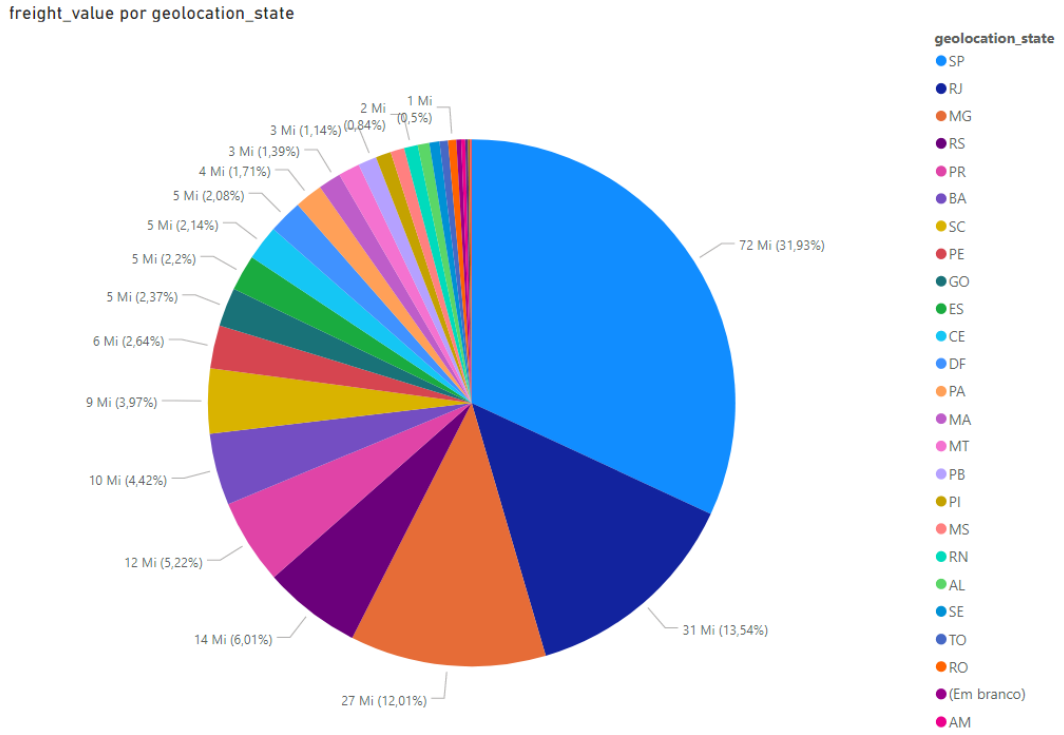

Figura 3. Valor acumulado de fretes para cada estado do Brasil.

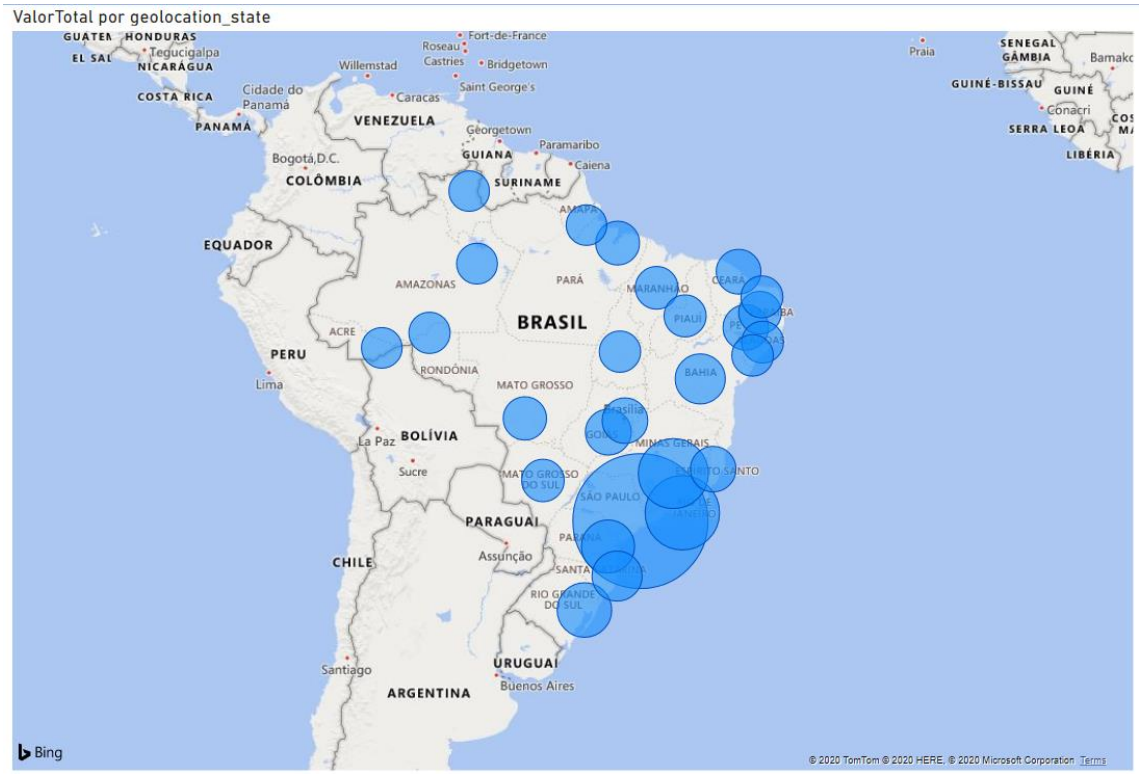

Figura 4. Valor total de pedidos (preço + frete) para cada estado do Brasil.

ETL

O processo de ETL basicamente: 1) lê os dados armazenados no Amazon S3; 2) transforma os dados organizando as relações de acordo com a modelagem fato-dimensão; e finalmente, 3) salva os dados transformados de volta no Amazon S3 para serem integrados a outras ferramentas e consumidos.
Todo o processamento foi feito em pyspark através da plataforma Databricks (código em anexo no *link* do GitHub).

Obs: Também é possível fazer a integração direta entre o Datrabricks e o Amazon Redshift, mas o nível gratuito de acesso à plataforma AWS não atende este recurso.

Data Warehouse

Foi utilizado o serviço Amazon Redshift para a criação do DW. As tabelas foram criadas de acordo com a modelagem proposta e copiadas dos dados armazenados no Amazon S3 (entregues após o processamento de ETL).
Os scripts de criação de todas as tabelas estão em anexo no *link* do GitHub.

Resultados

Como validação da modelagem proposta, foram comparados os tempos de execução de algumas *queries* (anexadas no *link* do GitHub) realizadas tanto nos dados do modelo original fornecido pelo Olist, quanto nos dados do modelo proposto.
A primeira análise refere-se ao tipo de pagamento mais escolhido entre os clientes. O tempo de execução da *query* da modelagem proposta (0,611 segundos) representa aproximadamente 10% do tempo de execução da modelagem original (6,113 segundos). Conforme observado nas Figuras 5 e 6.

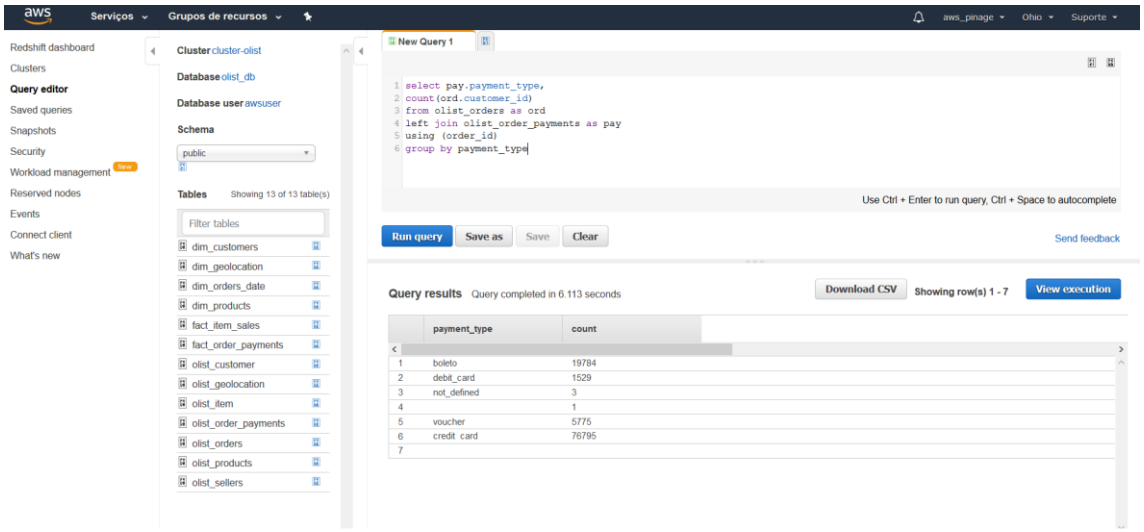

Figura 5. Query no modelo original - Número de clientes por tipo de pagamento.

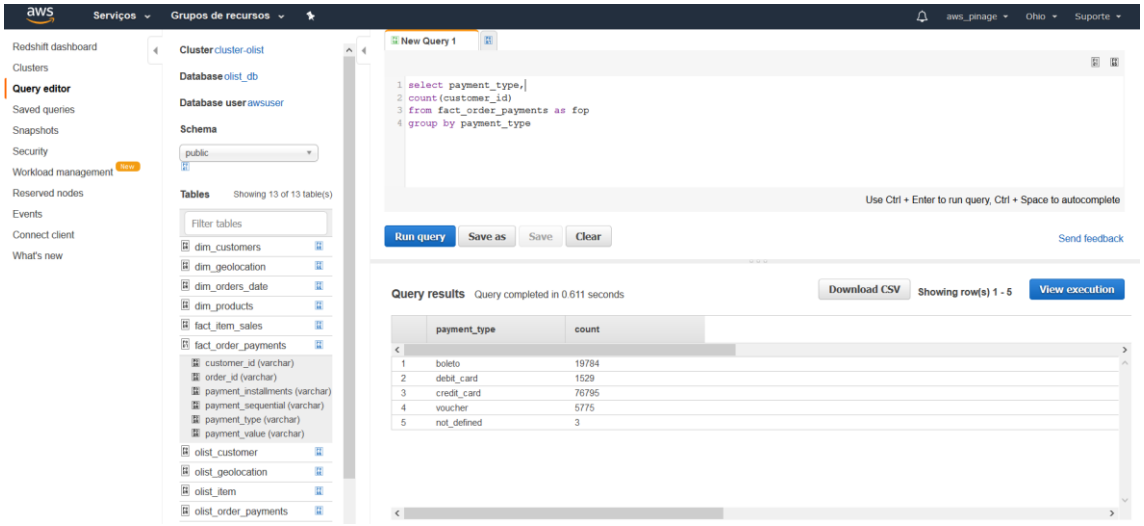

Figura 6. Query no modelo proposto - Número de clientes por tipo de pagamento.

A segunda análise refere-se aos totais de valores em itens e em fretes demandados por cada cliente, também chamados de *Gross Merchandise Volume* (GMV). O tempo de execução da modelagem proposta (1,508 segundos) representa aproximadamente 20% do tempo de execução da modelagem original (7,296 segundos). Conforme observado nas Figuras 7 e 8.

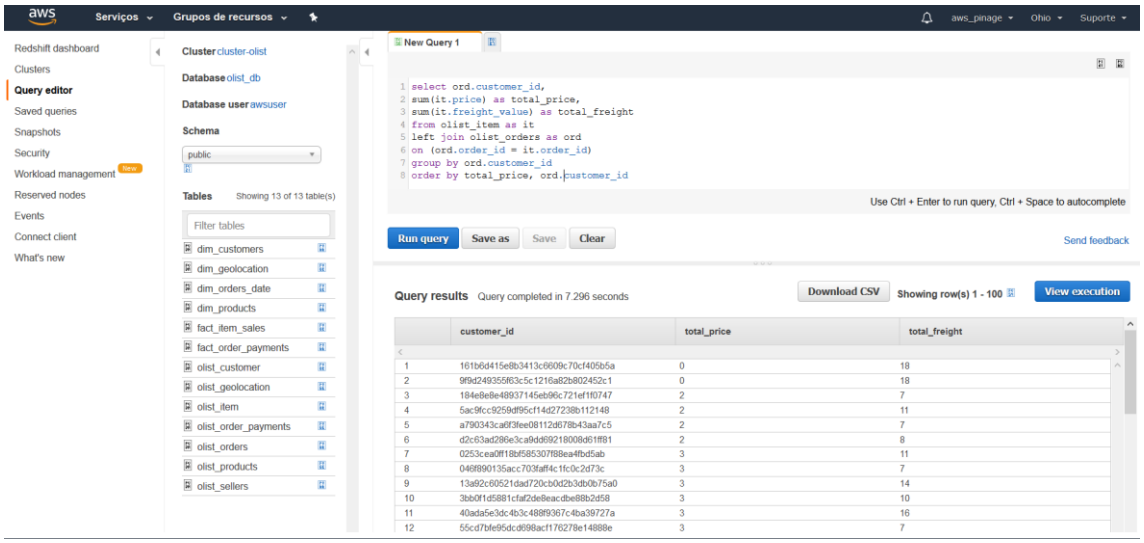

Figura 7. Análise no modelo original - GMV por cliente.

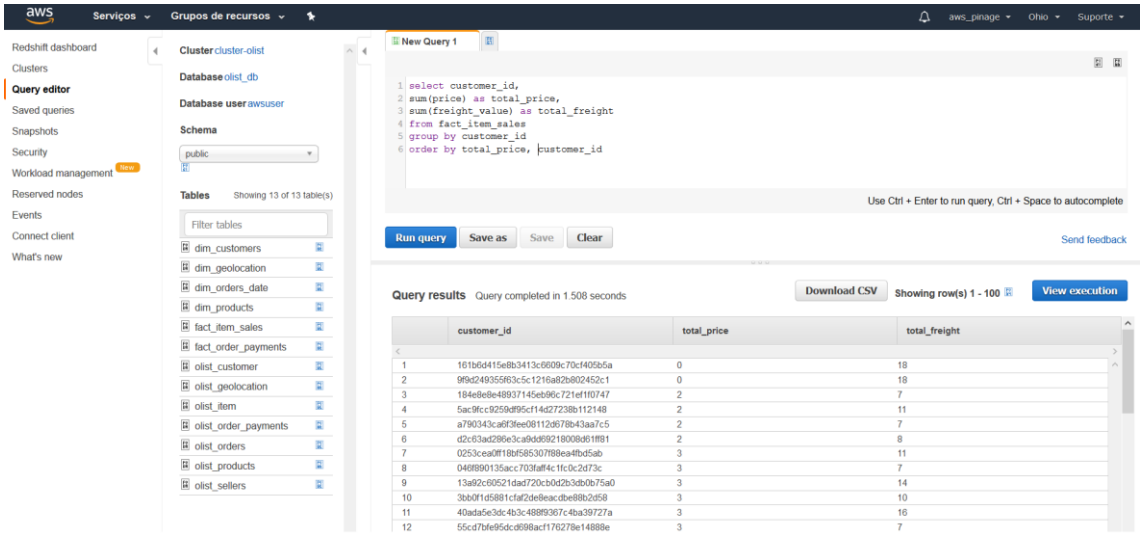

Figura 8. Análise no modelo proposto - GMV por cliente.

Considerações Finais

O modelo proposto foi todo projetado em ambiente *cloud*, majoritariamente pela plataforma da AWS. A nova modelagem se mostra mais adequada para os cientistas de dados e analistas de BI por resultar análises mais rápidas e *queries* mais “limpas”.

Além disso, analistas de BI podem utilizar tanto as ferramentas de visualização em *cloud* quanto às *on-premisse*, boa parte delas tem conectores com o

Amazon Redshift, acompanhando possíveis atualizações de infra-estrutura dos dados.

Para lidar com grande volume de dados, a plataforma da AWS já apresenta um modelo escalável, onde é possível adicionar ou remover nós do cluster sem perda dos dados e interrupção dos sistemas.

Finalmente, nesta plataforma, os dados podem ser disponibilizados em tempo real sem grandes custos, acredito que uma prova de conceito como no caso de implementações com Kafka podem ser apresentadas aos diretores.