PROJECT DOCUMENTATION

DATA LAKE ANALYTICS

1. PROJECT OVERVIEW

This project aims to design and implement a Data Lake Analytics solution for e-commerce transactional data using Azure Data Factory (ADF), Azure Data Lake Storage Gen2, and Azure Databricks. The solution follows the Medallion Architecture (Bronze → Silver → Gold) to systematically ingest, clean, transform, and aggregate data into analytics-ready formats.

2. OBJECTIVES

- Ingest raw e-commerce data into Azure Data Lake (Bronze Layer).
- Clean, validate, and standardize data using ADF Mapping Data Flows (Silver Layer).
- Aggregate and prepare business-ready datasets (Gold Layer).
- Automate workflows with ADF pipelines and triggers.
- Enable advanced analytics and reporting through Power BI and ML models.

3. ABOUT THE PROJECT

E-commerce generates massive data from customers, products, orders, and payments. To extract insights, the data needs to be processed through a structured pipeline.

- **Bronze Layer:** Stores raw ingested data (unaltered, source of truth).
- **Silver Layer:** Stores cleaned and standardized data (validated, structured).
- **Gold Layer:** Stores aggregated, business-ready datasets for dashboards and ML.

By using this layered approach, the project ensures data quality, scalability, lineage, and business value extraction.

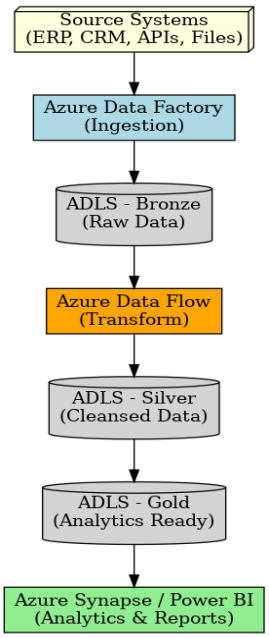
4. KEY BENEFITS

- **Centralized Storage:** All raw, cleaned, and business-ready data is managed in Azure Data Lake, ensuring a single source of truth.
- **Improved Data Quality:** Silver layer removes duplicates, nulls, and incorrect values, making the dataset reliable for analytics.

- **Business Insights:** Gold layer provides aggregated insights such as total sales by country, top customers, and revenue by month, supporting decision-making.
- **Customer Understanding:** Enables customer segmentation (e.g., frequent buyers, high-value customers) for targeted marketing.
- **Sales Optimization:** Identifies best-selling products, seasonal demand, and underperforming items.
- **Scalability:** Handles growing e-commerce data volumes without performance issues using Azure cloud resources.
- **Automation:** ADF pipelines and triggers automate the ETL process, reducing manual effort and errors.
- **Support for Advanced Analytics:** Gold layer data can be used for predictive models like customer churn prediction or recommendation systems.
- **Cost-Effectiveness:** Cloud-based pay-as-you-go model reduces infrastructure and maintenance costs.
- **Auditability & Lineage:** Bronze → Silver → Gold layering preserves data history and transformation steps, ensuring traceability.

5. ARCHITECTURE DIAGRAM

HIGH-LEVEL ARCHITECTURE FLOW:



5.1 ARCHITECTURE FLOW

Data Ingestion (Bronze Layer)

- E-commerce raw data (CSV files, databases, APIs) is ingested into Azure Data Lake Storage Gen2 using Azure Data Factory (ADF) pipelines.
- Data is stored in its original format (raw, unaltered) for audit and traceability.

Data Processing & Cleaning (Silver Layer)

- Raw data from the Bronze layer is cleaned, validated, and standardized.
- This step uses ADF Mapping Data Flows and Azure Databricks notebooks to remove duplicates, handle null values, and enforce data consistency.
- The output is structured, analytics-ready data stored in the Silver layer.

Data Transformation & Aggregation (Gold Layer)

- Silver data is transformed into aggregated, business-focused datasets.
- Examples include:
    - Total sales by country
    - Monthly revenue trends
    - Customer segmentation (high-value, frequent buyers)
- These curated datasets are stored in the Gold layer for reporting and advanced analytics.

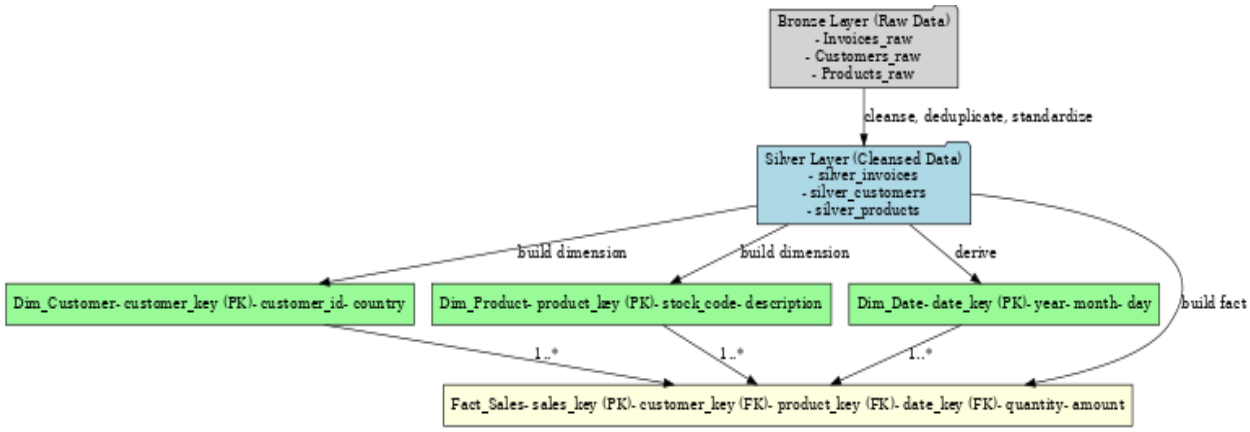
Analytics & Consumption

- Power BI connects to the Gold datasets for dashboards and interactive reporting.
- Machine Learning models (e.g., churn prediction, recommendation systems) are trained using Gold data in Azure Databricks.
- Business applications and APIs can also consume Gold data for operational insights.

Automation & Orchestration

- ADF pipelines and triggers orchestrate the entire process (Bronze → Silver → Gold).
- This ensures continuous data refresh, auditability, and minimal manual intervention.

6. ER DIAGRAM



6.1 ER MODEL

- **Customers** → CustomerKey (PK), CustomerID, Country
- **Products** → ProductKey (PK), StockCode (ProductID), Description, UnitPrice
- **Orders (Invoices)** → OrderKey (PK), InvoiceNo, InvoiceDate, CustomerKey (FK)
- **OrderDetails** → OrderDetailKey (PK), OrderKey (FK), ProductKey (FK), Quantity, UnitPrice, LineTotal

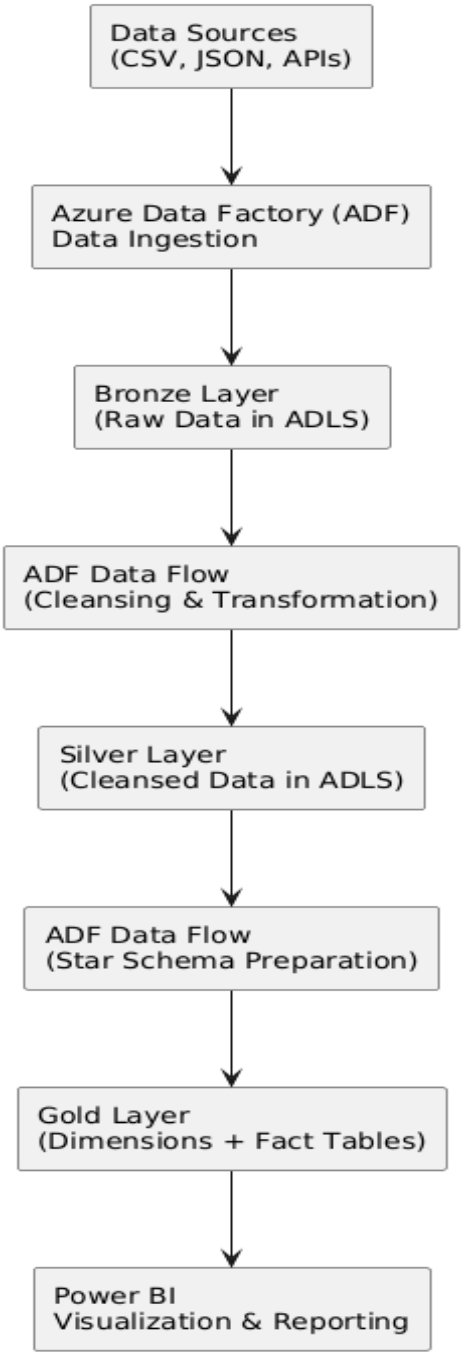
7. HOW IT WORKS

Data Sources (CSV, DB, API)

- Raw e-commerce data is collected from multiple sources such as CSV files (e.g., Kaggle dataset), databases, or APIs.
- This ensures that both structured and semi-structured data are available for downstream analytics.

Azure Data Factory (ADF)

- ADF orchestrates data ingestion pipelines.
- It automates the extraction of raw data from sources and loads it into the storage system.
- Scheduling and monitoring ensure continuous and reliable data movement.



Azure Data Lake Storage (ADLS Gen2)

- The ingested data is stored in the Medallion Architecture format:
    - Bronze Layer → Raw, unprocessed data (kept for traceability).
    - Silver Layer → Cleaned, validated, and standardized data.
    - Gold Layer → Aggregated and business-ready datasets for reporting.
- This layered structure helps maintain data quality and governance.

Azure Data Lake Analytics (ADLA)

- Performs processing and querying on curated datasets.
- Enables advanced analytics and large-scale computations without needing a dedicated cluster.
- Helps prepare data models for reporting and insights.

Power BI

- Connects to the Gold Layer datasets.
- Provides interactive dashboards, reports, and KPIs for business decision-making.
- Stakeholders can track sales trends, customer behavior, and other insights in real time.

8. CONCLUSION

The Data Lake Analytics project successfully implemented a Medallion Architecture (Bronze, Silver, Gold) using Azure Data Factory (ADF), Azure Data Lake Storage (ADLS), and Power BI.

- Raw retail data was ingested into Bronze, cleansed and standardized in Silver, and aggregated into Gold for analytics.
- Automated ADF pipelines ensured reliable ingestion, transformation, and governance of data.
- Power BI dashboards delivered clear business insights, including sales trends, top products, customer analysis, and revenue growth patterns.
- The solution achieved its goals of centralizing data, improving quality, enabling decision-making, and reducing manual effort.

This project demonstrates a scalable, cost-efficient, and business-ready data platform that can be extended in the future with real-time ingestion, advanced analytics, and AI/ML use cases.