olist Data Engineering Lab

E-COMMERCE DATA PIPELINE & DATA INSIGHTS

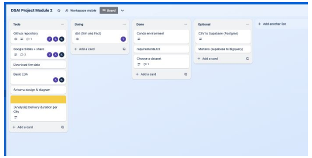
Data Engineering Team:

- 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



DSAI Project Module 2

Todo
- Github repository
- Google Slides + share
- Download the data
- Basic EDA
- Schema design & diagram
- [Analysis] Delivery duration per City

Doing
- dbt (Dim and Fact)

Done
- Conda environment
- requirements.txt
- Choose a dataset

Optional
- CSV to Supabase (Postgres)
- Meltano (supabase to bigquery)

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

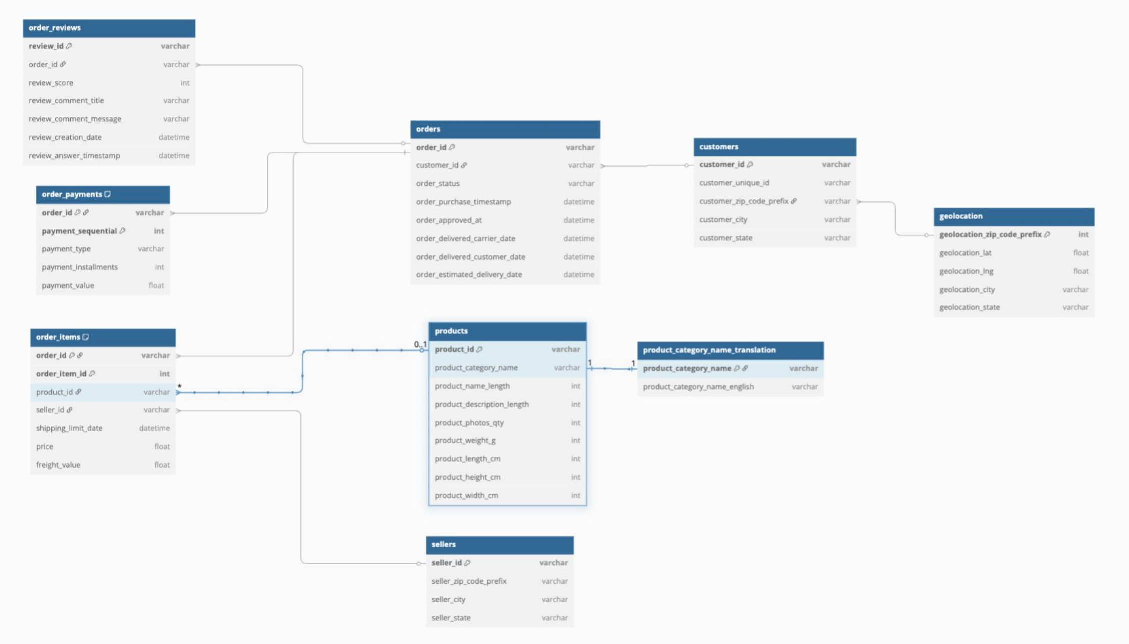
Benefits of Using Trello

- ❑ 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



order_reviews	
review_id	varchar
order_id	varchar
review_score	int
review_comment_title	varchar
review_comment_message	varchar
review_creation_date	datetime
review_answer_timestamp	datetime

order_payments	
order_id	varchar
payment_sequential	int
payment_type	varchar
payment_installments	int
payment_value	float

order_items	
order_id	varchar
order_item_id	int
product_id	varchar
seller_id	varchar
shipping_limit_date	datetime
price	float
freight_value	float

orders	
order_id	varchar
customer_id	varchar
order_status	varchar
order_purchase_timestamp	datetime
order_approved_at	datetime
order_delivered_carrier_date	datetime
order_delivered_customer_date	datetime
order_estimated_delivery_date	datetime

customers	
customer_id	varchar
customer_unique_id	varchar
customer_zip_code_prefix	varchar
customer_city	varchar
customer_state	varchar

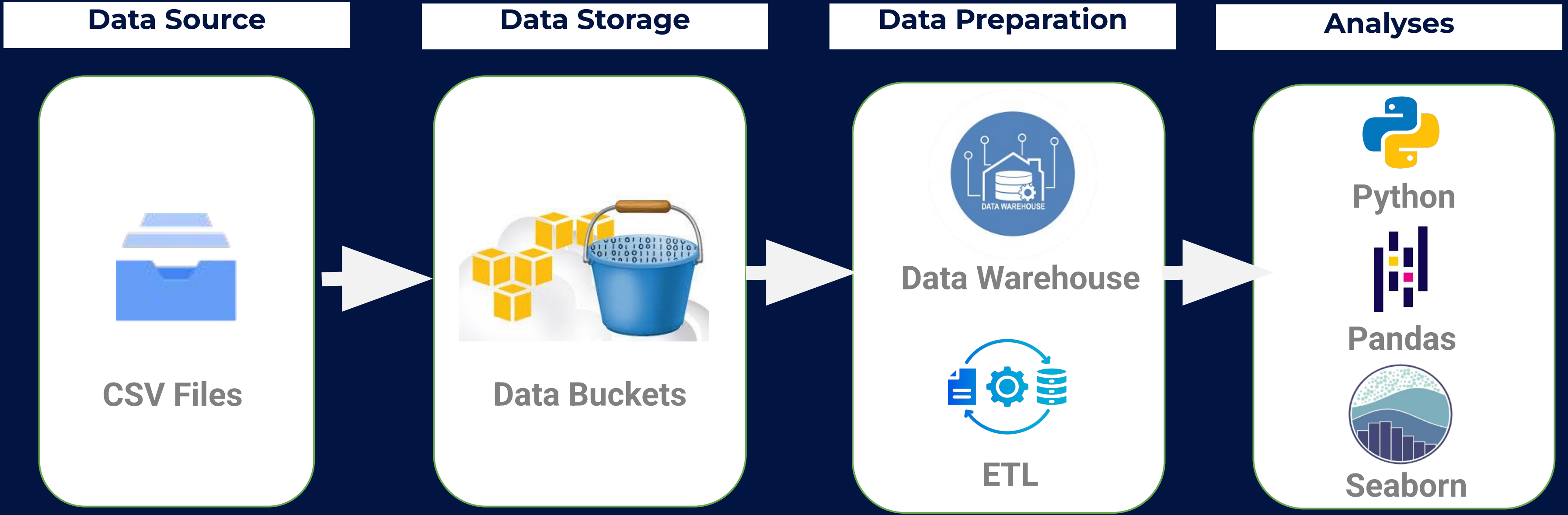
geolocation	
geolocation_zip_code_prefix	int
geolocation_lat	float
geolocation_lng	float
geolocation_city	varchar
geolocation_state	varchar

products	
product_id	varchar
product_category_name	varchar
product_name_length	int
product_description_length	int
product_photos_qty	int
product_weight_g	int
product_length_cm	int
product_height_cm	int
product_width_cm	int

product_category_name_translation	
product_category_name	varchar
product_category_name_english	varchar

sellers	
seller_id	varchar
seller_zip_code_prefix	varchar
seller_city	varchar
seller_state	varchar

Logical Data Pipeline Architecture



Data Source

CSV Files

Data Storage

Data Buckets

Data Preparation

Data Warehouse

ETL

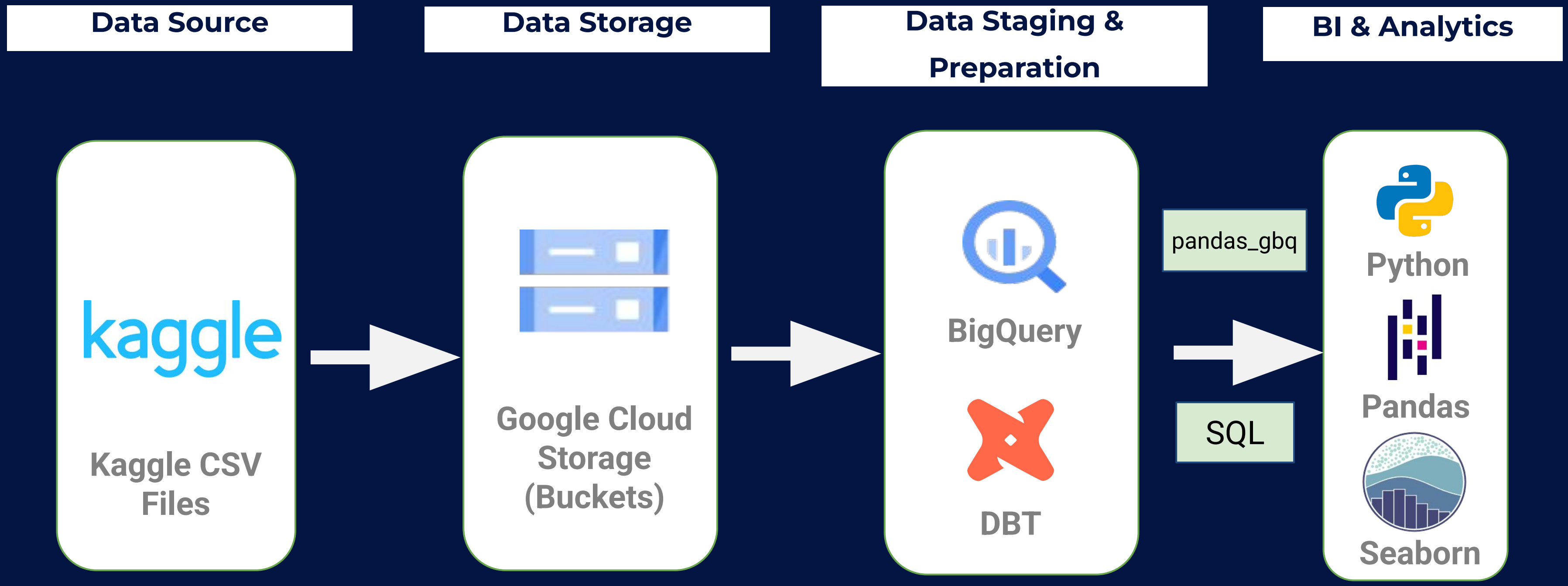
Analyses

Python

Pandas

Seaborn

Platform Data Pipeline Architecture v 1.0



Data Source

Data Storage

Data Staging & Preparation

BI & Analytics

Kaggle CSV Files

Google Cloud Storage (Buckets)

BigQuery

DBT

pandas_gbq

SQL

Python

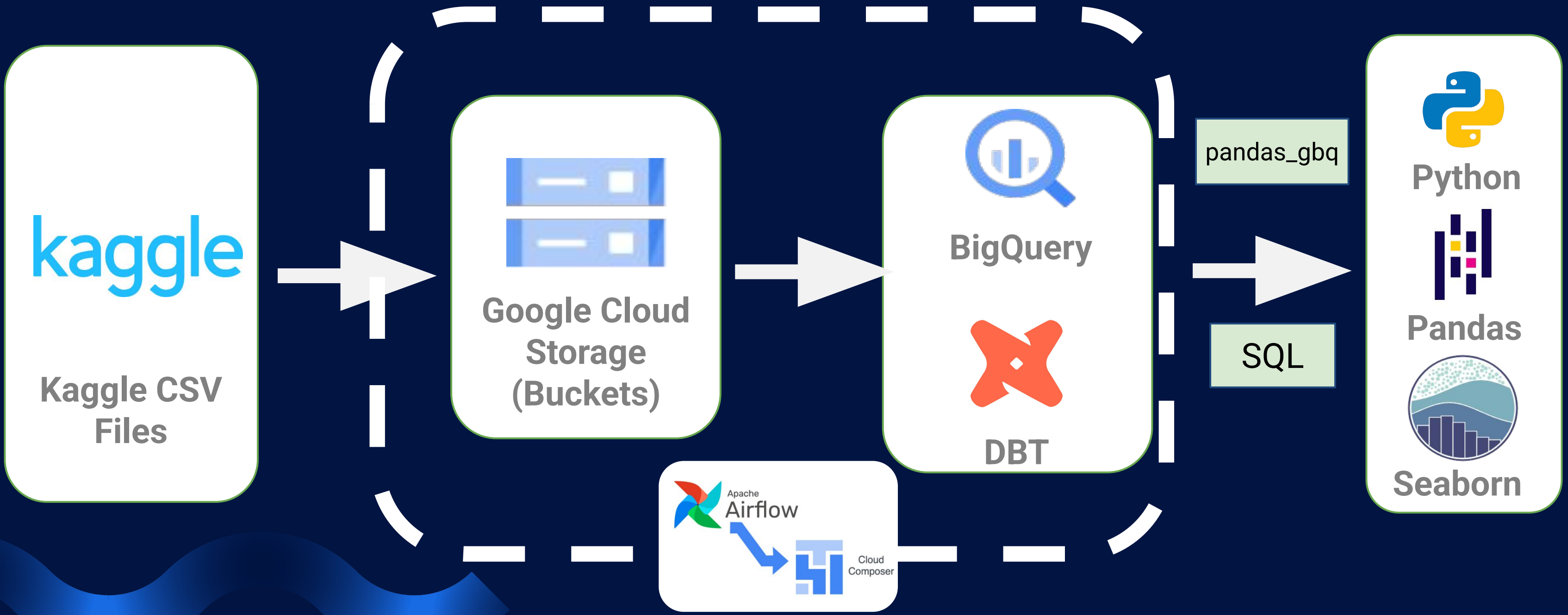
Pandas

Seaborn

olist Data Engineering Lab

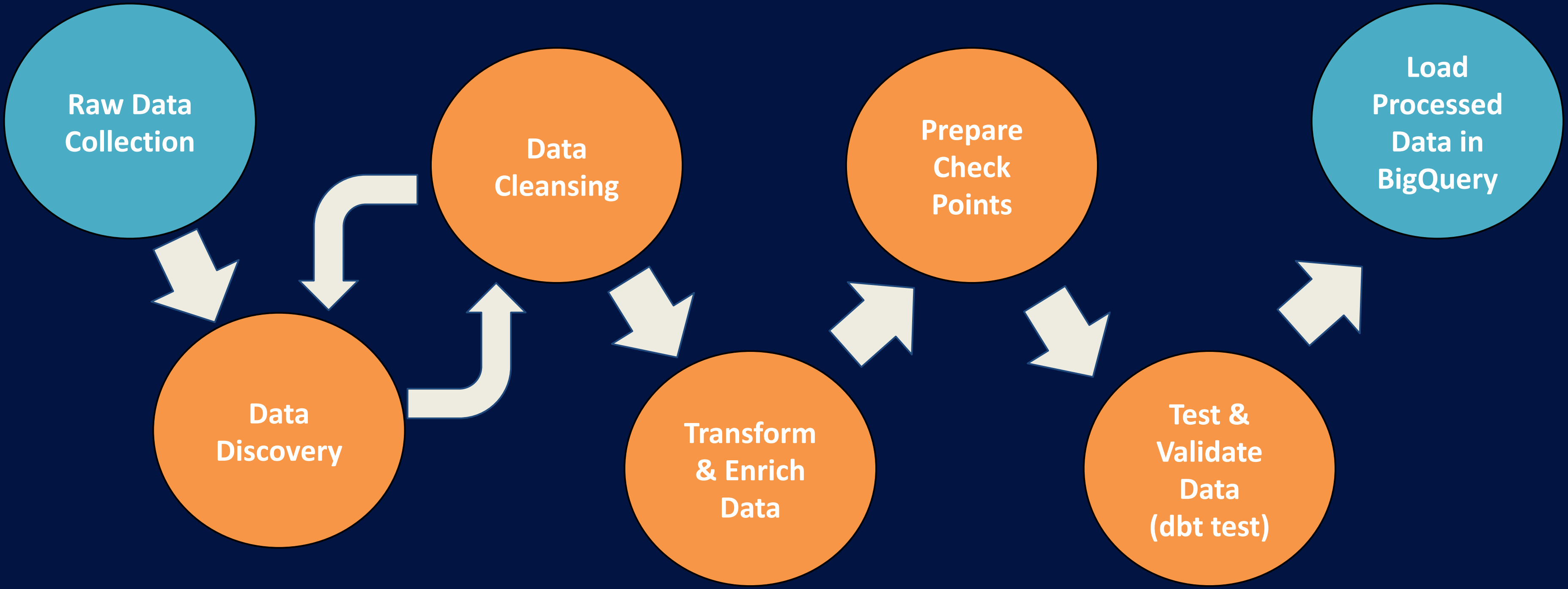
Platform Data Pipeline Architecture v 2.0

Data Source | **Data Storage** | **Data Staging & Preparation** | **BI & Analytics**



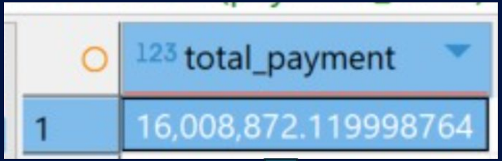
olist Data Engineering Lab

Data Preparation

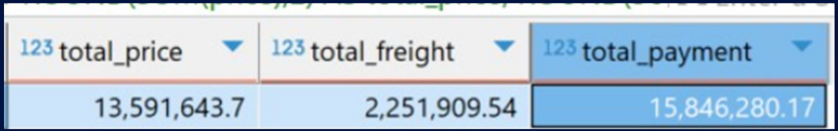


Test Case:

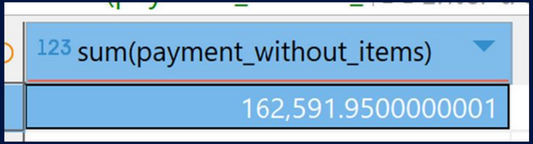
Does the delivered order value match our payment receipts?



	total_payment
1	16,008,872.119998764

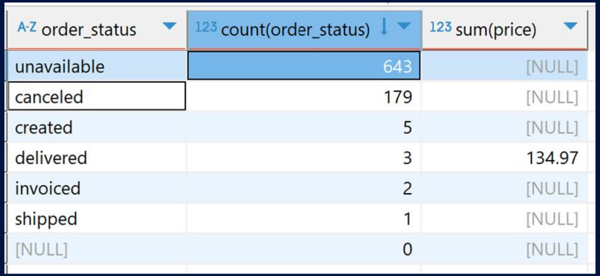


total_price	total_freight	total_payment
13,591,643.7	2,251,909.54	15,846,280.17



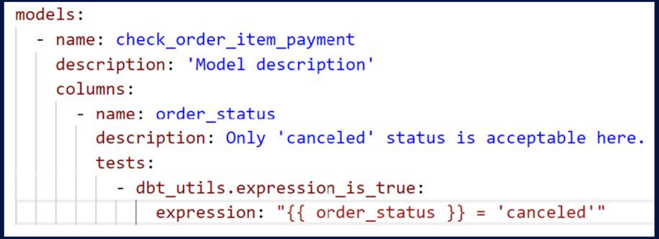
sum(payment_without_items)
162,591.9500000001

```
rows: 766 orders were paid without order items
```



order_status	count(order_status)	sum(price)
unavailable	643	[NULL]
canceled	179	[NULL]
created	5	[NULL]
delivered	3	134.97
invoiced	2	[NULL]
shipped	1	[NULL]
[NULL]	0	[NULL]

Unavailable?
Need to find out if link to deliveries.



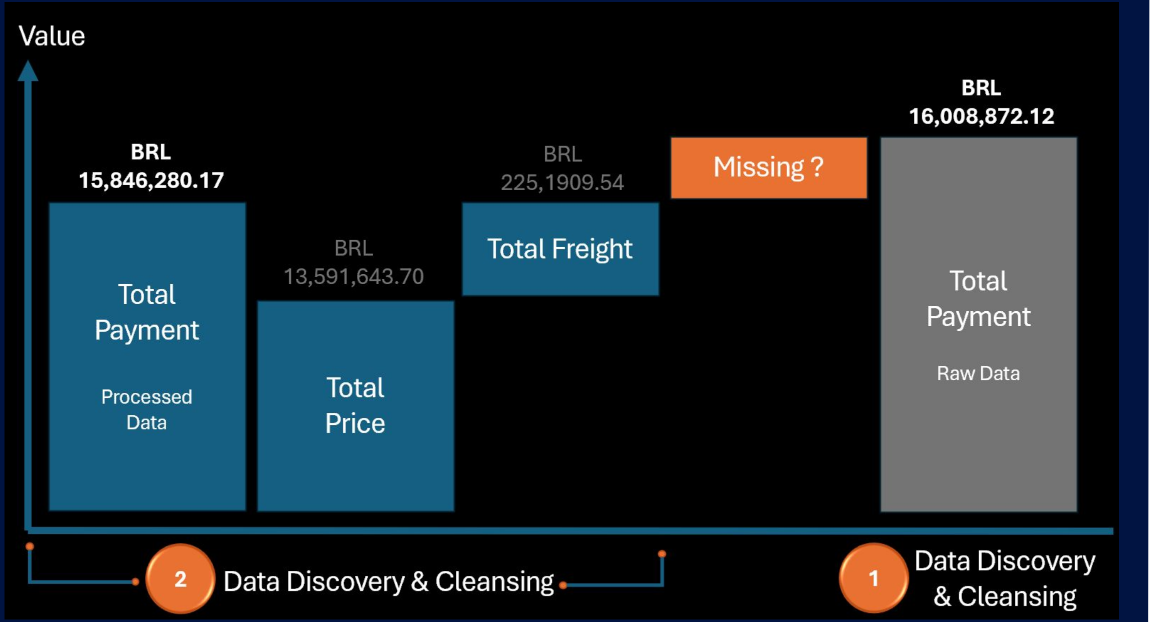
```
models:
  - name: check_order_item_payment
    description: 'Model description'
    columns:
      - name: order_status
        description: Only 'canceled' status is acceptable here.
        tests:
          - dbt_utils.expression_is_true:
              expression: "{{ order_status }} = 'canceled'"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

Here is the content for your review first, I want to add nicer snapshots to illustrate the process.

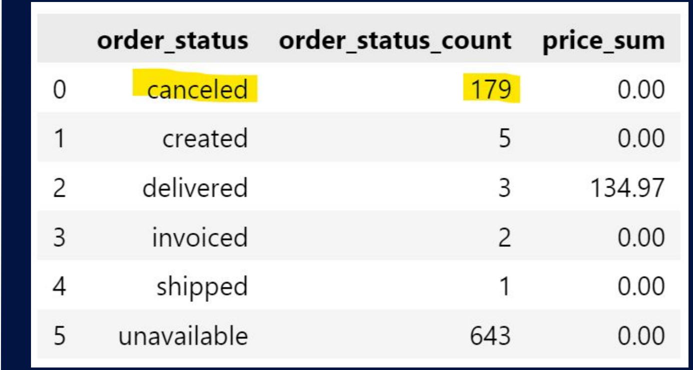
olist Data Engineering Lab

Test Case:

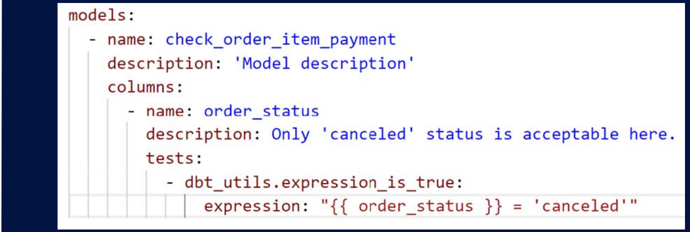
Does the delivered order value match our payment receipts?



New Query : Checking Missing Data



	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



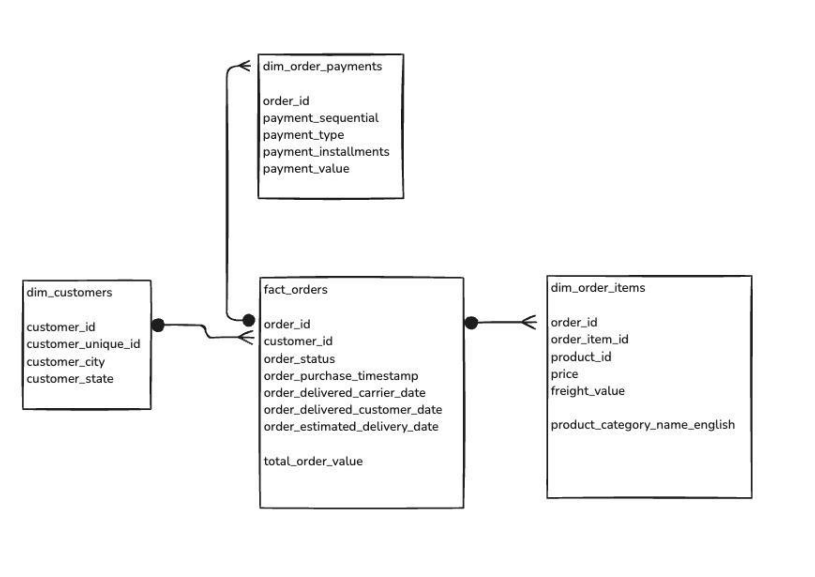
```
models:
  - name: check_order_item_payment
    description: 'Model description'
    columns:
      - name: order_status
        description: Only 'canceled' status is acceptable here.
        tests:
          - dbt_utils.expression_is_true:
              expression: "{{ order_status }} = 'canceled'"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

Test Case 1 working file link

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).



dim_order_payments

order_id
payment_sequential
payment_type
payment_installments
payment_value

dim_customers

customer_id
customer_unique_id
customer_city
customer_state

fact_orders

order_id
customer_id
order_status
order_purchase_timestamp
order_delivered_carrier_date
order_delivered_customer_date
order_estimated_delivery_date

total_order_value

dim_order_items

order_id
order_item_id
product_id
price
freight_value

product_category_name_english

olist Data Engineering Lab

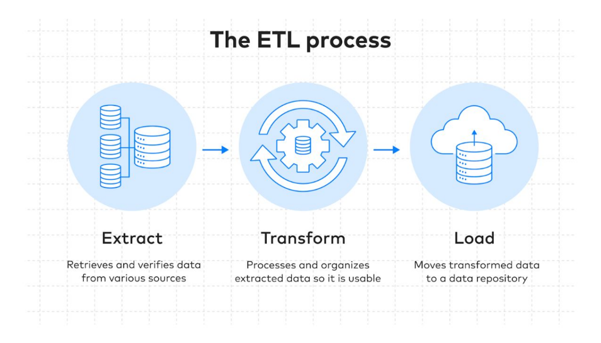
ELT Pipeline Features

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

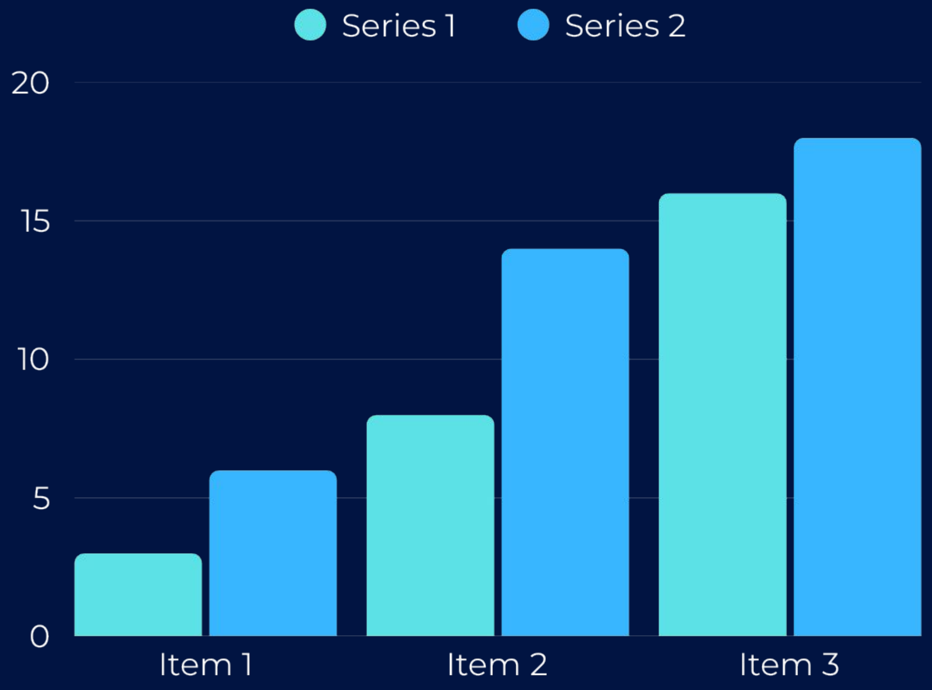
3. **Creation of Derived Columns:**
   - Developed additional metrics like **total_order_amount** within dbt models to enrich the dataset and provide deeper analytical insights.

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.

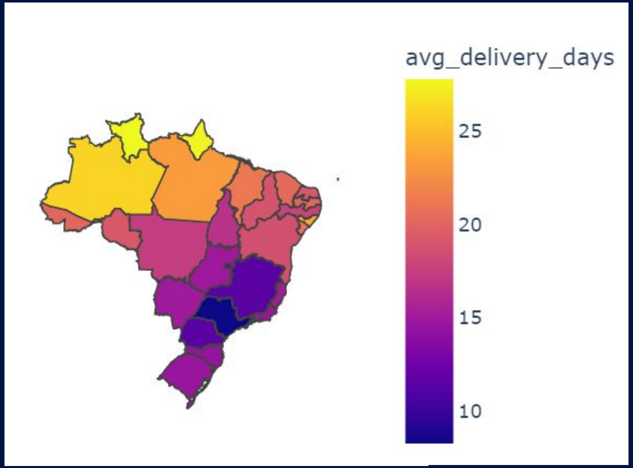

The ETL process
Extract
Retrieves and verifies data from various sources
Transform
Processes and organizes extracted data so it is usable
Load
Moves transformed data to a data repository

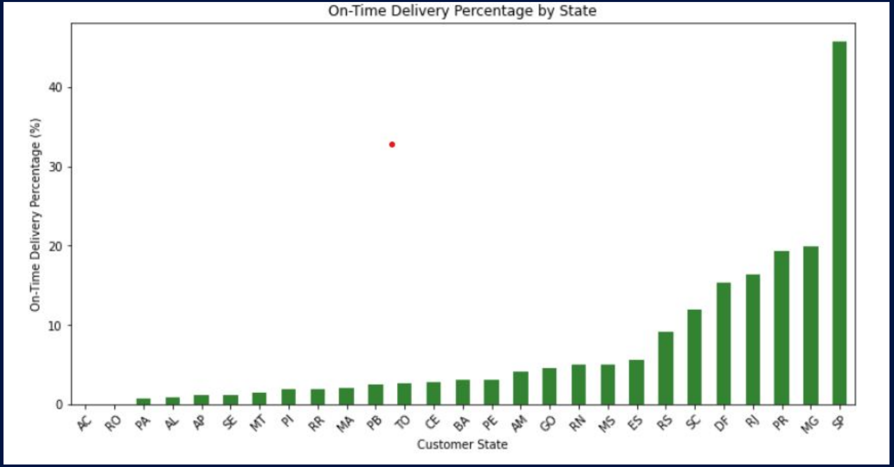
olist Data Engineering Lab

Data Analysis Insights

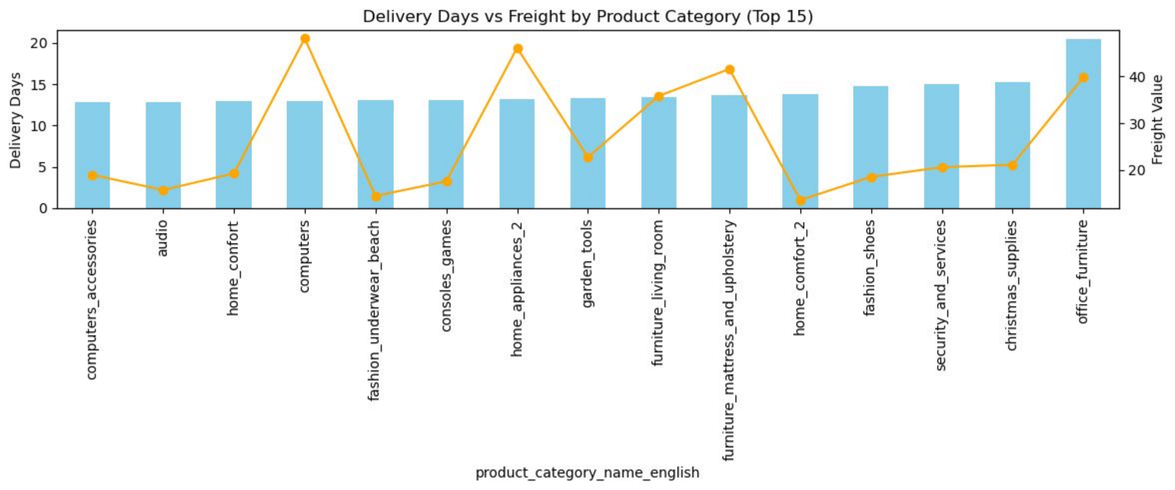


1) Northern States in Brazil had the highest average delivery dates

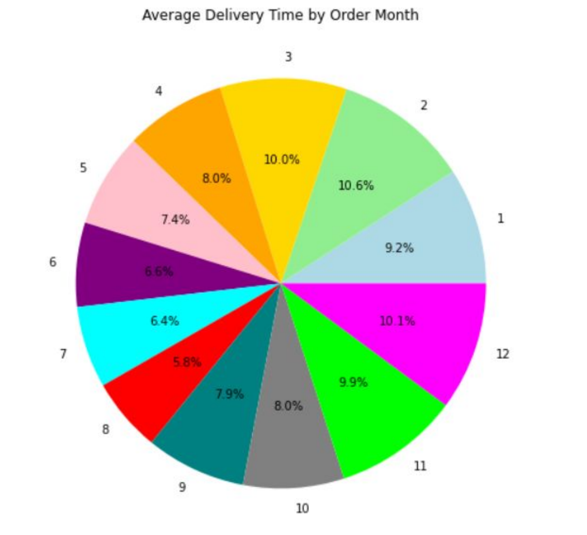


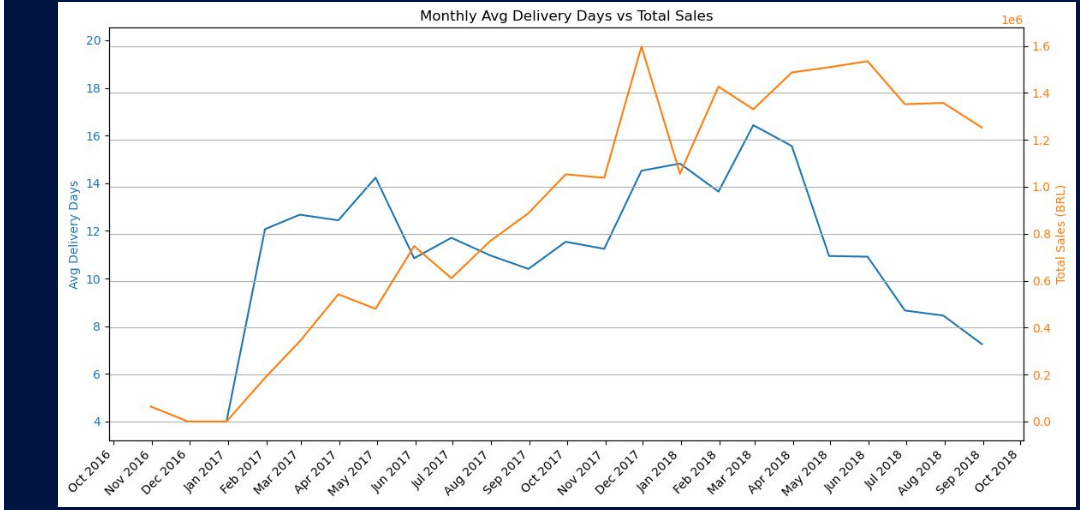


2) Bulkier Items show a stronger correlation to higher delivery times

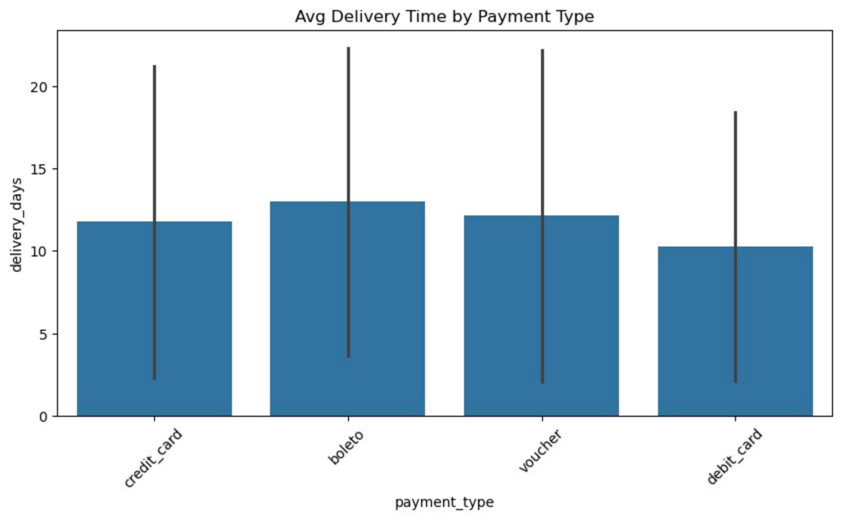


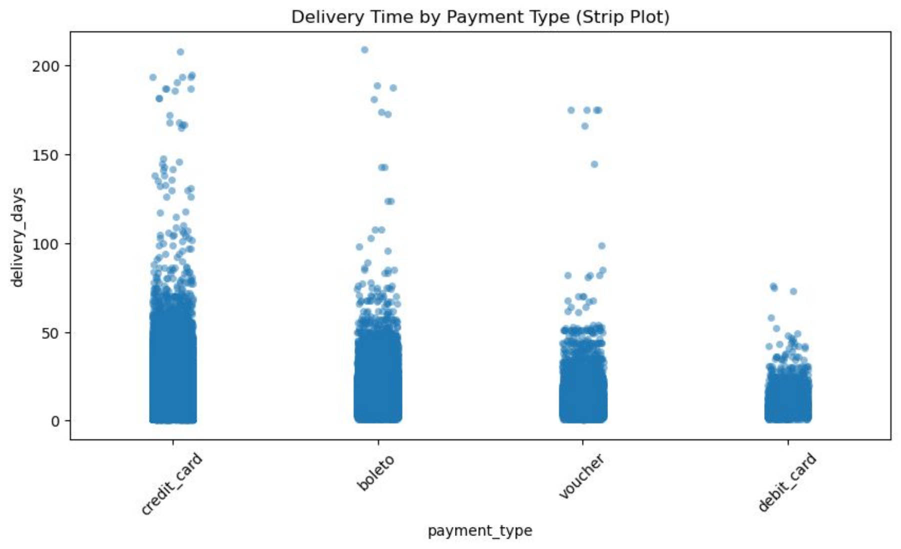
3) Delivery times peak in February, March, November, and December due to seasonal demand, holidays





4) Boleto & Credit Card Transactions have higher delivery times





Pipeline Orchestration - Google Cloud Composer

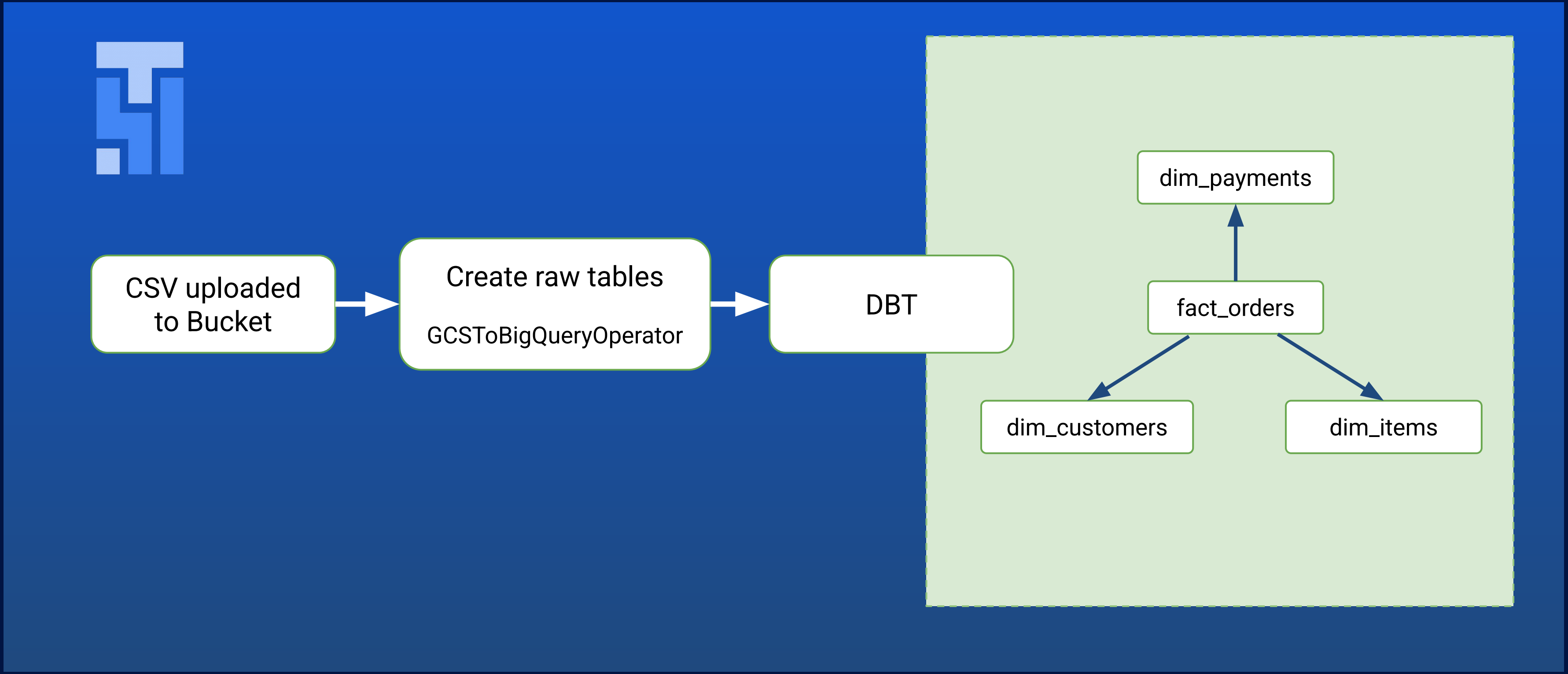
Managed service, using Apache Airflow.

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.

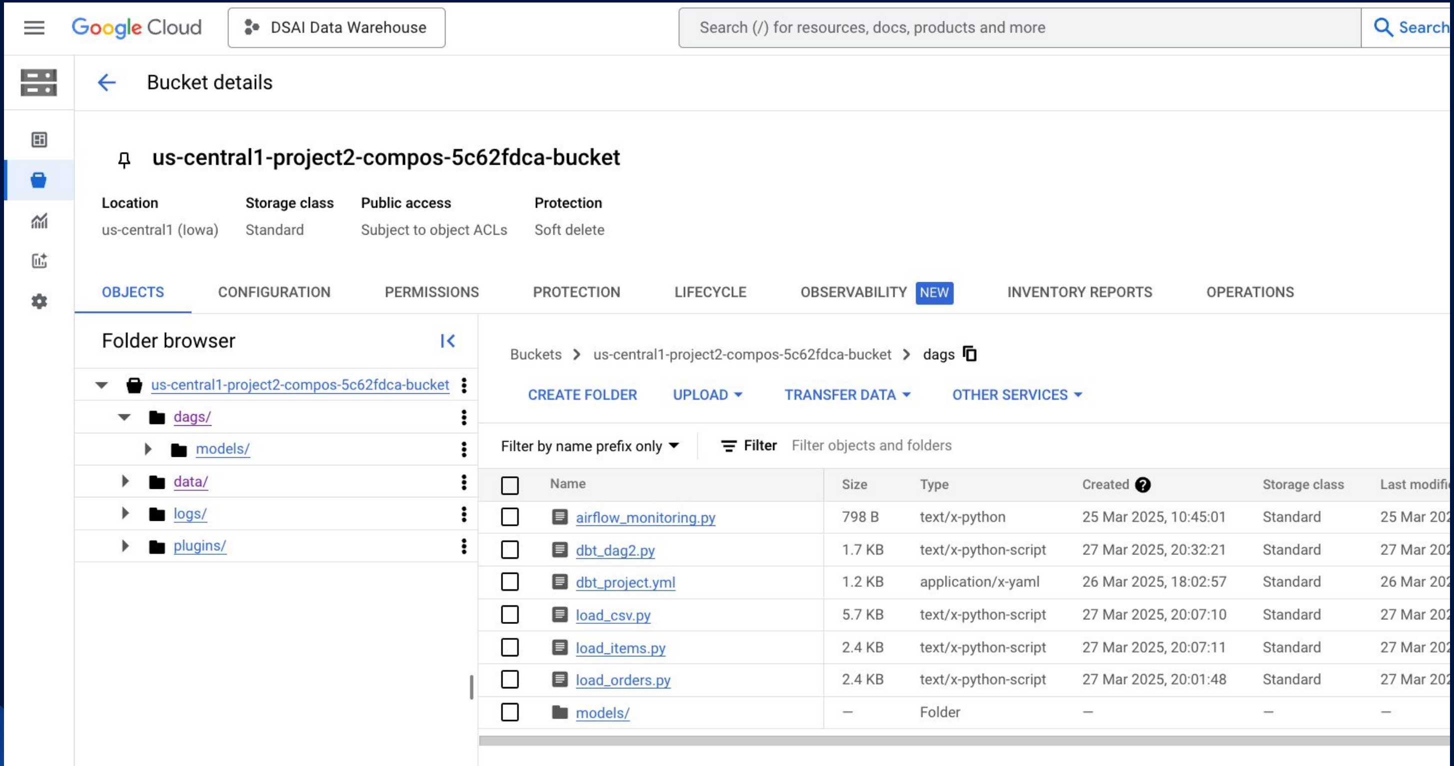
olist Data Engineer Lab

Data Pipeline in Google Cloud Composer



Data Engineer Lab

DAGs folder in Bucket



us-central1-project2-compos-5c62fdca-bucket

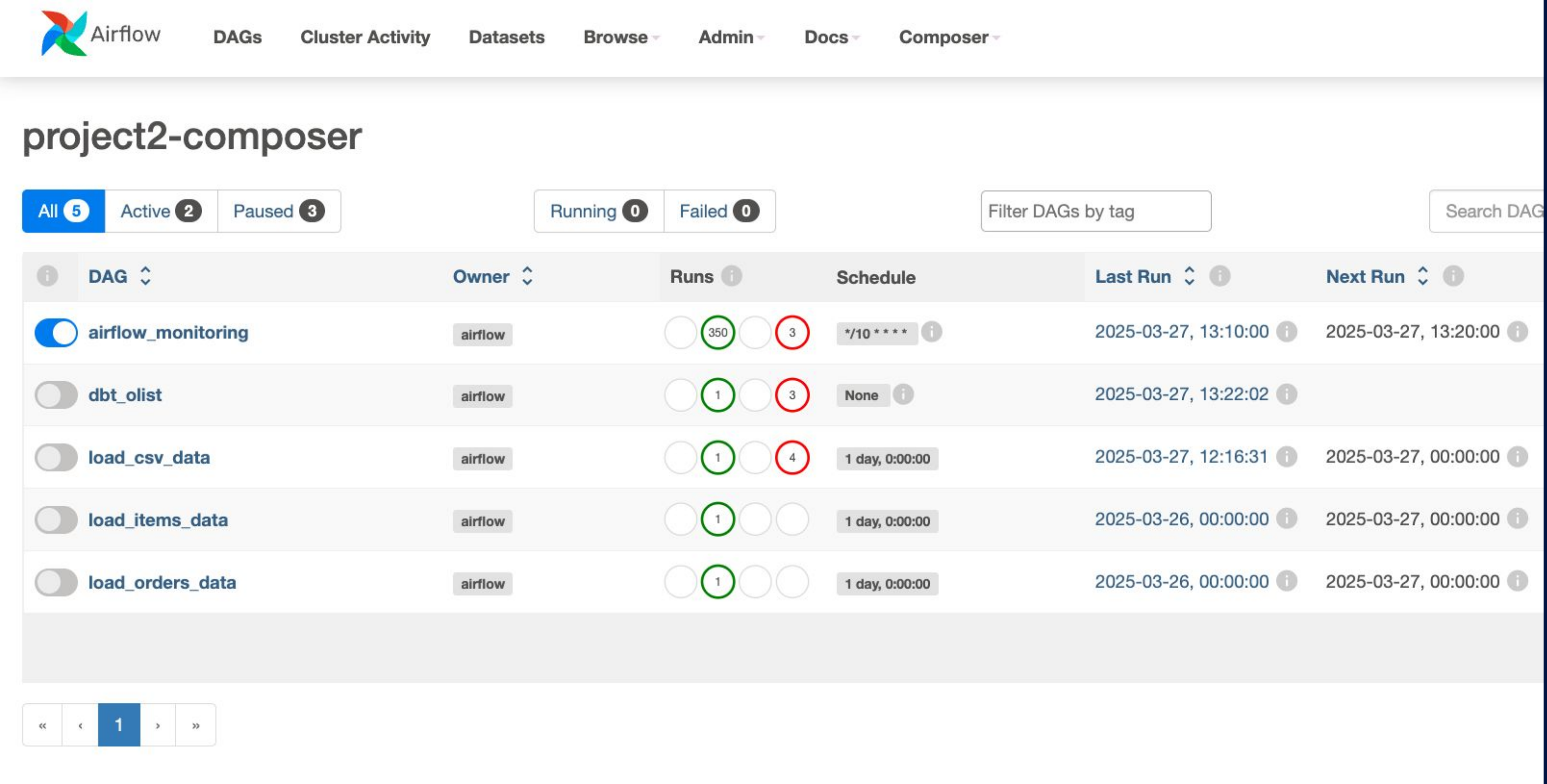
Location	Storage class	Public access	Protection
us-central1 (Iowa)	Standard	Subject to object ACLs	Soft delete

- us-central1-project2-compos-5c62fdca-bucket
  - dags/
    - models/
  - data/
  - logs/
  - plugins/

Buckets > us-central1-project2-compos-5c62fdca-bucket > dags

□	Name	Size	Type	Created	Storage class	Last modifi
□	airflow_monitoring.py	798 B	text/x-python	25 Mar 2025, 10:45:01	Standard	25 Mar 202
□	dbt_dag2.py	1.7 KB	text/x-python-script	27 Mar 2025, 20:32:21	Standard	27 Mar 202
□	dbt_project.yml	1.2 KB	application/x-yaml	26 Mar 2025, 18:02:57	Standard	26 Mar 202
□	load_csv.py	5.7 KB	text/x-python-script	27 Mar 2025, 20:07:10	Standard	27 Mar 202
□	load_items.py	2.4 KB	text/x-python-script	27 Mar 2025, 20:07:11	Standard	27 Mar 202
□	load_orders.py	2.4 KB	text/x-python-script	27 Mar 2025, 20:01:48	Standard	27 Mar 202
□	models/	—	Folder	—	—	—

Airflow DAGs in Google Cloud Composer

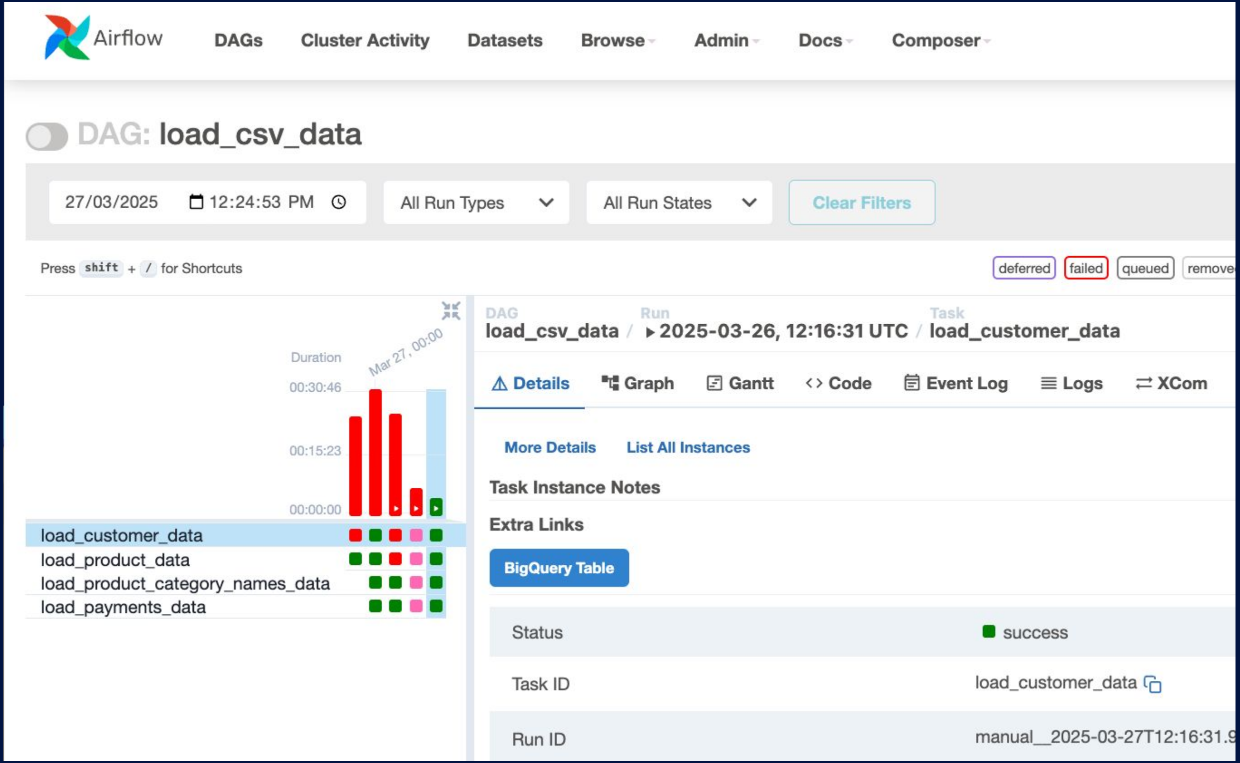


project2-composer

	DAG	Owner	Runs	Schedule	Last Run	Next Run
●	airflow_monitoring	airflow	○ 350 ○ 3	*/10 * * * *	2025-03-27, 13:10:00	2025-03-27, 13:20:00
○	dbt_olist	airflow	○ 1 ○ 3	None	2025-03-27, 13:22:02	
○	load_csv_data	airflow	○ 1 ○ 4	1 day, 0:00:00	2025-03-27, 12:16:31	2025-03-27, 00:00:00
○	load_items_data	airflow	○ 1 ○ ○	1 day, 0:00:00	2025-03-26, 00:00:00	2025-03-27, 00:00:00
○	load_orders_data	airflow	○ 1 ○ ○	1 day, 0:00:00	2025-03-26, 00:00:00	2025-03-27, 00:00:00

olist Data Engineer Lab

load_csv_data DAG run logs



DAG: load_csv_data

27/03/2025 12:24:53 PM | All Run Types | All Run States | Clear Filters

load_customer_data 🟥 🟩 🟥 🟪 🟩
load_product_data 🟩 🟩 🟥 🟪 🟩
load_product_category_names_data 🟩 🟩 🟪 🟩
load_payments_data 🟩 🟩 🟪 🟩

DAG load_csv_data / Run ▸ 2025-03-26, 12:16:31 UTC / Task load_customer_data

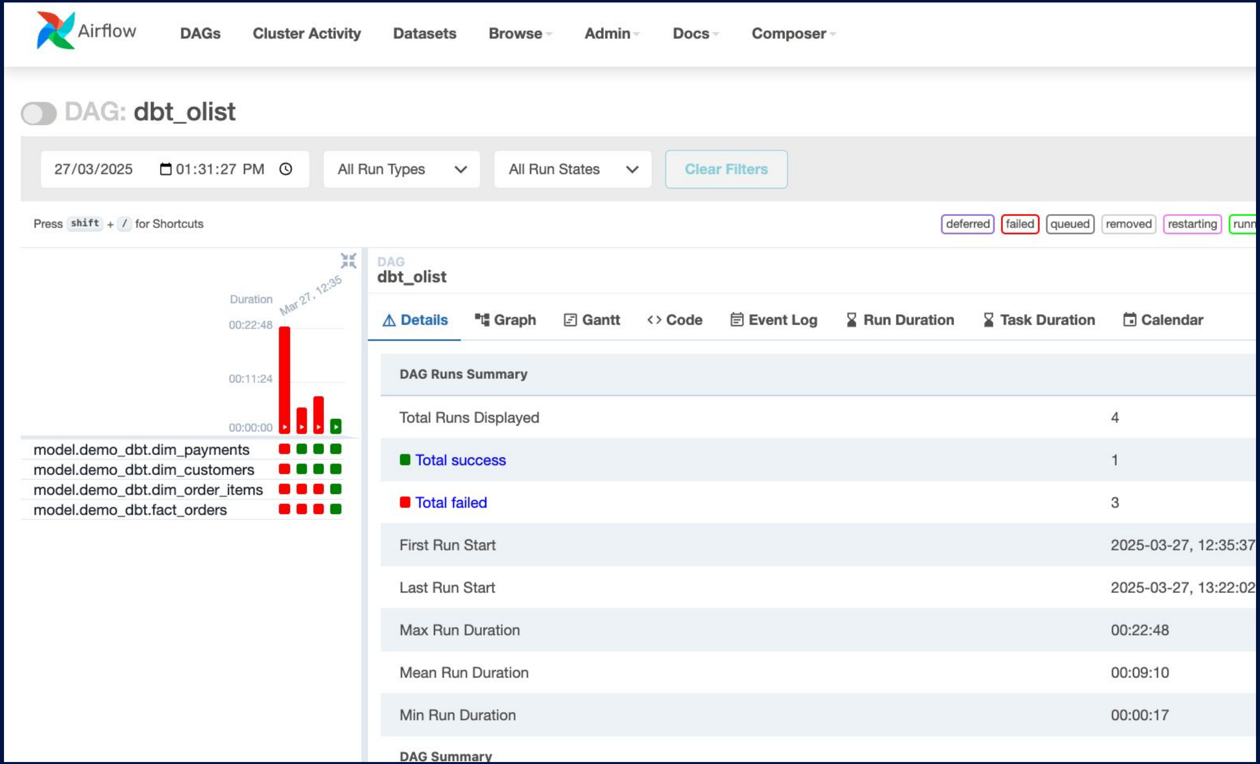
Task Instance Notes

Extra Links

Status	🟩 success
Task ID	load_customer_data
Run ID	manual__2025-03-27T12:16:31.9

olist Data Engineer Lab

DBT running as DAG



DAG: dbt_olist

27/03/2025 01:31:27 PM | All Run Types | All Run States

Press shift + / for Shortcuts

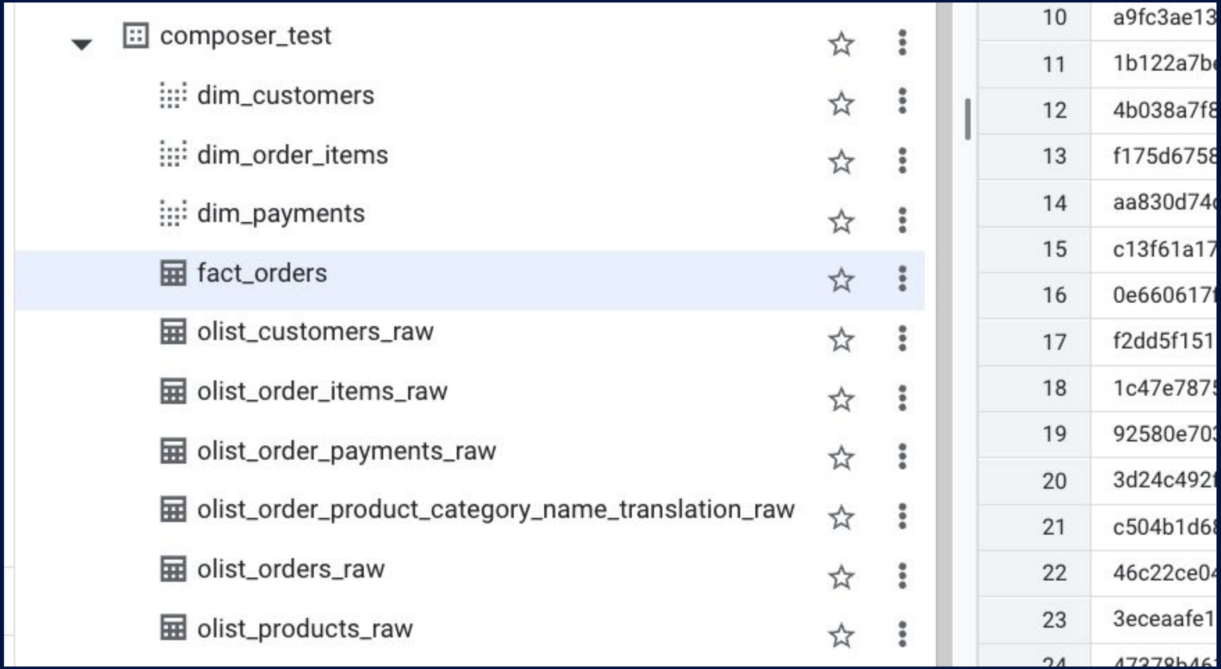
model.demo_dbt.dim_payments 🟥 🟩 🟩 🟩
model.demo_dbt.dim_customers 🟥 🟩 🟩 🟩
model.demo_dbt.dim_order_items 🟥 🟥 🟥 🟩
model.demo_dbt.fact_orders 🟥 🟥 🟥 🟩

DAG
dbt_olist

DAG Runs Summary	
Total Runs Displayed	4
🟩 Total success	1
🟥 Total failed	3
First Run Start	2025-03-27, 12:35:37
Last Run Start	2025-03-27, 13:22:02
Max Run Duration	00:22:48
Mean Run Duration	00:09:10
Min Run Duration	00:00:17

DAG Summary

Tables and Views in BigQuery



composer_test

- dim_customers
- dim_order_items
- dim_payments
- fact_orders
- olist_customers_raw
- olist_order_items_raw
- olist_order_payments_raw
- olist_order_product_category_name_translation_raw
- olist_orders_raw
- olist_products_raw

10	a9fc3ae13
11	1b122a7b
12	4b038a7f
13	f175d675
14	aa830d74
15	c13f61a17
16	0e660617
17	f2dd5f151
18	1c47e787
19	92580e70
20	3d24c492
21	c504b1d6
22	46c22ce0
23	3eceaafe1
24	47378b46

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