Olist

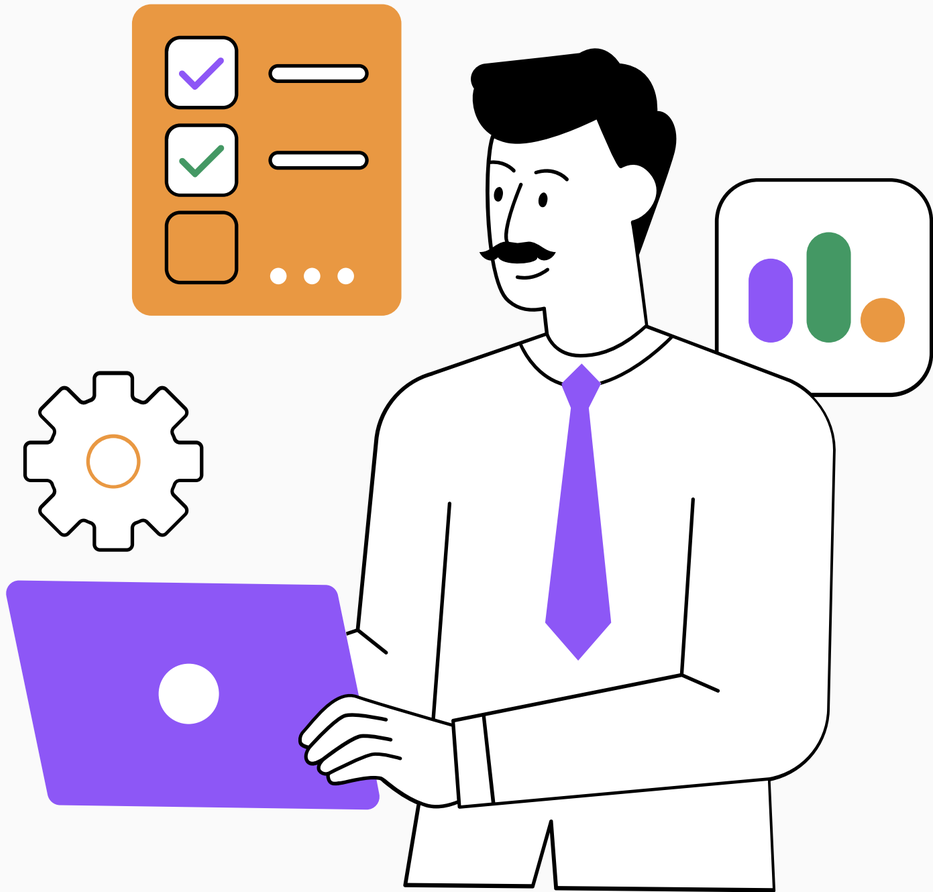
Brazilian E-Commerce Public Dataset

Team 1

- Foo Suan Joo
- Ruby Ferdianto
- Lim Vuthy (Tony)
- Alwyn Quek
- Tan Zhi Hao
- Jeff Lee

Agenda

1. Executive Summary
    a. High-level analysis purpose and focus
    b. Problem statement and scope
    c. Performance insights methodology
2. Methodology & Data
    a. Dataset description (Olist e-commerce data)
    b. Approach and assumptions
    c. Data description and exclusions
3. Implementation/Recommendations & Technical Details
4. Analysis, Results & Insights (Sales/Marketing)
    a. Key metrics & Insights
    b. Strategic Recommendations
5. Analysis, Results & Insights (Operations)
    a. Key metrics & Insights
    b. Strategic Recommendations
6. Conclusion
    a. Summary and strategic implications



Agenda

1. Executive Summary (1.2mins)
- High-level analysis purpose and focus

1.1. Introduction & Context
- Problem statement and scope
- Performance insights methodology

2. Methodology & Data (How we approached it) (4.6mins)
- Dataset description (Olist e-commerce data)
- Approach and assumptions
- Data description and exclusions

3. Implementation/Recommendations (What we do about it) & Technical Details (How it works) (4mins)

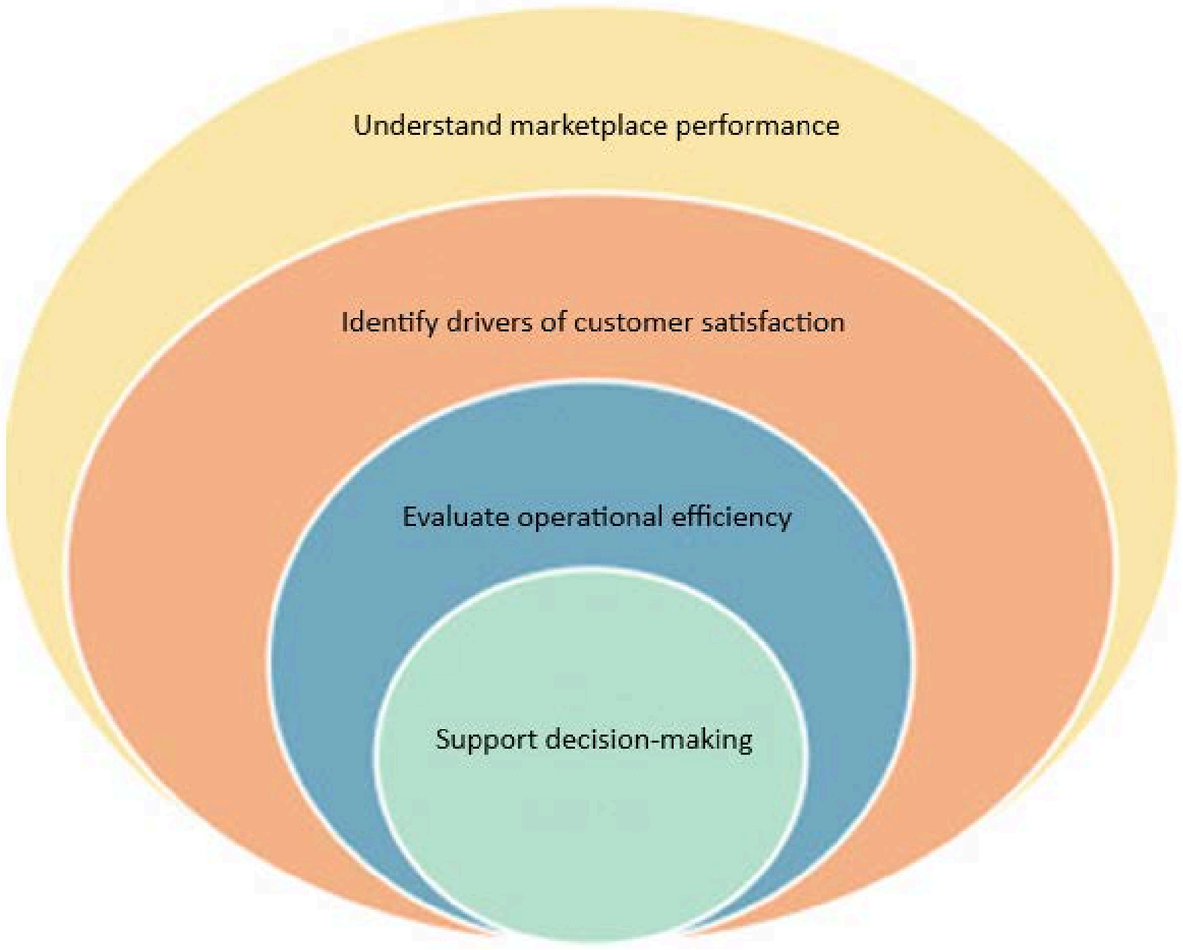
4. Analysis and findings. Results & Insights for Sales/Mktg (2mins)
- Key metrics dashboard
- Insights
- Strategic Recommendations - Actionable recommendations based on findings

5. Analysis and findings. Results & Insights for Operations (3mins)
- Key metrics dashboard
- Insights
- Strategic Recommendations - Actionable recommendations based on findings

6. Conclusion (1.2mins)
- Summary and strategic implications

1. Executive Summary

High-level analysis purpose and focus



Purpose

Focus

Introduction & Context

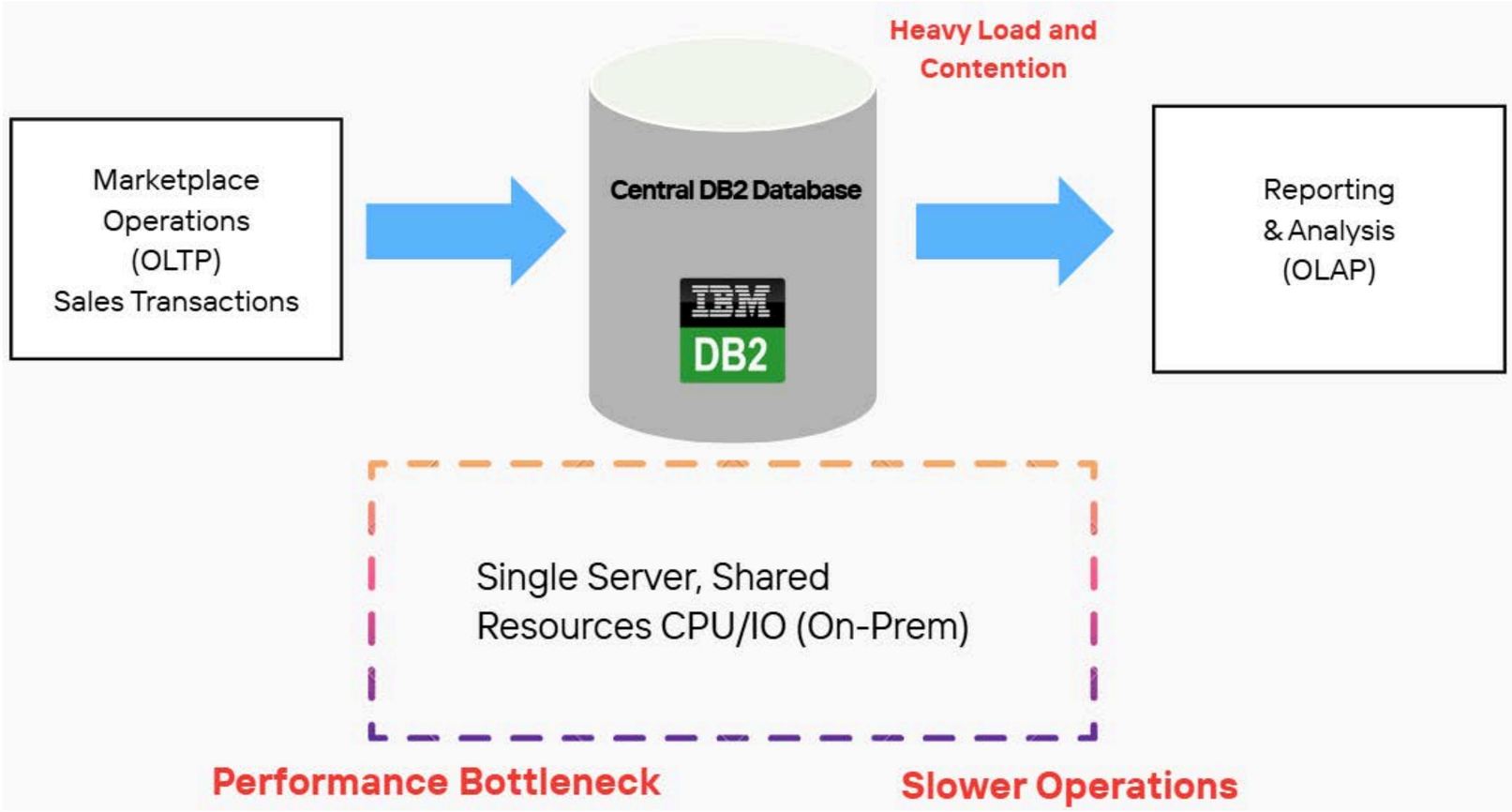
Problem Statement

Olist, a Brazilian e-commerce marketplace, needs to understand its struggles and optimize:

- Customer Experience (Customer Satisfaction)
- Seller Contributions (Uneven Seller Performance
- Operational Inefficiencies (Logistical issues)

Current Legacy System

Why are we doing this? What are we trying to do?
For data scalability, faster reporting and dashboarding

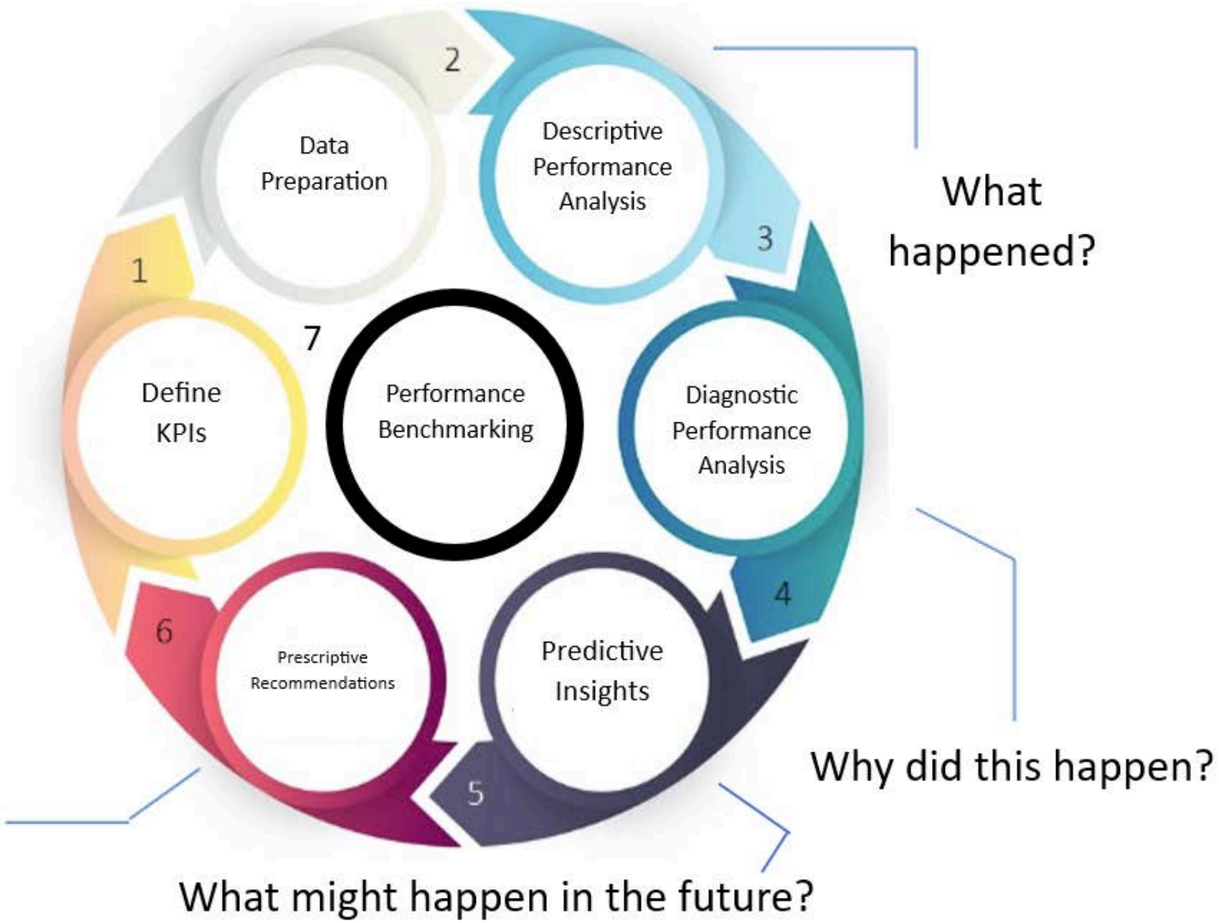


Introduction & Context

Scope of Analysis:

01	02	03	04
Customer Perspective	Seller & Product Perspective	Operational / Marketplace Perspective	Predictive & Prescriptive Insights (Time limitations)

Performance insights methodology



2. Methodology & Data

(How we approached it)

Methodology & Data (How we approached it)

- Understanding business objectives/problems and its' impact
- Our choice of the tools e.g. why did you choose Dagster over Airflow
- Performing Data quality checks - Ensuring quality data for transformation
- Ensuring an end-to-end Data Pipline (Staging → Warehouse layer → Analytical Layer)
  - ERD diagram (how data is linked and what its purpose)
- Architecture (Structures and tools we need to support the data pipeline)
- Orchestration – automating the processing
- What we will do/how we will resolve when the data pipeline breaks
- What are the business insights and business values

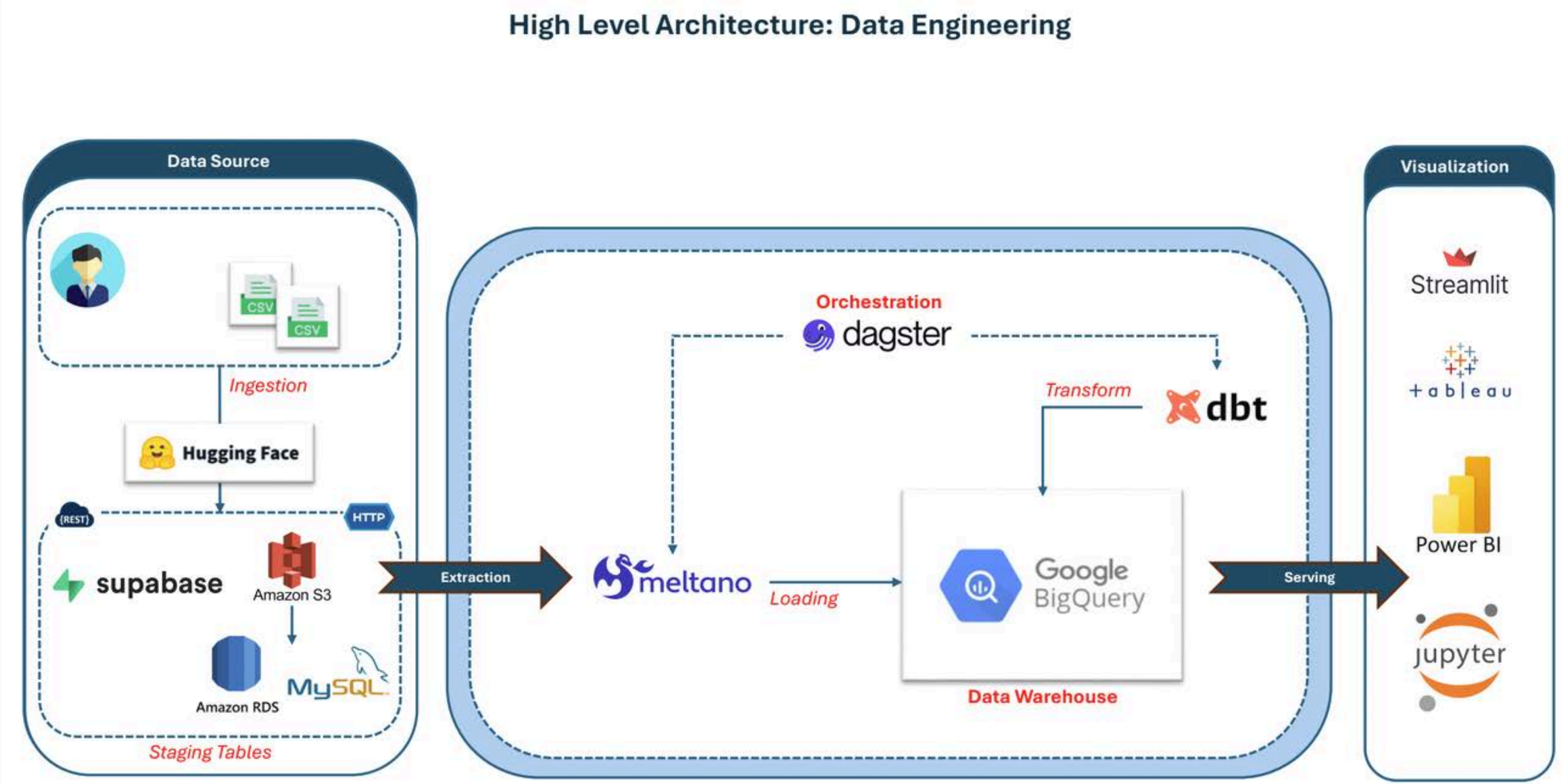
How are we going to achieve it?

• Replace DB2 on prem with Supabase, a managed Postgres service, for the operational system of record.

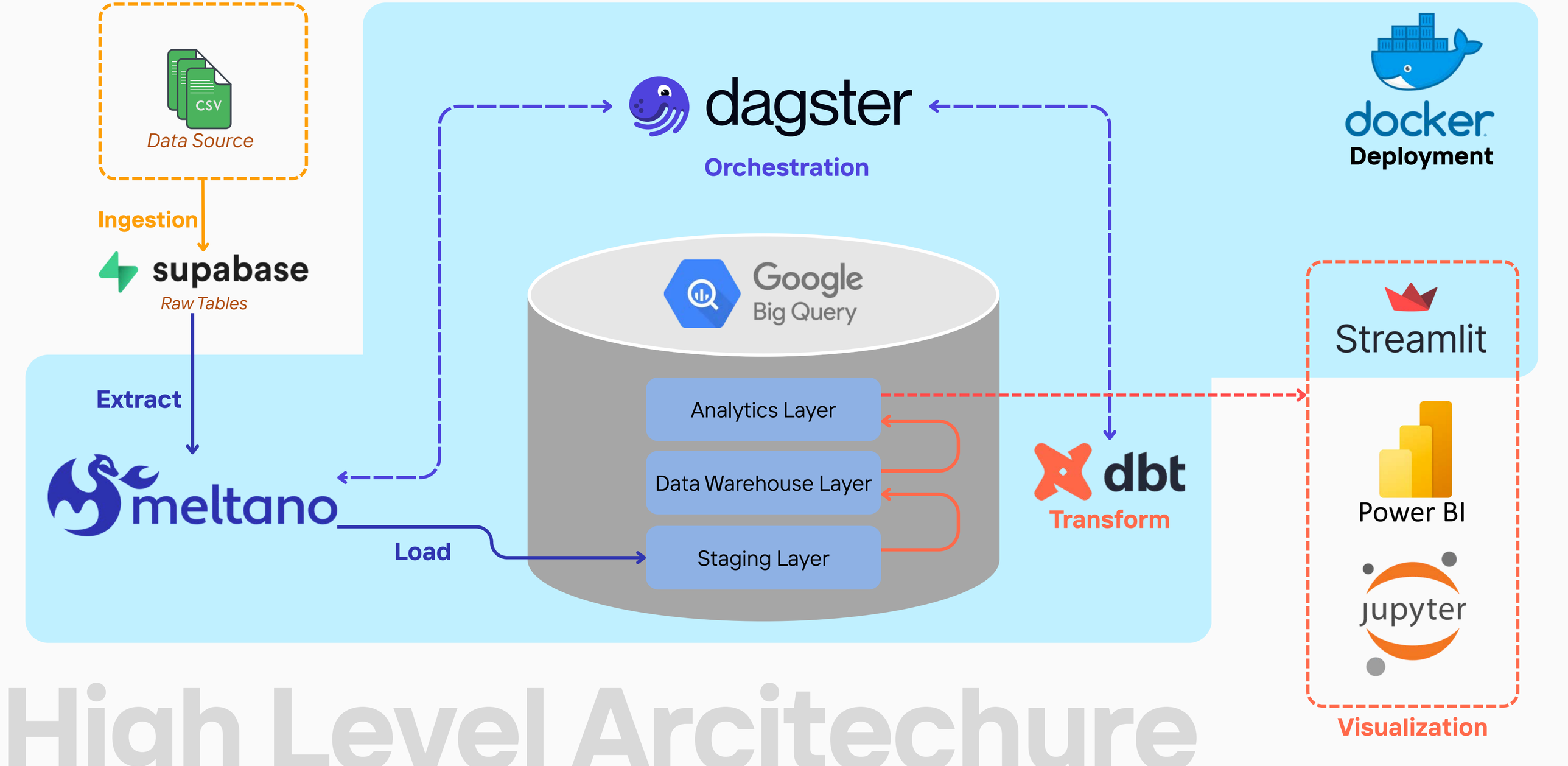
• Set up a separate system, GCP, for reporting (OLAP) so analytics no longer slow down daily operations.

• Build a dimensional data warehouse in BigQuery to serve reporting needs, following the Kimball approach (Staging → Warehouse layer → Analytical Layer).

• Use Meltano to extract and load data, dbt to transform and document the models, and Dagster to orchestrate the full pipeline with scheduling and monitoring.



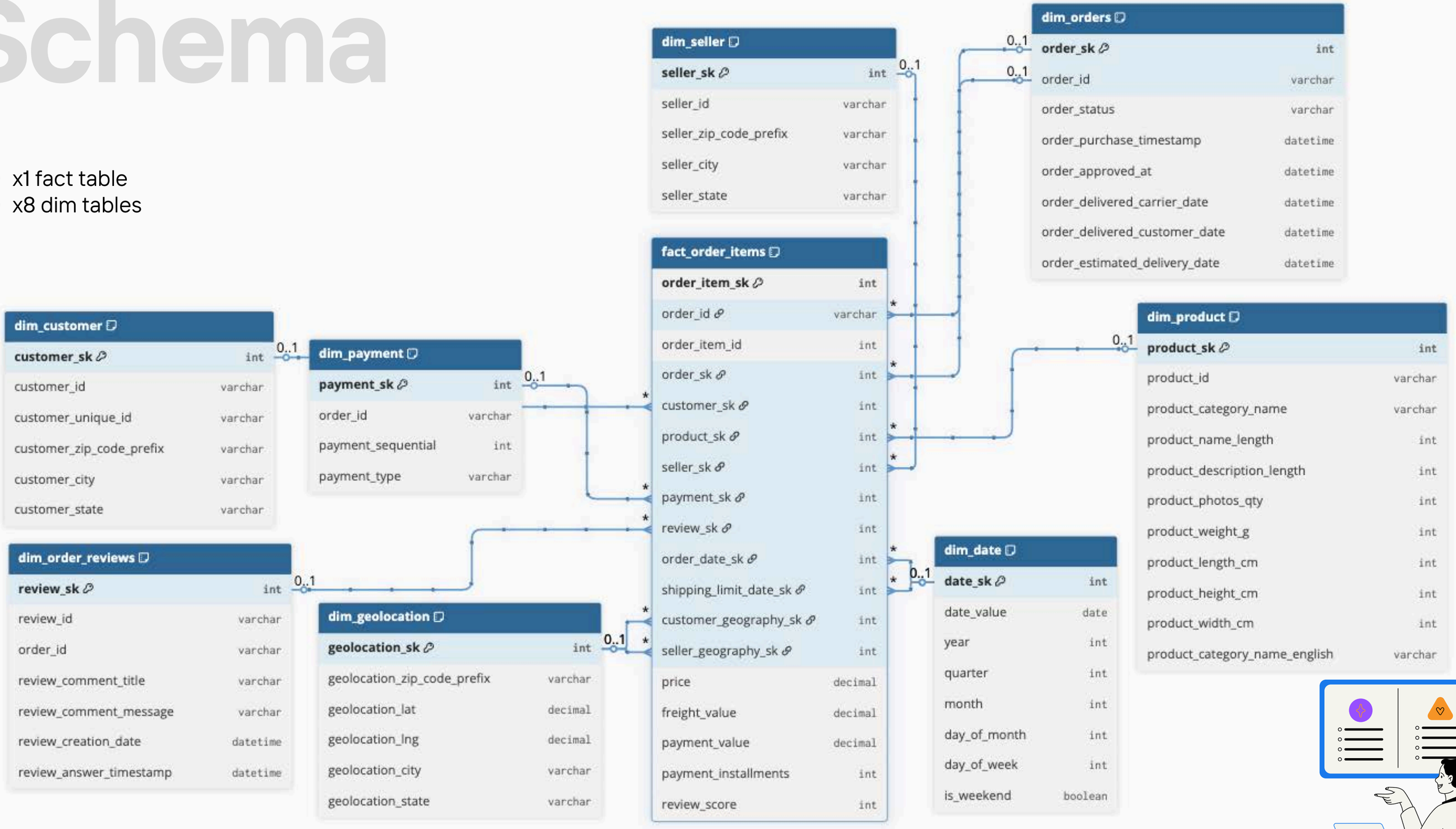
High Level Arcitechure



High Level Arcitechure

Schema

- x1 fact table
- x8 dim tables



Analytics Layer

orders_analytics_obt	
order_sk	varchar
order_id	varchar
order_date	date
customer_state	varchar
order_status	varchar
total_items	int
total_item_value	float
total_order_value	float
delivery_performance	varchar
order_complexity	varchar
customer_order_behavior	varchar
avg_review_score	float
total_fulfillment_days	int

revenue_analytics_obt	
revenue_sk	varchar
order_date	date
customer_state	varchar
item_price	float
allocated_payment	float
market_segment	varchar
shipping_complexity	varchar

seller_analytics_obt	
seller_sk	varchar
performance_tier	varchar
seller_segment	varchar
total_revenue	float

customer_analytics_obt	
customer_sk	varchar
customer_segment	varchar
total_spent	float
churn_risk_level	varchar
satisfaction_tier	varchar

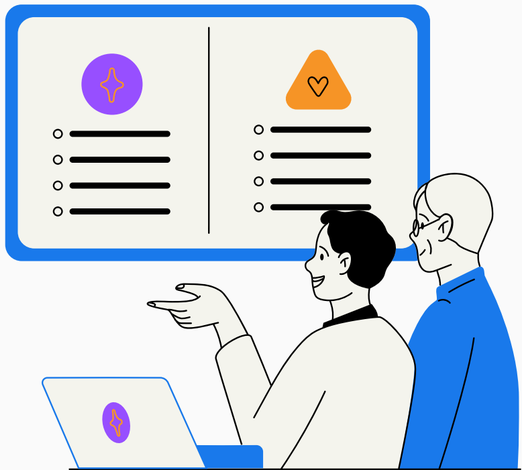
operation_analytics_obt	
order_id	varchar NN
order_item_id	int NN
order_status	varchar
order_purchase_timestamp	timestamp NN
order_approved_at	timestamp NN
order_delivered_carrier_date	timestamp NN
order_delivered_customer_date	timestamp NN
order_estimated_delivery_date	timestamp NN
approval_days	int
handling_days	int
in_transit_days	int
total_delivery_days	int
edd_horizon_days	int
late_to_edd_flag	bool NN
edd_delta_days	int
early_days	int
days_late_to_edd	int
price	numeric(12,2)
freight_value	numeric(12,2)
product_category_name	varchar
product_category_name_english	varchar

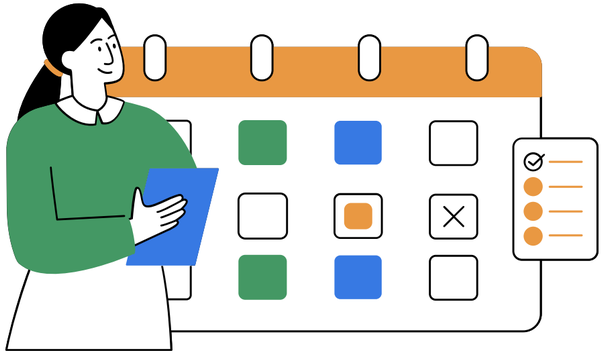
product_weight_g	int
product_length_cm	int
product_height_cm	int
product_width_cm	int
product_volume_liters	float
product_density	float
customer_state	varchar
customer_city	varchar
seller_state	varchar
seller_city	varchar
route_distance_km	float
route_type	varchar
order_year	int
order_month	int
order_dow	int
year_month	varchar
performance_category	varchar

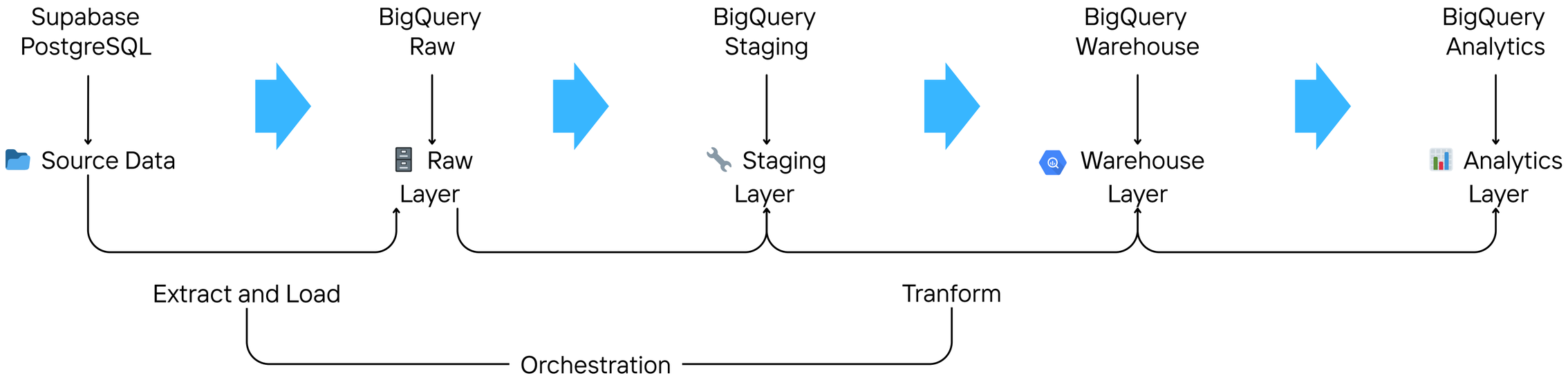
delivery_analytics_obt	
delivery_transaction_sk	varchar
delivery_performance_tier	varchar
delivery_speed_tier	varchar
delivery_accuracy_tier	varchar
order_status	varchar

payment_analytics_obt	
payment_transaction_sk	varchar
payment_type	varchar
installment_category	varchar
payment_risk_level	varchar
customer_payment_profile	varchar

geographic_analytics_obt	
state_code	varchar
market_development_tier	varchar
geographic_region	varchar
market_tier	varchar

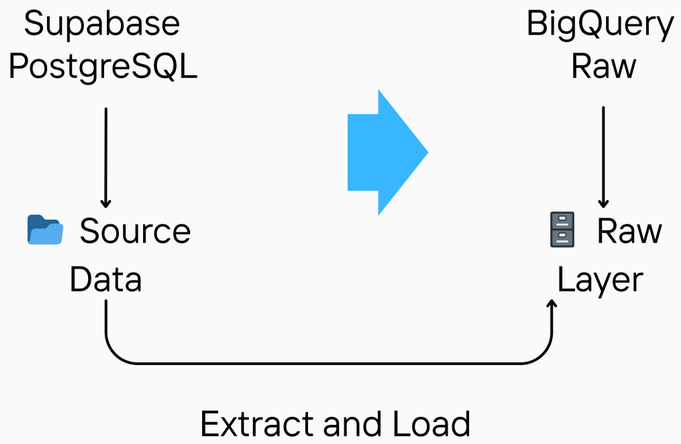






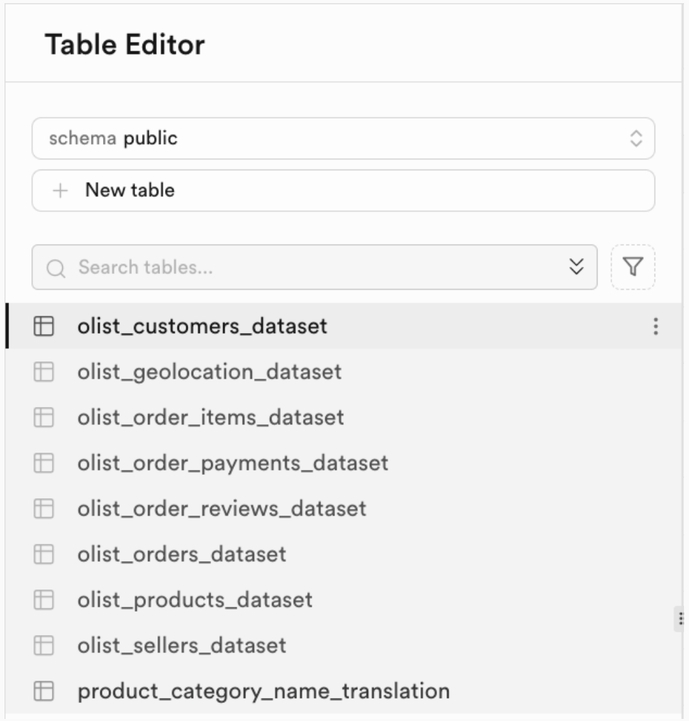
Data Pipeline

supabase



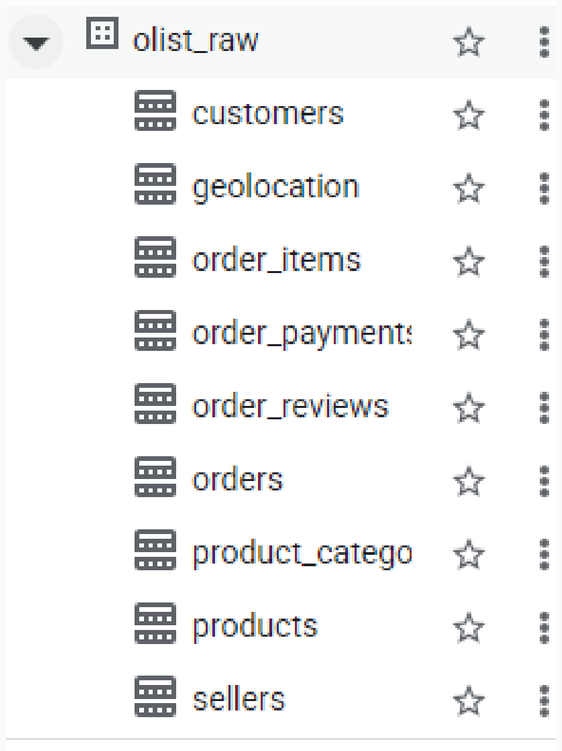
Google Big Query

Tap



meltano

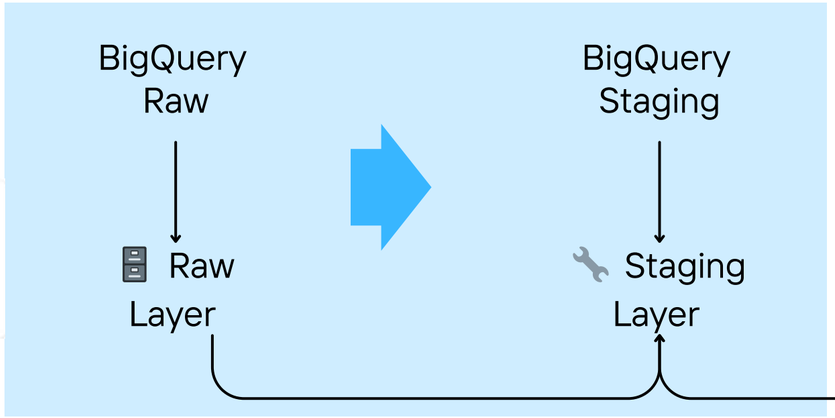
Target



dbt

dbt_utils

dbt_expectations



Staging Data Check

Preliminary Data Check

Text Normalisation
e.g. São Paulo → SAO PAULO

Data Type Casting
e.g. lat/lng to INT,
zip code as STRING

Data Preservation
add "_original" for comparison

Data Validation Flags

Geographic Validation

- valid zip code: Brazillian 5-digit zip code format
- valid lat/lng: bound to Brazil (lat : -35.0 to 5.0 , lng: -75.0 to -30.0
- valid city
- valid state

Business logic Validation

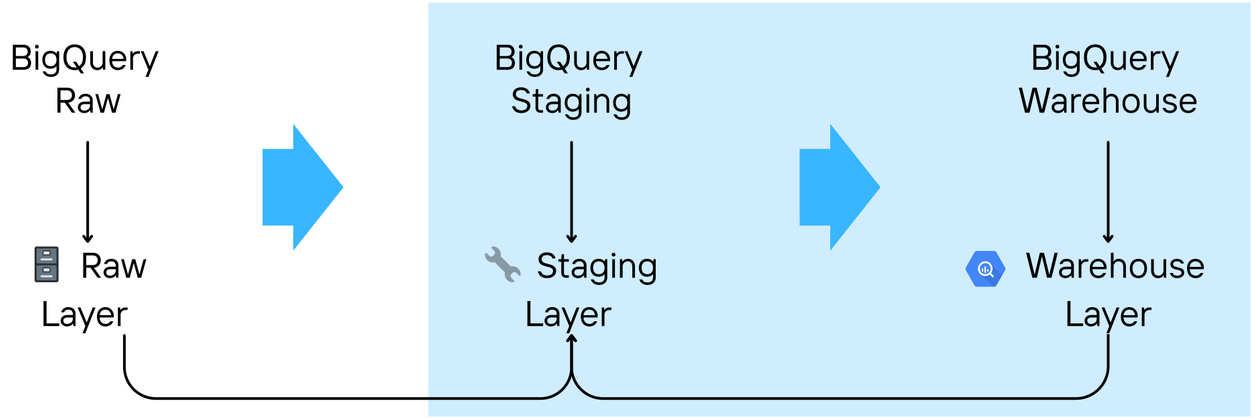
- Temporal logic from order → approval → carrier → customer
- Price/Freight/Payment value range
- Product weight/dimension range
- Payment value reconciliation (voucher payment type)
- Decimal Standardization

Data Quality Check

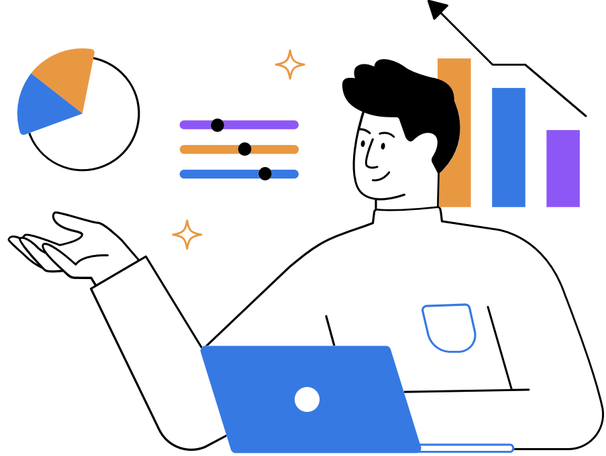
dbt

dbt_utils

dbt_expectations



Warehouse Data Cleaning / Transformation



Transformation

Star Schema

- Fact table
- Dimensional Modeling
- customer_id row de-duplication
- Surrogate key generation

Data Test:

Uniqueness Test

- unique
- not_null

Referential Integrity test

- relationships:
  to: ref('dim_customer')
  field: customer_sk

Geography Consistency Test

- zip code and city/state

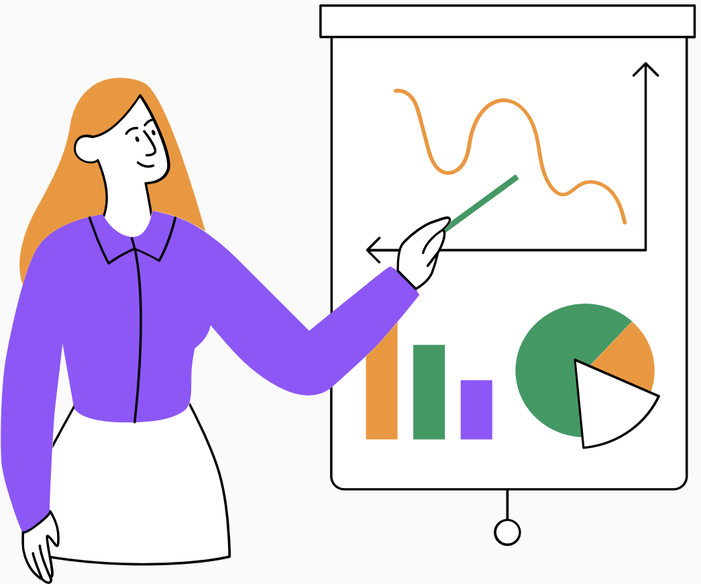
Business Logic test

- Payment value = Order value

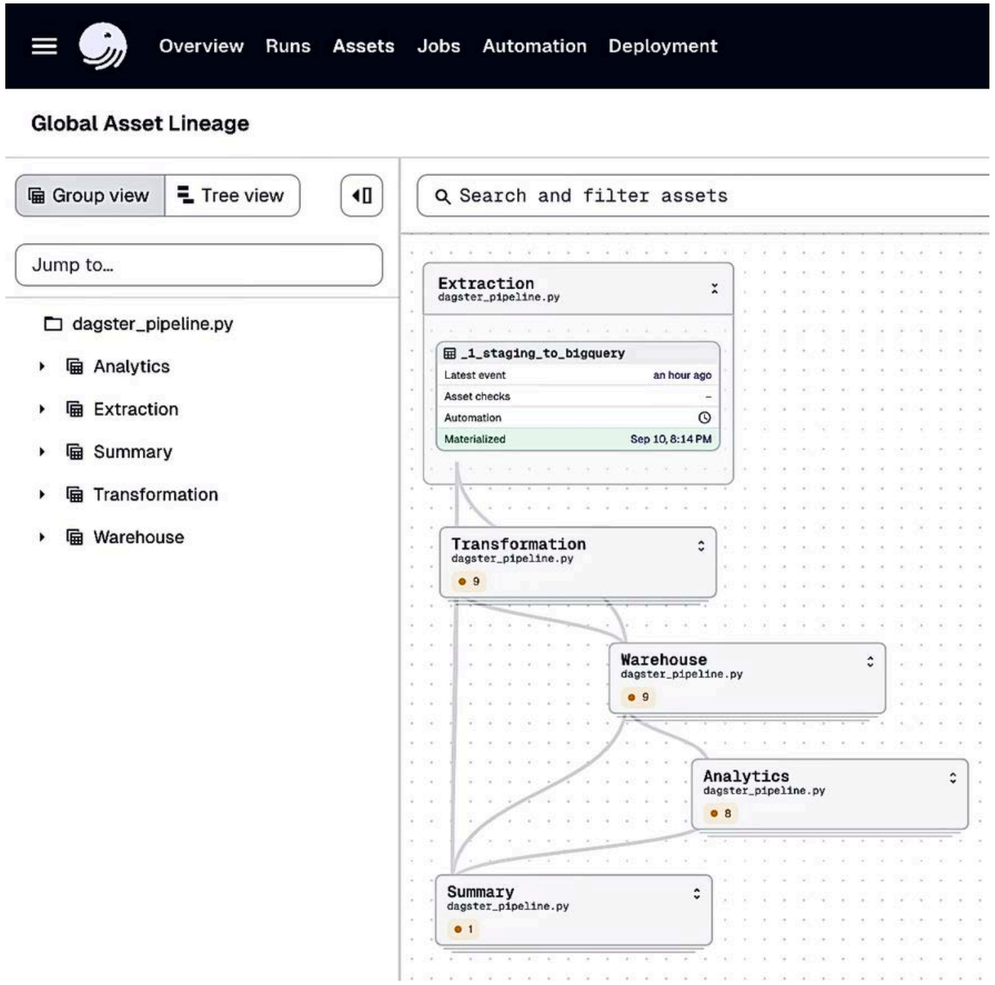
Data Quality Check

3. Implementation / Recommendations

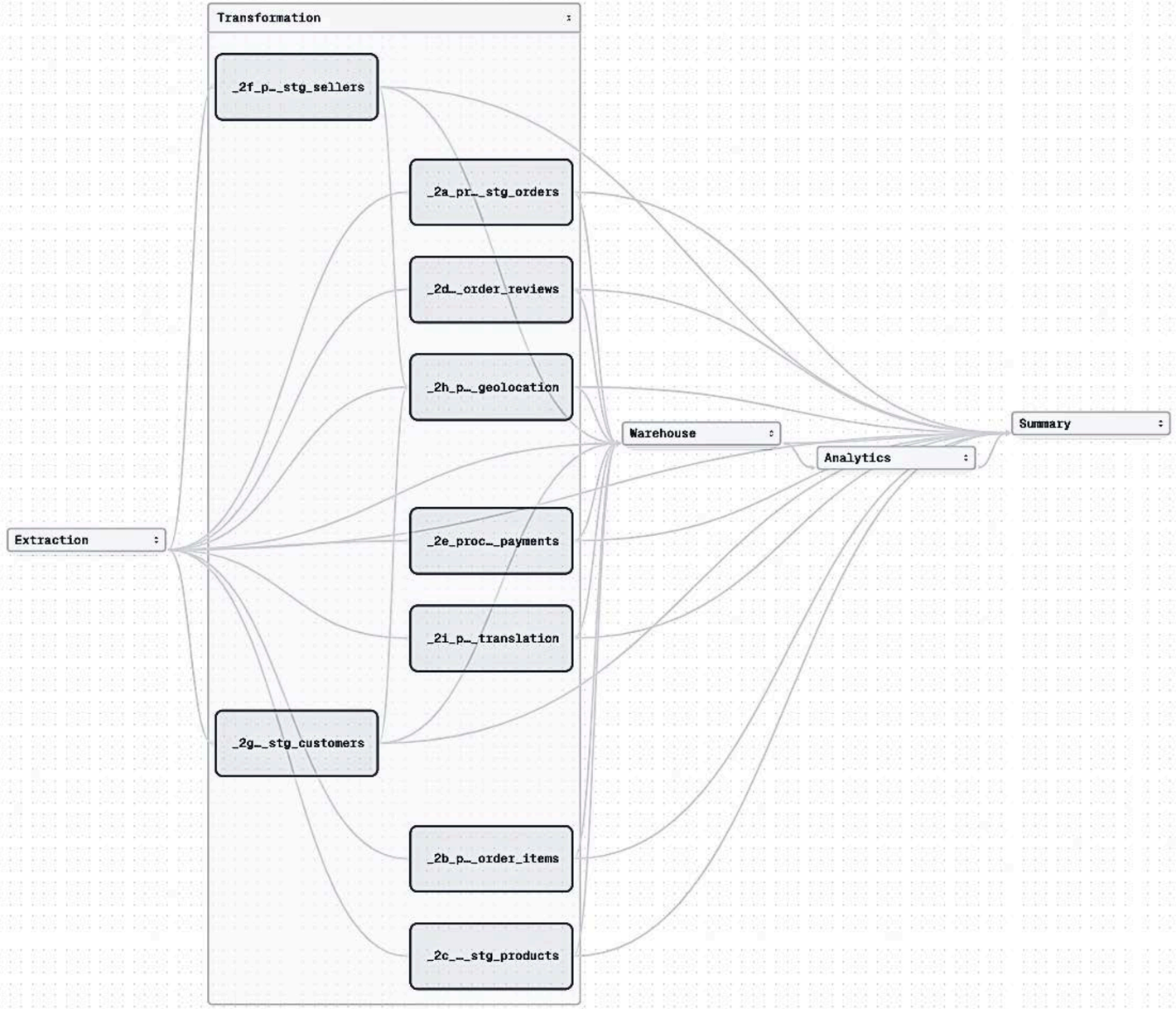
(What we do about it) & Technical Details (How it works)



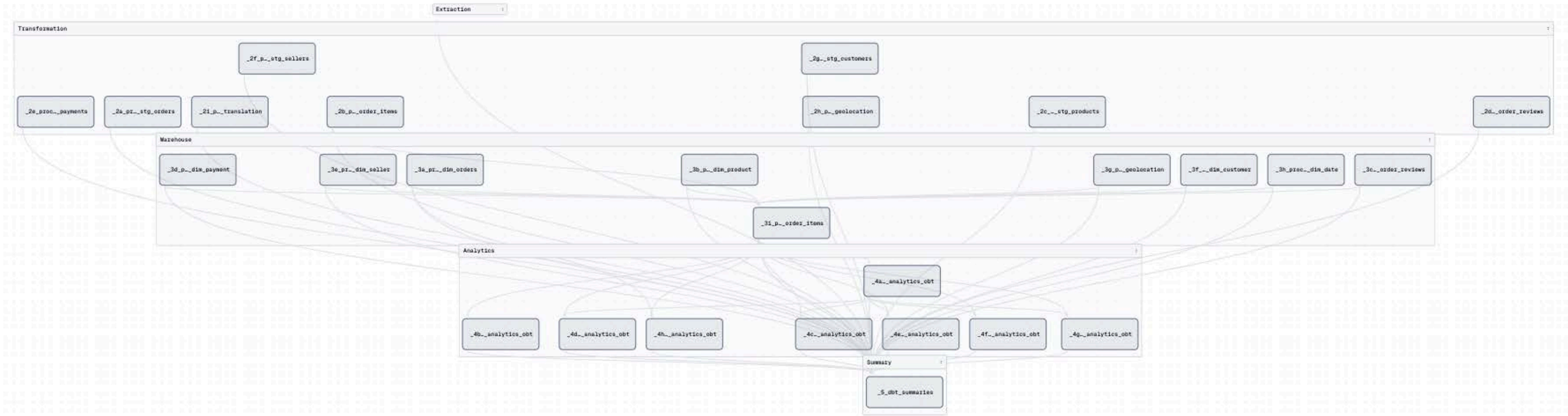
Orchestration



Dagster



Orchestration



Full Pipeline

Orchestration

Gmail

Ruby Ferdianto <rubyferdianto@gmail.com>

[Dagster Pipeline] SUCCESS - Function Status Report

rubyferdianto@gmail.com <rubyferdianto@gmail.com> 7 September 2025 at 14:43
To: Alwyn.quek07@gmail.com, Jeff.mailme@gmail.com, rubyferdianto@gmail.com, suanjoo@gmail.com, vuthylim@live.com, Zhihao@masterwugui.com

Pipeline Execution Summary

Overall Status: SUCCESS

Success Rate: 100.0%

Successful Functions: 26

Failed Functions: 0

Status Types Explained

- ✅ **SUCCESS:** Function completed successfully with no issues
- ✅ **COMPLETED:** Function finished successfully (may have minor warnings)
- ⚠️ **WARNING:** Function completed but with noted issues
- ❌ **FAILED:** Function failed with critical errors
- ❓ **UNKNOWN:** Status could not be determined

Function Status Details

Function	Status	Table	Record Count	Error Details
_1_staging_to_bigquery	✅ SUCCESS	customers → customers (Supabase: Unknown, BigQuery: 99441) | geolocation → geolocation (Supabase: Unknown, BigQuery: 1000163) | order_items → order_items (Supabase: Unknown, BigQuery: 112650) | order_payments → order_payments (Supabase: Unknown, BigQuery: 103886) | order_reviews → order_reviews (Supabase: Unknown, BigQuery: 99224) | orders → orders (Supabase: Unknown, BigQuery: 99441) | products → products (Supabase: Unknown, BigQuery: 32951) | sellers → sellers (Supabase: Unknown, BigQuery: 3095) | product_category_name_translation → product_category_name_translation (Supabase: Discovered via Meltano, BigQuery: 71)	N/A	
_2a_processing_stg_orders	✅ COMPLETED	stg_orders	99,441	
_2b_processing_stg_order_items	✅ SUCCESS	stg_order_items	112,650	
_2c_processing_stg_products	✅ COMPLETED	stg_products	32,951	
_2d_processing_stg_order_reviews	✅ COMPLETED	stg_order_reviews	98,410	
_2e_processing_stg_order_payments	✅ COMPLETED	stg_order_payments	103,886	
_2f_processing_stg_sellers	✅ COMPLETED	stg_sellers	3,095	
_2g_processing_stg_customers	✅ COMPLETED	stg_customers	99,441	
_2h_processing_stg_geolocation	✅ COMPLETED	stg_geolocation	19,177	
_2i_processing_stg_product_category_name_translation	✅ SUCCESS	stg_product_category_name_translation	73	
_3a_processing_dim_orders	✅ SUCCESS	dim_orders	99,441	
_3b_processing_dim_product	✅ SUCCESS	dim_product	32,951	
_3c_processing_dim_order_reviews	✅ SUCCESS	dim_order_reviews	98,410	
_3d_processing_dim_payment	✅ SUCCESS	dim_payment	103,886	
_3e_processing_dim_seller	✅ SUCCESS	dim_seller	3,095	
_3f_processing_dim_customer	✅ SUCCESS	dim_customer	99,441	
_3g_processing_dim_geolocation	✅ SUCCESS	dim_geolocation	19,177	
_3h_processing_dim_date	✅ SUCCESS	dim_date	3,653	
_3i_processing_fact_order_items	✅ SUCCESS	fact_order_items	112,650	
_4a_processing_revenue_analytics_obt	✅ SUCCESS	revenue_analytics_obt	112,647	
_4b_processing_orders_analytics_obt	✅ SUCCESS	orders_analytics_obt	98,665	
_4c_processing_delivery_analytics_obt	✅ SUCCESS	delivery_analytics_obt	112,647	
_4d_processing_customer_analytics_obt	✅ SUCCESS	customer_analytics_obt	95,419	
_4e_processing_geographic_analytics_obt	✅ SUCCESS	geographic_analytics_obt	27	
_4f_processing_payment_analytics_obt	✅ SUCCESS	payment_analytics_obt	112,647	
_4g_processing_seller_analytics_obt	✅ SUCCESS	seller_analytics_obt	3,095	

Failed Function Details

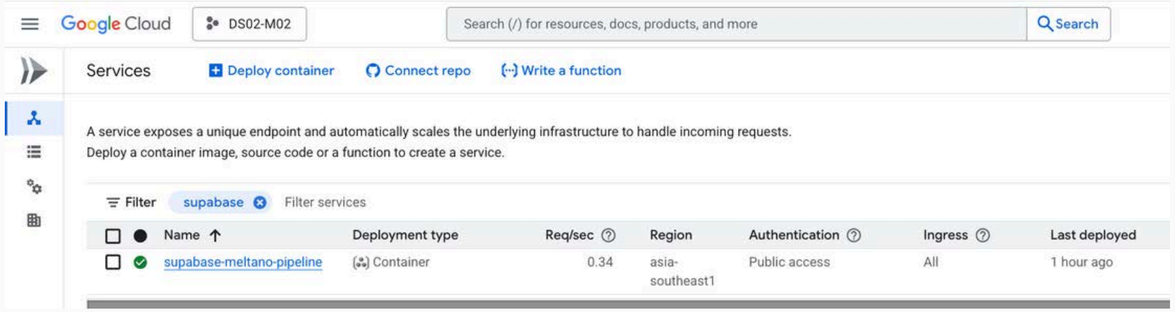
No failed functions!

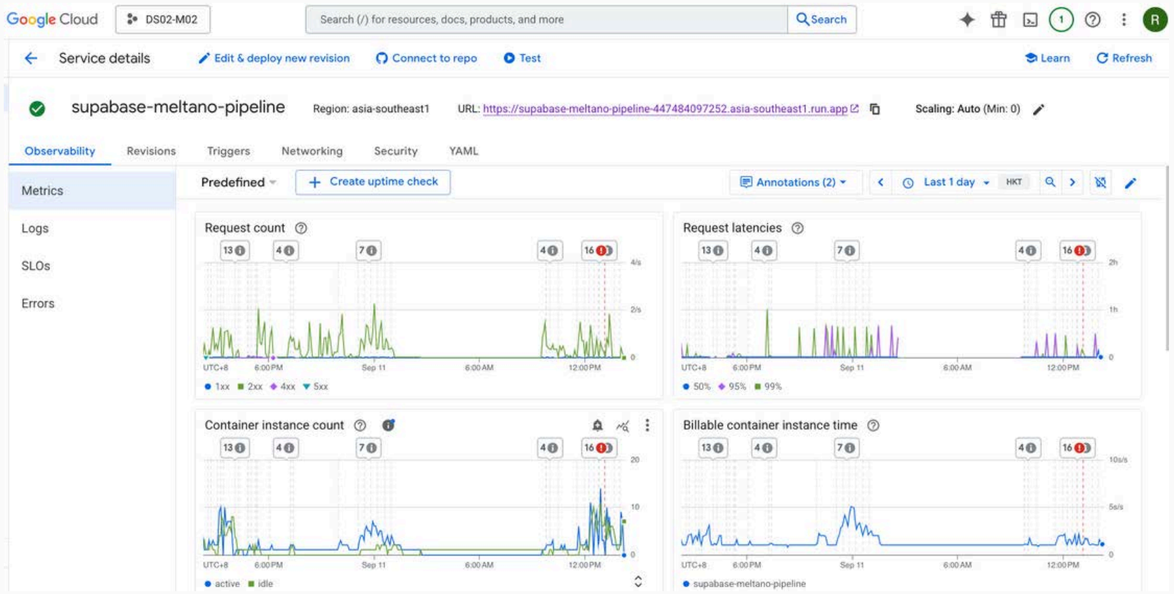
Last Pipeline - Summary

Orchestration

https://supabase-meltano-pipeline-447484097252.asia-southeast1.run.app/assets

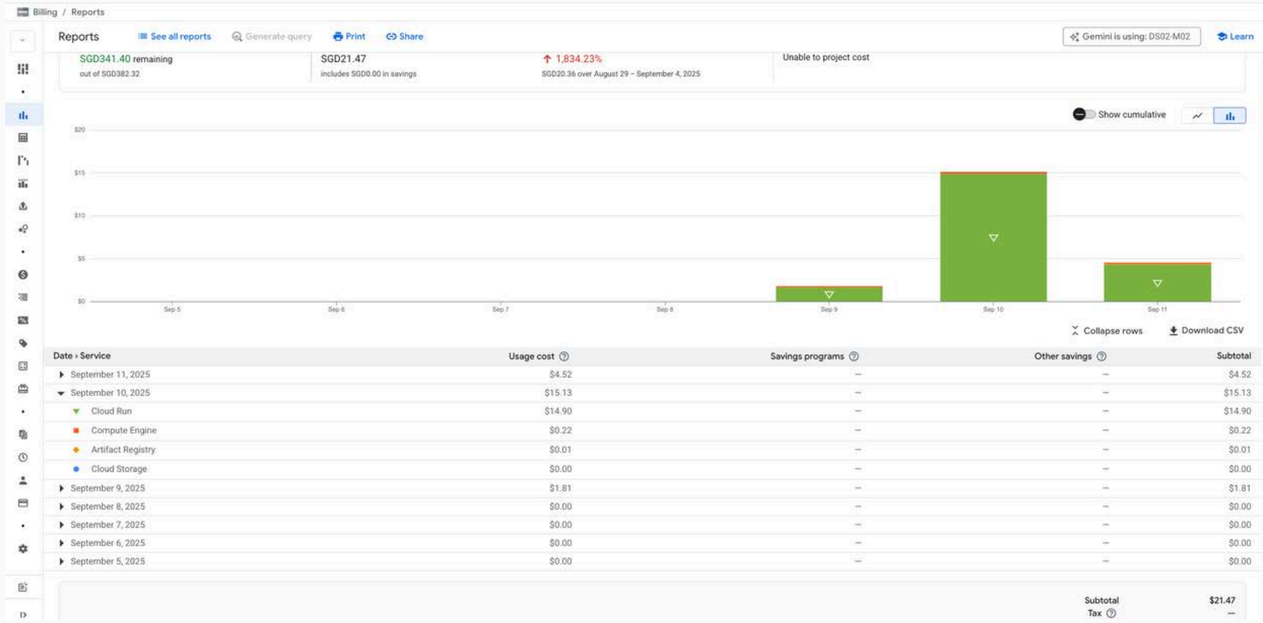
Docker @ GCP





Challenges:

- Google Cloud Run - could be expensive services
- Different behaviour between deployed locally against cloud services
- Processing time to deploy

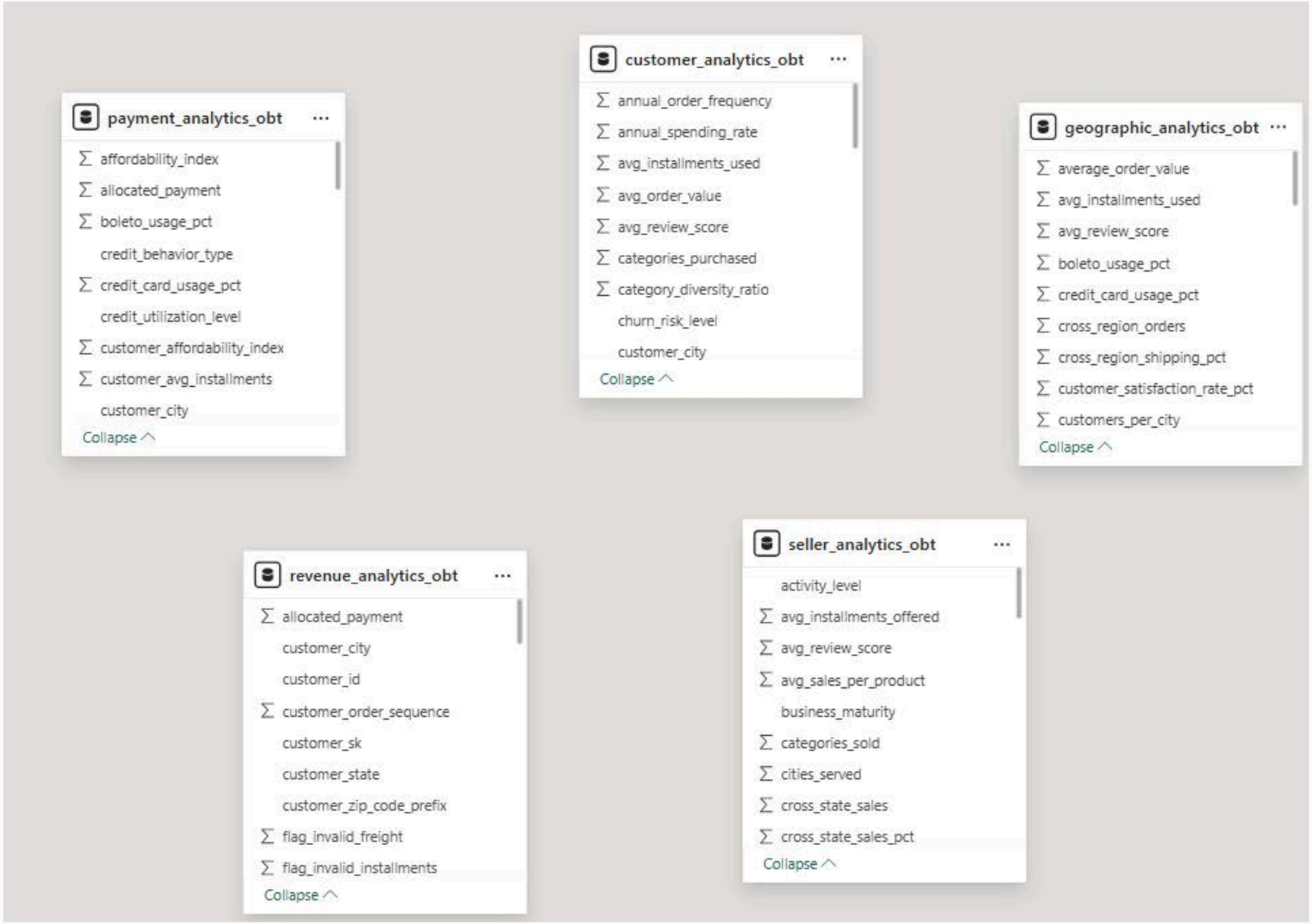


4. Analysis, Results & Insights (Sales / Marketing)

Key Metrics & Insights

Strategic Recommendations

PowerBI ERD - Ready for Serving



PowerBI Dashboard

13.59M
Sum of total_revenue

19.61K
Sum of daily_revenue_rate

227M
Sum of order_year

10.85bn
Sum of predicted_annual_clv

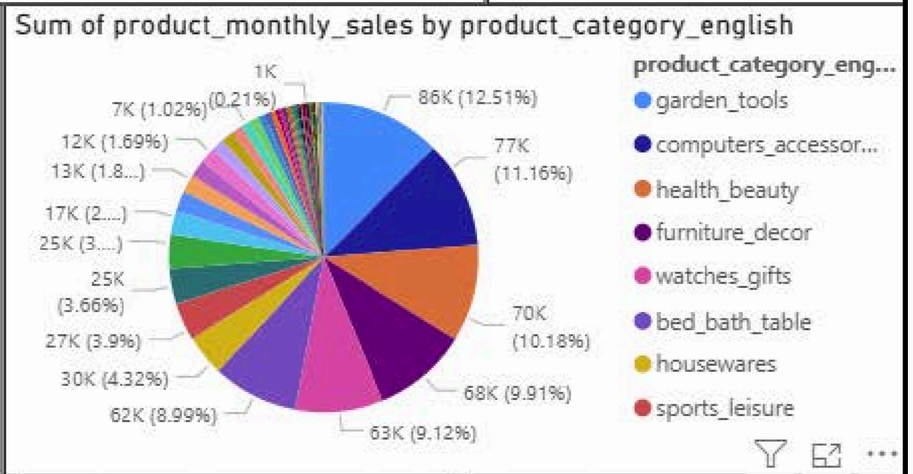
3.62M
Sum of payment_premium_pct

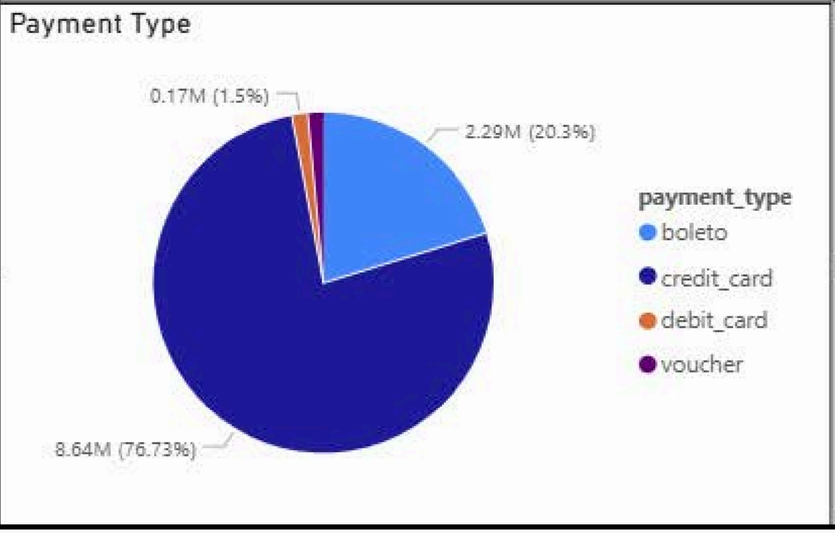
99K
Sum of total_customers

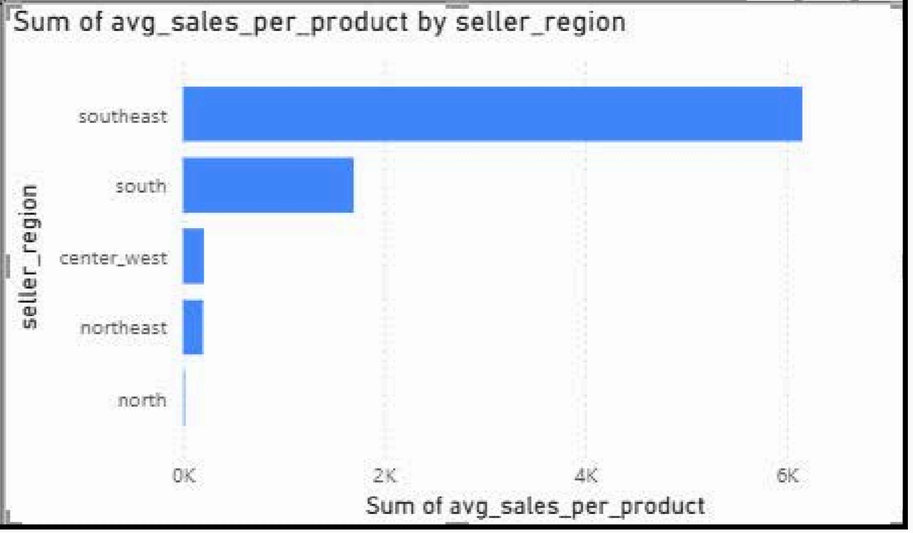
15.85M
Sum of total_payment_received

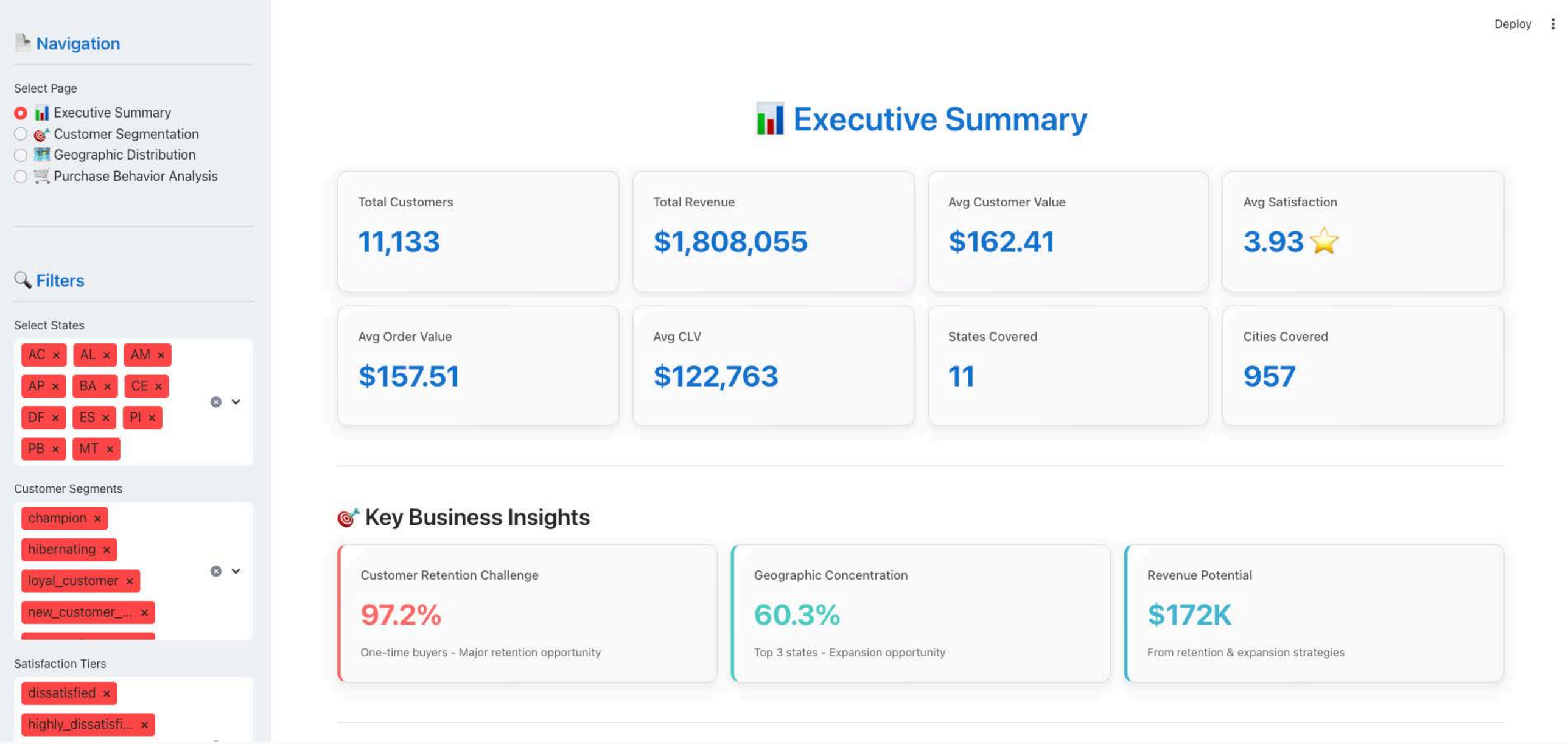
boleto
First payment_type

9.12M
Sum of customer_affordability_index









Streamlit Dashboard

🎯 Customer Segmentation Analysis

Segmentation uses an RFM (Recency / Frequency / Monetary) + behavior framework. Below is a consolidated view combining definitions, criteria, size, value and recommended action for each segment (ordered by average monetary value / strategic leverage).

Segment	Definition (Plain)	Core Criteria (Simplified)	Customers	% Base	Avg Spent	Primary Objective	Recommended Action
🏆 Champions	Highest value & most recently active multi-purchase customers; brand advocates	High recency; high frequency (≥4); monetary > $500	12	0.01%	$826.00	Retain & leverage advocacy	VIP perks, early access, referral incentives
💎 Loyal Customers	Consistent repeat purchasers with strong spend	Good recency; high frequency (≥3); monetary > $400	113	0.12%	$630.23	Deepen loyalty & referrals	Tiered rewards, exclusive bundles
🌱 Potential Loyalists	Emerging repeat buyers showing upward value trend	Recent purchase; medium frequency (2–3); rising spend	1,651	1.73%	$347.85	Accelerate to loyal	Personalized cross-sell, timed re-engagement
🆕 New (High Value)	Recent first/second purchase with high initial basket	High recency; frequency 1–2; first order > $200	38,580	40.42%	$258.10	Secure 2nd / 3rd purchase	Onboarding sequence, 2nd purchase incentive
😴 Hibernating	Previously active, now lapsed beyond recency threshold	Low recency; historical frequency ≥2; declining engagement	1,137	1.19%	$94.86	Reactivate or churn confirm	Win-back offer, reminder of value, feedback survey
🆕 New (Low Value)	Recent single low-spend purchase; low commitment	High recency; frequency 1; order < $100	53,926	56.52%	$53.24	Nurture & increase AOV	Educational content, bundles, progressive offers

Marketing Report

📈 Revenue Opportunities

1. Customer Retention Strategy
Impact: Convert 10% of one-time buyers to occasional buyers

- Potential: 9,251 customers × $116.79 additional spend = $1.08M revenue increase
- Actions: Welcome series, product recommendations, loyalty program

2. Geographic Expansion
Top 3 states (66.5%), SP holds 41.9% (customers), 38.5% (revenue)
Impact: Increase market penetration in underperforming states

- Potential: Focus on states with <1,000 customers
- Actions: Regional marketing campaigns, local partnerships

3. Segment Elevation
Impact: Move customers up the value ladder

- New Customer (Low Value) → High Value: 5% conversion = $1.11M revenue
- Potential Loyalists → Loyal: 25% conversion = $482K revenue

4. Average Order Value Optimization
Impact: Increase AOV across all segments

- 5% AOV increase = $679K additional revenue
- Focus: Cross-selling, bundle offers, minimum order incentives

🎨 Marketing Recommendations

Immediate Actions (Next 30 Days)
🔄 **Retention Campaign Launch**

1. Welcome Series: 3-email sequence for new customers
2. Second Purchase Incentive: 15% discount within 60 days
3. Product Recommendations: Based on purchase history

📧 **Segment-Specific Campaigns**
Champions/Loyal: VIP program launch

1. Potential Loyalists: Personalization increase
2. New Customers: Educational content series
3. Hibernating: Win-back offer (25% discount)

Medium-term Strategy (Next 90 Days)
🏗️. **Program Development**

1. Loyalty Program: Points-based system with tier benefits
2. Referral Program: Leverage satisfied customers
3. Geographic Expansion: Marketing in underperforming states

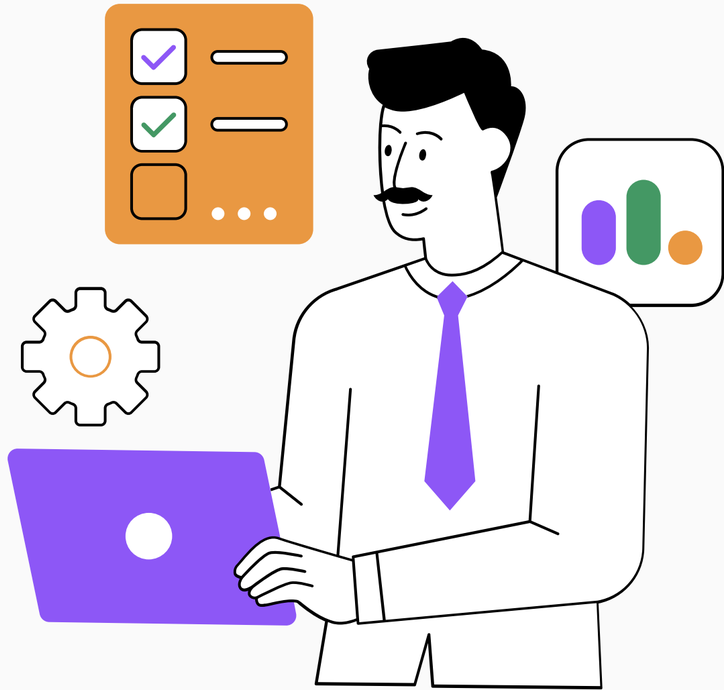
📊. **Data & Analytics**

1. Customer Journey Mapping: Identify friction points
2. Cohort Analysis: Track retention improvements
3. Predictive Modeling: Identify churn risk early

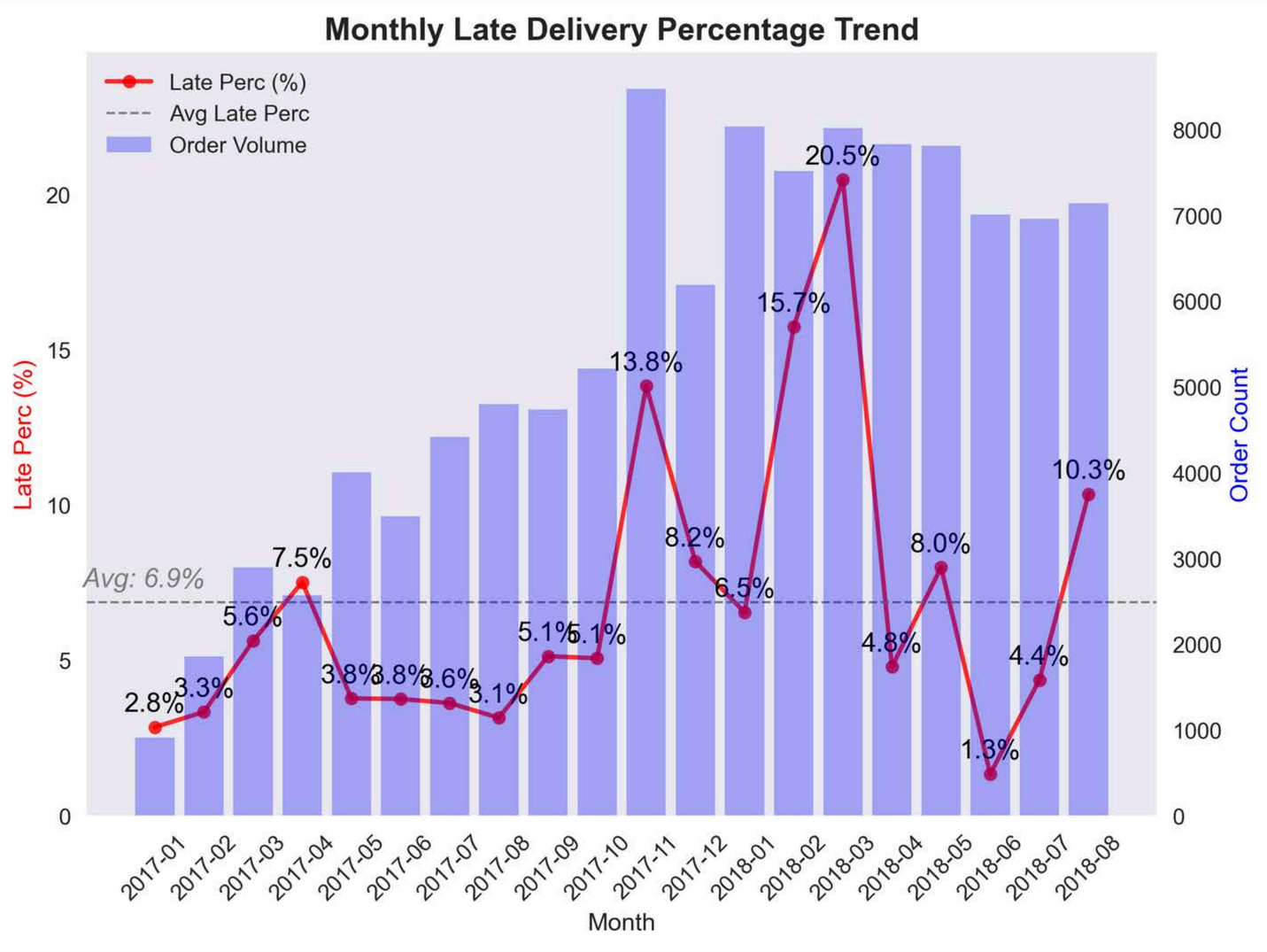
5. Analysis, Results & Insights (Operations)

Key Metrics & Insights

Strategic Recommendations



Operations



- Revenue has stagnated since Jan 2018
- Late deliveries are rising for the past 2 months

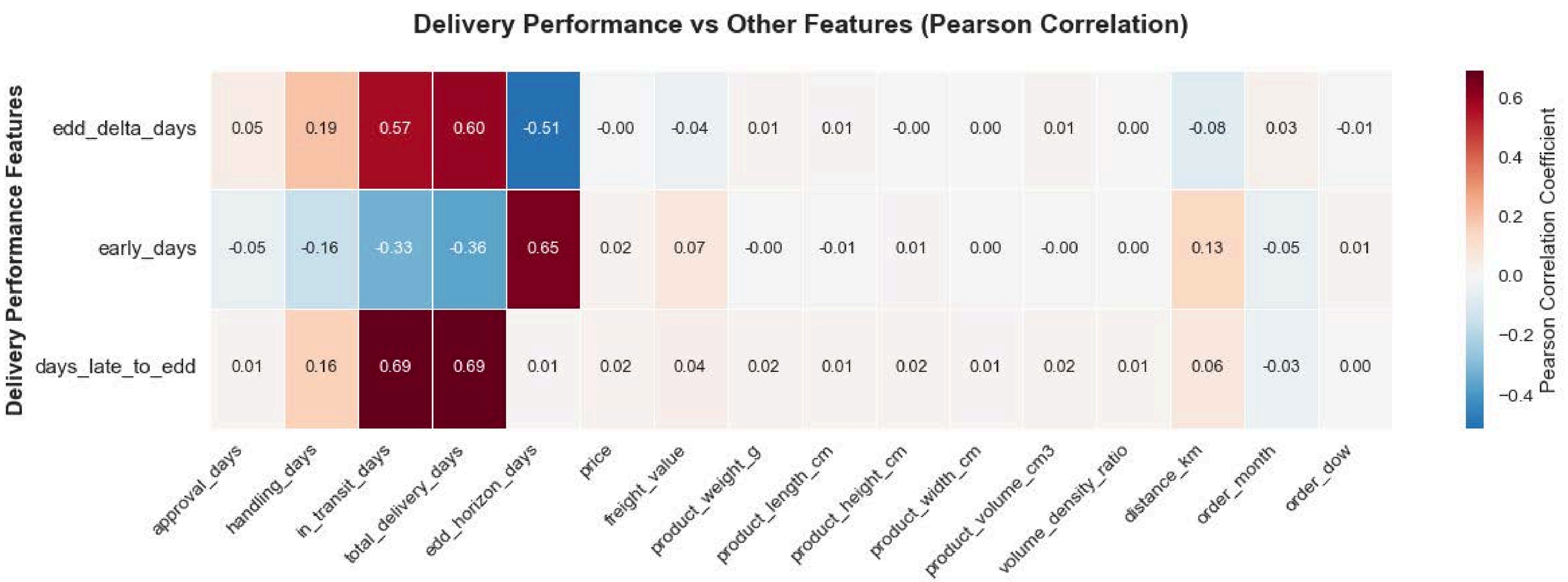
Task:

Identify the most impactful operational improvement.

Tool:

Parquet, Polars, Seaborn, Jupyter Notebook

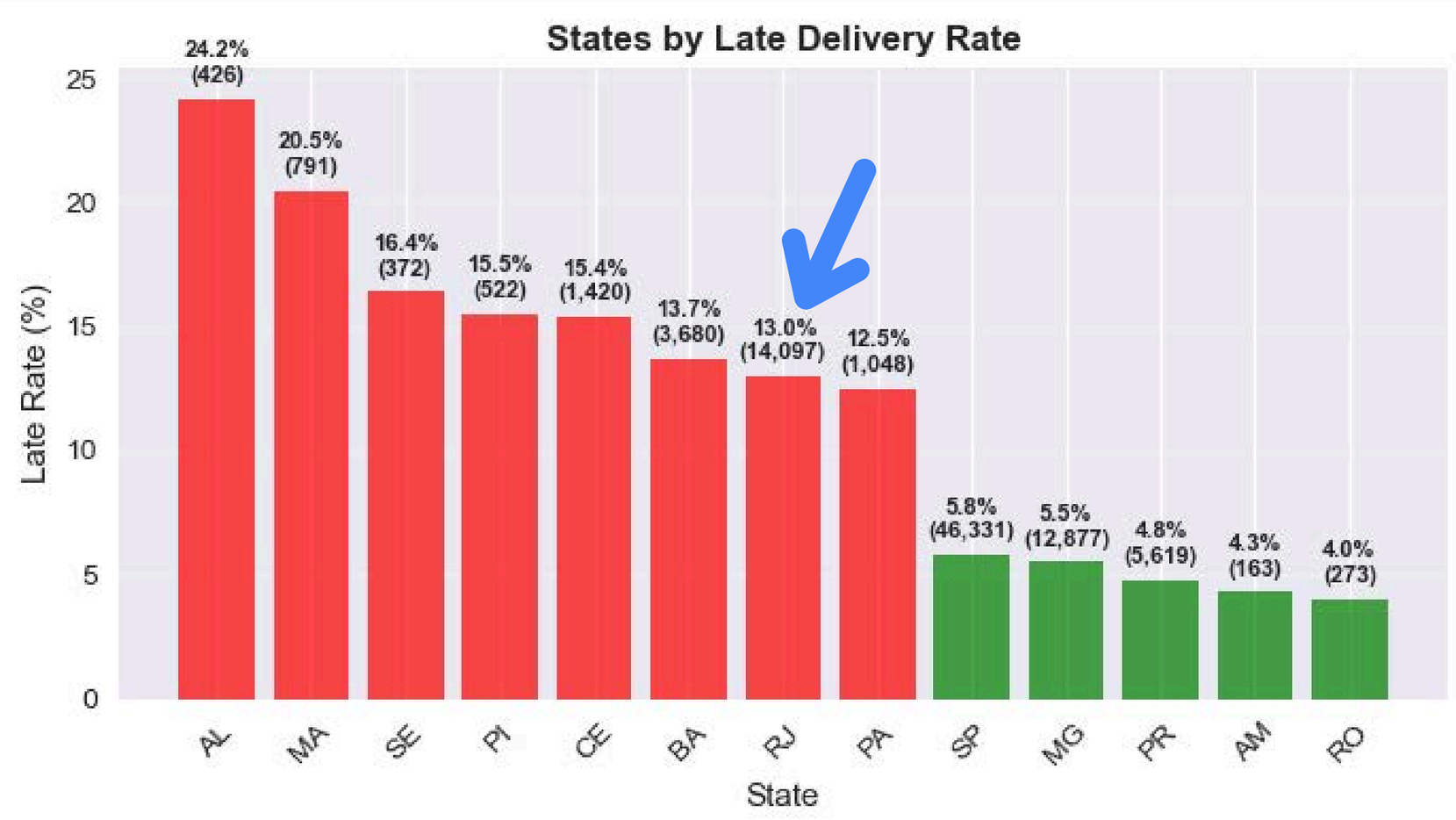
Operations

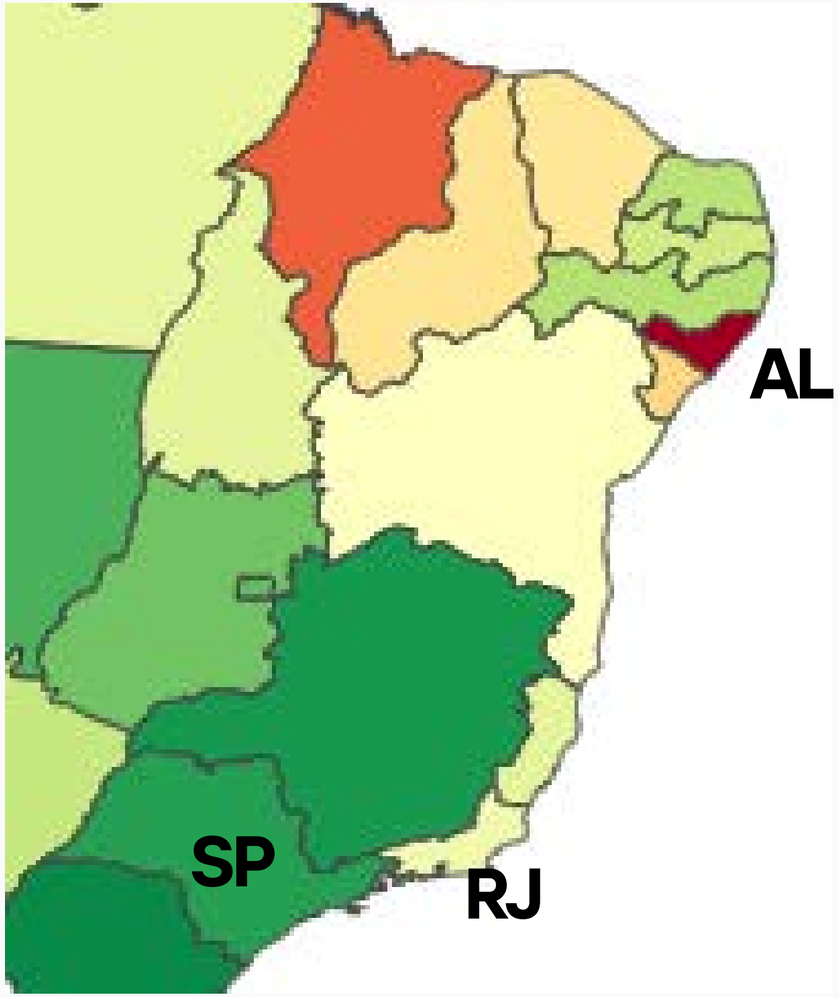


- EDD Delta Days: Actual vs Estimated Delivery Date
- Early Days: Days delivered before EDD
- Days Late: Days delivered after EDD

Operations

- No strong correlation between late delivery and other features.
- **Transit time** is the strongest driver of late deliveries.
- ⬇ Transit time, ⬇ Late delivery
- Proposal: Establishing distribution centres
- Shorten delivery routes and improve reliability





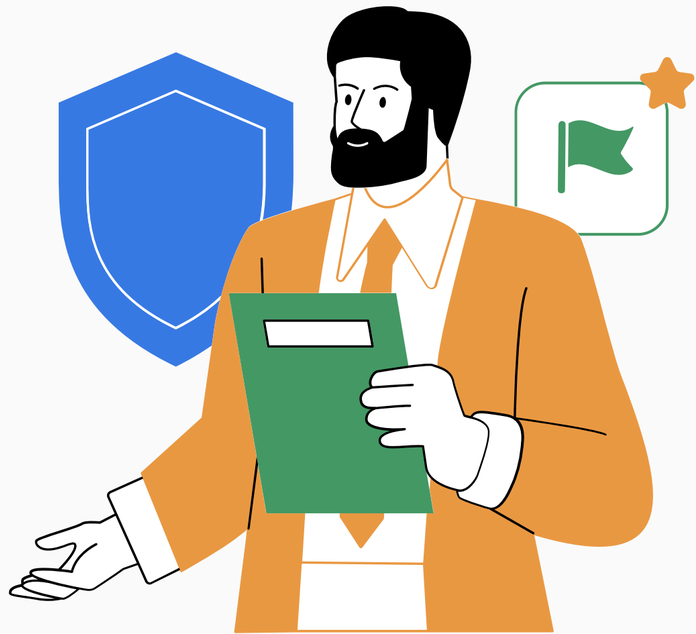
Recommendation for distribution centre: **Rio de Janeiro**

- Close proximity to Sao Paulo (The largest customer base)
- High demand
- High % of late deliveries

Operations

6. Conclusion

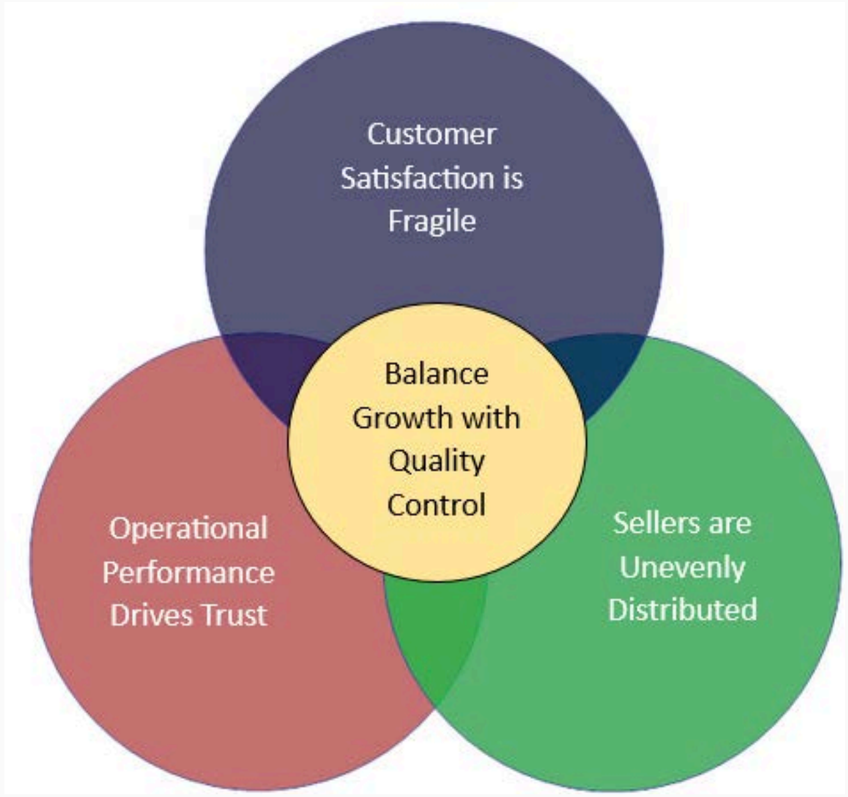
Summary and strategic implications



Conclusion

Summary

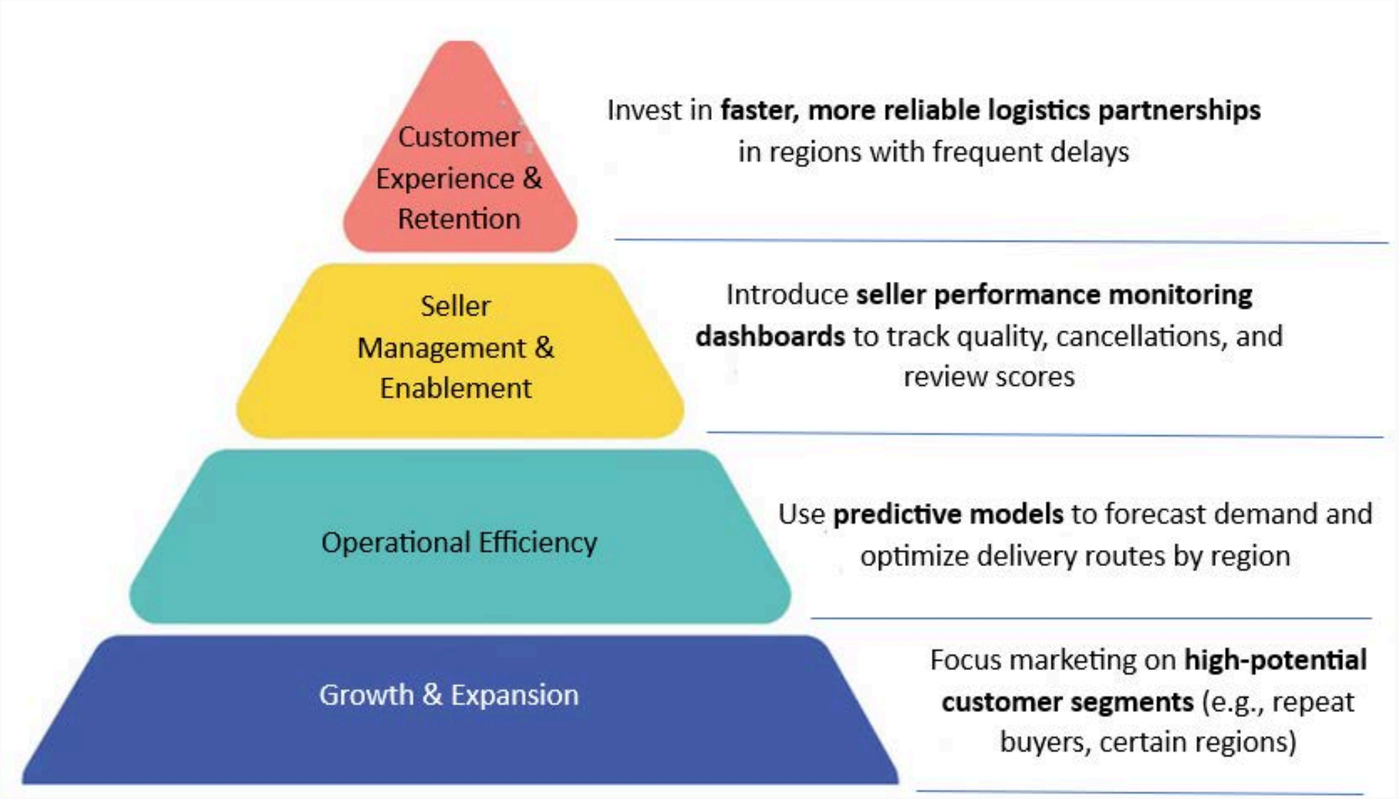
Dataset highlights three central findings:



Olist's growth depends on improving customer experience, enabling sellers, and fixing logistics inefficiencies. By focusing strategically on delivery, trust, and retention, Olist can transform insights from the Kaggle dataset into scalable marketplace improvements.

Strategic Implications

Recommended Strategic moves:



Conclusion

Technical: Formation from 1-2-3 to 1-3-2






YNWA