Development of an ELT Tool for Data Analysis of Kaggle Data

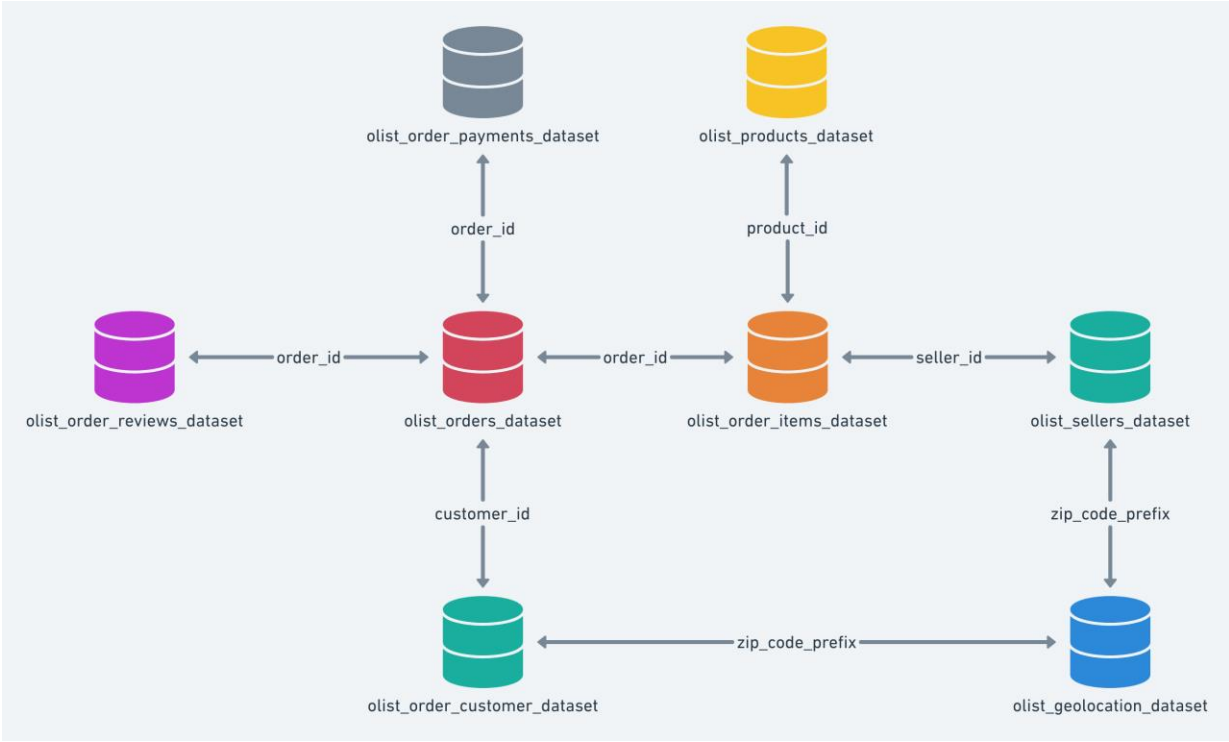
Using Python, PostgreSQL, Docker, Docker Compose, Airflow, DBT and BigQuery to Analyze Brazilian E-Commerce Data

August, 2024

Ifeanyi Franklin Ike

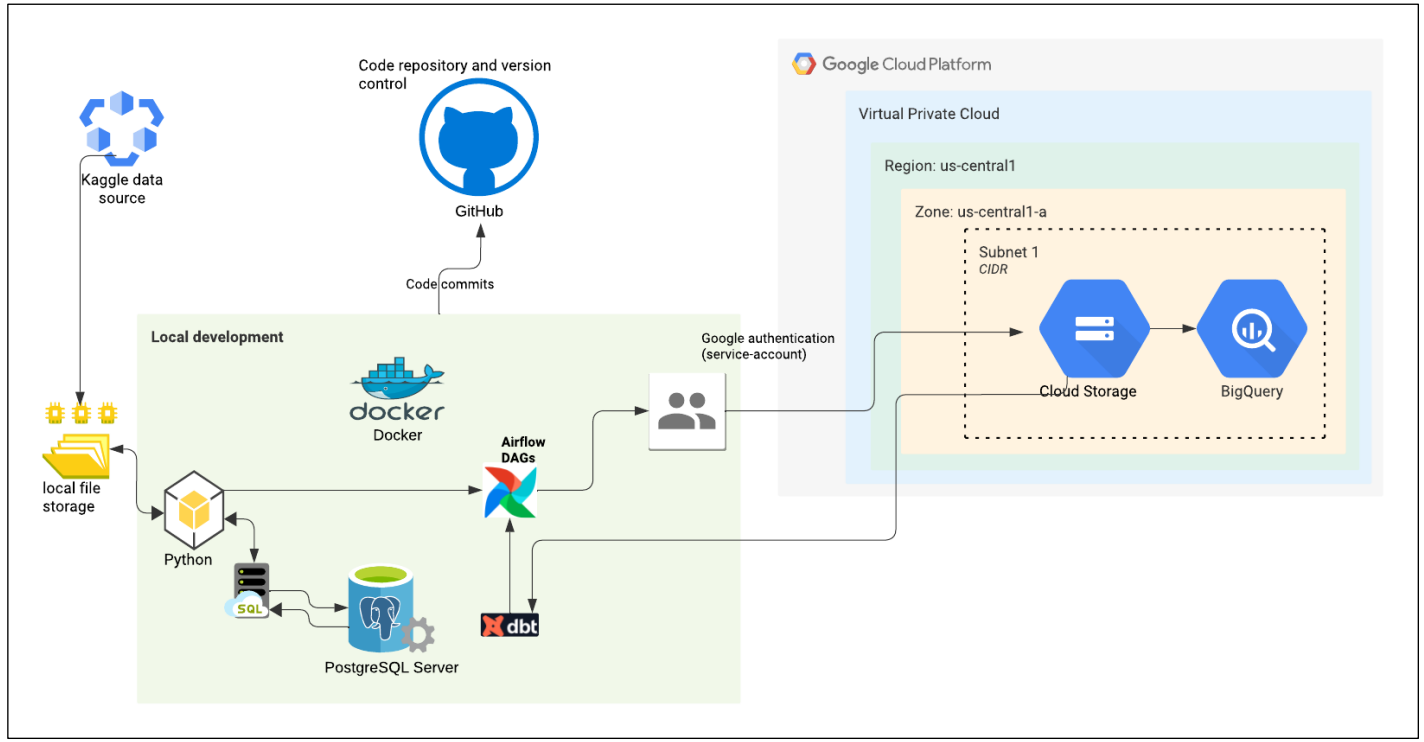
Project Overview

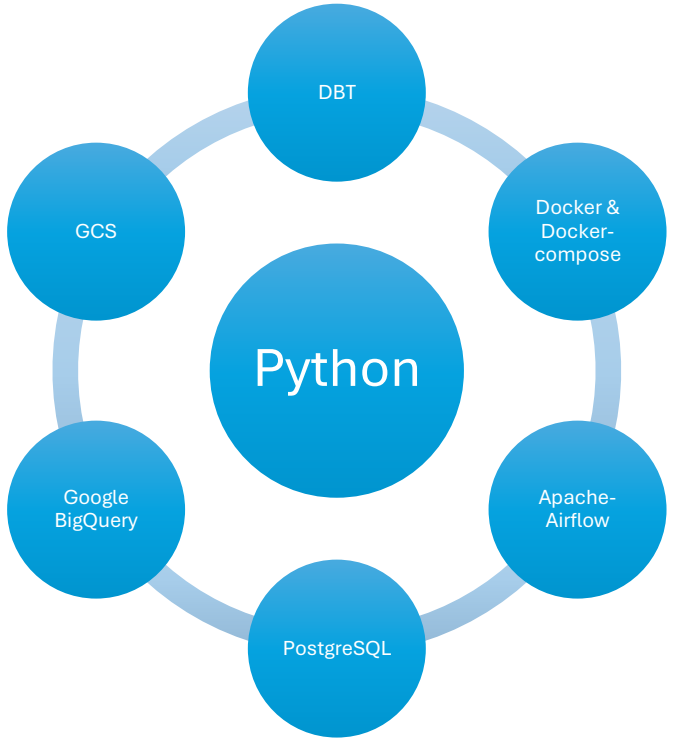
- Brazilian E-Commerce dataset is a public dataset of orders from Olist Store at multiple marketplaces in Brazil.
  - With 100,000 orders
  - From 2016 – 2018
  - It contains the following datasets (csv) used in this project:
    - Orders, Customers, Sellers, Products, Geolocation, Order payments, Order reviews and Order items



- This project is therefore set up to:
  - Develop an ELT tool for handling the csv data from Kaggle
  - Extract insights and answer key business questions regarding the Brazilian c-commerce data.

System Architecture & Tool Stack





Research Questions

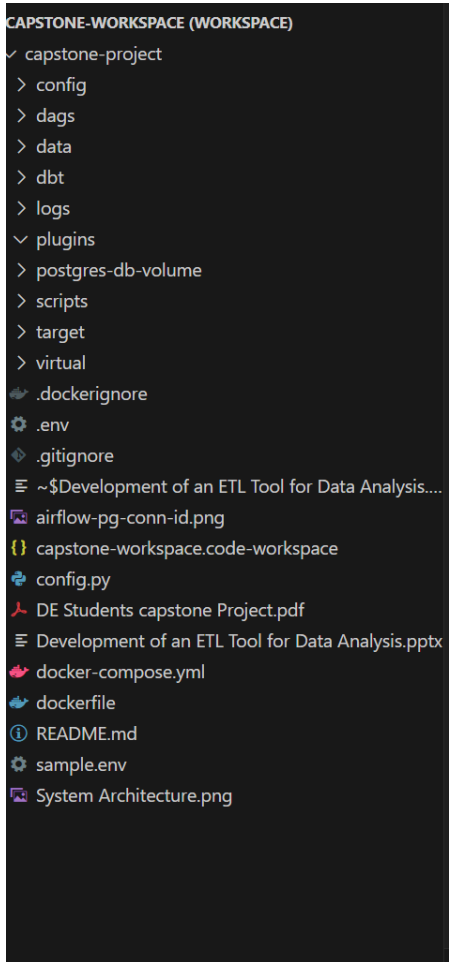
- Which product categories have the highest sales?
- What is the average delivery time for orders?
- Which states have the highest number of orders?

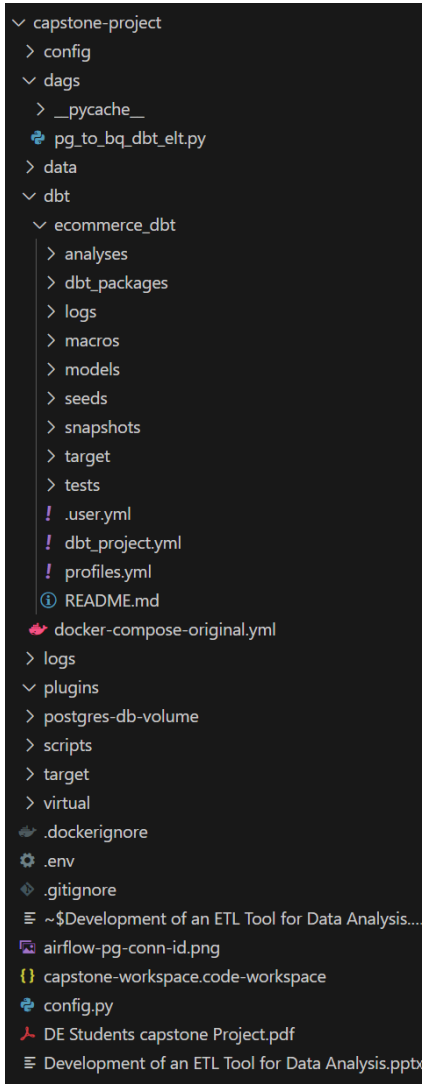
Methodology

- **Approach:**
  - Local file ➡ PostgreSQL ➡ GCS ➡ BigQuery ⇔ DBT
- **Data Pipeline:**
  - **Ingestion**: Raw data ingested into PostgreSQL using SQLAlchemy
  - **Loading/storage to cloud**: Data loaded to GCS and BigQuery from PostgreSQL.
  - **Transformation**: Data modeled using DBT in BigQuery.
  - **Analysis**: Queries run and saved on the transformed data in BigQuery.

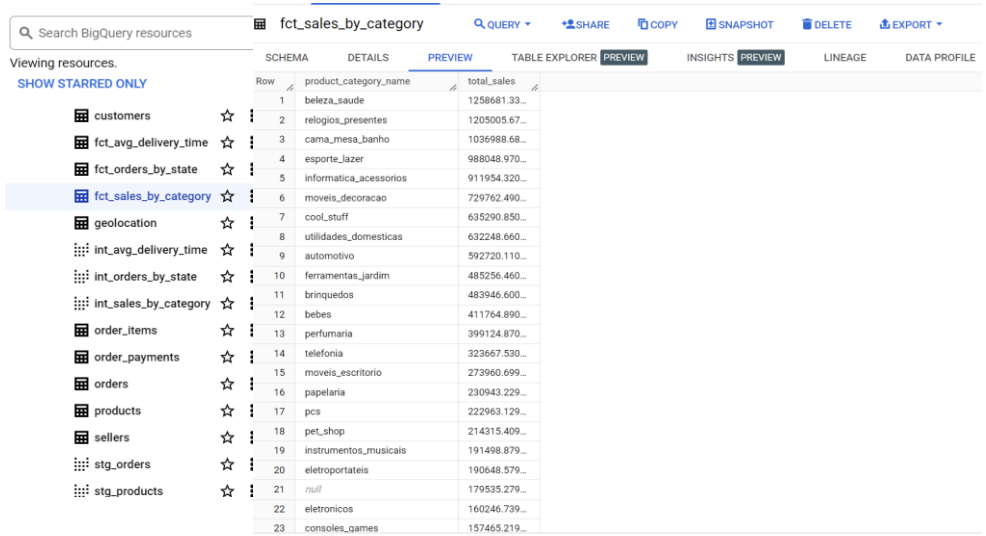
Snapshots of Code and Results

Directory management for the project

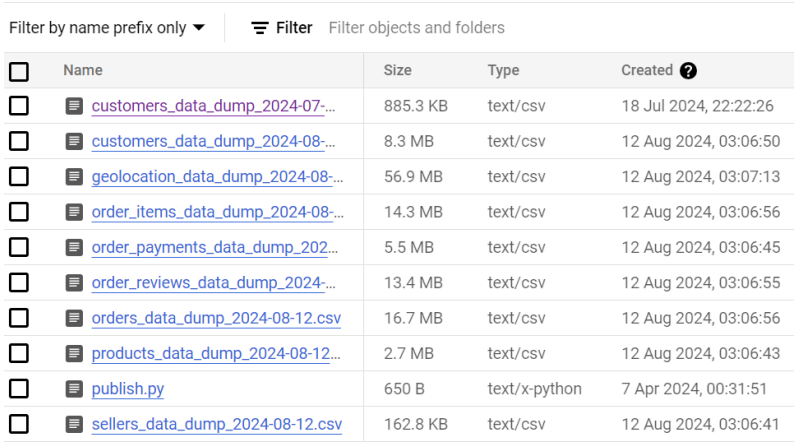




Data stored in BigQuery

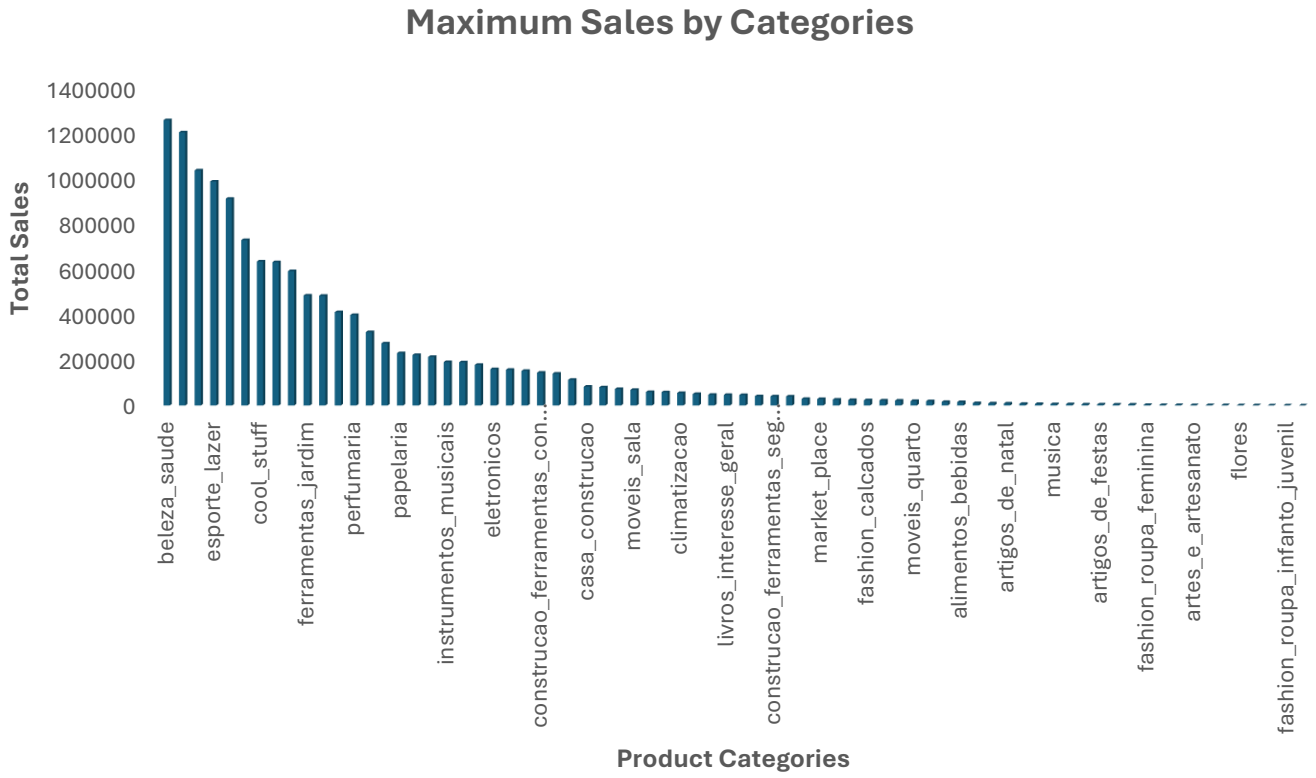


Data stored in GC Bucket



Answer to Question 1 - **Highest Sales by Product Category**

- **Key Insight:**
  - Product category with highest sales was **Beleza Saude** with **1258681.34** sales
  - Other 3 categories with very high sales :
    - Relogios Presentes (1205005.68 sales)
    - Cama Mesa Banho (1036988.68 sales) and
    - Esporte Lazer (988048.97 sales)



Answer to Question 2 - **Average Delivery Time**

- **Key Insight:**
  - Average delivery time: **301.4 hours**

Top 10 orders with highest delivery time (hours)

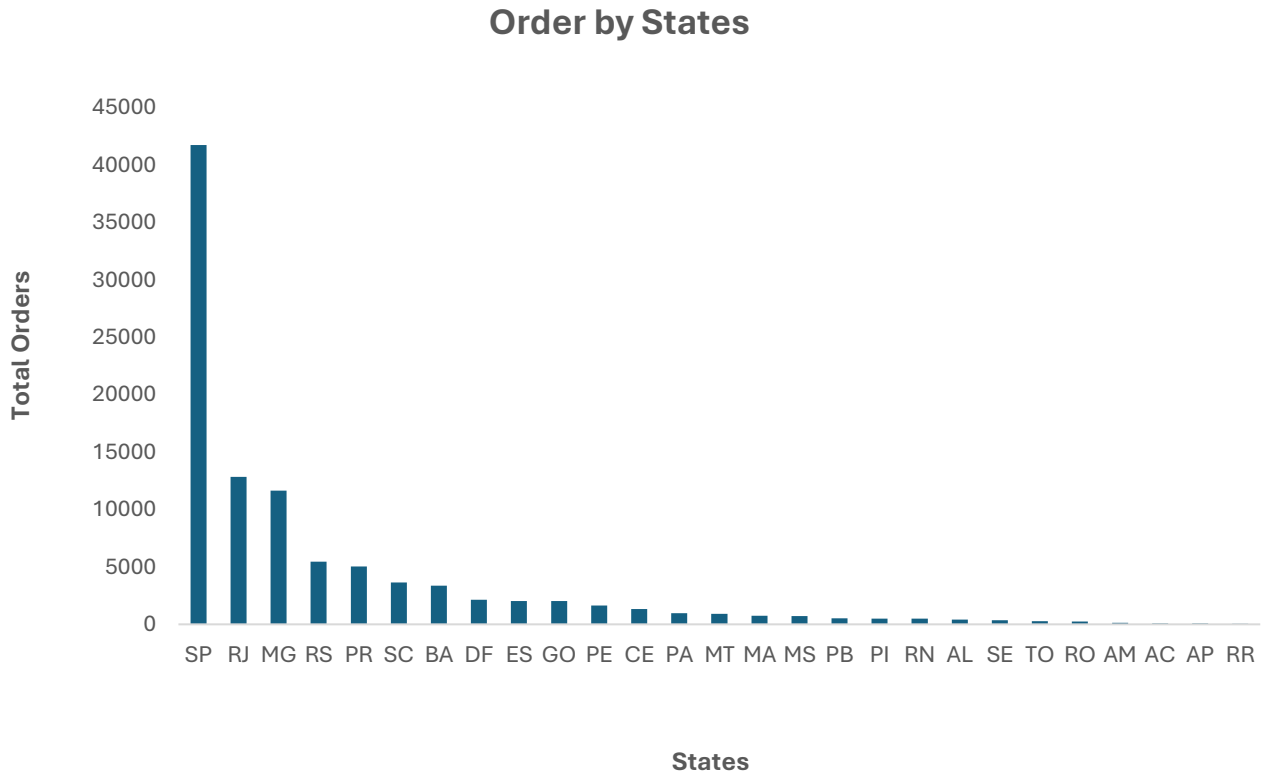
Row	order_id	avg_delivery_hours
1	ca07593549f1816d26a572e06...	5031.09
2	1b3190b2dfa9d789e1f14c05b...	5000.44
3	440d0d17af552815d15a9e41a...	4695.22
4	2fb597c2f772eca01b1f5c561b...	4676.4
5	285ab9426d6982034523a855f...	4671.21
6	0f4519c5f1c541ddec9f21b3bd...	4657.19
7	47b40429ed8cce3aee9199792...	4595.13
8	2fe324febf907e3ea3f2aa9650...	4556.72
9	2d7561026d542c8dbd8f0daea...	4515.23
10	c27815f7e3dd0b926b5855262...	4505.85

Top 10 orders with lowest delivery time (hours)

Row	order_id	avg_delivery_hours
1	1d893dd7ca5f77ebf5f59f0d20...	12.8
2	434cecee7d1a65fc65358a632...	18.75
3	f3c6775ba3d2d9fe2826f93b71...	20.53
4	8339b608be0d84fca9d8da68b...	20.72
5	bb5a519e352b45b714192a02f...	21.38
6	e65f1eeee1f52024ad1dcd034...	21.42
7	21a8ffca665bc7a1087d31751...	22.46
8	d5fbeedc85190ba88580d6f82...	22.52
9	f349cdb62f69c3fae5c4d7d3f3...	23.63
10	38c1e3d4ed6a13cd0cf612d4c...	23.66

Answer to Question 3 - **Orders by State**

- **Key Insight:**
  - The State with the highest order: **SP** with **41,746** orders
  - Top 3 States with the highest orders:
    - SP 41,746
    - RJ 12,852
    - MG 11,635



Conclusion

- This project has developed an end-to-end ELT tool using a combination of Docker, PostgreSQL, Airflow, DBT and GCP.
- The project also simplifies data modelling and querying to answer business intelligence questions.
- The project can therefore be adopted in order datasets by changing the configurations.

THANK YOU

Additional Resources

- Project repository:
  - https://github.com/ififrank2013/ELT-Tool-for-Data-Analysis-Capstone-Project
- Learn more about dbt
  - https://docs.getdbt.com/docs/introduction
- Check out [Discourse]
  - https://discourse.getdbt.com/
  - for commonly asked questions and answers