# Project Title: Healthcare Claims Data Processing and Analytics using PySpark on Databricks

**Domain: Healthcare Analytics**

**Difficulty Level: Intermediate to Advanced**

---

**Scenario:**

A hospital system wants to analyze patient data to reduce readmission rates, optimize treatment outcomes, and detect patterns in hospital operations. The existing data is large and semi-structured, and the hospital wants to build a scalable data engineering pipeline using **PySpark on Databricks Community Edition**.

As a Data Engineer trainee, your task is to ingest, clean, transform, and analyze this dataset and deliver meaningful insights to the analytics team.

---

**Dataset:**

**Dataset Name**: diabetic_data.csv
**Source**: Kaggle - Hospital Readmission Dataset

**Size**: ~100,000 patient records

---

**Data Details (Selected Columns):**

- race, gender, age
- admission_type_id, discharge_disposition_id, admission_source_id
- time_in_hospital, num_lab_procedures, num_procedures, num_medications
- number_outpatient, number_emergency, number_inpatient
- diag_1, diag_2, diag_3 (Diagnosis codes)
- readmitted (Yes/<30/>30/No)
- change, diabetesMed, insulin, A1Cresult

---

**Problem Statement:**

**"How can we build a scalable data pipeline that processes hospital records and generates insights to reduce readmission rates and improve patient care?"**

---

**Objectives & Tasks:**

**1. Data Ingestion:**

- Load the CSV into Databricks using PySpark.

**2. Data Cleansing:**

- Handle missing values like "?" in race, gender, diag_1, etc.
- Remove or analyze invalid entries (e.g., gender = "Unknown/Invalid").

**3. Data Transformation:**

- Convert categorical columns into meaningful values (e.g., map admission types).
- Derive new columns such as:
    - total_visits = number_outpatient + number_emergency + number_inpatient
    - readmission_flag = 1 if readmitted in ("<30", ">30") else 0

**4. Exploratory Analysis:**

- Average time in hospital per age group or diagnosis.
- Identify which diagnoses have high readmission rates.
- Relationship between insulin/A1C levels and readmission.

**5. Performance Optimization:**

- Implement data partitioning by age or readmission_flag.
- Cache intermediate results for repeated analysis.

**6. Data Export:**

- Save cleaned dataset as Parquet/Delta.
- Export analytical summaries as CSV for BI tools.

---

**Problems You Are Solving:**

- Automating hospital data processing pipelines.
- Improving data quality and preparing data for ML analysis.
- Identifying patterns in patient treatments and readmissions.
- Supporting hospital teams in reducing unnecessary readmissions.

---

**Expected Outcomes:**

- A structured, cleaned dataset in Delta format.
- Analytical reports (ppt) on:
    - Top causes of readmission
    - Effect of medication on readmission
    - Time in hospital by age group

- Hands-on understanding of PySpark-based ETL pipelines on Databricks.

---

**Skills & Tools Covered:**

- PySpark (DataFrame API, UDFs, Joins, Aggregations)
- Databricks Community Edition
- Delta Lake
- Data partitioning, caching, file formats
- Data analysis for healthcare domain