



# Northeastern University

## College of Engineering

Data Warehousing & Integration

IE 6750

FALL 2024

## [Retail sales and Inventory Analysis]

AWS Project Report

Group 11

Student 1- Jahnavi Redyy Ganesina

Student 2- Manasi Bondalapati

Student 1 [ganesina.j@northeastern.edu](mailto:ganesina.j@northeastern.edu)

Student 2 [bondalapati.m@northeastern.edu](mailto:bondalapati.m@northeastern.edu)

Submission Date: \_\_\_\_\_ 12/07/24 \_\_\_\_\_

## **Table Of Contents**

| <b>Topic</b>         | <b>Page No.</b> |
|----------------------|-----------------|
| 1. Problem statement | 2               |
| 2. ETL processes     | 3               |
| 3. Analyzing trends  | 11              |
| 5. Conclusion        | 24              |
| 6. References        | 25              |

## **Problem Statement**

The goal of this project is to analyze the Brazilian e-commerce dataset in sales, customer behavior, and operational efficiency. By implementing a data pipeline, we aim to process and analyze large volumes of retail data to identify trends, optimize inventory etc

The primary objective is to implement a cloud-based data pipeline using AWS services to process and analyze large volumes of retail data. This approach will enable the identification of trends, optimization of inventory management, and enhancement of overall business operations

The project utilizes various AWS services to create an efficient ETL (Extract, Transform, Load) pipeline. These services include AWS S3 for data storage, AWS Glue for data transformation, AWS Athena for data analysis, and AWS Step Functions for workflow orchestration

Here are the key KPIs and metrics that were analyzed.

### Sales Performance Metrics

- Total sales value by product category
- Revenue per seller
- Product-wise sales volume
- Sales trends over time

### Customer Behavior Metrics

- Average review scores
- Review distribution across product categories

### Operational Metrics

- Average freight costs per seller
- Seller performance by region

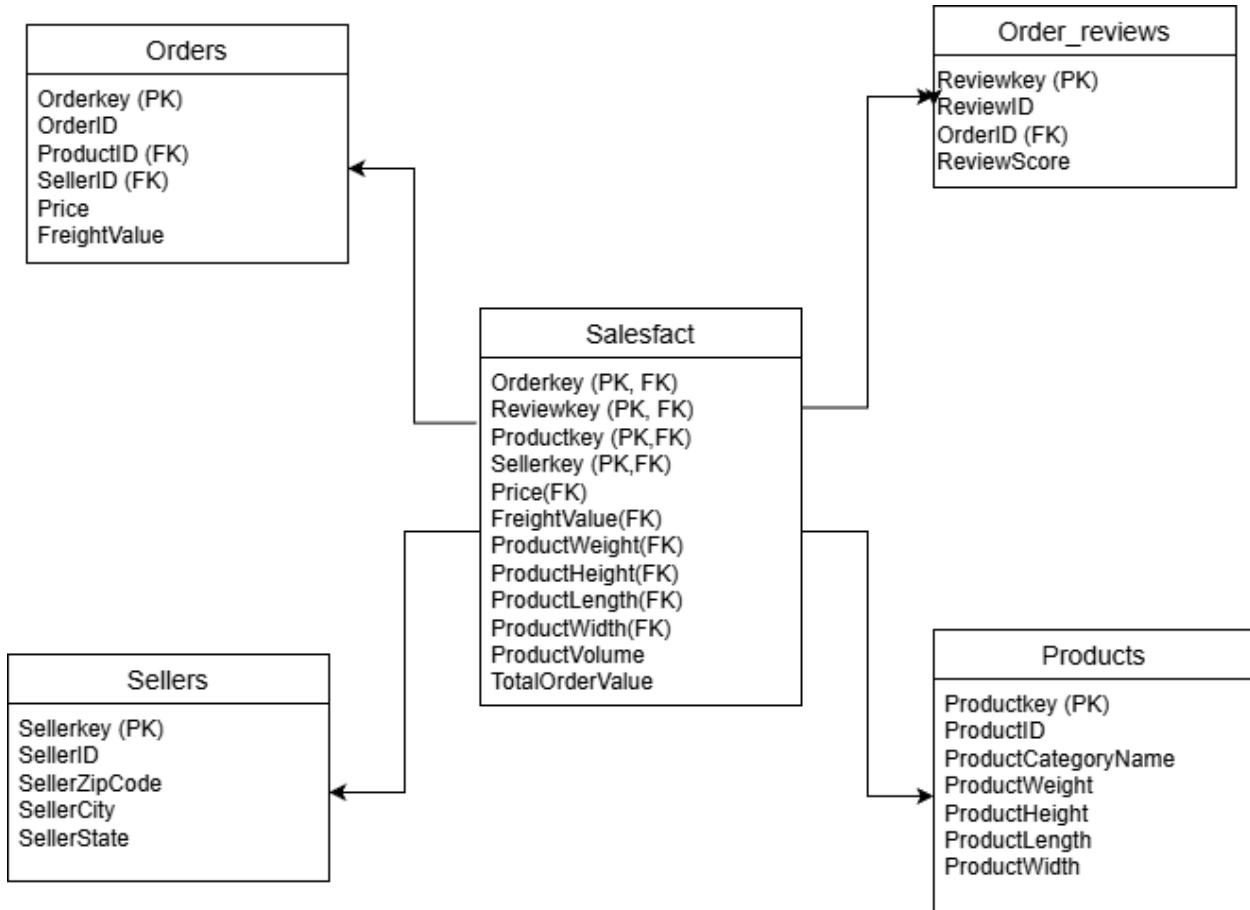
### Product Analytics

- Product dimensions (average length, width, height)
- Product weight metrics
- Category-wise product performance
- Top products by total sales value

### Seller Performance Metrics

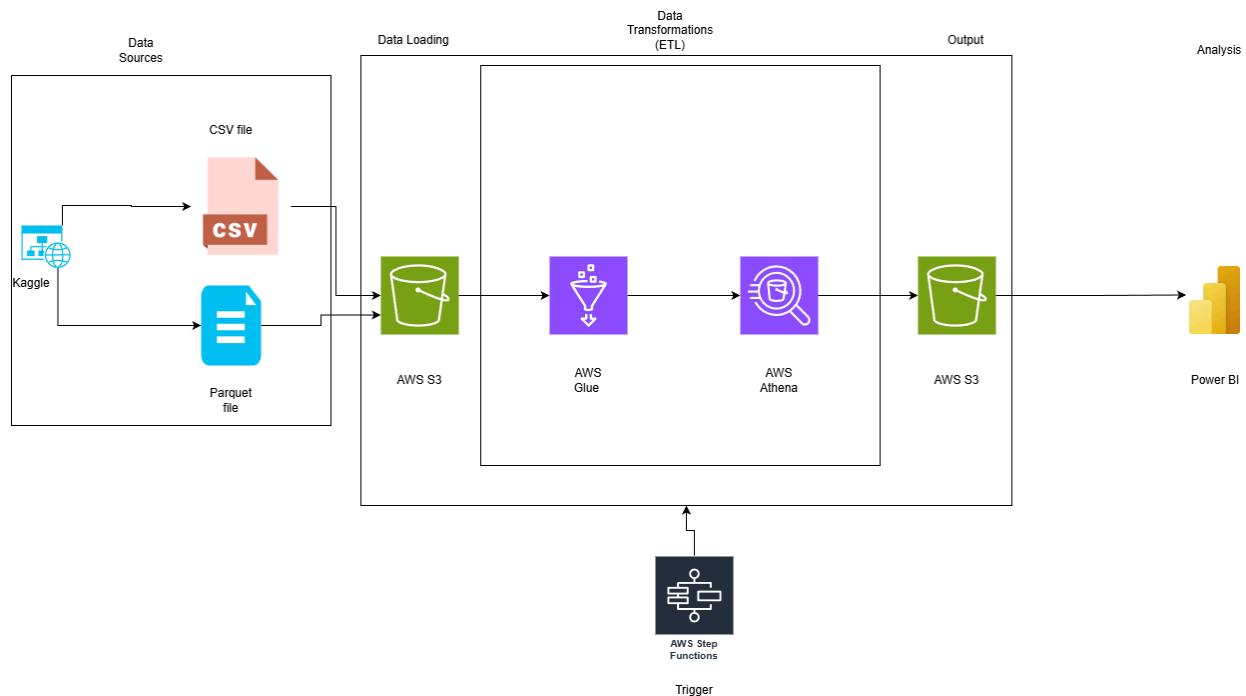
- Total sales value by seller
- Geographical distribution of sales
- Average review scores by seller
- Seller efficiency by state/region

## Logical Model



## ETL Process

### Data Architecture



### Data Sources

The pipeline processes multiple data files from the Brazilian e-commerce platform:

- Order Items (CSV format)
- Products (CSV format)
- Order Reviews (CSV format)
- Sellers (Parquet format)

### Data Loading Phase

#### AWS S3 Implementation

- Raw data files are uploaded to the initial S3 bucket
- AWS Step Functions orchestrates the entire ETL workflow

## S3 Bucket

Amazon S3

Buckets

Access Grants

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

Storage Lens groups

AWS Organizations settings

Feature spotlight [10]

oltp/

Objects (4) Info

| Name              | Type    | Last modified                           | Size     | Storage class |
|-------------------|---------|---|----------|---------------|
| order_reviews.csv | csv     | November 30, 2024, 12:17:15 (UTC-05:00) | 135.2 KB | Standard      |
| order.csv         | csv     | November 30, 2024, 12:17:15 (UTC-05:00) | 403.1 KB | Standard      |
| products.csv      | csv     | November 30, 2024, 12:17:16 (UTC-05:00) | 273.8 KB | Standard      |
| sellers.parquet   | parquet | November 30, 2024, 12:17:15 (UTC-05:00) | 26.7 KB  | Standard      |

## Data warehouse bucket (team11projectdw)

Amazon S3

Buckets

Access Grants

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

Storage Lens groups

AWS Organizations settings

olap/

Objects (5) Info

| Name               | Type   | Last modified | Size | Storage class |
|--------------------|--------|---------------|------|---------------|
| order_dimension/   | Folder | -             | -    | -             |
| product_dimension/ | Folder | -             | -    | -             |
| review_dimension/  | Folder | -             | -    | -             |
| salesfact/         | Folder | -             | -    | -             |
| seller_dimension/  | Folder | -             | -    | -             |

## Data Transformation Phase

### AWS Glue Job 1: Dimension Loading

- Processes four key dimensions:

- Order Items
- Products
- Reviews
- Sellers

- Performs transformations:

- Selects relevant columns
- Removes duplicate entries

- Assigns primary keys to dimensions
  - Outputs transformed dimensions as CSV files to S3

## AWS Glue Job 2: Fact Table Creation

- Creates the sales fact table through:
    - Joining dimension data using keys
    - Calculating business metrics (order value, product volume)
    - Assigning primary keys to fact rows
  - Stores fact table as CSV in output S3 bucket

AWS Glue

- 1) dim\_etl - executes the loading data into the dimensions of the data warehouse

The screenshot shows the AWS Glue job editor interface. The top navigation bar includes the AWS logo, a search bar, and various status indicators. On the left, a sidebar lists categories like AWS Glue, Data Catalog, Data Integration, and Legacy pages, along with links for What's New, Documentation, and the AWS Marketplace. The main content area is titled 'dim\_etl' and shows the 'Script' tab selected. The script code is as follows:

```
1 import boto3
2 import pandas as pd
3
4
5 # Input arguments for the Glue job
6 args = sys.argv
7 source_bucket_name = "team11project" # Source S3 bucket
8 target_bucket_name = "team11projectcdw" # Target S3 bucket
9 oltp_prefix = "oltp/" # Folder containing source data
10 olap_prefix = "olap/" # Folder for transformed data in the target bucket
11
12 # File paths for source data
13 order_items_file = f"s3://{source_bucket_name}/{oltp_prefix}order.csv"
14 order_reviews_file = f"s3:///{source_bucket_name}/{oltp_prefix}order_reviews.csv"
15 products_file = f"s3:///{source_bucket_name}/{oltp_prefix}products.csv"
16 sellers_file = f"s3:///{source_bucket_name}/{oltp_prefix}sellers.parquet"
17
18 # --- Step 1: Load the data into Pandas DataFrames ---
19 print("Loading data from S3...")
20 df_order_items = pd.read_csv(order_items_file)
21 df_order_reviews = pd.read_csv(order_reviews_file)
22 df_products = pd.read_csv(products_file)
23 df_sellers = pd.read_parquet(sellers_file)
24
```

Below the script, it says 'Python Ln 1, Col 1' and shows 0 errors and 0 warnings. The top right corner indicates the job was last modified on 11/30/2024 at 1:06:40 PM, with buttons for Actions, Save, and Run.

- 2) fact etl - executes loading data into fact table of the data warehouse

The screenshot shows the AWS Glue Data Catalog interface. On the left, there's a navigation sidebar with links like 'AWS Glue', 'Getting started', 'ETL jobs', 'Data Catalog tables', 'Data connections', 'Workflows (orchestration)', 'Data Catalog' (expanded), 'Databases', 'Tables', 'Stream schema registries', 'Schemas', 'Connections', 'Crawlers', 'Classifiers', 'Catalog settings', 'Data Integration and ETL' (expanded), and 'Legacy pages'. Below that are links for 'What's New', 'Documentation', and 'AWS Marketplace'. At the bottom left is a 'Enable compact mode' button. The main area has a search bar at the top right with '[Option+S]' and a 'fact\_etl' job card. The card includes tabs for 'Script' (selected), 'Job details', 'Runs', 'Data quality', 'Schedules', 'Version Control', and 'Upgrade analysis - preview'. It also shows 'Last modified on 11/30/2024, 2:10:42 PM' and buttons for 'Actions', 'Save', and 'Run'. The 'Script' tab contains the following Python code:

```
import sys
import boto3
import pandas as pd
#
# Input arguments for the Glue job
args = sys.argv
source_bucket_name = "team11projectdw" # Source S3 bucket for dimension tables
target_bucket_name = "team11projectdw" # Target S3 bucket for SalesFact table
olap_prefix = "olap/" # Folder for transformed data
#
# File paths for existing dimension tables
order_dimension_file = f"s3://{{source_bucket_name}}/{{olap_prefix}}order_dimension/order_dimension.csv"
review_dimension_file = f"s3://{{source_bucket_name}}/{{olap_prefix}}review_dimension/review_dimension.csv"
product_dimension_file = f"s3://{{source_bucket_name}}/{{olap_prefix}}product_dimension/product_dimension.csv"
seller_dimension_file = f"s3://{{source_bucket_name}}/{{olap_prefix}}seller_dimension/seller_dimension.csv"
#
# --- Step 1: Load the dimension data into Pandas DataFrames ---
print("Loading dimension tables from S3...")
df_order_dimension = pd.read_csv(order_dimension_file)
df_review_dimension = pd.read_csv(review_dimension_file)
df_product_dimension = pd.read_csv(product_dimension_file)
df_seller_dimension = pd.read_csv(seller_dimension_file)
#
# --- Step 2: Create SalesFact Table ---
Python Ln 1, Col 1 ⚙️ Errors: 0 ⚙️ Warnings: 0
```

# Data Cataloging

- Implements Glue Crawlers for:
  - Products
  - Orders
  - Sellers
  - Order reviews
- Updates the AWS Glue Data Catalog automatically

## AWS Crawler:

The screenshot shows the AWS Glue Crawler interface. On the left, there's a sidebar with navigation links like Getting started, ETL jobs, Data Catalog tables, and Data Catalog. The main area is titled "Crawlers" and contains a table with the following data:

| Name               | State | Last run  | Last run times...  | Log      | Table changes fr... |
|--------------------|-------|-----------|--------------------|----------|---------------------|
| order_dimension... | Ready | Succeeded | November 30, 20... | View log | -                   |
| product_dimension  | Ready | Succeeded | November 30, 20... | View log | -                   |
| review_dimension   | Ready | Succeeded | November 30, 20... | View log | -                   |
| salesfact          | Ready | Succeeded | November 30, 20... | View log | -                   |
| seller_dimension   | Ready | Succeeded | November 30, 20... | View log | -                   |

# Data Analysis Phase

## AWS Athena Implementation

- Executes SQL queries on the transformed data
- Generates insights on:
  - Sales trends
  - Seller performance
  - Product category metrics
  - Customer review analysis

## AWS Athena

Select queries for dimensions & facts:

### 1) Order dimension

Completed  
Time in queue: 69 ms Run time: 712 ms Data scanned: 450.87 KB

Copy Download results

| #  | orderkey | order_id | product_id | seller_id | price  | freight_value | shipping_limit_date |
|----|----------|----------|------------|-----------|--------|---------------|---------------------|
| 1  | 1        | 107778   | 5929       | 243       | 180.0  | 20.45         | 6/19/18 9:18        |
| 2  | 2        | 2392     | 24600      | 2211      | 10.99  | 16.05         | 3/9/17 14:35        |
| 3  | 3        | 77830    | 10949      | 2188      | 49.99  | 8.88          | 6/6/18 17:18        |
| 4  | 4        | 99820    | 29654      | 1685      | 117.3  | 14.43         | 5/18/17 8:02        |
| 5  | 5        | 41298    | 11898      | 2346      | 58.9   | 13.77         | 5/3/18 18:15        |
| 6  | 6        | 111434   | 6682       | 2662      | 79.99  | 17.81         | 2/8/18 19:28        |
| 7  | 7        | 18925    | 31440      | 69        | 169.99 | 17.76         | 10/6/17 18:14       |
| 8  | 8        | 102448   | 25876      | 1249      | 19.9   | 15.1          | 1/5/18 15:09        |
| 9  | 9        | 33508    | 8871       | 1183      | 159.94 | 41.11         | 10/23/17 14:05      |
| 10 | 10       | 49560    | 26916      | 2618      | 223.0  | 19.44         | 6/14/18 10:55       |
| 11 | 11       | 73815    | 18491      | 317       | 29.9   | 8.72          | 1/15/18 7:22        |
| 12 | 12       | 88254    | 15035      | 282       | 67.99  | 8.88          | 6/13/18 11:15       |
| 13 | 13       | 9790     | 16844      | 2909      | 98.0   | 29.09         | 9/15/17 11:45       |
| 14 | 14       | 99605    | 31610      | 988       | 29.99  | 7.78          | 3/8/18 12:30        |

### 2) Product dimension

Completed  
Time in queue: 110 ms Run time: 524 ms Data scanned: 304.93 KB

Copy Download results

| #  | productkey | product_id | product_category_name | product_weight | product_length | product_height |
|----|------------|------------|-----------------------|----------------|----------------|----------------|
| 1  | 1          | 2          | artes                 | 1000.0         | 30.0           | 18.0           |
| 2  | 2          | 25         | moveis_decoracao      | 1800.0         | 40.0           | 10.0           |
| 3  | 3          | 38         | eletronicos           | 263.0          | 18.0           | 12.0           |
| 4  | 4          | 40         | moveis_decoracao      | 600.0          | 25.0           | 25.0           |
| 5  | 5          | 43         | cama_mesa_banho       | 6350.0         | 45.0           | 15.0           |
| 6  | 6          | 47         | cool_stuff            | 15350.0        | 47.0           | 40.0           |
| 7  | 7          | 48         | moveis_decoracao      | 1300.0         | 25.0           | 16.0           |
| 8  | 8          | 60         | brinquedos            | 250.0          | 20.0           | 14.0           |
| 9  | 9          | 72         | relogios_presentes    | 250.0          | 16.0           | 11.0           |
| 10 | 10         | 83         | cama_mesa_banho       | 1750.0         | 30.0           | 18.0           |
| 11 | 11         | 96         | moveis_decoracao      | 800.0          | 53.0           | 8.0            |
| 12 | 12         | 101        | moveis_decoracao      | 150.0          | 35.0           | 2.0            |
| 13 | 13         | 110        | utilidades_domesticas | 1200.0         | 23.0           | 22.0           |
| 14 | 14         | 116        | moveis_decoracao      | 1000.0         | 30.0           | 18.0           |

### 3)Seller dimension

The screenshot shows the Amazon Athena Query editor interface. At the top, there are navigation icons, a search bar, and a status bar indicating "N. Virginia" and "jahnavig". Below the header, the path "Amazon Athena > Query editor" is visible. On the left, a sidebar shows "Views (0)". The main area displays a table titled "Results (1,626)" with the following columns: #, sellerkey, seller\_id, seller\_zipcode, seller\_city, and seller\_state. The table contains 13 rows of data. At the bottom right of the results table, there are "Copy" and "Download results" buttons.

| #  | sellerkey | seller_id | seller_zipcode | seller_city       | seller_state |
|----|-----------|-----------|----------------|-------------------|--------------|
| 1  | 1         | 1         | 13023          | campinas          | SP           |
| 2  | 2         | 2         | 13844          | mogi guacu        | SP           |
| 3  | 3         | 5         | 12914          | bragance paulista | SP           |
| 4  | 4         | 7         | 55325          | brejao            | PE           |
| 5  | 5         | 8         | 16304          | penapolis         | SP           |
| 6  | 6         | 9         | 1529           | sao paulo         | SP           |
| 7  | 7         | 10        | 80310          | curitiba          | PR           |
| 8  | 8         | 13        | 1222           | sao paulo         | SP           |
| 9  | 9         | 18        | 70740          | brasilia          | DF           |
| 10 | 10        | 19        | 45810          | porto seguro      | BA           |
| 11 | 11        | 23        | 4156           | sao paulo         | SP           |
| 12 | 12        | 24        | 13320          | salto             | SP           |
| 13 | 13        | 26        | 30494          | belo horizonte    | MG           |

### 4)Review dimension

The screenshot shows the Amazon Athena Query editor interface. At the top, there are navigation icons, a search bar, and a status bar indicating "N. Virginia" and "jahnavig". Below the header, the path "Amazon Athena > Query editor" is visible. On the left, a sidebar shows "Views (0)". The main area displays a table titled "Results (8,705)" with the following columns: #, reviewkey, order\_id, and score. The table contains 14 rows of data. At the bottom right of the results table, there are "Copy" and "Download results" buttons.

| #  | reviewkey | order_id | score |
|----|-----------|----------|-------|
| 1  | 1         | 109963.0 | 5     |
| 2  | 2         | 85888.0  | 5     |
| 3  | 3         | 19453.0  | 5     |
| 4  | 4         | 59703.0  | 1     |
| 5  | 5         | 95614.0  | 5     |
| 6  | 6         | 71198.0  | 1     |
| 7  | 7         | 18064.0  | 5     |
| 8  | 8         | 674.0    | 5     |
| 9  | 9         | 106637.0 | 3     |
| 10 | 10        | 94276.0  | 1     |
| 11 | 11        | 88664.0  | 5     |
| 12 | 12        | 21373.0  | 5     |
| 13 | 13        | 109181.0 | 2     |
| 14 | 14        | 34836.0  | 5     |

## 5) Sales dimension

The screenshot shows the Amazon Athena Query Editor interface. At the top, there's a search bar and navigation links for 'Amazon Athena' and 'Query editor'. Below the search bar, it says 'Views (0)' and 'Completed'. The main area is titled 'Results (10,000)' with a 'Search rows' input field. To the right of the table are buttons for 'Copy' and 'Download results'. The table has a header row with columns: '#', 'orderkey', 'reviewkey', 'productkey', 'sellerkey', 'price', 'freight\_value', 'product\_weight', and 'product\_length'. The data rows show various values for these fields.

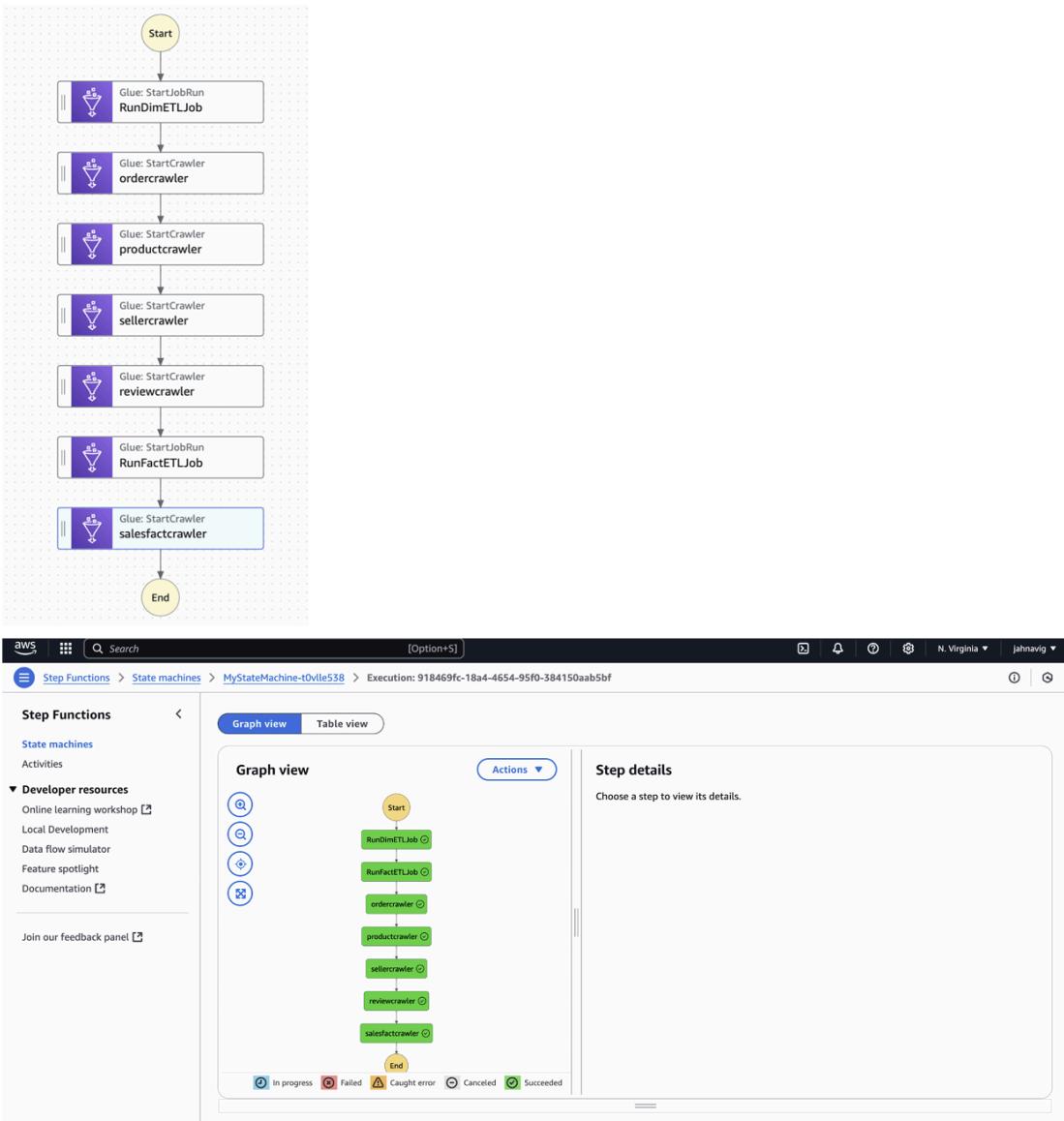
| #  | orderkey | reviewkey | productkey | sellerkey | price  | freight_value | product_weight | product_length |
|----|----------|-----------|------------|-----------|--------|---------------|----------------|----------------|
| 1  | 1        | 4207.0    | 1209       | 136       | 180.0  | 20.45         | 501.0          | 18.0           |
| 2  | 2        | 242.0     | 4925       | 1169      | 10.99  | 16.05         | 100.0          | 80.0           |
| 3  | 3        | 697.0     | 2190       | 1155      | 49.99  | 8.88          | 7875.0         | 60.0           |
| 4  | 4        | 5310.0    | 5968       | 885       | 117.3  | 14.43         | 4105.0         | 67.0           |
| 5  | 5        | 394.0     | 2380       | 1254      | 58.9   | 13.77         | 5700.0         | 39.0           |
| 6  | 6        | 169.0     | 1354       | 1407      | 79.99  | 17.81         | 650.0          | 25.0           |
| 7  | 7        | 6059.0    | 6303       | 42        | 169.99 | 17.76         | 433.0          | 21.0           |
| 8  | 8        | 6101.0    | 5192       | 646       | 19.9   | 15.1          | 125.0          | 20.0           |
| 9  | 9        | 4881.0    | 1778       | 611       | 159.94 | 41.11         | 13867.0        | 101.0          |
| 10 | 10       | 536.0     | 5414       | 1386      | 223.0  | 19.44         | 300.0          | 18.0           |
| 11 | 11       | 202.0     | 3732       | 173       | 29.9   | 8.72          | 475.0          | 20.0           |
| 12 | 12       | 8505.0    | 3020       | 152       | 67.99  | 8.88          | 450.0          | 16.0           |
| 13 | 13       | 8595.0    | 3387       | 1521      | 98.0   | 29.09         | 925.0          | 25.0           |
| 14 | 14       | 5170.0    | 6342       | 499       | 29.99  | 7.78          | 350.0          | 17.0           |

## Workflow Orchestration

### AWS Step Functions Process

1. Triggers dimension loading job
2. Executes fact table loading job
3. Runs Glue Crawlers for data catalog updates:
  - Product dimension
  - Review dimension

## AWS STEP FUNCTION



### Seller dimension

- Order dimension
- Sales fact

### Output Storage

- All transformed files are stored as CSV format
- Successfully created tables:
  - Dimension tables: Products, Sellers, Reviews, Orders
  - Fact table: Sales
- Data is made available for querying through AWS Athena

Power BI is used to analyze all of this

This ETL pipeline ensures efficient data processing, transformation, and analysis of the e-commerce dataset while maintaining data integrity and enabling comprehensive business insights.

## Analyzing Trends

### Analysis using AWS athena

#### 1. Top Products by Total Sales Value

```
SELECT
    p.product_category_name AS ProductCategory,
    p.product_id AS ProductID,
    SUM(sf.totalordervalue) AS TotalSalesValue
FROM
    salesfact sf
JOIN
    product_dimension p ON sf.Productkey = p.Productkey
GROUP BY
    p.product_category_name, p.product_id
ORDER BY
    TotalSalesValue DESC
LIMIT 10;
```

| Results (10)                     |                                   |           |                    |
|----------------------------------|-----------------------------------|-----------|--------------------|
| <input type="text"/> Search rows |                                   |           |                    |
| #                                | ProductCategory                   | ProductID | TotalSalesValue    |
| 1                                | beleza_saude                      | 19053     | 8592.489999999998  |
| 2                                | beleza_saude                      | 16152     | 7216.97            |
| 3                                | eletroportateis                   | 21198     | 4950.34            |
| 4                                | informatica_acessorios            | 8291      | 4491.309999999995  |
| 5                                | cama_mesa_banho                   | 29234     | 4350.120000000001  |
| 6                                | consoles_games                    | 30705     | 4175.26            |
| 7                                | esporte_lazer                     | 13562     | 4163.51            |
| 8                                | bebés                             | 25594     | 4037.89            |
| 9                                | relogios_presentes                | 794       | 3833.3800000000015 |
| 10                               | construcao_ferramentas_construcao | 30303     | 3830.66            |

## 2. Average Review Score by Seller

```
SELECT
    s.seller_id AS SellerID,
    s.seller_city AS SellerCity,
    s.seller_state AS SellerState,
    AVG(r.score) AS AverageReviewScore
FROM
    salesfact sf
JOIN
    seller_dimension s ON sf.Sellerkey = s.Sellerkey
JOIN
    review_dimension r ON sf.Reviewkey = r.Reviewkey
GROUP BY
    s.seller_id, s.seller_city, s.seller_state
ORDER BY
    AverageReviewScore DESC;
```

Results (1,533)

Copy   Download results

< 1 ... > |

| #  | SellerID | SellerCity         | SellerState | AverageReviewScore |
|----|----------|--------------------|-------------|--------------------|
| 1  | 1177     | santo andre        | SP          | 5.0                |
| 2  | 484      | curitiba           | PR          | 5.0                |
| 3  | 2559     | caxias do sul      | RS          | 5.0                |
| 4  | 2029     | sao paulo          | SP          | 5.0                |
| 5  | 2723     | sao caetano do sul | SP          | 5.0                |
| 6  | 3054     | sao paulo          | SP          | 5.0                |
| 7  | 1779     | sao paulo          | SP          | 5.0                |
| 8  | 1316     | colatina           | ES          | 5.0                |
| 9  | 150      | diadema            | SP          | 5.0                |
| 10 | 1492     | curitiba           | PR          | 5.0                |
| 11 | 2255     | curitiba           | PR          | 5.0                |

### 3. Distribution of Product Dimensions

```

SELECT
    p.product_category_name AS ProductCategory,
    AVG(p.product_length) AS AvgProductLength,
    AVG(p.product_width) AS AvgProductWidth,
    AVG(p.product_height) AS AvgProductHeight,
    AVG(p.product_weight) AS AvgProductWeight
FROM
    salesfact sf
JOIN
    product_dimension p ON sf.Productkey = p.Productkey
GROUP BY
    p.product_category_name
ORDER BY
    ProductCategory;

```

| #  | ProductCategory           | AvgProductLength   | AvgProductWidth    | AvgProductHeight   | AvgProductWeight |
|----|---------------------------|--------------------|--------------------|--------------------|------------------|
| 2  | agro_industria_e_comercio | 30.705882352941178 | 22.0               | 22.88235294117647  | 3371.0588235     |
| 3  | alimentos                 | 19.05128205128205  | 14.871794871794872 | 12.41025641025641  | 650.46153846     |
| 4  | alimentos_bebidas         | 22.11111111111111  | 17.333333333333332 | 17.40740740740741  | 1460.1851851     |
| 5  | artes                     | 44.705882352941174 | 33.588235294117645 | 9.705882352941176  | 1320.5882352     |
| 6  | artes_e_artesanato        | 18.666666666666668 | 13.666666666666666 | 9.666666666666666  | 866.666666666    |
| 7  | artigos_de_festas         | 29.666666666666668 | 24.166666666666668 | 19.83333333333332  | 1908.3333333     |
| 8  | artigos_de_natal          | 34.4               | 27.7               | 14.3               | 1955.0           |
| 9  | audio                     | 20.115384615384617 | 15.653846153846153 | 8.923076923076923  | 1593.7692307     |
| 10 | automotivo                | 33.44638403990025  | 24.98004987531172  | 16.01995012468828  | 2700.6059850     |
| 11 | bebes                     | 34.96989966555184  | 28.535117056856187 | 20.240802675585286 | 3255.8361204     |
| 12 | bebidas                   | 21.3181818181817   | 18.045454545454547 | 19.84090909090909  | 969.77272727     |
| 13 | beleza_saude              | 23.944380069524914 | 18.11239860950174  | 14.16106604866744  | 1045.5191193     |
| 14 | brinquedos                | 30.329479768786126 | 23.973988439306357 | 20.77456647398844  | 1577.1127167     |
| 15 | cama_mesa_banho           | 36.94449950445986  | 31.00792864222002  | 12.786917740336968 | 2050.3310208     |
| 16 | casa_comforto             | 45.84375           | 37.65625           | 16.0               | 3217.1875        |

#### 4. Seller Performance by Region

```
SELECT
    s.seller_state AS State,
    s.seller_id AS SellerID,
    SUM(sf.totalordervalue) AS TotalSalesValue
FROM
    salesfact sf
JOIN
    seller_dimension s ON sf.Sellerkey = s.Sellerkey
GROUP BY
    s.seller_state, s.seller_id
ORDER BY
    s.seller_state, TotalSalesValue DESC;
```

| #  | ▼   State         | ▼   SellerID | ▼   TotalSalesValue | ▼ |
|----|-------------------|--------------|---------------------|---|
| 1  | rio grande do sul | 552          | 90.03999999999999   |   |
| 2  | BA                | 902          | 15577.240000000002  |   |
| 3  | BA                | 2686         | 706.58              |   |
| 4  | BA                | 875          | 531.84              |   |
| 5  | BA                | 2454         | 510.22              |   |
| 6  | BA                | 2288         | 392.59              |   |
| 7  | BA                | 19           | 343.58              |   |
| 8  | BA                | 1971         | 224.45000000000002  |   |
| 9  | BA                | 406          | 223.44              |   |
| 10 | BA                | 2602         | 217.84              |   |

## 5. Review Distribution Across Product Categories

```
SELECT
    p.product_category_name AS ProductCategory,
    r.score AS ReviewScore,
    COUNT(r.score) AS ReviewCount
FROM
    salesfact sf
JOIN
    product_dimension p ON sf.Productkey = p.Productkey
JOIN
    review_dimension r ON sf.Reviewkey = r.Reviewkey
GROUP BY
    p.product_category_name, r.score
ORDER BY
    p.product_category_name, r.score;
```

Categories with Poor Ratings:

Results (282)

Search rows

| #   | ProductCategory                    | ReviewScore | ReviewCount |
|-----|------------------------------------|-------------|-------------|
| 99  | construcao_ferramentas_jardim      | 1           | 2           |
| 100 | construcao_ferramentas_jardim      | 2           | 1           |
| 94  | construcao_ferramentas_iluminacao  | 1           | 2           |
| 95  | construcao_ferramentas_iluminacao  | 2           | 3           |
| 96  | construcao_ferramentas_iluminacao  | 3           | 2           |
| 97  | construcao_ferramentas_iluminacao  | 4           | 7           |
| 98  | construcao_ferramentas_iluminacao  | 5           | 9           |
| 92  | construcao_ferramentas_ferramentas | 4           | 1           |
| 93  | construcao_ferramentas_ferramentas | 5           | 4           |
| 88  | construcao_ferramentas_construcao  | 1           | 5           |
| 89  | construcao_ferramentas_construcao  | 3           | 7           |
| 90  | construcao_ferramentas_construcao  | 4           | 11          |
| 91  | construcao_ferramentas_construcao  | 5           | 40          |

( Copy ) ( Download results )

< 1 ... > |

For construcao\_ferramentas\_jardim, there are 2 reviews with a score of 1 and 1 review with a score of 2. This suggests products in this category might need improvement.

Categories with Mixed Reviews:

For construcao\_ferramentas\_iluminacao, there are varying review scores (e.g., 1, 2, 3, 4) with differing counts. This indicates a mix of customer satisfaction levels.

## 6. Avg freight costs of sellers

```
SELECT
    s.seller_id AS SellerID,
    AVG(sf.freight_value) AS AverageFreightCost
FROM
    salesfact sf
JOIN
    seller_dimension s ON sf.Sellerkey = s.Sellerkey
GROUP BY
    s.seller_id
ORDER BY
    AverageFreightCost DESC;
```

Results (1,626)

Copy   Download results

Search rows

| #  | SellerID | AverageFreightCost |
|----|----------|--------------------|
| 54 | 3044     | 59.56              |
| 14 | 3039     | 119.99             |
| 86 | 3028     | 46.82333333333333  |
| 5  | 3018     | 143.78             |
| 34 | 3014     | 72.91              |
| 19 | 2973     | 115.14             |
| 17 | 2951     | 116.3              |
| 3  | 2950     | 146.3              |
| 8  | 2943     | 134.78             |
| 22 | 2893     | 108.74             |
| 15 | 2892     | 117.958            |
| 84 | 2887     | 47.25              |
| 55 | 2880     | 54.6               |

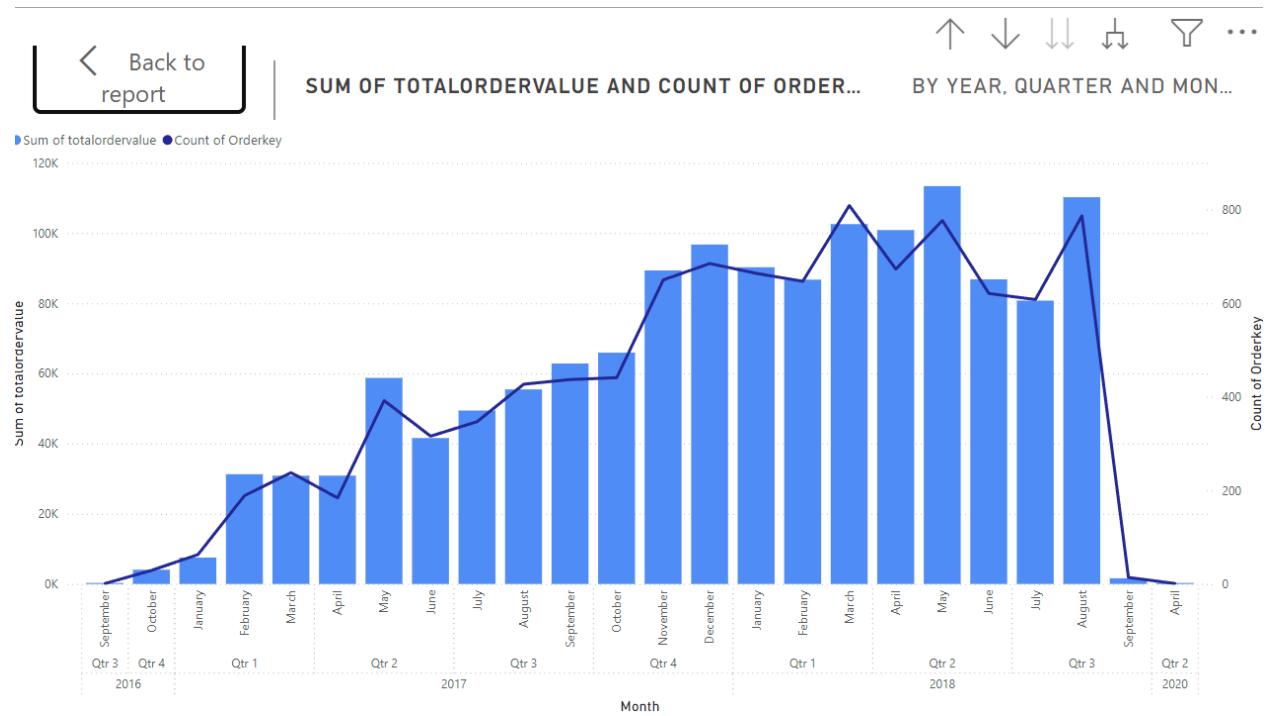
Based on the results of average freight costs by seller, there is a significant variation among sellers, ranging from efficient operations like Seller 3028 (with a low average of 46.82) to high-cost sellers such as 2950 (with an average cost of 146.3).

# Analysis Using Power BI

## 1. Sales Performance Analysis

- From Sales Fact Table: Orderkey, price, totalordervalue
- From Order Dimension Table: shipping\_limit\_date (if available)
- X-axis: Time (daily, weekly, or monthly depending on date range)
- Primary Y-axis: Total Sales Value (sum of totalordervalue)
- Secondary Y-axis: Number of Orders (count of Orderkey)
- Use a line for Total Sales Value and bars for Number of Orders

**Analysis:** This will show sales trends over time, allowing you to identify peak sales periods and overall growth.



## Key Performance Points

### Peak Performance

- Highest revenue point: ~120K (April-May 2018)
- Most consistent high-performance period: Q1-Q2 2018

### Growth Rate

- 2016-2017: Gradual upward trend
- 2017-2018: Steeper growth with higher volatility

## Business Cycles

- Monthly fluctuations show consistent patterns
- Order volume (blue line) closely tracks revenue trends
- Sharp decline after Q3 2018 indicates potential business disruption

## 2. Product Category Performance

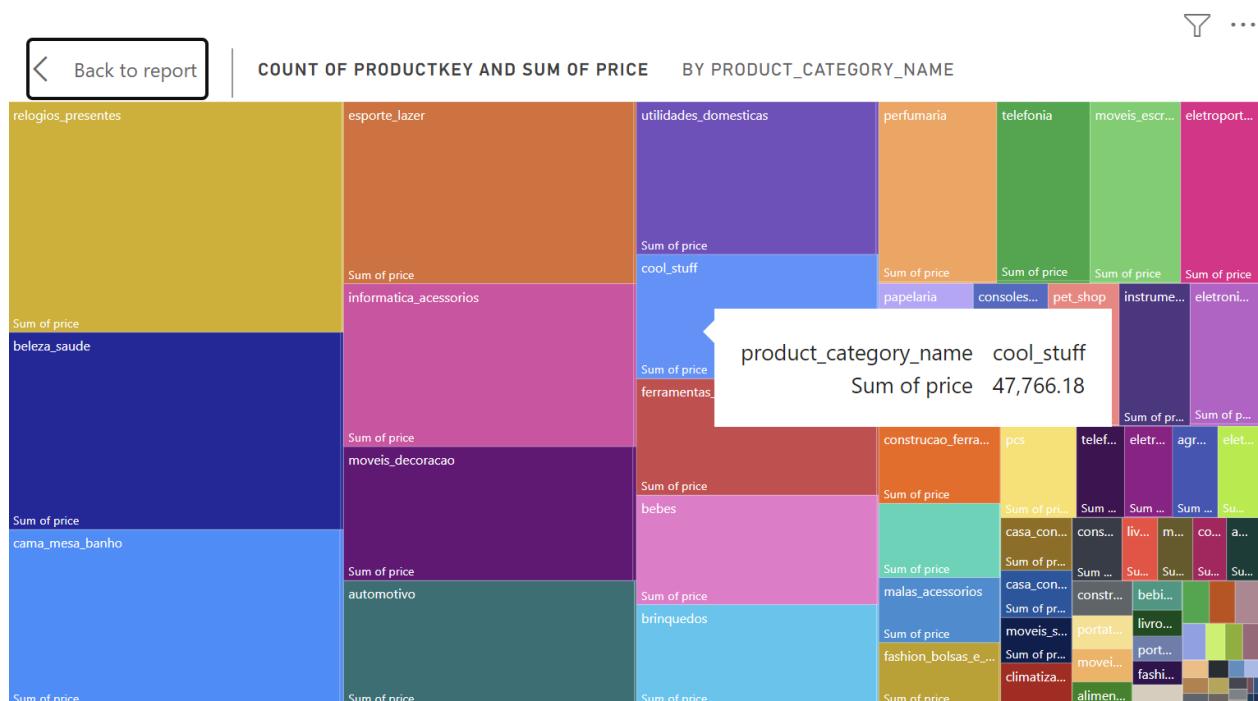
Fields:

- From Product Dimension Table: product\_category\_name
- From Sales Fact Table: price, Productkey

Design:

- Size of rectangles: Total Sales Value per category
- Color of rectangles: Number of products sold per category

**Analysis:** This visualization will quickly show which product categories are driving the most revenue and which are most popular in terms of units sold.



## Major Categories by Revenue

## Top Revenue Generators:

- "relogios\_presentes" (Watches/Gifts) shows the largest rectangle, indicating highest total sales value
- "esporte\_lazer" (Sports/Leisure) represents the second-largest revenue category
- "utilidades\_domesticas" (Household Utilities) ranks third in revenue generation

## Category Distribution

### Mid-Range Categories:

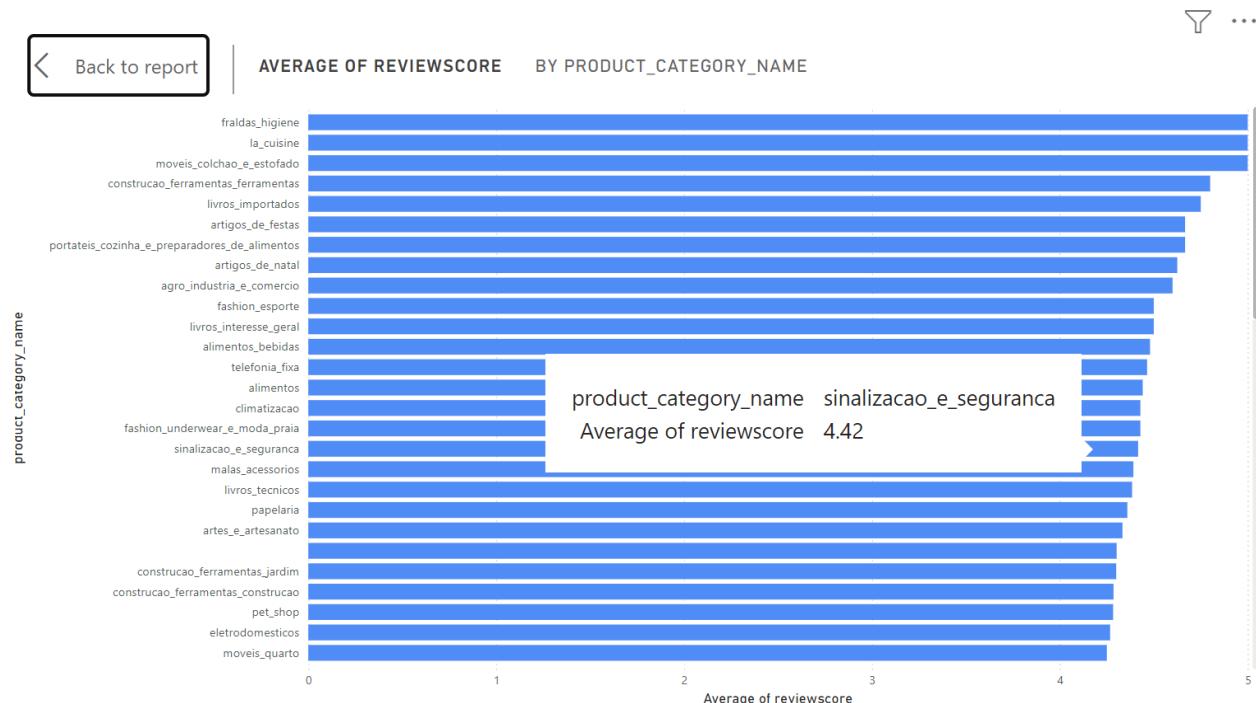
- "perfumaria" (Perfumes)
- "telefonia" (Telephony)
- "moveis\_escritorio" (Office Furniture)
- "eletroport" (Small Appliances)

### Notable Insights:

- "cool\_stuff" category shows a value of 47,766.18, visible in the tooltip

## 3. Customer Satisfaction Overview

**Analysis:** This will provide an at-a-glance view of customer satisfaction levels and the overall sentiment towards your products



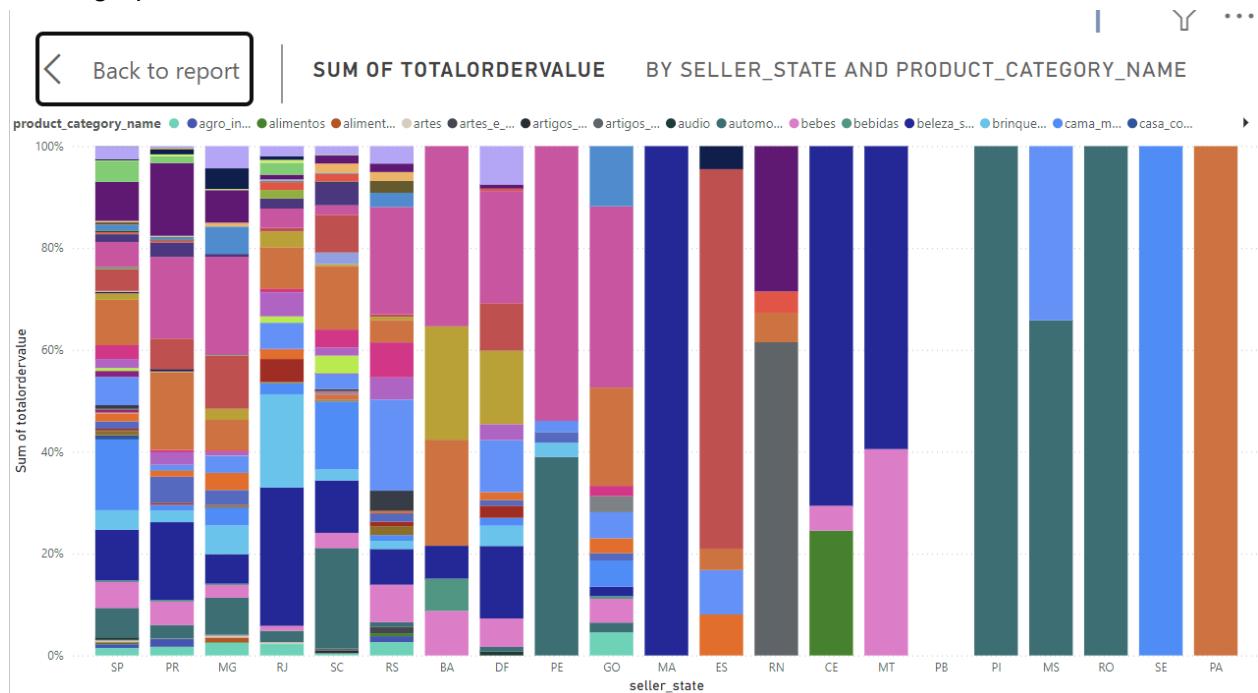
- Review Score Distribution:

- 5-star ratings: ~58% (dominant)
- 4-star ratings: ~19%
- 3-star ratings: ~8%
- 2-star ratings: ~3%
- 1-star ratings: ~11%

### Category Performance by Reviews

- Highest Rated Categories:
  - Fraldas\_higiene
  - La\_cuisine
  - Chao\_e\_estofado
- Most Reviewed Categories:
  - Utilidades\_domesticas
  - Esporte\_lazer
  - Moveis\_decoracao

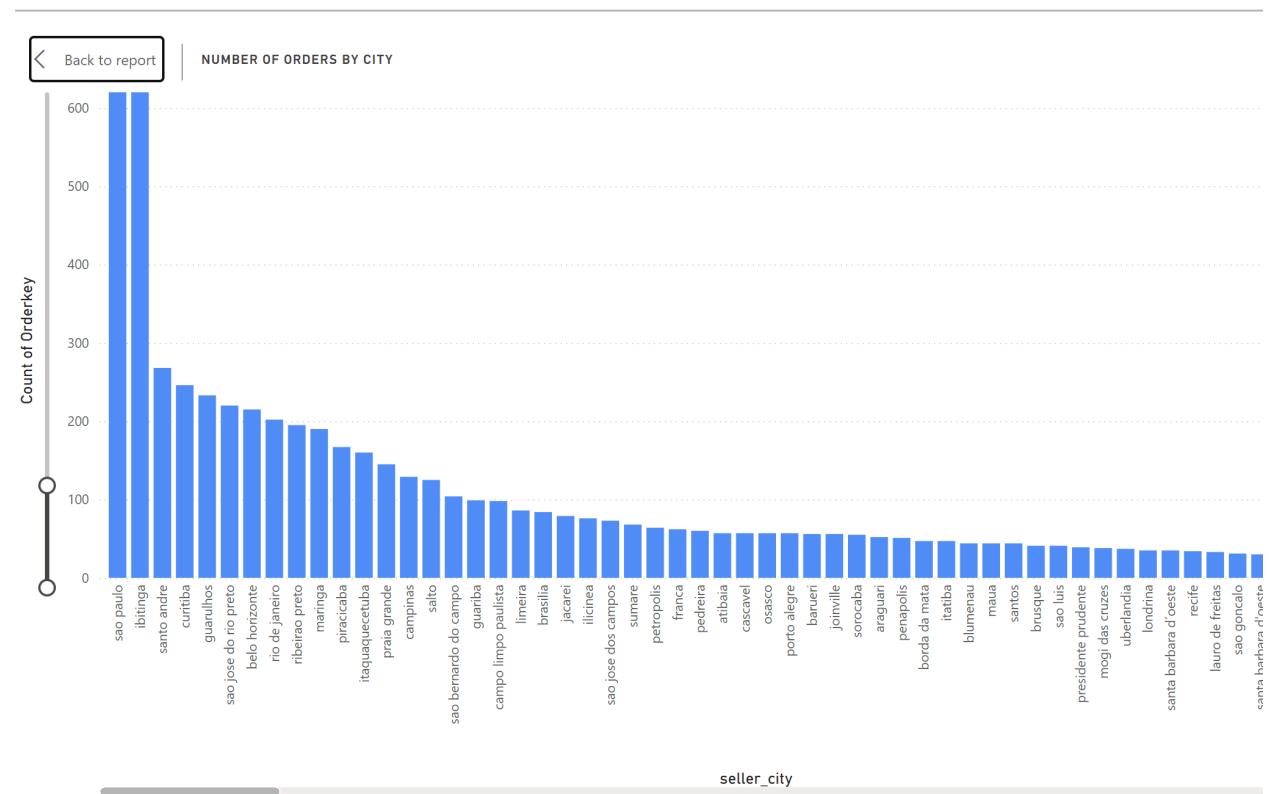
### 4. Geographical Sales Distribution



## Regional Distribution

### State-wise Performance

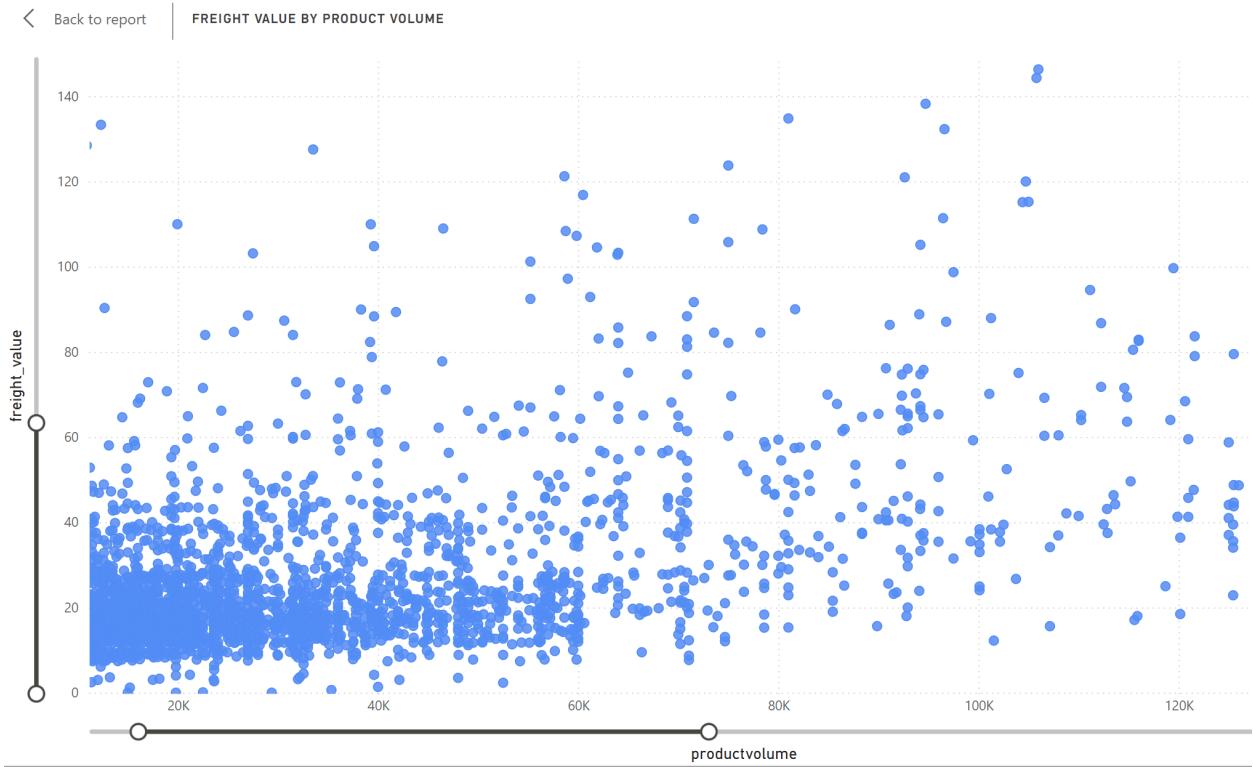
- SP (São Paulo) dominates the market
- PR (Paraná) and MG (Minas Gerais) show strong performance
- Smaller states like SE and PA have limited but focused category presence



### Order Distribution by City

- São Paulo and Ibirubá lead with the highest number of orders, both exceeding 600 orders.

- There's a significant drop after the top two cities, with the third-ranking city having less than half the orders of the leaders.
- The distribution shows a long tail, indicating a wide spread of orders across many cities with lower volumes.

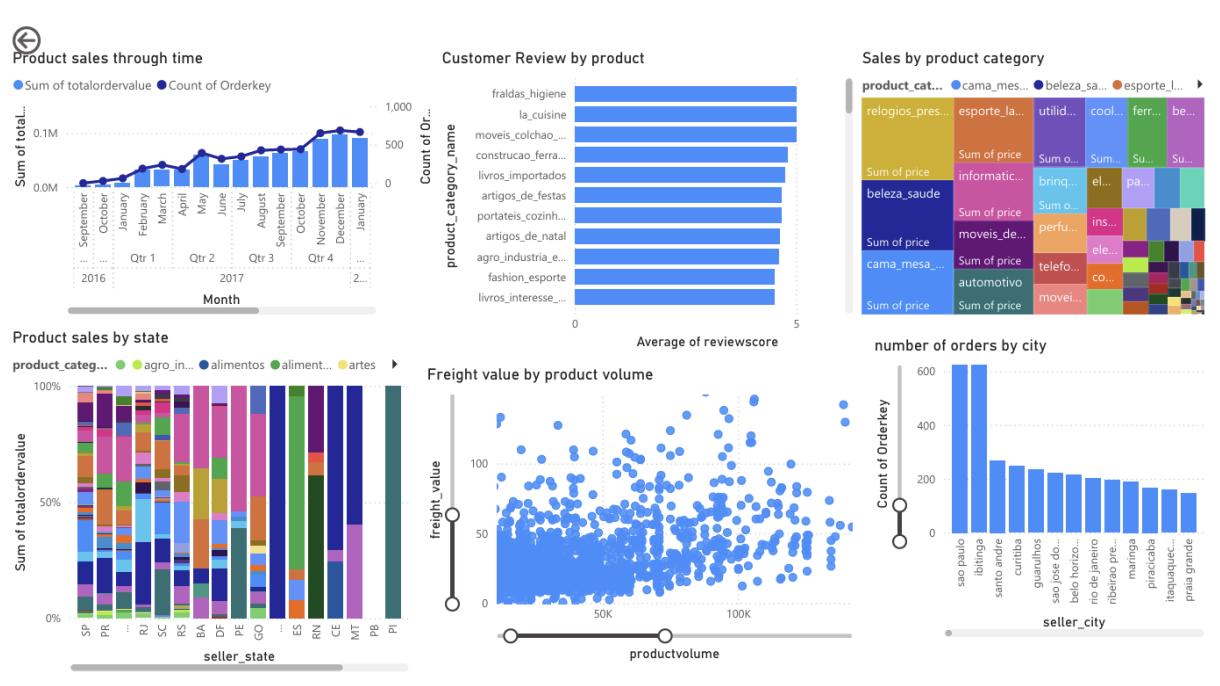


## Product Volume and Shipping Costs

### Freight Value Analysis

- Strong correlation between product volume and freight costs
- Majority of products cluster in 40K-80K volume range
- Freight values generally range from 20-60 units
- Several outliers with high freight costs (140+ units) for larger volumes
- Dense concentration of orders in lower volume range (40K-60K)

## Dashboard(Power BI)



## **Conclusion**

This data warehousing project leveraged AWS cloud services to analyze the Brazilian e-commerce platform's performance. The project successfully implemented a comprehensive ETL pipeline, efficiently utilizing multiple AWS services, including S3, Glue, Athena, and Step Functions.

A data lake was established using AWS S3, providing a secure and scalable storage solution for both raw and processed data. Automated ETL processes were implemented using AWS Glue, facilitating seamless data transformation. The use of Glue Crawlers enabled efficient data cataloging, while AWS Athena offered advanced SQL analytics capabilities. The entire workflow was orchestrated using AWS Step Functions, ensuring a streamlined and coordinated data processing pipeline.

The project's key achievements include the successful processing and transformation of multiple data sources, such as CSV and Parquet files. Dimensional and fact tables were created to support efficient querying, and automated data processing workflows were established to enhance operational efficiency. Through detailed SQL analysis, actionable business insights were generated, highlighting the platform's performance and opportunities for improvement.

From a business perspective, the data warehouse solution provided crucial insights. São Paulo emerged as the leading market with the highest sales concentration, while top-performing product categories such as Fraldas\_higiene, la\_cuisine, and chao\_e\_estofado were identified. Customer satisfaction patterns revealed a 3.43% rate of 5-star ratings, and regional sales distribution and shipping cost patterns were mapped. Additionally, the project identified expansion opportunities in emerging markets, paving the way for strategic growth.

## References

1. [Brazilian e-commerce Dataset](#)
2. [Amazon Web Services](#)
3. [Amazon S3](#)
4. [AWS glue](#)
5. [AWS crawler](#)
6. [AWS Athena](#)
7. [AWS Step Function](#)
8. [Power BI](#)