PROJECT OVERVIEW DOCUMENT

DATA WAREHOUSING INTEGRATION

1. PROJECT OVERVIEW

The Data Warehousing Integration project establishes a robust, scalable, and automated data pipeline on Microsoft Azure. It integrates Azure Data Factory (ADF) for orchestration, Azure Databricks for distributed transformations, and Azure Synapse Analytics for storage and querying.

The project is designed to enable organizations to convert raw, fragmented datasets into a unified, analytics-ready warehouse, empowering analysts and decision-makers with timely insights. By leveraging cloud-native services, the solution ensures performance, scalability, and cost optimization, aligning with enterprise data strategy goals.

2. OBJECTIVES

The specific objectives of this project include:

- **Ingestion & Orchestration:** Automate ingestion of raw datasets (customers, products, sales transactions, dates) into Azure Data Lake Storage via ADF pipelines.
- **Transformation at Scale:** Cleanse, validate, and enrich datasets using Spark-powered Databricks notebooks to ensure high data quality.
- **Loading into Warehouse:** Structure curated data into fact and dimension models within Azure Synapse Analytics.
- **Analytics-Readiness:** Deliver a star schema warehouse optimized for reporting and dashboarding.
- **Reliability & Monitoring:** Ensure automated scheduling, error handling, and monitoring of pipelines for production-grade operation.

3. BUSINESS CONTEXT

In the retail and e-commerce domain, organizations generate massive amounts of data daily—covering sales, customers, and product performance. Without a centralized warehouse, data silos lead to inefficiencies and hinder business insights.

This project addresses the following pain points:

- Fragmented data across systems delaying reporting.
- Inconsistent or inaccurate data limiting decision-making.
- Manual ETL processes are prone to error and inefficiency.

By implementing this end-to-end data warehousing pipeline, the organization gains a single source of truth for analytics, enabling answers to key business questions like:

- Which products generate the highest revenue?
- Who are the top customers by sales?
- What are the sales trends over time?

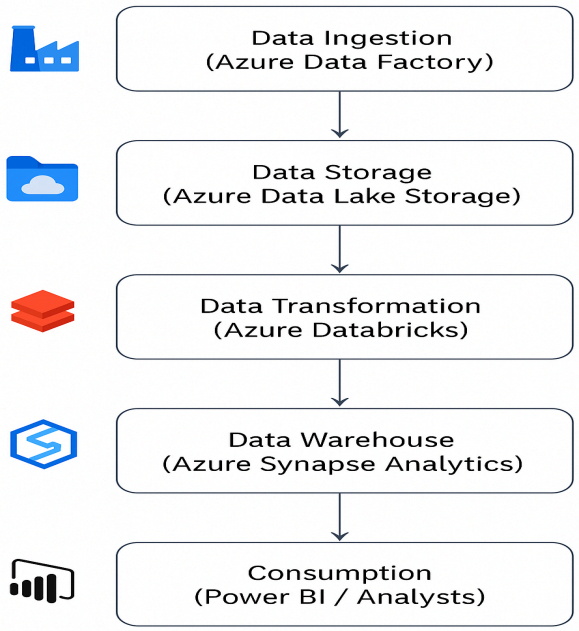
4. KEY BENEFITS

The implemented solution delivers the following benefits:

- **Scalability:** Databricks clusters scale dynamically to meet demand.
- **Performance:** Parallel data processing accelerates ingestion and transformation.
- **Data Quality:** Cleansing and validations ensure accurate, trustworthy warehouse data.
- **Cost Efficiency:** Autoscaling and auto-termination prevent unnecessary resource consumption.
- **Actionable Insights:** Synapse tables support BI dashboards that surface trends, customer behavior, and product performance.
- **Future-Readiness:** Modular design allows expansion into advanced analytics and ML integration.

5. HIGH-LEVEL ARCHITECTURE

The solution architecture is structured into clearly defined layers:



5.1. DATA INGESTION (AZURE DATA FACTORY):

- Ingests raw datasets into ADLS Raw Zone.
- Automates execution via pipelines and schedules.

5.2. DATA STORAGE (AZURE DATA LAKE STORAGE):

- Organizes data into Raw, Curated, and Analytics zones.
- Ensures separation of raw input and transformed output.

5.3. DATA TRANSFORMATION (AZURE DATABRICKS):

- Cleans, validates, and enriches data using PySpark.
- Derives measures like Total Sales = Quantity × Price.

5.4. DATA WAREHOUSE (AZURE SYNAPSE ANALYTICS):

- Hosts fact and dimension tables in a star schema (FactSales, DimCustomer, DimProduct, DimDate).
- Provides fast querying and analytics capabilities.

5.5. CONSUMPTION (POWER BI / ANALYSTS):

- Analysts and BI users query Synapse directly.
- Dashboards highlight KPIs such as sales by product, top customers, and monthly revenue trends.

6. ENTITY-RELATIONSHIP MODEL (ERD)

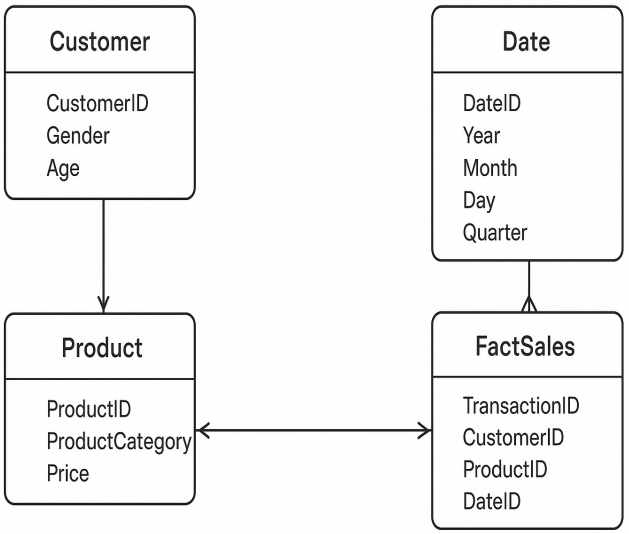
6.1. ENTITIES:

- **Customer:** CustomerID, Gender, Age
- **Product:** ProductID, ProductCategory, Price
- **Date:** DateID, Year, Month, Day, Quarter
- **FactSales:** TransactionID, CustomerID, ProductID, DateID, Quantity, UnitPrice, TotalAmount

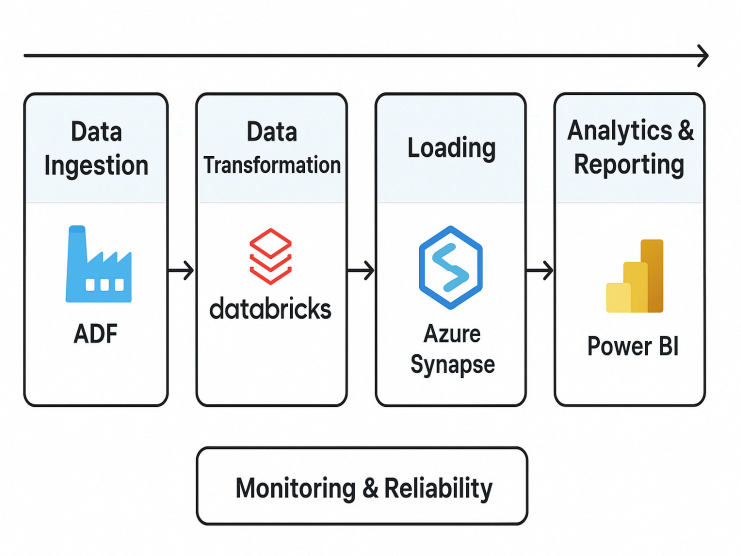
6.2. RELATIONSHIPS:

- One Customer → Many Sales
- One Product → Many Sales
- One Date → Many Sales

The FactSales table acts as the central fact table, linking all dimensions.



7. WORKFLOW (HOW IT WORKS)



7.1 DATA INGESTION (ADF → ADLS)

- ADF pipelines copy raw CSV datasets into ADLS Raw Zone.

- Metadata and monitoring logs are captured for traceability.

7.2 DATA TRANSFORMATION (DATABRICKS)

- ADF triggers Databricks notebooks.
- Data is cleansed, validated, enriched, and aggregated.
- Nulls, duplicates, and formatting errors are removed.
- Outputs are written to ADLS Curated Zone.

7.3 LOADING (ADF → SYNAPSE)

- Curated data is loaded into Synapse tables.
- Fact and dimension models are created and registered.

7.4 ANALYTICS & REPORTING (SYNAPSE → POWER BI)

- Business analysts run SQL queries on Synapse.
- Power BI dashboards visualize sales trends, customer segments, and product performance.

7.5 MONITORING & RELIABILITY

- ADF provides logs and alerts on pipeline execution.
- Retries ensure resiliency against transient failures.

8. EXPECTED OUTCOMES & SUCCESS CRITERIA

8.1. EXPECTED OUTCOMES:

- Fully automated ingestion, transformation, and loading of data.
- Clean, curated fact and dimension tables in Synapse.
- End-to-end monitoring and error handling in ADF.
- Power BI dashboards for interactive business insights.

8.2. SUCCESS CRITERIA:

- ≥ 99% pipeline success rate.
- Transformation and load of 1M+ records within 15 minutes.

- Accurate joins across FactSales and all dimensions.
- Verified reporting in Power BI connected to Synapse.
- Cost efficiency demonstrated via autoscaling clusters and optimized ADF runs.
