Healthcare Data Lakehouse Project

Project Type: End-to-End Data Engineering & Analytics (AWS + Databricks)

1. Executive Summary

This project builds a scalable Healthcare Data Lakehouse using AWS S3 and Databricks to transform raw patient CSV data into business-ready insights. The solution follows Medallion Architecture (Bronze, Silver, Gold) and delivers a single-page executive dashboard for decision-making.

2. Business Problem Statement

Hospitals collect patient-level data including Age, Gender, Medical Condition, Treatment, Insurance, Income, Smoking Status, Admission Type, and Length of Stay. Raw CSV data lacks validation, structure, and analytical capability. Management requires standardized reporting for compliance and strategic decisions.

3. Business Objectives

- Build cloud-based data pipeline using AWS and Databricks.
- Ensure data quality and compliance through validation rules.
- Store data externally in AWS S3 (Data Lake).
- Provide executive-level single-page dashboard.
- Enable data-driven healthcare decisions.

4. Medallion Architecture Design

Bronze Layer (Raw Data):

- Ingest CSV file into Databricks.
- Store unmodified data in S3 (bronze folder) as Delta format.
- Create external table referencing S3 location.

Silver Layer (Cleaned Data):

- Remove null values in critical columns.
- Validate Age > 0 and Length_of_Stay >= 0.
- Standardize categorical fields (Gender, Smoking Status, Admission Type).
- Cast numeric columns to correct data types.
- Store cleaned data in S3 silver folder

Gold Layer (Business Aggregations):

- Average Length of Stay by Medical Condition.
- Admission Type distribution percentage.
- Smoking Status impact on Length of Stay.
- Insurance Type vs Patient Outcome analysis.
- Region-wise patient distribution.
- Store aggregated tables in S3 gold folder.
- Register external Delta tables in Databricks metastore.

5. AWS S3 Storage Strategy

S3 Bucket Structure:

s3://healthcare-lakehouse-sarosh/

├── bronze/

├── silver/

└── gold/

All tables stored as Delta format and created as external tables in Databricks.

6. Data Compliance & Governance

- Enforced schema validation.
- Removed invalid and negative records.
- Standardized categorical values to avoid duplication.
- Delta Lake ensures ACID transactions.
- Audit-ready structured storage for healthcare analytics.

7. Executive Dashboard (Single Page Design)

KPIs Section (Top Row):

- Total Patients
- Average Length of Stay
- Emergency Admission %
- Most Common Disease

Visualizations Section:

- Bar Chart: Medical Condition vs Avg Length of Stay
- Pie Chart: Admission Type Distribution

- Column Chart: Insurance Type vs Outcome
- Region-wise Patient Distribution
- moking Status vs Avg Length of Stay

9. Technical Deliverables

- Databricks Notebooks (Bronze, Silver, Gold).
- AWS S3 Data Lake (External Delta Tables).
- Executive Dashboard (Single Page).
- Architecture Diagram.