DSAI Module 2 Group 2 Project

Brazilian E-commerce dataset

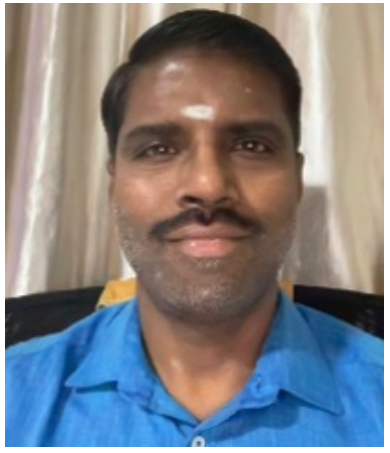
20 Dec 2025

Meet the Team

•Divya Maurya

•Ivan Han

•Wan Huan



•Govindan Dhanasekaran

•Kuppuswamy Mahalakshmi

Executive Summary

This project built a scalable, cloud-based analytics solution transforming raw e-commerce data into executive-ready insights. Using BigQuery, dbt, and Python, it revealed revenue trends, logistics impact on customer satisfaction, regional performance, and customer behavior, enabling faster, data-driven decisions across sales, operations, and customer experience.

Problem:

- Data scattered in raw CSVs, no central structure.
- Hard to track revenue, top products/regions/customers.
- Operational impact on satisfaction unclear; manual reporting is slow and error-prone.

Proposed Solution:

- Automated ingestion into BigQuery and star-schema warehouse.
- ELT pipeline with dbt and data quality checks.
- Python & Streamlit dashboards for analytics and executive insights.

Executive Summary

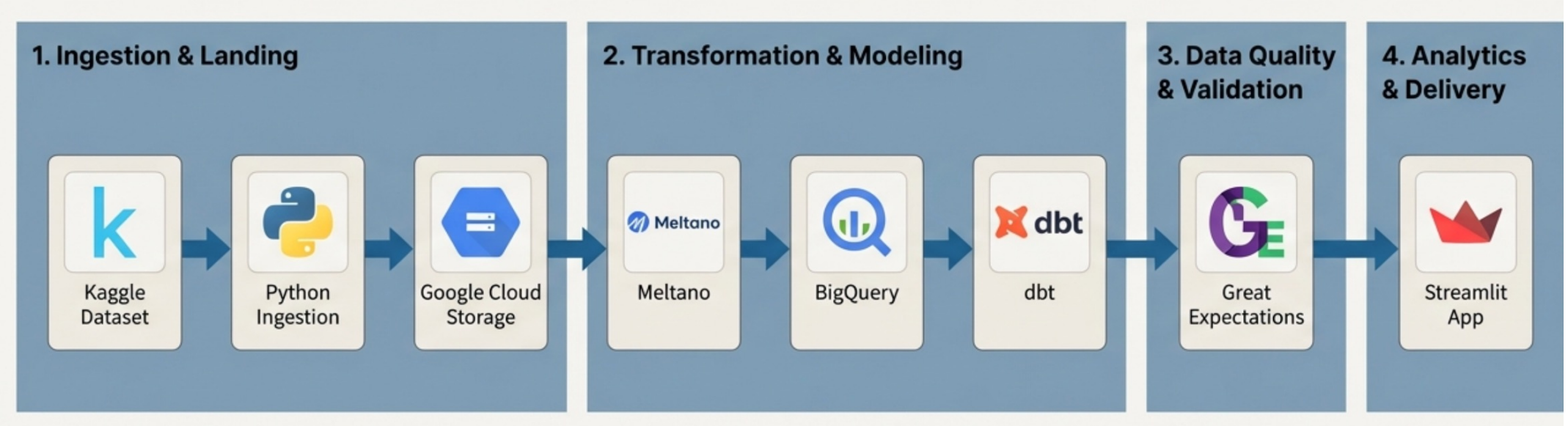
High-Level Results:

- Revenue & operations: Trends, top products/regions, delivery impact on satisfaction.
- Customer & loyalty: Segmentation and insights on retention.
- Reusable, analytics-ready data foundation for future BI use.

Business Impact:

- Reduced manual reporting; faster, confident decision-making.
- Enabled data-driven strategies across sales, operations, and customer experience.
- Improved visibility for executive planning and operational improvements.

Architecture Blueprint

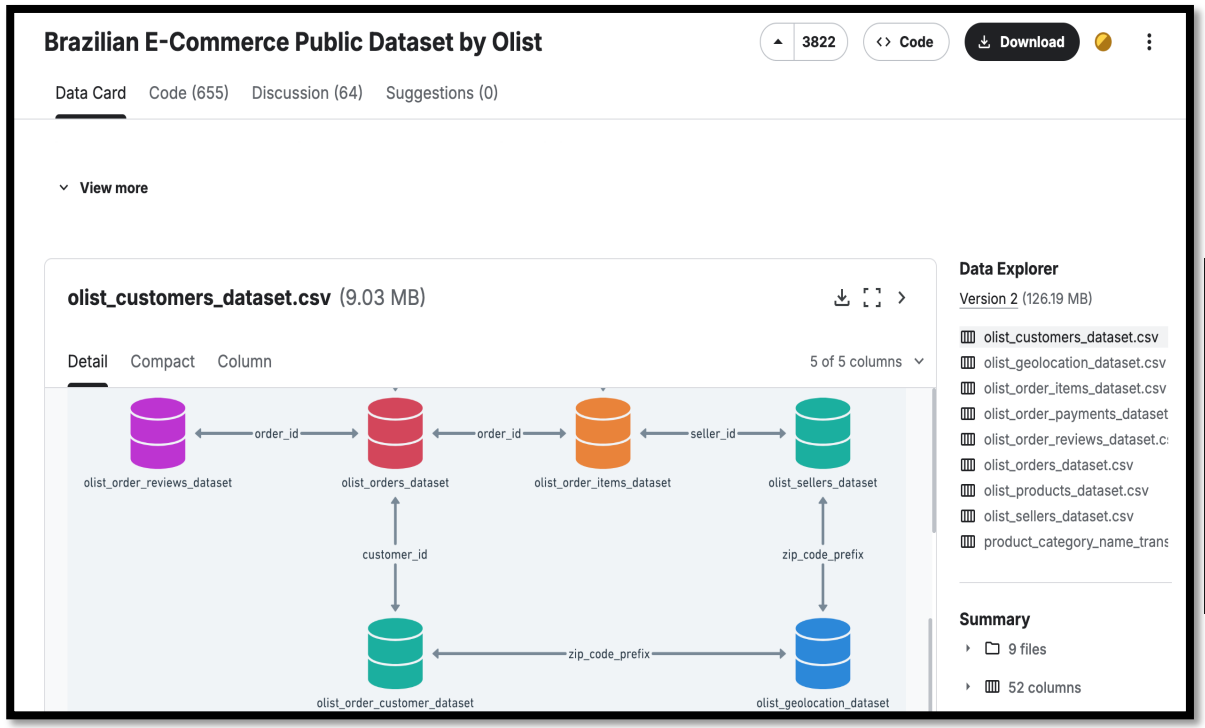


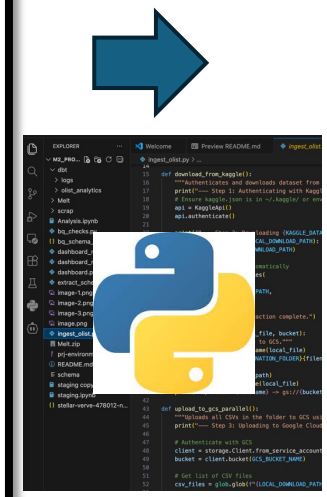
Technology Stack & Justification

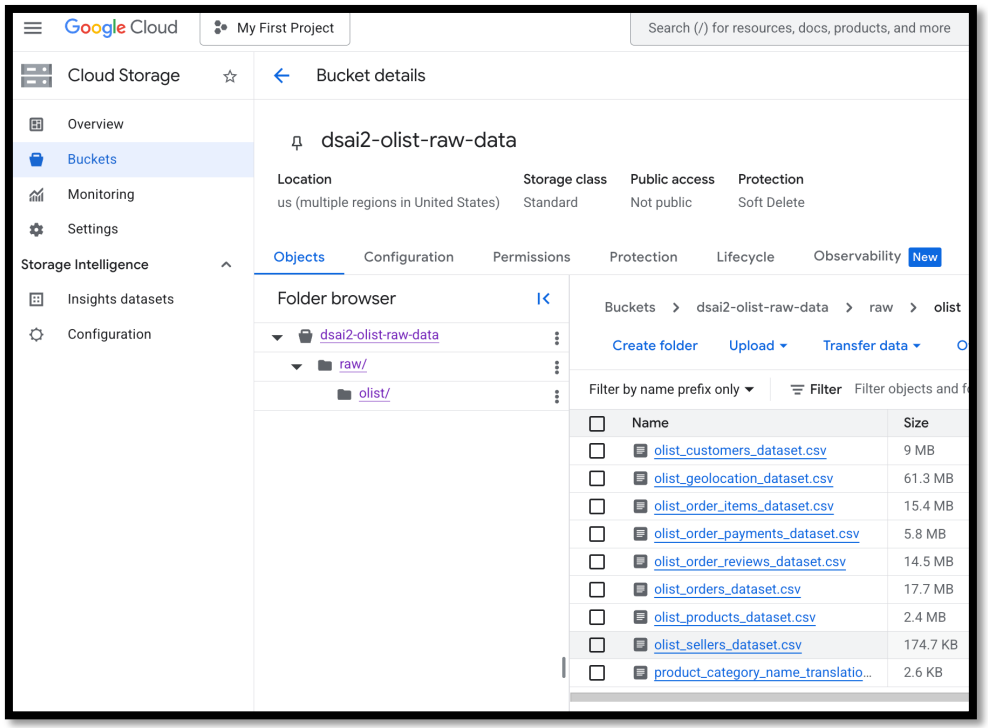
Layer	Tool	Reason for Choice
Ingestion	Python + Kaggle API	Flexible, reproducible, widely used
Storage	Google Cloud Storage	Cheap, scalable object storage
Warehouse	BigQuery	Serverless, scalable, SQL-based analytics
Transformation	Dbt	Version-controlled ELT, testing support
Orchestration	Meltano	Lightweight orchestration and dbt integration
Quality control	Greate Expectation	Automated Validation & Testing
Analytics	Python (pandas)	Exploratory analysis and metrics
Visualization	Streamlit	Rapid interactive dashboards

Data Ingestion and Loading

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

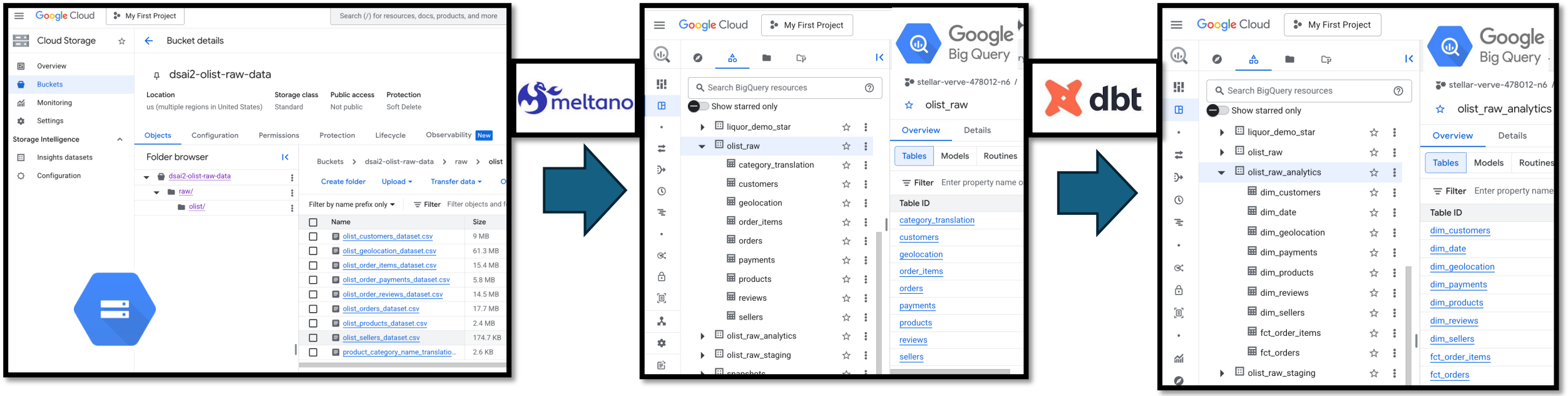






Automated the extraction of the Brazilian E-Commerce dataset from Kaggle using Python scripts and securely loaded the raw files into Google Cloud Storage to serve as our centralized data lake

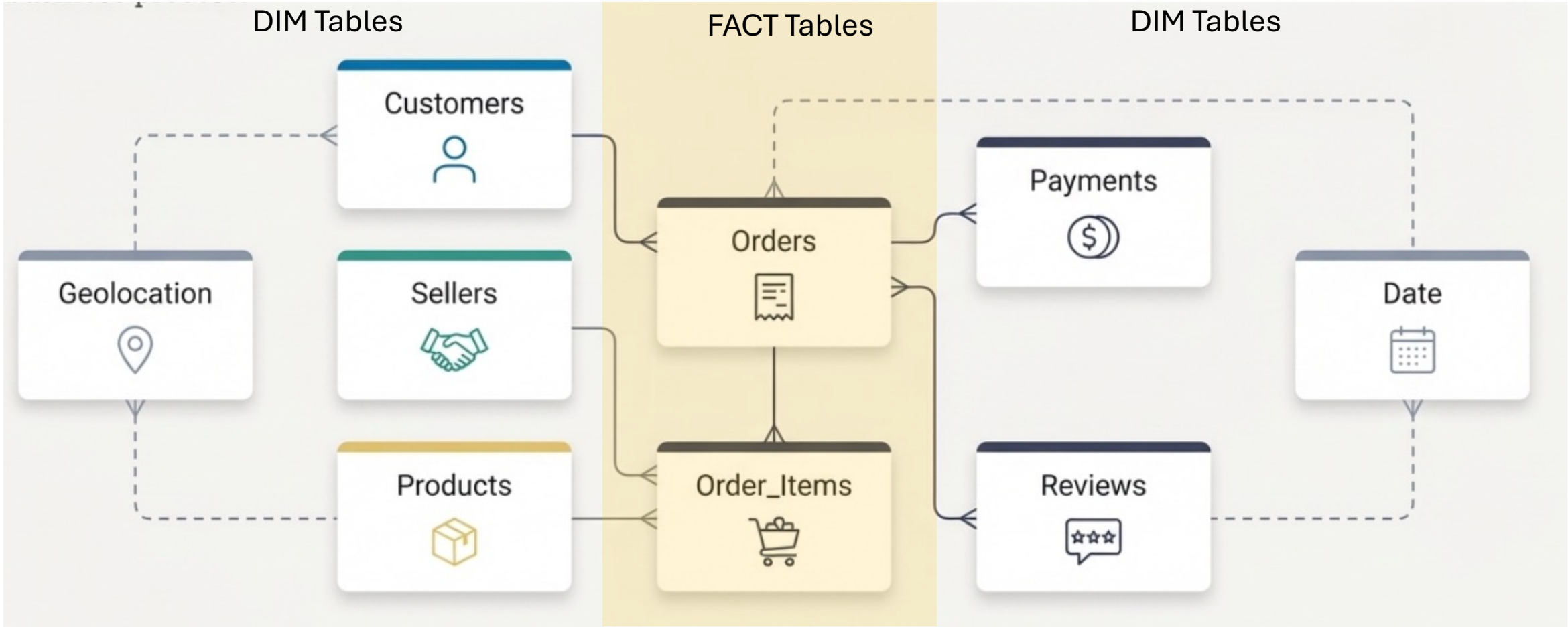
Transformation and Modeling



Utilized Meltano to orchestrate the pipeline, loading raw data directly into Google BigQuery to establish a scalable foundation for our transformation workflows

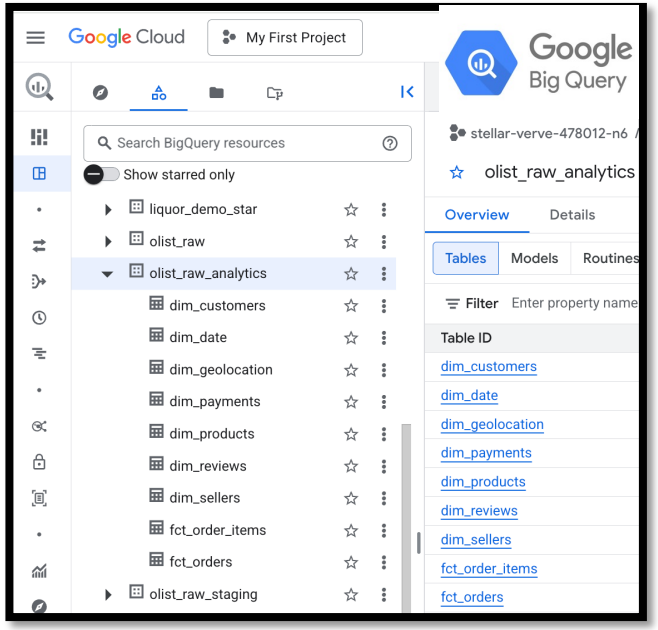
Leveraging dbt, we transformed raw data into a Star Schema format, creating optimized Fact and Dimension tables within BigQuery to support high-performance reporting

Transformation and Modeling



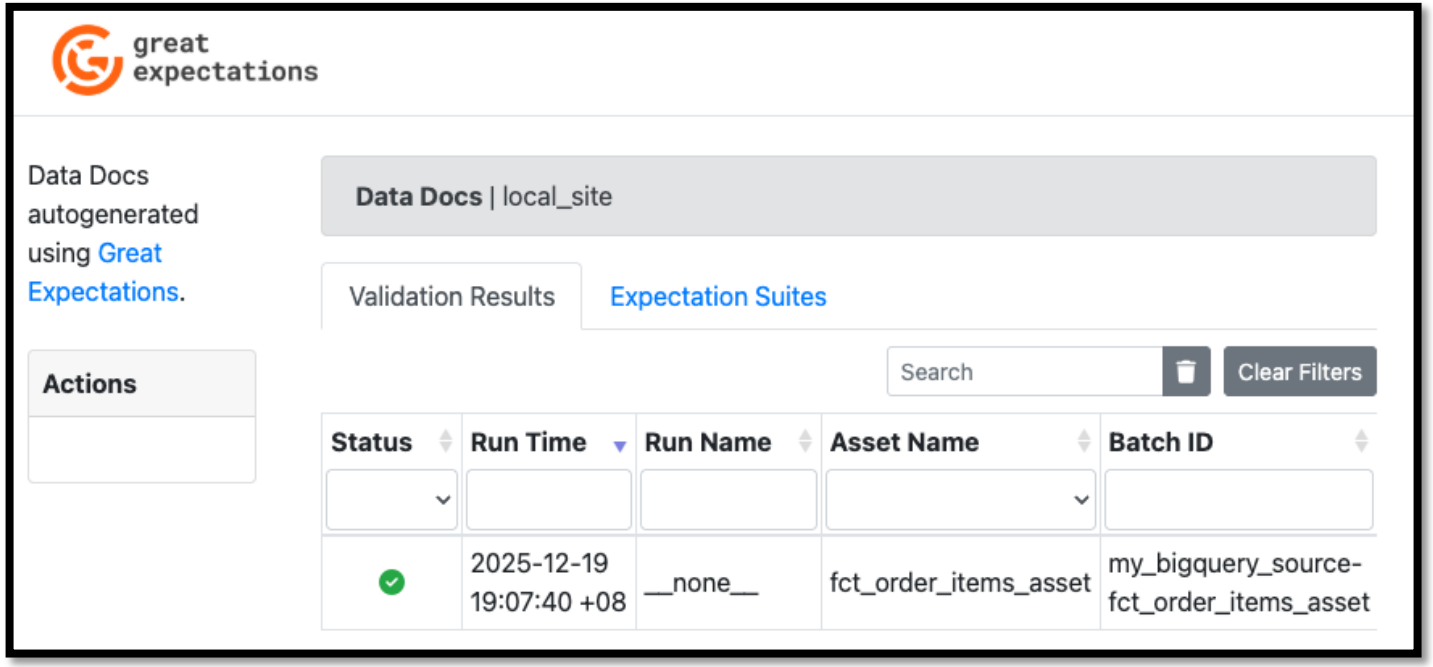
Designed a relational data model connecting key business entities—such as customers, orders, and products—to structure the data for efficient analytical querying

Data Quality and Validation



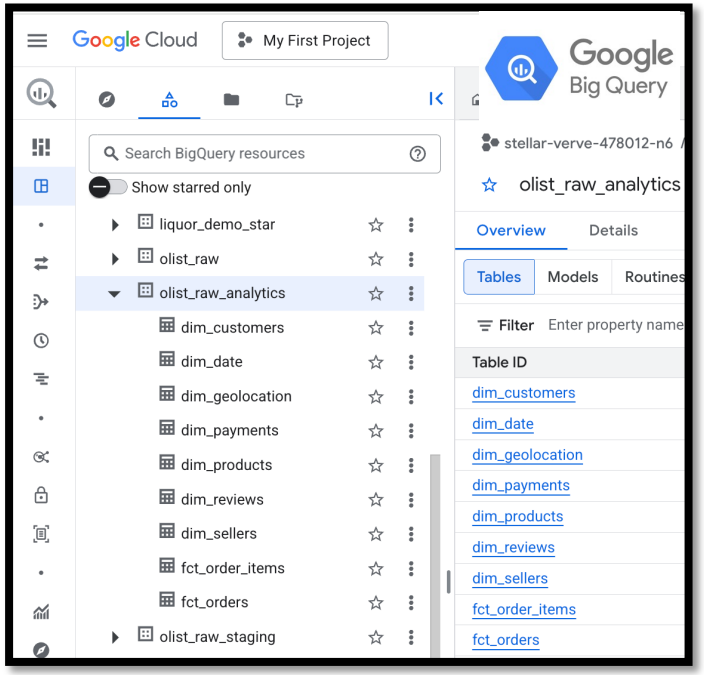
great expectations

GX

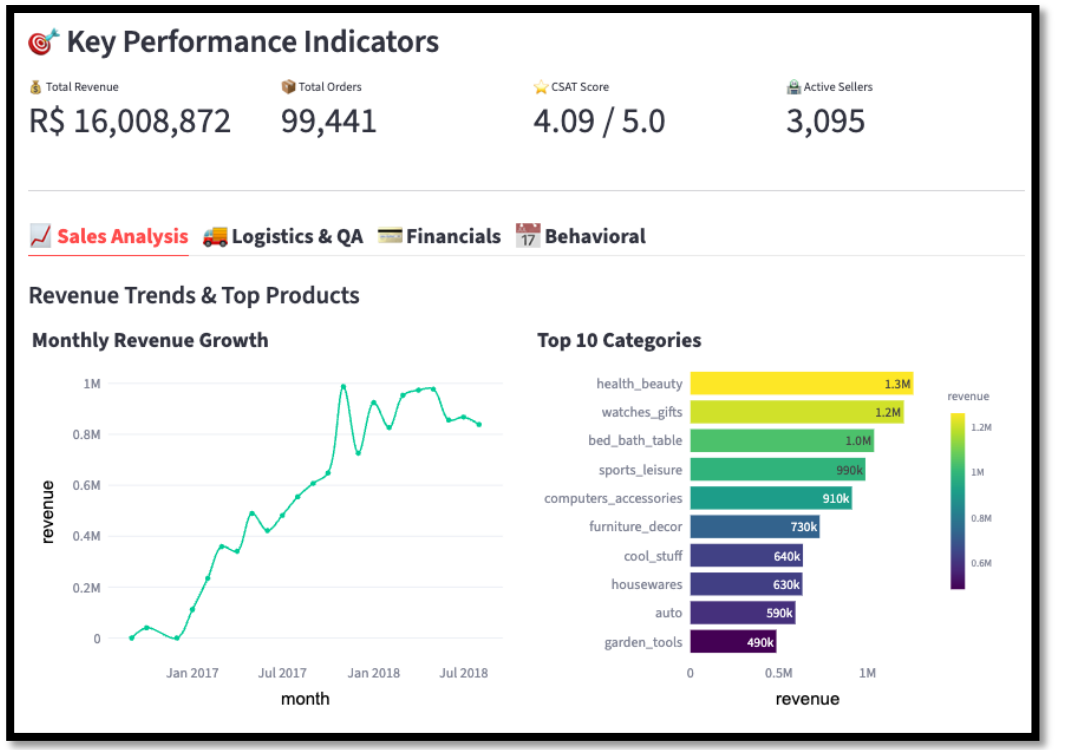


To ensure trust in our insights, we integrated Great Expectations to run automated validation tests, verifying data integrity and quality before it reaches the dashboard

Data Analysis and Business Insights



Streamlit



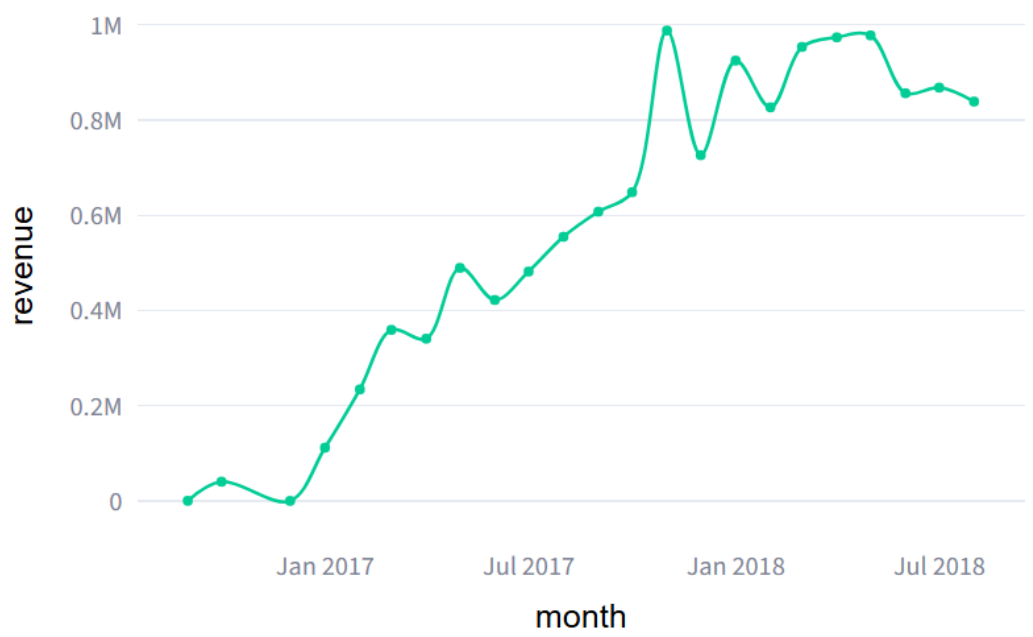
We utilized Python and Streamlit to develop a live, interactive dashboard that transforms complex data into accessible, real-time business insights for stakeholders

Business Insights

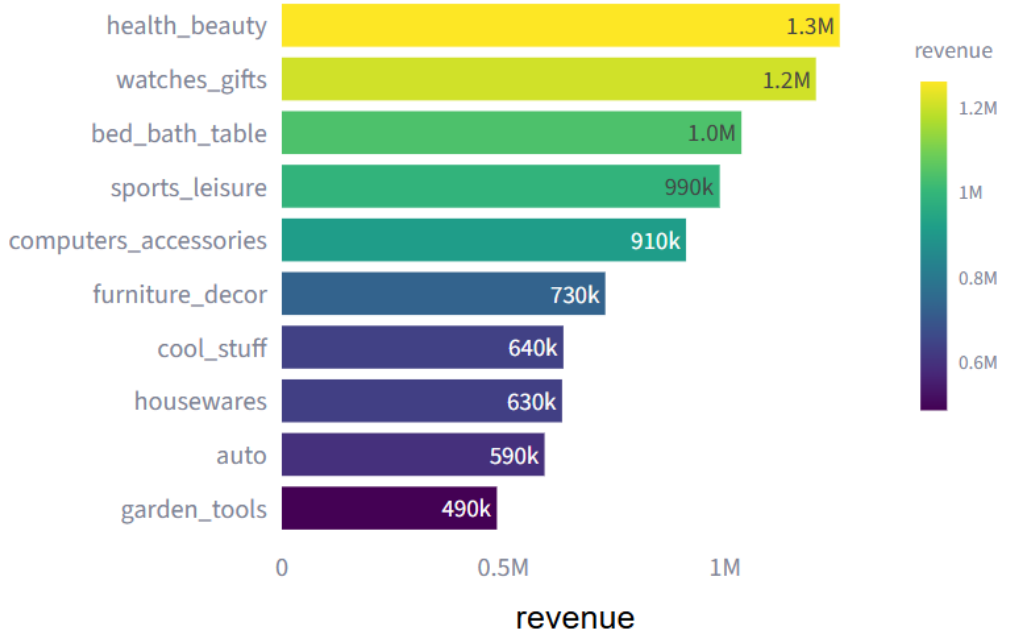
Key Performance Indicators

Total Revenue	Total Orders	CSAT Score	Active Sellers
R$ 16,008,872	99,441	4.09 / 5.0	3,095

Monthly Revenue Growth



Top 10 Categories

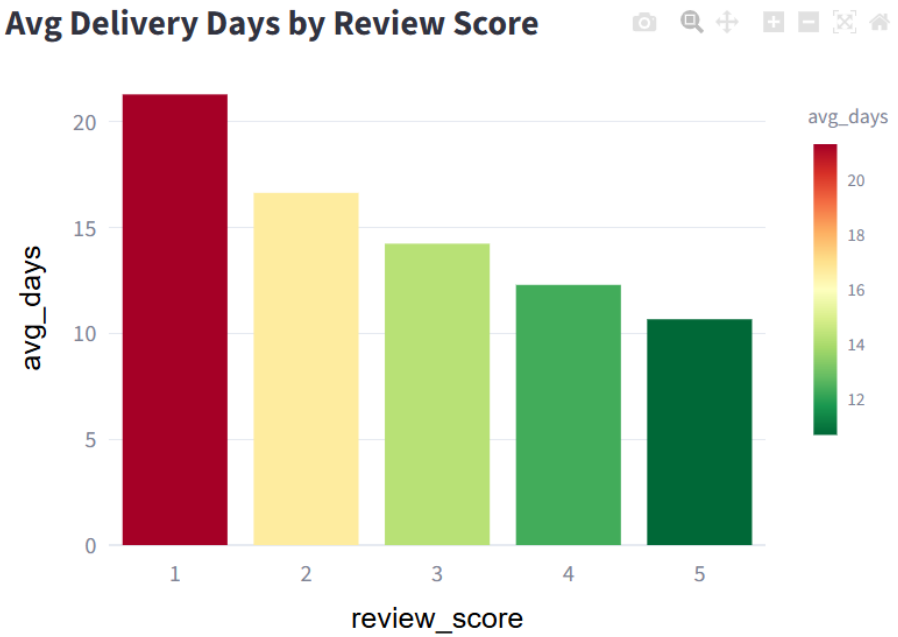


- The revenue shows a steady upward trend
- Revenue peaks in late 2017,reaching close to **R$1million/month**.
- Slight drop in mid-2018 but overall growth is strong.

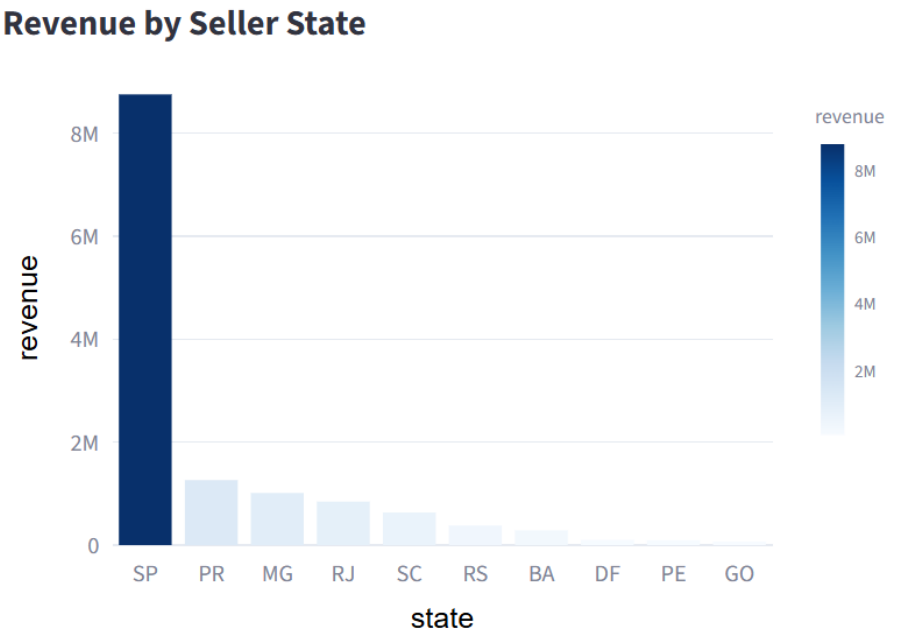
- Product mix – Health & beauty ,watches gifts are bread winner .Marketing spend need to double on these 2 categories

Business Insights

Speed vs. Satisfaction



Seller Distribution



Critical Insight:

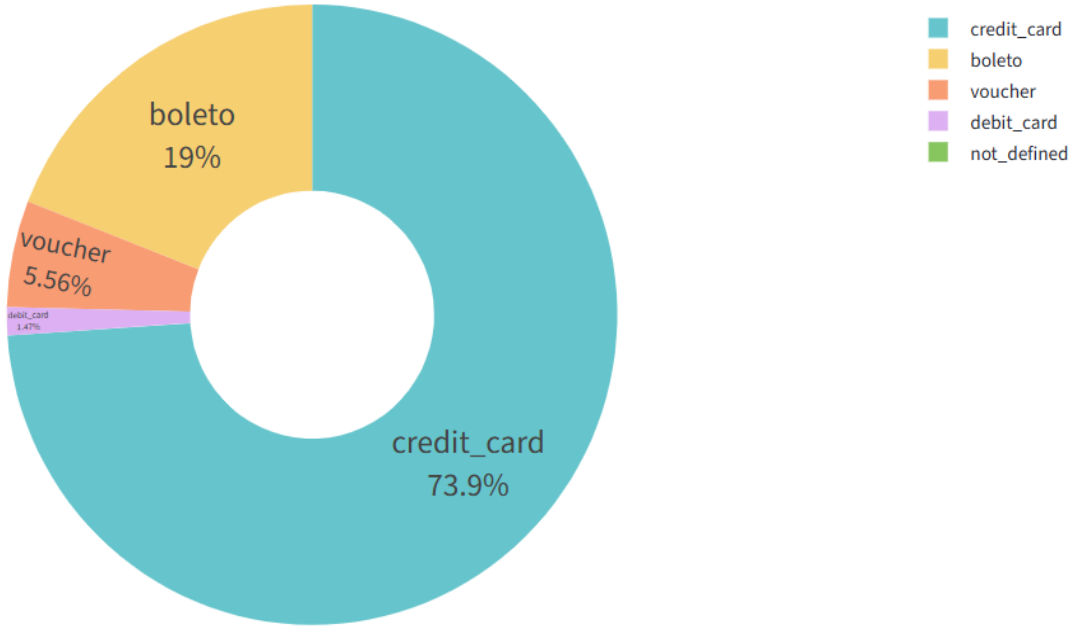
- Direct correlation between review score and avg delivery time
- 5 -star reviews are delivered in 10days while 1-star take 21 days

- São Paulo leads revenue, over twice Rio's; steep drops in other states show e-commerce concentration and growth potential elsewhere.

Business Insights

Payment Preferences

Order Count by Payment Method



Insight:

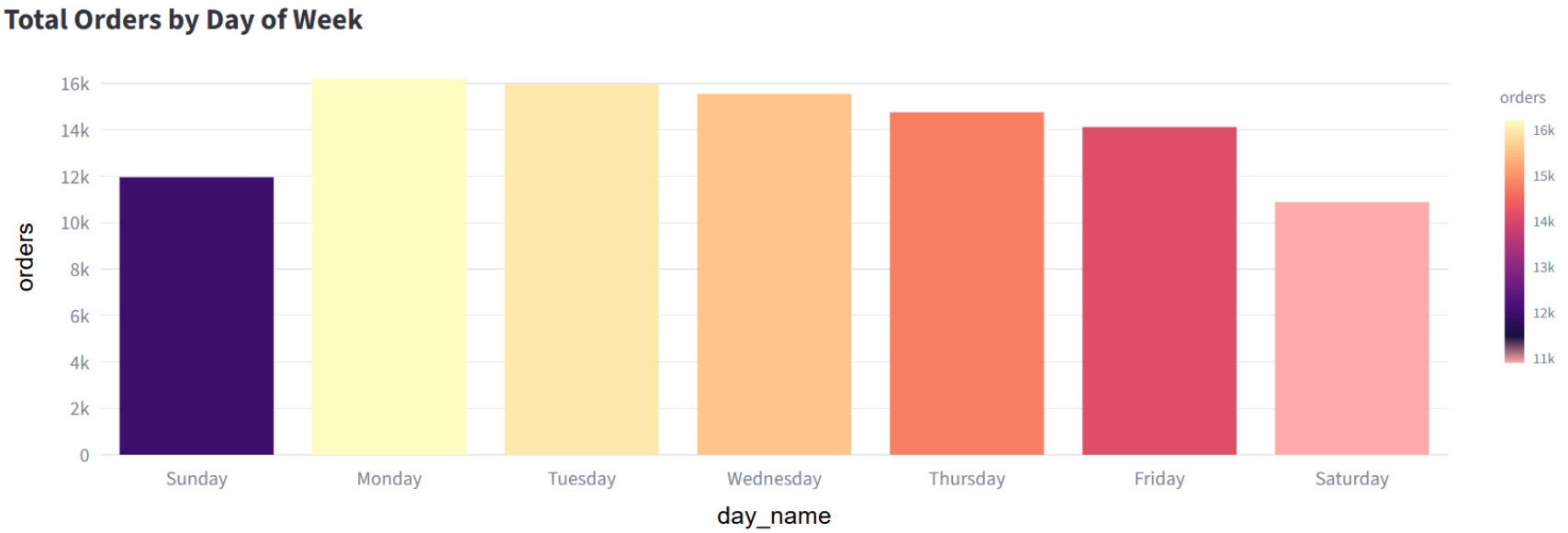
- The majority of customers prefer **credit cards**, showing mature digital customer base.
- **Boleto**, a local Brazilian payment method, still plays a significant role (19%).
- vouchers and debit cards are less used.

Business Implication:

- To ensure payment gateway is robust for smoother checkout.
- Consider **promotions or incentives** for other payment methods to diversify payment risk.

Business Insights

When do customers buy?



The New Insight:

- Behavioral analysis to see when people shop
- **Monday & Tuesday** tops with 16k orders & weekend on low side

Business Implication:

• Launching email campaigns & flash sales should be targeted on Monday , Tuesday

Risk & Mitigation

Data Quality Risks: Missing, inconsistent, or incorrect data may affect insights.
Mitigation: Automated dbt tests, uniqueness & referential integrity checks, and business logic validation.

Cost Overruns: Growing data volume could increase storage and compute costs.
Mitigation: Use external tables, selective materialization, and serverless BigQuery to control expenses.

Scalability Challenges: High data volumes may slow processing.
Mitigation: Serverless architecture and modular design ensure efficient scaling.

Source Data Changes: Schema or structure changes in raw datasets could break the pipeline.
Mitigation: Modular, version-controlled dbt models for easier updates and minimal disruption.

THANK YOU