

Data Warehouse Project: Brazilian E-Commerce Pipeline with Olist Dataset



olist

Contents

01 Introduction

02 Dataset Exploration

03 Data Pipeline Architecture

04 Data Transformation with dbt

05 Workflow Orchestration with Airflow

06 Data Visualization with Power BI

07 Conclusion



01

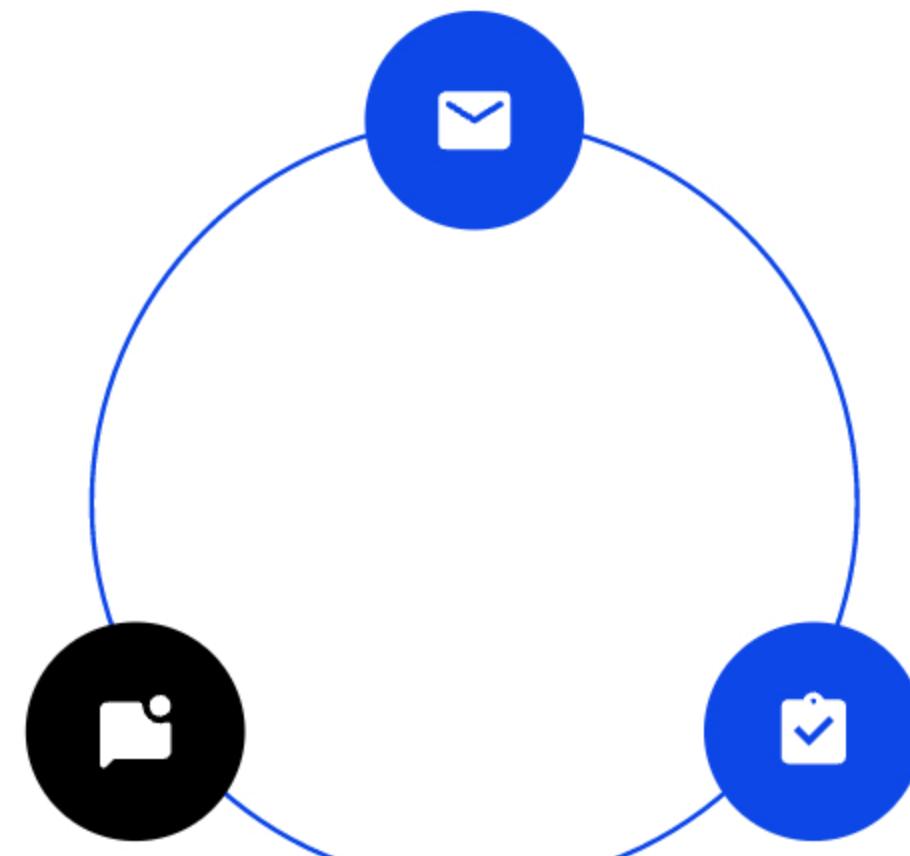
Introduction



Project Overview

Purpose and Goals

The project aims to build a robust data warehouse for Olist's Brazilian e-commerce data, enabling efficient data integration, analytics, and informed decision-making to enhance business performance.



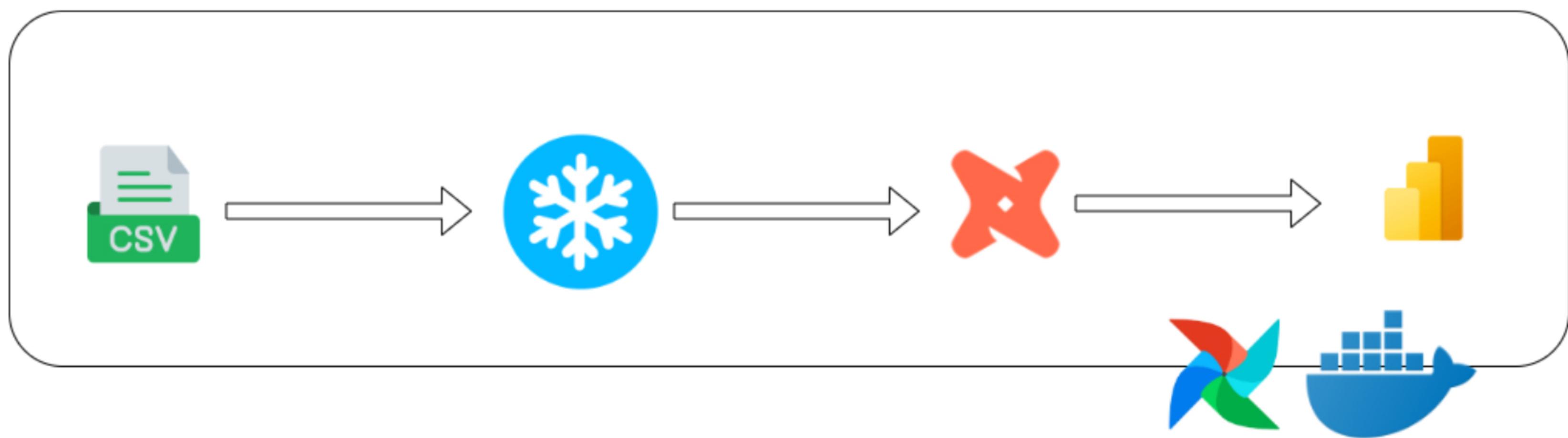
Dataset Introduction: Olist Brazilian E-Commerce

The Olist dataset captures diverse Brazilian e-commerce transactions, including orders, customers, products, and reviews, enabling comprehensive analysis of sales performance and customer behavior.

Problem Statement and Use Cases

Olist's fragmented e-commerce data hinders insights; this project consolidates sales, customer, and product data to enable improved demand forecasting, customer behavior analysis, and marketing strategies.

Technologies Used





02

Dataset Exploration



Overview of Olist Dataset

Data Sources and Tables

The Olist dataset comprises multiple tables, including orders, customers, products, sellers, and payments, sourced from Brazilian e-commerce transactions, enabling comprehensive analysis of sales and logistics.

Key Metrics and Dimensions

The Olist dataset includes over 100,000 orders with key metrics such as total sales, customer count, and order frequency. Dimensions cover product categories, geographical regions, and customer demographics.

Data Volume and Quality Assessment

The Olist dataset contains over 100,000 records with multiple tables. Data quality is generally high, though minor missing values and inconsistencies require cleaning before analysis.

Initial Data Profiling

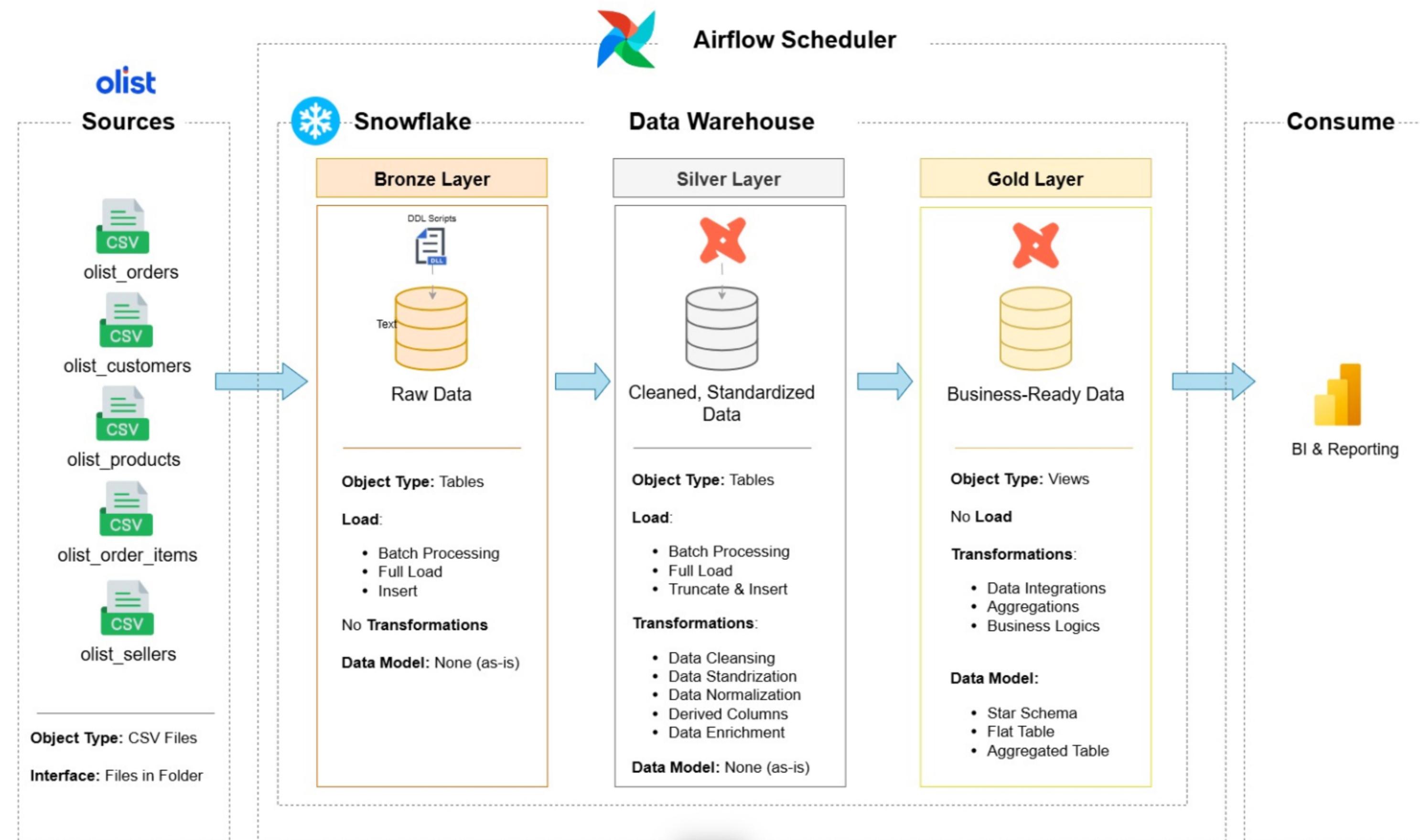
Exploring Customer Data

The customer dataset includes 100,000+ entries, highlighting diverse demographics, purchase behaviors, and regional distribution. Key metrics: average order frequency, customer retention, and geographic concentration.

Product and Category Insights

The dataset contains 7,000+ unique products across 73 categories. Top categories include electronics and home appliances, showing diverse inventory and sales potential within the Brazilian e-commerce market.

Pipeline Design Overview



Snowflake Implementation



Data ingestion leverages Snowflake's Snowpipe for continuous, automated loading from Olist's raw e-commerce data, ensuring real-time availability and scalability within the data warehouse environment.



Leveraged Snowflake's micro-partitioning and automatic clustering to optimize storage. Utilized result caching and query pruning to enhance performance, reducing query latency and storage costs significantly.



Role-based access ensures data protection by restricting permissions to authorized users; Snowflake's multi-factor authentication and encryption enhance security across the Brazilian e-commerce pipeline.

Snowflake Implementation

Step1

DataBase & Schemas Creation

The screenshot shows the Snowflake UI interface for creating a new database and its associated schemas. The database 'OLIST_DWH' is selected, and the 'Schemas' tab is active. The table lists six schemas: BRONZE, GOLD, GOLD_DBT_TEST_..., INFORMATION_S..., PUBLIC, and SILVER, all created by ACCOUNTADMIN just 19 hours ago.

NAME ↑	OWNER	CREATED
BRONZE	ACCOUNTADM...	19 hours ago
GOLD	ACCOUNTADM...	14 hours ago
GOLD_DBT_TEST_...	ACCOUNTADM...	1 hour ago
INFORMATION_S...	—	—
PUBLIC	ACCOUNTADM...	19 hours ago
SILVER	ACCOUNTADM...	19 hours ago

Snowflake Implementation

Step2

Extract Datasets in Bronze Layer

The screenshot shows the Snowflake UI interface for the schema **OLIST_DWH / BRONZE**. The schema was created by **ACCOUNTADMIN** 19 hours ago. The **Tables** tab is selected, displaying 6 tables:

NAME ↑	TYPE	CLASSIFICATION	OWNER	ROWS	BYTES	CREATED	...
MY_FIRST_DBT_MODEL	Table	—	ACCOUNTAD...	2	1.0KB	13 hours ago	...
OLIST_CUSTOMERS	Table	—	ACCOUNTAD...	99.4K	6.6MB	19 hours ago	...
OLIST_ORDERS	Table	—	ACCOUNTAD...	99.4K	5.4MB	19 hours ago	...
OLIST_ORDER_ITEMS	Table	—	ACCOUNTAD...	112.7K	8.4MB	18 hours ago	...
OLIST_PRODUCTS	Table	—	ACCOUNTAD...	33.0K	1.4MB	18 hours ago	...
OLIST_SELLERS	Table	—	ACCOUNTAD...	3.1K	120.5KB	18 hours ago	...

Snowflake Implementation

Step3

Data Transformation in Silver Layer

The screenshot shows the Snowflake Schema browser interface. The top navigation bar displays the schema name **OLIST_DWH / SILVER**. Below the navigation bar, there are tabs for **Schema Details** and **Tables**, with **Tables** being the active tab. A message indicates the schema was created by **ACCOUNTADMIN** 19 hours ago. On the left side, a sidebar contains various icons for database management tasks. The main content area displays a table titled **5 Tables**, listing the following information:

NAME ↑	TYPE	CLASSIFICATION	OWNER	ROWS	BYTES	CREATED	...
CUSTOMERS	Table	—	ACCOUNTADM...	99.4K	6.6MB	17 hours ago	...
ORDERS	Table	—	ACCOUNTADM...	99.4K	8.7MB	17 hours ago	...
ORDER_ITEMS	Table	—	ACCOUNTADM...	112.7K	8.8MB	17 hours ago	...
PRODUCTS	Table	—	ACCOUNTADM...	33.0K	1.4MB	17 hours ago	...
SELLERS	Table	—	ACCOUNTADM...	3.1K	120.5KB	16 hours ago	...

At the bottom right of the table, there are buttons for **Search**, **All Tables**, and a refresh icon.

Snowflake Implementation

Step4

Load Data in Gold Layer

The screenshot shows the Snowflake UI interface for the schema **OLIST_DWH / GOLD**. The top navigation bar includes a back arrow, a refresh icon, and a search bar. On the right, there are buttons for **...**, **Create**, and a dropdown menu. The left sidebar has icons for Schema, Tables, Views, and other database objects. The main content area displays the **Views** tab, which lists five views:

NAME ↑	TYPE	OWNER	CREATED
AGG_SALES_BY_DAY	View	ACCOUNTADMIN	23 minutes ago
DIM_CUSTOMERS	View	ACCOUNTADMIN	12 hours ago
DIM_PRODUCTS	View	ACCOUNTADMIN	12 hours ago
DIM_SELLERS	View	ACCOUNTADMIN	12 hours ago
FACT_ORDERS	View	ACCOUNTADMIN	12 hours ago

Below the table are buttons for **Search**, **All Views**, and a trash bin icon.



04 Data Transformation with dbt



Quality and Testing

Step1

dbt Tests Implementation

Step2

Documentation and Lineage

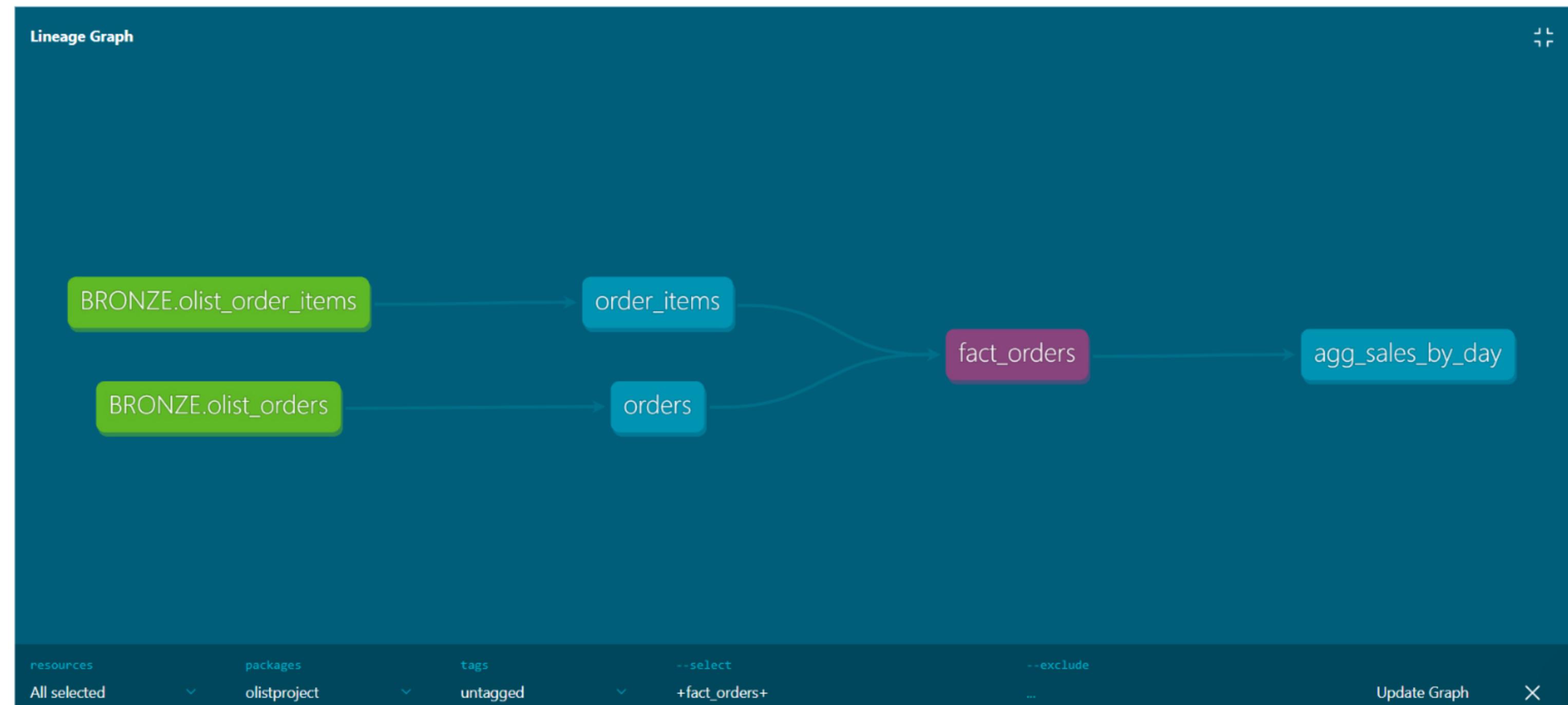
Step3

Error Handling and Debugging



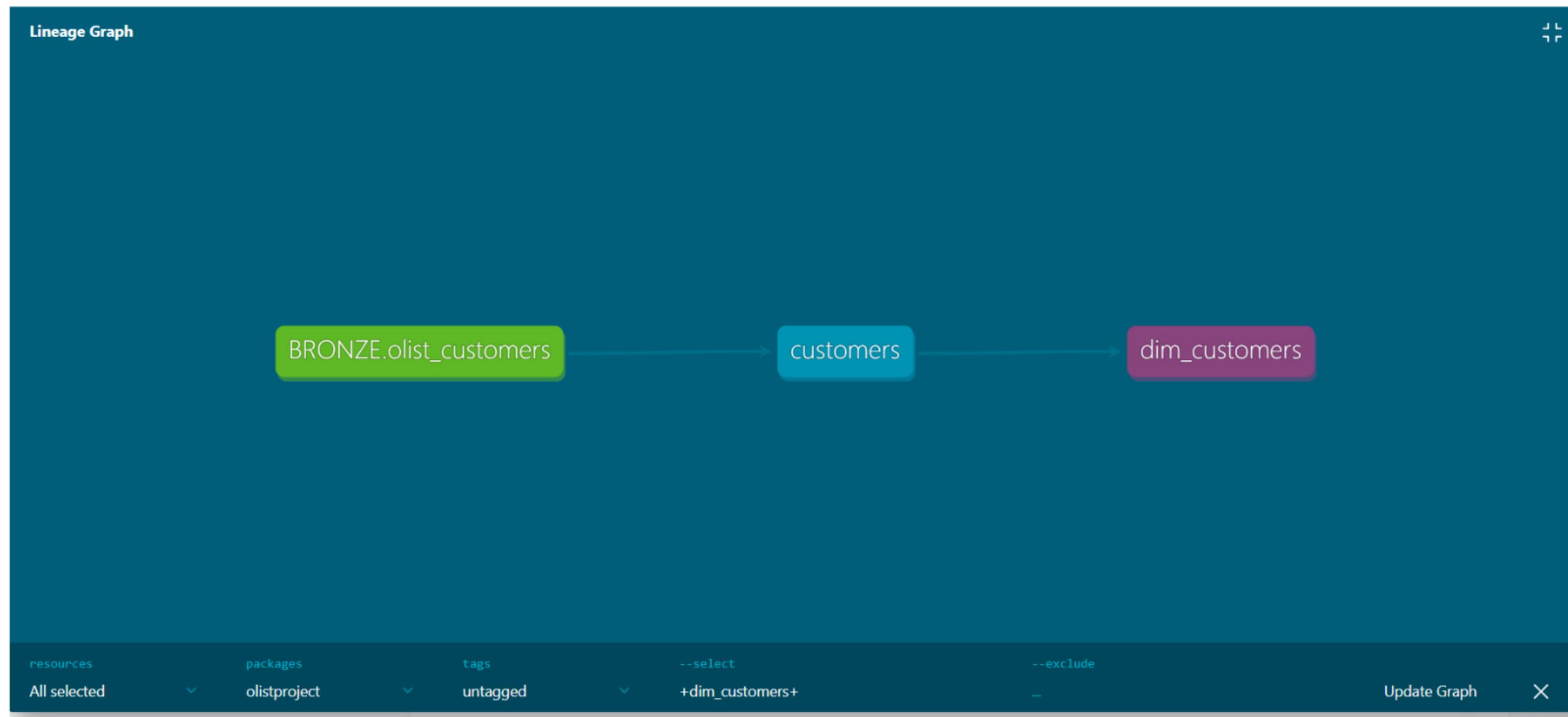
Modeling Approach

fact_orders dbt docs



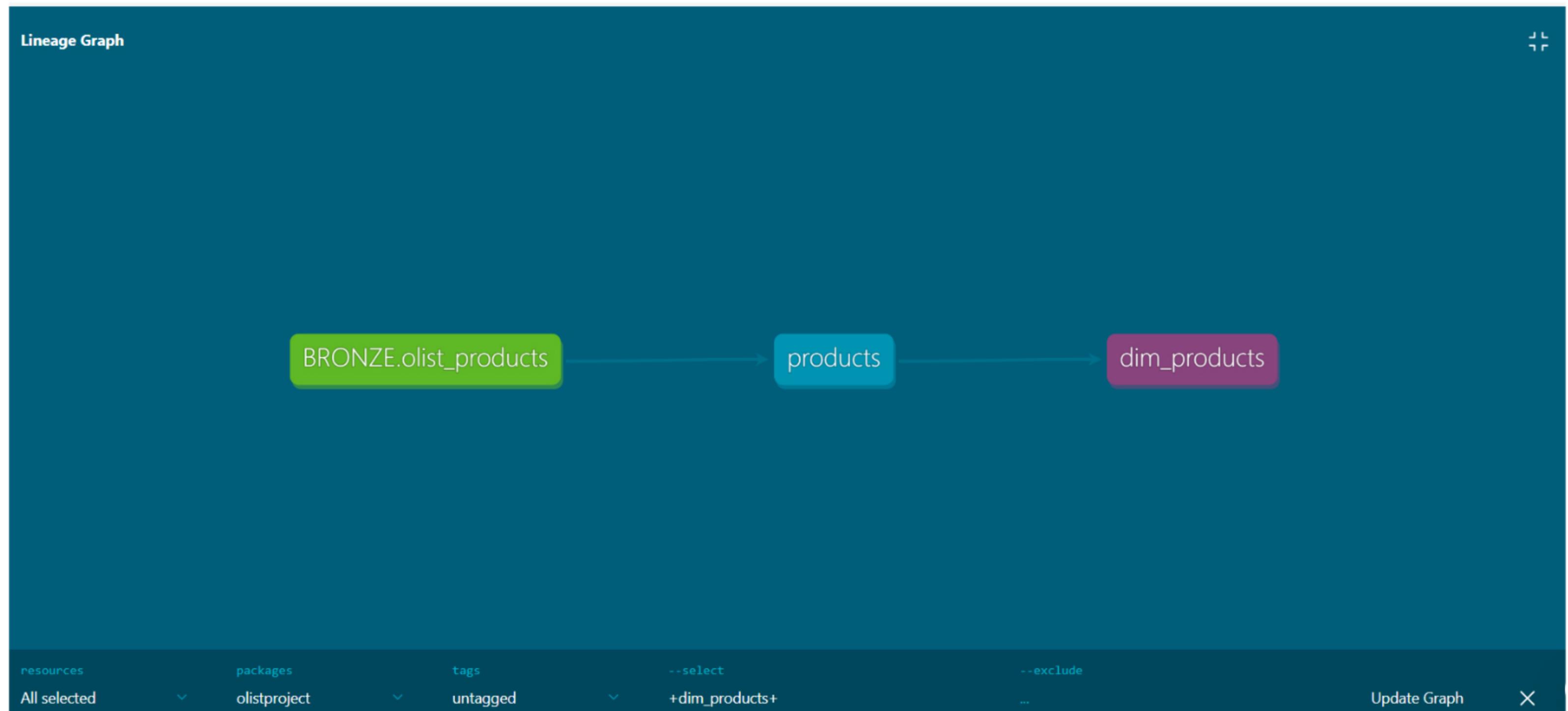
Modeling Approach

dim_customers dbt docs



Modeling Approach

dim_products dbt docs



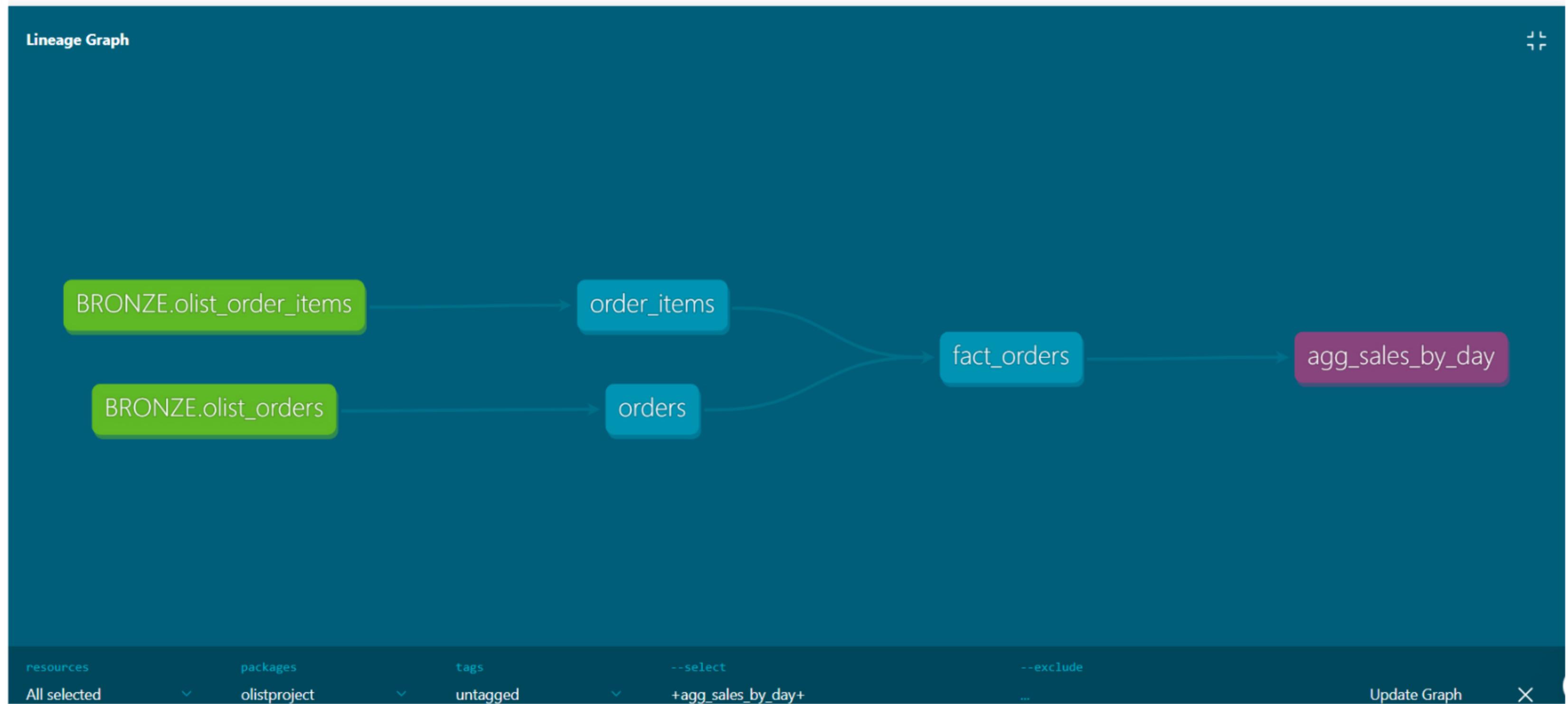
Modeling Approach

dim_sellers dbt docs



Modeling Approach

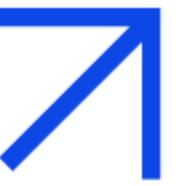
aggregation_sales_by_day dbt docs





05

Workflow Orchestration
with Airflow



DAG Design

01

Task Dependencies and Scheduling

Task dependencies in Airflow DAG ensure sequential execution, while scheduling automates timely pipeline runs, optimizing data flow and maintaining ETL consistency in the Brazilian E-Commerce project.

02

Handling Failures and Retries

In the DAG design, failure handling includes automatic task retries with configurable limits and exponential backoff, ensuring pipeline resilience and minimizing manual intervention for smoother data processing.

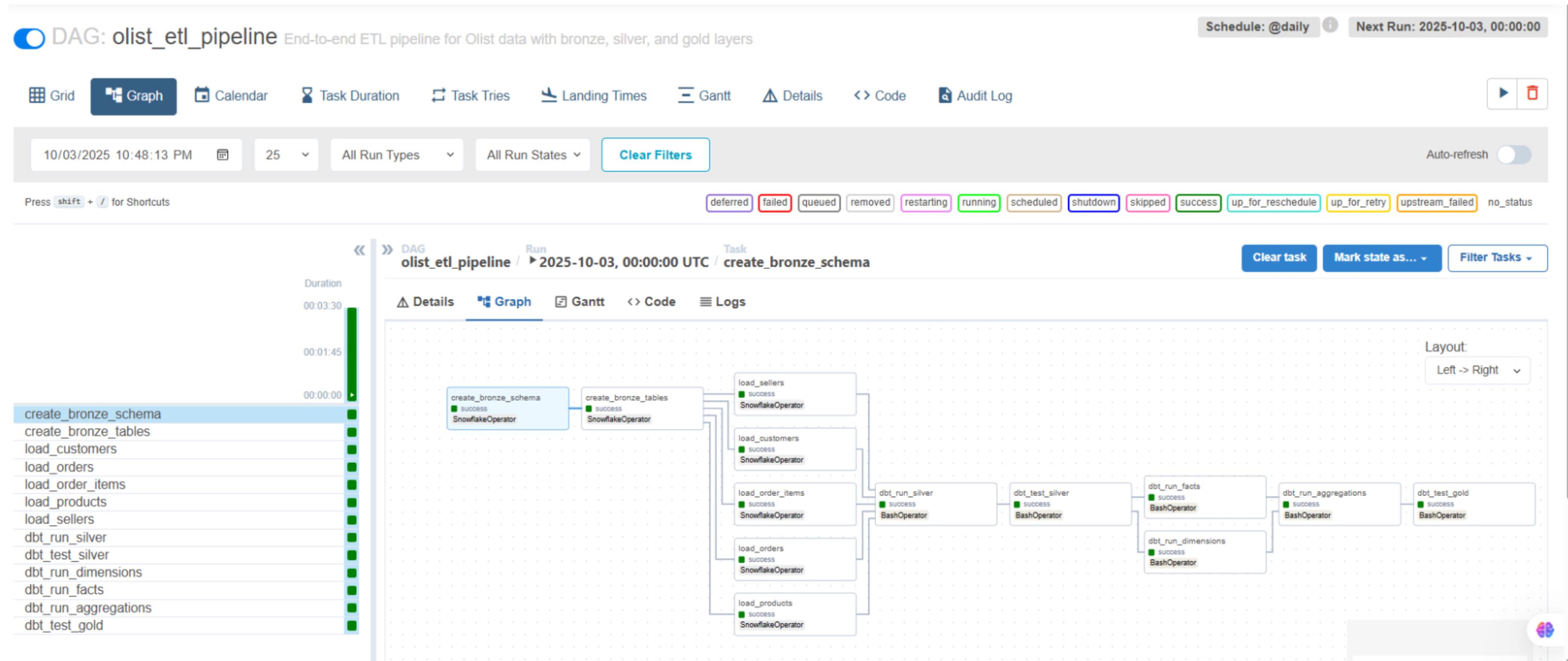


03

Monitoring and Alerts Setup

Implement comprehensive monitoring using Airflow's native tools and integrate alerts via email and Slack to promptly address DAG failures, ensuring pipeline reliability and timely issue resolution.

Integration with Snowflake and dbt





06 Data visualization with Power BI



Navigators

Executive Summary

Customer Deep Dive

Products & Sellers Performance

Operations & Logistics Efficiency

96K

Unique Customers

137.75

AOV

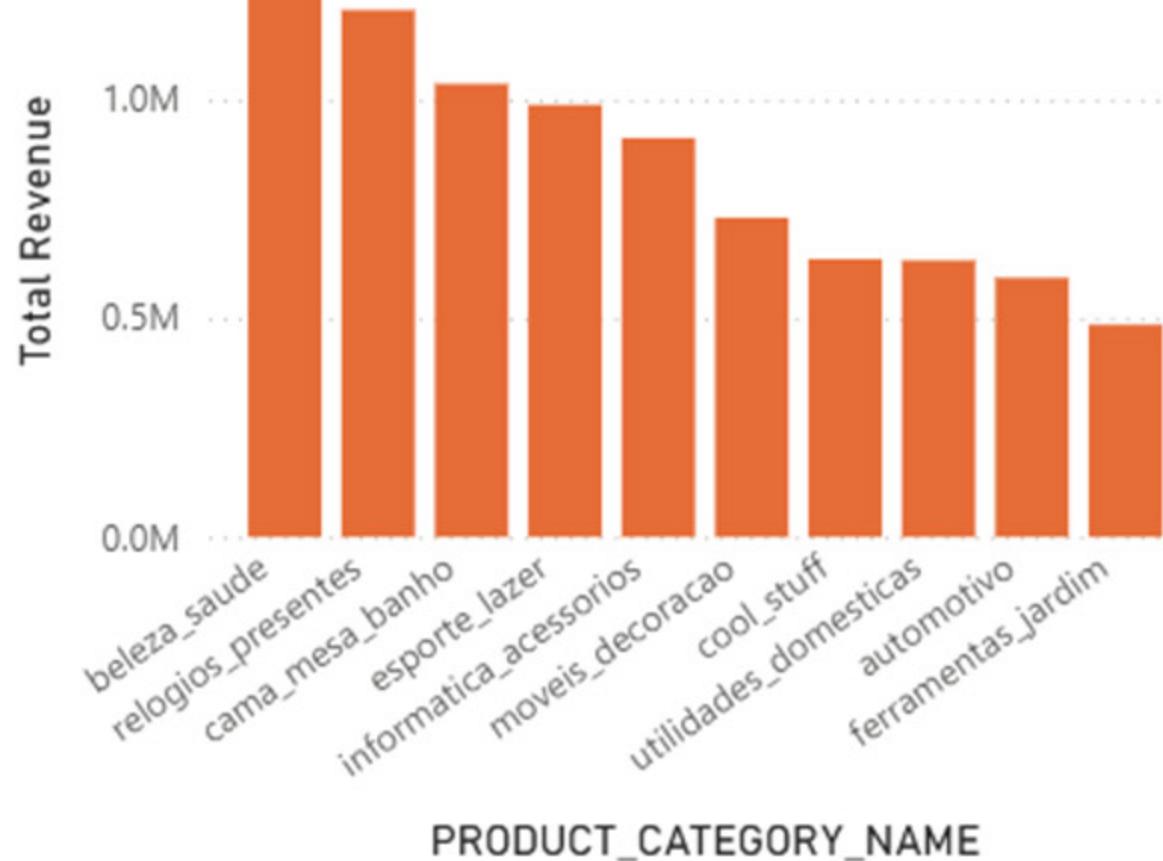
99K

Total Orders

13.59M

Total Revenue

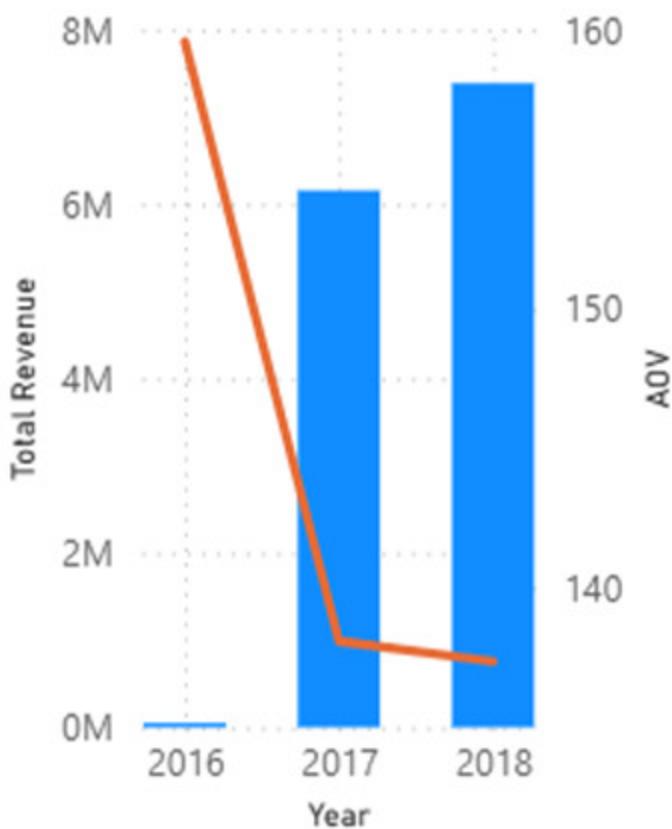
Top 10 Product Categories



Total Revenue by CUSTOMER_STATE



Monthly revenue performance and average order value



Navigators

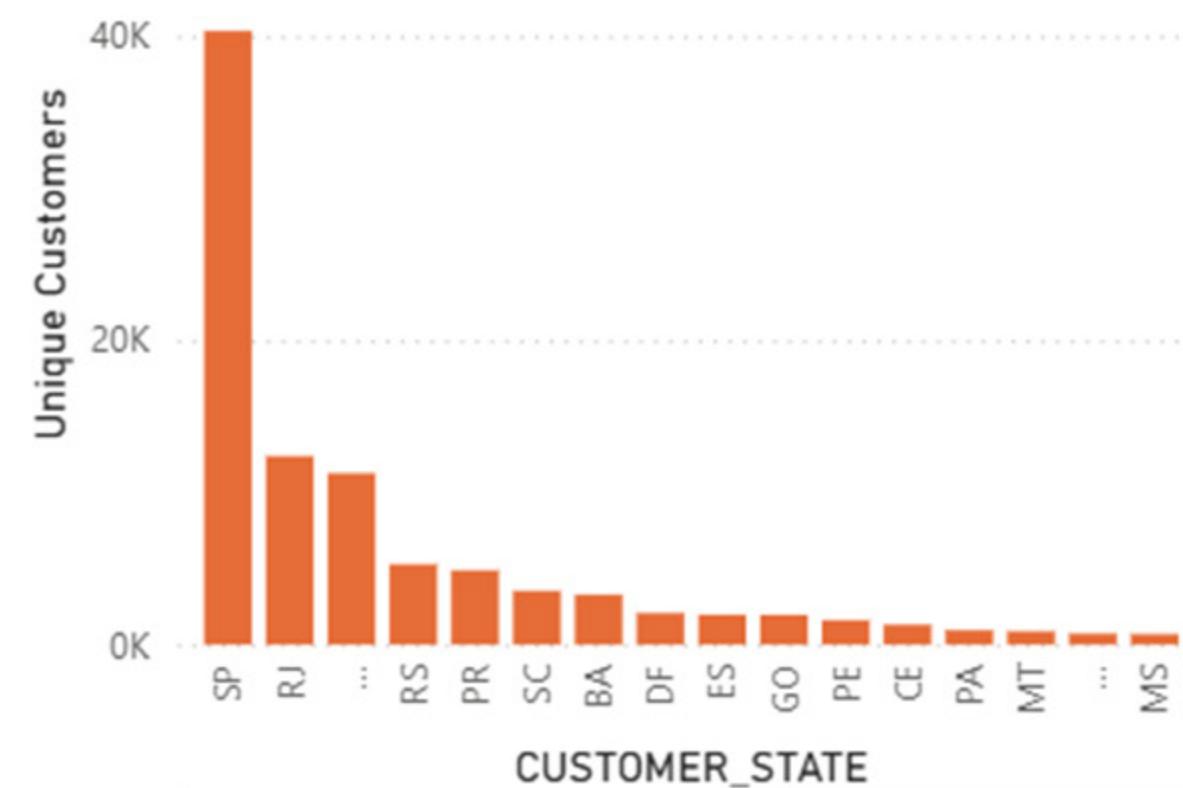
Executive Summary

Customer Deep Dive

Products & Sellers Performance

Operations & Logistics Efficiency

Top 10 Clients by States



First Order Date

141.44

Average Revenue per Customer

96K

Unique Customers

Unique Customers States

▲ 40302 RJ ▲ 12384 MG ▲ 11259 RS ▲

CUSTOMER_SK	Total Revenue	First Order Date
0	13,591,643.70	9/4/2016 9:15:19 PM
1	13,591,643.70	9/4/2016 9:15:19 PM
2	13,591,643.70	9/4/2016 9:15:19 PM
3	13,591,643.70	9/4/2016 9:15:19 PM
4	13,591,643.70	9/4/2016 9:15:19 PM
5	13,591,643.70	9/4/2016 9:15:19 PM
6	13,591,643.70	9/4/2016 9:15:19 PM
7	13,591,643.70	9/4/2016 9:15:19 PM
8	13,591,643.70	9/4/2016 9:15:19 PM
Total	13,591,643.70	9/4/2016 9:15:19 PM



Navigators

Executive Summary

Customer Deep Dive

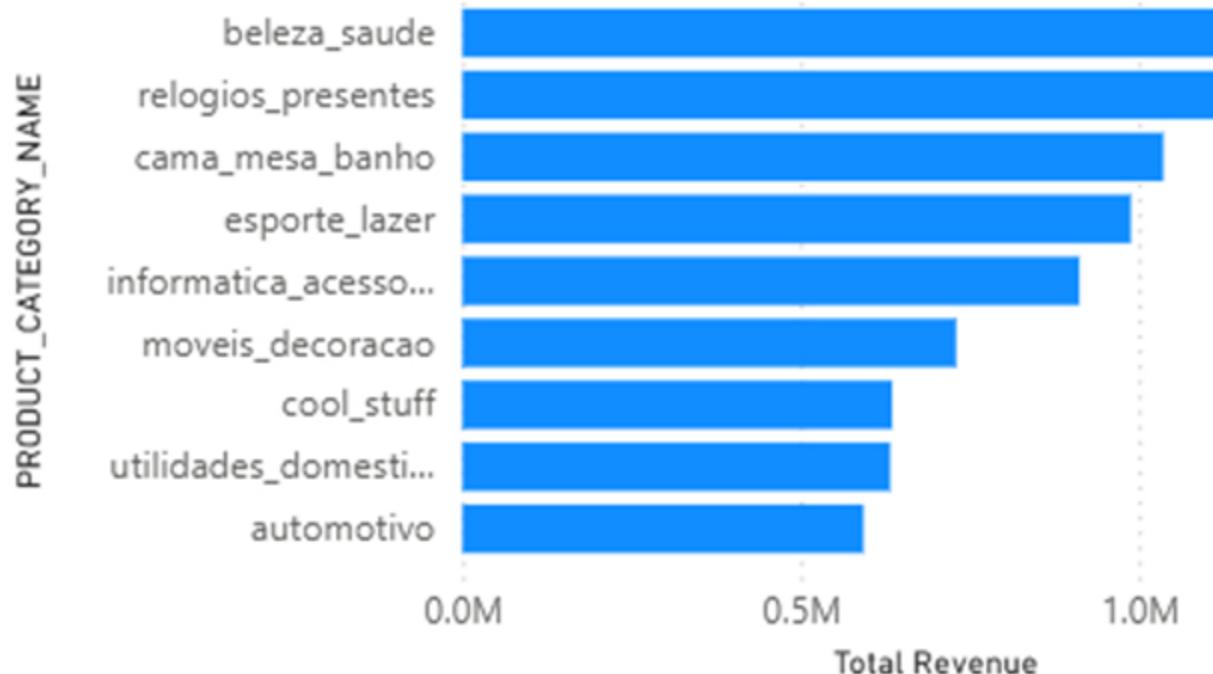
Products & Sellers Performance

Operations & Logistics Efficiency

Total Revenue by CUSTOMER_STATE



Analysis of the monthly changes in the ranking of top-selling product categories

3095
Total Sellers113K
Total Products Sold120.65
Average Price

Top 10 sellers by revenue



Navigators

Executive Summary

Customer Deep Dive

Products & Sellers Performance

Operations & Logistics Efficiency

0.02

On-Time Delivery %

1477

On-Time Orders

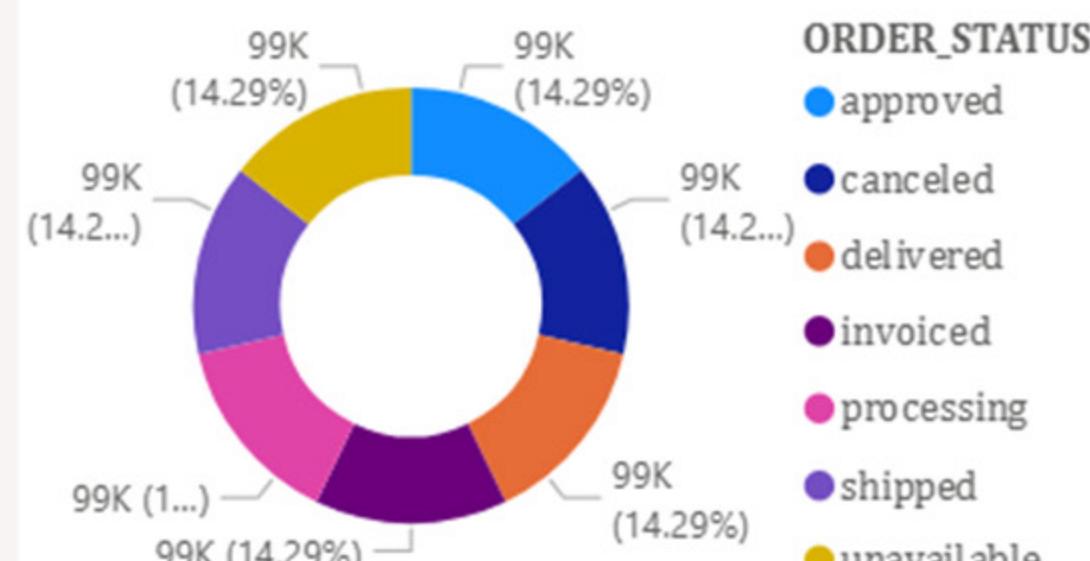
96K

Total Delivered Orders

0.52

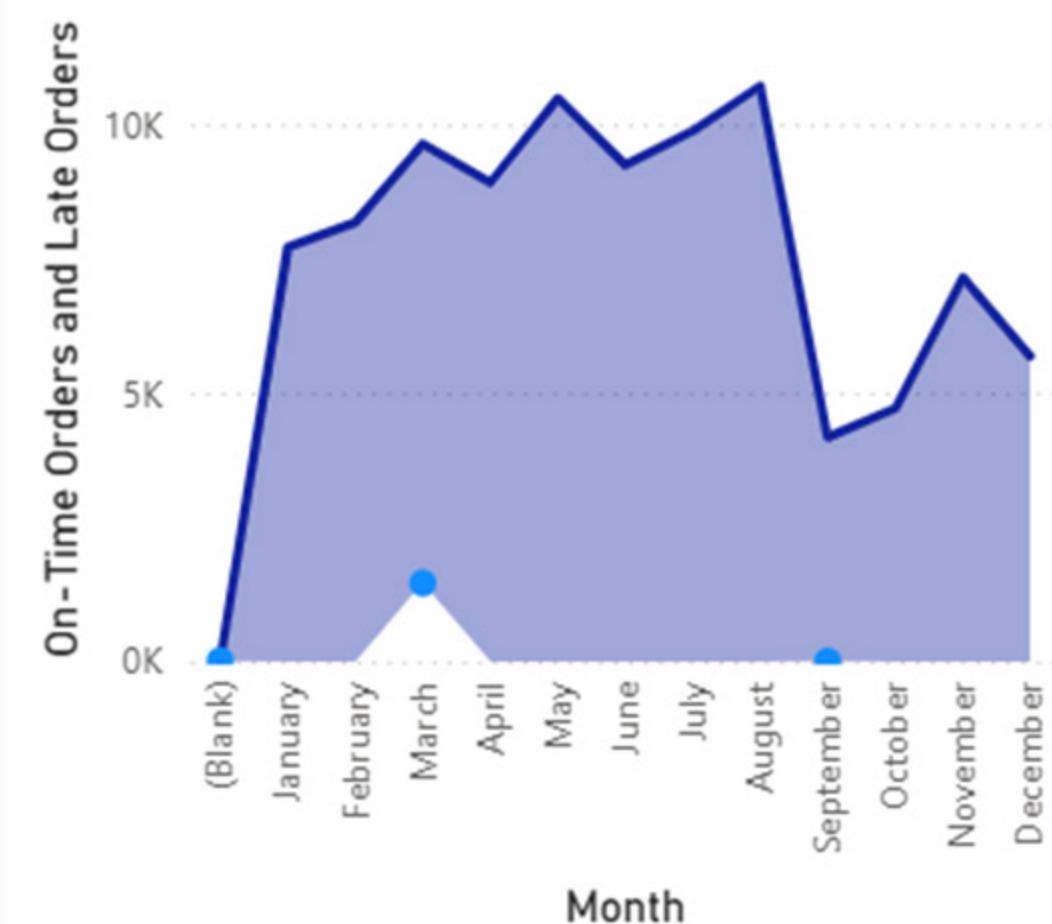
Average Delivery Time

Order Status Breakdown

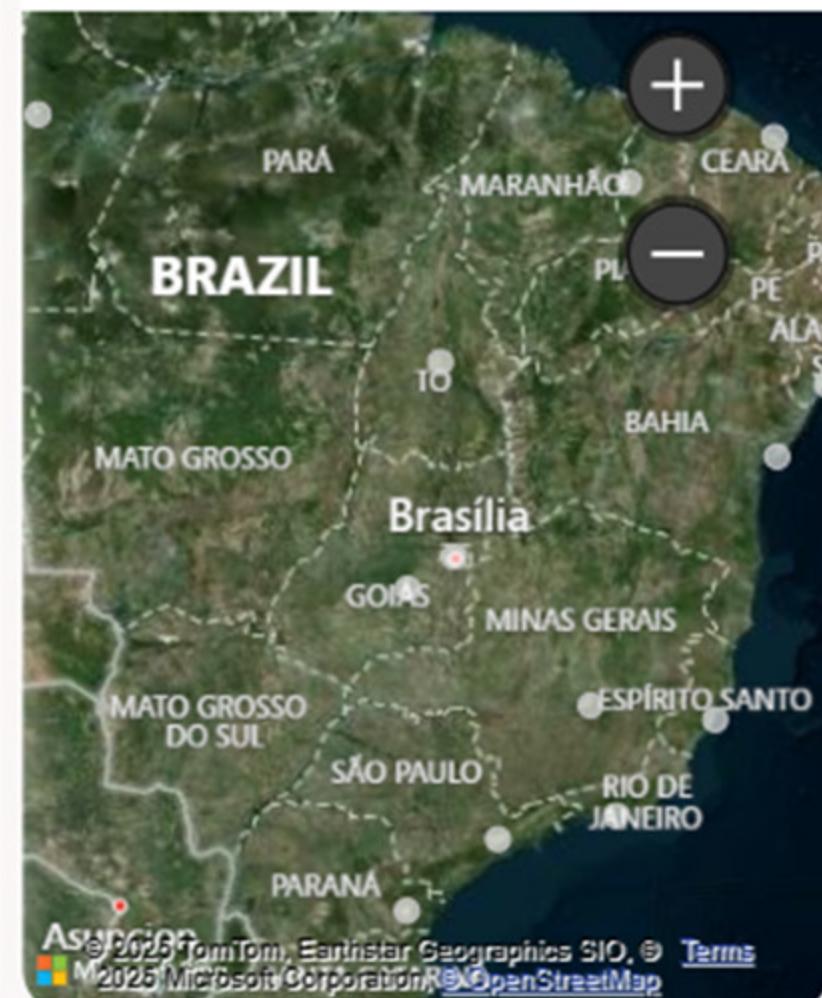


On-Time Orders and Late Orders by Month

On-Time Orders Late Orders



Total Revenue by CUSTOMER_STATE





Summary of Findings
and Next Steps

07 Conclusion



Project Summary

Achievements and Outcomes

Successfully integrated and optimized the Olist dataset for streamlined e-commerce analysis, enhancing data accessibility and supporting informed decision-making. Future work will focus on real-time data updates and advanced analytics integration.

Lessons Learned

The project highlighted the importance of robust data cleaning, efficient pipeline automation, and scalable storage. Future work includes enhancing real-time analytics and incorporating additional data sources for deeper insights.

Any Question ?



Thanks