ARCHITECTURE DOCUMENT

INTRODUCTION

This architecture document describes the design and implementation of a cloud-native analytics platform built on Azure Databricks and Delta Lake for a multinational retail enterprise. The implementation focuses on scaling, incremental processing, and analytic ready data. The solution follows enterprise data engineering best practices using a multi-layered Bronze Silver Gold architecture to ensure data quality, and performance.

TECHNOLOGY STACK

1) Core Processing

- Azure Databricks
- Apache Spark (PySpark)

2) Storage

- Azure Data Lake Storage (via Databricks Volumes)

3) Analytics & Visualization

- Power BI (connected to Gold Delta tables)

ARCHITECTURE COMPONENTS

1) Bronze layer

- The Bronze layer stores raw, unmodified data exactly as received from source systems.
- Data ingested from CSV files stored in Databricks Volumes
- No business transformations applied yet on the data
- Enables reprocessing

2) Silver layer

- The Silver layer transforms raw data into clean, reliable, and standardized datasets suitable for analytics.
- It performs data preprocessing techniques like handling null data, duplicate data, standardising and normalizing data.
- Timestamp normalization to UTC
- Data calibration (financial accuracy) for total_amount.

3) Gold layer

- The Gold layer provides business-friendly, aggregated datasets optimized for reporting and dashboards.
- Optimized for Power BI consumption
- Analyses the daily sales analysis by understanding metrics like revenue, transactions, units sold, average order value
- Analyses the product performance by understanding metrics like revenue, units sold, average selling price
- Analyses the regional sales summary by understanding metrics like total revenue, transaction count, regional performance.

HIGH LEVEL ARCHITECTURE