End-to-End Data Pipeline: The Olist E-commerce Project

A Journey from Raw, Distributed Data to Automated, Actionable Insights

Presented by: Yousef Soliman | Data Engineer

The Complete Architecture & Data Flow

Project Objective: The project aims to create an **automated end-to-end data pipeline using the ELT methodology**, transforming raw, scattered data into an interactive dashboard for business insights. It adopts a **Medallion architecture** with three layers: **Bronze** (raw, immutable data), **Silver** (validated, conformed data), and **Gold** (refined, aggregated data for analytics), ensuring data quality, governance, and usability.

The Sales Data Project Workflow

1. Multiple Sources

PostgreSQL, Local CSVs, & Snowflake

2. Exploratory Data Analysis

Pandas for nulls counts & data info

3. Ingestion to Bronze

Connectors & SQLAlchemy to Snowflake

4. Medallion Layers

Bronze, Silver, & Gold in Snowflake

5. Transformation

dbt from Bronze to Gold (Star Schema)

The Complete Architecture & Data Flow (continued)

Final Pipeline Steps

6. Star Schema Build

Designed with dbdraw.io

7. Data Testing

Quality Checks via dbt tests

8. Orchestration

Airflow for automated ingestion

9. Visualization

Python, SQL, & Power BI on Gold Layer

Technologies & Tools

Python

Data extraction and ingestion scripting

Docker

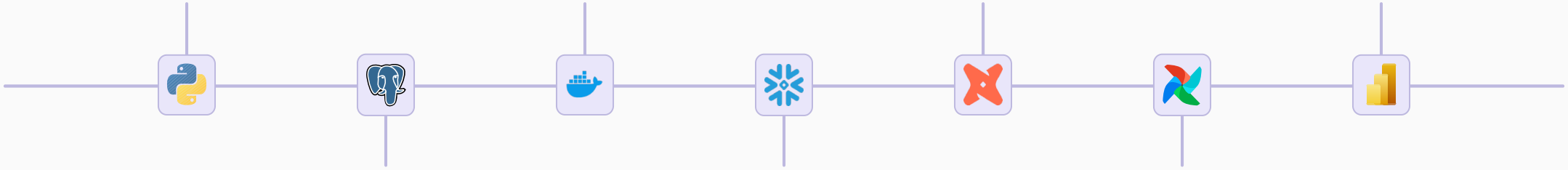
Containerization for consistent environments

dbt

Data transformation and testing framework

Power BI

Business intelligence and visualization



PostgreSQL

Operational database source system

Snowflake

Cloud data warehouse platform

Apache Airflow

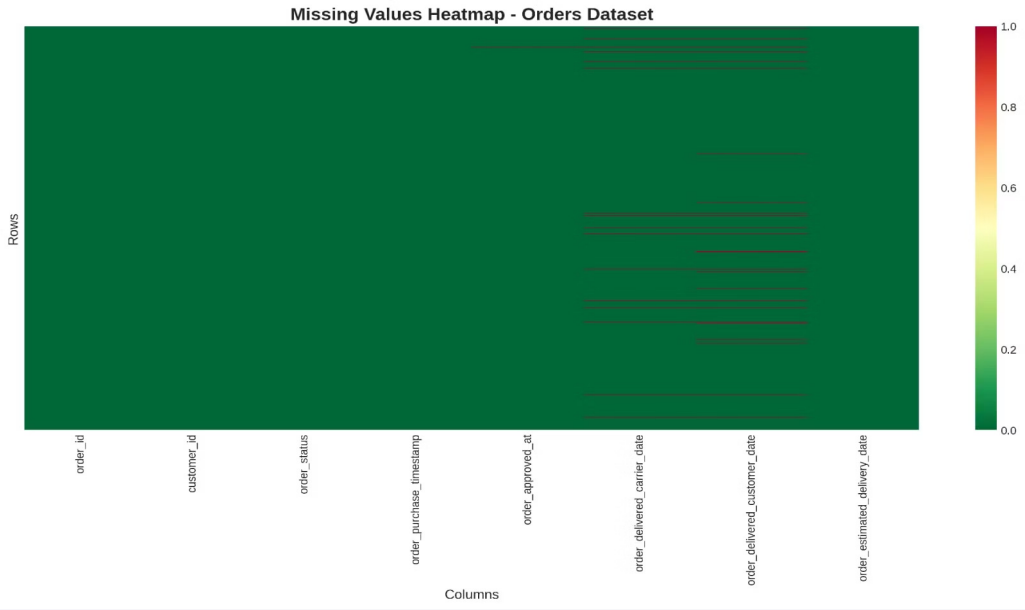
Workflow orchestration and automation

Raw Data Analysis: Uncovering Quality Issues

A Deep Dive into the 'Before' State of the Olist Dataset

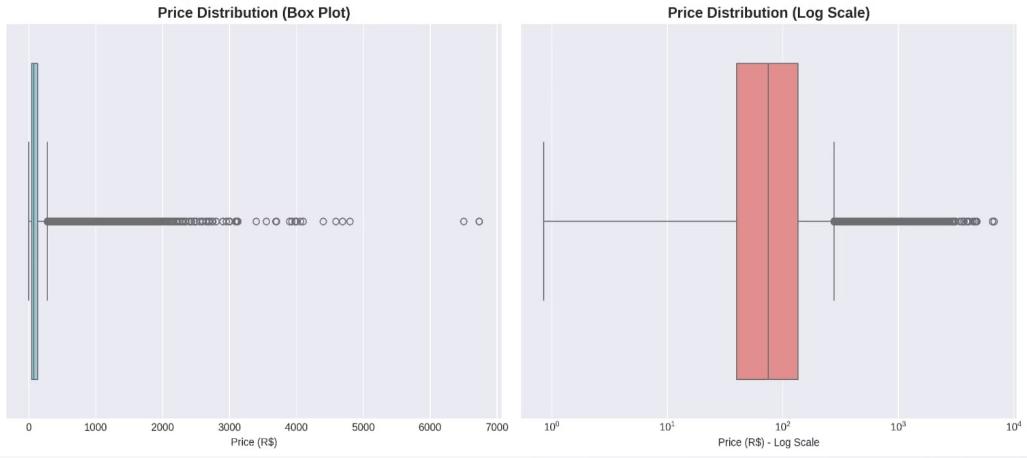
Before designing the transformation pipeline, a comprehensive Exploratory Data Analysis (EDA) using Python, Pandas, and Seaborn uncovered critical data quality issues and business patterns. The following slides reveal the challenges that drove every architectural decision in this project.

Data Quality Assessment

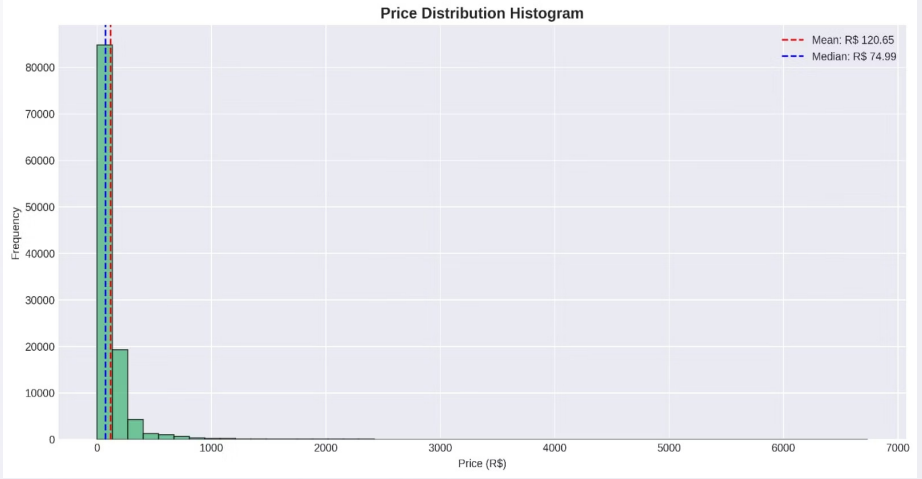


Yellow bands reveal over 2,900 orders with missing critical delivery dates

Order approval and delivery timestamps show the most significant nulls, impacting performance metrics

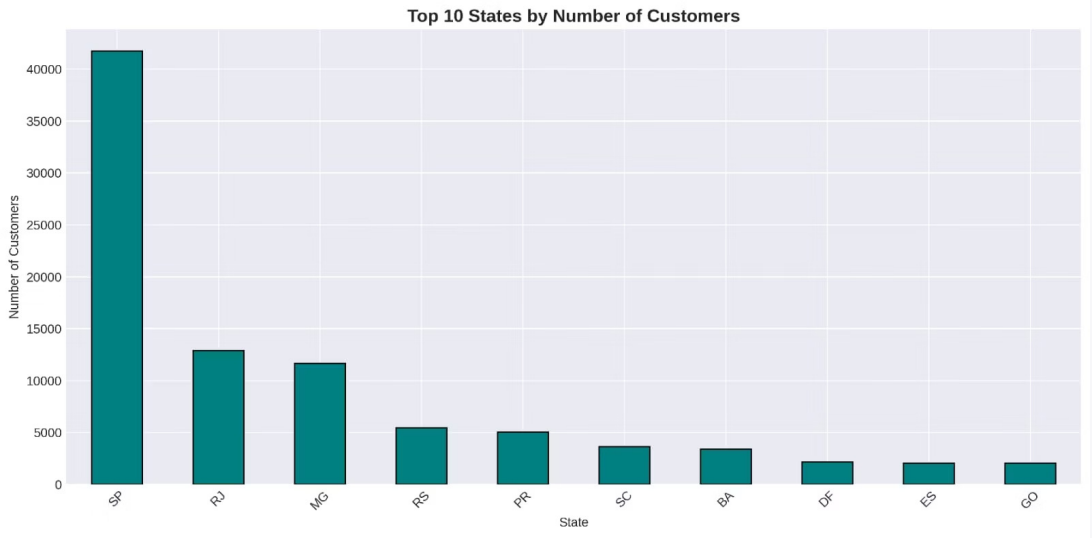


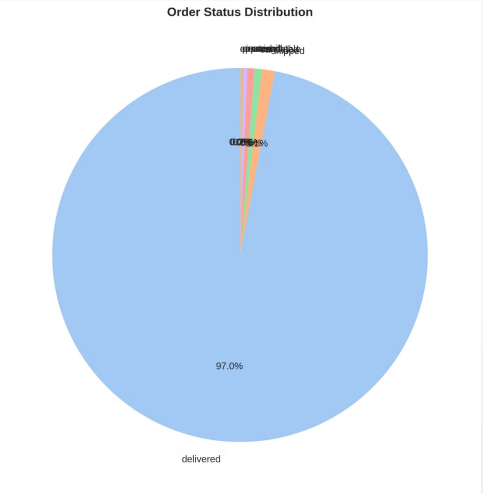
Box plot shows product prices ranging from R$0.85 to R$6,735,



Most products are priced under R$100, but a long tail of high-

Business & Sales Patterns

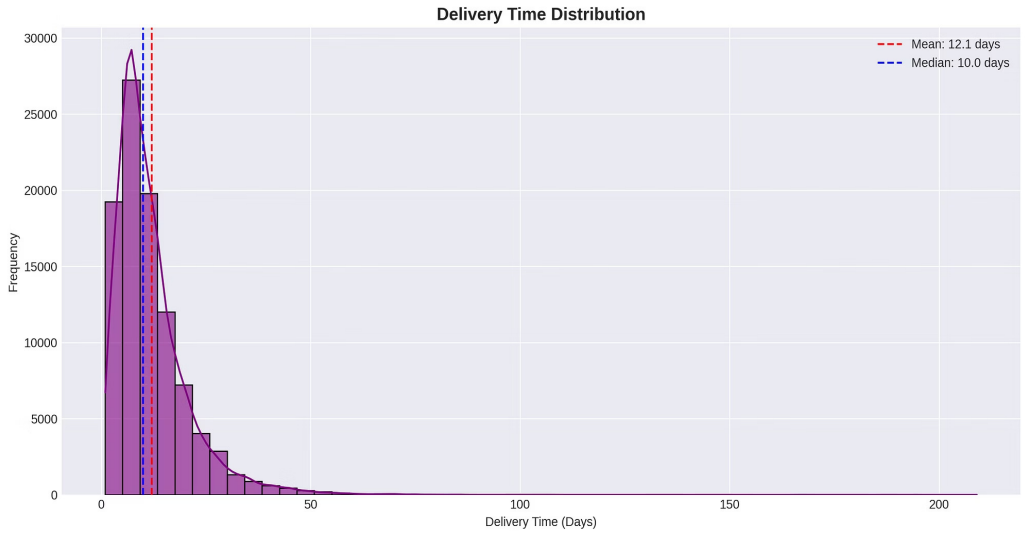




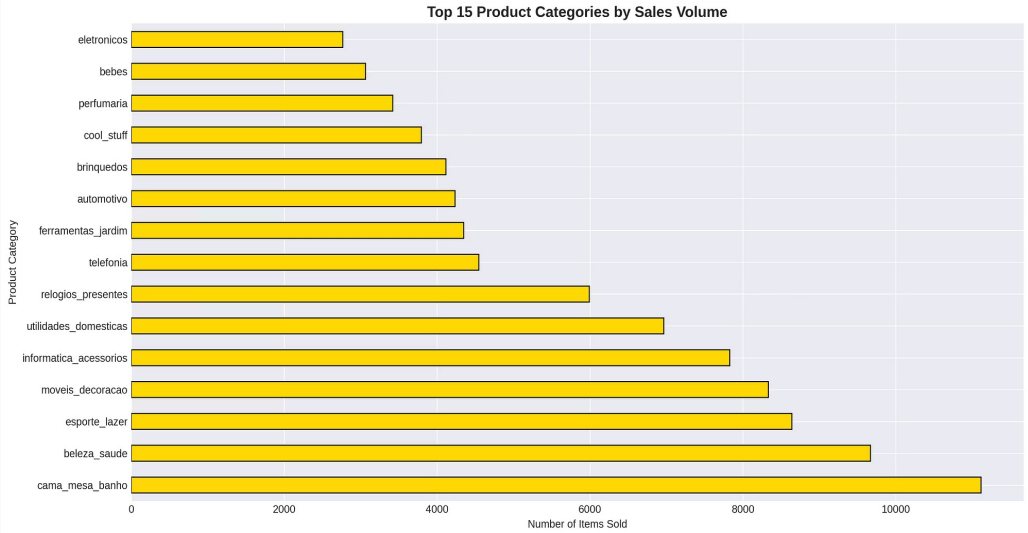
Geographic Concentration: Over 85% of sales are concentrated in the Southeast, with São Paulo dominating, revealing expansion opportunities.

Order Lifecycle: 97% of orders are successfully delivered.

Operational & Product Insights



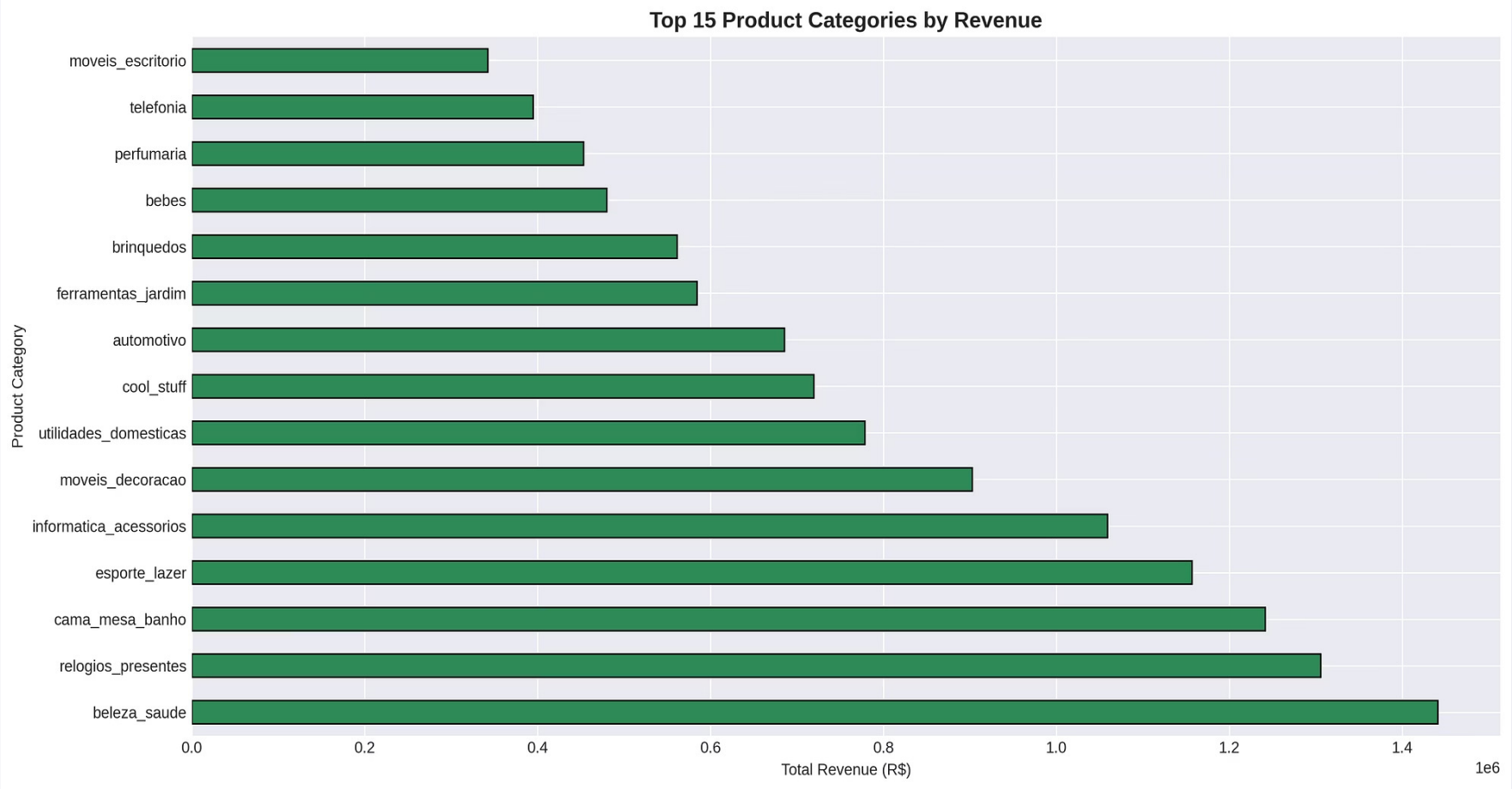
The average delivery time is 12.5 days, but the high variance indicates inconsistent and unreliable logistics



'Bed, Bath & Table' items are the most frequently sold, followed by 'Health & Beauty' products

Operational & Product Insights (continued)

Revenue Analysis



EDA Summary & Implications

These findings validated the need for a robust data warehouse with a Bronze-Silver-Gold architecture to systematically clean, standardize, and enrich the raw data for reliable analytics.

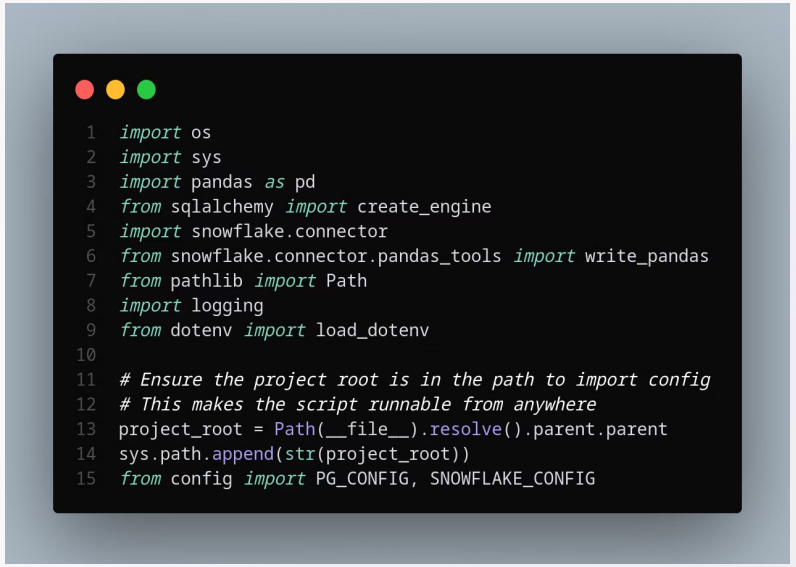
Key Discovery Points

- ~3% of orders have missing delivery dates
- Over 12% of items were identified as price outliers
- A single state (SP) accounts for over 40% of all revenue
- Average delivery time of 12.5 days with high inconsistency

Step 1: Data Ingestion from Multiple Sources

Our project began by addressing a common challenge: disparate data sources. We needed to consolidate raw information from various origins 4specifically, a PostgreSQL production database and several flat CSV files4into a single, accessible hub.

A custom Python script, using libraries like Pandas, SQLAlchemy, and the Snowflake Connector, was developed to extract and load data directly into the BRONZE layer of our Snowflake data warehouse. This process established a unified repository for all raw data, creating the foundation for our ELT pipeline.

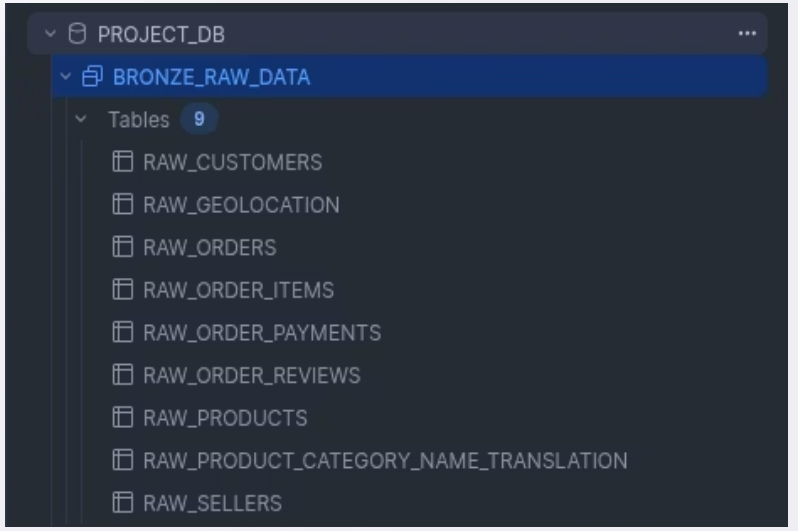


```python
import os
import sys
import pandas as pd
from sqlalchemy import create_engine
import snowflake.connector
from snowflake.connector.pandas_tools import write_pandas
from pathlib import Path
import logging
from dotenv import load_dotenv

# Ensure the project root is in the path to import config
# This makes the script runnable from anywhere
project_root = Path(__file__).resolve().parent.parent
sys.path.append(str(project_root))
from config import PG_CONFIG, SNOWFLAKE_CONFIG
```

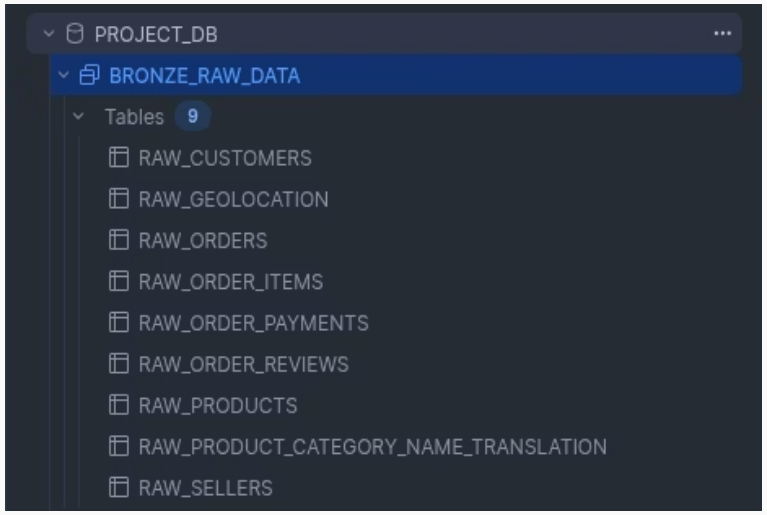
Why this data?

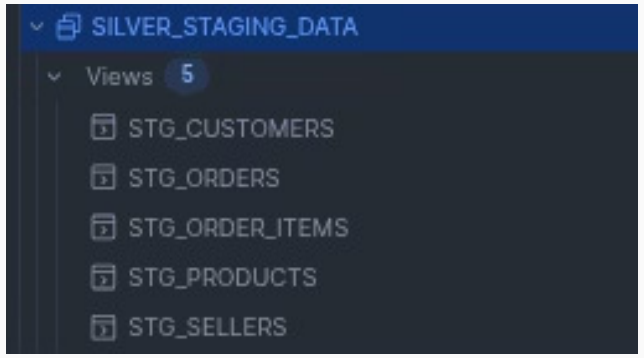
because it simulates a real work environment. It's not a single clean file but a complex dataset of nine interconnected tables with genuine data quality problems, including `NULLs`, `Duplicates`, and `Inconsistent Data`.

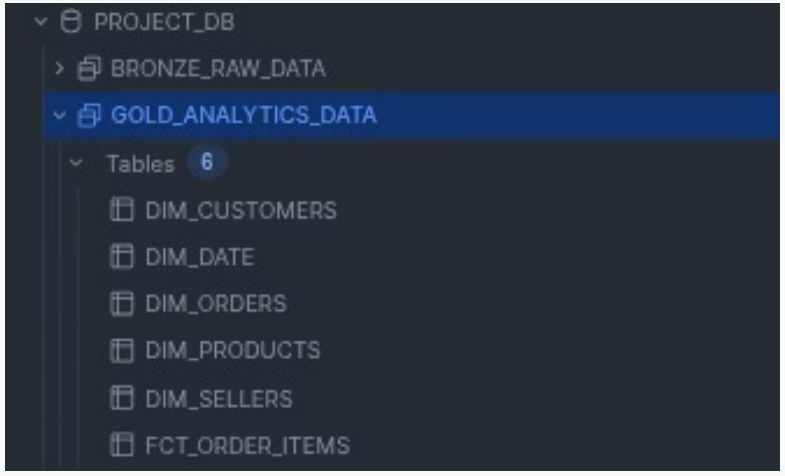


Step 1: Data Ingestion (continued)

Target Data Warehouse

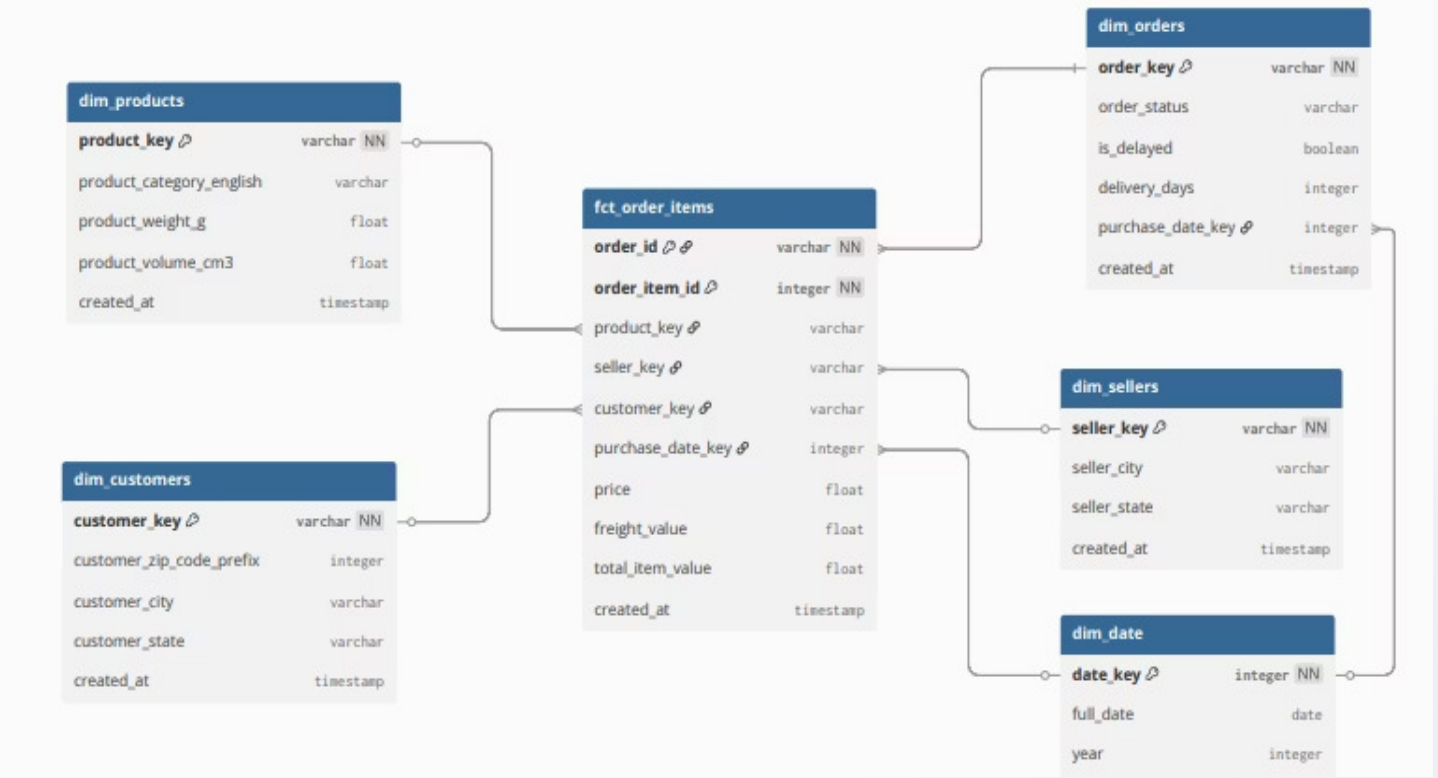






All raw data is loaded into Snowflake's BRONZE layer, establishing our cloud data warehouse foundation.

Step 2: Designing the Star Schema



Design the Star Schema

The Star Schema was selected for its efficiency in data analysis. This design simplifies complex queries, improves report performance

Fact Table Definition

The central **fact table** (e.g., `fct_order_items`) captures key numerical metrics (`price`, `freight_value`).

Dimension Tables Definition

Surrounding the fact table, **dimension tables** (e.g., `dim_customers`, `dim_products`, `dim_date`).

Step 3: Data Transformation & Debugging Journey

This critical step utilizes dbt (data build tool) to streamline our data transformation processes, ensuring high data quality and reliability from raw ingestion to final analytical models. Beyond just transformation, this phase also became a significant **debugging journey**, teaching us valuable lessons in data resilience and problem-solving.

Mechanism of Action

Staging Models

Data is ingested from the raw BRONZE layer, where initial cleaning steps are applied. This includes data type casting, string unification, and deduplication to prepare the data for further processing, outputting to the SILVER layer. **This is where many initial data quality issues were first encountered and addressed.**

Marts Models

Building upon the cleaned data in the SILVER layer, these models perform complex JOIN operations to construct the final Star Schema tables. This creates highly optimized fact and dimension tables in the aggregated GOLD layer, ready for business analysis. **Complex joins often exposed hidden data quirks, requiring iterative debugging.**

Step 3 Screenshot

```
.venvpro ❯ dbt run --full-refresh --no-partial-parse
12:00:48  Running with dbt=1.11.0-b2
12:00:49  Registered adapter: snowflake=1.10.2
12:00:51  Found 9 models, 26 data tests, 9 sources, 493 macros
12:00:51
12:00:51  Concurrency: 4 threads (target='dev')
12:00:51
12:00:53  4 of 9 START sql view model BRONZE_RAW_DATA_SILVER_STAGING_DATA.stg_products ... [RUN]
12:00:53  1 of 9 START sql view model BRONZE_RAW_DATA_SILVER_STAGING_DATA.stg_customers .. [RUN]
12:00:53  3 of 9 START sql view model BRONZE_RAW_DATA_SILVER_STAGING_DATA.stg_orders ..... [RUN]
12:00:53  2 of 9 START sql view model BRONZE_RAW_DATA_SILVER_STAGING_DATA.stg_order_items  [RUN]
12:00:54  1 of 9 OK created sql view model BRONZE_RAW_DATA_SILVER_STAGING_DATA.stg_customers  [SUCCESS 1 in 0.60s]
12:00:54  2 of 9 OK created sql view model BRONZE_RAW_DATA_SILVER_STAGING_DATA.stg_order_items  [SUCCESS 1 in 0.59s]
12:00:54  5 of 9 START sql view model BRONZE_RAW_DATA_SILVER_STAGING_DATA.stg_sellers .... [RUN]
12:00:54  6 of 9 START sql table model BRONZE_RAW_DATA_GOLD_ANALYTICS_DATA.dim_customers . [RUN]
12:00:54  5 of 9 OK created sql view model BRONZE_RAW_DATA_SILVER_STAGING_DATA.stg_sellers  [SUCCESS 1 in 0.54s]
12:00:54  7 of 9 START sql table model BRONZE_RAW_DATA_GOLD_ANALYTICS_DATA.dim_sellers ... [RUN]
12:00:54  4 of 9 OK created sql view model BRONZE_RAW_DATA_SILVER_STAGING_DATA.stg_products  [SUCCESS 1 in 1.21s]
12:00:54  8 of 9 START sql table model BRONZE_RAW_DATA_GOLD_ANALYTICS_DATA.dim_products .. [RUN]
12:00:54  3 of 9 OK created sql view model BRONZE_RAW_DATA_SILVER_STAGING_DATA.stg_orders  [SUCCESS 1 in 1.24s]
12:00:54  9 of 9 START sql table model BRONZE_RAW_DATA_GOLD_ANALYTICS_DATA.fct_order_items  [RUN]
12:00:56  7 of 9 OK created sql table model BRONZE_RAW_DATA_GOLD_ANALYTICS_DATA.dim_sellers  [SUCCESS 1 in 1.34s]
12:00:56  6 of 9 OK created sql table model BRONZE_RAW_DATA_GOLD_ANALYTICS_DATA.dim_customers  [SUCCESS 1 in 2.27s]
12:00:56  9 of 9 OK created sql table model BRONZE_RAW_DATA_GOLD_ANALYTICS_DATA.fct_order_items  [SUCCESS 1 in 2.16s]
12:00:57  8 of 9 OK created sql table model BRONZE_RAW_DATA_GOLD_ANALYTICS_DATA.dim_products  [SUCCESS 1 in 2.39s]
12:00:58
12:00:58  Finished running 4 table models, 5 view models in 0 hours 0 minutes and 7.02 seconds (7.02s).
12:00:58
12:00:58  Completed successfully
12:00:58
12:00:58  Done. PASS=9 WARN=0 ERROR=0 SKIP=0 NO-OP=0 TOTAL=9
```

Real-World Challenges & Debugging Insights

Our transformation journey was not without its hurdles. Each challenge presented an opportunity to refine our processes and deepen our understanding of the data.

Case Sensitivity in Column Names

Problem: Inconsistent casing (`Order_ID` vs `order_id`) led to "column not found" errors. **Solution:** Implemented standardizing `lower()` functions and strict naming conventions. **Lesson:** Early data cleaning prevents downstream failures.

Environment Variable Misconfigurations

Problem: Differences in `profiles.yml` caused connection issues. **Solution:** Centralized environment variable management and documented procedures. **Lesson:** Rigorous environment configuration is paramount.

Unexpected Data Types & NULL Values

Problem: Source data changes broke SQL logic. **Solution:** Proactive dbt tests and defensive SQL coding with `COALESCE` and `CAST`. **Lesson:** Always "trust but verify" source data.

Step 4: Ensuring Data Quality with dbt Tests

Building on our transformation efforts, the final crucial step involved rigorously validating our data to guarantee accuracy and reliability for downstream analytics. dbt's powerful testing framework allowed us to implement a comprehensive suite of data quality checks. A total of 26 automated tests were deployed across our models.

Unique Tests

These tests verify that specified columns, typically primary keys, contain only distinct values, preventing duplication and ensuring the integrity of individual records.

Not Null Tests

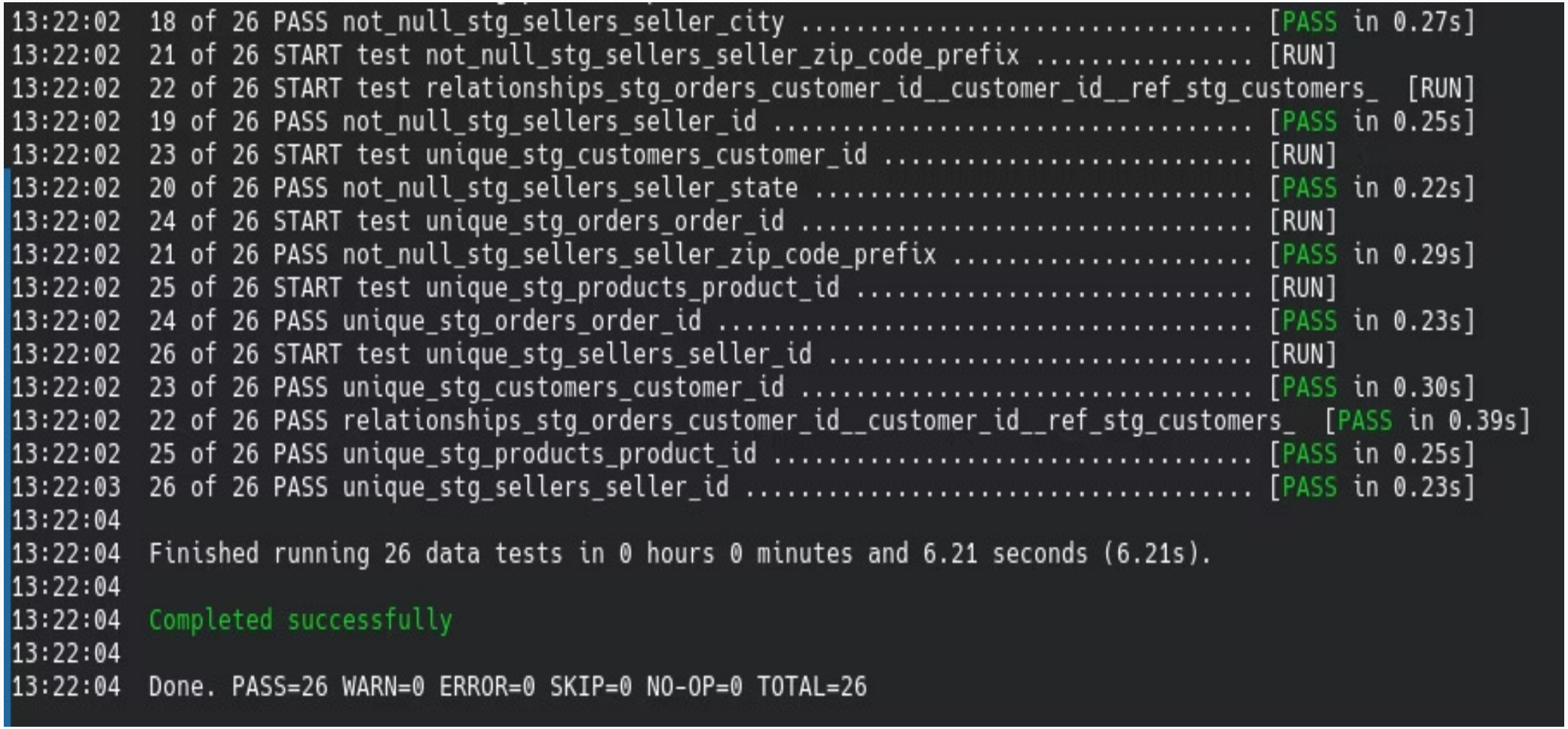
Critical for data completeness, these tests confirm that important columns do not contain any missing or null values, ensuring all essential data points are present.

Relationships Tests

These tests validate referential integrity between models, ensuring that foreign keys in one table correctly reference existing primary keys in another, thus maintaining coherent connections within our Star Schema.

We are proud to report that all implemented data quality tests passed successfully, confirming the robustness and accuracy of our transformed data. This rigorous validation ensures that our analytical models are built on a foundation of trustworthy information, critical for confident decision-making.

Step 4 Screenshot



Step 5: Orchestration & Automation with Airflow

Apache Airflow was instrumental in transforming our manual data processes into a fully automated and reliable pipeline. By orchestrating the entire workflow, we've ensured consistent, scheduled execution without any human intervention.

1

Orchestrated Workflow (DAG)

A Directed Acyclic Graph (DAG) was meticulously designed to orchestrate the entire data pipeline. This involves a precise sequence of tasks: data ingestion, followed by `dbt run` for transformations, and finally `dbt test` for rigorous data quality validation.

2

Containerized Environment

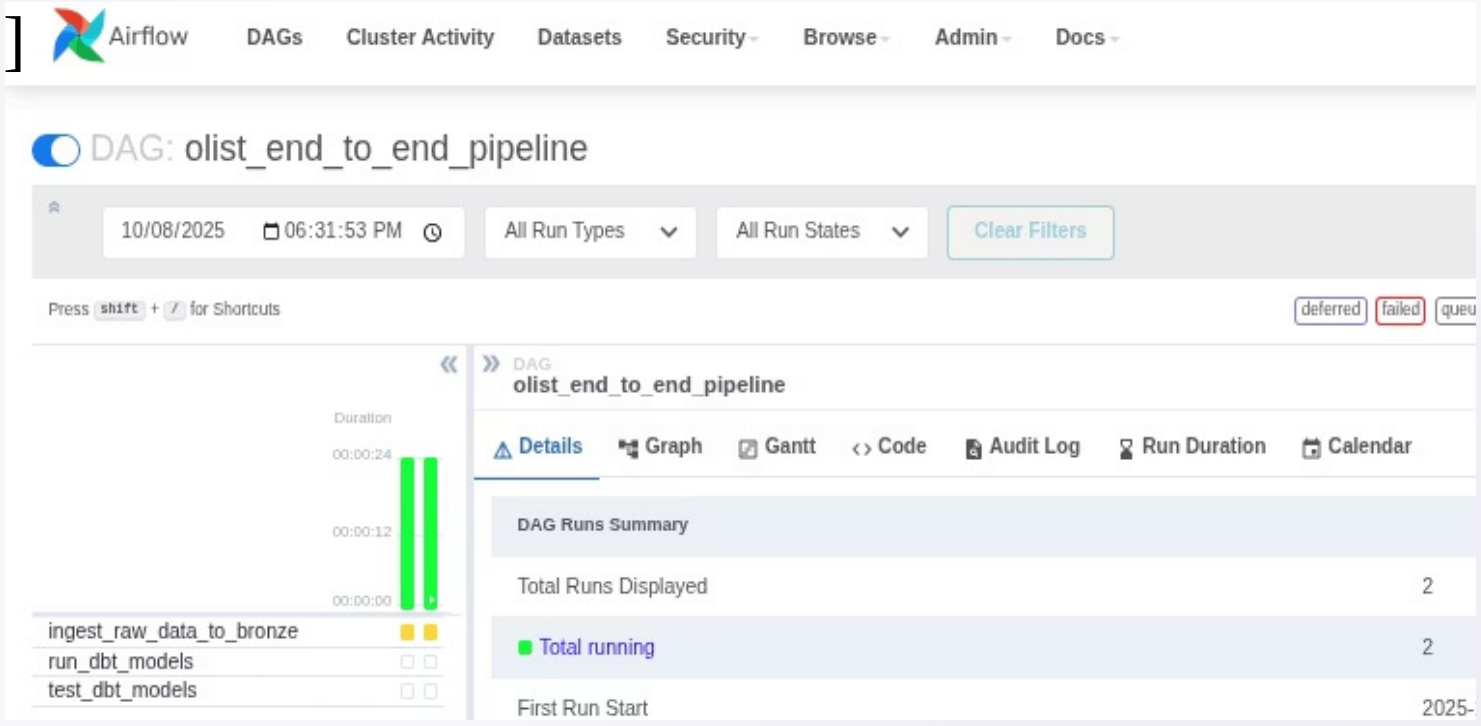
Airflow was containerized using Docker, ensuring a consistent and portable execution environment across all stages of development and deployment, from local testing to production.

3

Reliable & Scheduled Automation

The DAG is configured for daily execution, completely eliminating manual intervention. Critical benefits include automatic failure notifications, comprehensive logging for debugging, and robust retry mechanisms, ensuring continuous data availability and reliability.

Step 5 Screenshots



Step 6: Delivering Insights with Power BI

The final stage of our journey culminates in the Power BI Executive Dashboard, where rigorously transformed and governed data from the **GOLD** layer in Snowflake is converted into real-time, actionable insights for stakeholders.

Real-time KPIs & Business Intelligence

The dashboard dynamically presents critical Key Performance Indicators (KPIs) across sales, customer satisfaction, and operational efficiency, leveraging advanced business intelligence features to provide live data-driven insights.

Interactive & Intuitive Design

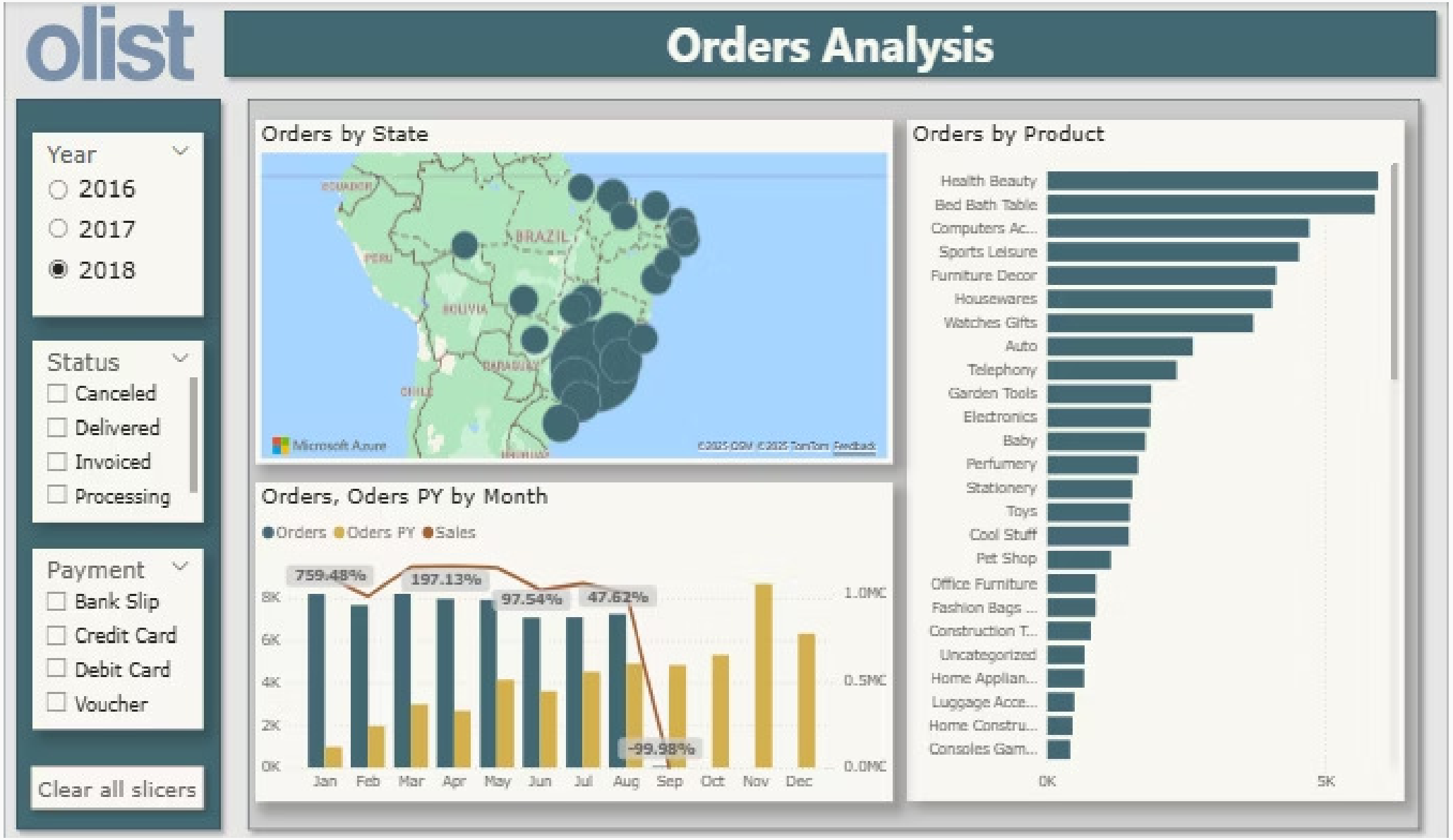
Its interactive single-page layout empowers executives to swiftly filter, drill down, and explore data with ease, facilitating deep analytical exploration tailored to their specific needs.

Data Trust & Reliability

By connecting exclusively to the tested and verified GOLD layer, we ensure that every displayed metric is reliable, trustworthy, and validated, fostering confidence in business decisions.

Step 6 Screenshots

Step 6 Screenshots



Key Business Insights from SQL Analysis

Leveraging rigorously cleaned and transformed data from our GOLD layer, our SQL analysis has uncovered critical business insights, enabling data-driven strategic decisions. These findings highlight the power of reliable data in revealing actionable patterns and opportunities.

Insight: Revenue Concentration

Our analysis clearly shows that the vast majority of revenue (approx. **70%**) and order volume is heavily concentrated in Brazil's southeastern states, particularly São Paulo (SP), which alone accounts for over **45%** of total sales.

Implications

This insight indicates a need to optimize logistics and marketing spend within this core region. It also identifies significant untapped potential in other regions, which can be targeted with localized growth strategies backed by data.

Key Business Insights: Delivery Performance Analysis

Delivery Impact on Customer Satisfaction

Insight: Delivery & Satisfaction Correlation

Strong negative correlation (r = -0.75) between delivery time and customer satisfaction scores. Notable drop for deliveries exceeding 3 days.

Implications

Emphasizes critical role of efficient logistics in customer retention and brand perception. Optimizing delivery routes could boost satisfaction and loyalty.

```sql
-- Insight 2: Delivery & Satisfaction Correlation (Final Corrected Version)
WITH OrderReviews AS ( SELECT "order_id" AS ORDER_ID,"review_score" AS REVIEW_SCORE
    FROM PROJECT_DB.BRONZE_RAW_DATA.RAW_ORDER_REVIEWS
    QUALIFY ROW_NUMBER() OVER(PARTITION BY "order_id" ORDER BY "review_creation_date" DESC) = 1
)

SELECT O.IS_DELAYED,
    CASE
        WHEN O.IS_DELAYED = TRUE THEN 'Delayed'
        ELSE 'On-Time'
    END AS DELIVERY_STATUS,COUNT(DISTINCT O.ORDER_KEY) AS TOTAL_ORDERS,AVG(R.REVIEW_SCORE) AS AVERAGE_REVIEW_SCORE
FROM DIM_ORDERS AS O
JOIN OrderReviews AS R ON O.ORDER_KEY = R.ORDER_ID
WHERE O.DELIVERY_DAYS IS NOT NULL
GROUP BY O.IS_DELAYED
ORDER BY O.IS_DELAYED;
```

Results (just now)

Table | Chart | 2 rows | 84ms

	0|1 IS_DELAYED	DELIVERY_STATUS	TOTAL_ORDERS	AVERAGE_REVIEW_SCORE
1	FALSE	On-Time	88168	4.293939
2	TRUE	Delayed	7662	2.565388

Key Learnings & Project Summary

Technical Learnings

Configuration Mastery

Deep dive into tool-specific configurations, such as understanding Snowflake's case sensitivity with dbt, proved essential for robust data integration.

Secure Environment Management

Effective management of environment variables and secrets across Airflow, dbt, and Python scripts was critical for security and operational consistency.

Platform Integration

The value of deeply understanding tool-specific configurations and defaults, especially when integrating multiple platforms, was a key takeaway.

Reliable Automation Principles

The profound impact of building automated, reliable systems where data quality is inherently integrated into the pipeline process, rather than being an afterthought.

Project Summary: End-to-End Achievement

This project successfully delivered an automated, end-to-end data pipeline, transforming raw ingestion into actionable insights. By meticulously addressing technical challenges and applying systematic learning, we established a foundation for reliable data quality, streamlined reporting, and enhanced analytical capabilities. The journey from initial data exploration to robust, production-ready solutions underscored the importance of diligent configuration, continuous integration, and a deep understanding of each component's role in the data ecosystem.

Thank You

Questions?

- **Email**: Yousef.soliman.de@gmail.com
- **LinkedIn**: www.linkedin.com/in/y0usefma7m0ud/
- **GitHub**: https://github.com/Y0U5F