# Databricks Olist Delivery Delay Prediction

End-to-End Data Engineering & Machine Learning Project

Build With Databricks – Codebasics Challenge

GitHub: Databricks-olist-delivery-delay-prediction

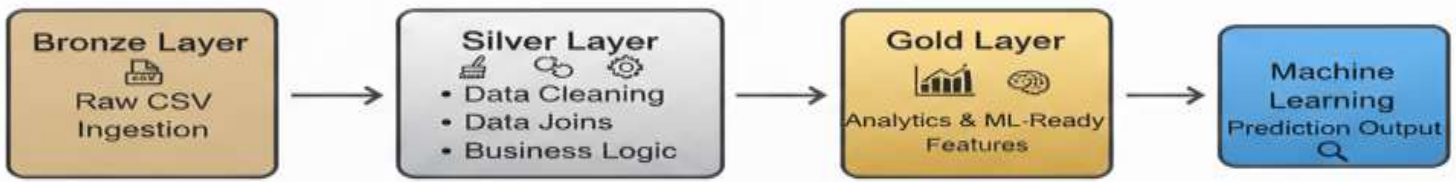
# Problem Statement

- Late deliveries impact customer satisfaction and seller performance
- Businesses need early signals to identify delivery risks
- Objective: Predict whether an e-commerce order will be delivered late

# Dataset Overview

- Olist Brazilian E-Commerce Dataset
- Real-world, multi-table dataset
- Includes orders, customers, and delivery timelines

# Databricks Medallion Architecture



# Data Transformation & Business Logic

- Deduplicated order records
- Joined orders with customers
- Implemented delivery delay logic
  - delivery_delay_days
  - is_delayed

# Gold Layer: Analytics & Features

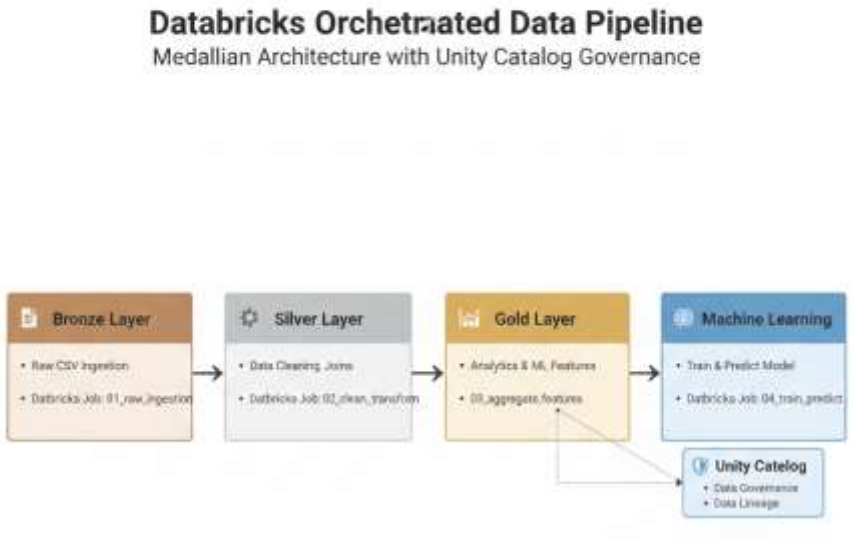
- Aggregated business metrics
- ML-ready feature tables
- Clean inputs for model training

# Machine Learning

- Problem type: Binary Classification
- Target variable: is_delayed
- Model: Logistic Regression
- Metrics: Accuracy, Precision, Recall
- Experiment tracking using MLflow

# Automation & Governance

- Automated pipeline using Databricks Jobs
- Data governance with Unity Catalog
- End-to-end data lineage



# Outcome & Business Impact

- Early identification of delayed deliveries
- Improved customer communication
- Better seller and logistics decisions

# Conclusion

- End-to-end Databricks Lakehouse solution
- Strong foundation in Data Engineering and Machine Learning
- Portfolio-ready project

Thank You