**MediHealth Hospitals – Healthcare Data Integration with Azure Data Factory**

**Problem Statement**

MediHealth Hospitals operates across multiple branches and stores patient-related information in different systems:

* **Azure SQL Database** → Pharmacy transactions.
* **Azure Blob Storage** → Lab results in CSV and radiology reports in JSON.

This fragmentation creates challenges:

* Doctors cannot access a unified patient record.
* Compliance reporting requires manual effort.
* Researchers struggle to combine clinical and pharmacy datasets.
* The hospital cannot easily implement predictive analytics.

The hospital needs a **centralized, automated integration platform** using Azure Data Factory (ADF) to ingest, transform, and unify medical data into **Azure SQL Database** for analytics and reporting.

**Skill Tower to Develop the Project**

This project helps develop expertise in:

1. **Azure Data Factory (ADF)** – pipelines, data flows, triggers, orchestration.
2. **Integration Runtimes** – Azure IR for cloud and Self-hosted IR for on-premises.
3. **Azure SQL Database & ADLS** – as target and staging areas.
4. **Data Engineering Concepts** – ETL vs ELT, incremental loads, data quality.
5. **DevOps** – GitHub integration for CI/CD.
6. **Healthcare Data Domain** – patient, lab, pharmacy, compliance datasets.

**Use Case / Architecture Diagram**

**Use Case:**  
Unify patient data from multiple sources into a **single Azure SQL Database** for doctors, administrators, and researchers.

**Architecture (Textual Description):**

* **Sources**
  + Azure SQL DB (EHR)
  + Azure SQL Database (Pharmacy)
  + Blob Storage (CSV lab results, JSON radiology)
* **ADF Pipelines**
  + Copy raw data → ADLS staging
  + Transform using Mapping Data Flows & Wrangling Data Flows
  + Load curated data → Azure SQL Database (final schema: Patient, LabResults, Pharmacy, Radiology)
* **Consumers**
  + Doctors → Patient dashboards (via Power BI over Azure SQL DB)
  + Compliance Officers → Automated HIPAA/GDPR reports
  + Researchers → Curated tables for analysis

**User Stories**

1. *As a doctor,* I want to see combined lab, pharmacy, and patient data in one dashboard.
2. *As a compliance officer,* I need automated compliance-ready reports in SQL DB.
3. *As a researcher,* I want curated datasets to analyze treatment outcomes.
4. *As a data engineer,* I need pipelines that handle errors and retries automatically.
5. *As a DevOps engineer,* I want GitHub integration to manage ADF changes.

**Expected Deliverables**

* Linked Services for Azure SQL DB, Blob, ADLS.
* Datasets for structured (SQL) and semi-structured (CSV/JSON) data.
* Copy activities to move raw → staging → curated data.
* Mapping Data Flows for joins, aggregates, conditional splits, derived columns.
* Wrangling Data Flows for analyst-driven data shaping.
* Parameterized pipelines for multi-branch ingestion.
* Triggers for daily, hourly, and event-based loads.
* Incremental load strategy with watermark column (LastModifiedDate).
* Error handling (retry, logging, error zone in ADLS).
* GitHub-integrated ADF project for CI/CD.
* Final **curated Azure SQL Database schema** for reporting.

**Implementation Notes**

* Create **ADLS zones**: staging, curated, error-handling.
* Store **raw copies** before transformation for audit purposes.
* Implement **incremental loading** via LastModifiedDate watermark.
* Use **event triggers** for real-time ingestion of lab results CSV/JSON.
* Optimize **ADF pricing** by limiting debug runs, batching files, and pushing aggregations into SQL DB were cheaper.