Oracle AI Pte Ltd

Data warehouse

Scalable Data Solutions - Big Data Engineering

NTU - PACE
December 2025

Team-7 Members

Joseph
Chen Chao
Guan Xun
Nathaniel
Darwin

olist

Executive Summary

e-commerce

Executive Summary (2017 ~ 2018)

Total number of customers	93,358
Total number of sellers	2,970
Total number of orders delivered	96,211
Total number of products	32,951
Total number of product categories	71

Serving 27 states, 4,085 cities

Late Delivery Rate 6.59%

Black Friday >8x multiple Late delivery rate up to 16.41%

16 out of 27 (~60%) states have above average late delivery rate

Key Product Categories based on GMV
- **Health and Beauty**
- **Watches and Gifts**
- **Bed, bath and Table**

Key Stakeholder concerns		
Management	Operations	IT
Delivery & Customer Satisfaction	Suppliers **Delayed** delivery	Delayed Reporting visibility Isolated Operational systems Inconsistent KPIS Delayed strategic decision making

Mission

Build a scalable **BigQuery-based Data Warehouse and Operational BI platform** to enable Olist eCommerce Operations and Strategic Business Units (SBUs) to make **timely, data-driven decisions**.

Key Business Benefits

- **Faster insights** through self-service reporting.
- **Consistent, trusted data** for operational and strategic decision-making
- **Improved agility** with flexible analytics for current and future needs
- **Enterprise-scale architecture** supporting growth and performance
- **Daily KPI visibility** to support management decision-support

Problem Statement & Scope

Area	Summary
Current State	Order fulfillment delays and customer retention challenges across a large geography are impacting revenue and profit margins.
Business Objectives	• Modernize IT and data platforms• Enable data-driven decisions• Grow and retain customers• Reduce delayed and cancelled orders• Strengthen trusted suppliers and logistics
Key Recommendations	• Improve KPI and management reporting across the operating model• Enable near real-time visibility into late deliveries and cancellations for management action• Build an enterprise Data Warehouse for consistent, trusted reporting
Scope – Business	• Orders and customer analytics• Customer ratings and reviews• Order cancellations and supplier performance• Migration from legacy reports to Operational DW reporting
Scope – Technology	• BigQuery Data Warehouse• Meltano, DBT, Dagster ELT pipeline• Star schema data models• Metabase reporting platform
Reporting Scope	• Up to 2 complex and 2 simple reports each for Orders and Customer Reviews• Prioritized enhancements and new reports
Key Deliverables	• Requirements & sign-offs• Data flow diagrams & dataset definitions• Star schemas (data provider models)• Reporting designs & specifications• ELT pipelines, validations, orchestration. Tech Lead, Data Engineering team, Reporting team.
Business Value	• Faster, more reliable KPI reporting• Improved management decision-making• Scalable, future-ready analytics platform

Methodology & Data

Kimballs four step process

Step	Step Description	Details
1	Business Process	Order Fulfillment & Delivery Lifecycle Customer purchase → Seller processing → Carrier shipping → Delivery → Customer review
2	Declare the grain	Individual transaction by Order line item.
3	Identify Dimensions Who, what , when , where -	Who : customers/ sellers What : product, When : 2017 ~ 2018 Where : States in Brazil.
4	Identify fact	1. Identify the number of late delivery 2. Relate the above with the customer score

Star Schema Design -Current Release

dim_customer
---
customer_unique_id
Number_of_orders
lifetime_value
average_order_value
First_order_at
Last_order_at
Customer_lifespan_days
Customer_segment
ltv_segment

Fact_order_items

fact_order_items
order_item_sk (PK, surrogate)
order_id (degenerate dimension)

-- Foreign Keys
purchase_date_key (FK)
delivery_date_key (FK)
seller_key (FK)
customer_key (FK)
product_id
order_status_key (FK)

-- Measures (numeric only)
price
freight_value
delivery_days
delivery_diff_from_estimated_days
seller_processing_hours
carrier_shipping_days
total_payment_value
total_payment_installments
review_score

dim_product
---
product_id
Product_category_name_english
Product_weight_g
Product_length_cm
Product_height_cm
Product_width_cm
Total_revenue_generated
Number_of_orders
Total_units_sold
average_review_score

dim_seller
---
seller_id
Total_revenue
total_orders
Total_items_sold
Average_review_score
avg_hours_to_ship

Snow flake - Star Schema Design - Planned for next release

dim_date
---
date_key (PK)
full_date
day
month
month_name
quarter
year
week_of_year
is_weekend

dim_order_status
---
order_status_key (PK)
order_status_code -- delivered, shipped, canceled, etc.
order_status_desc
is_final_status

Fact_order_items

fact_order_items
---
order_item_sk (PK, surrogate)
order_id (degenerate dimension)

-- Foreign Keys
purchase_date_key (FK)
delivery_date_key (FK)
seller_key (FK)
customer_key (FK)
product_key (FK)
order_status_key (FK)

-- Measures (numeric only)
price
freight_value
delivery_days
delivery_diff_from_estimated_days
seller_processing_hours
carrier_shipping_days
total_payment_value
total_payment_installments
review_score

dim_product
---
product_key (PK)
product_id (NK)
product_category_name
product_category_name_english

dim_customer
---
customer_key (PK)
customer_unique_id (NK)
customer_city
customer_state

dim_seller
---
seller_key (PK)
seller_id (NK)
seller_city
seller_state

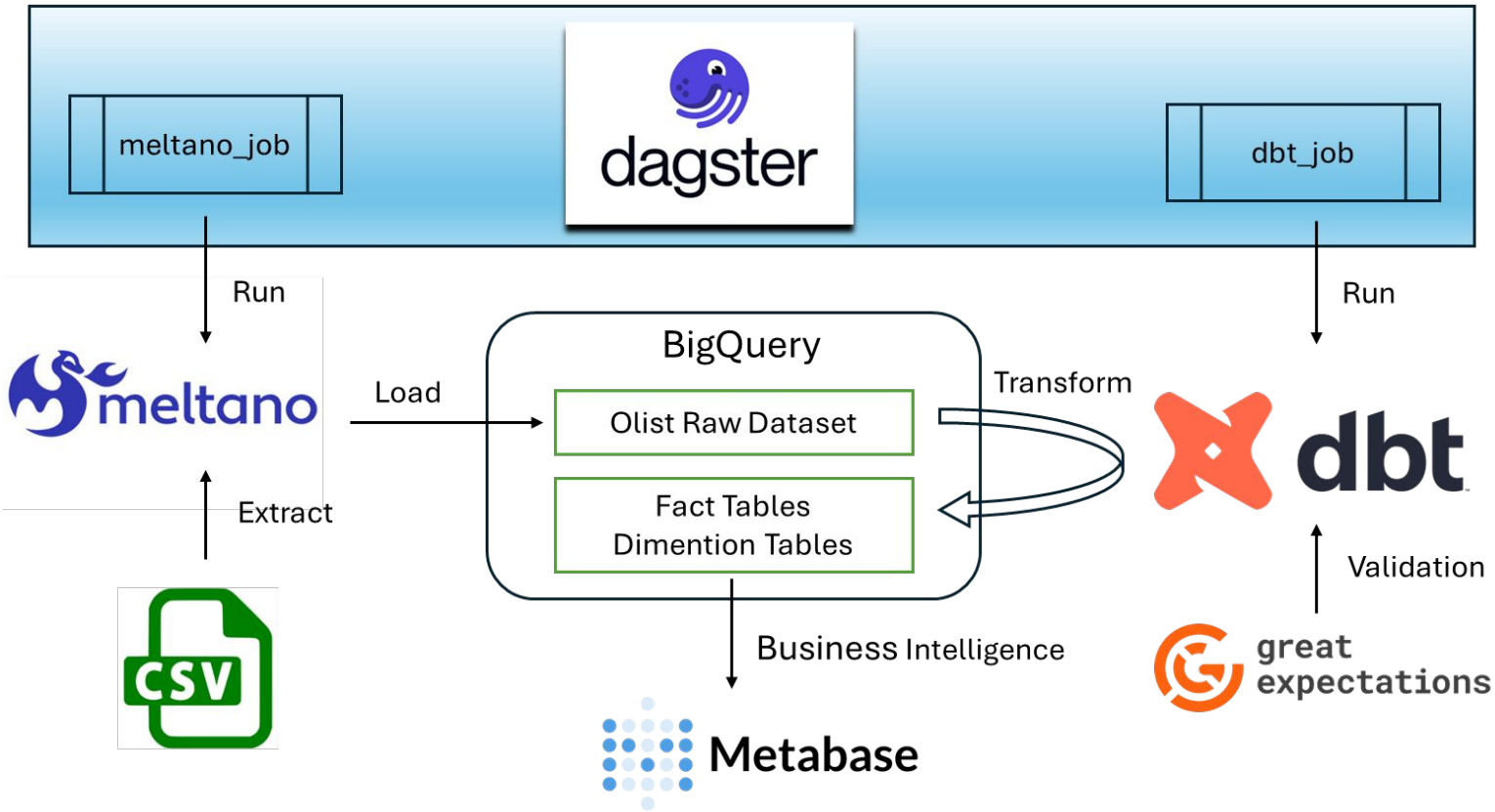
Architecture decisions

Build / Approach	Option 1	Option 2
Product Approach	SFTP , Meltano, DBT, Dagster, Big Query , Google Compute, The reporting platform must be a self service SaaS platform, Python based custom reporting	GCS for extraction and loading, Meltano, DBT, Dagster, Great Expectations,BigQuery , Google Compute, SaaS Metabase (The reporting platform must be a self service SaaS platform.)
Critical Requirements Satisfied	Yes	Yes
Critical Requirements Not Satisfied	Custom report building - requires Dev support	Reporting - Self service , enabled by the Metbase platform and scalable.
Other Considerations		
Recommendations	No	Yes

Area	Decisions
Build or Buy	Build
Application Vendor(s)	N/A
Application Style	Web Application
Server Technologies	All cloud platforms : Google BigQuery , Google Cloud Storage , Google Compute , SaaS Metabase.
Hosting	Google CLOUD
Network	HTTPS - Internet connectivity for the mobile/web clients
Security	AD User authentication for the Web User.
Integration	Meltano
Transformation	DBT & Great Expectations for Validations
Reporting	Metabase (SaaS)
Orchestration	Dagster

Solution Architecture

Data Pipeline Architecture



Risk and Mitigation

Risk	Description and Mitigation
Failure of Orchestration Dagster Web Server	Deploy dagster in multi-zone GKE and external HA cloud SQL
Duplicate Data loads	Incremental data loads with bookmarks, deduplication models in DBT
Observability & Operational risks	Pipeline failure not detected. Mitigation cloud , dagster monitoring alerts.

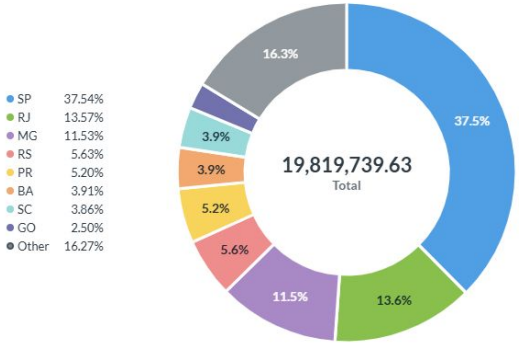
Operational data warehouse details

Dataset Description	https://console.cloud.google.com/bigquery?ws=!1m4!1m3!3m2!1sntusctp!2solist_dataset (Loaded by Meltano) https://console.cloud.google.com/bigquery?ws=!1m4!1m3!3m2!1sntusctp!2solist_dataset_stg (DBT) https://console.cloud.google.com/bigquery?ws=!1m4!1m3!3m2!1sntusctp!2solist_dataset (Data warehouse)
Approach & Assumptions	From data extraction to datawarehouse updates is fully automated via Dagster Orchestration. We support a daily (2.00 am) scheduling to provide accurate same day reporting of KPI's

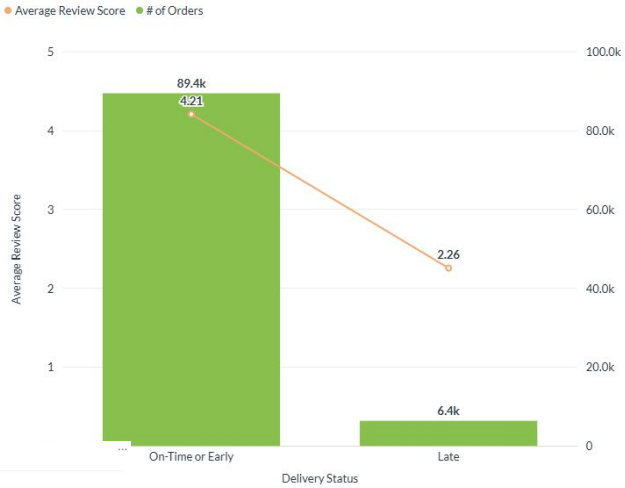
Results & Insights

Key metrics dashboard

Revenue Distribution in each State



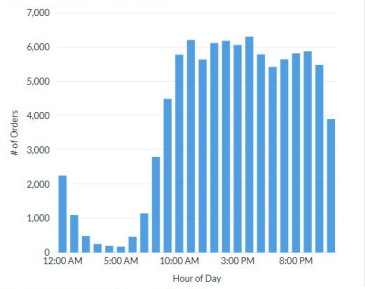
Review Score vs Delivery Status



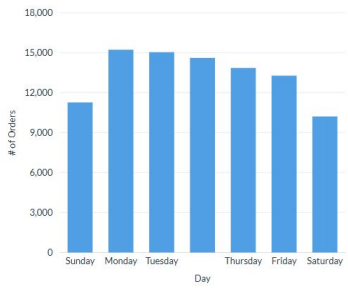
Late Delivery Rate by State

#	State	% of Late Delivery
1	AL	20.88%
2	MA	17.89%
3	SE	16.27%
4	CE	13.63%
5	PI	13.55%
6	BA	11.94%
7	RJ	11.62%
8	PA	11.22%
9	RR	10.87%
10	PB	10.73%
11	ES	10.69%
12	TO	9.68%
13	PE	9.04%
14	RN	8.99%
15	MS	8.87%
16	SC	8.2%
17	GO	6.27%
18	MT	6.06%
19	DF	5.92%
20	RS	5.86%
21	SP	4.38%
22	MG	4.36%
23	RO	4.03%
24	PR	3.88%
25	AP	3.7%
26	AC	3.3%
27	AM	3.05%

Number of Orders by Hour of the Day



Number of Orders by Day of the Week



Average Late Delivery Rate

6.59%

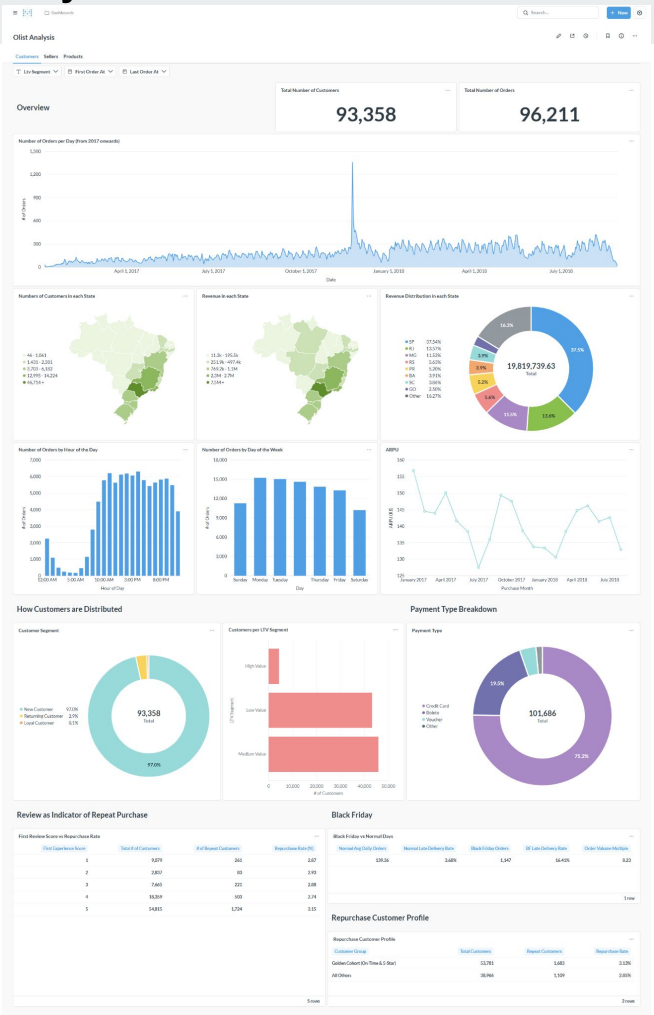
Black Friday vs Normal Days

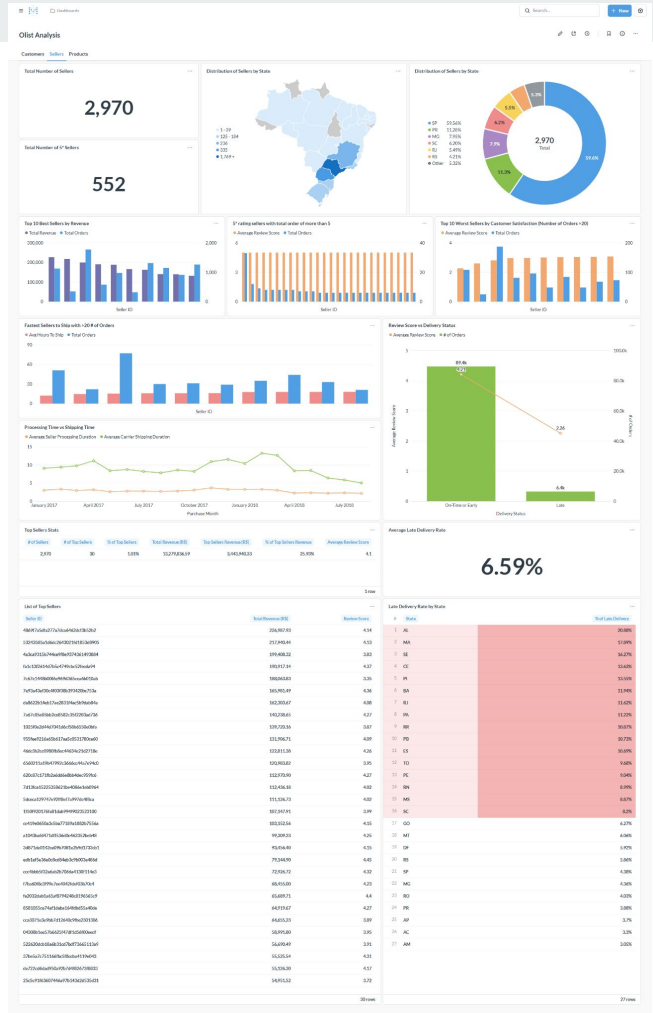
Normal Avg Daily Orders	Black Friday Orders	Order Volume Multiple	BF Late Delivery Rate
139.36	1,147	8.23	16.41%

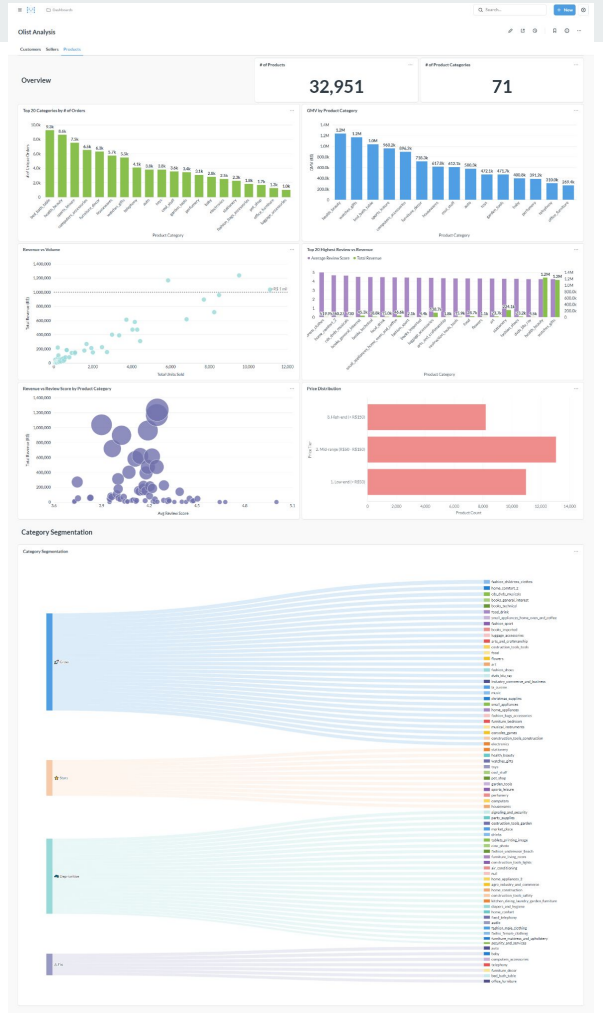
Repurchase Customer Profile

Customer Group	Total Customers	Repeat Customers	Repurchase Rate
Golden Cohort (On-Time & 5-Star)	53,781	1,683	3.13%
All Others	38,966	1,109	2.85%

Key metrics dashboard







Order Reports

Core Corridor
SP-RJ-MG

🕒 Peak Activity:
10 AM – 4 PM
Monday & Tuesday

Black Friday
8.23x multiple
More late delivery

Category Segmentation
“Grow”, “Start”, “Fix”, “Deprioritise”

Key Insights

1. Three core corridor identified: Sao Paulo, Rio de Janeiro and Minas Gerais
2. Order volume peaks during the middle of the day and on Monday & Tuesday
3. Black Friday sees order surge of more than 8 times with higher late delivery rate from 6.59% to 16.41%
4. Product categories can be segmented into **"Grow", "Stars", "Fix" and "Deprioritise"**

Recommended Actions

1. Prioritise marketing effort and warehouse placements in core corridor
2. Schedule more customer support and logistics staff during peak to ensure rapid order processing
3. Implement dynamic shipping deadlines and pre-emptive stock positioning
4. Focus promotional on **"Stars" and "Grow"** segments and price audit on **“Fix”** segment

Customer Review Reports

Delivery Impact
4.21 vs 2.26
(On-time vs Late)

5★ Sellers
18.6%
of sellers

Golden Cohort
+10% higher
repurchase

Regional Risk
AL & MA
>17% late

Key Insights

1. Delivery performance is the strongest driver of review scores
2. A small elite of sellers drives disproportionate revenue and satisfaction
3. Loyalty potential exists but is concentrated in 5★ delivery experiences
4. Northeast delivery failures are actively damaging brand perception

Recommended Actions

1. Deploy proactive late-delivery alerts to manage expectations
2. Launch 'Preferred Seller' badges for 5★ sellers
3. Target Golden Cohort with post-purchase loyalty incentives
4. Audit and replace underperforming carriers in AL, MA, SE

Final Deliverable

Project repo	https://github.com/chenchaosg/dsai3-m2-group7-project
Powerpoint PDF	Upload the pdf of presentation

Conclusion

Q&A