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

olist

Contents

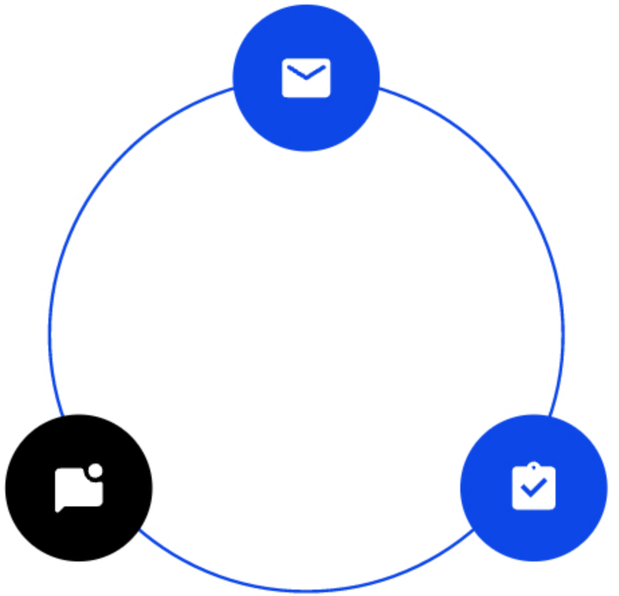
olist

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.



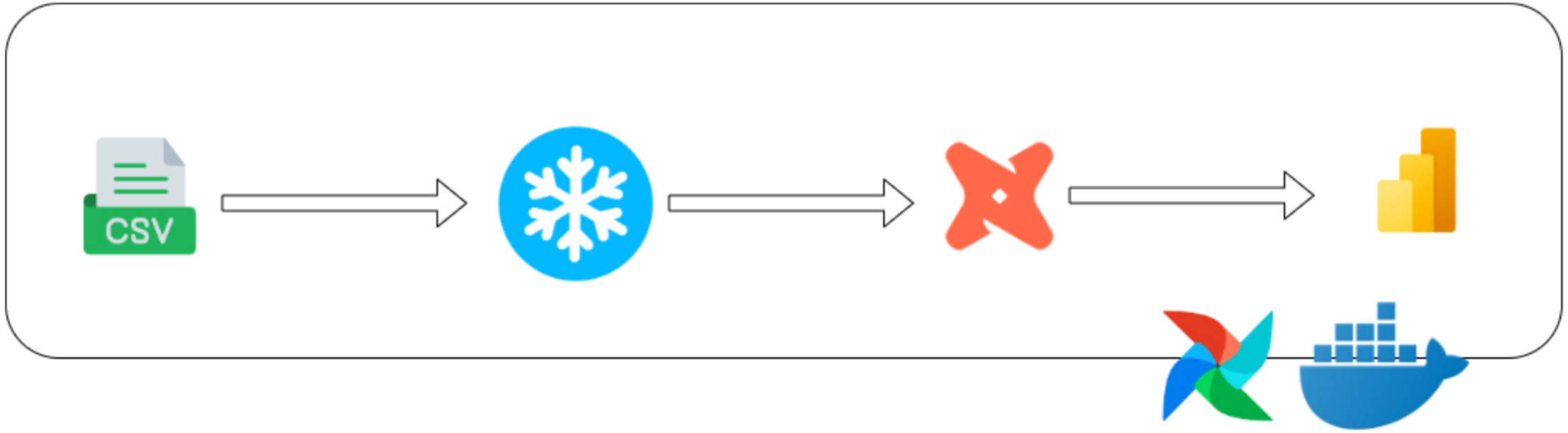
Dataset Introduction: Olist Brazilian E-Commerce

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



olist

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

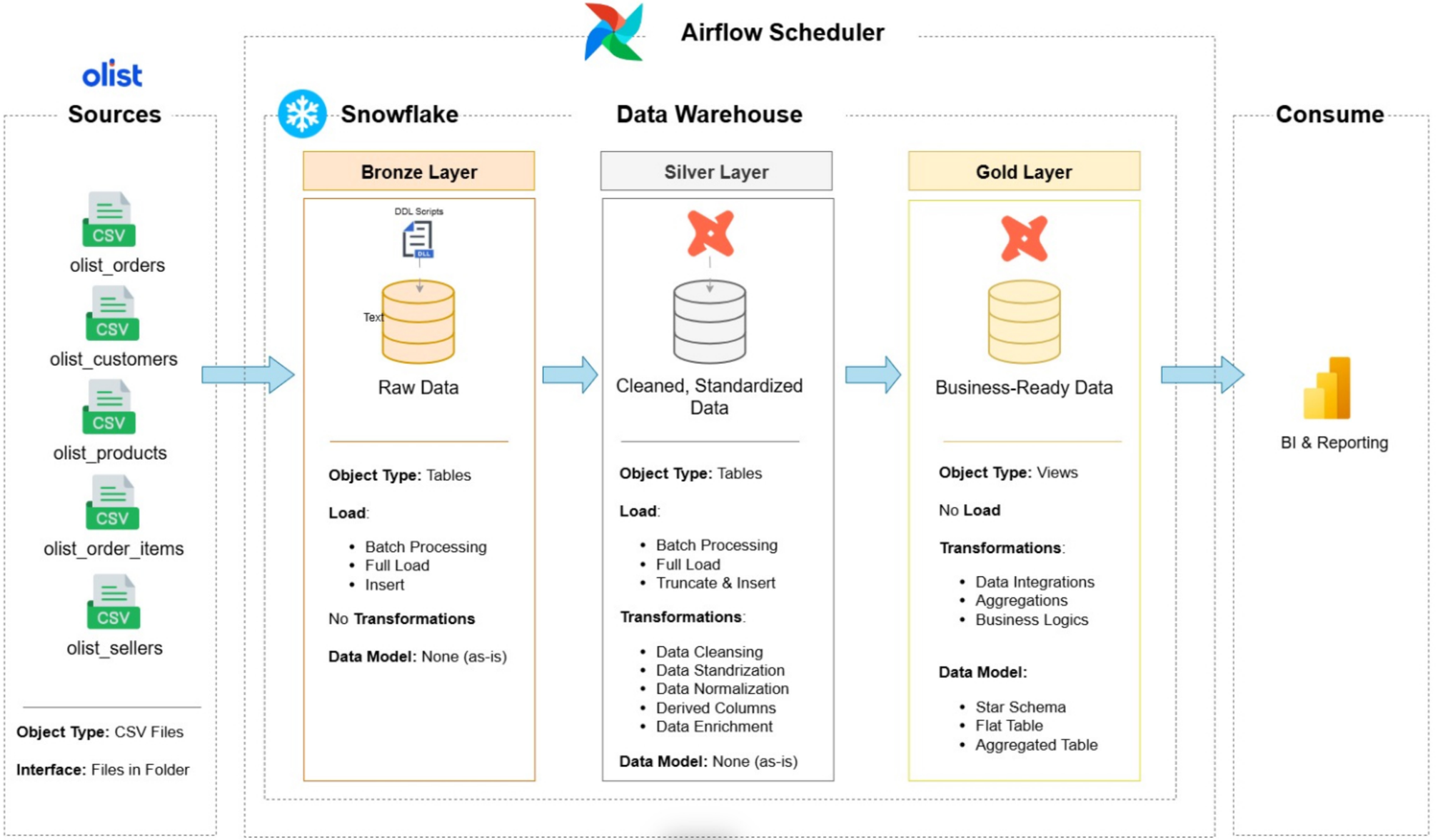
Exploring Customer Data

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 Setup

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.

Storage and Performance Optimization

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.

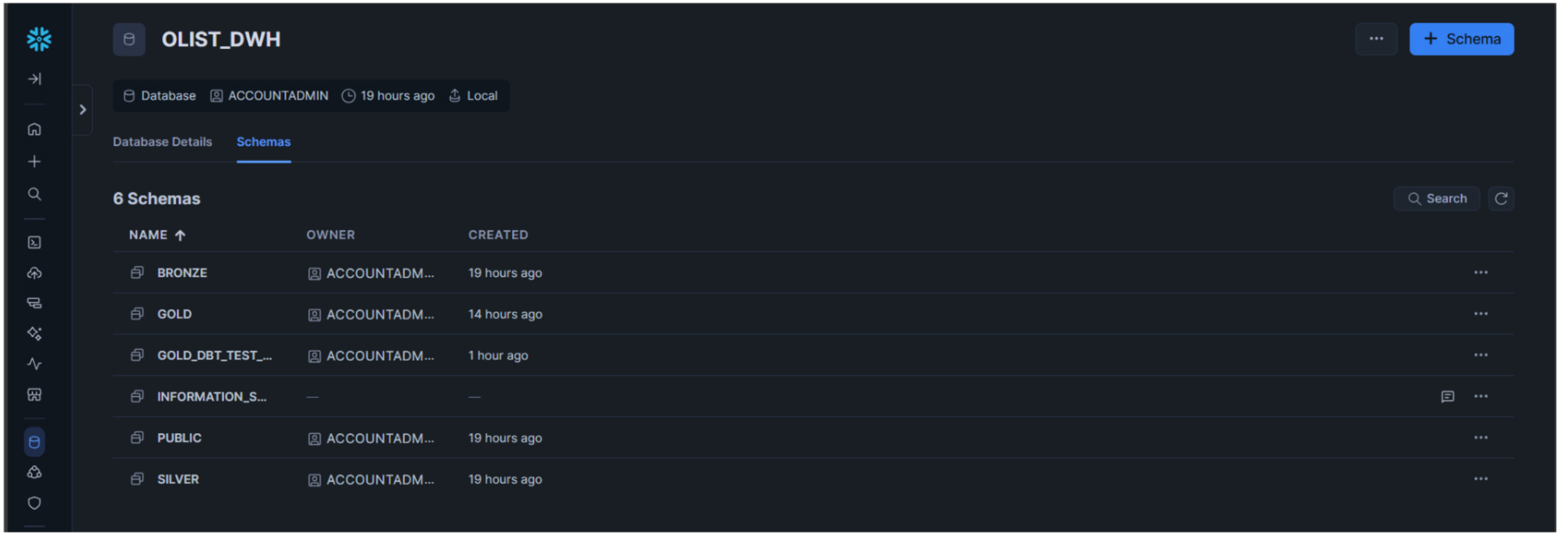
Security and Access Controls

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

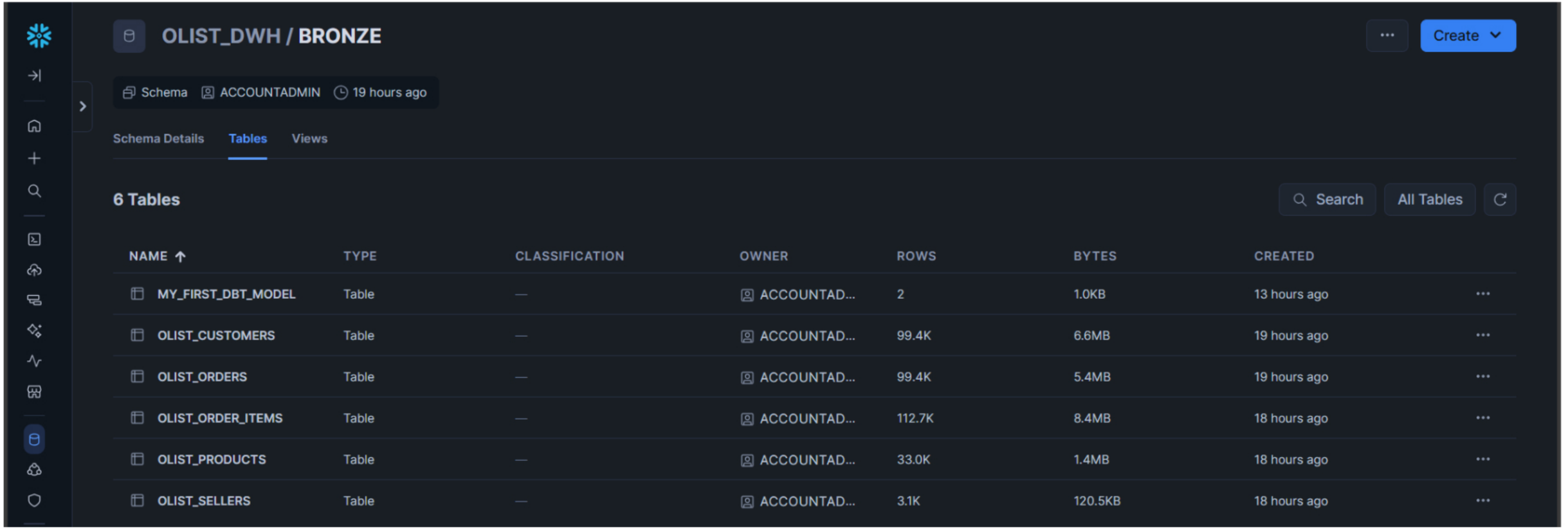
DataBase & Schemas Creation



Snowflake Implementation

Step2

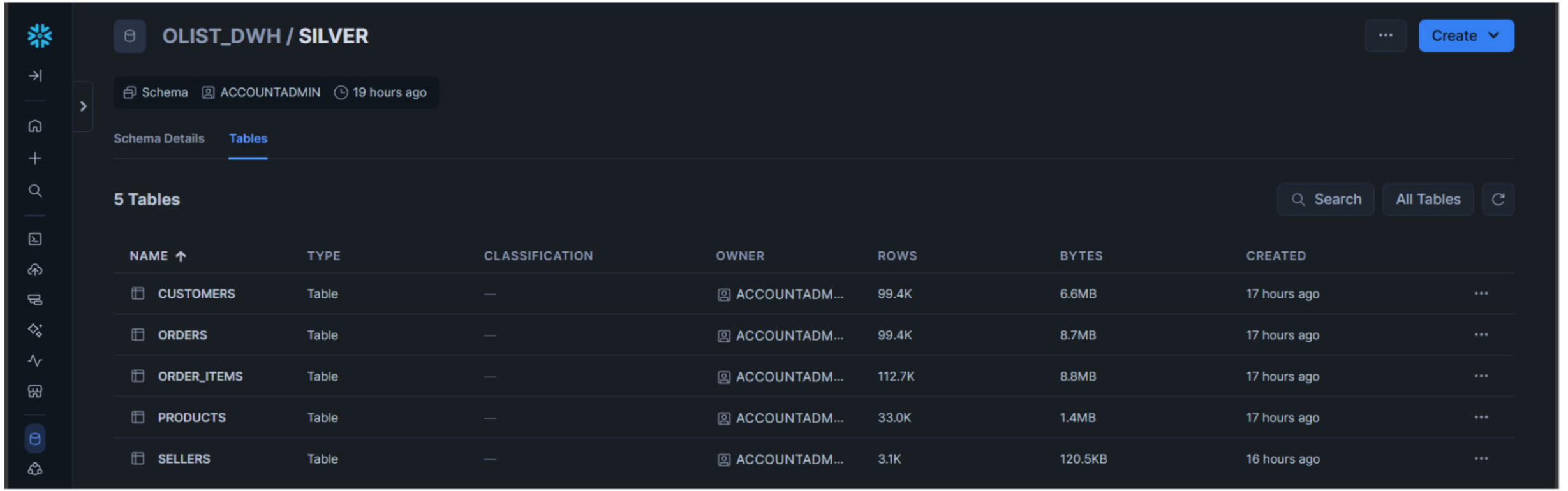
Exteract Datasets in Bronze Layer



Snowflake Implementation

Step3

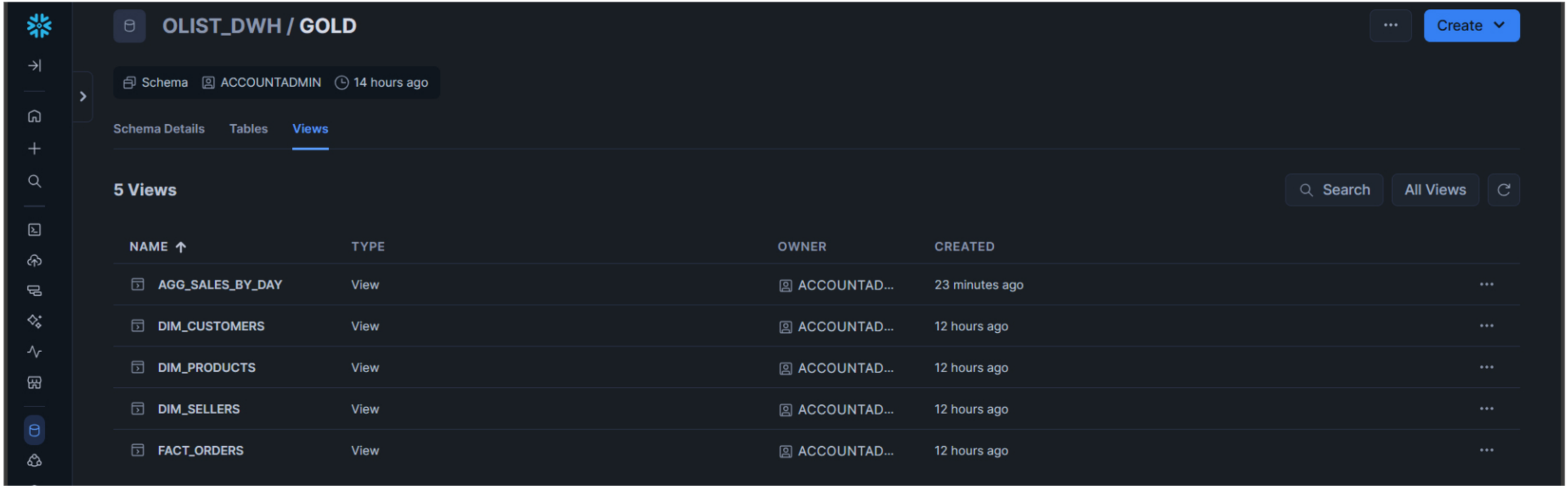
Data Transformation in Silver Layer

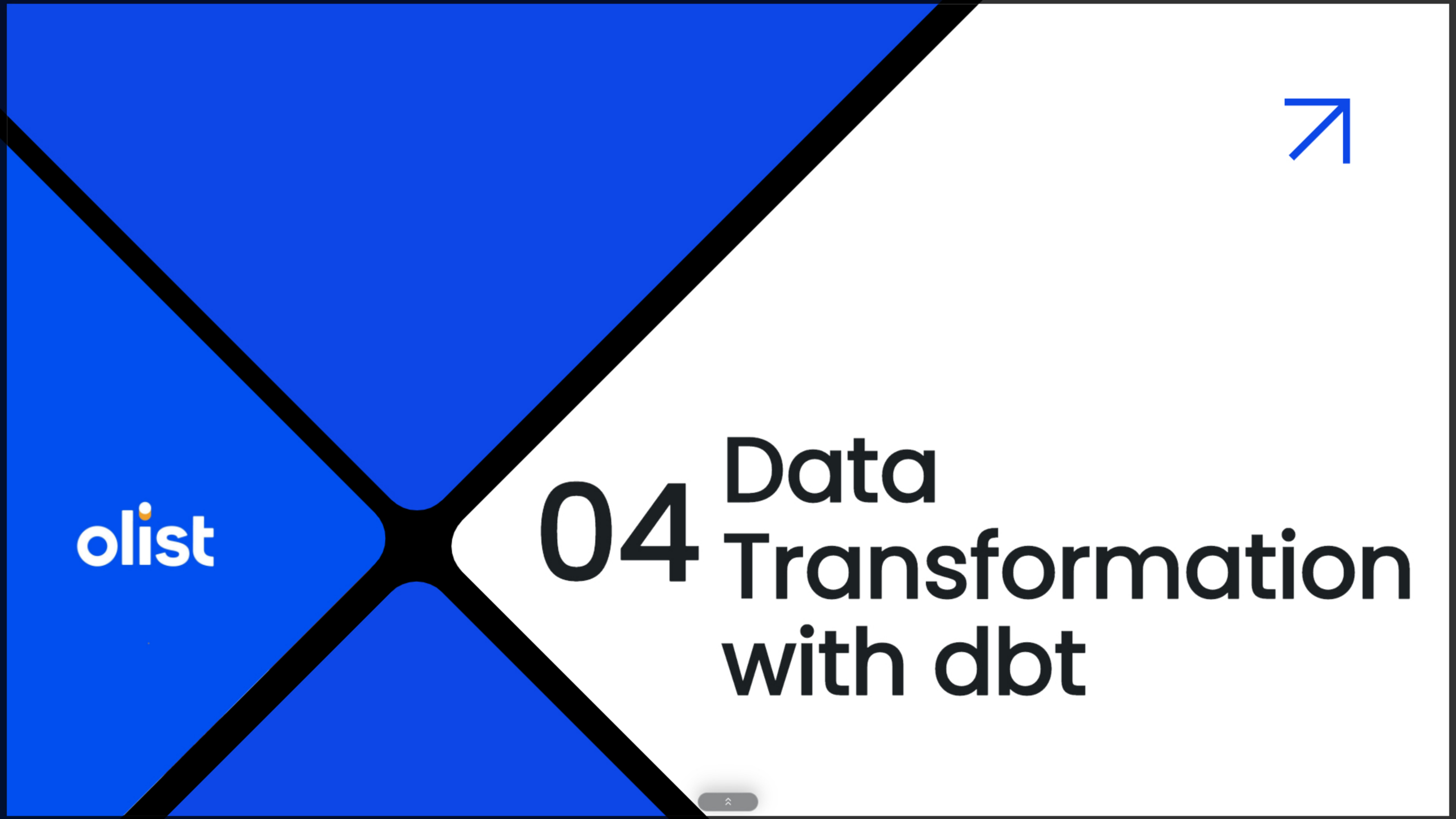


Snowflake Implementation

Step4

Load Data in Gold Layer




olist
04 Data Transformation with dbt

Quality and Testing

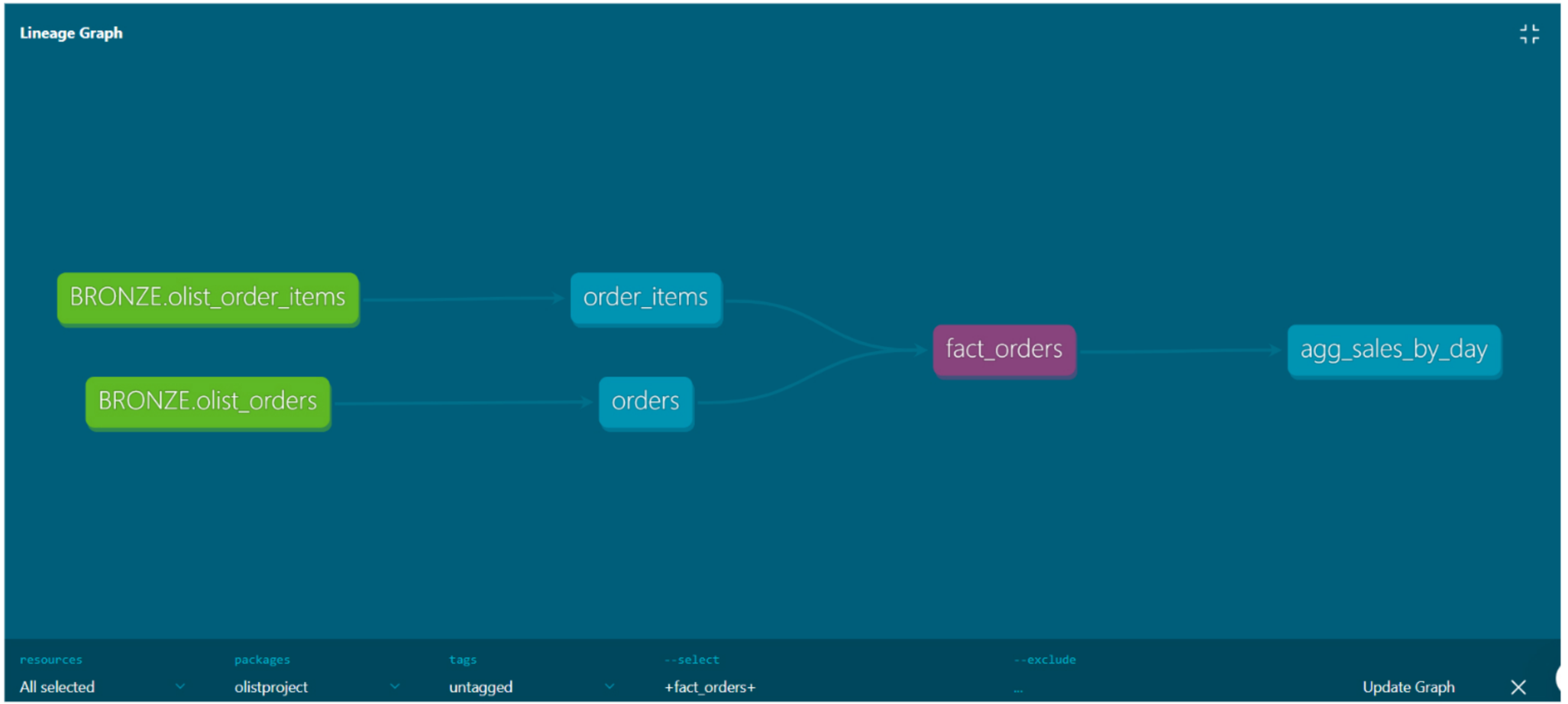
Step1
dbt Tests Implementation

Step2
Documentation and Lineage

Step3
Error Handling and Debugging

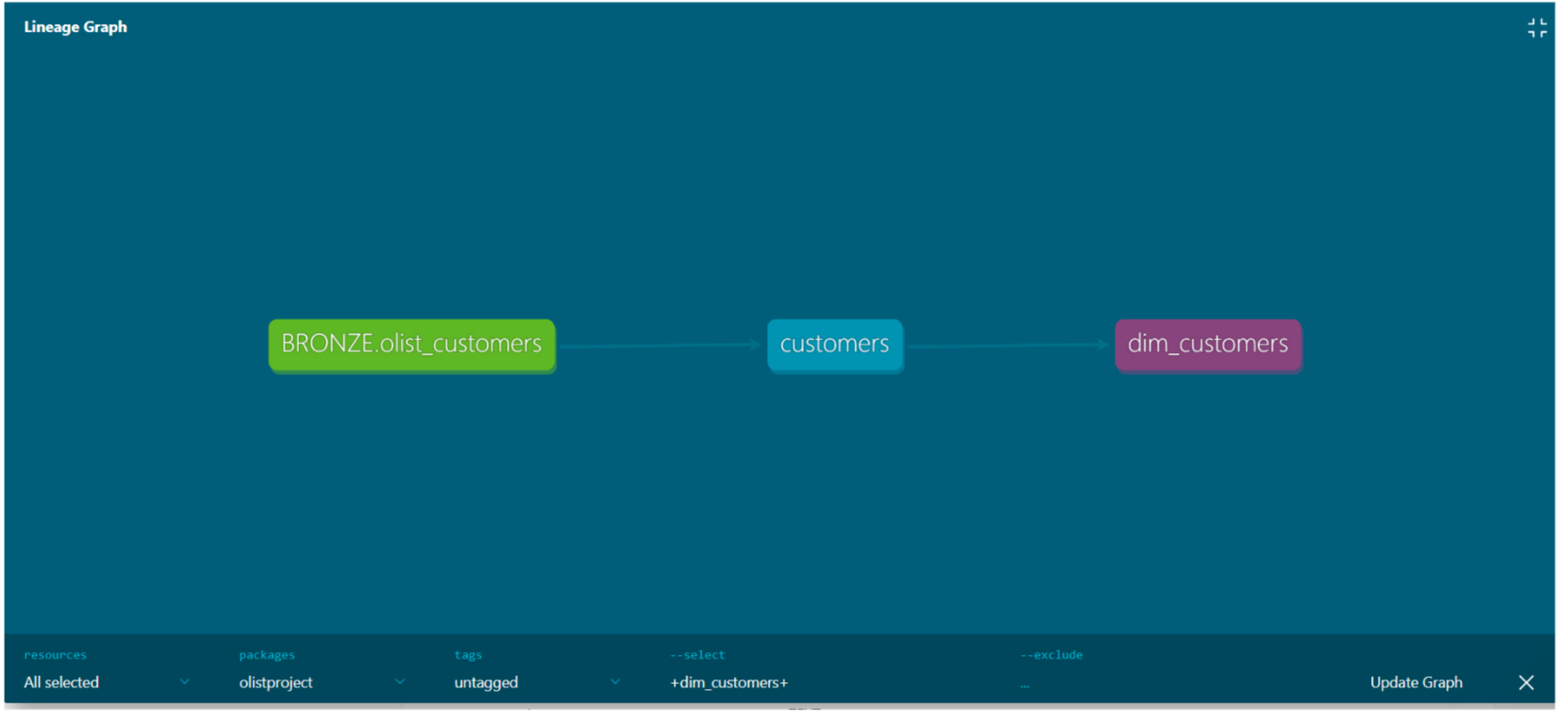
Modeling Approach

fact_orders dbt docs



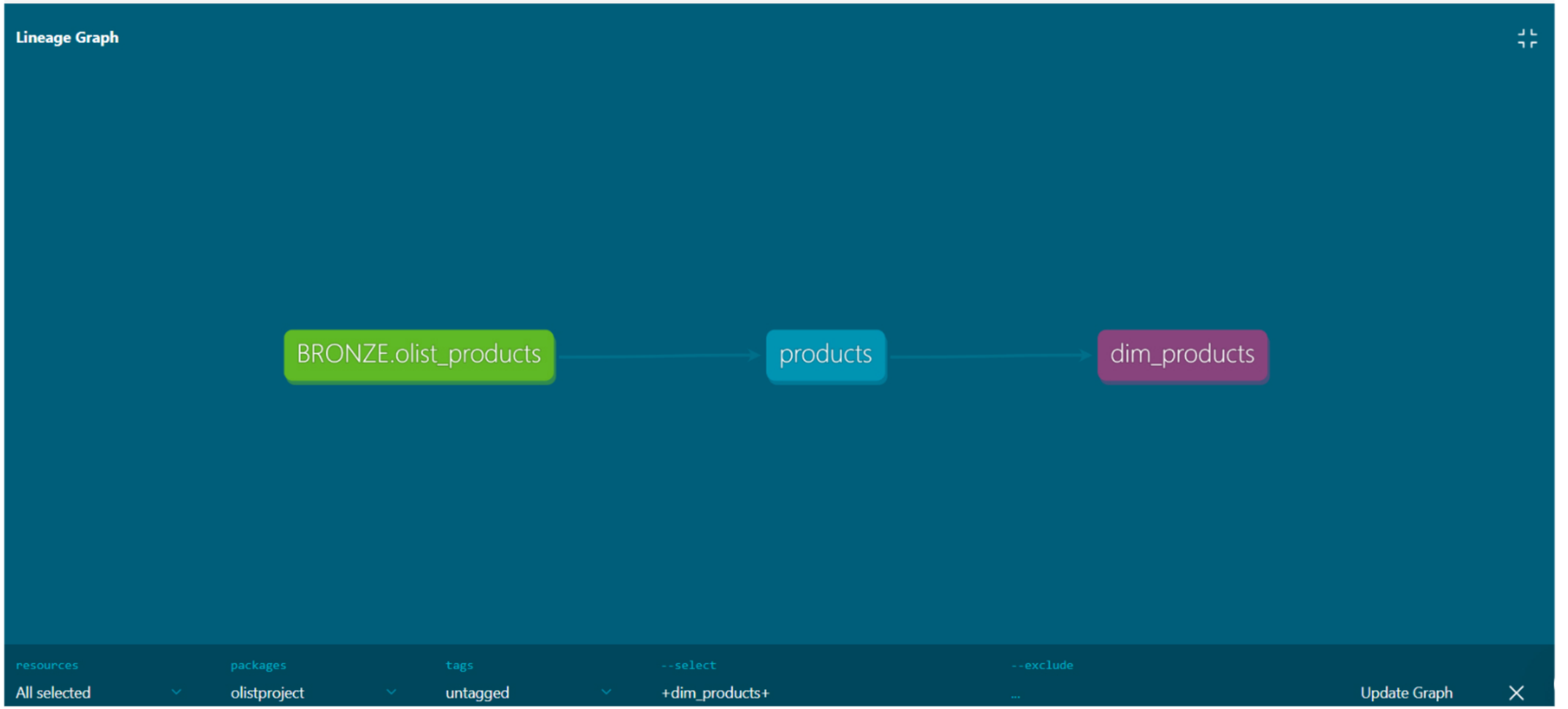
Modeling Approach

dim_customers dbt docs



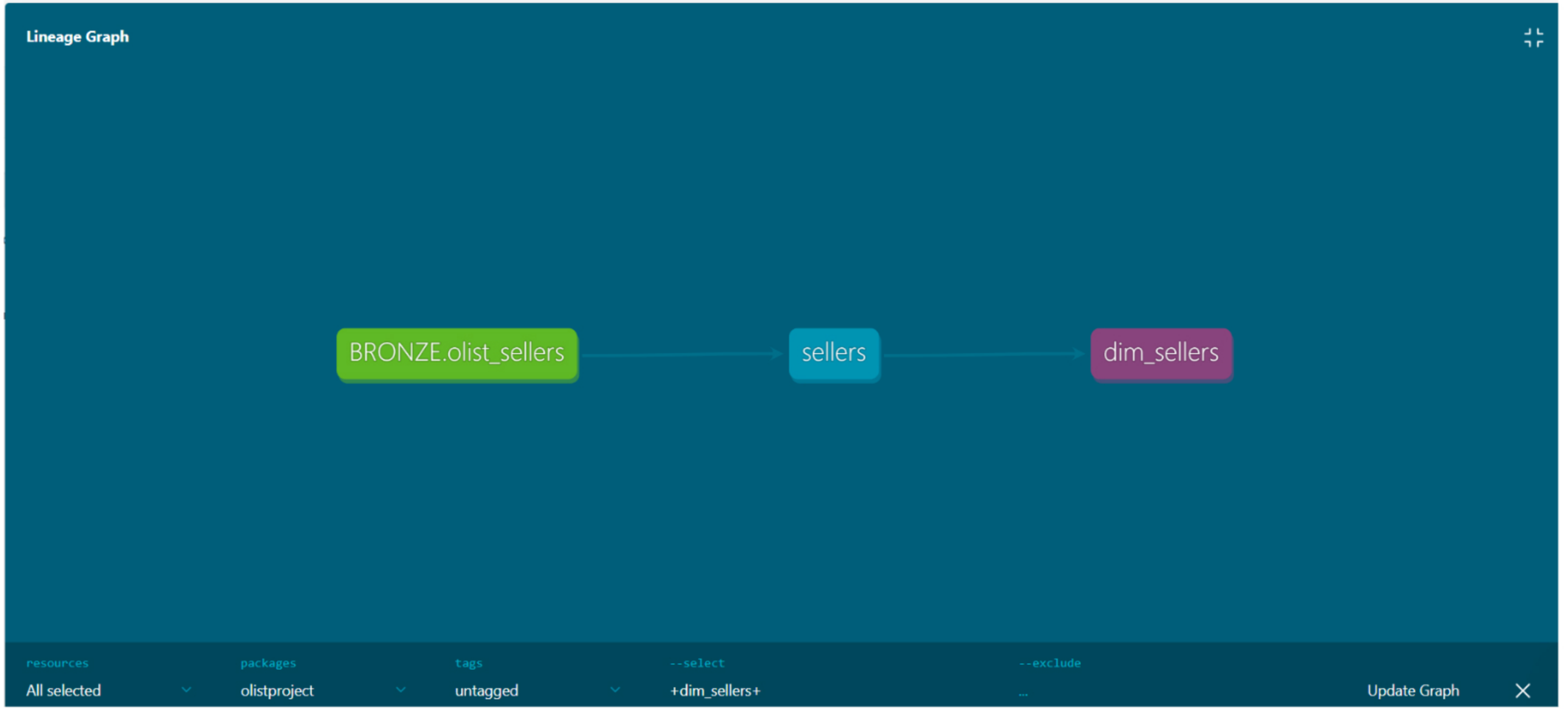
Modeling Approach

dim_products dbt docs



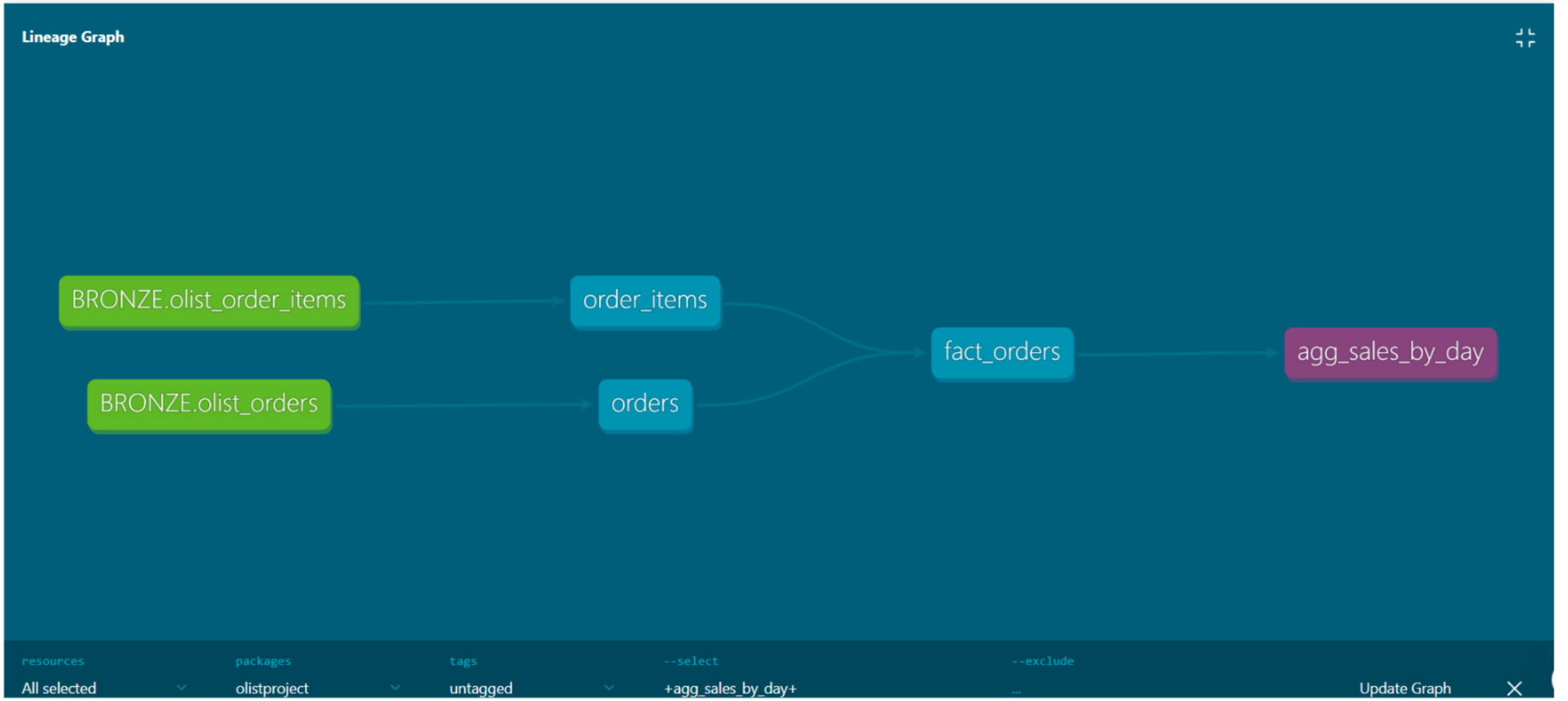
Modeling Approach

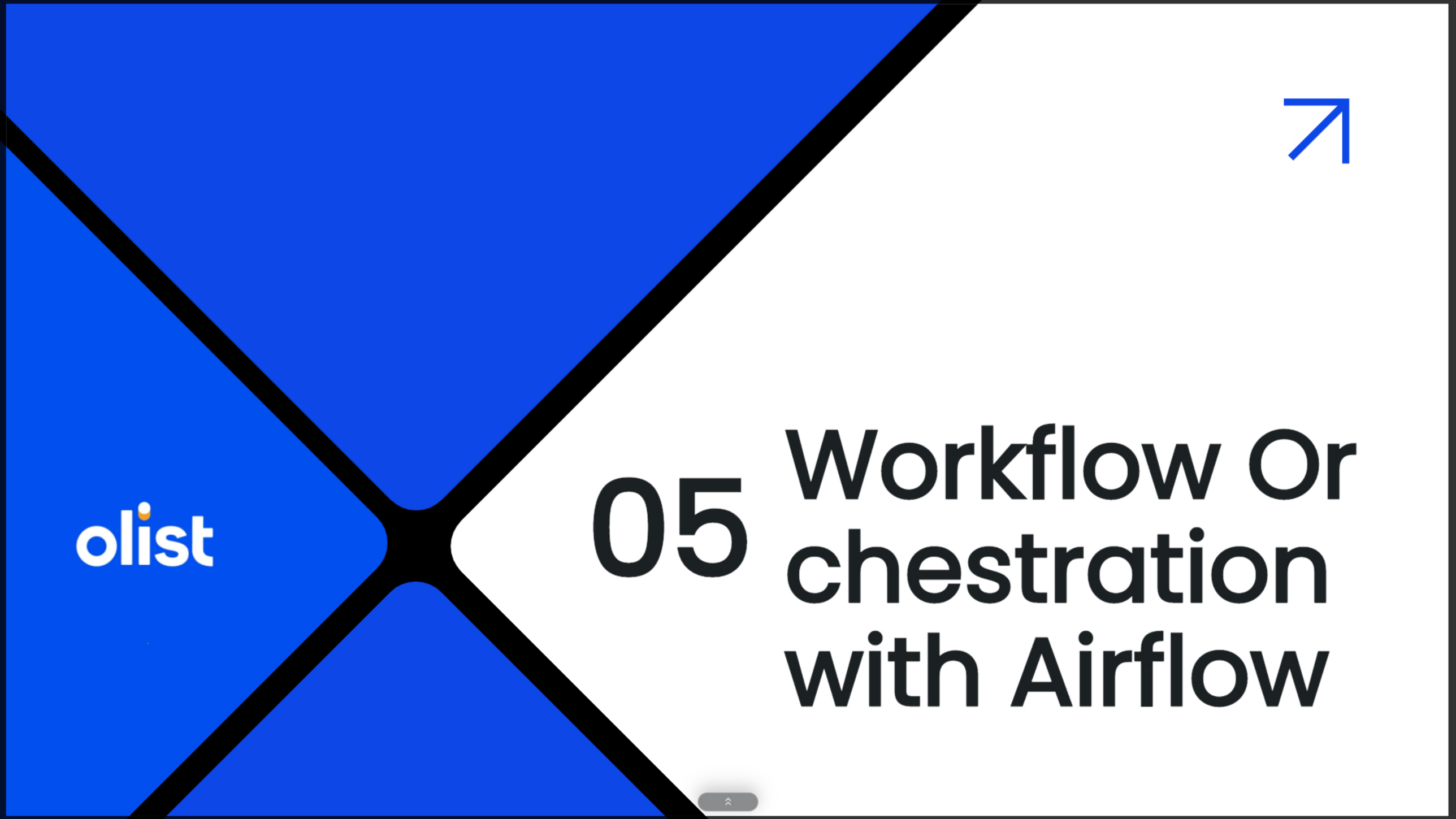
dim_sellers dbt docs



Modeling Approach

aggreation_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

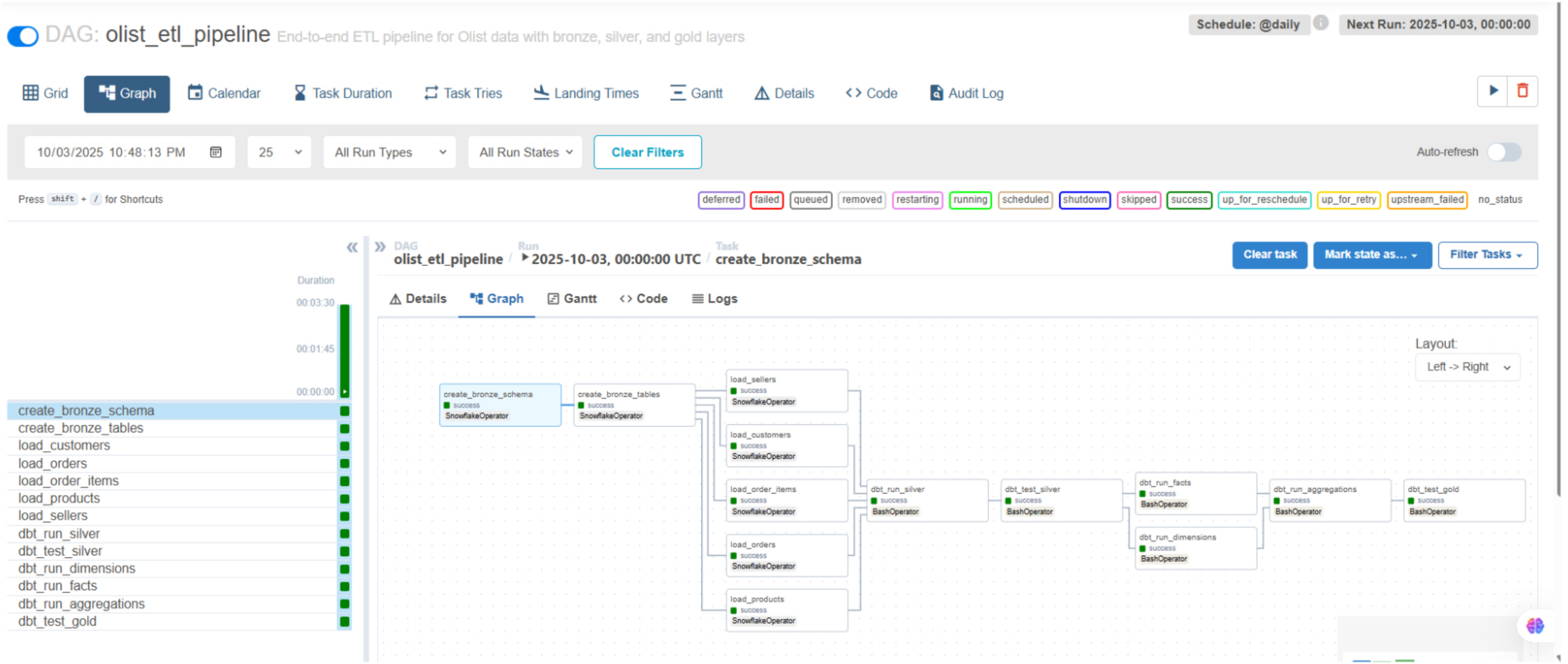
In the DAG design, failure handling includes automatic task retries with configurable limits and exponential backoff, ensuring pipeline resilience and minimizing manual intervention for smoother data processing.

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



olist

06 Data Visualization with Power BI

olist

Date Range | 2016 | 2017 | 2018

Navigators

- Executive Summary
- Customer Deep Dive
- Products & Sellers Performance
- Operations & Logistics Efficiency

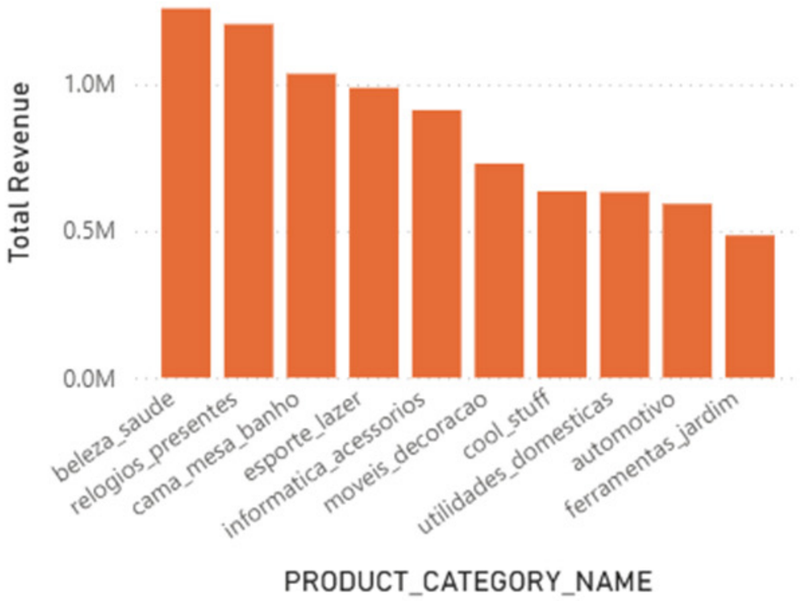
96K
Unique Customers

137.75
AOV

99K
Total Orders

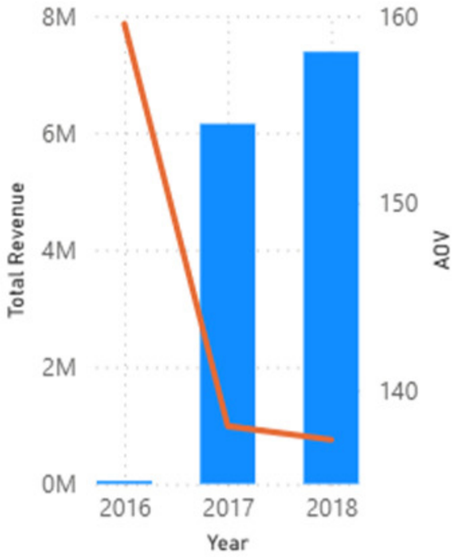
13.59M
Total Revenue

Top 10 Product Categories



Total Revenue by CUSTOMER_STATE

Monthly revenue performance and average order value



olist

Date Range | 2016 | 2017 | 2018

Navigators

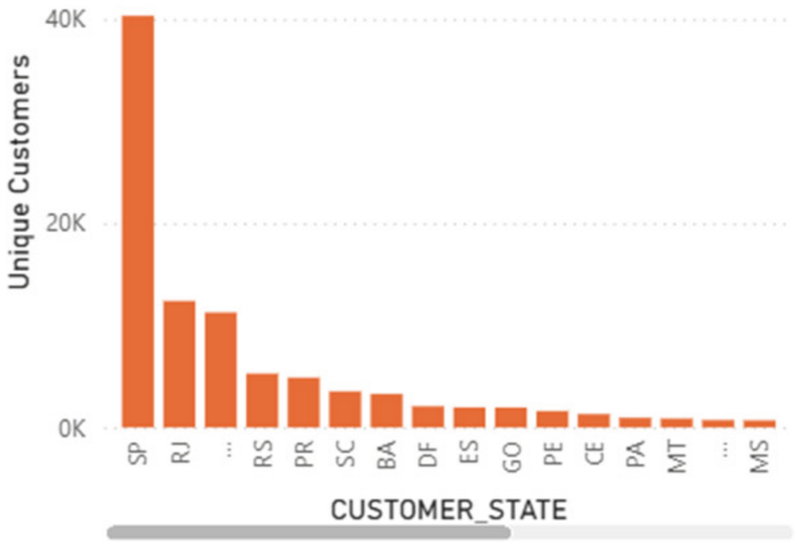
- Executive Summary
- Customer Deep Dive
- Products & Sellers Performance
- Operations & Logistics Efficiency

9/4/2016 9:15:19 PM
First Order Date

141.44
Average Revenue per Customer

96K
Unique Customers

Top 10 Clients by States



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**

olist

Date Range | 2016 | 2017 | 2018

Navigators

- Executive Summary
- Customer Deep Dive
- Products & Sellers Performance
- Operations & Logistics Efficiency

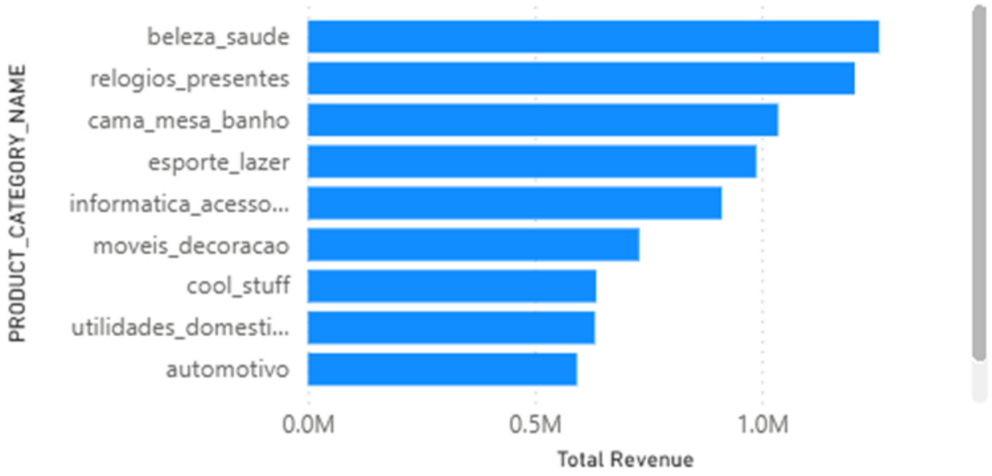
Total Revenue by CUSTOMER_STATE

3095
Total Sellers

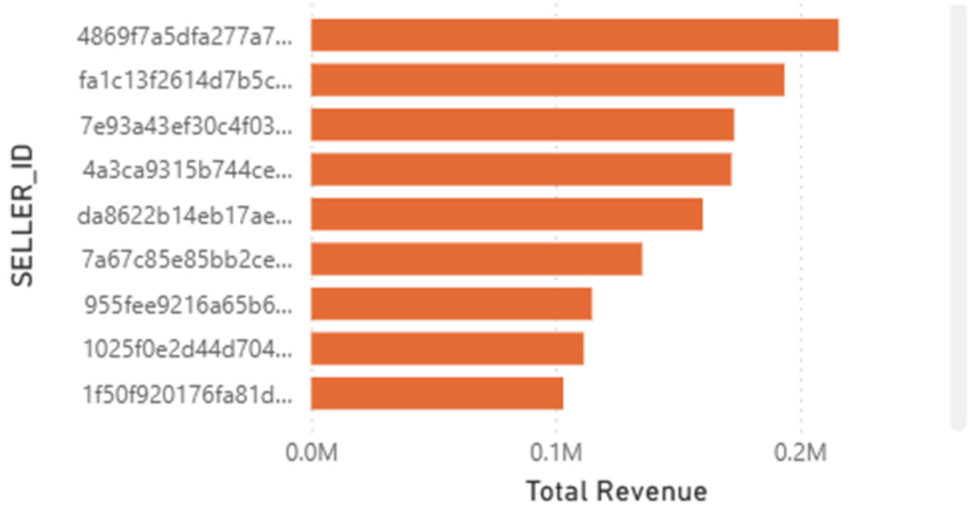
113K
Total Products Sold

120.65
Average Price

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



Top 10 sellers by revenue



olist

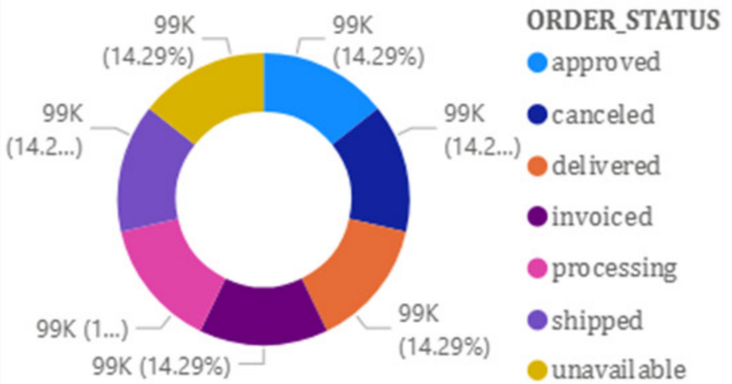
Date Range | 2016 | 2017 | 2018

Navigators

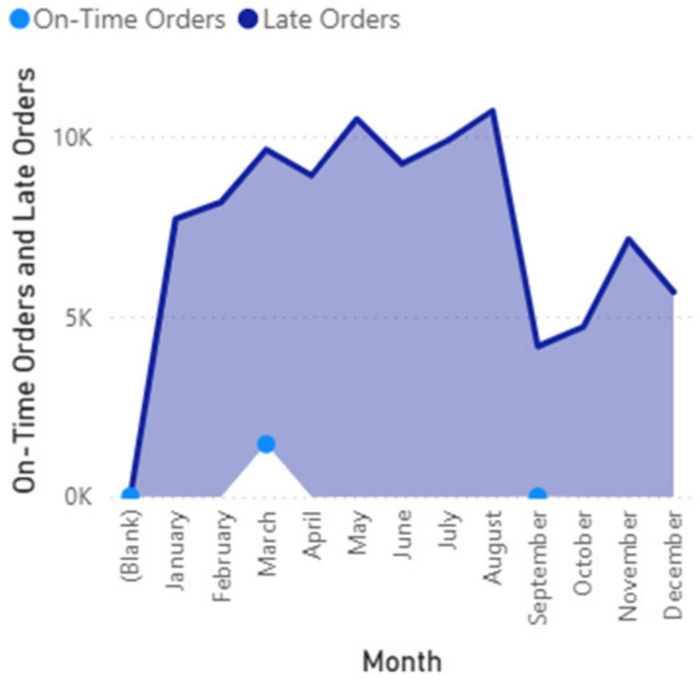
- Executive Summary
- Customer Deep Dive
- Products & Sellers Performance
- Operations & Logistics Efficiency

0.02	1477	96K	0.52
On-Time Delivery %	On-Time Orders	Total Delivered Orders	Average Delivery Time

Order Status Breakdown



On-Time Orders and Late Orders by Month



Total Revenue by CUSTOMER_STATE

Summary of Findings
and Next Steps

olist

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