olist Data Engineering Lab

E-COMMERCE DATA PIPELINE & DATA INSIGHTS

Data Engineering Team:
- Bandekar Anjali Ranjit
- Irene Lua
- M Logaraj
- Yee Soon Tuck

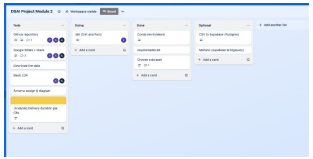
olist Data Engineering Lab

Agenda

1. Project Management Approach
2. Logical Data Pipeline Architecture
3. Platform Data Pipeline Architecture (2 Variations)
4. Data Warehouse Design
5. ELT Pipeline
6. Data Quality Testing
7. Data Analysis & Insights
8. Pipeline Orchestration

olist Data Engineering Lab

Trello



Project Management Strategy - Trello

- Trello was utilized for tracking project tasks and progress.
- Each team member has designated cards for their responsibilities.
- The board is organized into columns for 'To Do', 'In Progress', and 'Done'.
- Guide for daily meetings
- Able to give comments ensure clear communication and accountability

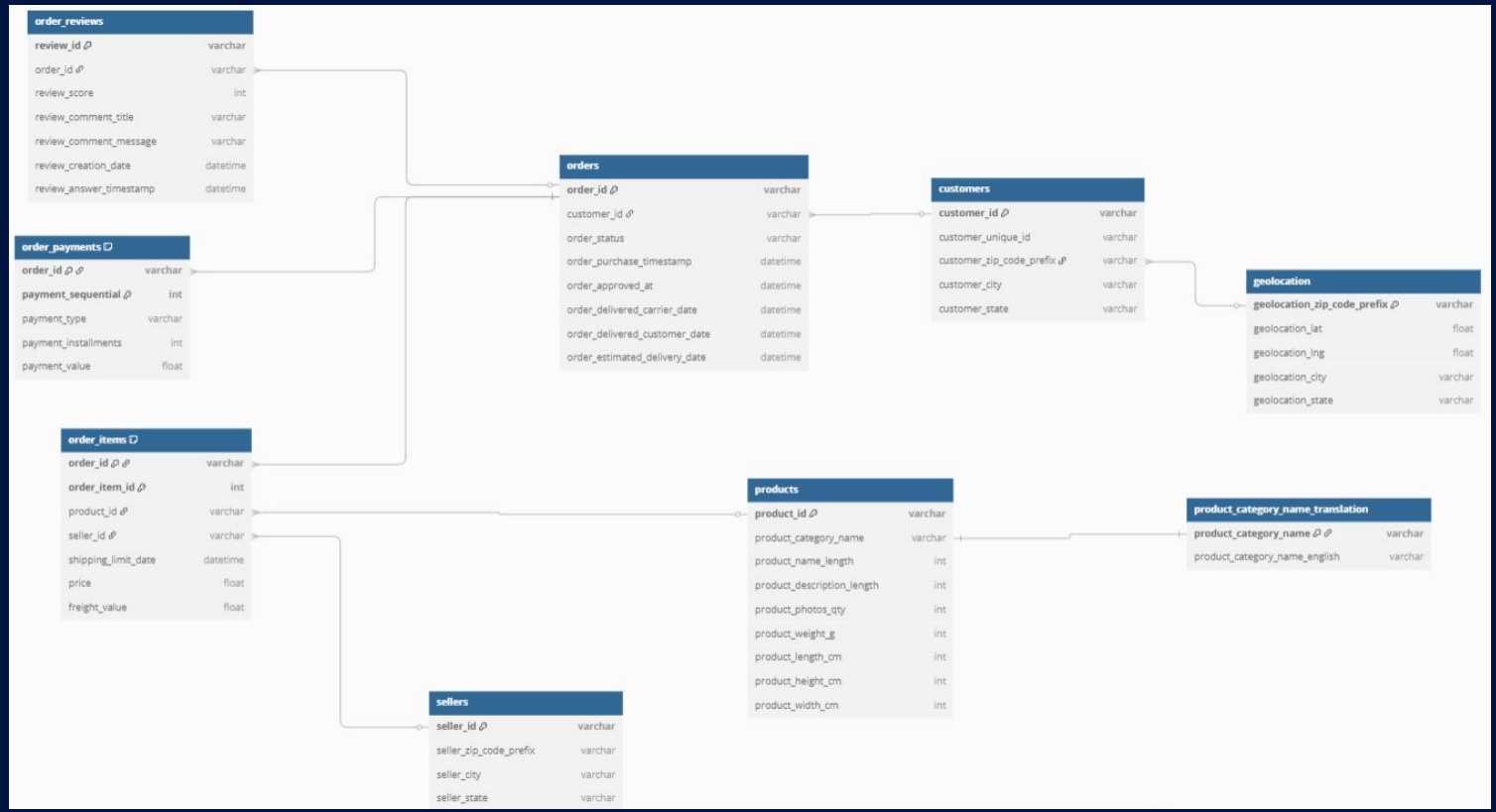
Benefits of Using Trello

- Visual organization enhances team collaboration and transparency.
- Easy to assign tasks and set deadlines within the platform.
- New ideas/links could be easily updated and read by team members
- Trello's mobile app allows team members to update tasks on the go.

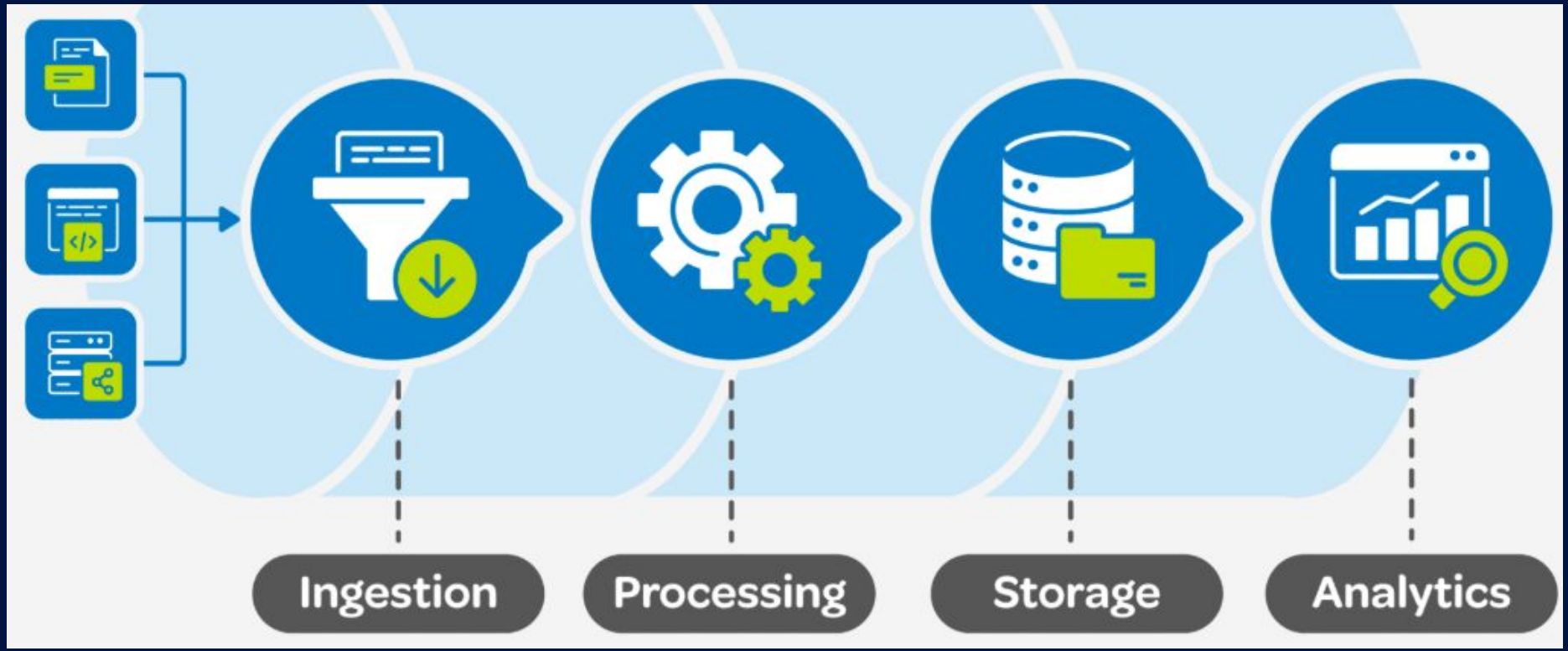
olist Data Engineering Lab

Source Data

Data set: Brazilian E-Commerce Public Dataset by Olist



Logical Data Pipeline Architecture



olist Data Engineering Lab

Data Pipeline Architecture v 1.0

Data Source

Raw Data

Data Storage

BigQuery

Data Preparation

DBT

pandas_gbq

SQL

BI & Analytics

Python

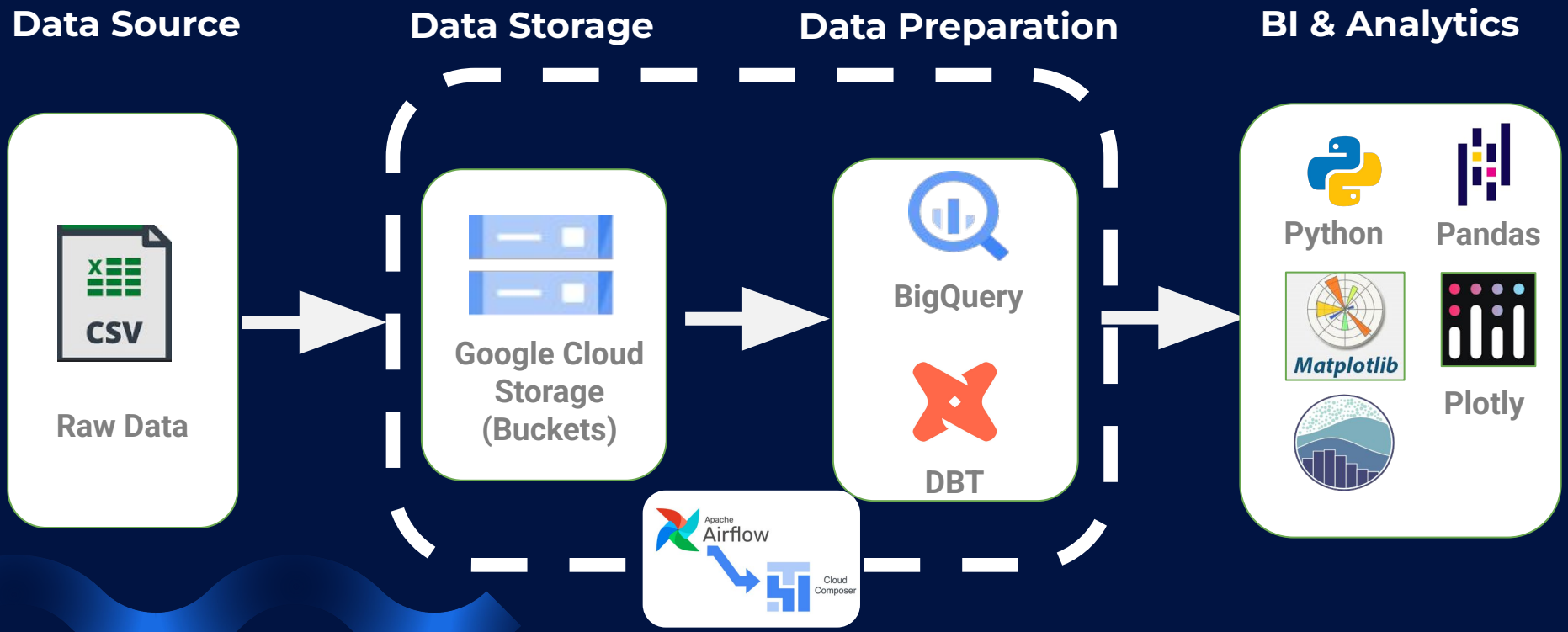
Pandas

Plotly

Seaborn

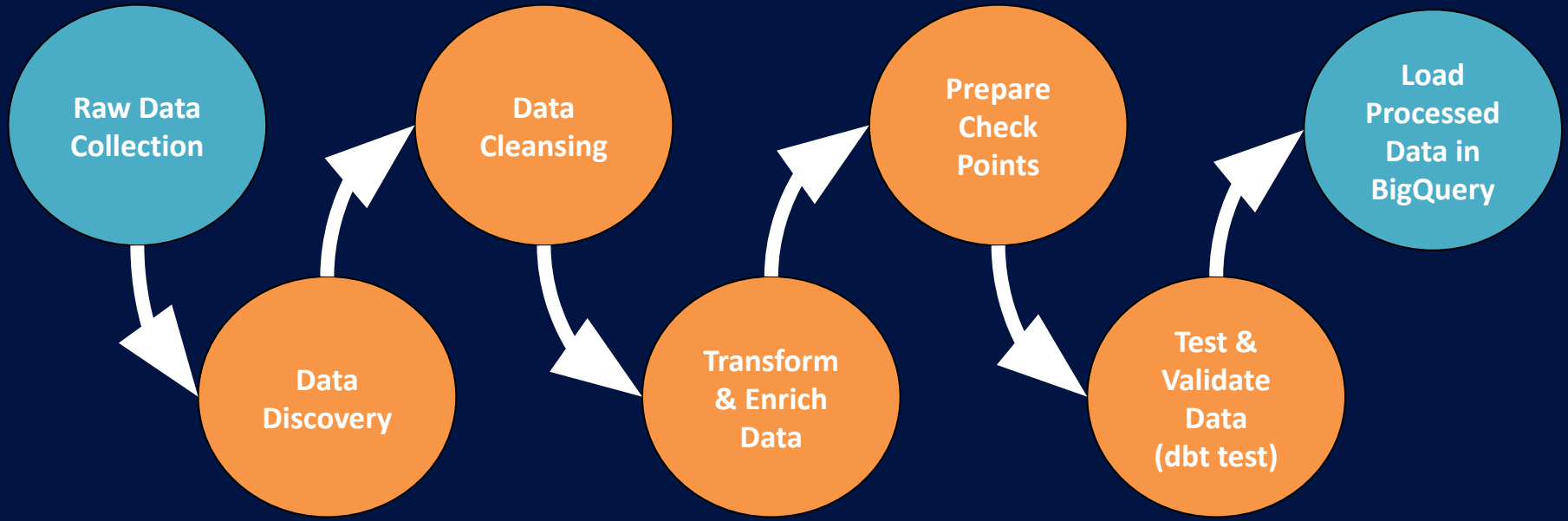
olist Data Engineering Lab

Data Pipeline Architecture v 2.0



olist Data Engineering Lab

Data Preparation

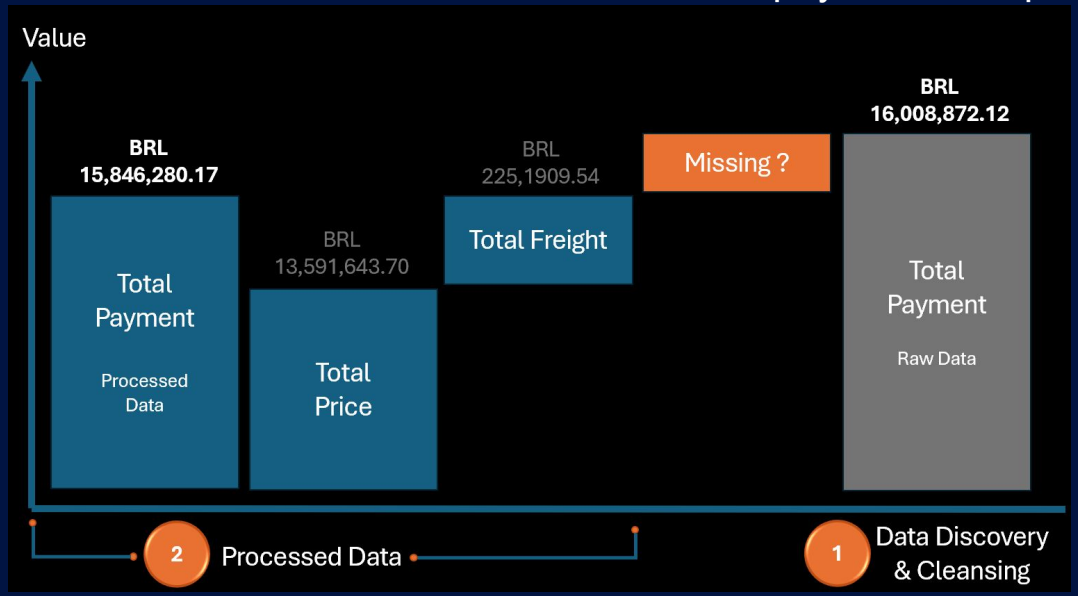


olist Data Engineering Lab

Data Exploration & Validation

Test Case:

Does the delivered order value match our payment receipts?



Test Case working files link

New Query : Checking **Missing Data**
Total : 833 payments transactions
776 distinct orders with BRL162,591.95

	order_status	order_status_count	price_sum
0	canceled	179	0.00
1	created	5	0.00
2	delivered	3	134.97
3	invoiced	2	0.00
4	shipped	1	0.00
5	unavailable	643	0.00

Check Point:
We accept ‘canceled’ status for dirty data.
But the rest of the transactions require to rectify and clean up up.

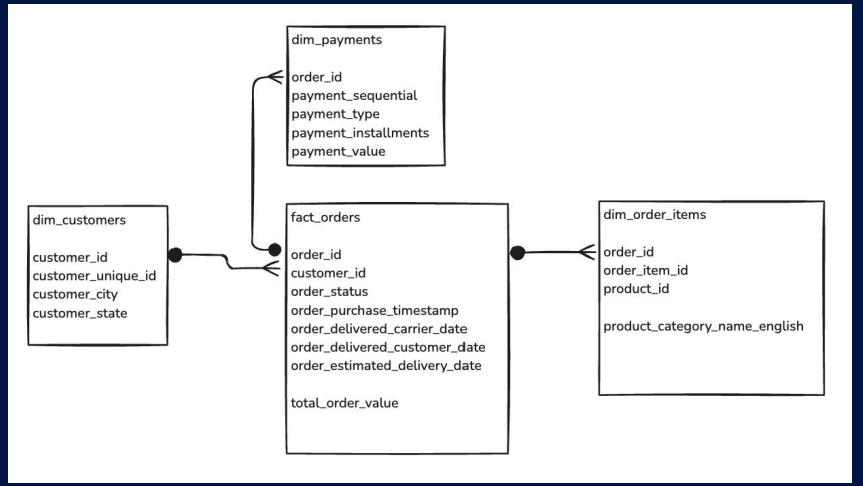
```yaml
models:
  - name: check_order_item_payment
    description: 'Model description'
    columns:
      - name: order_status
        description: Only 'canceled' status is acceptable here.
        tests:
          - dbt_utils.expression_is_true:
              expression: "{{ order_status }} = 'canceled'"
```

olist Data Engineering Lab

Data Warehouse Design (Star Schema)

Business Question: Providing Insights on Delivery Performance of Olist

1. Central Table — **fact_orders**
   a. Tracks every order placed, including key timestamps, delivery status, and total value.
      ➤ It's the heart of the schema where all delivery metrics are measured.
2. Customer Dimension — **dim_customers**
   a. Adds geographic context (city, state) to each order, enabling analysis of regional delivery trends and customer behavior.
3. Item Dimension — **dim_order_items**
   a. Breaks down orders into individual items with price, freight cost, and product category — crucial for understanding delivery cost drivers and product-based delays.
4. Payment Dimension — **dim_order_payments**
   a. Links payment methods and values to orders, allowing insights into how payment types affect delivery performance (e.g., delays with boleto payments).



olist Data Engineering Lab

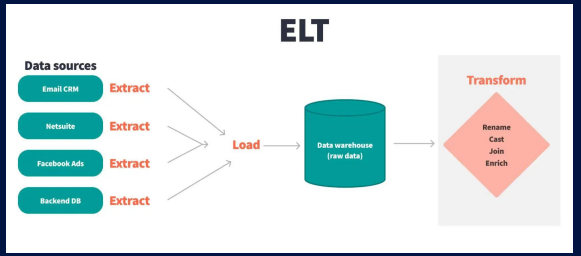
ELT Pipeline Features

1. **Transformation of Raw Data into a Star Schema Using dbt:**
   - Utilized dbt to structure the raw data into a star schema, enhancing query performance and simplifying data analysis.
   - Created SQL models to represent different dimensions and fact tables within the schema.

2. **Implementation of Data Cleaning and Validation Steps:**
   - Employed dbt's testing capabilities to ensure data quality by defining tests within the project, such as checking for uniqueness, non-null constraints, dbt_utilis functions

3. **Creation of Derived Columns:**
   - Developed additional metrics like **total_order_amount** within dbt models to enrich the dataset and provide deeper analytical insights.

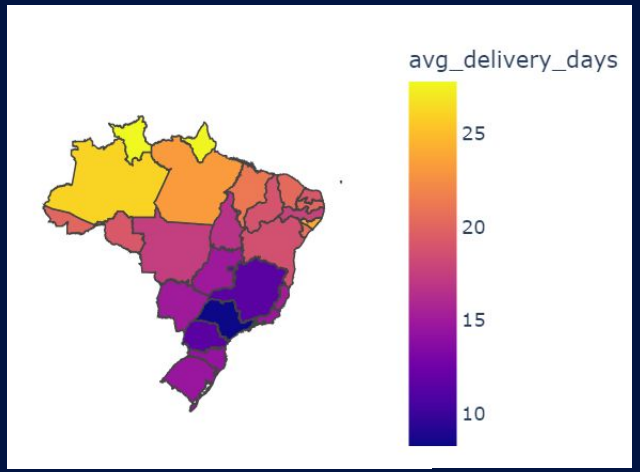
4. **Exploratory Data Analysis Using Jupyter Notebooks:**
   - Utilized Jupyter Notebooks to perform exploratory data analysis on the transformed data, allowing for interactive examination and visualization of trends and patterns.

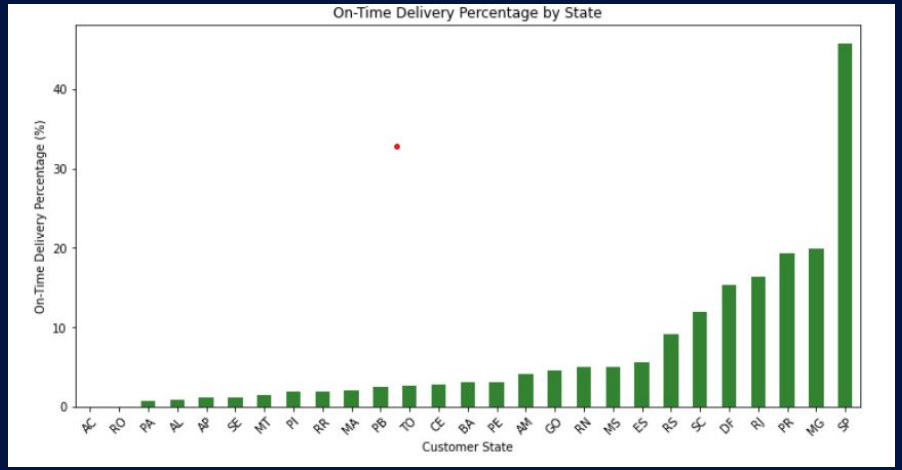


olist Data Engineering Lab

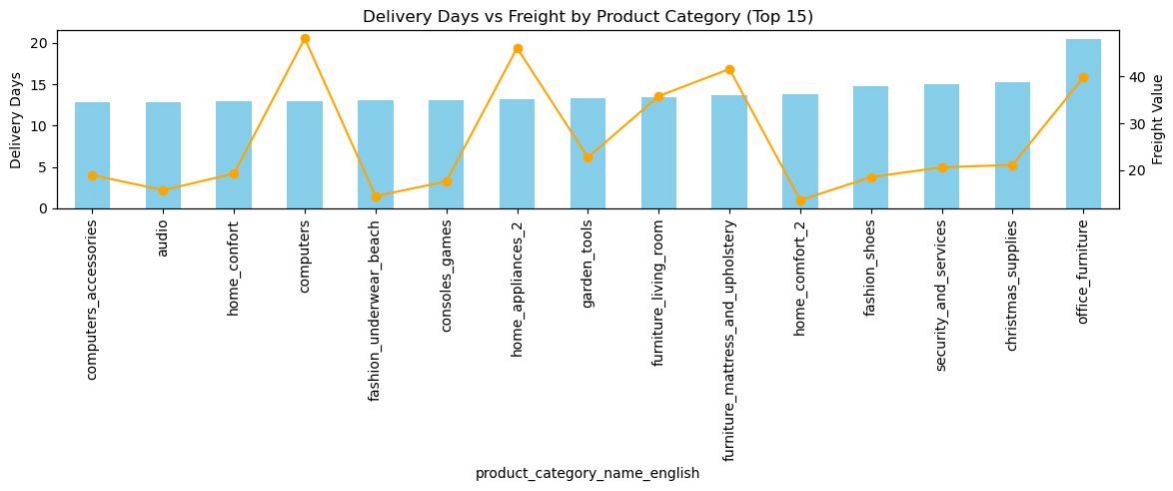
Data Analysis Insights

1) Northern States in Brazil had the highest average delivery dates



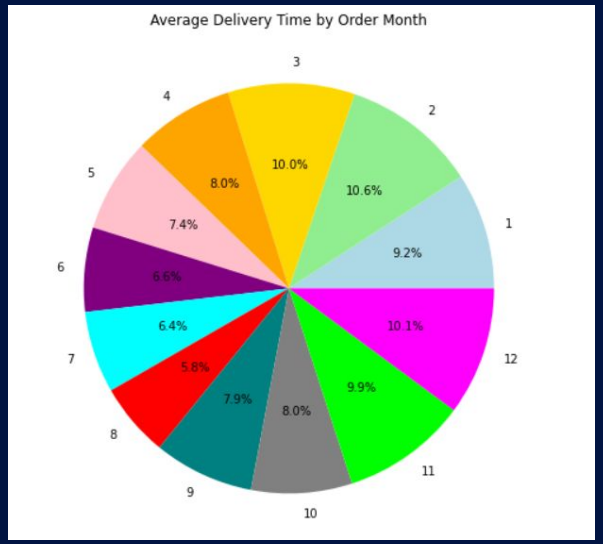


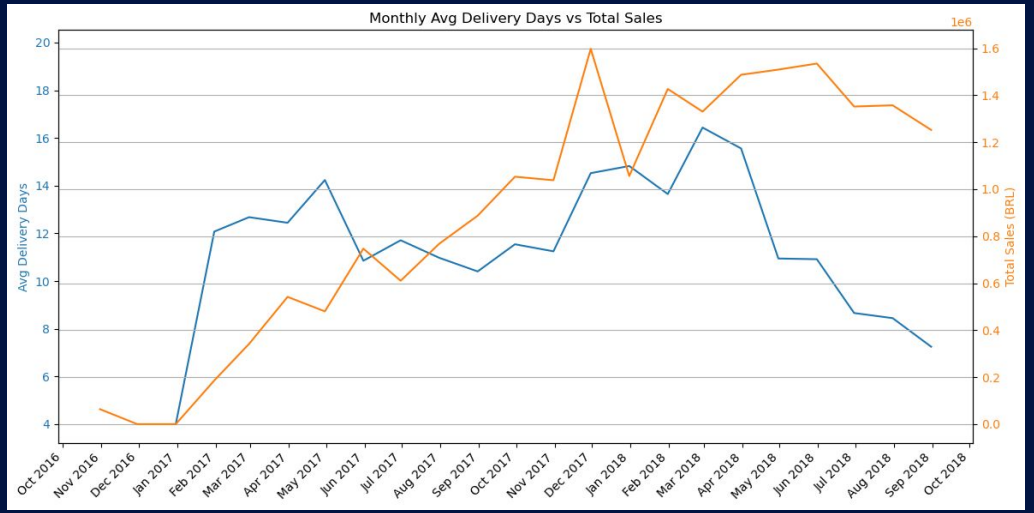
2) Bulkier Items show a stronger correlation to higher delivery times



olist Data Engineering Lab

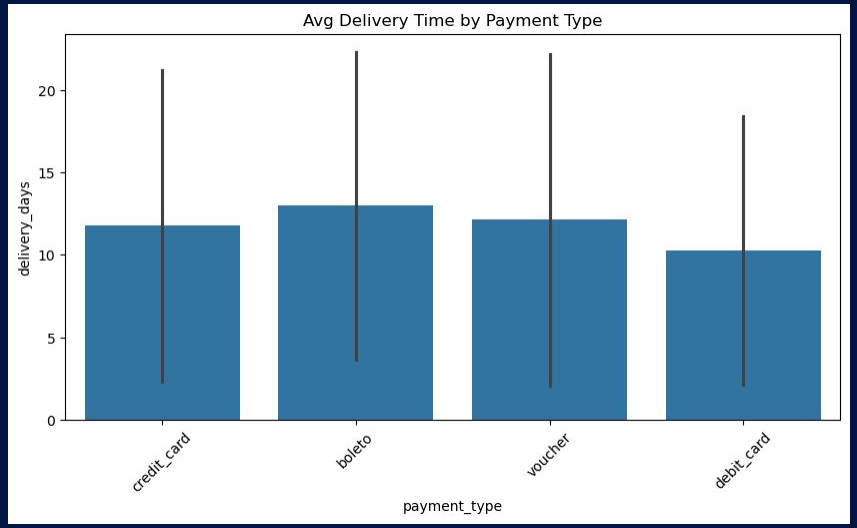
3) Delivery times peak in February, March, November, and December due to seasonal demand, holidays

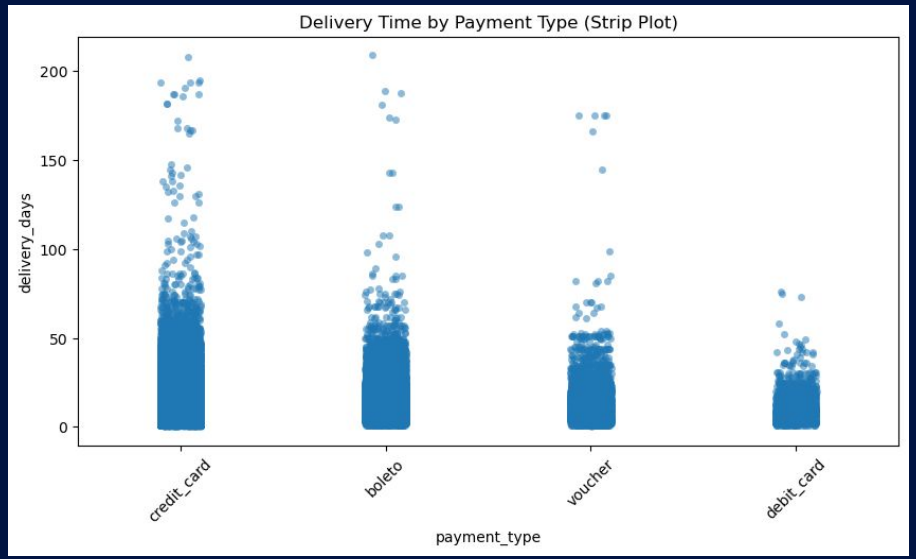




olist Data Engineering Lab

4) Boleto & Credit Card Transactions have higher delivery times





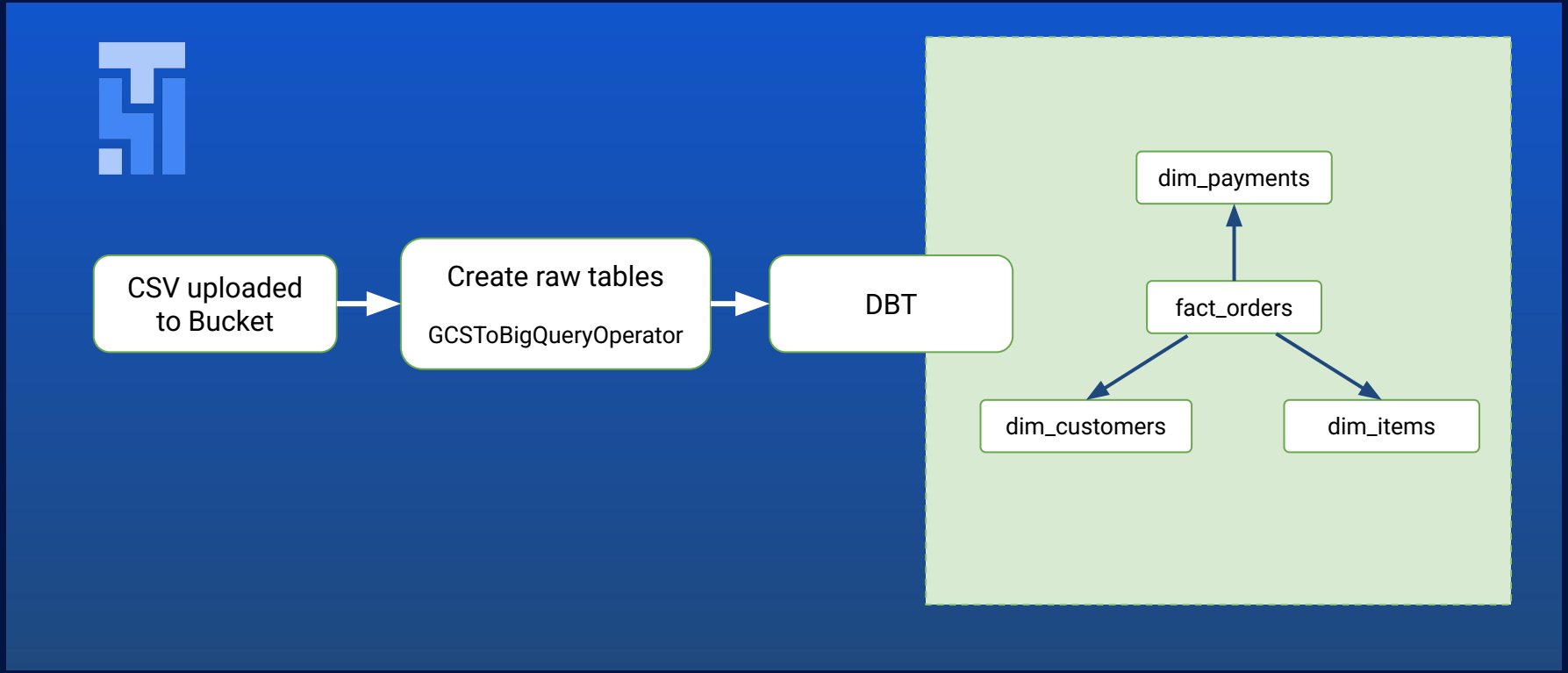
olist Data Engineer Lab

Pipeline Orchestration - Google Cloud Composer

Managed service, using Apache Airflow.
Note: this orchestration pipeline is connected to a different dataset in BigQuery (composer_test) for the purpose of presentation.
The other parts of this Project uses the “brazil_e_commerce” dataset.

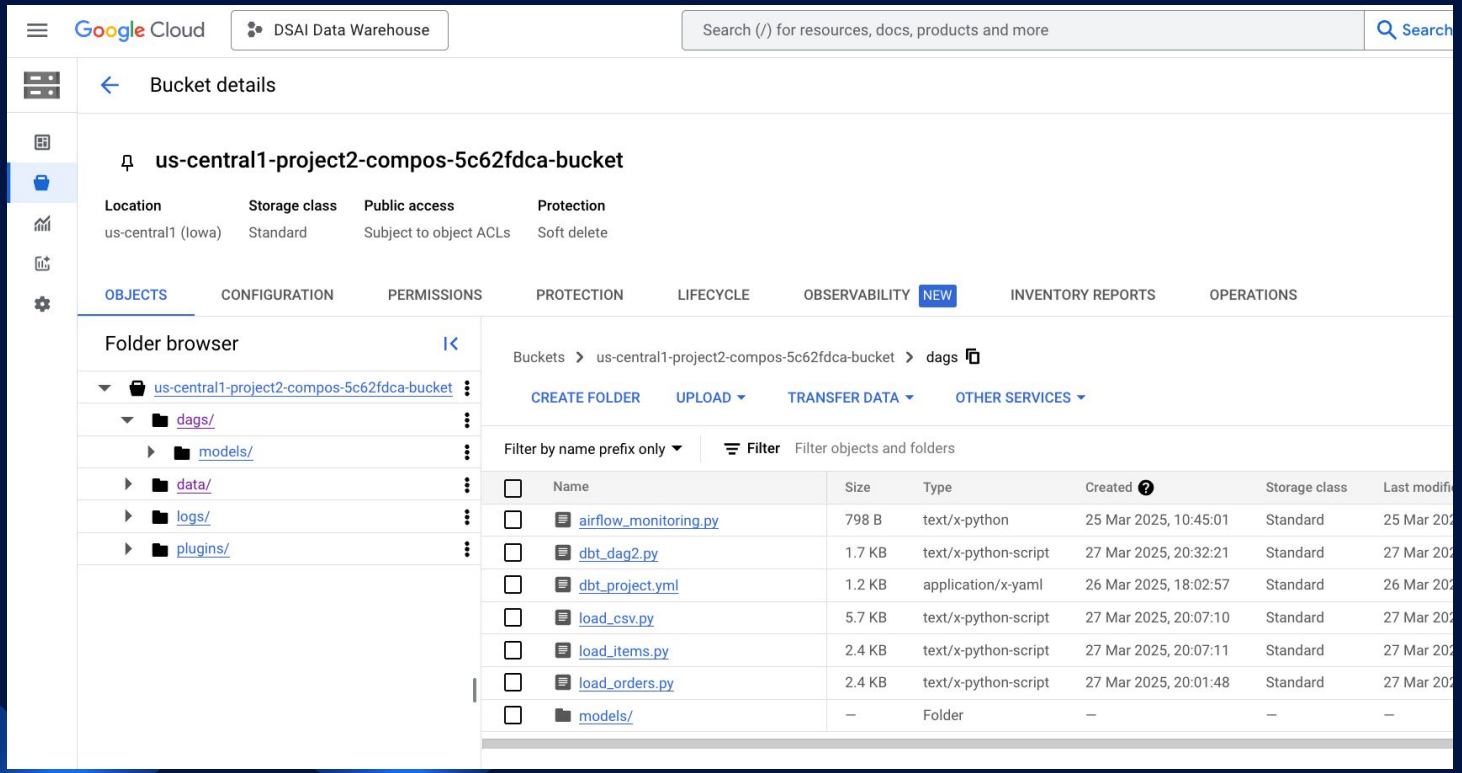
olist Data Engineer Lab

Data Pipeline in Google Cloud Composer



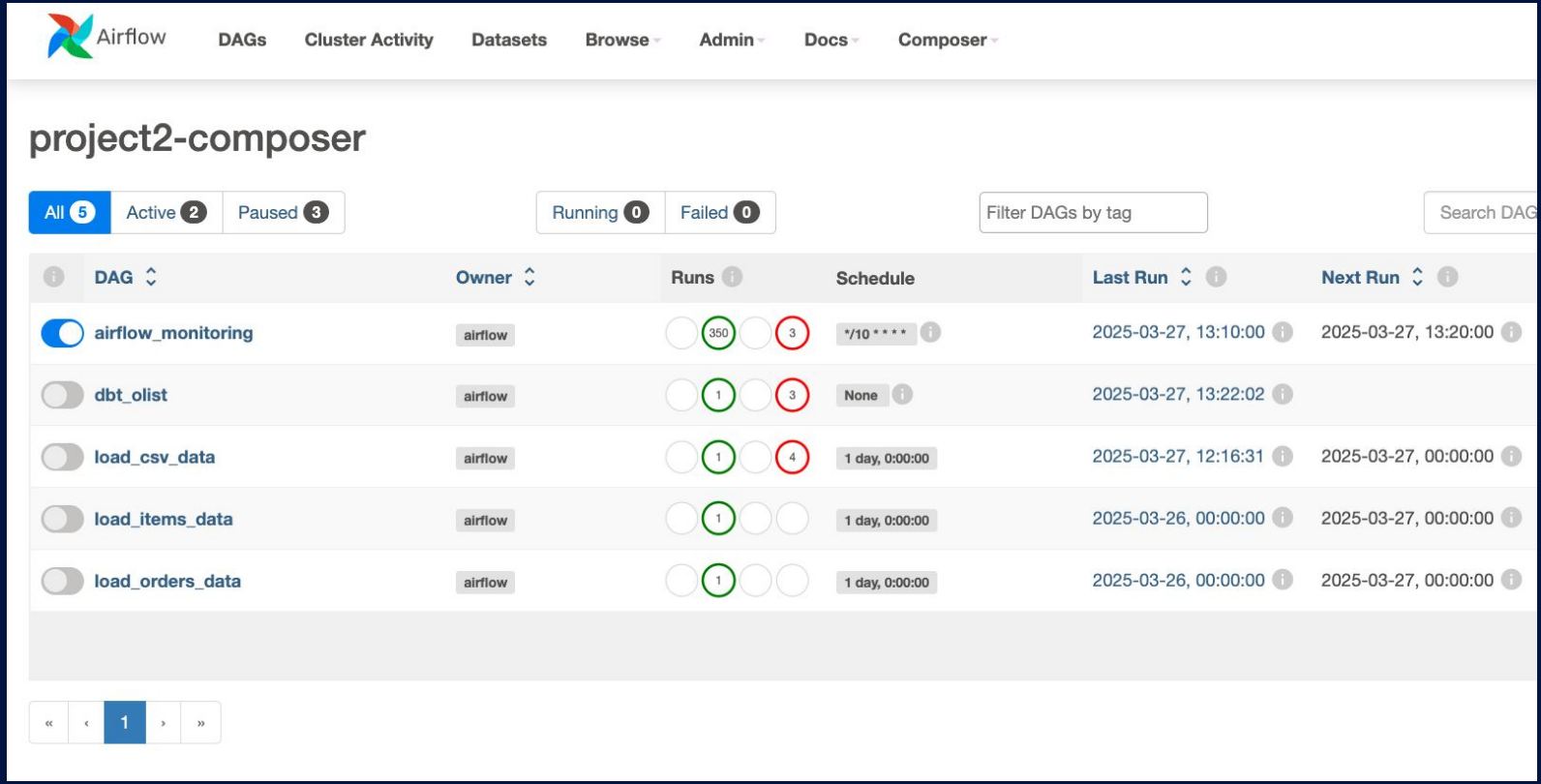
olist Data Engineer Lab

DAGs folder in Bucket

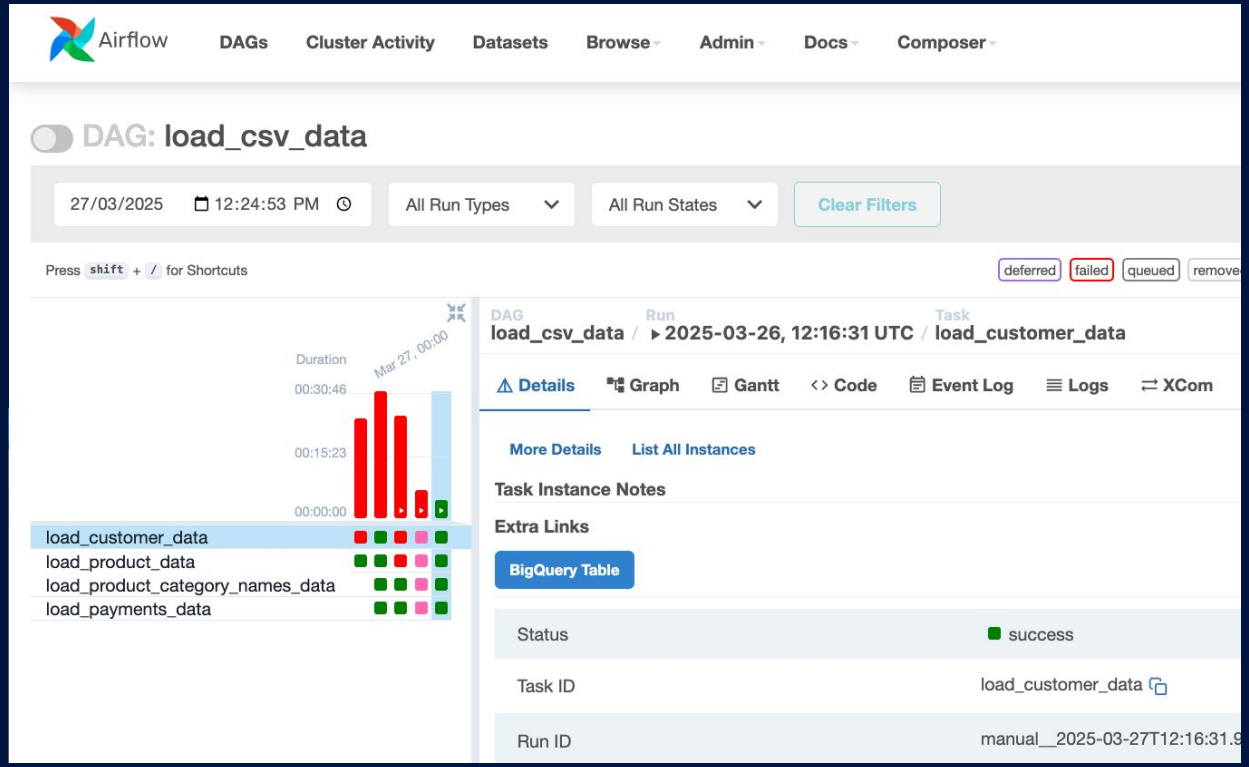


olist Data Engineer Lab

Airflow DAGs in Google Cloud Composer

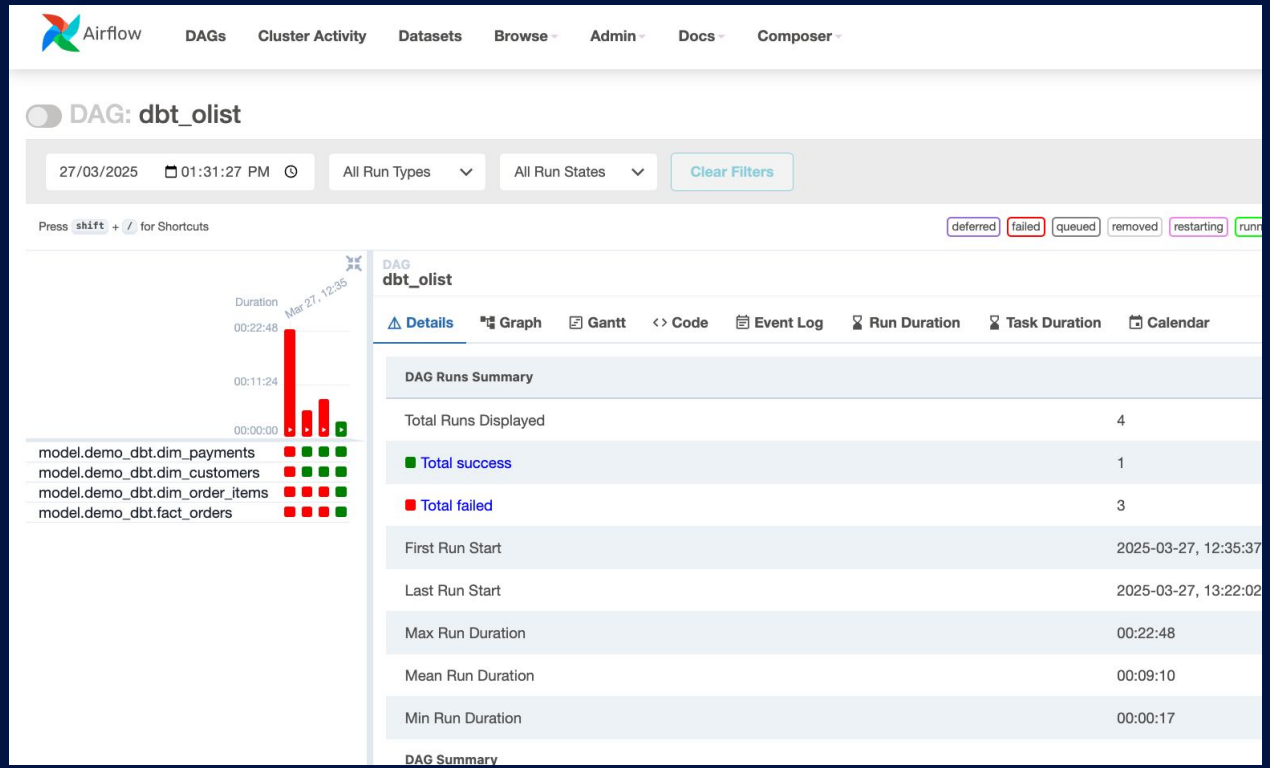


load_csv_data DAG run logs

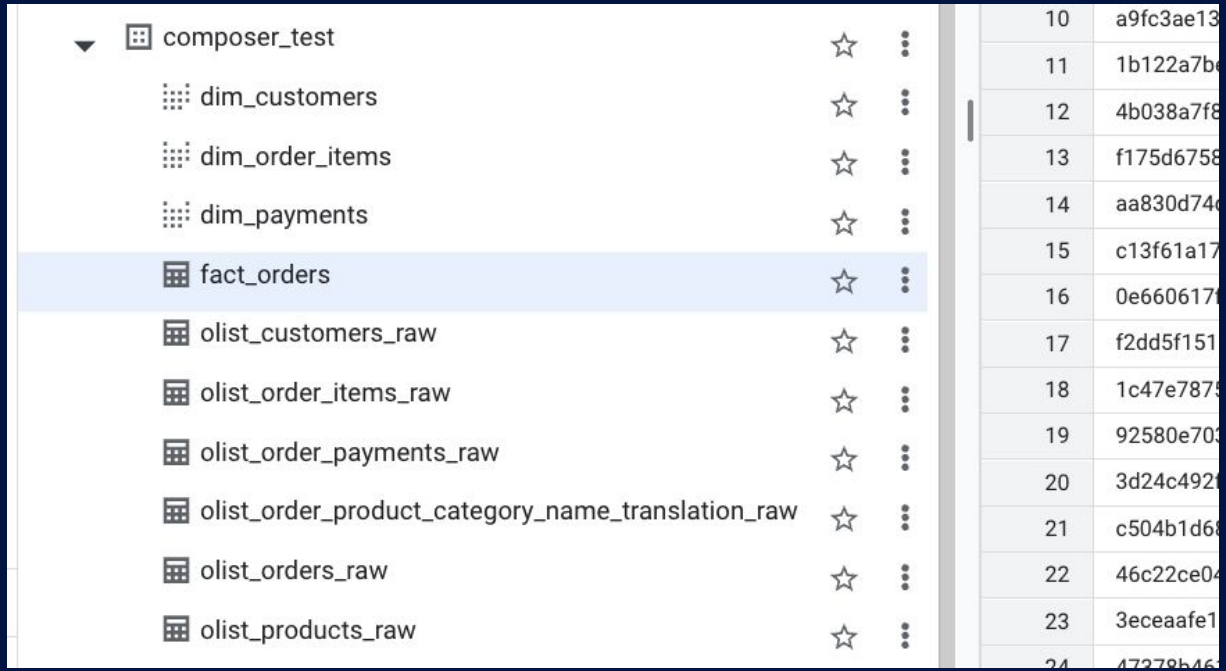


olist Data Engineer Lab

DBT running as DAG



Tables and Views in BigQuery



olist Data Engineer Lab

References

https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data

https://cloud.google.com/bigquery/docs/samples/bigquery-pandas-gbq-read-gbq-simple

https://stripe.com/en-sg/resources/more/boleto-an-in-depth-guide

https://www.olist.com/

https://docs.getdbt.com/guides/bigquery

https://schemas.getdbt.com/dbt/manifest/v12/index.html#nodes_additionalProperties_anyOf_i4

olist Data Engineer Lab

THANK YOU

+123-456-7890

HELLO@OLIST-DATAELAB.COM

WWW.OLIST-DATAELAB.COM

123 MONSERATT ST., SR, ST 12345