1. Giới thiệu

Trong lĩnh vực y tế, việc quản lý và phân tích dữ liệu bệnh nhân là yếu tố quan trọng để hỗ trợ nghiên cứu, dự báo và cải thiện chất lượng điều trị. Dữ liệu thô thường được lưu trữ phân tán và khó phân tích trực tiếp. Do đó, việc thiết kế kho dữ liệu (Data Warehouse) và xây dựng quy trình ETL (Extract – Transform – Load) bằng SQL Server Integration Services (SSIS) là cần thiết.

Trong báo cáo này, nhóm tiến hành thiết kế và triển khai một Data Warehouse cho bệnh nhân ung thư phổi, với mô hình dữ liệu 6 Dimension – 2 Fact, đồng thời trình bày chi tiết quy trình SSIS từ khâu trích xuất dữ liệu đến nạp vào hệ thống.

2. Dataset và ý nghĩa các cột

Dataset bao gồm các thuộc tính sau:

STT	Cột	Ý nghĩa
1	id	Mã định danh duy nhất cho mỗi bệnh nhân.
2	age	Tuổi bệnh nhân tại thời điểm chẩn đoán.
3	gender	Giới tính (male, female).
4	country	Quốc gia/khu vực cư trú.
5	diagnosis_date	Ngày bệnh nhân được chẩn đoán ung thư phổi.
6	cancer_stage	Giai đoạn ung thư (Stage I – IV).
7	family_history	Có tiền sử gia đình mắc ung thư (yes/no).
8	smoking_status	Tình trạng hút thuốc (current, former, never, passive smoker).
9	bmi	Chỉ số khối cơ thể (Body Mass Index).

10	cholesterol_level	Mức cholesterol của bệnh nhân (giá trị số).
11	hypertension	Có bị cao huyết áp (yes/no).
12	asthma	Có bị hen suyễn (yes/no).
13	cirrhosis	Có bị xơ gan (yes/no).
14	other_cancer	Có từng mắc bệnh ung thư khác (yes/no).
15	treatment_type	Loại điều trị (surgery, chemotherapy, radiation, combined).
16	end_treatment_date	Ngày kết thúc điều trị hoặc tử vong.
17	survived	Tình trạng sống sót sau điều trị (yes/no).

3. Thiết kế Data Warehouse

Để phục vụ phân tích đa chiều, dataset được chuẩn hóa thành mô hình Star Schema gồm 6 Dimension và 2 Fact.

3.1. Dimension Tables

1. **DimPatient**
    - patient_id: Mã bệnh nhân.
    - age: Tuổi bệnh nhân.
    - gender: Giới tính.
    - country: Quốc gia cư trú.

→ Giúp phân tích dữ liệu theo nhân khẩu học.

2. **DimLifestyle**
    - lifestyle_id: Khóa chính.
    - smoking_status: Tình trạng hút thuốc.
    - bmi: Chỉ số BMI.
    - cholesterol_level: Mức cholesterol.

- family_history: Tiền sử gia đình mắc ung thư.

→ Cho phép phân tích tác động của lối sống và yếu tố di truyền.

3. **DimMedicalHistory**
    - med_history_id: Khóa chính.
    - hypertension: Có/không cao huyết áp.
    - asthma: Có/không hen suyễn.
    - cirrhosis: Có/không xơ gan.
    - other_cancer: Có/không ung thư khác.

→ Giúp phân tích tác động của bệnh nền.

4. **DimCancerStage**
    - stage_id: Khóa chính.
    - stage_name: Giai đoạn ung thư (Stage I – IV).

→ Hỗ trợ so sánh kết quả điều trị theo từng giai đoạn bệnh.

5. **DimTreatment**
    - treatment_id: Khóa chính.
    - treatment_type: Loại điều trị (surgery, chemotherapy, radiation, combined).

→ Giúp phân tích hiệu quả của từng phương pháp điều trị.

6. **DimDate**
    - date_id: Khóa chính.
    - full_date: Ngày (YYYY-MM-DD).
    - day, month, quarter, year: Các cấp độ thời gian.

→ Cho phép phân tích dữ liệu theo chiều thời gian.

3.2. Fact Tables

1. **FactDiagnosis**
    - fact_diagnosis_id: Khóa chính.
    - patient_id (FK → DimPatient).
    - diagnosis_date_id (FK → DimDate).
    - stage_id (FK → DimCancerStage).
    - lifestyle_id (FK → DimLifestyle).

- med_history_id (FK → DimMedicalHistory).

→ Lưu trữ thông tin chẩn đoán ban đầu, hỗ trợ phân tích tỷ lệ mắc bệnh theo độ tuổi, giới tính, giai đoạn, lối sống.

2. **FactTreatmentOutcome**
    - fact_treatment_fact_id: Khóa chính.
    - patient_id (FK → DimPatient).
    - treatment_id (FK → DimTreatment).
    - start_date_id (FK → DimDate).
    - end_date_id (FK → DimDate).
    - survived: Kết quả điều trị (yes/no).
    - survival_days: Số ngày sống sót (end_date – diagnosis_date).

→ Hỗ trợ phân tích hiệu quả điều trị, thời gian sống sót theo stage, lifestyle, medical history.