
Capstone Project: Healthcare Data Engineering Platform on Azure Databricks

---

Objective

Build a complete data engineering solution for a fictional healthcare analytics company — **MediPulse Analytics** — using **Azure Databricks**. You will ingest, clean, transform, and analyze healthcare data from multiple sources, store it using Delta Lake, implement incremental loads, and prepare analytical views for downstream machine learning and dashboards.

Project Scenario

Company: MediPulse Analytics **Goal:** To process and analyze healthcare patient data, hospital data, and appointment records for building KPIs and dashboards.

Data Sources

You will simulate three raw datasets:

---

1 `patients.csv`

```
patient_id,name,age,gender,region
P001,Arjun Mehta,34,M,North
P002,Neha Sharma,29,F,South
P003,Rahul Gupta,40,M,East
P004,Sneha Nair,25,F,West
```

---

2 `hospitals.json`

```
[
  {"hospital_id": "H001", "hospital_name": "City Care", "region": "North"},
  {"hospital_id": "H002", "hospital_name": "LifePlus", "region": "South"},
  {"hospital_id": "H003", "hospital_name": "MediHope", "region": "East"},
  {"hospital_id": "H004", "hospital_name": "CureWell", "region": "West"}
]
```

---

3 `appointments_day1.csv`

```
appointment_id,patient_id,hospital_id,appointment_date,diagnosis,cost,status
A1001,P001,H001,2024-01-10,Diabetes,400,Completed
A1002,P002,H002,2024-01-11,Flu,250,Completed
A1003,P003,H003,2024-01-11,Heart Disease,1000,Pending
A1004,P004,H004,2024-01-12,Allergy,300,Completed
```

Step 1 — Bronze Layer: Raw Ingestion

Tasks:

- Read CSV and JSON data into DataFrames.
- Write them as Delta tables (`bronze_patients` , `bronze_hospitals` , `bronze_appointments`).

Step 2 — Silver Layer: Data Cleansing & Transformation

Tasks:

- Filter out `Pending` appointments.
- Join patients and hospitals to enrich appointment data.
- Add new calculated column: `year = year(appointment_date)` and `month` .
- Store output as `silver_appointments` .

Step 3 — Gold Layer: Analytical Aggregations

Tasks:

- Total revenue per hospital.
- Total patients per region.
- Top 3 most expensive diagnosis categories.
- Store as `gold_healthcare_summary` .

Step 4 — Incremental Load Simulation

Tasks:

- Create `appointments_day2.csv` with new data.
- Use `MERGE` or `Upsert` to update the silver table.
- Show how incremental data changes the gold table.

Step 5 — Delta Lake Features

Tasks:

- Use **Time Travel** to view the gold table before incremental load.
- Use **Vacuum** to clean up historical versions.
- Use **Optimize + Z-Ordering** on `hospital_id` .

Analytical Questions to Solve

1. Total revenue generated by each hospital.
2. Average cost per diagnosis category.
3. Number of patients served per region.
4. Trend of appointments month-over-month.
5. Top 5 most expensive treatments in the last 6 months.
