Github Repo: <https://github.com/divyagarigipati/datawarehouse/tree/main>

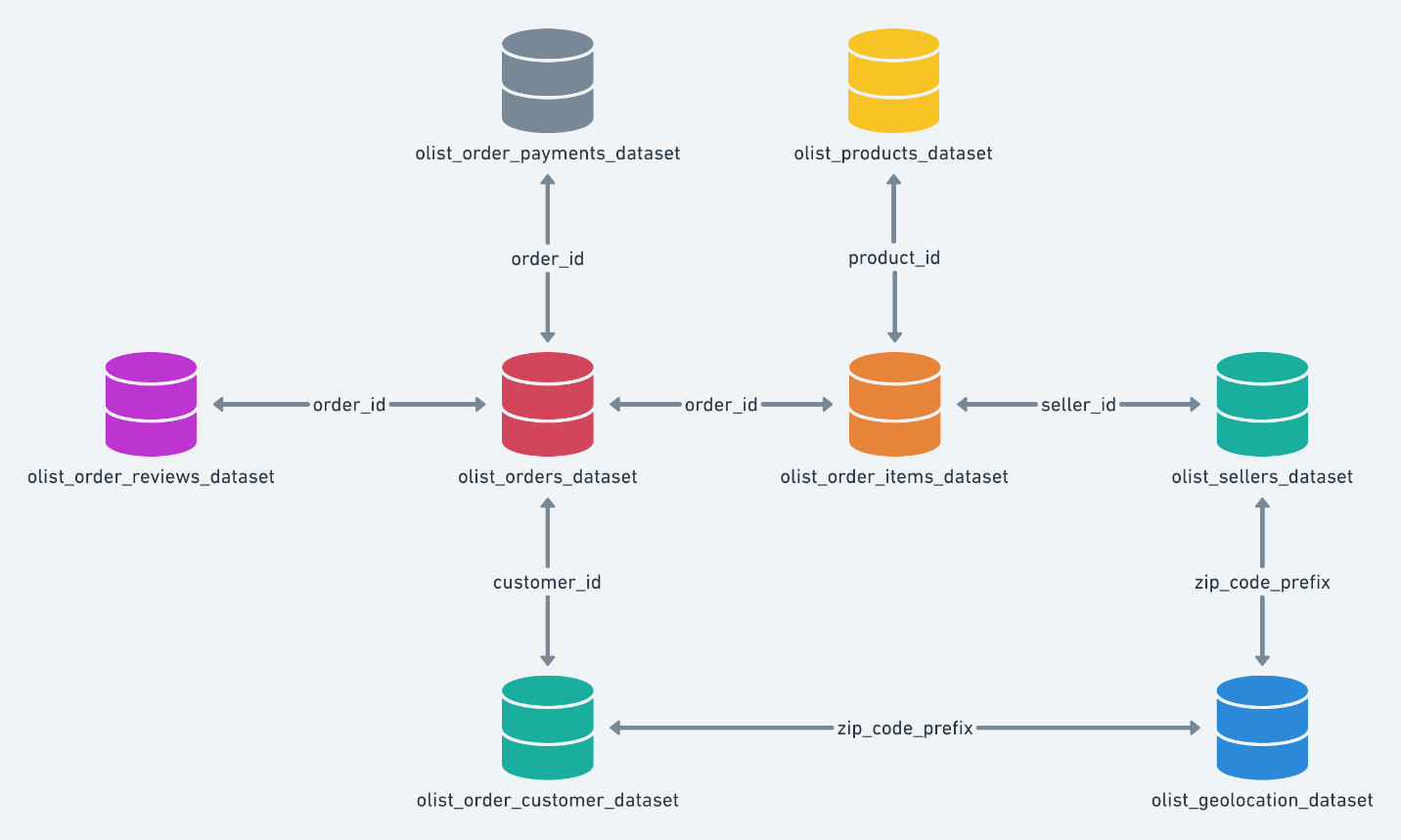
Notebook: <https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3969169828089009/2859625473491059/2340490761257204/latest.html>

Data Source:

Raw data files

<https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_customers_dataset.csv>

Initial Data Schema:



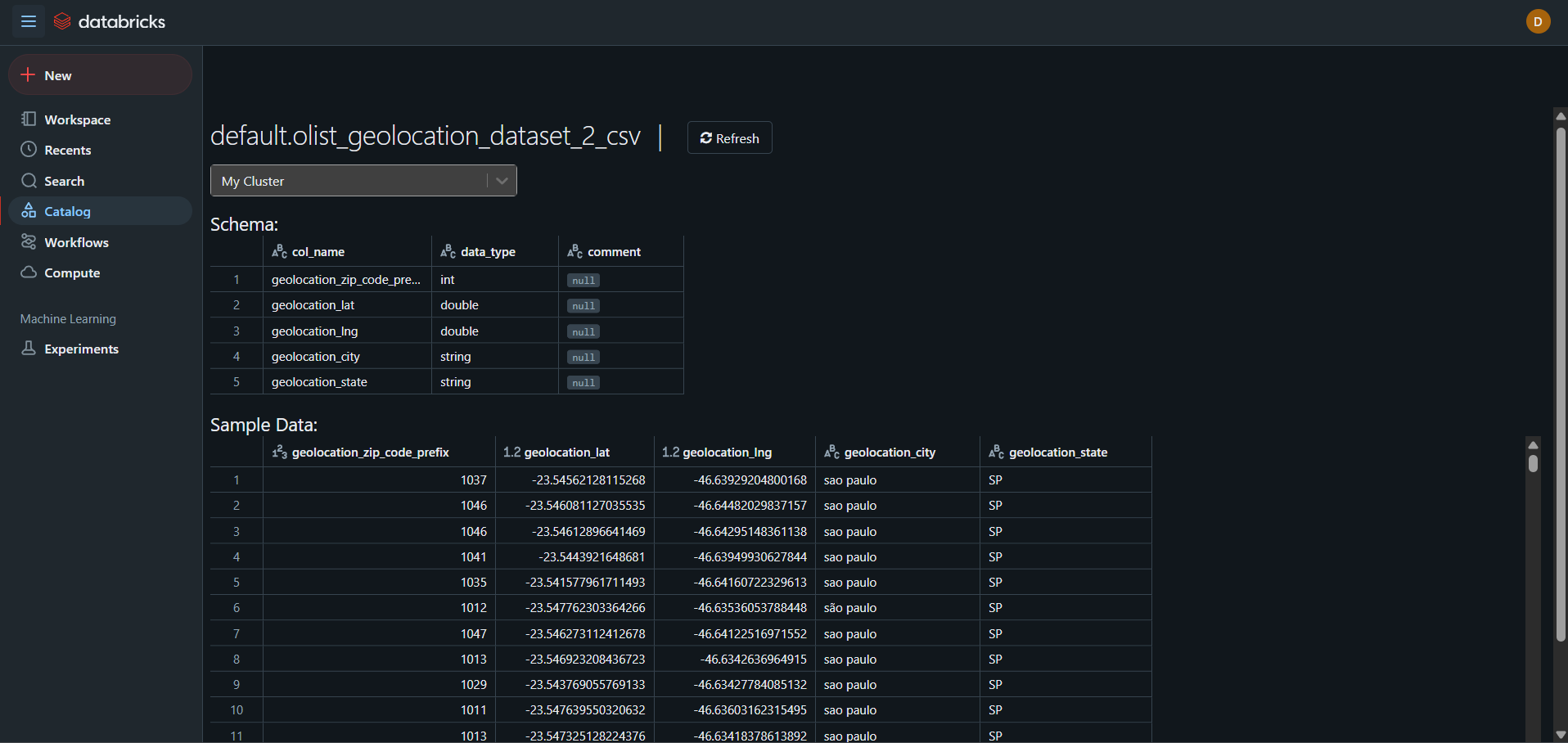
The raw e-commerce database has a relational structure, where the primary table is olist\_orders\_dataset which is an order lifecycle table combining different aspects. There is a olist\_order\_customer\_dataset table where customer information is stored, and it is linked by customer\_id. The olist\_order\_items\_dataset includes order details and references olist\_products\_dataset` for the product details, as well as olist\_sellers\_dataset for seller information. In olist\_geolocation\_dataset, zip\_code\_prefix connects sellers to geographic data. Orders are linked to their payments in olist\_order\_payments\_dataset and their reviews in olist\_order\_reviews\_dataset. This schema provides a structured way to analyze the customer transactions, sellers’ performance, delivery efficiency to fetch complete business insights.

Extracting Data:

Loading tables into data bricks hive metastore. The first row will be header, and schema will be inferred by databricks

The following images also contain the data dictionary (inferredschema) for each table.

Gelocations:



A screenshot of a computer

AI-generated content may be incorrect.

Order items:

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Order Payments:

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Order reviews

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Orders:

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Products:

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Sellers:

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Customers:

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Translation:

A screenshot of a computer

AI-generated content may be incorrect.

Loading the data into dimensions:

A screenshot of a computer program

AI-generated content may be incorrect.

Displaying Schema of dimension:

A screenshot of a computer

AI-generated content may be incorrect.

Customer\_dim

root

|-- customer\_id: string (nullable = true)

|-- customer\_unique\_id: string (nullable = true)

|-- customer\_zip\_code\_prefix: integer (nullable = true)

|-- customer\_city: string (nullable = true)

|-- customer\_state: string (nullable = true)

Geolocation\_dim

root

|-- geolocation\_zip\_code\_prefix: integer (nullable = true)

|-- geolocation\_lat: double (nullable = true)

|-- geolocation\_lng: double (nullable = true)

|-- geolocation\_city: string (nullable = true)

|-- geolocation\_state: string (nullable = true)

Order\_itmes\_dim

root

|-- order\_id: string (nullable = true)

|-- order\_item\_id: integer (nullable = true)

|-- product\_id: string (nullable = true)

|-- seller\_id: string (nullable = true)

|-- shipping\_limit\_date: timestamp (nullable = true)

|-- price: double (nullable = true)

|-- freight\_value: double (nullable = true)

Order\_payments\_dim

root

|-- order\_id: string (nullable = true)

|-- payment\_sequential: integer (nullable = true)

|-- payment\_type: string (nullable = true)

|-- payment\_installments: integer (nullable = true)

|-- payment\_value: double (nullable = true)

Order\_reviews\_dim

root

|-- review\_id: string (nullable = true)

|-- order\_id: string (nullable = true)

|-- review\_score: string (nullable = true)

|-- review\_comment\_title: string (nullable = true)

|-- review\_comment\_message: string (nullable = true)

|-- review\_creation\_date: string (nullable = true)

|-- review\_answer\_timestamp: timestamp (nullable = true)

Orders\_dim

root

|-- order\_id: string (nullable = true)

|-- customer\_id: string (nullable = true)

|-- order\_status: string (nullable = true)

|-- order\_purchase\_timestamp: timestamp (nullable = true)

|-- order\_approved\_at: timestamp (nullable = true)

|-- order\_delivered\_carrier\_date: timestamp (nullable = true)

|-- order\_delivered\_customer\_date: timestamp (nullable = true)

|-- order\_estimated\_delivery\_date: timestamp (nullable = true)

Products\_dim

root

|-- product\_id: string (nullable = true)

|-- product\_category\_name: string (nullable = true)

|-- product\_name\_lenght: integer (nullable = true)

|-- product\_description\_lenght: integer (nullable = true)

|-- product\_photos\_qty: integer (nullable = true)

|-- product\_weight\_g: integer (nullable = true)

|-- product\_length\_cm: integer (nullable = true)

|-- product\_height\_cm: integer (nullable = true)

|-- product\_width\_cm: integer (nullable = true)

Sellers\_dim

root

|-- seller\_id: string (nullable = true)

|-- seller\_zip\_code\_prefix: integer (nullable = true)

|-- seller\_city: string (nullable = true)

|-- seller\_state: string (nullable = true)

Prod\_transalation\_dim

root

|-- product\_category\_name: string (nullable = true)

|-- product\_category\_name\_english: string (nullable = true)

Getting Max and Min dates to generate date dimension:

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

Generating Date Dim:

A screenshot of a computer program

AI-generated content may be incorrect.

Creating fact table by joining all the dimensions:

A computer screen shot of a program

AI-generated content may be incorrect.

Fact and dimension ERD in star Schema:

**A screenshot of a computer

AI-generated content may be incorrect.**

The ERD represents a star schema designed for an e-commerce sales analysis system. At the center is the **fact\_table** table, which records transactional data, including order details, prices, freight costs, payment values, and review scores. It is linked to multiple dimension tables that provide descriptive context.

The **customer\_dim** stores customer details, while **sellers\_dim** contains seller information. **products\_dim** holds product attributes, and its relationship with **prod\_translation\_dim** allows category name translations. **orders\_dim** tracks various order-related timestamps, enabling order lifecycle analysis. **date\_dim** facilitates time-based reporting, and **geolocation\_dim** maps geographic details using zip code prefixes. The structure optimizes analytical queries by reducing data redundancy and improving performance.

Analytical Queries:

Average Fulfillment Time by Product Category

From the plot we can see that the artes\_e\_artesanato category have the lowest fulfillment within 5 days and then movies\_escritorio have the highest fulfillment may take up to 20 days. With average fulfillment in 10 days range for all other categories.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Year-Over-Year Revenue:

From the below graph we can see that there is huge gap between 2016 and 2017 this might be due to the fact that we are missing data form 2016. From 2017 to 2018 the revenue has increased from around 6500000 to about 8750000 that’s equal to 30% increase.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Average Order Value By Month To Get Seasonal Trends:

We can see that our data starts from 9/2016 this might’ve been contributed to the low revenue in previous analysis. We can see that from 1-2017 every month we have average orders of around 120/month.

A screenshot of a computer

AI-generated content may be incorrect.

A screen shot of a graph

AI-generated content may be incorrect.

Summary:

The project implements a data warehouse by converting a raw, normalized e-commerce dataset into a star schema dimensional model. The central fact table (fact\_sales) stores key transactional data, while surrounding dimension tables (e.g., customer\_dim, products\_dim, orders\_dim) provide descriptive context for each dimension. Using PySpark, I’ve efficiently processed complex queries such as Year-over-Year revenue, top-selling categories, and fulfillment times. This approach streamlined the data model, improved performance, and enabled flexible business intelligence reporting.

Challenges included data quality issues, ambiguous column names, and data type mismatches, which were resolved by renaming columns and casting types. The star schema reduced query complexity and improved performance. Future improvements include implementing surrogate keys, enhancing the ETL pipeline for better data quality and optimizing query performance and report automation for better scalability and maintenance.