Bộ thông tin và truyền thông

Học viện Công Nghệ Bưu Chính Viễn Thông
Cơ sở tại thành phố Hồ Chí Minh

Khoa Công nghệ thông tin 2

KHO DỮ LIỆU VÀ KHAI PHÁ DỮ LIỆU

XÂY DỰNG KHO DỮ LIỆU VỀ TRIỆU CHỨNG CỦA MỘT SỐ BỆNH THÔNG THƯỜNG. KHAI PHÁ KHO DỮ LIỆU NÀY CHO MỤC ĐÍCH CHẨN ĐOÁN BAN ĐẦU CHO NGƯỜI BỆNH

Nhóm 07 - D21CQCNHT01-N - PTITHCM

N21DCCN038	:	Hà Gia Huy
N21DCCN057	:	Lê Trung Nguyên
N21DCCN059	:	Trần Bình Phương Nhã

NHẬN XÉT CỦA GIẢNG VIÊN

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

LỜI MỞ ĐẦU

- Trong những năm gần đây, việc thu thập, phân tích và khai thác thông tin đã trở thành một yếu tố quan trọng trong nhiều lĩnh vực, đặc biệt là trong y tế. Khả năng sử dụng dữ liệu để hỗ trợ chẩn đoán, đưa ra quyết định chính xác và nhanh chóng có thể góp phần nâng cao chất lượng chăm sóc sức khỏe và giảm thiểu rủi ro cho người bệnh.

- Cùng với sự phát triển của công nghệ thông tin, kho dữ liệu y tế ngày càng được xây dựng và mở rộng, chứa đựng một lượng lớn thông tin về triệu chứng, bệnh lý và các phương pháp điều trị. Tuy nhiên, việc khai thác tri thức từ các kho dữ liệu này vẫn chưa được tận dụng đầy đủ. Đa phần dữ liệu y tế chỉ được sử dụng để lưu trữ hoặc báo cáo thống kê mà chưa thực sự được phân tích để hỗ trợ chẩn đoán, đặc biệt là trong giai đoạn ban đầu của quá trình khám bệnh.

- Khai phá dữ liệu (Data Mining) và phát hiện tri thức từ dữ liệu (Knowledge Discovery in Database - KDD) đã trở thành những hướng nghiên cứu quan trọng, giúp trích xuất các mẫu thông tin hữu ích từ kho dữ liệu, từ đó hỗ trợ các chuyên gia y tế và bệnh nhân trong việc nhận diện triệu chứng, đánh giá nguy cơ mắc bệnh và đưa ra các khuyến nghị phù hợp.

- Xuất phát từ thực tế này, chúng em đã chọn đề tài: "Xây dựng kho dữ liệu về triệu chứng của một số bệnh thông thường. Khai phá kho dữ liệu này cho mục đích chẩn đoán ban đầu cho người bệnh". Nghiên cứu này nhằm tạo ra một hệ thống dữ liệu có tổ chức, hỗ trợ phân tích, trích xuất thông tin và ứng dụng vào việc chẩn đoán sơ bộ, góp phần nâng cao hiệu quả chăm sóc sức khỏe cộng đồng.

- Nội dung nghiên cứu bao gồm các chương:

- **CHƯƠNG 1:** Tổng quan về khai phá dữ liệu.
- **CHƯƠNG 2:** Quy trình ETL
- **CHƯƠNG 3:** Xây dựng kho dữ liệu và khai phá dữ liệu

- Trong quá trình thực hiện đề tài, khối lượng kiến thức trong lĩnh vực này rất rộng và không ngừng được cập nhật. Do đó, chắc chắn không thể tránh khỏi những thiếu sót. Chúng em kính mong nhận được ý kiến đóng góp của thầy để có thể hoàn thiện đề tài một cách tốt nhất.

CHƯƠNG I
TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

I. Khai phá dữ liệu

- Khai phá dữ liệu (Data Mining) là một lĩnh vực nghiên cứu xuất hiện từ cuối thập niên 1980, bao gồm một tập hợp các kỹ thuật giúp phát hiện thông tin có giá trị tiềm ẩn trong các tập dữ liệu lớn. Quá trình này nhằm tìm ra các quy luật (patterns), xu hướng hoặc mối quan hệ giữa các dữ liệu, giúp người dùng có thể khai thác tri thức một cách hiệu quả.

- Năm 1989, Fayyad, Piatetsky-Shapiro và Smyth đã đưa ra khái niệm Phát hiện tri thức trong cơ sở dữ liệu (Knowledge Discovery in Databases - KDD), chỉ toàn bộ quá trình từ thu thập, xử lý đến khai thác dữ liệu để trích xuất thông tin có giá trị. Trong đó, khai phá dữ liệu là một bước quan trọng, sử dụng các thuật toán để xác định các mẫu dữ liệu có ý nghĩa.

- Khai phá dữ liệu tập trung vào hai khía cạnh chính:

- Phát hiện thông tin tự động hoặc bán tự động (Automated & Semi-Automated): Giúp trích xuất các quy luật và tri thức tiềm ẩn mà không cần kiểm tra thủ công từng dữ liệu.
- Hỗ trợ ra quyết định và dự đoán xu hướng (Decision Support & Trend Prediction): Cung cấp thông tin hữu ích cho quá trình phân tích và lập kế hoạch.

- Là một lĩnh vực liên quan chặt chẽ đến việc xây dựng và khai thác kho dữ liệu, khai phá dữ liệu có sự kết nối với các công nghệ sau:

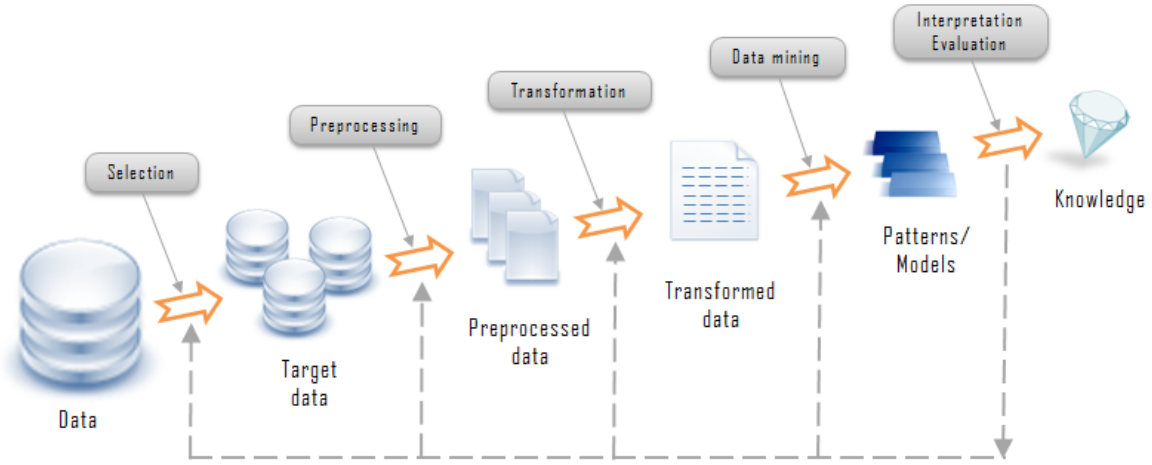
- Công nghệ cơ sở dữ liệu (Database Technology): Là nền tảng giúp lưu trữ và tổ chức dữ liệu một cách khoa học, tạo điều kiện thuận lợi cho quá trình truy vấn và khai thác thông tin.
- Kho dữ liệu (Data Warehousing): Cung cấp một môi trường tập trung, nơi dữ liệu được thu thập, xử lý và tối ưu hóa để phục vụ cho việc phân tích.
- Trực quan hóa dữ liệu (Data Visualization): Hỗ trợ hiển thị dữ liệu dưới dạng bảng, biểu đồ hoặc mô hình giúp người dùng dễ dàng nhận diện các mẫu và xu hướng quan trọng.
- Các phương pháp khai thác dữ liệu phổ biến: Bao gồm phân cụm (clustering), phân loại (classification), phát hiện quy luật (association rule mining) và phân tích xu hướng (trend analysis).

- Việc áp dụng khai phá dữ liệu trong kho dữ liệu về triệu chứng bệnh có thể giúp xác định mối liên hệ giữa các triệu chứng, hỗ trợ xây dựng cơ sở dữ liệu có cấu trúc, từ đó phục vụ tốt hơn cho quá trình phân tích và tra cứu thông tin y tế.

II. Quy trình khai phá dữ liệu

- Quy trình khai phá dữ liệu là một chuỗi lặp và tương tác gồm nhiều bước liên kết chặt chẽ, bắt đầu từ dữ liệu thô (raw data) và kết thúc với tri thức có giá trị (knowledge of interest) nhằm hỗ trợ người dùng đưa ra quyết định.

- Trong bối cảnh đề tài này, quá trình khai phá sẽ giúp xác định mối quan hệ giữa các triệu chứng của bệnh, hỗ trợ chẩn đoán ban đầu cho người bệnh dựa trên dữ liệu được thu thập và tổ chức trong kho dữ liệu.



Quy trình khai phá dữ liệu

- Quy trình khai phá dữ liệu bao gồm các bước sau:

+ Làm sạch dữ liệu (Data Cleaning)

- Loại bỏ dữ liệu nhiễu, không đầy đủ hoặc không nhất quán.
- Xử lý các giá trị bị thiếu trong tập dữ liệu về triệu chứng bệnh.

+ Tích hợp dữ liệu (Data Integration)

- Kết hợp dữ liệu từ nhiều nguồn khác nhau như hồ sơ y tế, tài liệu y khoa, cơ sở dữ liệu bệnh viện để tạo ra một kho dữ liệu thống nhất về triệu chứng bệnh.

+ Lựa chọn dữ liệu (Data Selection)

- Lọc và trích xuất các thông tin quan trọng phục vụ mục tiêu phân tích, chẳng hạn như triệu chứng phổ biến, triệu chứng đặc trưng theo nhóm bệnh, hoặc các yếu tố ảnh hưởng đến triệu chứng.

+ Biến đổi dữ liệu (Data Transformation)

- Chuẩn hóa dữ liệu về triệu chứng bệnh để đồng nhất định dạng, giúp quá trình khai phá hiệu quả hơn.
- Ánh xạ dữ liệu về một dạng thích hợp, chẳng hạn như mã hóa triệu chứng theo danh mục (categorical encoding).

+ Khai phá dữ liệu (Data Mining)

- Áp dụng các thuật toán khai phá dữ liệu để phát hiện mối quan hệ giữa các triệu chứng.
- Xác định các mẫu (patterns), chẳng hạn như triệu chứng nào thường xuất hiện cùng nhau, hoặc dấu hiệu nào có thể gợi ý một nhóm bệnh cụ thể.

+ Đánh giá mẫu (Pattern Evaluation)

- Phân tích mức độ quan trọng và ý nghĩa của các mẫu khai phá được.
- Lọc ra các mẫu hữu ích dựa trên các tiêu chí như tính chính xác, tính dễ hiểu và giá trị thực tiễn trong chẩn đoán bệnh.

+ Biểu diễn tri thức (Knowledge Presentation)

- Sử dụng các phương pháp trực quan hóa dữ liệu như bảng thống kê, biểu đồ, cây quyết định để giúp người dùng dễ dàng nhận diện tri thức từ kho dữ liệu.
- Trình bày các kết quả khai phá theo cách có thể hỗ trợ bác sĩ, chuyên gia y tế hoặc các hệ thống hỗ trợ chẩn đoán ban đầu.

III. Quy trình xây dựng mô hình khai phá dữ liệu

- Bước 1: Thu thập và tổ chức dữ liệu (Data Collection & Organization)

- Tổng hợp dữ liệu về triệu chứng của các bệnh thông thường từ các nguồn khác nhau (sách y khoa, tài liệu nghiên cứu, website y tế đáng tin cậy).
- Chuẩn hóa dữ liệu và lưu trữ vào kho dữ liệu (Data Warehouse) để dễ dàng khai thác sau này.

- Bước 2: Tiền xử lý và làm sạch dữ liệu (Data Cleaning & Preprocessing)

- Loại bỏ dữ liệu không hợp lệ, lỗi nhập liệu.
- Chuẩn hóa định dạng dữ liệu (ví dụ: các triệu chứng giống nhau nhưng cách viết khác nhau cần được đồng nhất).

- Xử lý dữ liệu bị thiếu (Missing Values), ví dụ: điền giá trị trung bình, loại bỏ dòng dữ liệu thiếu quá nhiều thông tin.

- Bước 3: Tích hợp và biến đổi dữ liệu (Data Integration & Transformation)

- Kết hợp dữ liệu từ nhiều nguồn khác nhau (nếu có).
- Chuyển đổi dữ liệu về định dạng phù hợp để dễ khai thác. Ví dụ:
    - Mã hóa triệu chứng dưới dạng số hoặc danh mục.
    - Tạo bảng dữ liệu phù hợp để phân tích (bảng triệu chứng - bệnh, bảng mức độ triệu chứng...).

- Bước 4: Khai phá dữ liệu (Data Mining)

- Sử dụng các kỹ thuật khai phá dữ liệu để tìm ra quy luật, xu hướng trong dữ liệu, bao gồm:
    - Phân cụm (Clustering): Nhóm các triệu chứng có xu hướng xuất hiện cùng nhau để hỗ trợ nhận diện bệnh.
    - Luật kết hợp (Association Rule Mining - Apriori, FP-Growth): Xác định mối liên hệ giữa các triệu chứng và các bệnh thường gặp (VD: "Nếu bệnh nhân có triệu chứng sốt và ho, thì có khả năng bị cảm cúm với độ hỗ trợ 80%").
    - Thống kê và trực quan hóa dữ liệu: Biểu đồ tần suất triệu chứng theo bệnh, mức độ phổ biến của các triệu chứng, v.v.

- Bước 5: Biểu diễn tri thức và báo cáo kết quả (Knowledge Presentation & Visualization)

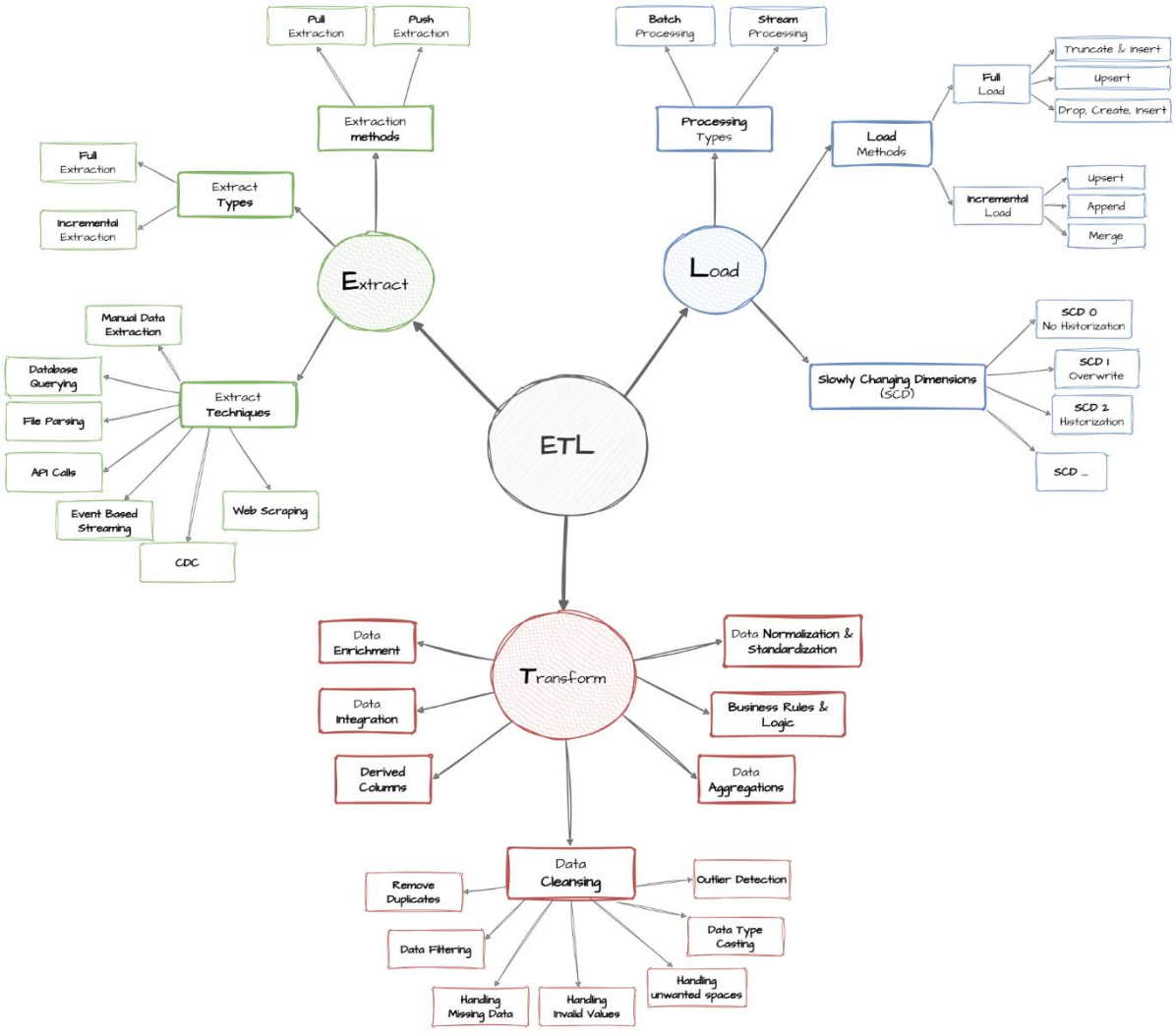
- Trình bày kết quả khai phá dữ liệu bằng:
    - Báo cáo thống kê (ví dụ: bệnh nào phổ biến nhất, triệu chứng nào thường gặp nhất...).
    - Biểu đồ trực quan (dùng biểu đồ cột, biểu đồ tròn, heatmap để hiển thị mối quan hệ giữa các triệu chứng).
    - Bảng quy tắc khai phá được, giúp dễ dàng tra cứu mối liên hệ giữa triệu chứng và bệnh.

CHƯƠNG II

QUY TRÌNH ETL

I. Khái niệm ETL

- ETL (Extract, Transform, Load) là một quá trình thu thập, xử lý và lưu trữ dữ liệu phổ biến trong hệ thống kho dữ liệu (Data Warehouse). Mô hình này được sử dụng rộng rãi trong các hệ thống phân tích dữ liệu, giúp doanh nghiệp có cái nhìn tổng quan và chính xác về hoạt động của mình.



Sơ đồ giới thiệu về quá trình ETL

II. Các giai đoạn trong quá trình ETL

- Extract (Trích xuất dữ liệu):

- Mục tiêu của bước này là thu thập dữ liệu từ nhiều nguồn khác nhau, có thể là:
    - Cơ sở dữ liệu quan hệ (MySQL, PostgreSQL, SQL Server, Oracle,...)
    - API từ các hệ thống khác
    - Dữ liệu từ các file CSV, XML, JSON, TXT,…
    - Hệ thống logs (Apache Logs, Server Logs,...)
- Thách thức:
    - Dữ liệu có thể nằm rải rác ở nhiều hệ thống khác nhau.
    - Dữ liệu có thể có nhiều định dạng khác nhau, đòi hỏi khả năng chuyển đổi.

- Transform (Chuyển đổi dữ liệu):

- Sau khi trích xuất, dữ liệu thô sẽ được làm sạch và chuyển đổi để phù hợp với mục đích phân tích. Các bước phổ biến trong giai đoạn này gồm:
    - Làm sạch dữ liệu (Data Cleaning): Loại bỏ dữ liệu trùng lặp, xử lý giá trị null, lỗi định dạng.
    - Chuyển đổi dữ liệu (Data Transformation): Đổi kiểu dữ liệu, chuẩn hóa dữ liệu theo một quy chuẩn chung, hợp nhất dữ liệu từ nhiều nguồn khác nhau,…
    - Tính toán dữ liệu mới: Tạo các cột mới dựa trên các quy tắc kinh doanh (Business Rules).
    - Tổng hợp dữ liệu: Gộp dữ liệu thành các mức phân tích cao hơn (ví dụ: tổng doanh thu theo tháng).
- Thách thức:
    - Xử lý dữ liệu lớn có thể tốn tài nguyên.
    - Yêu cầu các thuật toán chuyển đổi tối ưu để không làm chậm hệ thống.

- Load (Tải dữ liệu vào Data Warehouse):

- Sau khi dữ liệu đã được làm sạch và chuyển đổi, nó sẽ được tải lên kho dữ liệu (Data Warehouse) để phục vụ báo cáo và phân tích. Có hai phương pháp phổ biến:
    - Full Load: Tải toàn bộ dữ liệu vào kho mỗi lần chạy.
    - Incremental Load: Chỉ tải dữ liệu mới hoặc dữ liệu thay đổi để tối ưu hiệu suất.
- Thách thức:
    - Phải đảm bảo dữ liệu không bị trùng lặp hoặc mất mát.
    - Đảm bảo hiệu suất khi tải dữ liệu lớn.

III. So sánh với mô hình ELT

- ELT (Extract, Load, Transform) là một mô hình xử lý dữ liệu mới, đặc biệt phù hợp với các nền tảng đám mây. Khác với ETL, ELT có trình tự: Extract → Load → Transform.

- Do hạn chế về chi phí và công nghệ trước đây, xu hướng ETL được sử dụng rộng rãi hơn, tuy nhiên với các tiến bộ hiện tại, ngày càng nhiều doanh nghiệp chuyển sang áp dụng ELT để tận dụng khả năng lưu trữ và xử lý dữ liệu mạnh mẽ của các nền tảng đám mây như BigQuery, Snowflake, hoặc Amazon Redshift.

- So sáng ETL và ELT:

Tiêu chí	**ETL**	**ELT**
Trình tự xử lý	Trích xuất → Chuyển đổi → Tải	Trích xuất → Tải → Chuyển đổi
Cách xử lý dữ liệu	Dữ liệu được xử lý trước khi lưu vào Data Warehouse	Dữ liệu thô được tải lên trước, rồi xử lý sau
Hiệu suất	Có thể chậm khi xử lý dữ liệu lớn	Nhanh hơn nhờ tận dụng sức mạnh Cloud
Môi trường phù hợp	Hệ thống truyền thống, Data Warehouse	Cloud Data Warehouse như BigQuery, Snowflake
Khả năng mở rộng	Hạn chế	Cao, dễ dàng mở rộng

CHƯƠNG III
XÂY DỰNG KHO DỮ LIỆU VÀ KHAI PHÁ DỮ LIỆU

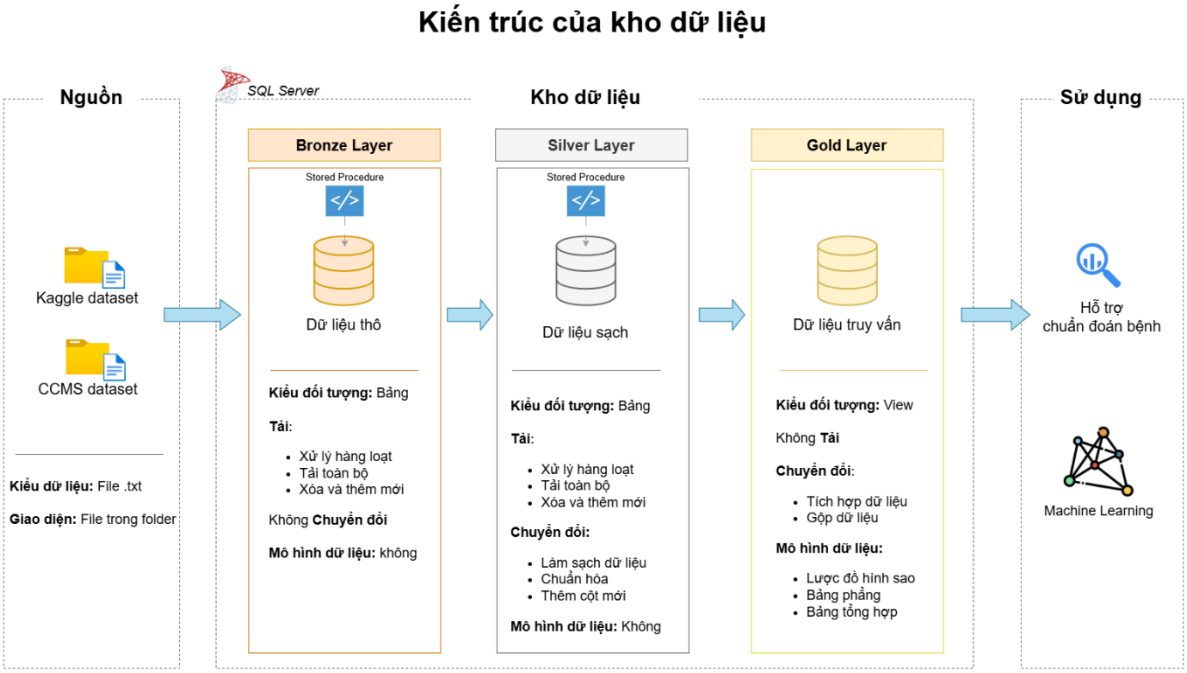
I. Kiến trúc huy chương (Medallion Architecture)

- "Kho dữ liệu về triệu chứng của một số bệnh thông thường" được chúng em thiết kế và xây dựng dựa trên kiến trúc huy chương (Medallion Architecture). Đây là một mô hình tổ chức dữ liệu thường được sử dụng trong Data Lake, Data Warehouse, Data Lakehouse,… kiến trúc này giúp chuẩn hóa và chuyển đổi dữ liệu theo từng giai đoạn để hỗ trợ phân tích dữ liệu hiệu quả.

- Mô hình này phân chia dữ liệu thành ba lớp chính, được đặt tên theo thứ hạng huy chương:

- Bronze layer (lớp Đồng): Chứa dữ liệu thô.
    - Chứa dữ liệu nguyên bản từ các nguồn như API, IoT, logs, database, files...
    - Dữ liệu chưa qua xử lý, có thể chứa lỗi, giá trị trùng lặp.
    - Thường lưu trữ ở dạng Parquet, JSON, Avro, Delta Lake,…
- Silver layer (lớp Bạc): Chứa dữ liệu đã làm sạch và chuẩn hóa.
    - Làm sạch dữ liệu: xử lý trùng lặp, giá trị null, chuẩn hóa schema.
    - Chuẩn bị dữ liệu cho phân tích nhưng vẫn giữ lại các chi tiết quan trọng.
- Gold layer (lớp Vàng): Chứa dữ liệu đã tổng hợp, sẵn sàng cho phân tích.
    - Dữ liệu được tổng hợp theo mô hình Kim tự tháp thông tin (Information Pyramid), Star Schema, Snowflake Schema.
    - Sẵn sàng để sử dụng trong BI (Power BI, Tableau, Looker) hoặc Machine Learning.
    - Tối ưu hóa hiệu suất truy vấn cho các báo cáo và phân tích chuyên sâu.

- Medallion Architecture giúp tổ chức và tối ưu hóa dữ liệu trong Data Lakehouse, mang lại khả năng xử lý dữ liệu mạnh mẽ, linh hoạt và có thể mở rộng. Đây là một phương pháp hiệu quả để xây dựng hệ thống dữ liệu hiện đại phục vụ phân tích, báo cáo, và AI/ Machine Learning.



Áp dụng kiến trúc huy chương vào bài toán xây dựng kho dữ liệu

II. Chuẩn bị dữ liệu và xây dựng lớp Đồng (Bronze layer)

- Dữ liệu thô được chúng em thu thập từ 2 nguồn:

- Dữ liệu về một số bệnh thường gặp, kèm mô tả và một số triệu chứng của bệnh: Disease Symptom dataset từ các tác giả Karthik Udyawar, Pranay Patil và Pratik Rathod tại website Kaggle đã được dịch sang tiếng Việt một phần.
- Dữ liệu về thông tin của bệnh và phân loại bệnh theo chuẩn quốc tế ICD: Hệ Thống Quản Lý Mã Hoá Lâm Sàng Khám Chữa Bệnh (Clinical Coding Management System - CCMS) của Cục quản lý Khám chữa bệnh, thuộc Bộ Y Tế Việt Nam và dữ liệu về ICD-10 2025 của Centers for Medicare & Medicaid Services - CMS thuộc Bộ Y tế và Dịch vụ Nhân sinh Hoa Kỳ.

- Tổng quan về dữ liệu: Các dữ liệu thu thập được bao gồm:

- **disease_description**: chứa thông tin về tên, mã ICD-10 và mô tả của một số bệnh cơ bản. Ví du:

disease	icd_code	description
AIDS	b24	Hội chứng suy giảm miễn dịch mắc phải (AIDS) là một tình tr
Acne	l70_0	Mụn trứng cá thông thường là sự hình thành của mụn đầu đe
Alcoholic hepatitis	k70_1	Viêm gan do rượu là tình trạng viêm gan do bệnh lý gây ra d
Allergy	t78_4	Dị ứng là phản ứng của hệ miễn dịch với một chất lạ thường
Arthritis	m13_9	Viêm khớp là tình trạng sưng và đau ở một hoặc nhiều khớp.

- **symptom_severity**: chứa thông tin về tên và mức độ nghiêm trọng của một số triệu chứng thường gặp. Ví dụ:

english_name	vietnamese_name	symptom_severity
abdominal_pain	Đau bụng	4
abnormal_menstruation	Rối loạn kinh nguyệt	6
acidity	Chứng axit dạ dày	3
acute_liver_failure	Suy gan cấp tính	6
altered_sensorium	Thay đổi nhận thức	2

- **disease_diagnosis**: chứa thông tin một tập (mảng) các triệu chứng kèm theo bệnh đã được chẩn đoán ra từ tập triệu chứng đó. Ví dụ:

disease	symptoms
AIDS	["muscle_wasting", "patches_in_throat", "high_fever", "extra_marital_contacts"]
AIDS	["patches_in_throat", "high_fever", "extra_marital_contacts"]
AIDS	["muscle_wasting", "high_fever", "extra_marital_contacts"]
AIDS	["muscle_wasting", "patches_in_throat", "extra_marital_contacts"]
AIDS	["muscle_wasting", "patches_in_throat", "high_fever"]

- **disease_category**: chứa thông tin danh mục bệnh theo chuẩn quốc tế ICD-10 bao gồm mã danh mục và tên danh mục. Ví dụ:

category_key	category_name
A00	Bệnh tả
A01	Bệnh thương hàn và phó thương hàn
A02	Nhiễm Salmonella khác
A03	Bệnh lỵ trực khuẩn
A04	Nhiễm khuẩn đường ruột do vi khuẩn khác

- **disease_icd_10**: chứa thông tin bệnh theo chuẩn quốc tế ICD-10 bao gồm mã bệnh và tên bệnh. Ví dụ:

code	english_name	vietnamese_name
A000	Cholera due to Vibrio cholerae 01, biovar cholerae	Bệnh tả do Vibrio cholerae 01, type sinh học cholerae
A001	Cholera due to Vibrio cholerae 01, biovar eltor	Bệnh tả do Vibrio cholerae 01, type sinh học eltor
A009	Cholera, unspecified	Bệnh tả, không đặc hiệu
A010	Typhoid	Thương hàn
A011	Paratyphoid fever A	Bệnh phó thương hàn A

- Các dữ liệu thu thập được sẽ được thêm vào SQL Server thông qua câu lệnh BULK INSERT kèm đường dẫn đến các file tương ứng.

- Lớp Đồng (Bronze layer) sẽ bao gồm các bảng sau:

- Bảng **kg_ disease_description**: chứa dữ liệu thô từ file **disease_description**, đến từ nguồn Kaggle.

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú
disease	nvarchar(255)	Tên bệnh
icd_code	nvarchar(10)	Mã ICD-10
description	nvarchar(max)	Mô tả bệnh

- Bảng **kg_symptom_severity**: chứa dữ liệu thô từ file **symptom_severity**, đến từ nguồn Kaggle.

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú
english_name	nvarchar(255)	Tên triệu chứng (tiếng Anh)
vietnamese_name	nvarchar(255)	Tên triệu chứng (tiếng Việt)
symptom_severity	int	Mức độ nghiêm trọng

- Bảng **kg_disease_diagnosis**: chứa dữ liệu thô từ file **disease_diagnosis**, đến từ nguồn Kaggle.

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú
disease	nvarchar(255)	Bệnh được chẩn đoán ra
symptoms	nvarchar(max)	Tập các triệu chứng xuất hiện

- Bảng **ccms_disease_category**: chứa dữ liệu thô từ file **disease_category**, đến từ nguồn CCMS.

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú
category_key	nvarchar(10)	Mã danh mục ICD-10
category_name	nvarchar(max)	Tên danh mục (tiếng Việt)

- Bảng **ccms_disease_ icd_10**: chứa dữ liệu thô từ file **disease_ icd_10**, đến từ nguồn CCMS.

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú
code	nvarchar(10)	Mã bệnh ICD-10
english_name	nvarchar(max)	Tên bệnh (tiếng Anh)
vietnamese_name	nvarchar(max)	Tên bệnh (tiếng Việt)

III. Xây dựng lớp Bạc (Silver layer) và xử lý dữ liệu thô

- Nhận thấy dữ liệu thô đang được lưu trữ ở dạng phi cấu trúc và bán cấu trúc, ta cần phải chuyển chúng về dạng có cấu trúc để tiện cho việc lưu trữ và xử lý về sau (Ví dụ để tạo các bảng dimension, fact, data mart,…).

- Lớp Bạc (Silver layer) sẽ bao gồm các bảng sau:

- Bảng **kg_ disease**: chứa dữ liệu sau khi làm sạch, xử lý và biến đổi từ bảng **kg_ disease_description**, đến từ lớp Đồng (Bronze layer).

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú	Áp dụng xử lý
disease_id	int	Mã bệnh (SQL PK)	Sử dụng hàm tự động tăng IDENTITY(1,1)
disease	nvarchar(255)	Tên bệnh	Sử dụng hàm TRIM để loại bỏ khảng trắng thừa
icd_code	nvarchar(10)	Mã ICD-10	Sử dụng hàm TRIM để loại bỏ khoảng trắng thừa, hàm REPLACE để thay

			dấu "_" thành "." và hàm UPPER để in hoa
category	nvarchar(10)	Mã danh mục	Sử dụng hàm LEFT để tách 3 ký tự đầu tiên của mã ICD-10 và hàm UPPER để in hoa
description	nvarchar(max)	Mô tả bệnh	Sử dụng hàm TRIM và hàm REPLACE để loại bỏ khoảng trắng thừa
dwh_create_date	datetime2	Ngày tạo data warehouse	Sử dụng hàm thời gian hiện tại GETDATE()

- Bảng **kg_ symptom**: chứa dữ liệu sau khi làm sạch, xử lý và biến đổi từ bảng **kg_symptom_severity**, đến từ lớp Đồng (Bronze layer).

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú	Áp dụng xử lý
symptom_id	int	Mã triệu chứng (SQL PK)	Sử dụng hàm tự động tăng IDENTITY(1,1)
english_name	nvarchar(255)	Tên triệu chứng (tiếng Anh)	Sử dụng hàm TRIM để loại bỏ khoảng trắng thừa, hàm REPLACE để thay dấu "_" thành "." và hàm các hàm CONCAT, LEFT, LEN, LOWER, UPPER, SUBSTRING để in hoa chữ cái đầu tiên và in thường các chữ cái còn lại
vietnamese_name	nvarchar(255)	Tên triệu chứng (tiếng Việt)	Sử dụng hàm TRIM và hàm REPLACE để loại bỏ khoảng trắng thừa

symptom_severity	int	Mức độ nghiêm trọng	Sử dụng hàm CASE, WHEN, ELSE để đặt các giá trị < 1 về 1
dwh_create_date	datetime2	Ngày tạo data warehouse	Sử dụng hàm thời gian hiện tại GETDATE()

- Bảng **kg_diagnosis**: chứa dữ liệu về các lần chẩn đoán và bệnh được chẩn đoán ra từ bảng **kg_disease_diagnosis**, đến từ lớp Đồng (Bronze layer).

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú	Áp dụng xử lý
diagnosis_id	int	Mã chẩn đoán (SQL PK)	Sử dụng hàm tự động tăng IDENTITY(1,1)
disease_id	int	Mã bệnh (SQL FK → kg_disease)	Sử dụng sub query để tìm ra mã bệnh tương ứng từ bảng **kg_disease** dựa trên sự giống nhau về tên bệnh
dwh_create_date	datetime2	Ngày tạo data warehouse	Sử dụng hàm thời gian hiện tại GETDATE()

- Bảng **kg_diagnosis_symptoms**: chứa dữ liệu về các lần chẩn đoán và các triệu chứng xuất hiện từ bảng **kg_disease_diagnosis**, đến từ lớp Đồng (Bronze layer).

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú	Áp dụng xử lý
diagnosis_id	int	Mã chẩn đoán (SQL FK → kg_diagnosis)	Sử dụng biến global @@IDENTITY để lấy mã chẩn đoán vừa được thêm từ bảng **kg_diagnosis**
symptom_id	int	Mã triệu chứng (SQL FK → kg_symptom)	Sử dụng sub query để tìm ra mã bệnh tương ứng từ bảng **kg_symptom** dựa trên sự giống nhau về tên triệu chứng

dwh_create_date	datetime2	Ngày tạo data warehouse	Sử dụng hàm thời gian hiện tại GETDATE()

- Bảng **ccms_disease_category**: chứa dữ liệu sau khi làm sạch, xử lý và biến đổi từ bảng **ccms_disease_category**, đến từ lớp Đồng (Bronze layer).

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú	Áp dụng xử lý
category_key	nvarchar(10)	Mã danh mục ICD-10	Sử dụng hàm TRIM và để loại bỏ khoảng trắng thừa và hàm UPPER để in hoa
category_name	nvarchar(max)	Tên danh mục (tiếng Việt)	Sử dụng hàm TRIM và hàm REPLACE để loại bỏ khoảng trắng thừa
dwh_create_date	datetime2	Ngày tạo data warehouse	Sử dụng hàm thời gian hiện tại GETDATE()

- Bảng **ccms_disease_icd_10**: chứa dữ liệu sau khi làm sạch, xử lý và biến đổi từ bảng **ccms_disease_icd_10**, đến từ lớp Đồng (Bronze layer).

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú	Áp dụng xử lý
code	nvarchar(10)	Mã bệnh ICD-10	Sử dụng hàm TRIM và để loại bỏ khoảng trắng thừa, hàm UPPER để in hoa, hàm CASE, WHEN, ELSE, LEN để đưa ra xử lý phù hợp dựa vào chiều dài của mã bệnh ICD-10. Khi chiều dài > 3 thì kết hợp các hàm CONCAT, LEFT và SUBSTRING để thêm dấu “.” vào sau kí tự thứ 3

english_name	nvarchar(max)	Tên bệnh (tiếng Anh)	Sử dụng hàm TRIM và để loại bỏ khoảng trắng thừa
vietnamese_name	nvarchar(max)	Tên bệnh (tiếng Việt)	Sử dụng hàm TRIM và hàm REPLACE để loại bỏ khoảng trắng thừa
dwh_create_date	datetime2	Ngày tạo data warehouse	Sử dụng hàm thời gian hiện tại GETDATE()

IV. Xây dựng lớp Vàng (Gold layer) và khai phá dữ liệu

- Dữ liệu sau khi được biến đổi ở lớp Bạc (Silver layer) đã trở nên sạch hơn và có thể được dùng cho các mục đích kinh doanh (Business Purposes).

- Dữ liệu tại lớp Vàng (Gold layer) sẽ được tổ chức thành các bảng chiều (dimension) chứa các thông tin đầy đủ về các thực thể và bảng sự kiện (fact) kết hợp các dữ liệu và phục vụ cho yêu cầu khai phá dữ liệu, hỗ trợ ra quyết đinh, tạo biểu đồ, báo cáo,… đã được định sẵn từ trước.

- Lớp Vàng (Gold layer) sẽ bao gồm các view sau:

- View **dim_common_diseases**: chứa dữ liệu về các bệnh thường gặp. Dữ liệu được kết hợp từ các bảng khác nhau, bao gồm:
    - Thông tin cơ bản về bệnh từ các bảng **kg_disease** đến từ lớp Bạc (Silver layer).
    - Thông tin danh mục bệnh từ bảng **ccms_disease_category** đến từ lớp Bạc (Silver layer).
    - Thông tin tên bệnh quốc tế và tiếng Việt từ bảng **ccms_disease_icd_10** đến từ lớp Bạc (Silver layer).

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú
Mã bệnh	int	Mã bệnh (SQL PK)
Mã bệnh theo ICD	nvarchar(10)	Mã ICD-10
Tên bệnh quốc tế	nvarchar(max)	Tên bệnh (tiếng Anh)
Tên bệnh tiếng Việt	nvarchar(max)	Tên bệnh (tiếng Việt)

Mô tả bệnh	nvarchar(max)	Mô tả bệnh
Mã phân loại	nvarchar(10)	Mã danh mục
Phân loại	nvarchar(max)	Tên danh mục (tiếng Việt)

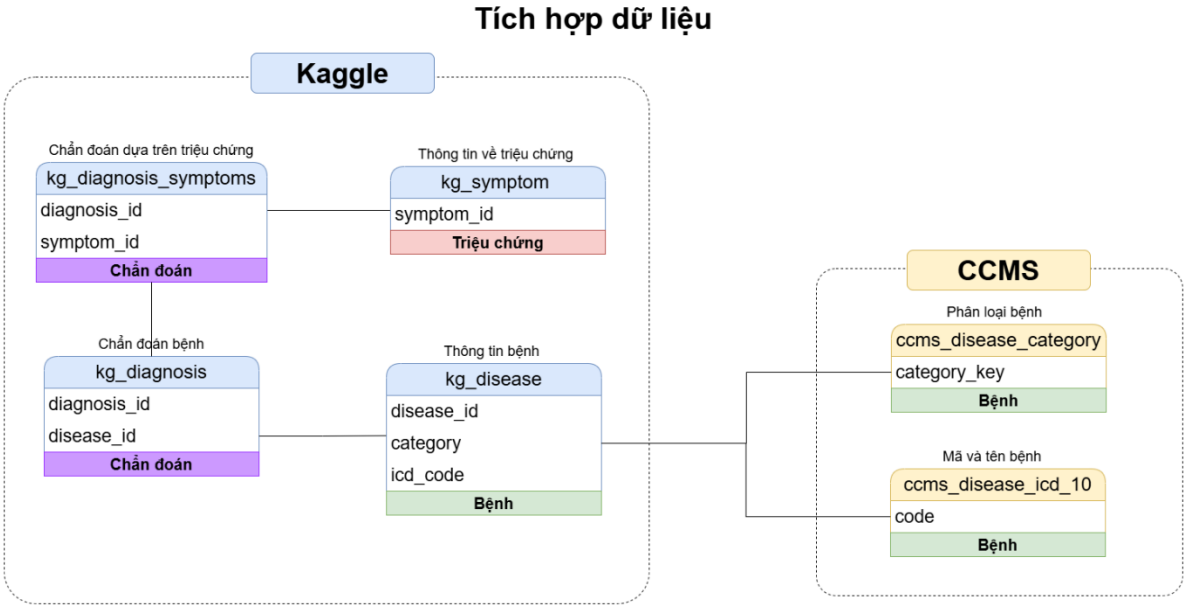
- View **dim_common_symptoms**: chứa dữ liệu về thông tin các triệu chứng thường gặp. Dữ liệu được tổng hợp từ:
    - Thông tin về triệu chứng từ bảng **kg_symptom** đến từ lớp Bạc (Silver layer).

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú
Mã triệu chứng	int	Mã triệu chứng (SQL PK)
Tên triệu chứng tiếng Anh	nvarchar(255)	Tên triệu chứng (tiếng Anh)
Tên triệu chứng tiếng Việt	nvarchar(255)	Tên triệu chứng (tiếng Việt)
Mức độ nghiêm trọng	int	Mức độ nghiêm trọng

- View **fact_likely_diseases_base_on_symptoms**: chứa thông tin các bệnh hay gặp khi xuất hiện một triệu chứng nào đó, kèm tỉ lệ (%) các bệnh ấy sẽ xuất hiện. Dữ liệu được kết hợp từ các bảng khác nhau, bao gồm:
    - Thông tin về các bệnh thường gặp từ view **dim_common_diseases** đến từ lớp Vàng (Gold layer).
    - Thông tin về triệu chứng từ view **dim_common_symptoms** đến từ lớp Vàng (Gold layer).
    - Thông tin về các lần chẩn đoán bệnh, tập hợp các triệu chứng, bệnh được chẩn đoán ra từ các bảng **kg_diagnosis** và **kg_diagnosis_symptoms** đến từ lớp Bạc (Silver layer).
- Ví dụ: xét triệu chứng **đau đầu**, tỉ lệ (%) các bệnh có thể mắc phải là:
    - Bệnh sốt đăng gơ (13.51%)
    - Bệnh thuỷ đậu không biến chứng (12.16%)
    - Bệnh đau nửa đầu [migraine], không đặc hiệu (12.16%)
    - Hạ đường máu không đặc hiệu (10.81%)
    - Viêm mũi họng cấp [cảm thường] (10.81%)
    - Thương hàn (10.81%)
    - Bệnh sốt rét do Plasmodium malariae không kèm theo biến chứng (9.46%)
    - Chóng mặt kịch phát lành tính (8,11%)
    - Tăng huyết áp vô căn (nguyên phát) (6.76%)

- Xuất huyết nội sọ, không đặc hiệu (5.41%)

Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa/ Ghi chú
Tên triệu chứng	nvarchar(255)	Tên triệu chứng (tiếng Việt)
Mã bệnh theo ICD	nvarchar(10)	Mã ICD-10
Tên bệnh	nvarchar(max)	Tên bệnh (tiếng Việt)
Số lần triệu chứng được ghi nhận	int	Đếm tổng số lần triệu chứng này đã xuất hiện (không kể bệnh nào được chẩn đoán ra)
Số lần bệnh được chẩn đoán	int	Đếm tổng số lần triệu chứng này đã xuất hiện và bệnh này được chẩn đoán ra
Tỉ lệ (%) mắc bệnh khi có triệu chứng này	decimal(10, 2)	Tỉ lệ (%) xuất hiện bệnh khi có triệu chứng, tính bằng công thức sau: $\frac{\text{Số lần bệnh được chẩn đoán}}{\text{Số lần triệu chứng được ghi nhận}} * 100.0\%$

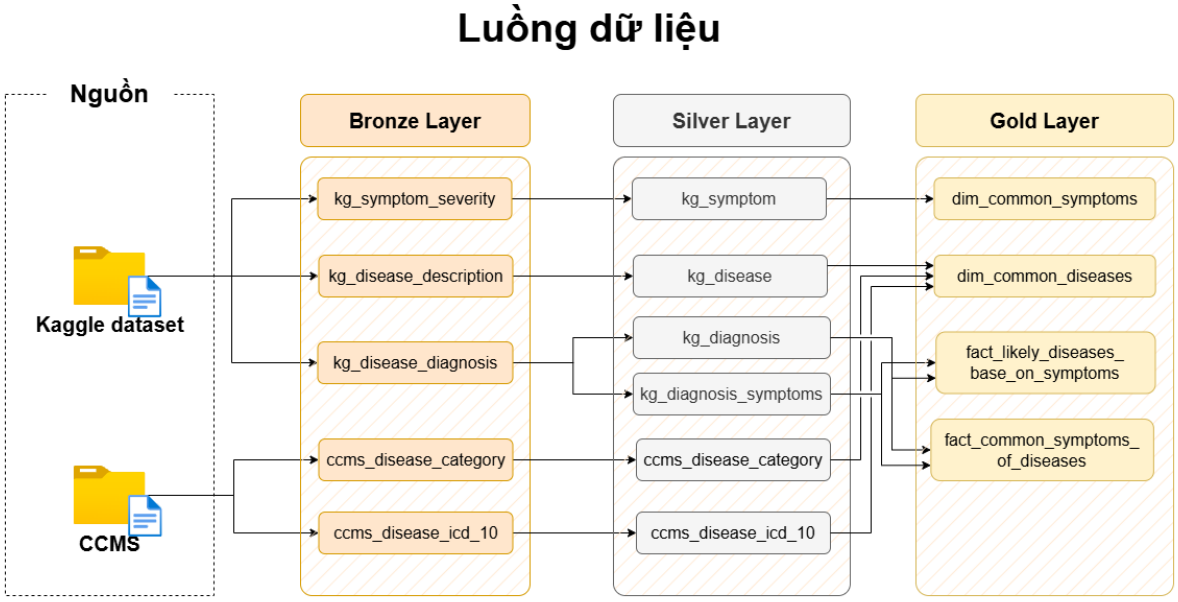


Quá trình tích hợp dữ liệu giữa các bảng dimension

- View **fact_common_symptoms_of_diseases**: chứa thông tin các triệu chứng hay gặp nhất ở một bệnh nào đó, kèm tỉ lệ (%) các triệu chứng ấy sẽ xuất hiện. Dữ liệu được kết hợp từ các bảng khác nhau, bao gồm:

    - Thông tin về các bệnh thường gặp từ view **dim_common_diseases** đến từ lớp Vàng (Gold layer).
    - Thông tin về triệu chứng từ view **dim_common_symptoms** đến từ lớp Vàng (Gold layer).
    - Thông tin về các lần chẩn đoán bệnh, tập hợp các triệu chứng, bệnh được chẩn đoán ra từ các bảng **kg_diagnosis** và **kg_diagnosis_symptoms** đến từ lớp Bạc (Silver layer).
- Ví dụ: xét bệnh **COVID- 19, xác định có vi rút**, tỉ lệ (%) các triệu chứng có thể xuất hiện là:
    - Ho (100.00%)
    - Sốt cao (100.00%)
    - Mệt mỏi (100.00%)
    - Mất vị giác (77.78%)
    - Mất khứu giác (22.22%)

Tên thuộc tính	**Kiểu dữ liệu**	**Ý nghĩa/ Ghi chú**
Mã bệnh theo ICD	nvarchar(10)	Mã ICD-10
Tên bệnh	nvarchar(max)	Tên bệnh (tiếng Việt)
Tên triệu chứng	nvarchar(255)	Tên triệu chứng (tiếng Việt)
Số lần bệnh được chẩn đoán	int	Đếm tổng số lần bệnh này đã được chẩn đoán (không kể triệu chứng nào được ghi nhận)
Số lần xuất hiện triệu chứng	int	Đếm tổng số lần bệnh này đã được chẩn đoán và triệu chứng này được ghi nhận
Tỉ lệ (%) xuất hiện triệu chứng	decimal(10, 2)	Tỉ lệ (%) xuất hiện triệu chứng khi mắc bệnh, tính bằng công thức sau: $\frac{\text{Số lần triệu xuất hiện triệu chứng}}{\text{Số lần bệnh được chẩn đoán}} * 100.0\%$



Luồng dữ liệu qua ba lớp Đồng - Bạc - Vàng