Healthcare Readmission Data Pipeline

Abstract:

This project is an end-to-end data engineering and analytics solution to analyze hospital readmission patterns using Azure cloud services and visualize insights with Power BI.

Problem Statement:

Hospitals face high costs and inefficiencies due to frequent patient readmissions. This project aims to build a scalable pipeline to:

- Clean and transform raw healthcare data

- Store it in a secure SQL database

- Provide visual insights into readmission trends by age, specialty, and diabetes medication

Solution Architecture:

Azure Data Lake Storage (Raw CSV)
↓
Azure Data Factory (Copy Activity)
↓
Azure Databricks (PySpark Data Cleaning)
↓
Azure SQL Database (Structured Table)
↓
Power BI Dashboard (Insight Visualization)

Tools & Technologies

Layer	Tools
Storage	Azure Data Lake Gen2
Data Processing	Azure Databricks (PySpark)
Orchestration	Azure Data Factory (ADF)
Database	Azure SQL Database

Steps to Reproduce:

1. Ingest Raw Data

   - Upload `hospital_readmissions.csv` to ADLS Gen2

2. ETL with Databricks

   - Clean data using PySpark

   - Output: `hospital_readmissions_cleaned.csv`

3. Load to SQL

   - Use ADF to copy cleaned CSV to Azure SQL table

4. Power BI

- Connect to SQL DB

- Create visuals like:

- Readmission by Specialty

- Age vs Diabetes Medication

- Avg. Time in Hospital

Sample Insights:

- Highest readmissions occur in Circulatory and Internal Medicine
- 60–80 age group is the most vulnerable
- Patients with diabetes medication = yes have higher return rates

Optional: Email Alerts & Automation

- ADF pipeline triggers on schedule (daily)
- Future enhancements:
  - Add email alerts via Azure Monitor
  - Log events into Log Analytics

Future Improvements:

- Build ML model to predict readmission risk
- Use Azure Synapse or DataBricks Delta Lake
- Add role-based dashboards for doctors/admin

Credits:

This project is for learning Azure Data Engineering and visualization in the healthcare domain.