Hospital Patient Visit ETL Project - Problem Statement

Objective:

Design and implement a data engineering pipeline that extracts, transforms, and loads (ETL) hospital patient visit data from multiple source files into a relational database. The goal is to create a clean, queryable data warehouse that supports SQL-based analytics for healthcare operations.

Data Sources:

You are provided with three datasets in CSV and JSON format:

- patients.csv: Contains basic patient demographics.
- doctors.csv: Contains doctor profiles and specialties.
- visits.json: Contains visit records including timestamps, diagnosis codes, and billing amounts.

Tasks:

1. Extract:
   - Load raw data from patients.csv, doctors.csv, and visits.json.
   - Validate input formats (dates, IDs, cost).

2. Transform:
   - Clean and normalize data:
     - Standardize gender and visit types.
     - Handle missing or invalid entries.
     - Convert date strings to proper date formats.
   - Join and validate foreign key relationships:
     - Each visit must have a valid patient_id and doctor_id.

3. Load:

- Store the cleaned data into three PostgreSQL tables:

  - patients(patient_id, name, gender, birth_date)

  - doctors(doctor_id, name, specialty)

  - visits(visit_id, patient_id, doctor_id, visit_date, diagnosis_code, duration_min, visit_type, cost_usd)

4. Analysis:

- Write 15 SQL queries to answer key operational questions (e.g., patient counts, revenue trends, doctor performance).

Success Criteria:

- All tables are normalized and enforce referential integrity.
- ETL process runs without error.
- Queries produce accurate and timely insights.
- Documentation includes data dictionary, sample queries, and assumptions.