Armazém de dados para análises do mercado brasileiro de e-commerce

Eduardo Assis, Gabriel Abílio, Gabriel Martins, Michel Barros, Victor Prates

30 de junho de 2025

Resumo

Este trabalho consiste na criação de um data warehouse que armazena dados reais e anonimizados de transações de comércio eletrônico no Brasil. Além de criar o armazenamento destes dados, é nosso objetivo conseguir extrair informações importantes para a análise, como criação de gráficos e emissão de relatórios.

1. Introdução

O comércio eletrônico tem se consolidado como uma das principais formas de consumo no Brasil, com milhares de transações sendo realizadas diariamente por meio de diferentes plataformas e marketplaces. Com o objetivo de permitir uma análise mais profunda sobre o comportamento de compra dos consumidores brasileiros, este trabalho utiliza um conjunto de dados públicos disponibilizado pela Olist, uma plataforma de e-commerce que atua em diversos marketplaces nacionais.

O dataset em questão contém informações de aproximadamente 100 mil pedidos realizados entre 2016 e 2018, incluindo dados sobre pagamentos, preços, status dos pedidos, fretes, localização dos clientes, características dos produtos e avaliações feitas pelos consumidores. Além disso, há um conjunto complementar de dados de geolocalização que relaciona os códigos postais brasileiros a suas respectivas coordenadas geográficas.

Com a intenção de organizar e estruturar essas informações de forma que facilite sua exploração analítica, este trabalho propõe a construção de um data warehouse. Através de técnicas de modelagem multidimensional e uso de ferramentas apropriadas, foi possível transformar os dados brutos em uma base integrada e otimizada para análise, permitindo a geração de insights relevantes sobre o e-commerce brasileiro.

2. Processo de Desenvolvimento

O processo de desenvolvimento do Data Warehouse seguiu uma abordagem estruturada, envolvendo diferentes etapas que vão desde a compreensão da origem dos dados até o pré-processamento necessário para garantir sua qualidade e usabilidade. Para lidar com a variedade de arquivos e informações presentes no dataset da Olist, foi necessário realizar uma análise exploratória inicial, seguida por um processo de extração, limpeza e transformação dos dados em um formato mais adequado à modelagem dimensional. Esta seção apresenta as principais etapas envolvidas nesse processo.

2.1 Origem dos Dados

Os dados utilizados neste projeto foram extraídos de um dataset público disponibilizado pela Olist, uma plataforma brasileira de vendas online que opera em diversos marketplaces. O conjunto de dados compreende cerca de 100 mil pedidos realizados entre os anos de 2016 e 2018 e foi disponibilizado de forma anonimizada para fins educacionais e analíticos.

O dataset é composto por múltiplos arquivos em formato CSV, cada um representando uma dimensão específica do processo de compra. Entre os principais arquivos, destacam-se:

- `olist_customers_dataset.csv`: informações sobre os clientes, como localização (estado e cidade).
- `olist_geolocation_dataset.csv`: coordenadas geográficas associadas aos CEPs brasileiros.
- `olist_order_payments_dataset.csv`: detalhes das transações financeiras, incluindo tipo de pagamento e parcelas.
- `olist_order_items_dataset.csv`: dados sobre os itens vendidos em cada pedido.
- `olist_orders_dataset.csv`: informações gerais sobre os pedidos, como status, data de compra, e prazo de entrega.
- `olist_products_dataset.csv`: características dos produtos comercializados.
- `olist_sellers_dataset.csv`: informações sobre os vendedores, como localização (estado e cidade).

2.2 Extração dos Dados

A extração dos dados foi realizada a partir dos arquivos CSV disponibilizados publicamente no dataset da Olist. Cada arquivo representava uma tabela ou entidade relevante no fluxo de pedidos de e-commerce, como pedidos, clientes, pagamentos, produtos, vendedores, entre outros.

Esses arquivos foram importados para um ambiente de análise com o uso da linguagem Python e a biblioteca `pandas`, permitindo uma leitura eficiente e o início da exploração relacional entre os dados.

As tabelas foram integradas por meio de suas chaves naturais, como `order_id`, `product_id`, `customer_id`, `seller_id`, e utilizadas para alimentar tanto a tabela fato `fct_sales` quanto as dimensões do modelo estrela.

2.3 Pré-processamento

O pré-processamento dos dados teve como objetivo garantir integridade, consistência e formato adequado para a carga no Data Warehouse modelado em estrela. Abaixo, descrevem-se os principais procedimentos realizados:

- **Conversão de dados temporais**: os campos de data e hora foram convertidos e desmembrados em componentes como ano, mês, dia, hora e dia da semana para alimentar a `dim_date`.
- **Tratamento de localização**: informações de CEP foram agregadas e associadas às coordenadas geográficas (latitude e longitude) para preencher as dimensões `dim_customer` e `dim_seller`.
- **Normalização de produtos**: categorias de produto e dimensões físicas (peso e medidas) foram organizadas para compor a `dim_product`.
- **Agregação de pagamentos**: os métodos de pagamento, número de parcelas e valores pagos foram agrupados por pedido e codificados para formar a `dim_payment`.
- **Construção da tabela fato**: a `fct_sales` foi construída consolidando os relacionamentos entre todas as dimensões e armazenando métricas como preço, valor do frete e total da venda.

Todas essas transformações foram conduzidas com scripts em Python, garantindo flexibilidade e rastreabilidade.

2.4 Ferramentas Usadas para o Desenvolvimento

A construção do Data Warehouse foi realizada com foco em desempenho e escalabilidade, utilizando ferramentas modernas que suportam grandes volumes de dados e operações analíticas. Abaixo estão as principais tecnologias empregadas:

- **Python (Pandas)**: utilizada para as etapas de extração e transformação de dados, compondo o pipeline ETL.
- **ClickHouse**: banco de dados analítico colunar de alto desempenho, escolhido para armazenar a tabela fato `fct_sales` e as tabelas dimensionais. Sua arquitetura permite consultas analíticas em tempo quase real mesmo com grandes volumes de dados.
- **Apache Superset**: plataforma de código aberto para visualização e exploração de dados, utilizada para criação dos dashboards interativos conectados ao ClickHouse, permitindo a análise dos dados modelados de forma dinâmica e eficiente.

O uso combinado dessas ferramentas proporcionou uma arquitetura orientada a dados, capaz de integrar, armazenar e analisar grandes volumes de informação com agilidade e flexibilidade.

2.5 Ambiente de execução do banco de dados

Após o processamento dos dados, uma instância do ClickHouse e do Apache Superset foram instaladas em um VPS (Virtual Private Server), onde os membros do grupo tiveram acesso às ferramentas de análise e criação de gráficos do Apache Superset.

Foi utilizado os contêineres em Docker do ClickHouse e do Apache Superset, conectados em uma mesma rede Docker e ambos conectados com o proxy reverso Caddy responsável por encaminhar as requisições corretamente para o serviço do Apache Superset.

Após instalação dos ambientes, foram criadas tabelas que representam os dados modelados no ClickHouse e os dados foram importados a partir de um arquivo `csv`. Após isso, foi necessário configurar a conexão entre Apache Superset e ClickHouse e criar os datasets a partir das tabelas criadas para começar a fazer a análise exploratória dos dados e criação dos gráficos.

3. Modelagem do DW

A modelagem do Data Warehouse teve como base o modelo em estrela, que visa otimizar consultas analíticas e facilitar a análise multidimensional. Neste capítulo, são apresentados os principais aspectos da modelagem, desde a análise das características da base até a definição do esquema estrela implementado no ClickHouse.

3.1 Características da Base

A base utilizada neste projeto possui dados reais e anonimizados do marketplace brasileiro Olist. Os registros abrangem aproximadamente 100 mil pedidos, contendo informações detalhadas de pedidos, pagamentos, produtos, clientes, vendedores, avaliações e geolocalização.

Algumas das características relevantes da base incluem:

- Presença de dados históricos entre 2016 e 2018.
- Estrutura relacional, porém com arquivos separados para cada entidade.
- Variedade de dados categóricos e numéricos.
- Necessidade de tratamento de dados faltantes e inconsistentes.

Essas características exigiram uma abordagem de integração e transformação robusta, que culminou na criação de uma base unificada e preparada para análise OLAP.

3.2 Modelagem de 4 Passos

A construção do Data Warehouse seguiu a metodologia de modelagem dimensional em quatro etapas, conforme proposto por Ralph Kimball. A seguir, detalhamos cada um desses passos conforme aplicado ao contexto do dataset da Olist:

1. **Fato: vendas**
   O objetivo central do modelo é medir o processo de venda de produtos. A tabela fato `fct_sales` registra cada item vendido, permitindo análise detalhada de métricas como valor da venda, frete e total pago pelo cliente.

2. **Granularidade: Um item vendido em uma venda**
   A granularidade da tabela fato é definida como o nível mais detalhado possível: cada linha representa um item individual vendido em um determinado pedido. Essa granularidade permite flexibilidade para análises tanto no nível de item quanto agregações por pedido, cliente, produto ou tempo.

3. **Dimensões: Contexto da análise**
   As dimensões foram escolhidas com base nas principais perspectivas pelas quais as vendas podem ser analisadas. As dimensões utilizadas são:

   - `dim_product`: informações sobre o produto vendido.
   - `dim_seller`: dados do vendedor responsável pela venda.
   - `dim_customer`: dados do cliente que realizou o pedido.

- `dim_date`: data e hora da venda.
- `dim_payment`: método e condições de pagamento.

4. **Fatos: Métricas quantitativas**
   As principais métricas armazenadas na tabela fato incluem:

   - `price`: valor do item vendido.
   - `freight_value`: custo do frete associado ao item.
   - `total`: soma do preço do item e do valor do frete.

3.3 Esquema Estrela

O esquema estrela implementado é composto por uma tabela fato e cinco tabelas dimensionais, conforme a seguir:

Figura 1: Modelagem estrela

Esse modelo permite consultas rápidas e flexíveis, como agregações por tempo, categoria de produto, região geográfica ou método de pagamento.

3.4 Bus do DW

Este esquema tem como finalidade mapear os principais processos ou eventos do negócio, permitindo visualizar interseções entre eles. Essa análise facilita a identificação de dimensões compartilhadas, servindo como base para a definição e construção de Data Marts especializados.

Tabela 1: Matriz Bus do DW

Fato / Dimensão	**Vendas**	**Vendendor**	**Produto**	**Data**	**Pagamento**	**Cliente**
`Registro de vendas`	X	X	X	X	X	X
`Cadastro de produto`			X			
`Cadastro de vendendor`		X				
`Cadastro de cliente`						X
`Processamento de pagamento`				X	X	X
`Envio do Pedido`	X	X	X	X		X
`Entrega do Produto`	X	X	X	X		X
`Comportamento do Cliente`	X		X	X	X	X
`Desempenho do Vendedor`	X	X	X	X		

4. Insights e Visualizações

Considerando os dados utilizados, pode-se extrair algumas conjunturas relevantes para a análise, como os seguintes temas: Vendas e Receita; Produtos e Categorias; Geolocalização de Clientes e Vendedores. A partir disso, os dados foram visualizados, a fim de realizar a extração de insights e informações relevantes sobre essas áreas.

4.1 - Vendas e Receita

Análises temporais do volume de vendas, tanto por horários do dia quanto ao longo dos meses, revelaram alguns padrões comportamentais e tendências de crescimento.

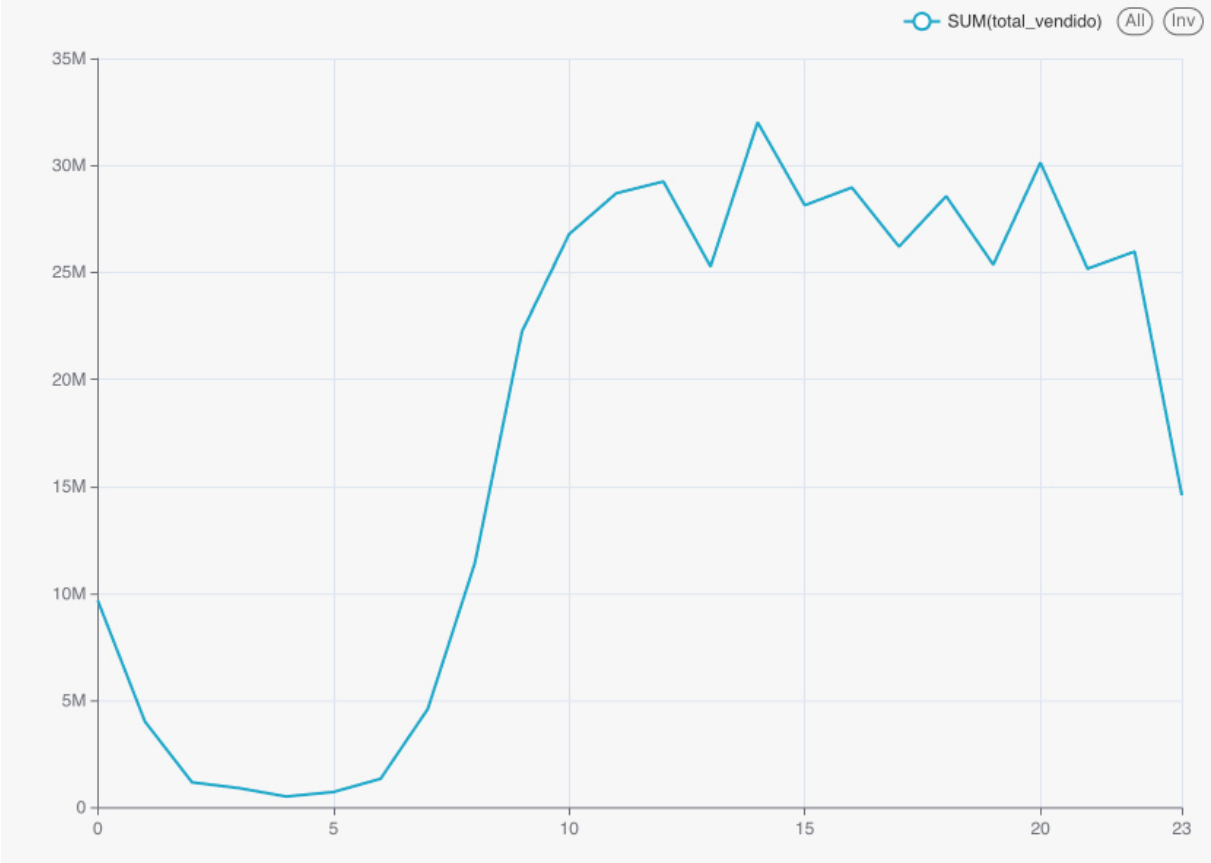

Figura 2: Volume de vendas por hora do dia

Vendas por horário Os dados mostram dois picos principais:

- No início da tarde, por volta das 14h, concentrando cerca de (12%) do volume diário total;
- Já ao redor das 20h, ocorre um novo aumento, com cerca de (10%) do volume diário.

Esses dois horários representam juntos cerca de (22%) do faturamento diário, sugerindo que campanhas de marketing e promoções podem ter maior impacto se disparadas nesses horários de maior conversão.

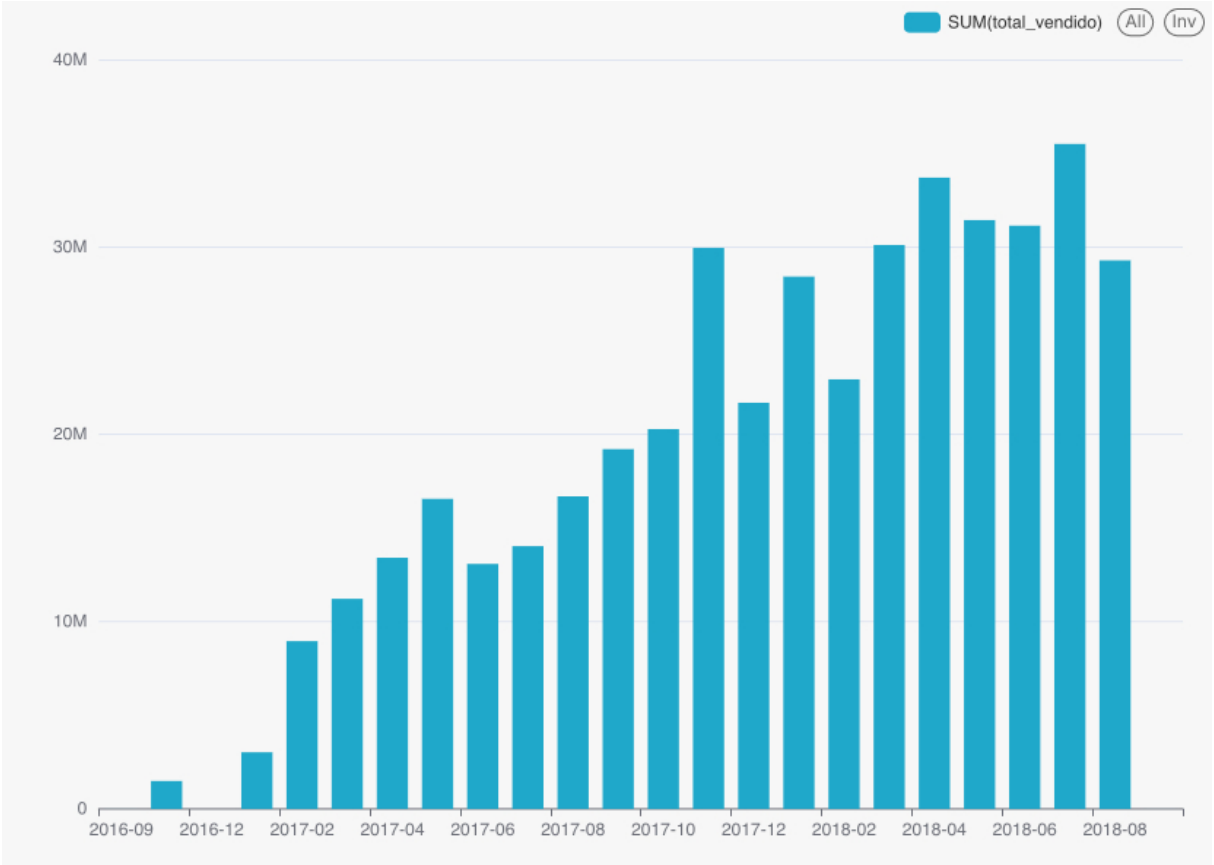

Figura 3: Total de vendas mensais (set/2016 – ago/2018)

Observa-se um crescimento constante da receita ao longo do período (setembro/2016–agosto/2018), com picos em outubro/2017 (crescimento de +25% em relação ao mês anterior), abril/2018 (+30%) e julho/2018 (+20%). Isso sugere:

- Consolidação da plataforma Olist, com expansão de sua base de consumidores e diversidade de produtos;
- Influência de eventos sazonais — como Black Friday, Páscoa e promoções de meio de ano — que potencializam a receita nesses meses específicos.
- Consolidação da plataforma Olist, com expansão de sua base de consumidores e diversidade de produtos;
- Influência de eventos sazonais – como Black Friday, Páscoa e promoções de meio de ano – que potencializam a receita nesses meses específicos.

4.2 - Produtos e Categorias

A Figura 2 mostra o fluxo de valor de vendas de acordo com as diferentes categorias. É possível perceber, segundo tal visualização, as seguintes informações relevantes:

- "Beleza e Saúde"lidera com o maior valor de vendas (38,2M), seguida por "Esporte e Lazer"(31,4M), indicando uma forte demanda dos consumidores nessas áreas.

- Categorias como "Construção"e "Eletrônicos"apresentam valores significativos, mas menores (5,24M e 4,9M, respectivamente), sugerindo mercados de nicho com potencial de crescimento.
- Categorias menores, como "Seguros e Serviços"e "CDs, DVDs e Mídias"têm vendas mínimas (por volta de 10k), apontando para um interesse limitado ou saturação de mercado.

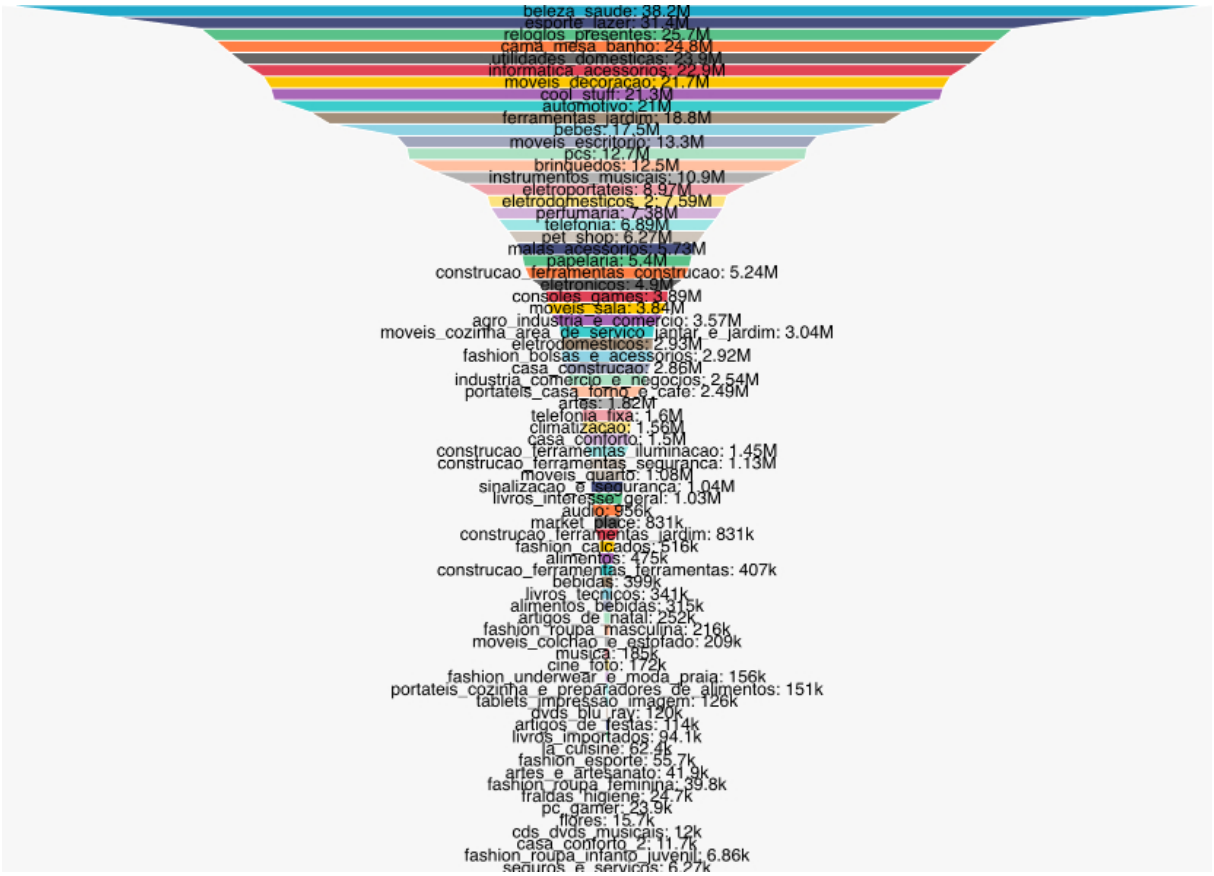

Figura 4: Fluxo de valor de vendas para diferentes categorias.

4.3 - Geolocalização de Clientes e Vendedores

O grafo ilustrado na Figura 3 demonstra a conectividade entre os estados brasileiros com base na origem e no destino de vendas. Os principais insights incluem:

- São Paulo (SP) destaca-se como o hub central, com o maior número de conexões com outros estados, indicando seu papel dominante nas transações de e-commerce.
- Estados como Minas Gerais (MG), Rio de Janeiro (RJ) e Paraná (PR) também apresentam significativa conectividade, sugerindo fortes redes regionais de comércio.
- Alguns estados do Norte e Nordeste, como Acre, Amapá e Roraima, exibem menos conexões, apontando para uma menor penetração de e-commerce nessas regiões.

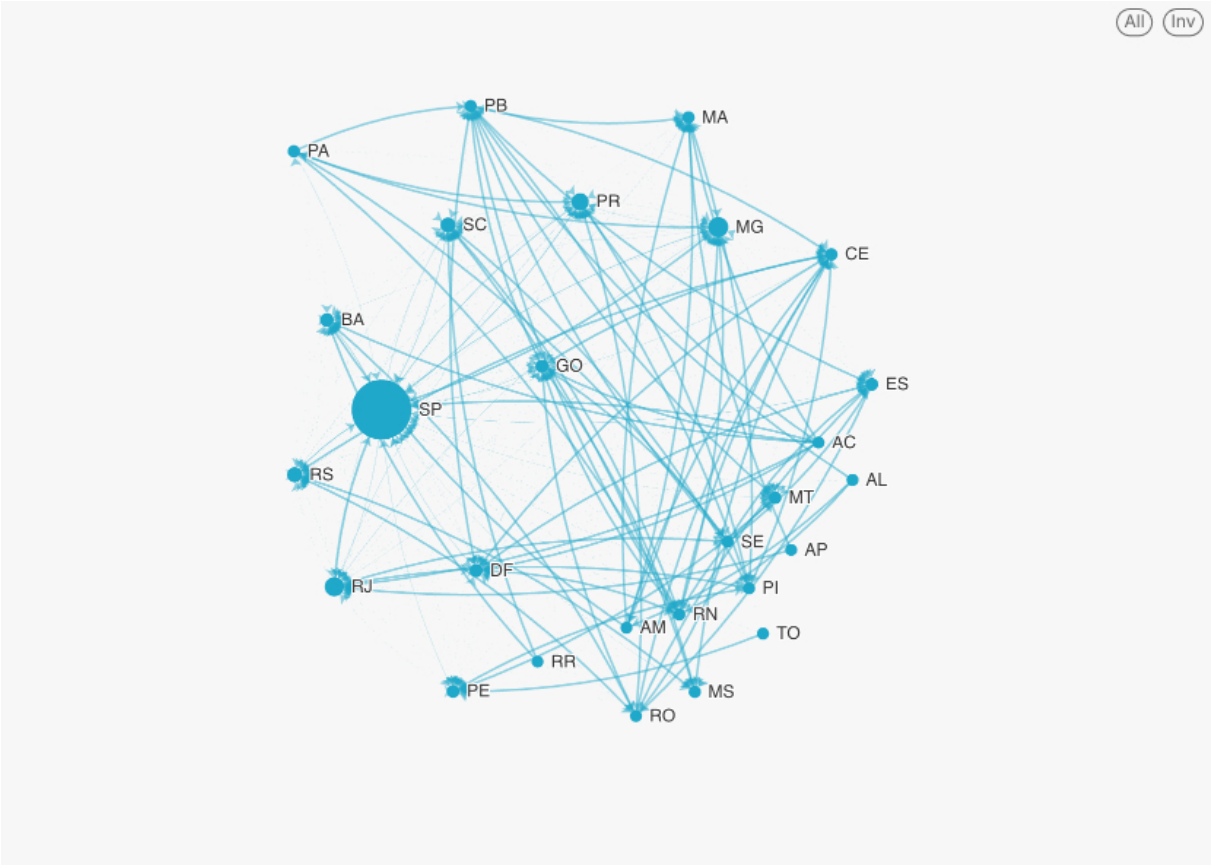

Figura 5: Grafo ilustrando as interações entre os estados brasileiros com base nas transações de vendas.

5. Conclusão

Com o crescimento exponencial na geração de dados, torna-se cada vez mais necessário adotar abordagens estruturadas para armazenamento e análise dessas informações. A construção de um Data Warehouse se mostra essencial nesse cenário, permitindo a integração de dados de diversas origens e viabilizando análises consistentes e em larga escala.

Neste trabalho, desenvolvemos um armazém de dados a partir de um conjunto real de informações de vendas do comércio eletrônico brasileiro. Utilizando o ClickHouse como banco de dados colunar, foi possível garantir alta performance nas consultas analíticas. As visualizações dos dados foram realizadas no Apache Superset, que se destacou por sua facilidade de uso e capacidade de criar dashboards interativos com diferentes perspectivas sobre os dados.

Apesar da complexidade envolvida nas etapas de preparação, transformação e modelagem dos dados, os resultados obtidos foram significativos. As análises permitiram identificar padrões de comportamento dos consumidores, horários de maior volume de vendas, preferências por categorias de produtos, além de insights sobre regiões e formas de pagamento mais utilizadas.

Esse processo evidencia a importância do uso de soluções modernas de Business Intelligence e da aplicação de boas práticas na modelagem dimensional, reforçando o valor de projetos de Data Warehousing no apoio à tomada de decisões baseada em dados.