Cart Catalysts

Generation Singapore

Group 5

Members

Alex Yong

Fahmy Mahamud

Ho Choo Geok

Janet Keh

Agenda Outline

1. Problem Statement
2. Project Objectives
3. Project Architecture Design
4. Market Basket Analysis
5. Monitoring, Testing and Validation
6. Project Deliverable
7. Challenges and Future Roadmap

Problem Statement

- Olist's business leaders struggle to monitor performance due to inconsistent and messy data.
- Lack of centralized structure makes it difficult to generate reliable insights.
- Existing reporting lacks clarity, drill-down capability, and real-time visibility.

Project Objectives

1. Establish a Unified Data Foundation

- Consolidate disparate data sources into a centralized Lakehouse architecture using Microsoft Fabric.

2. Design a Scalable and Modular Data Pipeline

- Build reusable ETL components using Fabric Notebooks.
- Ensure modularity by abstracting transformation logic, validation checks, and error handling into reusable pipeline blocks.

3. Enable Drill-Down Reporting

- Deliver real-time dashboards with drill-through capabilities and semantic modeling.

PROJECT ARCHITECTURE DESIGN

Tech Stack

Storage

- Azure Blob Storage

Fabric Workspace

Infrastructure and Workflow Orchestration

- Microsoft Fabric
- OneLake
- Lakehouse

Ingestion Transformation

- Apache PySpark
- Fabric Notebook
- pandas

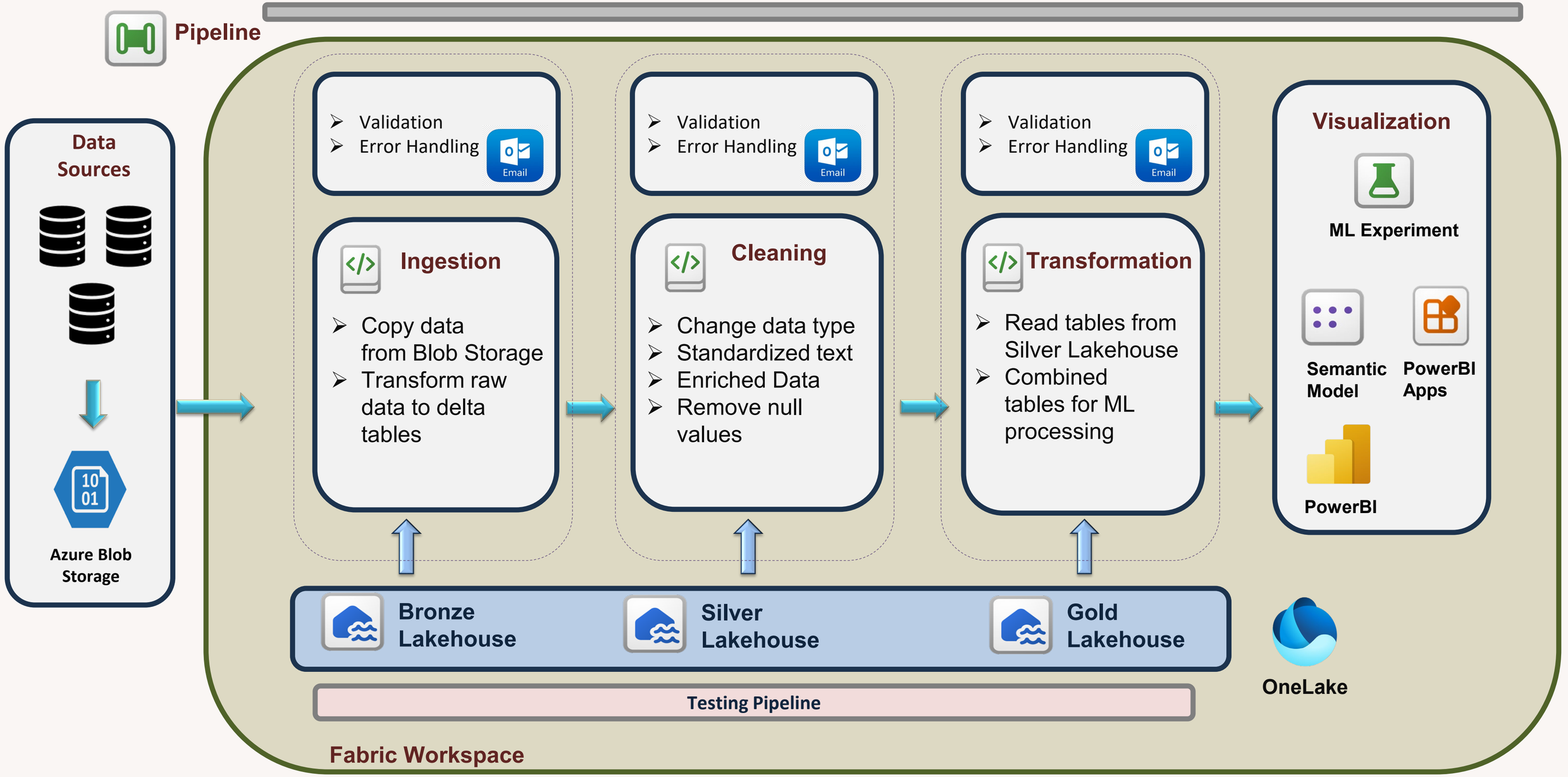
Visualization Modeling

- PowerBI
- Semantic Model
- Fabric Experiment
- PowerBI App

Notification Monitoring Testing

- Microsoft Outlook
- Monitoring Hub
- Pytest

Pipeline Architecture



Medallion Architecture

1. Improved Data Quality & Integrity

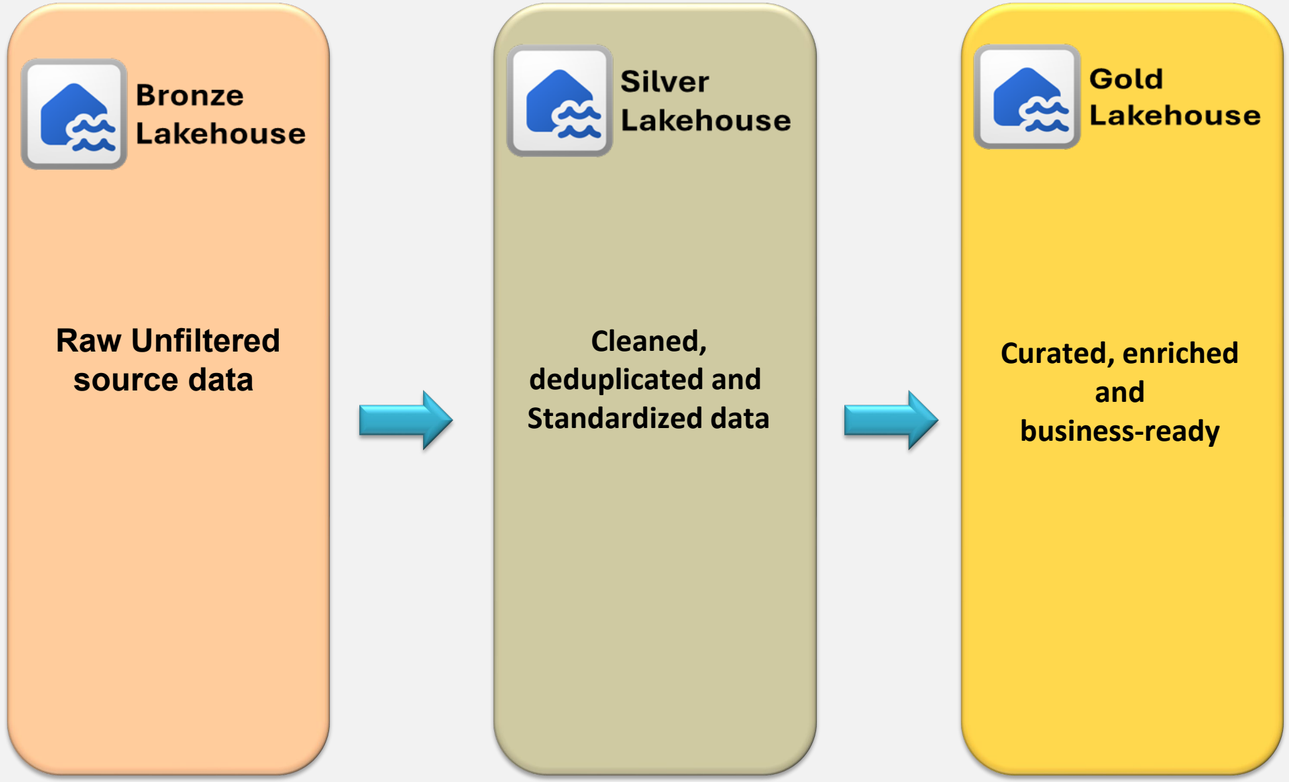
Each layer progressively refines the data. Ensures consistency, traceability, and trust in downstream analytics.

2. Flexibility & Reusability

Raw data in Bronze can be reused to regenerate Silver or Gold when business logic changes.
Modular design makes it easier to plug in new sources or update logic without breaking downstream flows.

3. Enhanced Data Governance

Each layer provides a checkpoint for auditing, validation, and compliance. It is easier to trace errors or anomalies back to their origin.



Bronze Lakehouse

The primary goal is to ingest raw data from source, and this layer provides a starting point for downstream transformations.

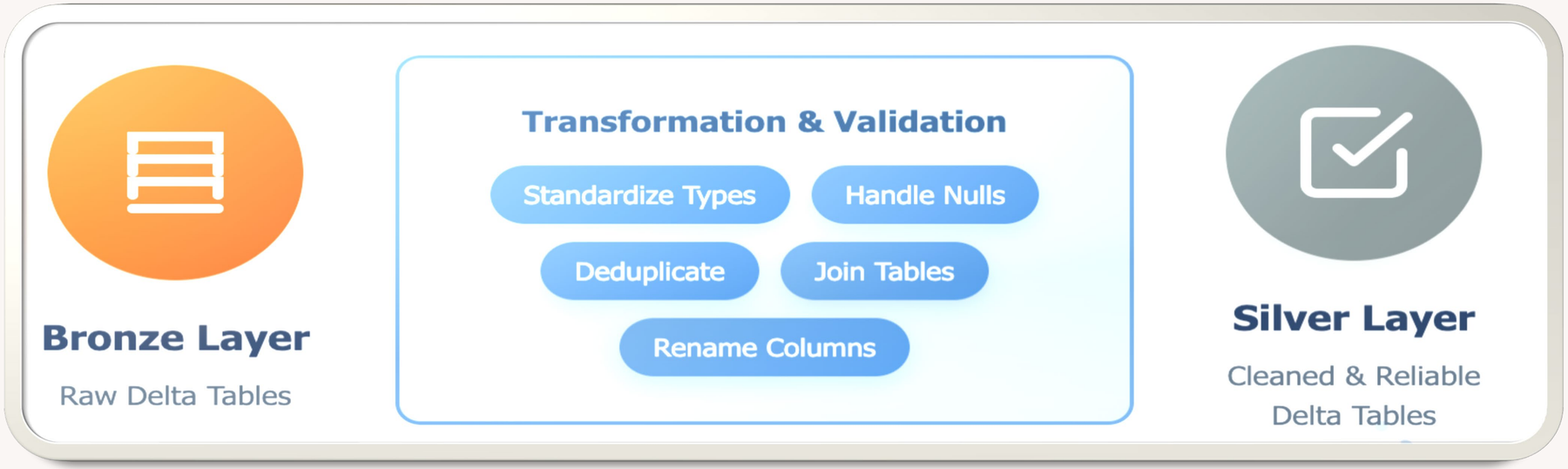
Process Flow:

- **Ingestion**: Raw CSV files are read from the Files section of the Fabric Lakehouse.
- **Conversion to Delta**: CSV files are converted to Delta format and saved as tables.
- **Schema & Structure**: The schema is inferred from the CSV headers and structure matches the source data 1:1.

Key Characteristics:

- **Overwrite Mode**: Data is replaced completely each time the pipeline runs.
- **Schema-on-Read**: Schema is inferred dynamically from the CSV files during ingestion, allowing for flexibility.
- **Simple Conversion**: Straightforward CSV-to-Delta transformation without additional metadata or audit columns.

Silver Lakehouse



Transforming raw data from the Bronze layer into cleaner Delta tables with type corrections and basic quality checks to prepare data for further processing.

Process Flow:

- **Data Sourcing**: Data is read from the Bronze Delta tables.
- **Cleaning & Conforming**: **Data Type Standardization**: Timestamps are cast to a consistent format, and numerical/string types are corrected. **Null Value Handling**: Nulls in critical columns are dropped for orders and reviews tables. **Other Transformations**: Additional columns like purchase_year and purchase_month are added to the orders table. Column names in the products table have two typos corrected (lenght → length).

Key Characteristics:

- **Type Consistency**: Data types are standardized across tables for numerical and timestamp fields.
- **Basic Quality Checks**: Simple validation to ensure critical fields are not null and data types are correct.

Gold Lakehouse

The primary goal is to prepare a unified, ML-ready dataset by transforming and consolidating multiple Silver Lakehouse tables.

- **Data Sourcing**: Data is read from the Silver Delta tables.
- **Transformation Logic**: Merge the above tables into a single, enriched table suitable for machine learning workflows.

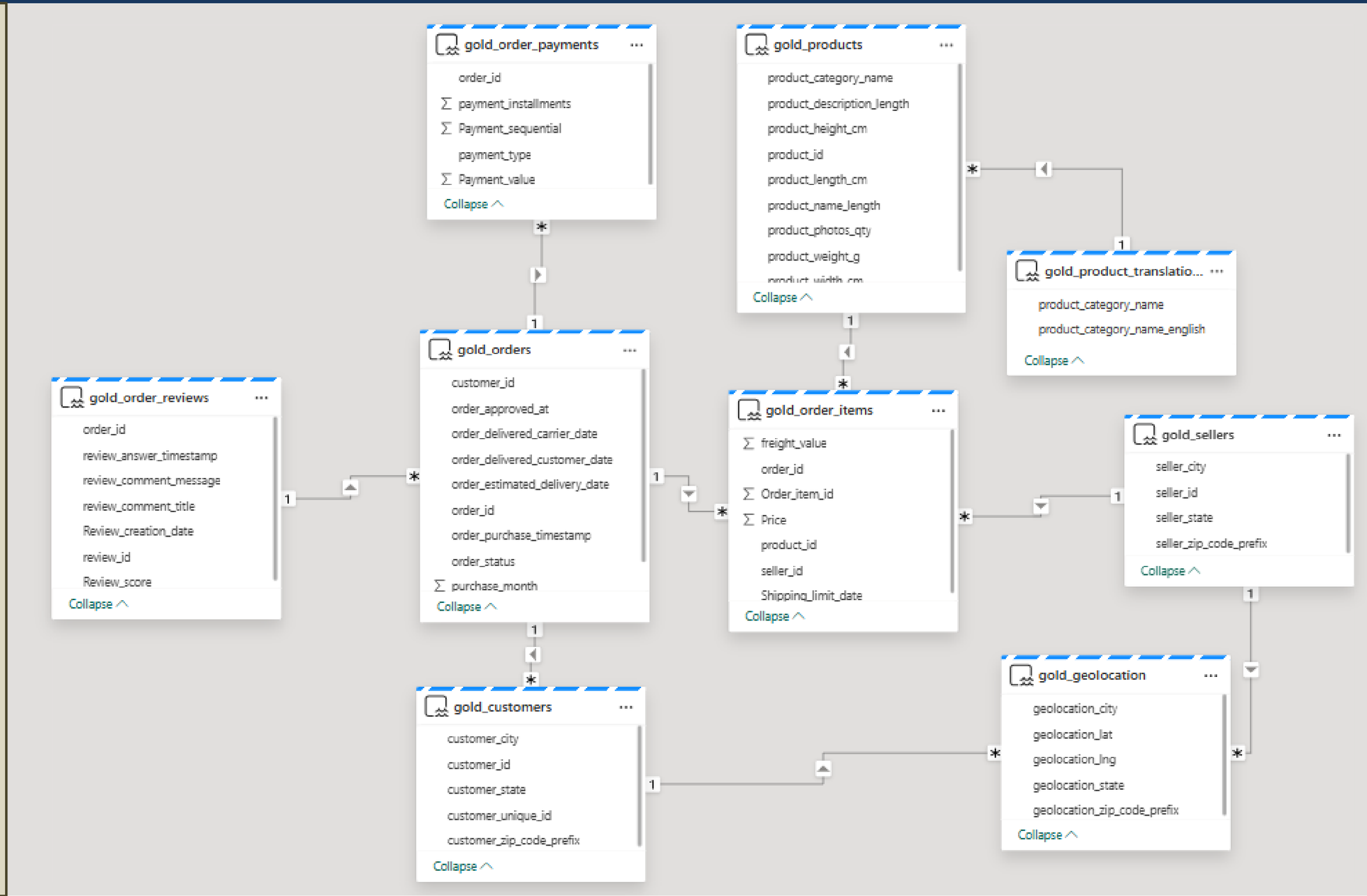
Key Characteristics:

- **Basic Quality Checks**: Validate that the schema of the Gold Lakehouse table mirrors the Silver Lakehouse source post-transformation.
- **Data Integrity Checks**: Ensure row counts between Silver and Gold tables remain consistent to confirm no data loss during transformation.

Gold Lakehouse

Dimensional Schema

To support machine learning and analytics by organizing data into a dimensional model—optimized for querying, slicing, and aggregating.



MONITORING TESTING AND VALIDATION

Monitoring Hub

Monitoring using Microsoft Fabric's native capabilities for end-to-end pipeline visibility

Monitoring Hub Dashboard

- Centralized monitoring for pipeline runs, notebooks, and semantic model refreshes
- Real-time status tracking with historical trend analysis

Automated Activity Tracking

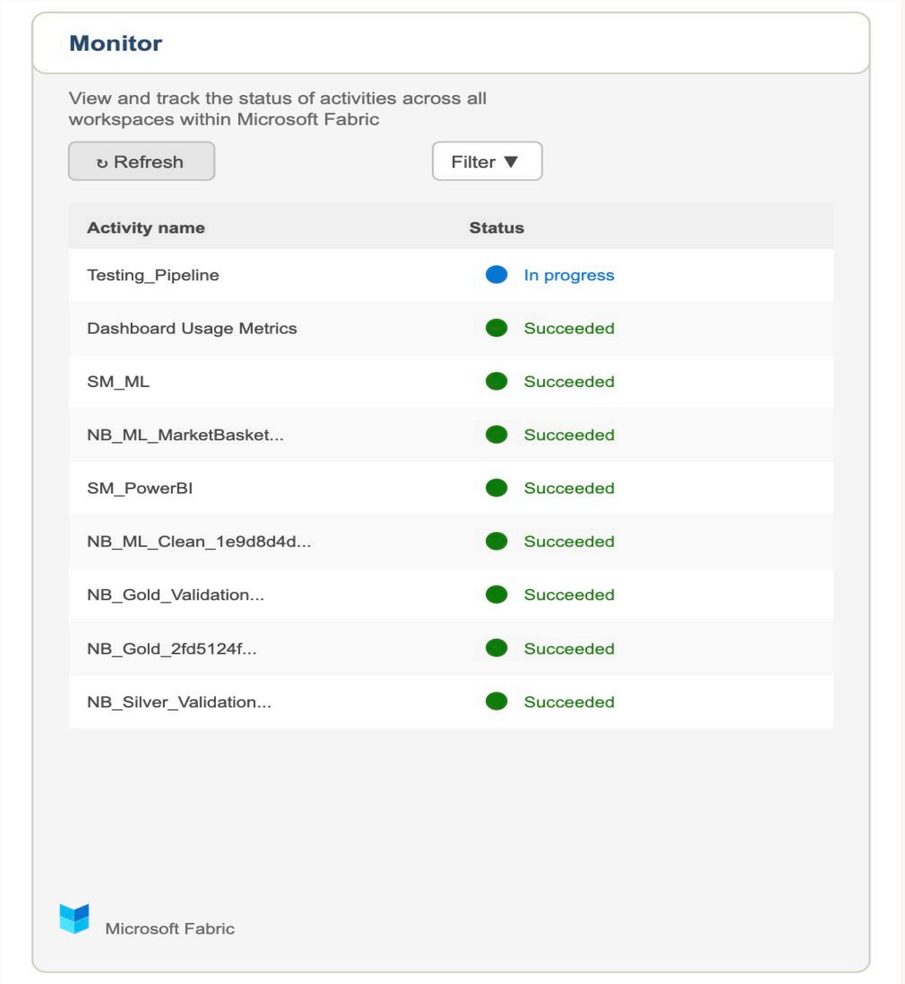
- Automatically capture execution metrics, duration, row counts, and error details
- Notebook runs log cell execution times, validation results, and stack traces

Intelligent Alerting

- Office 365 Outlook integration for failure and success notifications
- Customizable alerts with workspace context and actionable error information

Semantic Model Monitoring

- Power BI refresh status and duration tracking
- Automated alerts on refresh failures to ensure dashboard currency



Key Benefits: Single pane visibility • No additional infrastructure • Seamless Fabric integration • Built-in retention and analysis

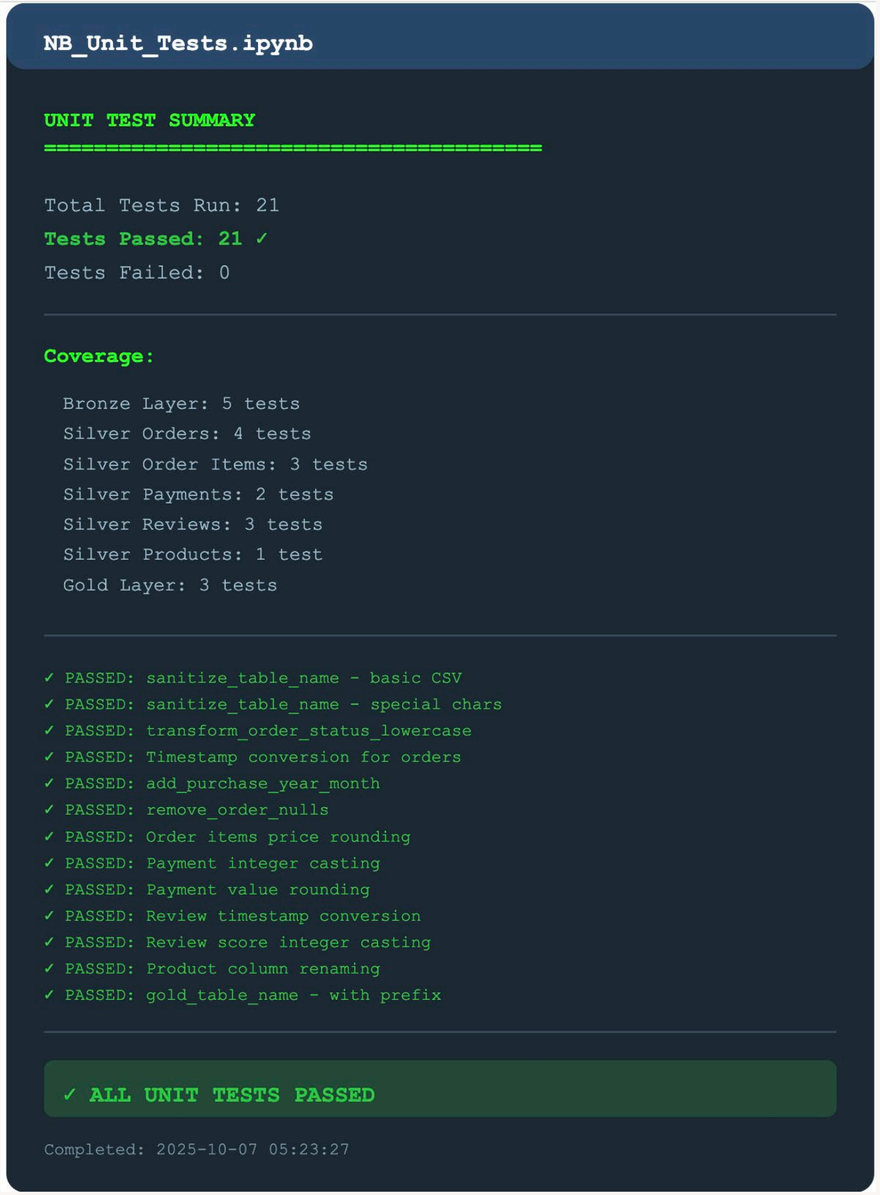
Testing Activities

💻 Code Validation: Unit Testing (Pytest)

- **Framework: Pytest** is leveraged to test **small, isolated functions** using fabricated data
- **Focus:** Validates the core logic of **21 individual, reusable transformation functions** (e.g., data type casting, rounding, sanitization) across all data layers
- **Isolation:** Functions are tested using small, fabricated **Spark DataFrames** or direct inputs, fully **isolating the transformation code** from the full pipeline environment

⚙️ Framework Validation: Integration Testing

- **Focus: Code-to-Engine Interaction**, validating the **seamless integration** between custom code and the **PySpark execution environment**
- **Goal:** Confirms functions correctly **interact with the Spark API** itself
- **Verification:** Ensures successful invocation of and data passing to core PySpark primitives (e.g., `to_timestamp()`, `spark_round()`, and `.na.drop()`)
- **Outcome:** Guarantees that transformations produce the **intended Spark DataFrame structure** correctly at the framework level



Validation of Pipeline Activities

Pipeline

Multi-stage validation framework across the Lakehouse architecture, ensure data integrity and schema consistency

Azure Blob Storage

NB_Bronze_ Copy_Validation

1. Expected files copy from Blob Storage to Lakehouse.
2. Primary and Foreign key are present in fact table.

Bronze Lakehouse

NB_Bronze_ Validation

1. Files loaded from Blob Storage converted to delta tables and save into Lakehouse.
2. No schema shift. Compare schema of delta tables with schema declared.

Silver Lakehouse

NB_Silver_ Validation

1. All tables copied from Bronze to Silver Lakehouse.
2. No schema shift.
3. Data integrity check that Primary key is not null

Gold Lakehouse

NB_Gold_ Validation

1. All tables copied from Silver to the Gold Lakehouse.
2. Schema of tables in Silver and Gold Lakehouse is the same.
3. Row count of tables in both Silver and Gold Lakehouse is the same.

Olist_Workspace

Validation of Pipeline Flow

Screenshot of the successful execution of the pipeline flow across the Bronze, Silver and Gold Lakehouse.

Each stage includes automated email notifications triggered upon failure / success, ensure traceability and rapid response.

Bronze Lakehouse

Activity name	Activity status
Error_Email_Notification_Bronze	Succeeded
Bronze_DeltaTable	Failed
Bronze_CopyBlob	Succeeded

Silver Lakehouse

Activity name	Activity status
Error_Email_Notification_Silver	Succeeded
Silver_Cleaning	Failed
Bronze_Validation	Succeeded
Bronze_DeltaTable	Succeeded
Bronze_CopyBlob	Succeeded

Gold Lakehouse

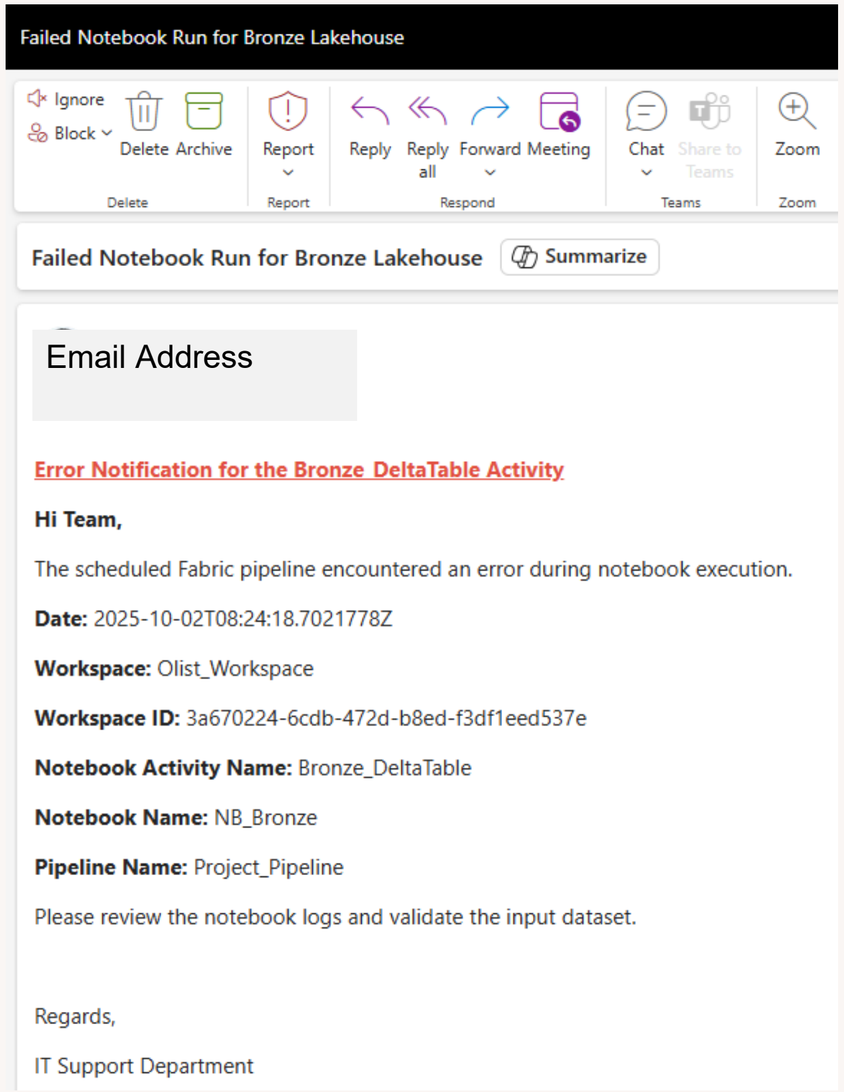
Activity name	Activity status
Error_Email_Notification_Gold	Succeeded
Gold_Transform	Failed
Silver_Validation	Succeeded
Silver_Cleaning	Succeeded
Bronze_Validation	Succeeded
Bronze_DeltaTable	Succeeded
Bronze_CopyBlob	Succeeded

Semantic Model Refresh

Activity name	Activity status
New_Report_Notification	Succeeded
ML_Semantic_model	Succeeded
Gold_Validation	Succeeded
Gold_Transform	Succeeded
Silver_Validation	Succeeded
Silver_Cleaning	Succeeded
Bronze_Validation	Succeeded
Bronze_DeltaTable	Succeeded
Bronze_CopyBlob	Succeeded

Validation of Pipeline Flow

Error Email Notification when the activity of the pipeline fails



Failed Notebook Run for Bronze Lakehouse

Ignore Block Delete Archive Report Reply Reply all Forward Meeting Chat Share to Teams Zoom

Delete Report Respond Teams Zoom

Failed Notebook Run for Bronze Lakehouse Summarize

Email Address

Error Notification for the Bronze_DeltaTable Activity

Hi Team,

The scheduled Fabric pipeline encountered an error during notebook execution.

Date: 2025-10-02T08:24:18.7021778Z

Workspace: Olist_Workspace

Workspace ID: 3a670224-6cdb-472d-b8ed-f3df1eed537e

Notebook Activity Name: Bronze_DeltaTable

Notebook Name: NB_Bronze

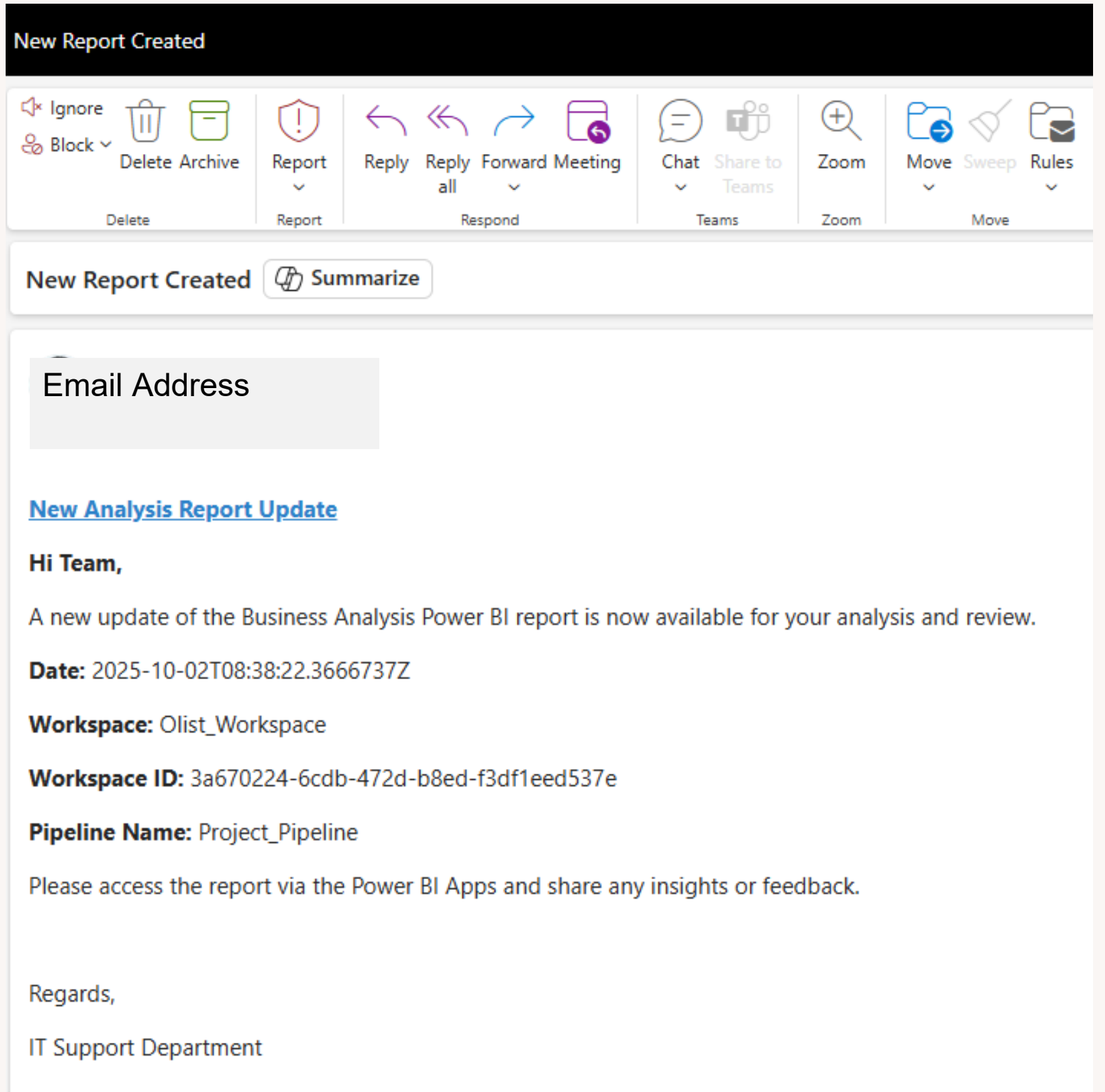
Pipeline Name: Project_Pipeline

Please review the notebook logs and validate the input dataset.

Regards,

IT Support Department

New Report Email Notification when the semantic model refresh is completed



New Report Created

Ignore Block Delete Archive Report Reply Reply all Forward Meeting Chat Share to Teams Zoom Move Sweep Rules

Delete Report Respond Teams Zoom Move

New Report Created Summarize

Email Address

New Analysis Report Update

Hi Team,

A new update of the Business Analysis Power BI report is now available for your analysis and review.

Date: 2025-10-02T08:38:22.3666737Z

Workspace: Olist_Workspace

Workspace ID: 3a670224-6cdb-472d-b8ed-f3df1eed537e

Pipeline Name: Project_Pipeline

Please access the report via the Power BI Apps and share any insights or feedback.

Regards,

IT Support Department

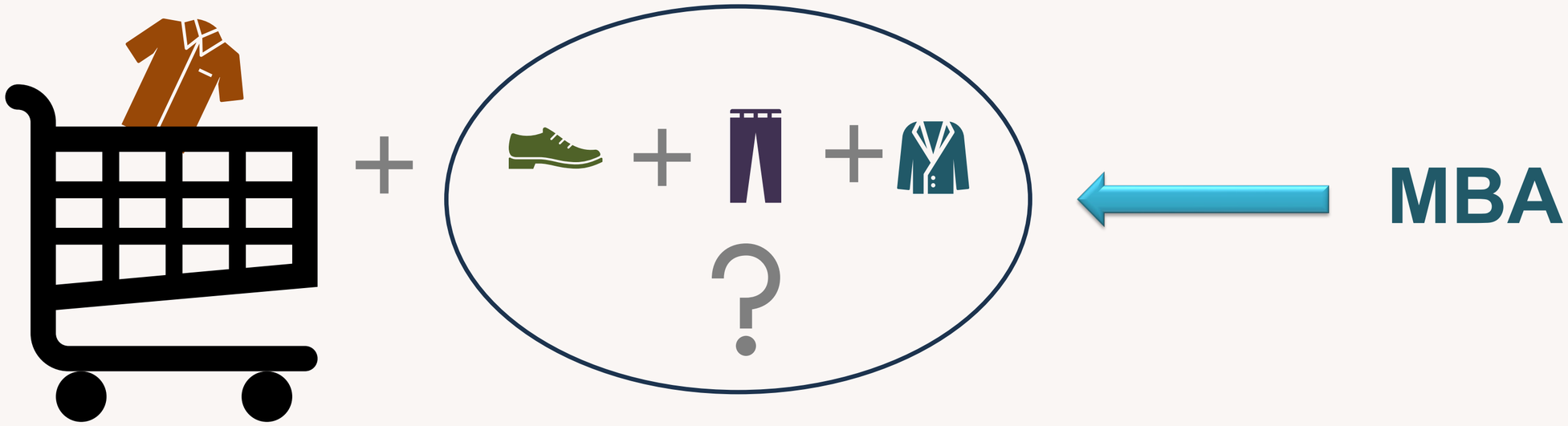
MARKET BASKET ANALYSIS

Market Basket Analysis (MBA)

olist

Ecommerce platform which connects sellers to multiple marketplaces and centralizes ecommerce operations

Business Goal: **$ Profit**



“Igniting every cart with smarter, bigger buys.”

Market Basket Analysis (MBA)

What is MBA?

How likely is a customer buying product B if the shopping cart has product A.

Frequent item sets: Which products are often bought together

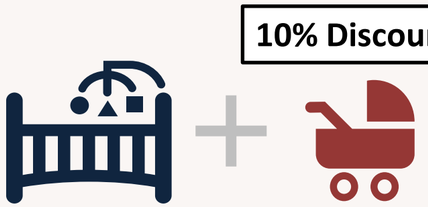
What is the use cases and impact to Olist ?

1. **Cross-Selling and Upselling Recommendations**
   - Identify frequently co-purchased items.
   - Recommend bundles or complementary products during checkout.

Buddle purchase

Upselling products

2. **Marketplace Insights and Seller Intelligence**
   - Provide sellers with actionable insights on which is the best product combinations.
   - Help optimize listings, bundle strategies, and pricing to increase visibility and conversion.



Smart Promotions

Market Basket Analysis (MBA)

How to implement MBA ?

Apriori Algorithm

1) Generate **Frequent Item Sets** in the database
2) Generate **Association Rules** from frequent item sets

What is Association Rules ?

Order Number					
001	✓			✓	1
002	✓	✓			
003	✓		✓	✓	2
004	✓				
005		✓	✓		
006	✓				
007	✓			✓	3
008	✓	✓			
009			✓		
010	✓			✓	4

LHS (antecedents)	RHS (consequents)	FREQUENCY	SUPPORT	CONFIDENCE	LIFT
Shirts	Pants	4	0.4	0.5	1.25

Frequency: No of times shirts and pants occurred in the transaction.
= Freq(LHS, RHS) = 4

Support: Frequency of item sets (LHS + RHS) in total transactions.
= Freq (LHS,RHS) / Total Transactions
= 4/10 = 0.4

Confidence: Likelihood of RHS being bought when LHS is bought.
= Freq (LHS,RHS) / Freq (LHS)
= 4/8 = 0.5

Lift: Strength of association beyond random chance.
= Support / (Support (LHS) x Support (RHS))
= 0.4 / (8/10 x 4/10) = 1.25

Lift > 1 -> both items appear together more often than expected by chance

Market Basket Analysis (MBA)

What is the process flow in Microsoft Fabric ?

Environment

- Set up new notebook environment for mlxtend external module.
- Import Apriori, association rules.
- Set up fabric experiment for ML flow and enable logging.

Load Tables

- Load tables from Gold Lakehouse. Combine required table.
- Create market basket matrix for association rule mining and remove low frequency columns from the matrix.

Apriori Algorithm

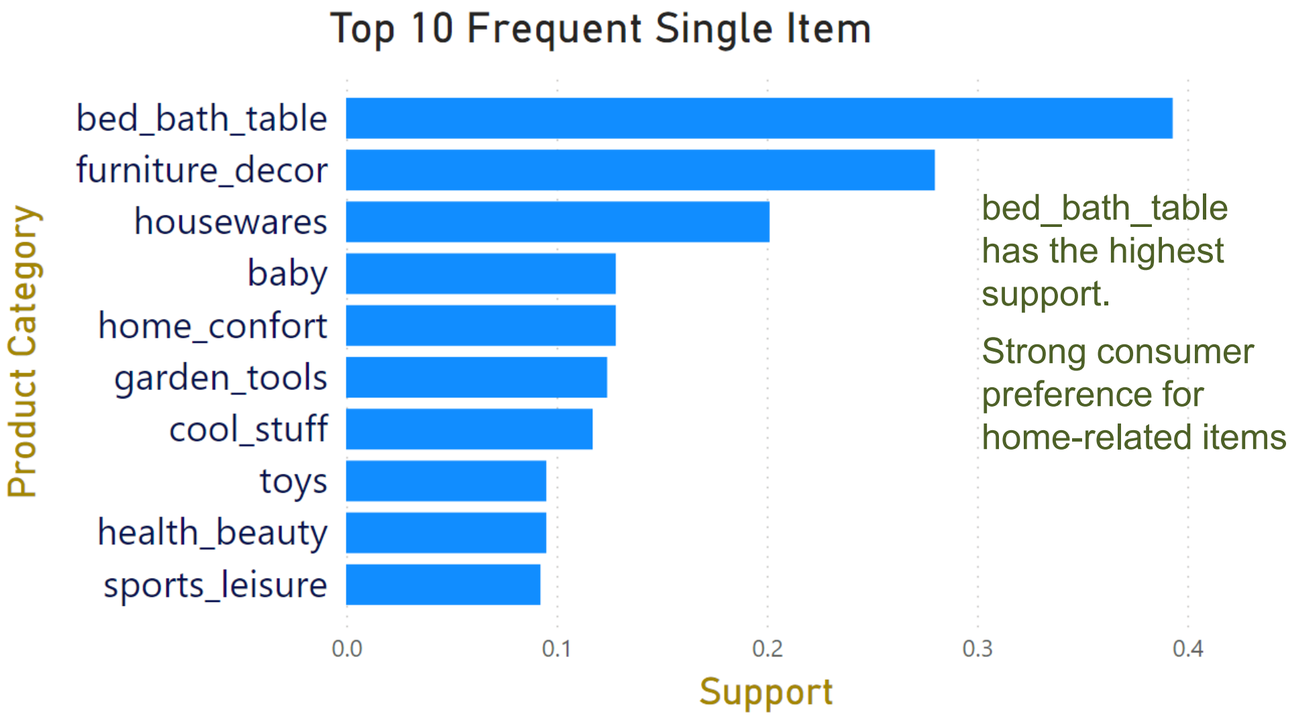
- Generate frequent item sets and association rules. Save both as Delta table in Gold Lakehouse.

Semantic Model

- Create Semantic model and generate new report with Power BI.
- Analyze the frequent item sets and association rules.

Market Basket Analysis

Frequent Single Item



bed_bath_table has the highest support.

Strong consumer preference for home-related items

IMPLICATIONS

- Categories for cross-selling and bundle promotions.
- Prioritize in home page placement, recommendation engines or email campaigns.

Frequent Item Sets

item_1	item_2	support
bed_bath_table	furniture_decor	0.10
home_confort	bed_bath_table	0.06
housewares	furniture_decor	0.03
cool_stuff	baby	0.03
housewares	bed_bath_table	0.03
toys	baby	0.03
bed_bath_table	baby	0.02
garden_tools	furniture_decor	0.02

bed_bath_table -> 4 out of 8 pairs. It's a Hub item

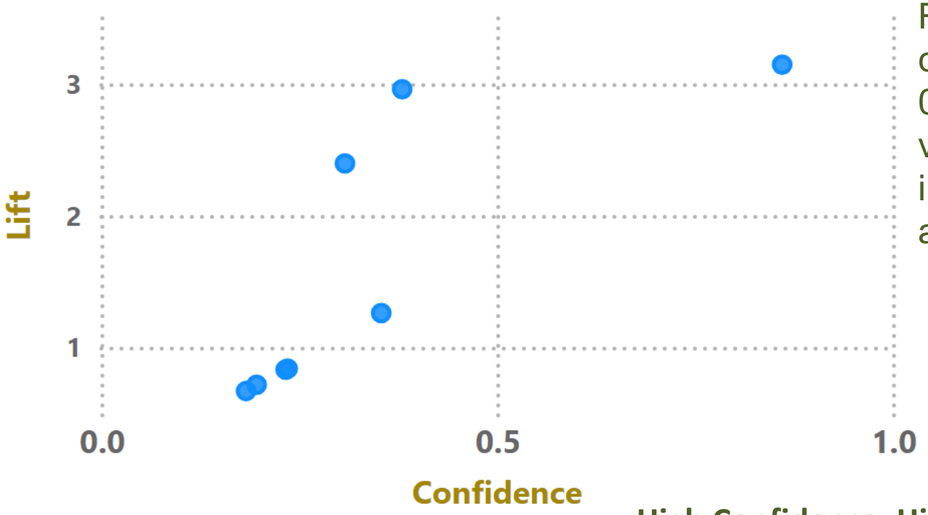
(cool_stuff, baby) and (toys, baby) -> family-oriented purchasing pattern. Target parents or gift buyers.

IMPLICATIONS

- Use these associations to design combo deals or discount bundles.
- Trigger real-time recommendations during check out.
- Inform home page design to group related items.

Market Basket Analysis

Confidence and Lift



Rules cluster in the confidence range of 0.2–0.4, with lift values above 1, indicating meaningful associations.

High Confidence, High Lift represent strong, reliable rules

- Recommendation engines
- Cross-selling strategies
- Personalized promotions

Association Rules

Association Rules with Lift > 1

antecedents	consequents	confidence	lift	
home_confort	bed_bath_table	0.86	3.15	1
toys	baby	0.38	2.96	2
cool_stuff	baby	0.31	2.40	3
bed_bath_table	furniture_decor	0.35	1.26	4

Rules with **lift greater than 1**, identify item pairings that occur more frequently together than by random chance.

IMPLICATIONS

1. Prioritize pair for bundling, homepage recommendation and inventory planning.
2. Targeted promotions for baby products when toys are purchased.
3. Lower confidence but still a strong lift—this might reflect seasonal or novelty-driven purchases. Consider event-driven campaigns.
4. This rule has the lowest lift indicating a mild positive association. Useful for low-risk cross-selling.

ASSESS THE QUALITY OF RULES

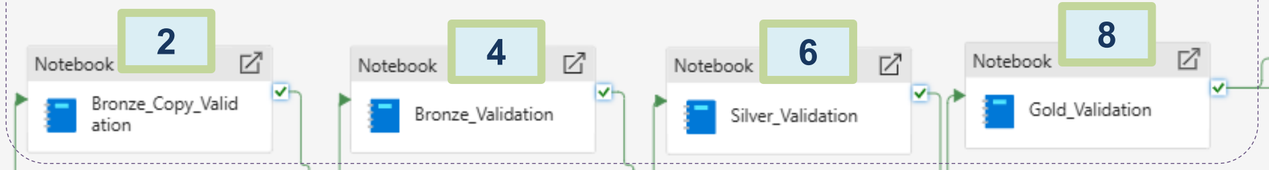
Confidence: Measures how often the rule has been found to be true.
Lift: Indicates how much more likely the consequents is to appear within the antecedent than by random chance.

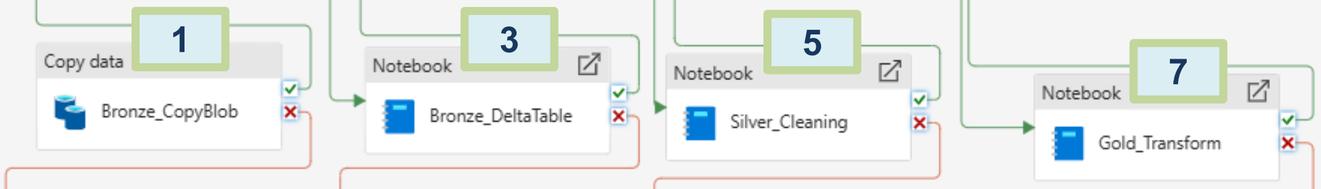
PROJECT DELIVERABLES

PROJECT PIPELINE

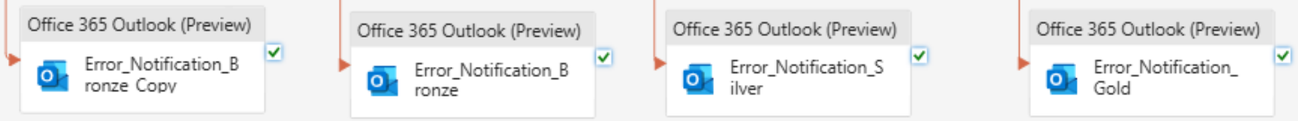
End-to-end pipeline from validation to reporting, engineered with a robust, modular, and scalable architecture.
It spans the full lifecycle from Medallion-based transformations.
Standalone testing framework for comprehensive quality assurance.

Validation: Schema consistency and data integrity



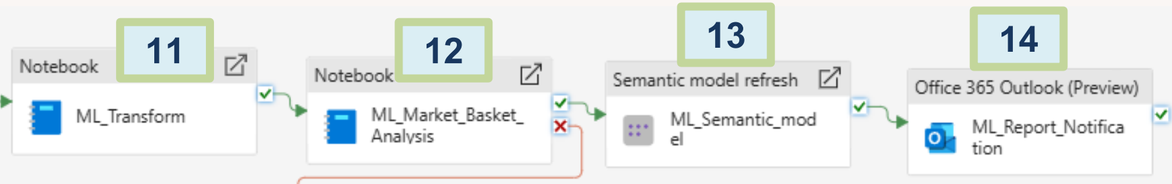


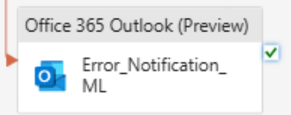
Medallion architecture: Separation between raw ingestion and refined transformations



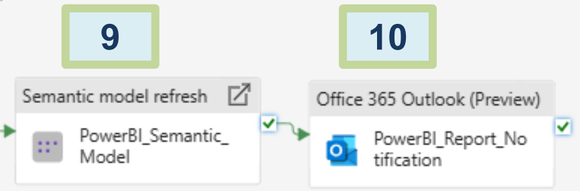
Errors Handling: Errors are captured and route through outlook for notification

Market Basket Analysis: Generate frequent item sets and association rules

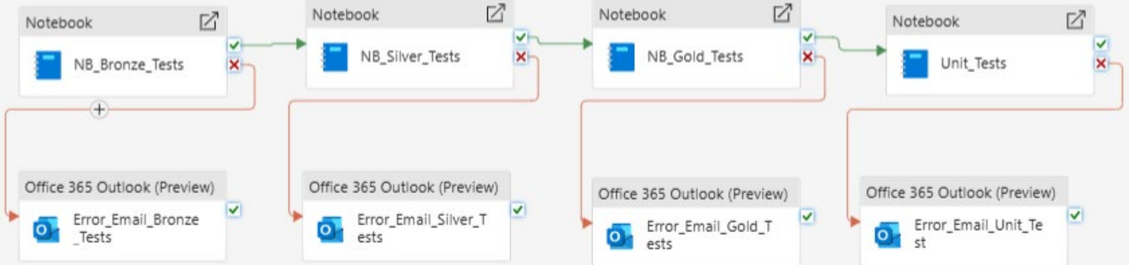




Errors Handling



Power BI: integrates with semantic layer to deliver stakeholder ready dashboards



Standalone Testing Frame: validates schema and data integrity across all Medallion layers

PROJECT PIPELINE

Screenshot showing the pipeline status with all activities marked as succeeded

Parameters | Variables | Settings | **Output** | Library variables

Pipeline run ID 764453da-711d-4212-b2d5-1a6d112f8abf View run detail **Pipeline status** Succeeded

Filter by keyword Showing 1 - 14 items

Activity name	Activity status	Run start
ML_Report_Notification	Succeeded	10/8/2025, 11:43:09 AM
ML_Semantic_model	Succeeded	10/8/2025, 11:42:25 AM
ML_Market_Basket_Analysis	Succeeded	10/8/2025, 11:33:53 AM
PowerBI_Report_Notification	Succeeded	10/8/2025, 11:33:04 AM
ML_Transform	Succeeded	10/8/2025, 11:32:23 AM
PowerBI_Semantic_Model	Succeeded	10/8/2025, 11:32:23 AM
Gold_Validation	Succeeded	10/8/2025, 11:30:37 AM
Gold_Transform	Succeeded	10/8/2025, 11:28:35 AM
Silver_Validation	Succeeded	10/8/2025, 11:27:23 AM
Silver_Cleaning	Succeeded	10/8/2025, 11:25:09 AM
Bronze_Validation	Succeeded	10/8/2025, 11:24:13 AM
Bronze_DeltaTable	Succeeded	10/8/2025, 11:21:58 AM
Bronze_Copy_Validation	Succeeded	10/8/2025, 11:21:12 AM
Bronze_CopyBlob	Succeeded	10/8/2025, 11:20:49 AM

VISUALIZATION AND REPORTING

Setting the context right…

👔 **Persona 4: Matthew Taylor – Strategy Director**

As Strategy Director, Matthew ’s role is to ***drive marketplace growth***, ***profitability, and customer trust***. He needs a single, consolidated view *of* ***business performance across sales, customer retention, delivery efficiency, and seller reliability.***

Problem Statement

- Leadership **lacks a consolidated view** of sales, customer retention, delivery, and seller performance on one page.

Project Objective

- To deliver an **executive-level summary** that brings together all marketplace KPIs into a single, intuitive dashboard.

1. Are we growing in the right places?

“As Strategy Director, I need to quickly see whether our marketplace is expanding in the right categories and regions — not just topline revenue.”

2. Are customers staying with us?

“Our growth is hollow if customers don’t return. I need clarity on churn vs. loyalty.”

3. Are our goods delivered timely as promise?

“Delays erode trust fast. I want an early-warning system on delivery performance.”

4. Are sellers strengthening or harming trust?

“Our sellers *are* the marketplace. Good ones build trust; weak ones destroy it.”

5. Where are the biggest risks this month?

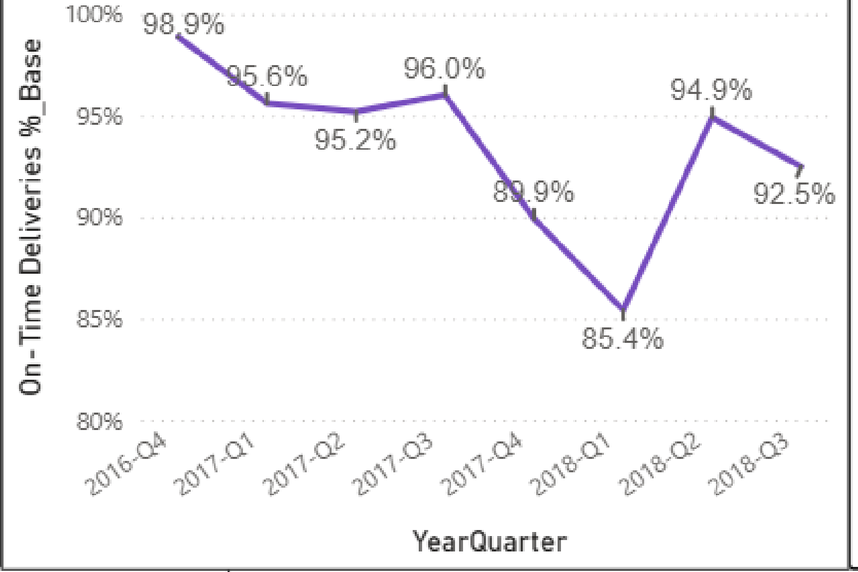
“I don’t want 20 pages of KPIs — I need a red flag list of risks now.”

Executive Summary (Oct 2016 -Aug 2018)

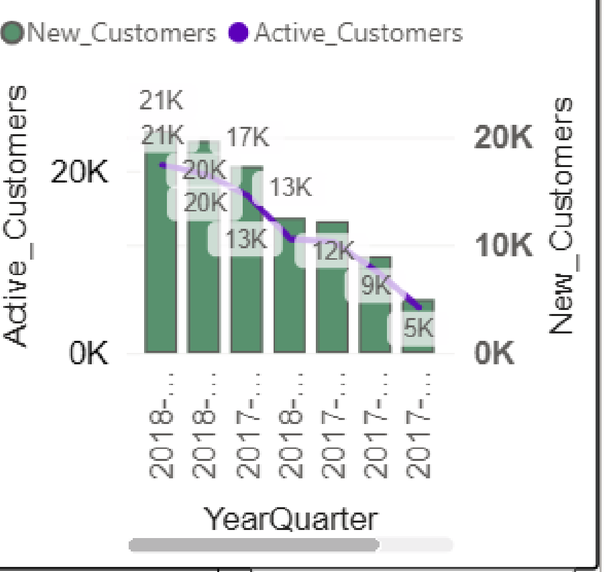
Gross Sales ($M)	Total Orders #counts	Avg Order Value ($)	On-Time Deliveries %	Avg Review Score_Seller
16M	99K	161	92%	4.1

Gross Sales Trend Line

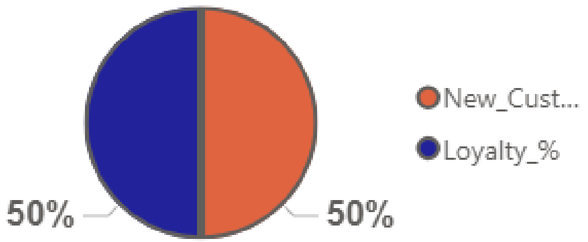
On-Time Delivery Trend



Customer Growth



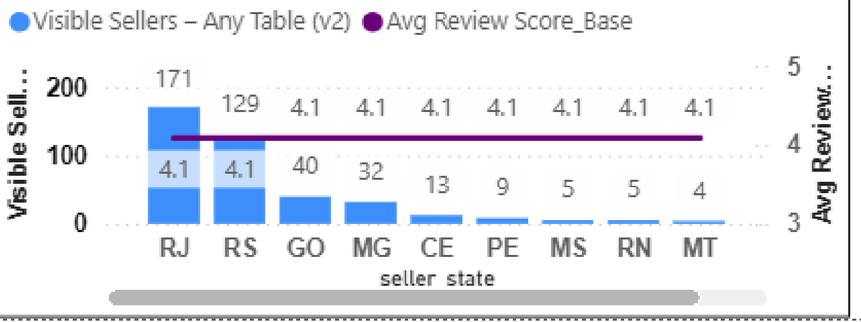
Customer Retention



Executive Risk Snapshot

Risk Factor	Severity %
Low Customer Ratings	17.90
Worst On-Time %	8.10
Worst Cancellation %	0.60
Low Sales Volume	0.00

Seller Review Distribution - Top States



YearQ...

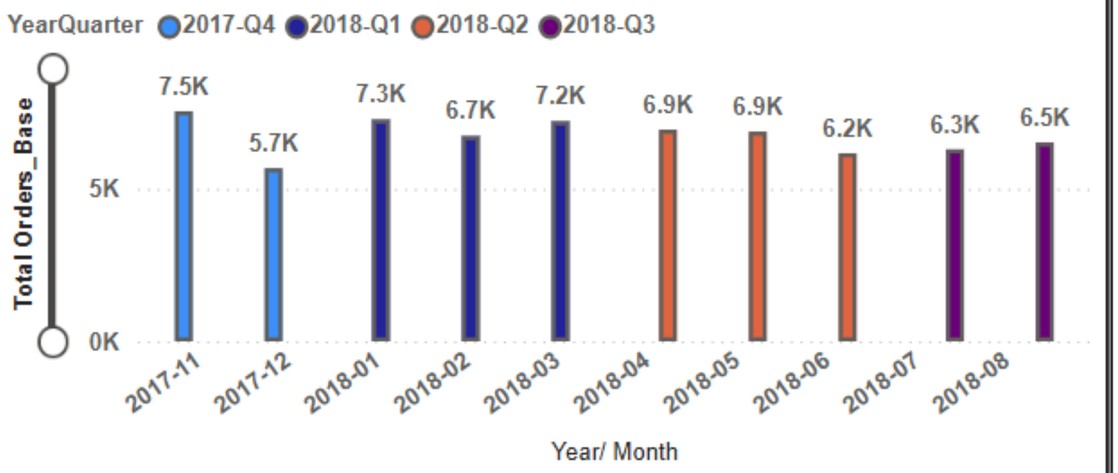
All

10
Top N (Selected)

01 Growth KPIs (Oct 2016 -Aug 2018)

Total Orders #counts	Avg Order Value ($)	Gross Sales ($M)
99K	161	16M

Gross Sales Trend by Year/Month ($)



Top 10 Cities by Gross Sales

Delivered orders (Oct 2016 – Aug 2018)

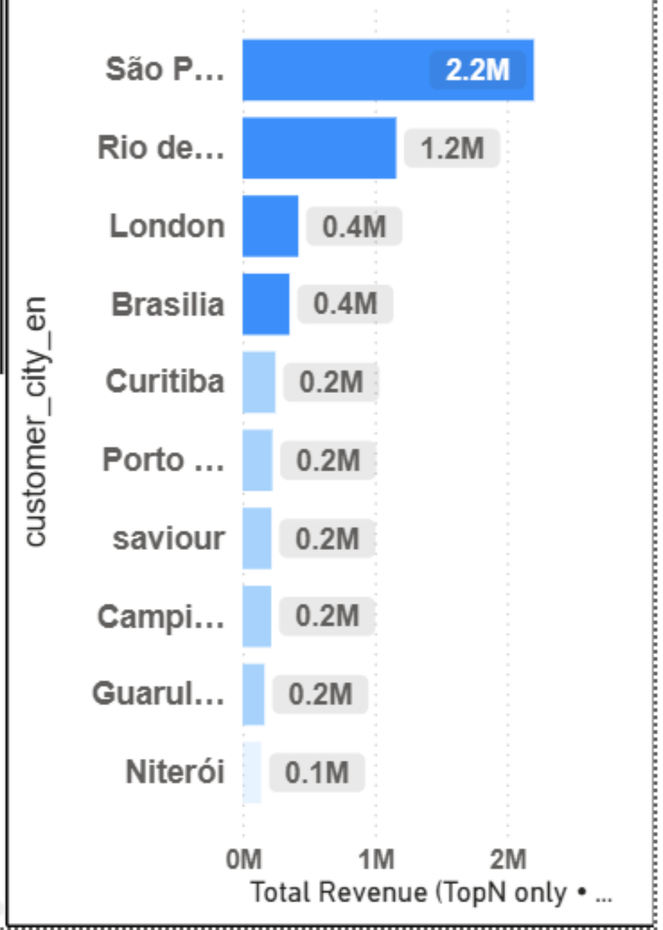
Top N: 10

YearQuar…: All

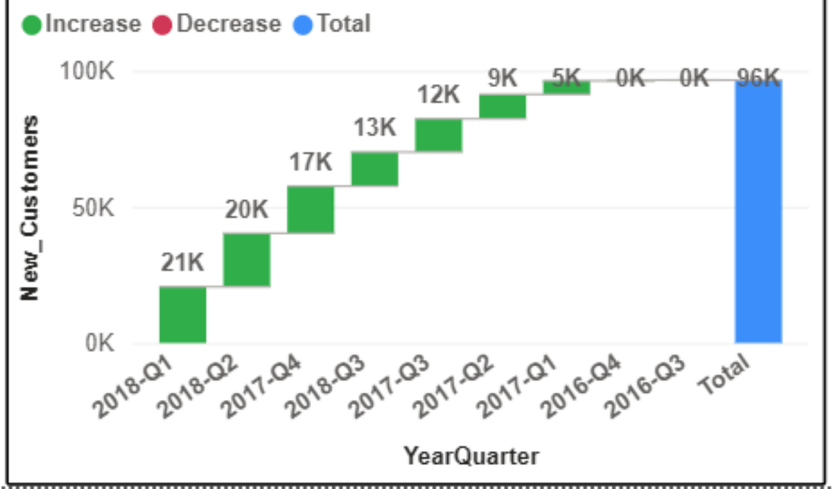
YearMonth: All

Top 10 Cities by Gross Sales

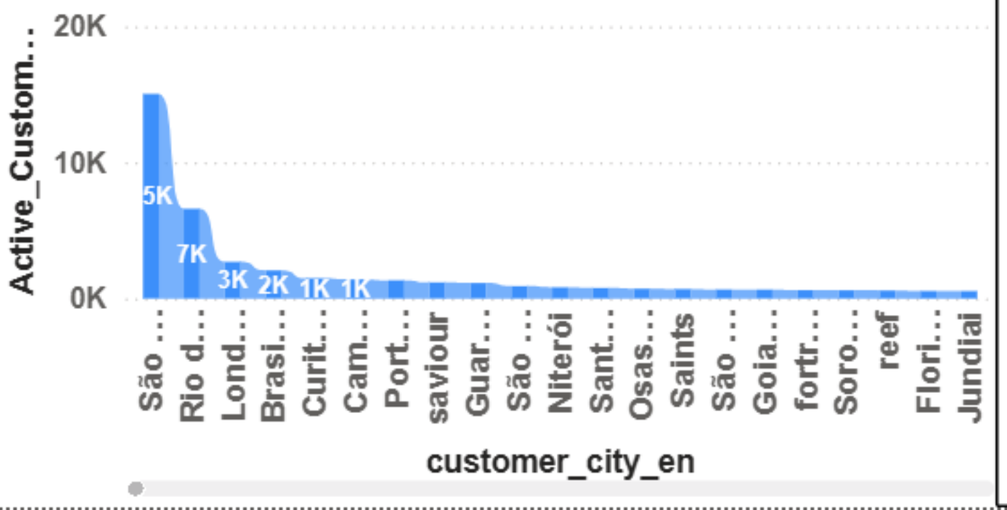
Delivered orders (Oct 2016 – Aug 2018)



Customer Growth by Quarter



Active Customers by City

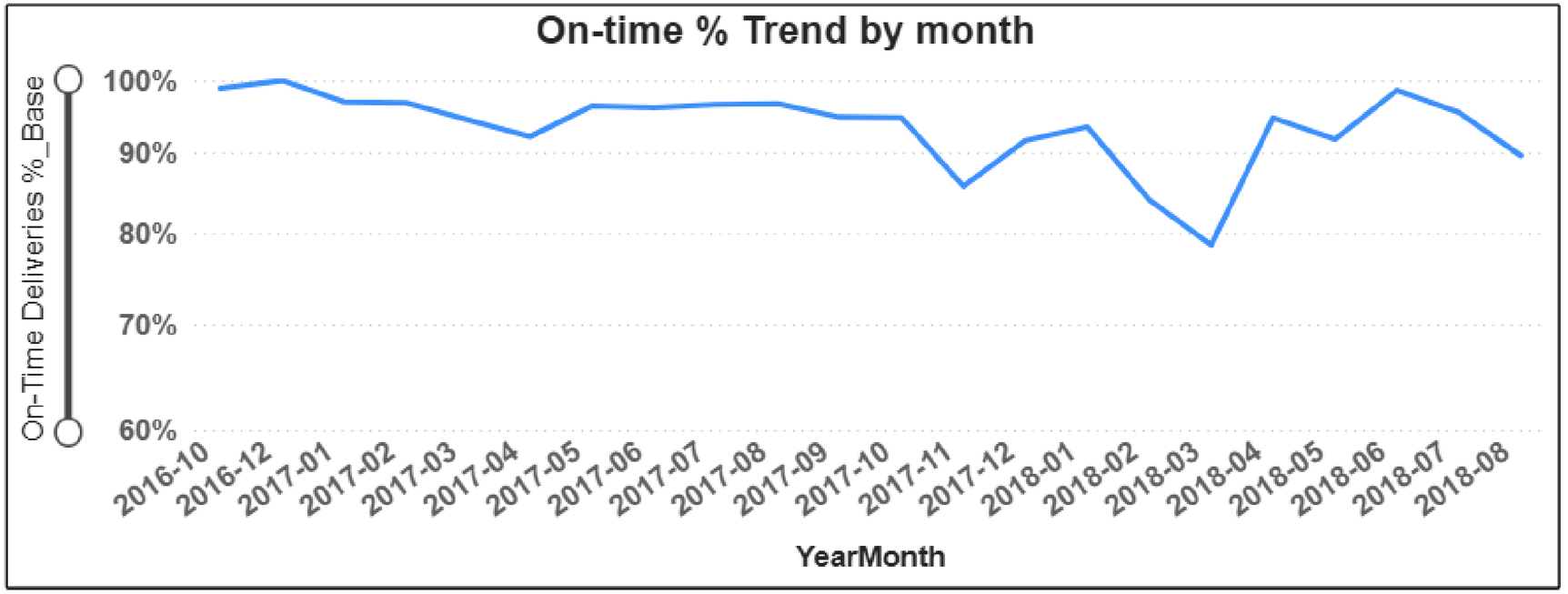


02 Delivery Performance (Oct 2016 -Aug 2018)

Delivered orders	On-Time Orders Delivered	On-Time Deliveries %	Avg Delivery Days (Purchase Date → Actual Delivery Date)	Avg Delivery Delay (Late Orders Only)
96K	89K	92%	12.56	0.77

- **Delivered Orders**
  *Total number of orders that were **successfully delivered** within the analysis period.*
- **On-Time Orders**
  *Number of delivered orders that **arrived on or before** the promised delivery date.*
- **On-Time Deliveries %**
  *Percentage of delivered orders that were on time = **(On-Time Orders ÷ Delivered Orders).***
- **Avg Delivery Days (Purchase → Actual Delivery)**
  *Average number of calendar days between when **a customer places an order and when it is delivered.***
- **Avg Delivery Delay (Late Orders Only)**
  *Average number of days delayed, but **only counting orders that arrived after the promised date.***

YearMonth
- Select all
- 2018-10
- 2018-09
- 2018-08
- 2018-07
- 2018-06
- 2018-05
- 2018-04
- 2018-03
- 2018-02
- 2018-01
- 2017-12



03 Customer Retention (Oct 2016 -Aug 2018)

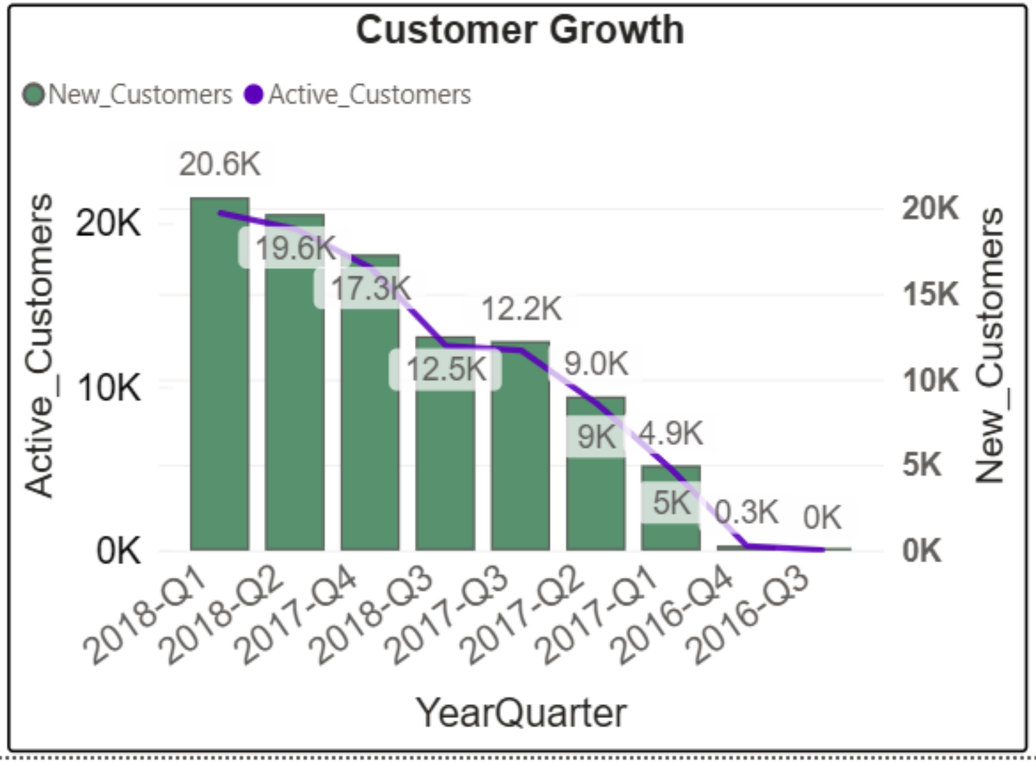
Quarterly new-customer acquisition is slowing.

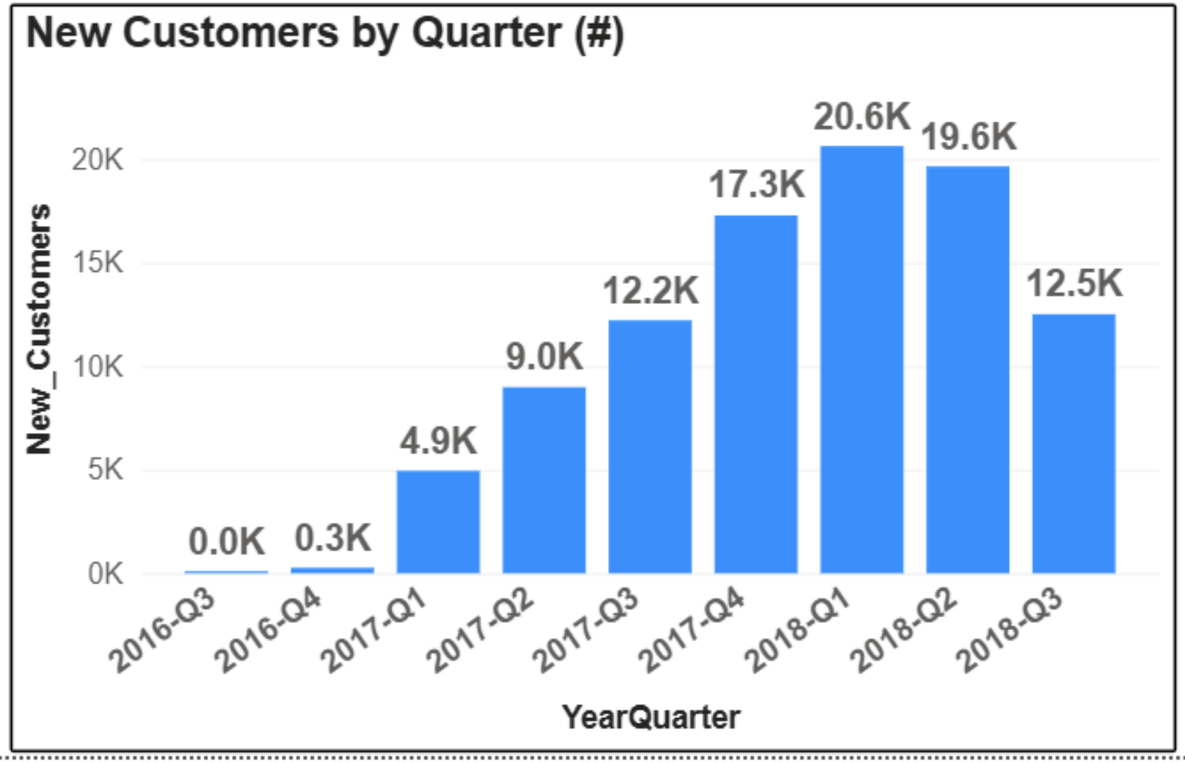
At the same time, the real story in the data is that active customers average 6–7K per month, but are trending downward into 2018.

Repeat purchases are rare, so overall lift depends on **winning back inactive customers and securing second purchases quickly**.

Possible Reasons For Decline in New Customers

- **Market Saturation** – Rapid early growth left fewer untapped customers.
- **Seasonality** – Q3/Q4 shifts in demand (holidays, school, budget cycles).
- **Operational Issues** – Bottlenecks in logistics or onboarding slowed growth.
- **Marketing Shift** – Less ad spend, more focus on retention/profitability.
- **Competition** – New rivals and promotions diverted potential customers.





04 Seller Trust & Ranking (Oct 2016 -Aug 2018)

On-Time Deliveries %	Delivered Orders
92%	96K

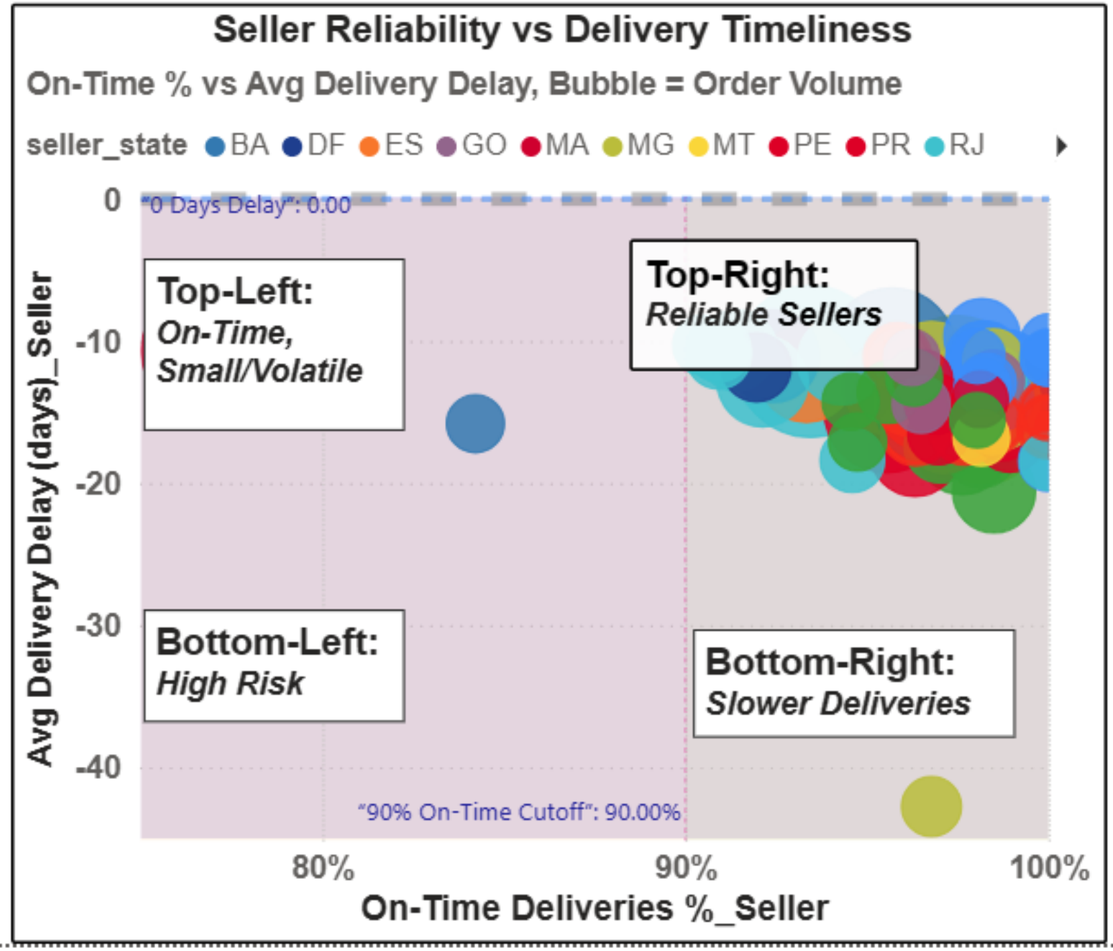
% Orders Cancelled	Delivery Delay (days)
0.5%	-11.20

YearMonth

Multiple s…

Seller Reliability & Order Performance

seller_id	Delivered Orders	On-Time Deliveries %
6560211a19b47992c3666cc44a7e94c0	1819	0.94
4a3ca9315b744ce9f8e9374361493884	1772	0.89
cc419e0650a3c5ba77189a1882b7556a	1651	0.94
1f50f920176fa81dab994f9023523100	1399	0.89
da8622b14eb17a	1311	0.92
Total	32788	0.92



Top N

10

seller_id

All

05 TOP 5 Risks

Worst On-Time % — Top 10 Sellers

seller_id	Visible Sellers – Any Table (v2)	On-Time Deliveries %_Seller	seller_state
001cca7ae9ae17fb1caed9dfb1094831	1	0.93	ES
001e6ad469a905060d959994f1b41e4f	1		RJ
02a2272692e13558373c66db98f05e2e	1		RJ
02d35243ea2e497335cd0f076b45675d	1	0.64	RN
Total	**424**	**0.83**	

Worst Avg Delay (days) — Top 10 Sellers

seller_id	Visible Sellers – Any Table (v2)	Avg Delivery Delay (days)_Seller	seller_state
001e6ad469a905060d959994f1b41e4f	1		RJ
02a2272692e13558373c66db98f05e2e	1		RJ
02d35243ea2e497335cd0f076b45675d	1	-5.45	RN
0417b067eeab773d2f7061a726dc477f	1	-3.21	SC
04843805947f0fc584fc1969b6e50fe7	1	0.63	RS
Total	**183**	**-10.47**	

Worst Cancellation Rate % — Top 10 Sellers

seller_id	Visible Sellers – Any Table (v2)	Cancellation Rate %_Seller	seller_state
001cca7ae9ae17fb1caed9dfb1094831	1		ES
001e6ad469a905060d959994f1b41e4f	1	1.00	RJ
003554e2dce176b5555353e4f3555ac8	1		GO
01bcc9d254e0143f0cc9791b960b2e47	1		GO
Total	**581**	**0.02**	

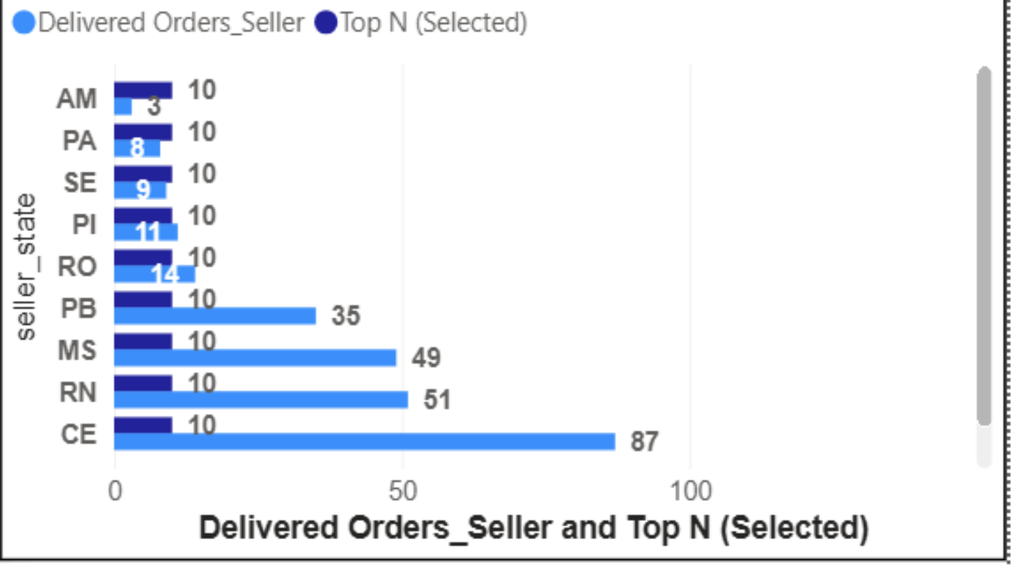
Total No. of Sellers

3095

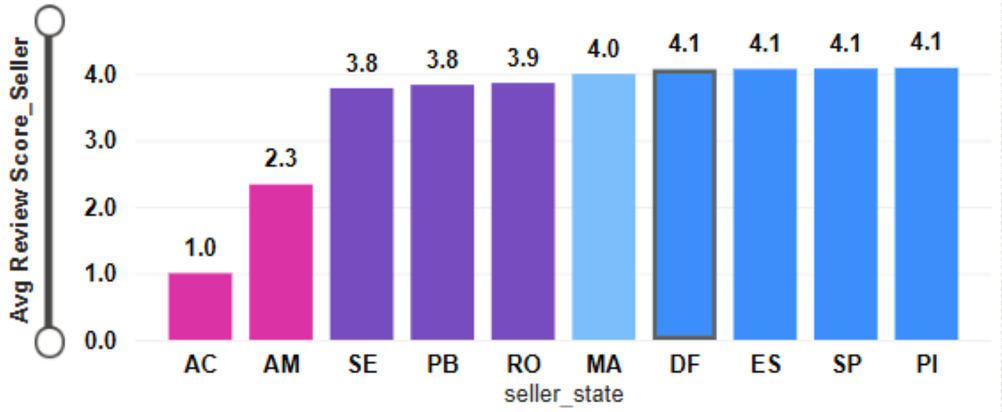
10
Top N (Selected)

Show Worst N
Top N
10

Low Sales Volume – Showing Bottom 10 Sellers



Low Customer Ratings — Showing Bottom 10 Sellers



Risks Identified & Action Steps

The Risk Snapshot shows three hot spots: **low ratings, on-time delivery gaps, and cancellations**.

Action: roll out an **Earned Trust program** — coaching and incentives tied to OTD, cancellation, and review lifts.

Seller reviews are concentrated in top states but flat at ~4.1.

Focus needs to be on lifting underperforming high-volume states.

Net: growth is real, but we protect it by pushing OTD ≥95% and reviews ≥4.3."

Monthly sales confirm seasonality spikes — we should replicate peak-month playbooks two months early. Revenue is highly concentrated in a handful of states.

Action:

Strategy: **land-and-expand** in the top five, and pilot two adjacent states with similar demographics.

Quarterly new customers are slowing, so overall lift hinges on retention. We must win back inactive customers and secure second purchases within 30 days.

By city, we see a long tail of small but meaningful customer bases.

Our next $1M comes from targeting the **next 10 cities** with lightweight demand gen and faster delivery promises.

Decision: Double down on top-5 states and surgically activate the next 10 cities."

Growth is strong, retention is steady, but seller reliability is our weak spot To scale, we must *protect trust* with OTD ≥95%, reviews ≥4.3, and a ruthless focus on reliable sellers."

Risks Identified & Action Steps

“Here’s execution reliability.

We delivered **96K orders**, **89K on time** → **92% OTD**. Average delivery is **12.6 days** from purchase, and late orders are only **0.77 days delayed**.

Good news: delays are shallow; bad news: OTD is inconsistent.

Trend line shows months in the mid-80s OTD, then recovery to ~95%. This volatility is the biggest threat to trust.

Action: enforce stricter SLAs with lagging sellers/carriers, and fund predictive monitoring to flag risks before they damage NPS.”

“Here’s highlights our seller-level reliability.

Overall: **92% OTD, 96K orders, cancellation only 0.5%**. But within that, large sellers dip to **89% OTD**, below our 90% cutoff.

The scatterplot shows clusters of sellers consistently underperforming while still driving volume — structural risk.

Action: implement a **Seller Reliability Program** with a 95% OTD minimum threshold, tiered performance recognition, and penalties for laggards.”

“Now to loyalty.

Headline metrics show **100% retention, new, and repeat purchase** (due to setup quirk). The real story is in the detail: **6–7K active customers per month, trending downward** into 2018.

New vs returning looks balanced, but churn is high — we’re replacing lost customers with new ones instead of compounding loyalty.

Action: approve retention programs. Specifically:

- Loyalty reward for **second purchase within 30 days**
- Win-back campaigns for customers inactive beyond 90 days.”

“Finally, here’s where exposure lives.

- **Worst OTD %**: Some sellers as low as **83%**, concentrated in RJ, ES, RN.
- **Worst Avg Delays**: >10 days in RJ and RN.
- **Worst Cancellation Rates**: up to 100%, abandoned sellers.
- **Low Sales Volume**: 3K+ sellers contribute negligible orders.
- **Low Ratings**: states AC and AM average **1.0–2.3 ratings**, damaging trust.

Action: approve a **‘Cut or Coach’ program**. We offboard bottom-decile sellers, retrain salvageable ones, and protect expansion states by removing sub-3.5 rating sellers.

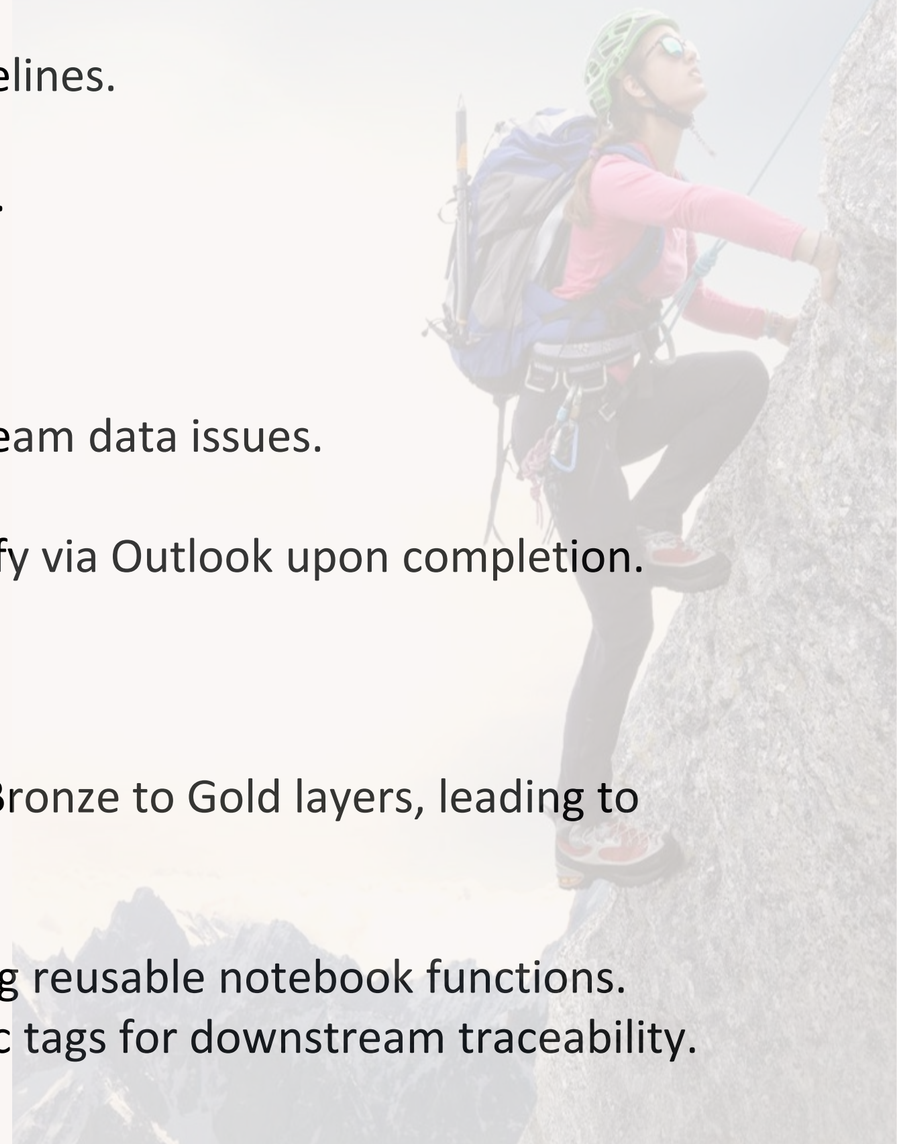
CHALLENGES
AND
FUTURE ROADMAP

Challenges and Solutions

1. **Ensuring robust errors handling in notebooks**
   Notebook failures (e.g., schema mismatch, missing files) can silently break pipelines.
   Solution:
   Implement try-except blocks and trigger Outlook alert with actionable context.

2. **Power BI report fail to update with new data**
   Power BI semantic models may fail to refresh due to schema changes or upstream data issues.
   Solution:
   Automate refresh using Fabric pipeline triggers semantic model fresh and notify via Outlook upon completion.

3. **Inconsistent data validation across pipeline stages**
   Data anomalies (e.g., nulls, outliers, schema drift) may pass undetected from Bronze to Gold layers, leading to misleading insights or report failures.
   Solution:
   Design modular validation checkpoints at each layer (Bronze, Silver, Gold) using reusable notebook functions.
   Trigger alerts via Outlook when anomalies exceed thresholds, and log semantic tags for downstream traceability.



Impact and Future Roadmap

IMPACT

1. Operational Efficiency

- Modular notebooks and reusable validation logic reduce manual intervention.
- Automated ingestion, transformation, and alerting pipeline execution.

2. Data Trust & Quality

- Multi-step validation ensures schema and data integrity.

FUTURE ROADMAP

1. Alerting & Observability

- Use of Power Automate integration to include Slack, Teams, or SMS alerts.

THANK YOU