Project Report Summary: Azure Healthcare Data Engineering & Analytics

This project is a comprehensive end-to-end data engineering and analytics solution for healthcare data from two major hospitals—Hospital A and Hospital B. The project exhibits a deep understanding of modern cloud data platforms and analytics, demonstrating expertise in orchestrating complex data pipelines, robust transformation processes, and insightful analytics reporting.

Project Workflow Overview

```
Raw Data (Hospital A & B)
   ↓
Azure Data Factory Pipelines
   ↓
SQL Stored Procedures (Staging Cleanup & Validation)
   ↓
Databricks Python Notebook
   ├─ Bronze Layer (Raw to Semi-structured)
   ├─ Silver Layer (Cleansed & Enriched)
   └─ Gold Layer (Unified A + B)
   ↓
Azure Synapse (Openrowset + External Tables)
   ↓
Power BI Dashboard (Final Reporting)
```

📊 Project Summary

1. Data Acquisition & Ingestion

Raw datasets were received separately from Hospital A and Hospital B, each containing critical healthcare information such as patients, providers, departments, encounters, and transactions. Azure Data Factory (ADF) pipelines were built to ingest these CSV files into Azure Data Lake Storage Gen2. The pipelines ensured proper folder structuring, validation, and successful ingestion checks to maintain data integrity.

2. Transformation & Primary Processing

Initial data cleansing and schema validations were performed using SQL-based stored procedures within Azure SQL Database. These transformations standardized the data, corrected structural issues, and prepared it for deeper analytical processing.

3. Advanced Processing with Azure Databricks

Using Apache Spark in Azure Databricks, Python notebooks were developed to implement a medallion architecture:

- **Bronze Layer:** Captured raw ingested data with minimal cleaning.
- **Silver Layer:** Applied business logic and performed in-depth cleaning and enrichment.
- **Gold Layer:** Unified and merged Hospital A and Hospital B datasets into a final analytical model using Delta Lake tables.

4. Data Serving with Azure Synapse

Delta tables from the Gold Layer were queried using Openrowset SQL queries in Azure Synapse Analytics. External tables were then created on top of these queries to expose a clean, structured, and SQL-accessible version of the healthcare data for downstream analysis.

5. Analytics and Reporting in Power BI

The final unified dataset was exported to Power BI Desktop, where a two-page interactive dashboard was created. This dashboard included DAX-based KPI cards, visual analytics across departments, and insights segmented by visit types. A significant focus was placed on identifying the financial burden on patients, revealing how much was paid out-of-pocket even after insurance coverage.

🔹 Conclusion

This solution integrates Azure Data Factory, Azure SQL, Databricks, Synapse, and Power BI into a cohesive healthcare analytics pipeline. It enabled a deep dive into patient costs and departmental trends, empowering stakeholders with actionable insights. This project reflects my passion towards modern cloud data engineering and analytics, from ingestion to visualization.