Supply Chain Data Engineering Pipeline

DS 5110

Team Member: Ameen Shaik

NUID: 002534243

1 Dataset Description

Dataset Source:

- Public E-commerce Supply Chain dataset from Kaggle: https://www.kaggle.com/datasets/shashwat
  smart-supply-chain-for-big-data-analysis

Description: The dataset represents a large-scale e-commerce supply chain system containing sales, customer, shipping, and product-level details. It includes features such as `Order Date`, `Shipping Date`, `Delivery Status`, `Sales per Customer`, `Profit per Order`, `Product Price`, `Category`, and `Customer Location`.
Structure: Data will be organized into three layers following the Medallion Architecture:

- **Bronze Layer:** Raw ingested data from the source file into ADLS.
- **Silver Layer:** Cleaned, standardized data using Databricks (PySpark).
- **Gold Layer:** Aggregated business insights for analytics and reporting.

Why Suitable: This dataset captures multiple dimensions of a real-world retail supply chain and supports analytics such as delivery risk prediction, profitability tracking, and product category performance — making it ideal for demonstrating an end-to-end cloud data engineering pipeline.

2 Tools and Methodologies

Tools:

- **Azure Data Factory (ADF):** Data ingestion and orchestration.
- **Azure Data Lake Storage (ADLS Gen2):** Centralized data lake for Medallion architecture.
- **Databricks (PySpark):** Data transformation, cleaning, and modeling.

- **Power BI:** Visualization and business intelligence layer.
- **GitHub:** Version control and project collaboration.

Methodologies:

1. **Medallion Architecture:** Organize data into Bronze (raw), Silver (refined), and Gold (curated) layers.
2. **Unified Data Model (UDM):** Integrate sales, customer, product, and logistics data for consistent analytics.
3. **ETL/ELT Pipeline:** Automate ingestion → transformation → loading using ADF and Databricks.
4. **Cloud-Native Design:** Leverage Azure ecosystem for scalability, reliability, and reproducibility.

3 Preliminary Timeline

Week	Milestone / Task	Deliverable
Nov 3–Nov 9	Cloud infrastructure setup (ADLS, ADF, Databricks) — environment provisioning and connectivity validation	Azure resources provisioned and validated (setup started Nov 3)
Nov 10–Nov 16	Configure ADF pipelines and ingest dataset into ADLS (Bronze layer)	Raw dataset available in Bronze container
Nov 17–Nov 23	Develop Silver layer transformations in Databricks using PySpark (cleaning, validation, schema refinement)	Cleaned and structured Silver dataset
Nov 24–Nov 30	Design Gold layer with business-level aggregations (sales, delivery, profitability metrics)	Curated Gold layer ready for analytics
Dec 1–Dec 7	Build Power BI dashboards connected to Gold layer; finalize UDM	Interactive dashboards and analytical insights
Dec 8–Dec 14	Testing, performance validation, documentation, and final submission	Final report and demo-ready deliverables

4 Progress and Next Steps

Progress So Far:

- Cloud environment setup **started on Nov 3** and completed: Resource Group, ADLS Gen2, ADF, and Databricks workspace provisioned.
- Verified workspace connectivity between ADF and Databricks.
- Dataset identified and schema explored; ingestion yet to begin.

Next Steps:

- Configure ADF pipelines to load dataset into ADLS (Bronze layer).

- Begin transformation logic in Databricks notebooks for Silver layer.
- Implement Gold layer business KPIs and validate Unified Data Model.
- Connect Power BI for visualization and create dashboards.
- Document the entire process and prepare final presentation.