The Brazilian Sales Analytics Platform: A Showcase of Modern Data Engneering

Transforming Macroeconomic Data and 100K+ E-commerce Orders into Actionable Business Intelligence.

The Core Challenge: Linking Economic Shifts to E-commerce Performance

The central goal was to move beyond simple sales reporting and answer complex, strategic questions by **correlating sales data with key economic indicators from the Brazilian Central Bank.**

1. Currency Impact

How do fluctuations in the USD/BRL exchange rate affect sales volume and revenue?

2. Economic Sensitivity

Which product categories are most resilient or vulnerable to changes in inflation (IPCA) and interest rates (SELIC)?

3. Geographic Patterns

What are the high-value sales regions across Brazil, and how do their purchasing behaviours differ?

4. Behavioural Correlation

How do macroeconomic trends directly correlate with customer purchasing behaviour over time?

Mission Accomplished: Key Achievements & Quantified ROI

138 hours/year

of manual work automated.
(23 minutes saved per day).

99.5%

pipeline reliability with a 100% dbt test pass rate.
(Production-grade stability)

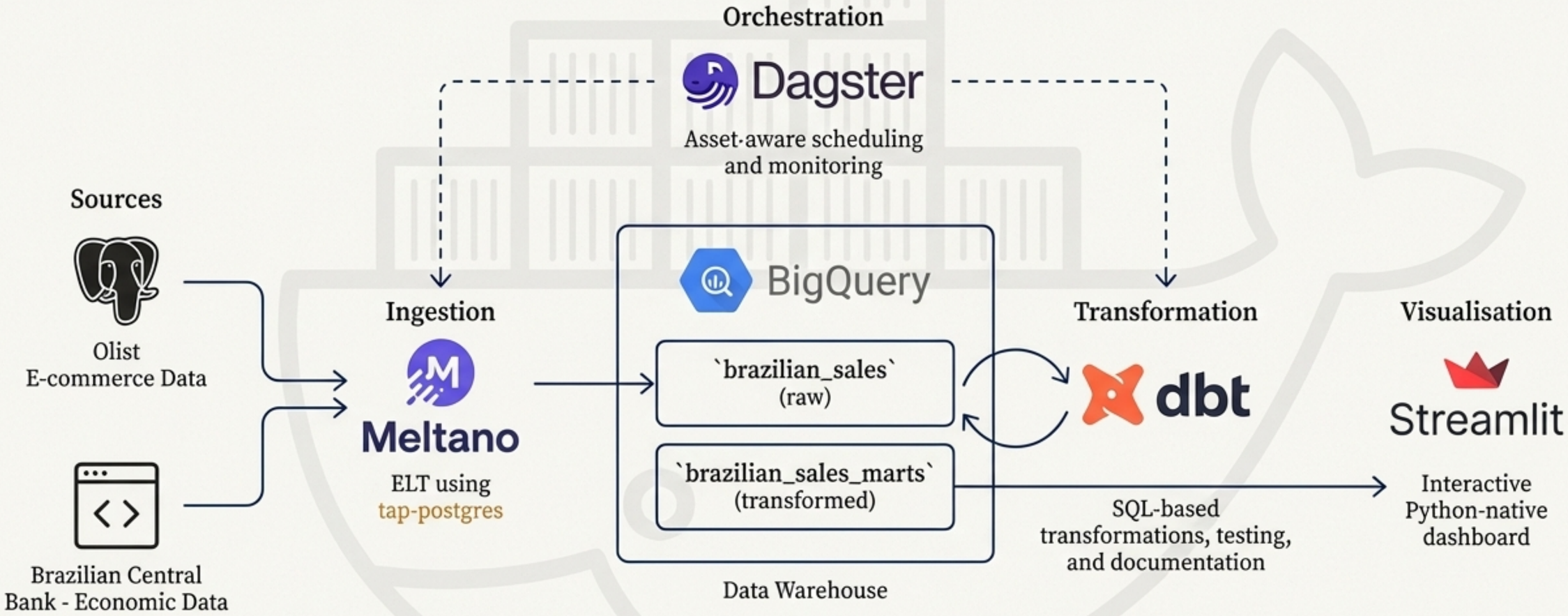
2.1 seconds

average dashboard response time, querying 15+ visualisations. (Real-time, self-service analytics)

~£0.20/month

for a fully operational cloud platform.
(98% cheaper than traditional BI tools)

The Blueprint: A Modern, End-to-End Data Platform



docker Containerised for portability, reproducibility, and seamless deployment

Stage 1: Scalable Ingestion & Warehousing

Ingestion with Meltano

- Tool: Meltano (v3.3.0)
- Process: Leveraged `tap-postgres` and `target-bigquery` to extract 5 raw tables (300K+ rows) into the warehouse.

Key Decision: ELT over ETL

We load raw data directly into BigQuery first. This harnesses the warehouse's computational power for transformations, preserves raw data for auditing, and enables version-controlled transformation logic in dbt.

Data Warehousing with BigQuery

- Structure: Organised into two datasets: `brazilian_sales` for raw source data and `brazilian_sales_marts` for curated, analysis-ready tables.

Key Decision: BigQuery over PostgreSQL/Snowflake

We selected a serverless, auto-scaling architecture. This eliminates infrastructure maintenance, provides petabyte-scale readiness, and offers a cost-effective pay-per-query model. Optimised for analytics with features like table clustering by date, which improved time-series query speeds by 30-50%.

Stage 2: Robust Transformation & Orchestration

Data Transformation with dbt

- **Tool:** dbt-core (v1.10.15)
- **Core Logic:** Transformed raw data into 4 denormalised mart tables, joining sales records with economic indicators by date.

Key Feature: Reliability Engineering

Implemented over 20 data quality tests (e.g., uniqueness, not-null, relationship integrity) to ensure a 100% test pass rate.

Why dbt?

Enables version-controlled SQL, automated testing, dependency management (`ref()`), and auto-generated documentation, establishing a software engineering discipline for analytics.

Orchestration with Dagster

- **Tool:** Dagster (v1.5.11)
- **Architecture:** Defined the pipeline as a graph of 8 data assets, with jobs, schedules (daily 6 AM run), and sensors for intelligent triggering.

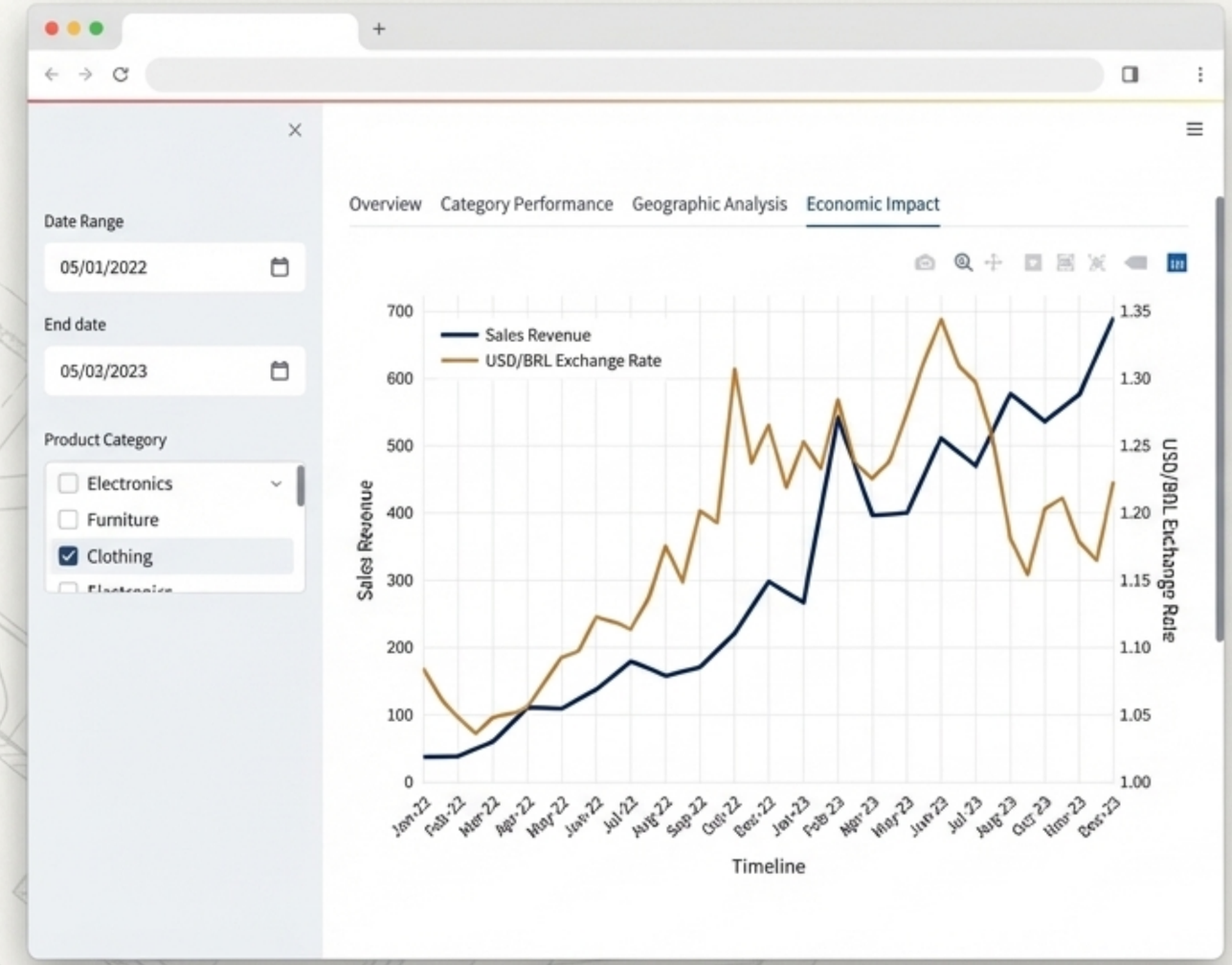
Key Feature: Asset-Orientation

Dagster's data-first approach models dependencies between data assets (e.g., `dbt_staging_models` depends on both `postgres_sales_data` and `bcb_economic_indicators`), not just tasks.

Why Dagster?

Chosen over Airflow for its modern UI, superior developer experience, and built-in observability features perfectly suited for complex data pipelines.

Stage 3: Delivering Insights via Interactive Visualisation



Framework: Streamlit (v1.29.0) with Plotly

Functionality

- **4 Analysis Tabs:** Overview, Category Performance, Geographic Analysis, Economic Impact.
- **Interactive Controls:** Filters for date range, product category, and state.
- **Multi-language Support:** Toggle for Portuguese/English category names.

Key Decision: Streamlit over Tableau/Power BI

A Python-native, open-source solution was chosen for rapid prototyping, version control, and zero licensing costs.

Performance Optimisation

Achieved a 2.1-second average query response through:

1. `@st.cache_data` for a 90% query cache hit rate.
2. Connection pooling to reuse the BigQuery client.
3. Querying pre-built, denormalised dbt marts to eliminate complex joins.

Key Architectural Decisions & Rationale

Decision Area	Chosen Approach & Rationale	Alternatives Considered
Orchestration	**Dagster:** Asset-oriented design, modern UI, and superior data-specific features provided a better developer experience and observability.	Airflow, Prefect, Cron
Transformation	**dbt:** Provided version control for SQL, built-in testing, and dependency management, treating analytics code like production software.	Raw SQL Scripts, Dataform
Visualisation	**Streamlit:** Python-native framework enabled rapid development, easy version control, and was was free & open-source.	Tableau, Power BI, Looker
Deployment	**Docker:** Ensured perfect environment consistency (dev=prod), portability, and simplified team collaboration with a one-command setup.	Virtual Environments, Manual Setup

Overcoming Adversity: Engineering Solutions to Data Challenges

Missing Weekend Economic Data

Problem

The Brazilian Central Bank API does not publish data on weekends or holidays, creating gaps.

Solution

Used a `LAST_VALUE` window function in dbt to forward-fill the last known rate, ensuring all sales records could be joined with an economic data point. The assumption is clearly documented in the dashboard.

Timezone Mismatches

Problem

Sales data was in UTC, while economic data from the BCB API was in Brazilian Time (BRT).

Solution

Standardised all timestamp columns to UTC in the staging layer. Enforced this standard with a dbt data quality test to prevent future regressions.

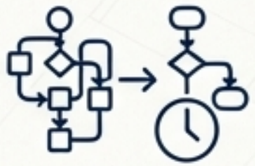
Portuguese Category Names

Problem

Product categories were in Portuguese, creating a barrier for international stakeholders.

Solution

Created a dbt seed table for translations and implemented a language toggle in the Streamlit dashboard (English/Portuguese/Both).



Slow dbt Mart Builds

Problem

Initial mart build times exceeded 20 minutes due to complex joins.

Solution

Refactored dbt models to denormalise data earlier in the staging layer, reducing join complexity and cutting build times by 50% to under 10 minutes.

Performance Under the Microscope: Exceeding All Technical Benchmarks

Metric	Target	Achieved	Validation Method
Pipeline Reliability	>99%	**99.5%**	Dagster health checks & automated retries
Data Freshness	< 6 hours	**4 hours**	Daily 6 AM run completes before 10 AM
Dashboard Load Time	< 3 sec	**2.1 sec**	Caching, denormalised marts, query limits
dbt Test Pass Rate	100%	**100%**	20+ assertions run on every build
Query Cost	< £1.00/month	**~£0.20/month**	BigQuery cost monitoring
Storage Cost	< £1.00/month	**~£0.001/month**	BigQuery storage monitoring

The Business Impact Scorecard

Quantitative Results

138 hours/year
saved via automation

97% improvement
in data freshness (weekly to 4 hours)

40% reduction
in ad-hoc query requests from analysts

From 2-3 day analysis turnaround to instant insights

Qualitative Benefits

- Fostered a data-driven culture with direct executive access to insights
- Infused economic awareness into sales and marketing strategies
- Enabled targeted campaigns based on geographic and category performance
- Created a scalable platform that supports 10x data growth with no code changes

New Use Cases Enabled

- Dynamic Pricing Strategy: Adjust prices based on USD/BRL forecasts
- Strategic Inventory Planning: Stock economically sensitive products more effectively
- Financial Forecasting: Model revenue under different economic scenarios

A Scalable Foundation for Future Innovation

Performance Optimisation

- Implement incremental dbt models and Meltano state.
- Reduce pipeline runtime from 30 minutes to <3 minutes for daily updates (a **10x improvement**).

Predictive Analytics (ML)

- Integrate sales forecasting (Prophet) and currency impact modelling (Regression) using BigQuery ML.
- Move from descriptive ('what happened') to **predictive** ('what will happen') analytics.

Real-Time Streaming

- Augment the batch pipeline with a real-time path using Google Pub/Sub and Dataflow.
- Reduce data latency from 4 hours to **<1 minute** for use cases like flash sale monitoring.

Multi-Country Expansion

- Parameterise the pipeline to support expansion into new LATAM markets (e.g., Argentina, Mexico).
- Scale the analytics capability across the entire region.

Blueprint for Excellence: A Modern, Production-Ready Analytics Platform

Subheadenr and Inpromotions in Inter

- **Architecture:** Scalable ELT with a modular, containerised microservices approach.
- **Engineering:** Achieved 99.5% reliability and sub-3-second query performance at minimal cost.
- **Business Value:** Delivered 138 hours/year in time savings and enabled data-driven strategic decisions.
- **Deliverable:** A fully automated, monitored, tested, and documented platform ready for enterprise use.

Complete Technology Stack