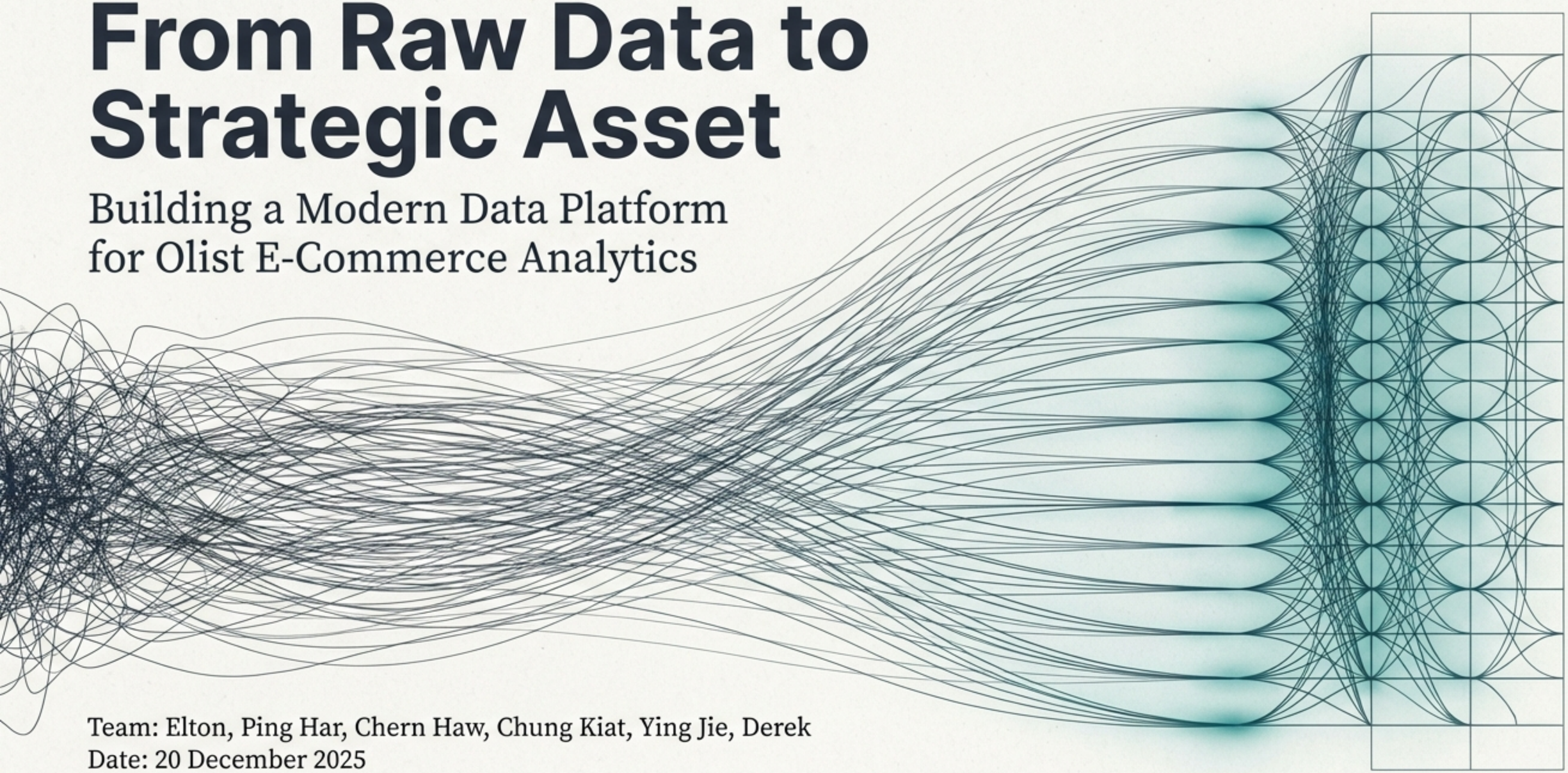
From Raw Data to Strategic Asset

Building a Modern Data Platform for Olist E-Commerce Analytics



Team: Elton, Ping Har, Chern Haw, Chung Kiat, Ying Jie, Derek

Date: 20 December 2025

NotebookLM

The Opportunity: Turning Dormant Data into a Competitive Edge

Problem

Olist's vast e-commerce data was disconnected and difficult to analyze, hindering strategic decision-making and creating "analysis paralysis."

Solution

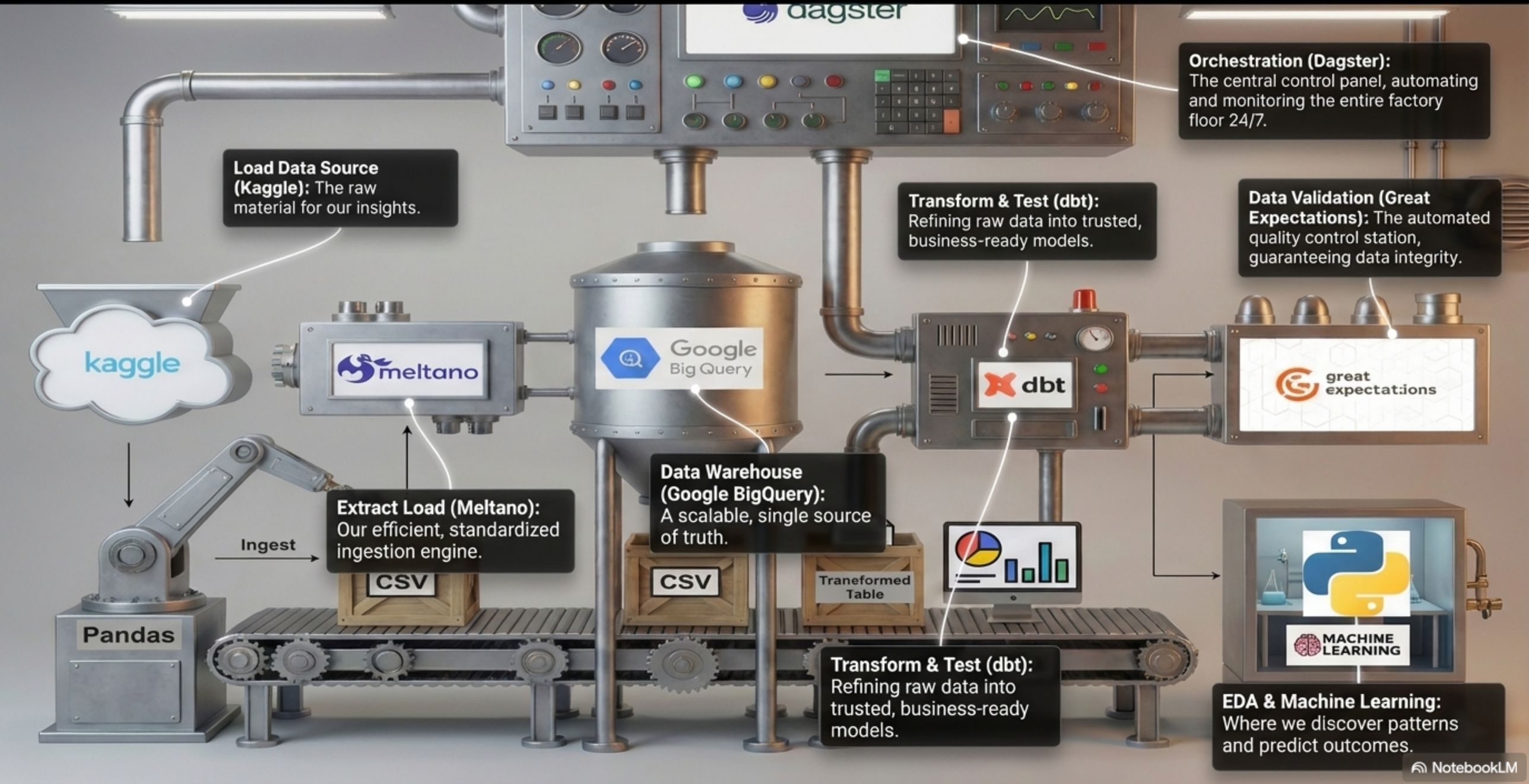
We engineered an automated, end-to-end modern data platform to transform raw data into a reliable, accessible, and predictive strategic asset.

Business Impact

Empower data-driven decisions across the organization, uncover hidden growth opportunities, and significantly improve operational efficiency.

NotebookLM

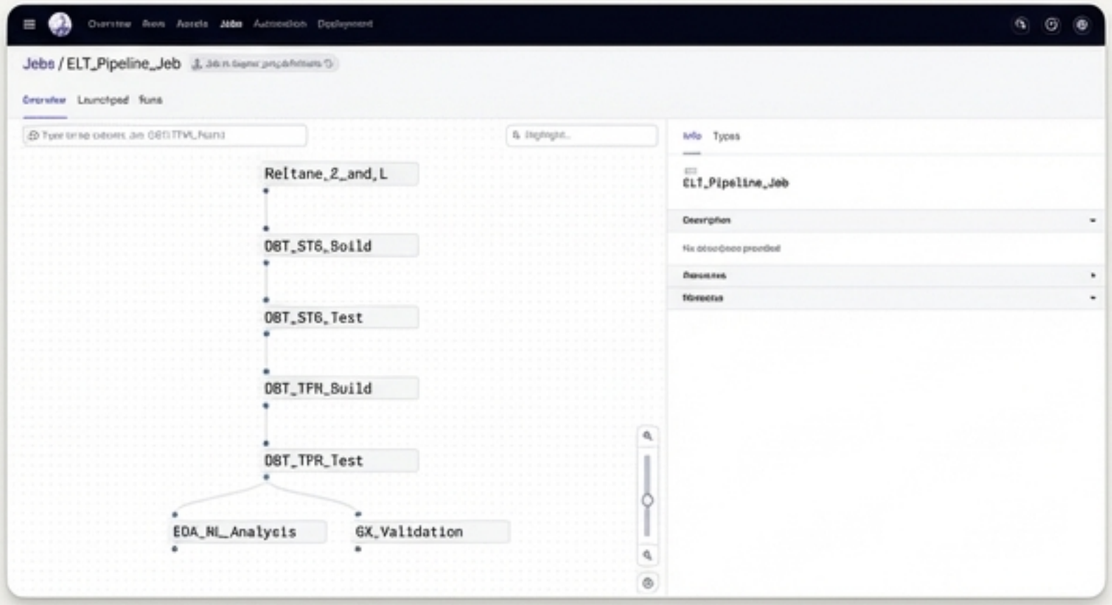
The Blueprint: Our Automated 'Insights Factory'



Full Automation and Visibility with Dagster

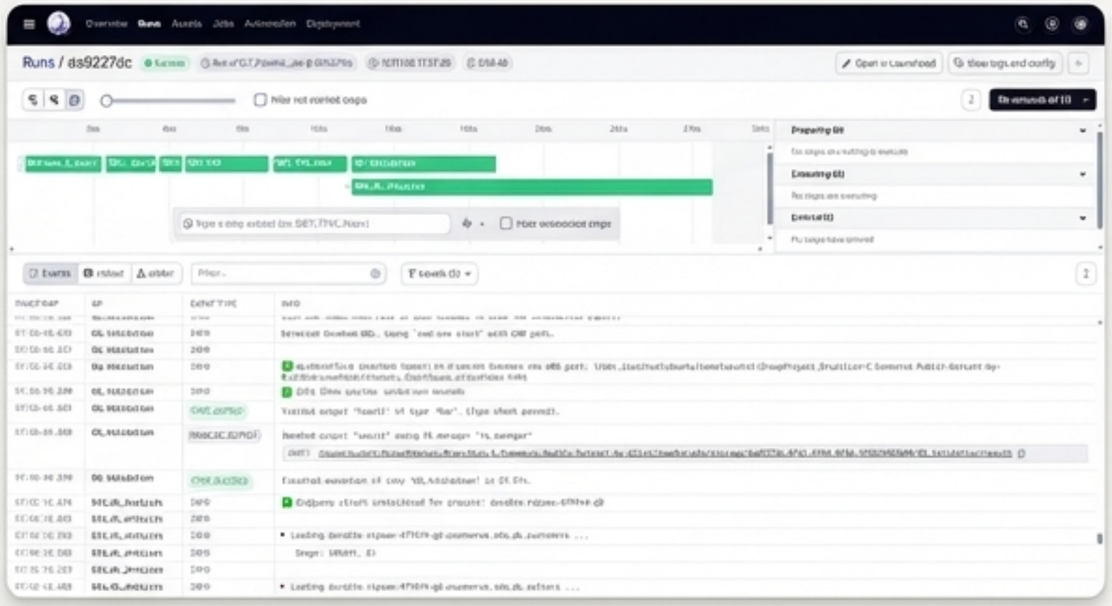
Our platform runs itself. It's an automated system that delivers fresh, reliable data daily without manual intervention, providing complete transparency into the process.

The Workflow Blueprint



The defined job ensures every step, from data loading to validation, is executed in the correct sequence, every time.

Real-Time Production Monitoring



We have a live view into every pipeline run, confirming success and enabling rapid troubleshooting if issues arise.

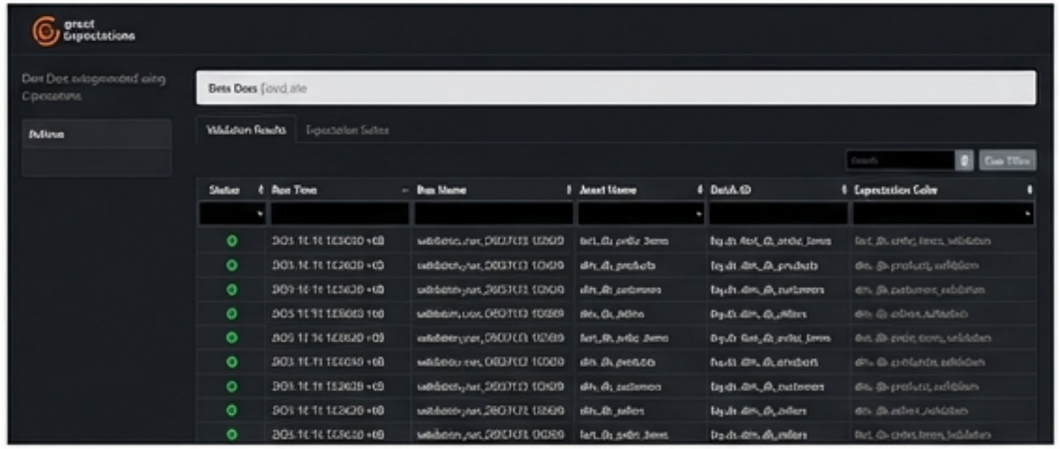
Key Takeaway: Reliability & Efficiency.
Reduced manual errors, faster data delivery, and a fully observable, auditable process.

NotebookLM

Building Unshakeable Trust in Our Data with Great Expectations

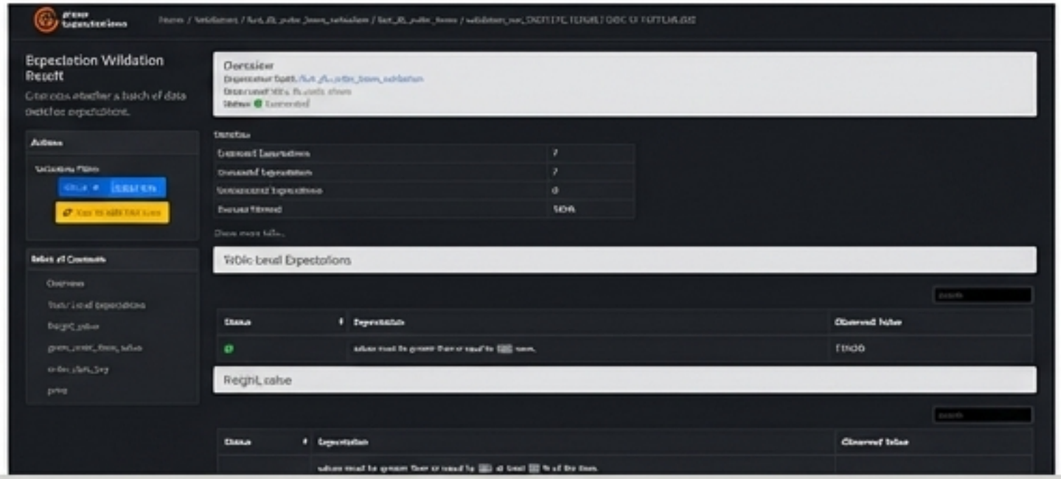
We don't just move data; we validate it. Every piece of data entering our analytics environment is automatically checked against business rules, ensuring every decision is based on a foundation of truth.

Comprehensive Validation Suites



Automated test suites run against all core data tables (customers, products, orders) with every pipeline execution.

Successful Quality Control Results



We enforce specific business rules—like checking for valid prices (`freight_value >= 0.0`) and ensuring data completeness—and consistently achieve 100% success.

Key Takeaway: Data Integrity & Governance.
Proactive error detection, increased confidence in reporting, and trustworthy analytics.

NotebookLM

The Platform in Action

We will now switch to a **brief live demonstration** of the **Dagster** and **Great Expectations** user interfaces to showcase the platform's automation and data validation capabilities.

NotebookLM

From Data to Discovery: Key Analytical Findings

The platform is already revealing critical insights into our customers, products, and operations.

Geographic Concentration: The top 5 states, led by São Paulo, drive 75% of our customer base, highlighting key markets for targeted campaigns.

High-Value Products: Top categories like Bed & Bath, Health & Beauty, and Sports & Leisure consistently generate the most revenue, informing product strategy.

Sales & Revenue Trends: We've identified clear seasonality in orders and revenue, with peaks in May, August, August, and November, enabling better inventory and marketing planning.

Seller Distribution: A high concentration of sellers in the Southeast presents a strategic opportunity to expand our seller network in other regions.

NotebookLM

Beyond Reporting: Predicting Order Delivery Times

We’ve moved from reactive analysis to proactive prediction. Our first machine learning model provides a baseline for forecasting order delivery times, a critical factor in customer satisfaction.

The Goal

To predict the delivery duration of an order based on factors like origin, destination, product type, and freight value. This helps manage customer expectations and optimize logistics.

The Baseline Performance

Mean Absolute Error (MAE): 8.52 Days

On average, our prediction is off by ~8.5 days.

R-squared (R^2): 0.255

The model currently explains 25.5% of the variance in delivery times, a solid foundation for future improvement.

Key Takeaway: We have established a foundational predictive capability. The next step is to refine this model to reduce error and enhance its business value.

NotebookLM

Empowering the Business with Self-Service Analytics

We will now demonstrate the interactive dashboard, which puts the power of this data directly into the hands of our business leaders.

Key Metrics Available:

- Customer Growth by Year
- Geographic Customer & Seller Density Maps
- Revenue & Freight Cost Trends
- Top Product Categories & Sellers by Revenue

Live Dashboard:
https://pinghar.github.io/Brazilian-E-Commerce-Public-Dataset-by-Olist/

NotebookLM

Strategic Learnings from Our Journey

The Modern Stack Delivers

Our chosen tools (Meltano, dbt, Dagster, GX) create a powerful, scalable, and cost-effective platform that is greater than the sum of its parts.

Automation is Non-Negotiable

Automating the pipeline and data quality checks is the only way to build a scalable foundation of trust and operational efficiency.

Data Quality is a Continuous Process

Great Expectations provides the framework, but true data quality requires ongoing partnership with the business to refine and expand rules.

This is a Foundation, Not a Destination

The platform's true value is unlocked by the strategic questions the business can now confidently ask and answer.

NotebookLM

Proactively Managing Risks

Risk	Mitigation Strategy
Data Scalability & Cost Data volume will increase with business growth, potentially impacting performance and cloud costs.	Leverage Google BigQuery's native auto-scaling; implement rigorous cost monitoring and optimization routines within our dbt models.
ML Model Drift The predictive accuracy of our models may degrade over time as market dynamics change.	Establish a continuous MLOps monitoring process for model performance and schedule regular, automated retraining cycles.
Upstream Data Quality "Garbage in, garbage out." Data quality issues at the source can still impact the pipeline.	Partner with source system owners to improve data entry standards; expand our Great Expectations suites to catch more edge cases proactively.

NotebookLM

From Platform to Profit: Our Recommendations

1.

Drive Business Adoption & Enablement

* **Action:** Roll out dashboard access and targeted training to Marketing, Sales, and Operations teams to embed data-driven decision-making into their daily workflows.

2.

Enhance Predictive Capabilities

* **Action:** Invest in further feature engineering to improve our delivery time model. Launch new initiatives for customer churn prediction and demand forecasting.

3.

Create a 360° Customer View

* **Action:** Expand the platform by integrating new, high-value data sources, such as marketing campaign data (e.g., Google Analytics) and customer support logs.

NotebookLM

Q&A

Thank you.

Presented by:
Elton, Ping Har, Chern Haw, Chung Kiat, Ying Jie, Derek

NotebookLM