

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

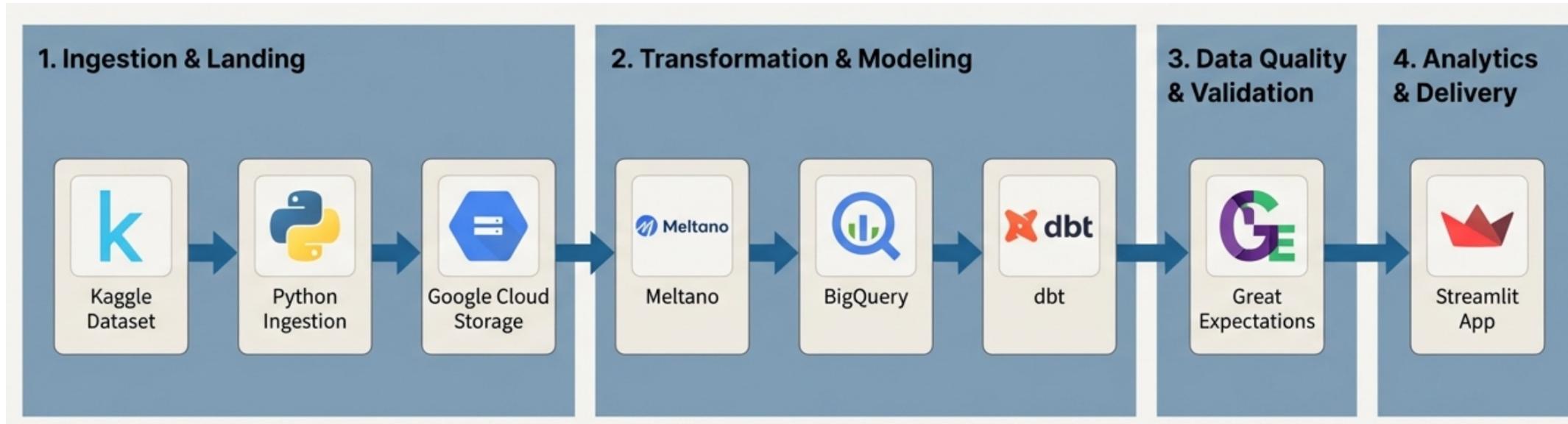
High-Level Results:

- 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	Great Expectations	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-e-commerce>

Brazilian E-Commerce Public Dataset by Olist

3822 Code (655) Discussion (64) Suggestions (0)

View more

olist_customers_dataset.csv (9.03 MB)

Detail Compact Column

5 of 5 columns

Data Explorer

Version 2 (126.19 MB)

- olist_customers_dataset.csv
- olist_geolocation_dataset.csv
- olist_order_items_dataset.csv
- olist_order_payments_dataset.csv
- olist_order_reviews_dataset.csv
- olist_orders_dataset.csv
- olist_products_dataset.csv
- olist_sellers_dataset.csv
- product_category_name_translation.csv

Summary

- 9 files
- 52 columns



Google Cloud My First Project

Cloud Storage Bucket details

dsai2-olist-raw-data

Location us (multiple regions in United States) Storage class Standard Public access Not public Protection Soft Delete

Objects Configuration Permissions Protection Lifecycle Observability

Folder browser

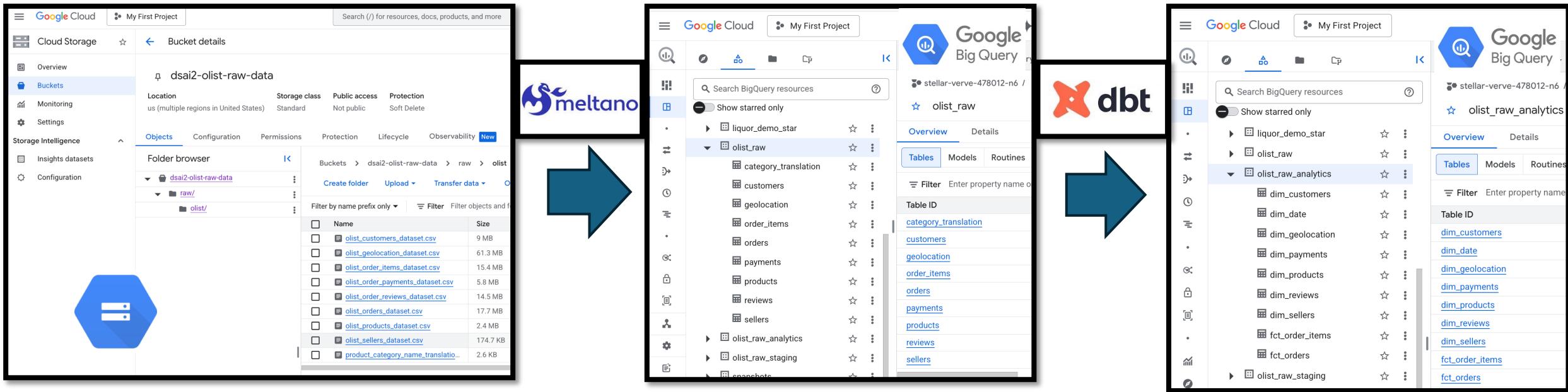
- dsai2-olist-raw-data
- raw/
- olist/

Filter by name prefix only

Name	Size
olist_customers_dataset.csv	9 MB
olist_geolocation_dataset.csv	61.3 MB
olist_order_items_dataset.csv	15.4 MB
olist_order_payments_dataset.csv	5.8 MB
olist_order_reviews_dataset.csv	14.5 MB
olist_orders_dataset.csv	17.7 MB
olist_products_dataset.csv	2.4 MB
olist_sellers_dataset.csv	174.7 KB
product_category_name_translatio...	2.6 KB

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

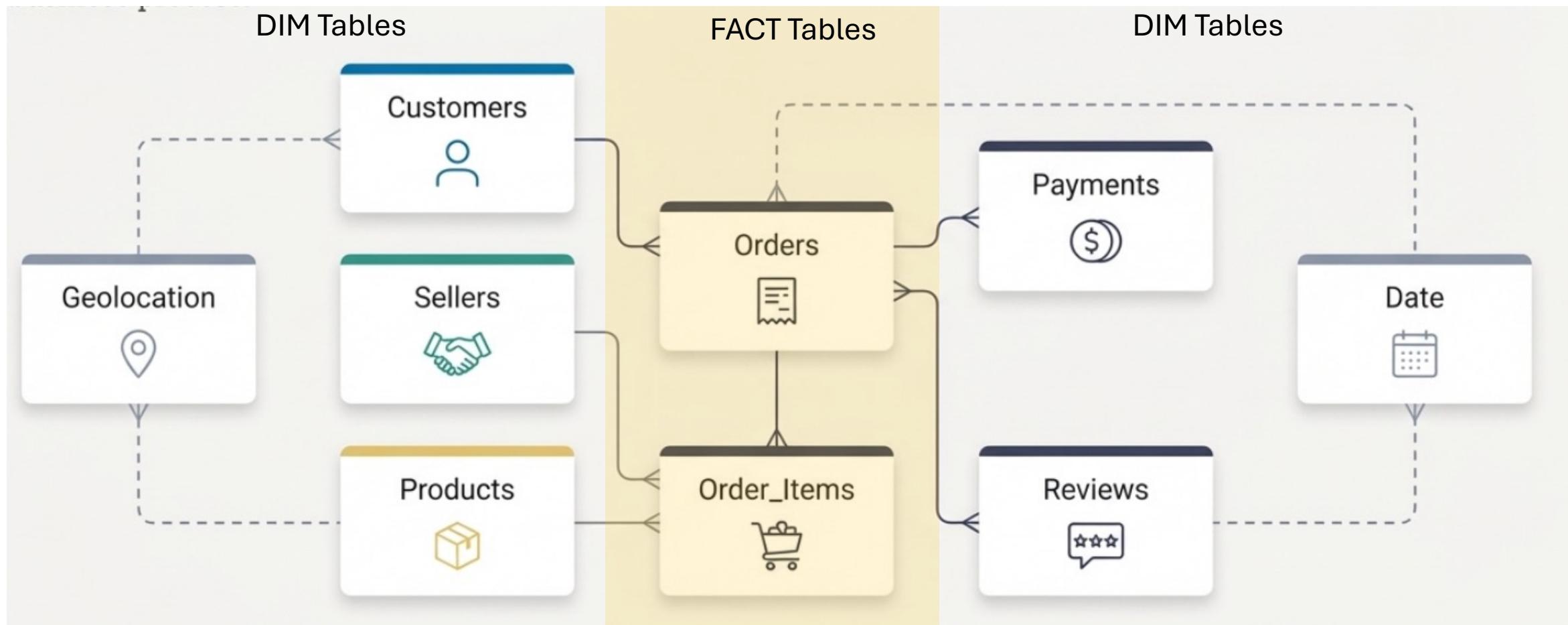
Transformation and Modeling



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

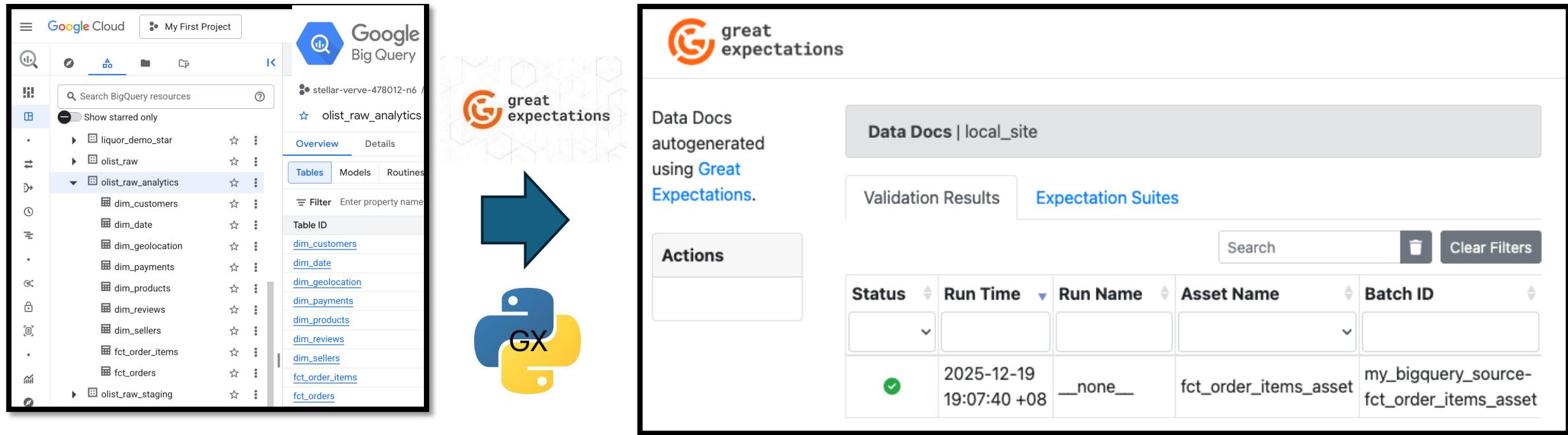
Leveraging dbt, we transformed raw data into a Star Schema format, creating optimized Fact and Dimension tables within BigQuery to support high-performance reporting

Transformation and Modeling



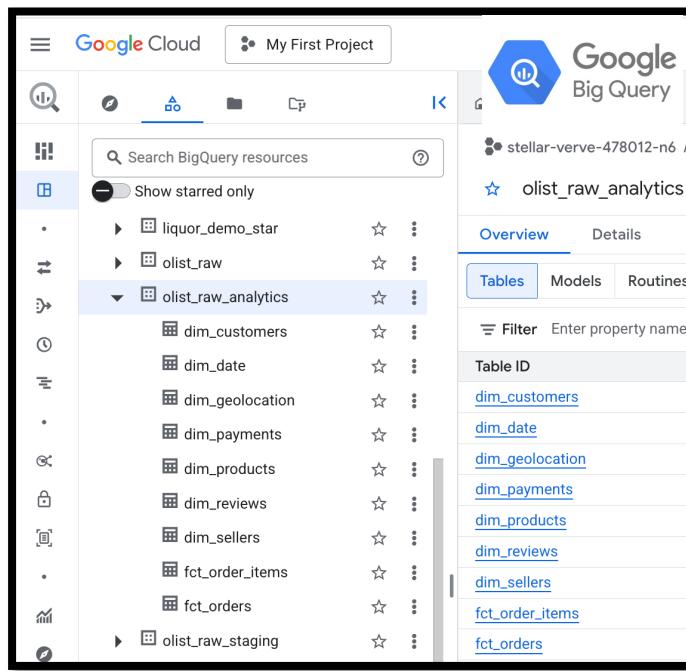
Designed a relational data model connecting key business entities—such as customers, orders, and products—to structure the data for efficient analytical querying

Data Quality and Validation

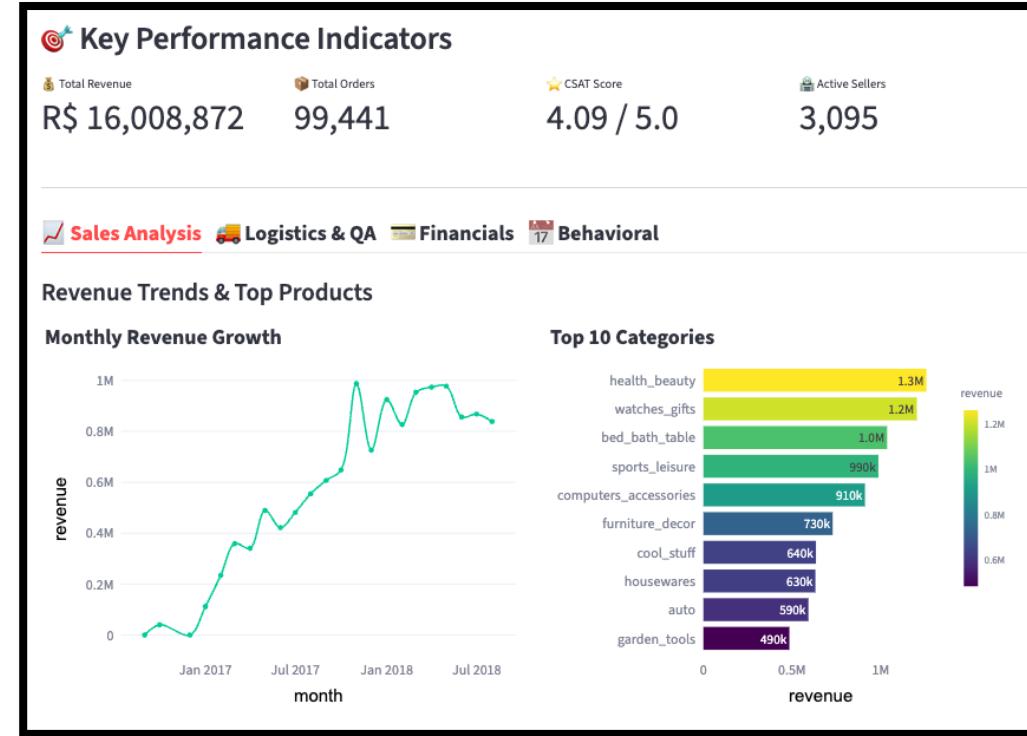
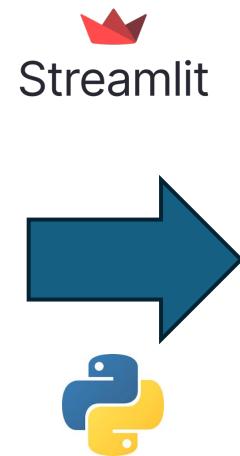


To ensure trust in our insights, we integrated Great Expectations to run automated validation tests, verifying data integrity and quality before it reaches the dashboard

Data Analysis and Business Insights



The screenshot shows the Google Cloud BigQuery interface. At the top, it says "Google Cloud" and "My First Project". On the left, there's a sidebar with various icons and a search bar. The main area shows a list of datasets: "liquor_demo_star", "olist_raw", "olist_raw_analytics", and "olist_raw_staging". "olist_raw_analytics" is currently selected, indicated by a blue border. Below it, under "Tables", there's a list of tables: "dim_customers", "dim_date", "dim_geolocation", "dim_payments", "dim_products", "dim_reviews", "dim_sellers", "fct_order_items", "fct_orders", and "fct_orders".



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
R\$ 16,008,872



Total Orders
99,441



CSAT Score
4.09 / 5.0

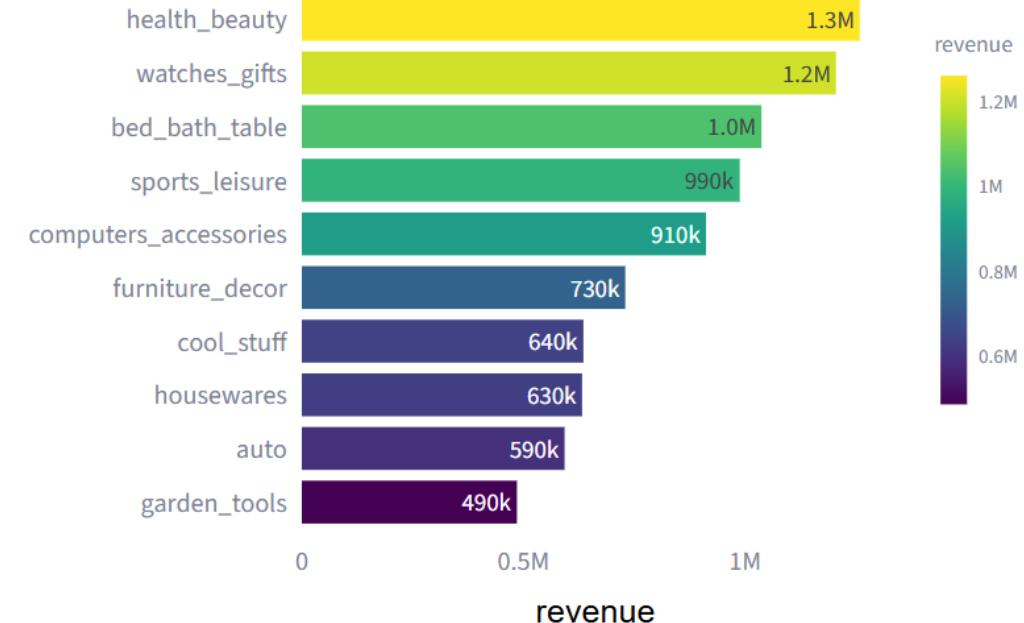


Active Sellers
3,095

Monthly Revenue Growth



Top 10 Categories



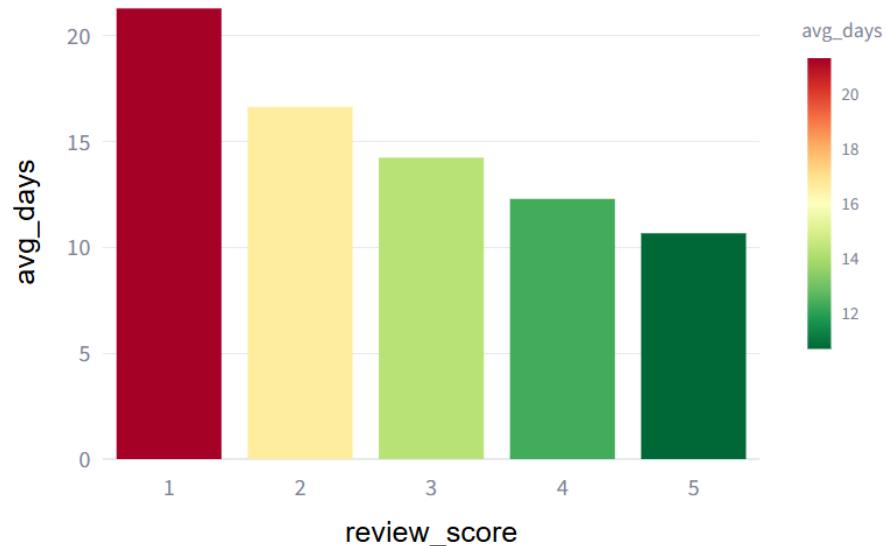
- The revenue shows a steady upward trend
- Revenue peaks in late 2017, reaching close to **R\$1million/month**.
- Slight drop in mid-2018 but overall growth is strong.

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

Business Insights

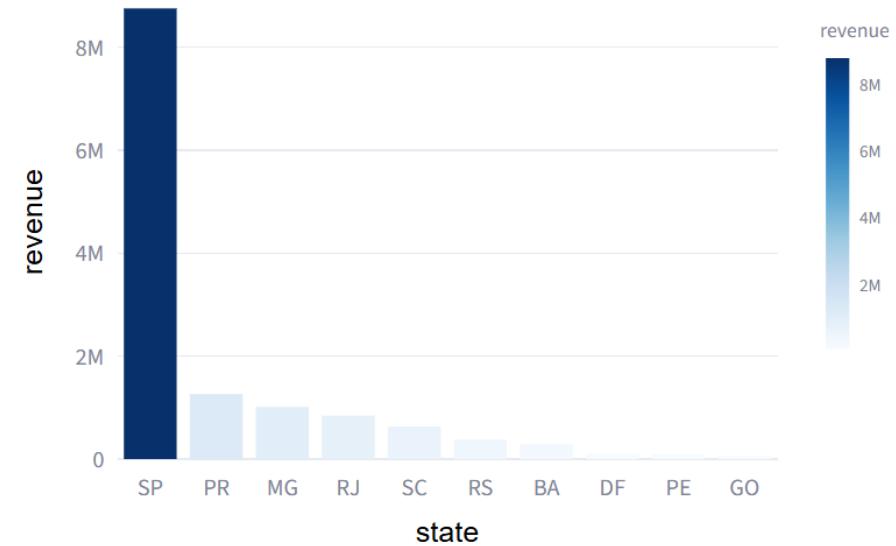
🐢 Speed vs. Satisfaction

Avg Delivery Days by Review Score



🌐 Seller Distribution

Revenue by Seller State



Critical Insight:

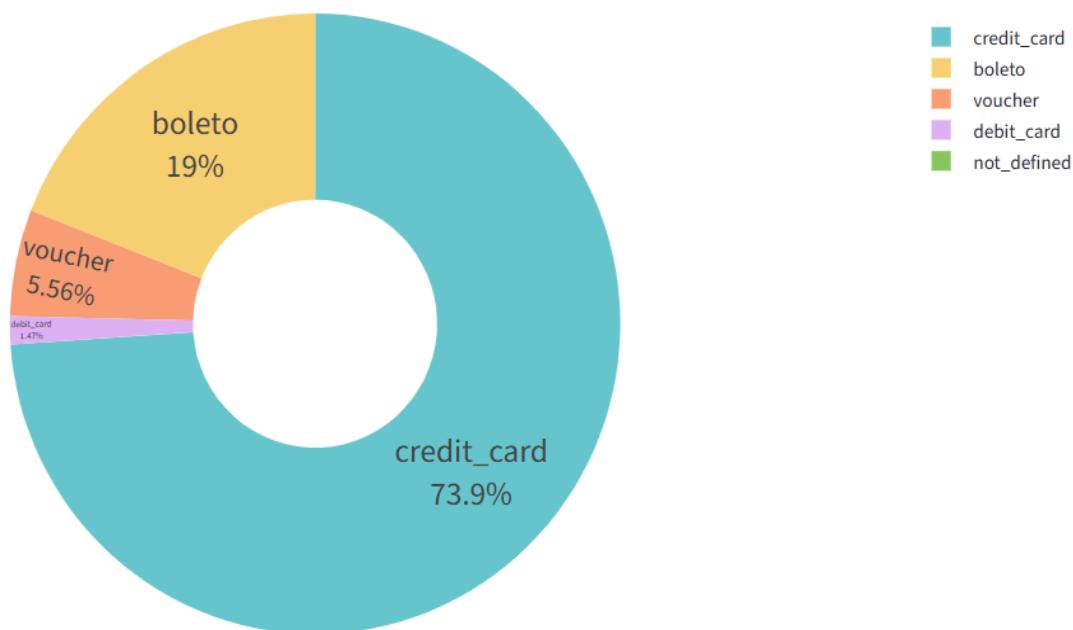
- Direct correlation between review score and avg delivery time
- 5 -star reviews are delivered in 10days while 1-star take 21 days

- 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.

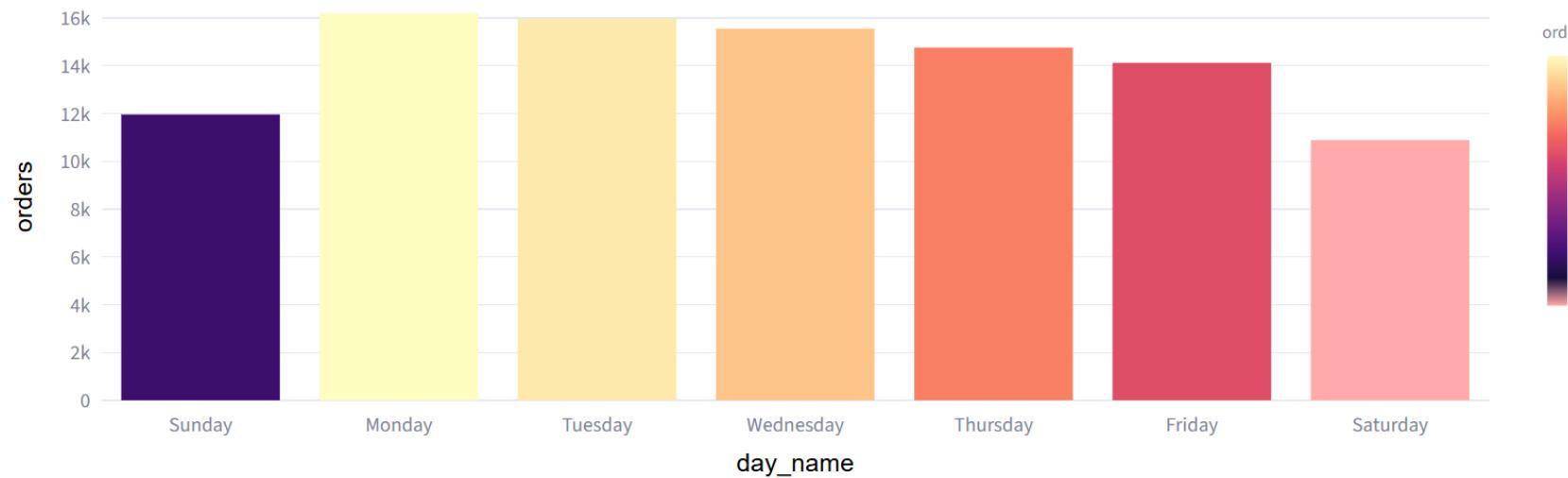
Business Implication:

- 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?

Total Orders by Day of Week



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