# **Project Title:** Scalable Data Lakehouse for Pharmaceutical Sales

**Domain:** Healthcare / Pharma  
**Client:** **GlobalPharma Inc.**  
**Tech Stack:** Azure Data Factory (ADF), Azure Databricks, Delta Lake, ADLS Gen2, Python, PySpark, SQL, Pandas, ETL

### ****🔍 Problem Statement****

GlobalPharma’s pharmaceutical sales data arrived from multiple external sources (APIs, CSV, XML) with no centralized, automated pipeline. Manual processing caused delays, inconsistent data quality, and limited analytics capability. The system lacked proper validation, transformation, and orchestration for scalable delivery to business tools and CRM platforms.

### ****🎯 Goal****

To build a fully **automated, scalable, and fault-tolerant data lakehouse** on Azure using **Medallion Architecture (Bronze → Silver → Gold)** to deliver clean, enriched, and analytics-ready datasets for pharmaceutical sales insights.

### ****🏗️ Solution & Architecture****

* **Orchestration:** Azure Data Factory pipelines manage ingestion, validation, transformation, and delivery workflows.
* **Processing:** Azure Databricks runs PySpark ETL jobs to cleanse, enrich, and aggregate data.
* **Storage:** ADLS Gen2 used for staging raw, validated, and curated datasets.
* **Delivery:** Gold-layer datasets delivered to **Azure Synapse Analytics, Power BI dashboards**, and CRM platforms.

### ****🟠 Bronze Layer – Raw Data Ingestion (Azure)****

* External data (API, CSV, XML) ingested into **ADLS Gen2 Raw Zone**.
* **Azure Functions** validate file format, check schema, and handle errors.
* Raw data stored for auditing and reproducibility.

### ****⚪ Silver Layer – Validation & Enrichment****

* **Databricks PySpark jobs** validate data, standardize fields, and handle nulls.
* Joins with reference tables for enrichment (e.g., product, territory, and sales rep information).
* Errors logged and handled automatically using ADF error-handling workflows.

### ****🟡 Gold Layer – Aggregation & Delivery****

* Aggregated, analytics-ready datasets created using **Databricks ETL jobs**.
* Gold-layer datasets **optimized using Delta Lake features** (Z-ordering, partitioning).
* Delivered to **Synapse Analytics** for reporting and **Power BI dashboards**.

### ****👨‍💻 Role & Responsibilities****

* Designed and implemented **end-to-end ETL workflows** in Azure Data Factory.
* Developed **PySpark transformation logic** in Databricks for business rules.
* Applied **Medallion Architecture** to structure data layers (Bronze → Silver → Gold).
* Ensured **data quality, deduplication, validation, and enrichment**.
* Managed **schema evolution, error logging, and workflow automation**.
* Optimized data storage and queries using Delta Lake best practices.

### ****📈 Business Impact****

* Reduced **manual data handling by 80%**.
* Improved **data processing efficiency by 40%**.
* Enabled **near real-time visibility** into sales data by territory and product.
* Scalable solution processed **~1 TB of monthly data**.
* Delivered high-quality datasets to **BI dashboards and CRM systems**.

### ****🧗‍♂️ Challenges Faced****

* Schema drift and inconsistencies across multiple external sources.
* Large file sizes and performance optimization of ETL jobs.
* Integration of legacy systems with modern Azure services.
* Ensuring cost-efficiency during large-scale deployments.

### ****✅ Outcome****

* Built a **fully automated, scalable Azure data pipeline**.
* Ensured **data reliability, accuracy, and governance**.
* Enabled **real-time insights and reporting** for pharmaceutical sales teams.
* Delivered structured datasets suitable for **analytics, dashboards, and CRM integration**.

### ****📄 Description****

This project developed a **cloud-native Azure pipeline** to automate ingestion, validation, transformation, and delivery of sales data. Following **Medallion Architecture**, the pipeline created raw (Bronze), enriched (Silver), and analytics-ready (Gold) datasets stored in ADLS Gen2 and delivered to Synapse Analytics and Power BI.

### ****📄 Summary****

Implemented a **scalable, fault-tolerant Azure data pipeline** to process large pharmaceutical datasets from multiple sources. Leveraged **ADF pipelines, Databricks, and Delta Lake** to improve processing speed, data quality, and automated delivery for BI dashboards and CRM applications.