Ho Chi Minh city University of Technology
Faculty of Computer Science and Engineering

BÁO CÁO LUẬN VĂN THẠC SĨ

Đề tài:

Xây dựng mô hình Data Lakehouse trên đám mây lai cho lưu trữ và xử lý dữ liệu y tế lớn

Developing a Data Lakehouse for Healthcare Big Data in a Hybrid Cloud and On-Premise Environment

Giảng viên hướng dẫn:
PGS.TS Thoại Nam
TS. Nguyễn Lê Duy Lai

Học viên thực hiện:
Đinh Thanh Phong – 2270243

Hồ Chí Minh - 12/2025

ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC BÁCH KHOA

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên: ĐINH THANH PHONG.....MSHV: 2270243.....

Ngày, tháng, năm sinh: 13/09/1994.....Nơi sinh: Vĩnh Long.....

Chuyên ngành: Khoa Học Máy Tính..... Mã số : 8480101.....

I. TÊN ĐỀ TÀI: Xây dựng mô hình Data Lakehouse trên đám mây lai cho lưu trữ và xử lý dữ liệu y tế lớn.

II. NHIỆM VỤ VÀ NỘI DUNG:

Nhiệm vụ và nội dung của luận văn bao gồm các công việc chính sau đây:

1. Nghiên cứu tổng quan về dữ liệu y tế lớn và các mô hình lưu trữ, xử lý dữ liệu hiện có như Data Warehouse, Data Lake và Data Lakehouse; phân tích ưu, nhược điểm của các mô hình này trong bối cảnh ứng dụng cho lĩnh vực y tế.
2. Phân tích đặc thù của các loại dữ liệu y tế trong bệnh viện, bao gồm dữ liệu sinh hiệu thời gian thực từ các thiết bị IoT y tế, dữ liệu hình ảnh y tế dung lượng lớn và dữ liệu hồ sơ bệnh nhân có cấu trúc.
3. Đề xuất mô hình kiến trúc Data Lakehouse trên nền tảng đám mây lai (Hybrid Cloud–On-Premise) phù hợp với yêu cầu bảo mật, chủ quyền dữ liệu và khả năng mở rộng cho ứng dụng trong lĩnh vực y tế.
4. Thiết kế kiến trúc tổng thể hệ thống theo Medallion Architecture, bao gồm các lớp Ingestion, Bronze, Silver và Gold; phân tách rõ ràng vai trò giữa hạ tầng on-premise và môi trường đám mây.
5. Xây dựng và triển khai các pipeline xử lý dữ liệu y tế, bao gồm:
    - Pipeline xử lý dữ liệu ECG thời gian thực trong phòng hồi sức tích cực (ICU).
    - Pipeline xử lý dữ liệu hình ảnh MRI theo cơ chế batch.
    - Pipeline đồng bộ dữ liệu đã xử lý và ẩn danh lên môi trường đám mây phục vụ phân tích và nghiên cứu.
6. Đánh giá hiệu năng và khả năng vận hành của hệ thống thông qua các kịch bản thử nghiệm, bao gồm độ trễ xử lý, mức sử dụng tài nguyên và tính ổn định của các pipeline dữ liệu.

III. NGÀY GIAO NHIỆM VỤ : 25/08/2025.....

IV. NGÀY HOÀN THÀNH NHIỆM VỤ: 15/12/2025.....

V. CÁN BỘ HƯỚNG DẪN (Ghi rõ học hàm, học vị, họ, tên): PGS.TS Thoại Nam – TS. Nguyễn Lê Duy Lai.....

.....

Tp. HCM, ngày . . . . tháng .. . . năm 20....

CÁN BỘ HƯỚNG DẪN	**CHỦ NHIỆM BỘ MÔN ĐÀO TẠO**
(Họ tên và chữ ký)	(Họ tên và chữ ký)

TRƯỞNG KHOA KH&KT MT
(Họ tên và chữ ký)

***Ghi chú**: Học viên phải đóng tờ nhiệm vụ này vào trang đầu tiên của tập thuyết minh LV*

TÓM TẮT

Sự gia tăng nhanh chóng của dữ liệu y tế, bao gồm tín hiệu sinh lý thời gian thực và dữ liệu hình ảnh y khoa dung lượng lớn, đặt ra nhiều thách thức trong lưu trữ, xử lý và phân tích dữ liệu. Các kiến trúc kho dữ liệu truyền thống khó đáp ứng yêu cầu về khả năng mở rộng, xử lý thời gian thực và quản trị dữ liệu trong bối cảnh hiện nay. Kiến trúc Data Lakehouse, kết hợp ưu điểm của Data Lake và Data Warehouse, cùng với mô hình triển khai đám mây lai (hybrid cloud–on-premise), được xem là hướng tiếp cận phù hợp cho bài toán dữ liệu y tế lớn.

Luận văn này nghiên cứu và đề xuất một mô hình Data Lakehouse trên nền tảng đám mây lai (Hybrid Cloud–On-Premise) nhằm phục vụ lưu trữ và xử lý dữ liệu y tế lớn đa dạng về định dạng và yêu cầu nghiêm ngặt về bảo mật cũng như chủ quyền dữ liệu.

Trên cơ sở phân tích đặc thù của các loại dữ liệu y tế phổ biến trong bệnh viện, bao gồm dữ liệu sinh hiệu thời gian thực từ thiết bị IoT y tế, dữ liệu hình ảnh y tế dung lượng lớn và dữ liệu hồ sơ bệnh nhân có cấu trúc, luận văn chỉ ra những hạn chế của các mô hình lưu trữ và xử lý truyền thống, cũng như các kiến trúc chỉ thuần on-premise hoặc thuần cloud. Đặc biệt, trong bối cảnh pháp lý tại Việt Nam, việc đưa toàn bộ dữ liệu y tế lên môi trường đám mây gặp nhiều rào cản liên quan đến bảo mật và tuân thủ quy định.

Từ đó, luận văn đề xuất một kiến trúc Data Lakehouse đám mây lai, trong đó hạ tầng on-premise của bệnh viện đóng vai trò trung tâm trong việc thu thập, lưu trữ và xử lý dữ liệu y tế gốc, còn môi trường đám mây được sử dụng để mở rộng khả năng phân tích và nghiên cứu trên các tập dữ liệu đã được xử lý và ẩn danh. Kiến trúc được thiết kế theo Medallion Architecture với các lớp Bronze, Silver và Gold, hỗ trợ đồng thời xử lý dữ liệu thời gian thực và dữ liệu theo cơ chế batch.

Luận văn triển khai và mô phỏng các pipeline xử lý dữ liệu ECG thời gian thực trong phòng ICU và pipeline xử lý dữ liệu hình ảnh MRI theo cơ chế batch, sử dụng các công nghệ như MQTT, Apache NiFi, Apache Spark Structured Streaming, Delta Lake, Apache Airflow và Databricks. Kết quả đánh giá cho thấy hệ thống đáp ứng tốt các yêu cầu về độ trễ, tính toàn vẹn dữ liệu và khả năng mở rộng, đồng thời đảm bảo các nguyên tắc bảo mật và truy vết nguồn gốc dữ liệu y tế.

Các kết quả đạt được chứng minh tính khả thi và giá trị thực tiễn của mô hình Data Lakehouse đám mây lai trong lĩnh vực y tế, đồng thời cung cấp một mô hình tham chiếu có thể áp dụng cho các bệnh viện quy mô vừa và lớn tại Việt Nam trong quá trình hiện đại hóa hệ thống lưu trữ và xử lý dữ liệu y tế.

Mục Lục

1. MỞ ĐẦU

1.1. Bối cảnh và động lực nghiên cứu

Sự bùng nổ của cuộc cách mạng công nghiệp 4.0 đã thúc đẩy mạnh mẽ quá trình chuyển đổi số trong lĩnh vực y tế, dẫn đến sự gia tăng nhanh chóng của dữ liệu y tế cả về khối lượng lẫn độ phức tạp. Dữ liệu này không chỉ bao gồm các hồ sơ bệnh án có cấu trúc mà còn chứa một lượng lớn dữ liệu phi cấu trúc và bán cấu trúc như hình ảnh y khoa dung lượng cao (MRI, CT) và các tín hiệu sinh hiệu thời gian thực từ thiết bị IoT y tế (ECG, EEG). Việc lưu trữ và khai thác hiệu quả nguồn tài nguyên này đang trở thành một thách thức lớn đối với hạ tầng công nghệ thông tin của các bệnh viện hiện nay.

Về mặt pháp lý và quy định lưu trữ, bối cảnh tại Việt Nam đặt ra những yêu cầu vô cùng khắt khe. Theo Nghị định số 42/2025/NĐ-CP và Luật Lưu trữ năm 2024, các cơ sở y tế có nghĩa vụ phải lưu trữ hồ sơ, tài liệu ngành y tế với thời hạn từ 10 năm đến vĩnh viễn. Đồng thời, Luật Bảo vệ dữ liệu cá nhân 2025 (có hiệu lực từ 01/01/2026) quy định rất nghiêm ngặt về việc chuyển dữ liệu cá nhân ra nước ngoài, yêu cầu phải có báo cáo đánh giá tác động và sự kiểm soát chặt chẽ của cơ quan chuyên trách. Những rào cản pháp lý này khiến việc triển khai các giải pháp lưu trữ thuần túy trên điện toán đám mây công cộng (Public Cloud) gặp nhiều khó khăn do lo ngại về chủ quyền dữ liệu và an ninh quốc gia.

Về mặt kỹ thuật và thực tiễn vận hành, các hệ thống y tế hiện nay đang đối mặt với ba vấn đề cốt lõi:

• Hạn chế của thiết bị y tế: Các thiết bị đầu cuối như máy điện tâm đồ hay máy chụp cộng hưởng từ thường có bộ nhớ giới hạn và sẽ tự động ghi đè dữ liệu cũ khi đầy, dẫn đến nguy cơ mất mát các thông tin lâm sàng quý giá nếu không có phương án thu thập tự động và tập trung kịp thời.

• Sự đa dạng về định dạng: Dữ liệu y tế đến từ nhiều nhà sản xuất khác nhau (GE, Siemens, Philips, Toshiba) với các định dạng đặc thù, đòi hỏi một hệ thống có khả năng chuẩn hóa và tích hợp dữ liệu đa nguồn.

• Sự bất cập của kiến trúc truyền thống: Các mô hình kho dữ liệu (Data Warehouse) truyền thống thường tốn kém khi mở rộng và khó xử lý dữ liệu phi cấu trúc, trong khi các hồ dữ liệu (Data Lake) lại thiếu các tính năng quản trị, kiểm soát chất lượng và đảm bảo tính ACID (Atomicity, Consistency, Isolation, Durability) cần thiết cho dữ liệu y tế nhạy cảm.

Từ thực trạng đó, nhu cầu xây dựng một hệ thống lưu trữ có khả năng mở rộng linh hoạt, hỗ trợ phân tích AI nhưng vẫn đảm bảo tuyệt đối tính bảo mật và tuân thủ pháp luật là vô cùng cấp thiết. Kiến trúc Data Lakehouse, kết hợp giữa sự linh hoạt của Data Lake và khả năng quản trị của Data Warehouse, khi được triển khai trên nền tảng đám mây lai (Hybrid Cloud), xuất hiện như một giải pháp tối ưu. Trong mô hình này, hạ tầng On-Premise sẽ đảm nhận việc lưu trữ và xử lý dữ liệu gốc nhằm đảm bảo chủ quyền, trong khi môi trường Cloud được tận dụng để mở rộng năng lực tính toán và chia sẻ dữ liệu nghiên cứu đã được ẩn danh với các tổ chức quốc tế.

Động lực chính của nghiên cứu này là nhằm đề xuất một mô hình tham chiếu thực tế, giúp các bệnh viện quy mô vừa và lớn tại Việt Nam hiện đại hóa hệ thống dữ liệu, hỗ trợ đắc lực cho công tác chẩn đoán, điều trị và nghiên cứu khoa học trong tương lai.

1.2. Thách thức trong quản lý dữ liệu y tế

Việc quản lý dữ liệu trong ngành y tế không chỉ đơn thuần là bài toán lưu trữ mà còn là sự giao thoa phức tạp giữa công nghệ xử lý dữ liệu lớn, giới hạn phần cứng và các rào cản pháp lý khắt khe. Cụ thể, các thách thức chính bao gồm:

1.2.1. Sự bùng nổ và tính đa dạng của dữ liệu (Volume, Velocity, Variety)

Dữ liệu y tế hiện đại tăng trưởng với tốc độ chóng mặt và tồn tại dưới nhiều hình thái khác nhau, tạo ra áp lực cực lớn lên hạ tầng lưu trữ truyền thống.

• Dữ liệu sinh hiệu thời gian thực (Velocity): Các hệ thống tại phòng hồi sức tích cực (ICU) yêu cầu thu thập dữ liệu từ hàng chục thiết bị điện tâm đồ (ECG) liên tục 24/7. Với tần số lấy mẫu lên tới 500Hz và độ phân giải 12-bit, một hệ thống ICU 20 giường tạo ra dòng dữ liệu khổng lồ, đòi hỏi khả năng xử lý streaming với độ trễ cực thấp để không gây ra tình trạng nghẽn (backlog).

• Dữ liệu hình ảnh dung lượng lớn (Volume): Các kỹ thuật chẩn đoán hình ảnh tiên tiến như MRI tạo ra các bộ dữ liệu cực kỳ chi tiết. Ví dụ, một bộ dữ liệu MRI thắt lưng có thể chứa tới hơn 48.000 lát cắt với độ phân giải cao, lưu trữ phân tán trong hàng ngàn tệp DICOM (.ima). Việc quản lý và truy xuất nhanh chóng các khối dữ liệu phi cấu trúc này là một thách thức đối với các hệ thống tệp thông thường.

1.2.2. Giới hạn của thiết bị đầu cuối và sự phân mảnh định dạng

Một rào cản thực tế rất lớn nằm ở chính các thiết bị y tế hiện có tại bệnh viện.

• Nguy cơ mất mát dữ liệu: Nhiều thiết bị y tế có bộ nhớ đệm hạn chế và được thiết kế theo cơ chế tự động ghi đè dữ liệu cũ khi dung lượng đầy. Nếu không có

quy trình thu thập tự động và tập trung kịp thời, các dữ liệu lâm sàng quý giá sẽ vĩnh viễn bị mất đi.

• Sự thiếu thống nhất: Dữ liệu y tế bị phân mảnh do phụ thuộc vào công nghệ và tiêu chuẩn riêng biệt của từng nhà sản xuất như GE, Siemens, Philips hay Toshiba. Điều này đòi hỏi hệ thống quản lý phải có khả năng chuẩn hóa (schema enforcement) và chuyển đổi dữ liệu đa nguồn về một định dạng thống nhất để phục vụ phân tích.

1.2.3. Hạn chế của các kiến trúc lưu trữ truyền thống

Các mô hình lưu trữ cũ đang dần bộc lộ những khiếm khuyết khi đối mặt với dữ liệu y tế lớn:

• Data Warehouse (Kho dữ liệu): Mặc dù hỗ trợ quản trị tốt nhưng lại rất tốn kém khi mở rộng quy mô và chỉ làm việc hiệu quả với dữ liệu có cấu trúc, hoàn toàn không phù hợp để lưu trữ hình ảnh hay tín hiệu sinh hiệu thô.

• Data Lake (Hồ dữ liệu): Cho phép lưu trữ mọi loại dữ liệu với chi phí rẻ nhưng lại thiếu các tính năng quản trị dữ liệu nhạy cảm, dễ rơi vào tình trạng "data swamp" (đầm lầy dữ liệu), nơi dữ liệu bị mất kiểm soát về chất lượng và không đảm bảo tính toàn vẹn (ACID).

1.2.4. Thách thức về bảo mật và chủ quyền dữ liệu

Dữ liệu y tế là loại dữ liệu đặc biệt nhạy cảm, chịu sự điều chỉnh của nhiều đạo luật như HIPAA, GDPR quốc tế và các quy định mới của Việt Nam.

• Rào cản đám mây: Việc đưa toàn bộ dữ liệu y tế lên môi trường đám mây thuần túy (Public Cloud) gặp nhiều rào cản liên quan đến tuân thủ pháp lý về chủ quyền dữ liệu cư trú tại lãnh thổ quốc gia.

• Bài toán chia sẻ và bảo mật: Nhu cầu chia sẻ dữ liệu giữa các viện nghiên cứu quốc tế (như WHO, Oxford) là rất lớn, nhưng phải đảm bảo dữ liệu đã được ẩn danh hoàn toàn và có khả năng truy vết nguồn gốc (lineage) để tránh rò rỉ thông tin cá nhân của bệnh nhân.

Những thách thức này đòi hỏi một hướng tiếp cận mới: một kiến trúc không chỉ mạnh về hiệu năng xử lý mà còn phải linh hoạt trong việc phối hợp giữa hạ tầng nội bộ (On-premise) an toàn và hạ tầng đám mây (Cloud) mạnh mẽ.

1.3. Sự nổi lên của kiến trúc Data Lakehouse

Trong những năm gần đây, cùng với sự bùng nổ của dữ liệu lớn (Big Data), các tổ chức và doanh nghiệp ngày càng phải đối mặt với yêu cầu lưu trữ và xử lý khối

lượng dữ liệu ngày càng đa dạng về loại hình, cấu trúc và tốc độ phát sinh. Đặc biệt trong lĩnh vực y tế, dữ liệu không chỉ bao gồm dữ liệu có cấu trúc truyền thống như hồ sơ bệnh án, thông tin bệnh nhân, mà còn bao gồm dữ liệu bán cấu trúc và phi cấu trúc như tín hiệu sinh học thời gian thực (ECG, EEG), hình ảnh y khoa (MRI, CT, X-ray), log thiết bị và dữ liệu IoT y tế. Điều này đặt ra những thách thức lớn đối với các mô hình kiến trúc dữ liệu truyền thống.

Mô hình Data Warehouse, vốn được thiết kế để xử lý dữ liệu có cấu trúc và phục vụ các truy vấn phân tích định kỳ, tỏ ra kém linh hoạt khi phải mở rộng sang các dạng dữ liệu phi cấu trúc và bán cấu trúc. Quá trình thiết kế schema cứng nhắc, chi phí lưu trữ cao và độ trễ trong xử lý khiến Data Warehouse không còn phù hợp trong bối cảnh dữ liệu lớn hiện đại. Ngược lại, mô hình Data Lake cho phép lưu trữ dữ liệu ở dạng thô, đa định dạng và chi phí thấp, tuy nhiên lại gặp phải các vấn đề nghiêm trọng về quản lý schema, chất lượng dữ liệu, tính nhất quán và khả năng truy vấn phân tích phức tạp. Những hạn chế này thường được mô tả bằng khái niệm “data swamp”, khi Data Lake trở nên khó khai thác và khó kiểm soát.

Trước những hạn chế của cả Data Warehouse và Data Lake, kiến trúc Data Lakehouse đã được đề xuất như một hướng tiếp cận kết hợp ưu điểm của hai mô hình này. Data Lakehouse là kiến trúc dữ liệu cho phép lưu trữ dữ liệu đa dạng trên nền tảng Data Lake, đồng thời cung cấp các đặc tính quản trị, hiệu năng và độ tin cậy tương tự như Data Warehouse. Thông qua việc áp dụng các công nghệ lưu trữ hiện đại như Delta Lake, Apache Hudi hoặc Apache Iceberg, kiến trúc Data Lakehouse hỗ trợ các tính năng quan trọng như ACID transactions, schema enforcement, schema evolution, versioning và time travel ngay trên nền tảng lưu trữ chi phí thấp.

Một đặc điểm quan trọng của Data Lakehouse là khả năng thống nhất các pipeline xử lý dữ liệu batch và streaming trong cùng một kiến trúc. Thay vì duy trì các hệ thống riêng biệt cho xử lý dữ liệu thời gian thực và dữ liệu theo lô, Data Lakehouse cho phép xử lý đồng thời cả hai loại dữ liệu trên cùng một lớp lưu trữ và cùng một hệ sinh thái công cụ phân tích. Điều này đặc biệt phù hợp với các hệ thống y tế hiện đại, nơi dữ liệu thời gian thực từ thiết bị theo dõi bệnh nhân cần được xử lý song song với dữ liệu lịch sử phục vụ nghiên cứu và phân tích chuyên sâu.

Bên cạnh đó, kiến trúc Data Lakehouse còn hỗ trợ tốt cho các nhu cầu phân tích nâng cao như học máy (Machine Learning), trí tuệ nhân tạo (AI) và phân tích dự đoán. Dữ liệu được lưu trữ tập trung, có kiểm soát chất lượng và truy xuất hiệu quả giúp giảm đáng kể độ phức tạp trong việc xây dựng các mô hình phân tích và đào tạo thuật toán. Đây là một lợi thế quan trọng trong lĩnh vực y tế, nơi các mô hình AI cần truy cập đồng thời dữ liệu lịch sử lớn và dữ liệu cập nhật liên tục.

Sự nổi lên của kiến trúc Data Lakehouse không chỉ xuất phát từ nhu cầu kỹ thuật mà còn gắn liền với xu hướng chuyển đổi số và điện toán đám mây. Trong bối cảnh nhiều tổ chức áp dụng mô hình hybrid cloud hoặc multi-cloud, Data Lakehouse cho phép triển khai linh hoạt trên cả hạ tầng on-premise và cloud, tận dụng khả năng mở rộng của đám mây trong khi vẫn đảm bảo yêu cầu bảo mật và chủ quyền dữ liệu. Điều này đặc biệt quan trọng đối với dữ liệu y tế, vốn chịu sự quản lý chặt chẽ bởi các quy định pháp lý và yêu cầu bảo mật nghiêm ngặt.

1.4. Giải pháp đám mây lai (Hybrid Cloud)

Điện toán đám mây đã và đang đóng vai trò quan trọng trong quá trình chuyển đổi số của các tổ chức nhờ khả năng cung cấp tài nguyên tính toán linh hoạt, mở rộng nhanh chóng và tối ưu chi phí đầu tư ban đầu. Các nền tảng đám mây công cộng cho phép triển khai các hệ thống phân tích dữ liệu lớn, xử lý song song và học máy với hiệu năng cao mà không cần đầu tư hạ tầng vật lý phức tạp. Tuy nhiên, đối với lĩnh vực y tế, việc đưa toàn bộ dữ liệu lên đám mây công cộng vẫn tồn tại nhiều rào cản đáng kể.

Trước hết, dữ liệu y tế là loại dữ liệu nhạy cảm, liên quan trực tiếp đến thông tin cá nhân, tình trạng sức khỏe và lịch sử điều trị của bệnh nhân. Các yêu cầu về bảo mật, quyền riêng tư và tuân thủ quy định pháp lý (như bảo vệ dữ liệu cá nhân và chủ quyền dữ liệu) khiến nhiều cơ sở y tế không thể hoặc không được phép lưu trữ và xử lý toàn bộ dữ liệu trên môi trường đám mây công cộng. Bên cạnh đó, các rủi ro liên quan đến kiểm soát truy cập, phụ thuộc nhà cung cấp và khả năng đáp ứng các tiêu chuẩn bảo mật đặc thù của ngành y tế cũng là những thách thức lớn.

Ngược lại, các hệ thống triển khai hoàn toàn theo mô hình on-premise mang lại mức độ kiểm soát cao đối với dữ liệu và hạ tầng, giúp các tổ chức y tế dễ dàng đáp ứng các yêu cầu về bảo mật và tuân thủ pháp lý. Tuy nhiên, mô hình này lại bộc lộ nhiều hạn chế về khả năng mở rộng, đặc biệt khi khối lượng dữ liệu tăng nhanh và nhu cầu phân tích ngày càng phức tạp. Việc đầu tư, vận hành và bảo trì hạ tầng tính toán lớn tại chỗ đòi hỏi chi phí cao, thời gian triển khai dài và nguồn nhân lực kỹ thuật chuyên sâu.

Trong bối cảnh đó, mô hình đám mây lai (Hybrid Cloud) được xem là giải pháp cân bằng, kết hợp ưu điểm của cả hai hướng tiếp cận on-premise và đám mây công cộng. Theo mô hình này, các dữ liệu y tế nhạy cảm, dữ liệu gốc và các hệ thống lõi (core systems) được lưu trữ và xử lý tại môi trường on-premise của bệnh viện hoặc trung tâm dữ liệu nội bộ. Điều này giúp đảm bảo quyền kiểm soát dữ liệu, tăng cường bảo mật và đáp ứng các yêu cầu pháp lý hiện hành.

Song song với đó, môi trường đám mây được sử dụng để phục vụ các tác vụ đòi hỏi tài nguyên lớn như xử lý dữ liệu batch, phân tích dữ liệu lớn, đào tạo mô hình học máy và lưu trữ dữ liệu đã được làm sạch hoặc ẩn danh. Việc phân tách rõ ràng vai trò giữa on-premise và cloud giúp các tổ chức y tế tận dụng được sức mạnh tính toán và khả năng mở rộng của đám mây mà không làm ảnh hưởng đến an toàn dữ liệu nhạy cảm.

Giải pháp đám mây lai cũng tạo điều kiện thuận lợi cho việc triển khai các kiến trúc dữ liệu hiện đại như Data Lakehouse. Trong mô hình này, Data Lakehouse có thể được triển khai phân tán trên cả hai môi trường, với các cơ chế kiểm soát luồng dữ liệu, quản lý quyền truy cập và đồng bộ dữ liệu giữa on-premise và cloud. Điều này cho phép xây dựng một hệ thống dữ liệu thống nhất, linh hoạt và có khả năng mở rộng, đồng thời vẫn đảm bảo các yêu cầu đặc thù của lĩnh vực y tế.

1.5. Mục tiêu của luận văn

Luận văn này hướng đến việc đề xuất, xây dựng và đánh giá một kiến trúc Data Lakehouse trong môi trường đám mây lai (Hybrid Cloud) nhằm phục vụ cho lưu trữ và xử lý dữ liệu y tế lớn. Trọng tâm của nghiên cứu không chỉ dừng lại ở việc trình bày một mô hình kiến trúc mang tính lý thuyết, mà còn tập trung vào việc triển khai một hệ thống thực nghiệm có khả năng vận hành, mở rộng và đáp ứng các yêu cầu thực tế của lĩnh vực y tế.

Xuất phát từ đặc thù của dữ liệu y tế, bao gồm sự đa dạng về loại hình dữ liệu (dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc), tốc độ phát sinh cao và yêu cầu nghiêm ngặt về bảo mật, luận văn đặt mục tiêu xây dựng một kiến trúc dữ liệu vừa đảm bảo tính linh hoạt và khả năng mở rộng, vừa đáp ứng các yêu cầu về an toàn thông tin và tuân thủ pháp lý. Trong bối cảnh đó, việc kết hợp kiến trúc Data Lakehouse với mô hình Hybrid Cloud được xem là một hướng tiếp cận phù hợp và có tính khả thi cao.

Bên cạnh việc đề xuất kiến trúc tổng thể, luận văn còn hướng đến việc hiện thực hóa mô hình thông qua các pipeline xử lý dữ liệu tiêu biểu, đại diện cho các kịch bản dữ liệu phổ biến trong lĩnh vực y tế. Thông qua việc triển khai thực nghiệm, luận văn đánh giá khả năng vận hành của hệ thống, từ đó làm rõ các ưu điểm, hạn chế và tiềm năng áp dụng trong thực tế.

Cụ thể, các mục tiêu của luận văn bao gồm:

• Xây dựng một kiến trúc Data Lakehouse dựa trên mô hình đám mây lai (Hybrid Cloud), trong đó phân tách rõ ràng vai trò của hạ tầng on-premise và môi trường đám mây, nhằm đảm bảo an toàn dữ liệu y tế nhạy cảm đồng thời tận dụng

được khả năng mở rộng và tài nguyên tính toán của đám mây.

• Thiết kế và triển khai các pipeline thu thập và xử lý dữ liệu y tế theo hai cơ chế chính: xử lý dữ liệu thời gian thực (streaming) từ các thiết bị IoT y tế như tín hiệu ECG, và xử lý dữ liệu theo lô (batch) từ các hệ thống hình ảnh y tế như MRI. Các pipeline được xây dựng theo nguyên tắc Medallion Architecture (Bronze – Silver – Gold) nhằm đảm bảo chất lượng và khả năng truy vết dữ liệu.

• Triển khai hệ thống thực nghiệm trên hạ tầng on-premise dựa trên các thành phần mã nguồn mở phổ biến, kết hợp với môi trường đám mây để thực hiện các tác vụ xử lý và lưu trữ mở rộng. Việc lựa chọn các công nghệ mã nguồn mở giúp giảm phụ thuộc vào nhà cung cấp và tăng khả năng áp dụng trong thực tế.

• Đánh giá hiệu năng, khả năng mở rộng và tính khả thi của mô hình thông qua các kịch bản thử nghiệm mô phỏng gần với môi trường vận hành thực tế. Các tiêu chí đánh giá tập trung vào độ ổn định của pipeline, khả năng chịu lỗi, độ trễ xử lý và khả năng đáp ứng khi khối lượng dữ liệu tăng.

Thông qua việc đạt được các mục tiêu trên, luận văn kỳ vọng góp phần làm rõ tính khả thi của việc áp dụng kiến trúc Data Lakehouse trong môi trường đám mây lai cho lĩnh vực y tế, đồng thời cung cấp một mô hình tham khảo cho các nghiên cứu và triển khai thực tế trong tương lai.

1.6. Phạm vi nghiên cứu và bài toán đặt ra

Trong phạm vi của luận văn này, tác giả tập trung nghiên cứu và triển khai một mô hình Data Lakehouse cho lĩnh vực y tế trong môi trường đám mây lai (Hybrid Cloud), với mục tiêu giải quyết các vấn đề liên quan đến thu thập, lưu trữ, xử lý và chia sẻ dữ liệu y tế lớn phát sinh từ nhiều nguồn khác nhau. Trọng tâm của nghiên cứu là kiến trúc hệ thống dữ liệu, các quy trình xử lý và khả năng vận hành của hạ tầng dữ liệu trong bối cảnh thực tế của lĩnh vực y tế, nơi tồn tại đồng thời các yêu cầu cao về bảo mật, độ tin cậy và khả năng mở rộng.

Luận văn không đi sâu vào việc phát triển hoặc cải tiến các thuật toán chẩn đoán y sinh, mô hình học máy hay các phương pháp phân tích lâm sàng chuyên sâu. Thay vào đó, nghiên cứu tập trung vào việc xây dựng một nền tảng dữ liệu có khả năng hỗ trợ hiệu quả cho các bài toán phân tích và nghiên cứu y sinh trong tương lai, thông qua việc đảm bảo dữ liệu được thu thập đầy đủ, lưu trữ có tổ chức, xử lý ổn định và sẵn sàng cho khai thác.

Bài toán đặt ra trong luận văn được xác định là xây dựng một hệ thống Data Lakehouse trong môi trường Hybrid Cloud có khả năng tiếp nhận và xử lý đồng thời dữ liệu y tế theo hai cơ chế chính: dữ liệu thời gian thực (streaming) và dữ liệu theo lô (batch). Hệ thống cần đảm bảo các yêu cầu về khả năng mở rộng, hiệu năng xử lý và bảo mật dữ liệu, đồng thời cho phép chia sẻ và khai thác dữ liệu phục vụ các hoạt động phân tích, nghiên cứu và ra quyết định trong lĩnh vực y tế.

Để giải quyết bài toán tổng quát trên, luận văn tập trung vào các vấn đề kỹ thuật chính sau:

• Thu thập và xử lý dữ liệu y tế thời gian thực từ các thiết bị IoT y tế, với trường hợp điển hình là dữ liệu điện tâm đồ (ECG) có tần số lấy mẫu cao (500Hz). Dữ liệu này phát sinh liên tục với tốc độ lớn, đòi hỏi hệ thống phải hỗ trợ cơ chế xử lý streaming có độ trễ thấp, đảm bảo không mất mát dữ liệu và có khả năng chịu lỗi trong quá trình vận hành.

• Xử lý và lưu trữ dữ liệu hình ảnh y tế theo cơ chế batch, bao gồm các dữ liệu CT, MRI và X-ray được lưu trữ ở định dạng DICOM. Đây là nhóm dữ liệu có dung lượng lớn, cấu trúc phức tạp và yêu cầu đặc thù về lưu trữ cũng như quản lý metadata, đòi hỏi hệ thống phải có khả năng tổ chức dữ liệu hiệu quả và hỗ trợ truy xuất phục vụ phân tích.

• Thiết kế và triển khai kiến trúc on-premise dựa trên các thành phần mã nguồn mở, nhằm đảm bảo tính chủ động trong vận hành hệ thống, kiểm soát chặt chẽ dữ liệu y tế nhạy cảm và đáp ứng các yêu cầu về bảo mật cũng như tuân thủ pháp lý.

• Kết nối và mở rộng hệ thống on-premise lên môi trường đám mây để phục vụ các tác vụ xử lý dữ liệu lớn, phân tích nâng cao và đánh giá khả năng mở rộng của kiến trúc Data Lakehouse trong mô hình Hybrid Cloud. Việc phân tách vai trò giữa on-premise và cloud được xem là yếu tố then chốt nhằm cân bằng giữa bảo mật và hiệu năng.

Trong khuôn khổ luận văn, dữ liệu được sử dụng bao gồm dữ liệu thời gian thực được mô phỏng từ các thiết bị IoT y tế, dữ liệu hình ảnh y tế theo lô và dữ liệu định danh bệnh nhân được tạo lập phục vụ mục đích thử nghiệm và đánh giá hệ thống. Việc đánh giá mô hình tập trung vào các tiêu chí kỹ thuật như khả năng thu thập dữ liệu, hiệu năng xử lý, khả năng mở rộng, khả năng chịu lỗi và tính khả thi trong triển khai thực tế, thay vì đánh giá độ chính xác của các mô hình phân tích hoặc chẩn đoán y sinh.

Mục này nhằm làm rõ phạm vi nghiên cứu và định hướng giải quyết bài toán của luận văn, đồng thời tạo cơ sở cho việc trình bày chi tiết kiến trúc hệ thống, các pipeline xử lý dữ liệu và kết quả thực nghiệm trong các chương tiếp theo.

2. TỔNG QUAN CÁC NGHIÊN CỨU LIÊN QUAN

2.1. Các nghiên cứu liên quan Hybrid Cloud Data Lakehouse

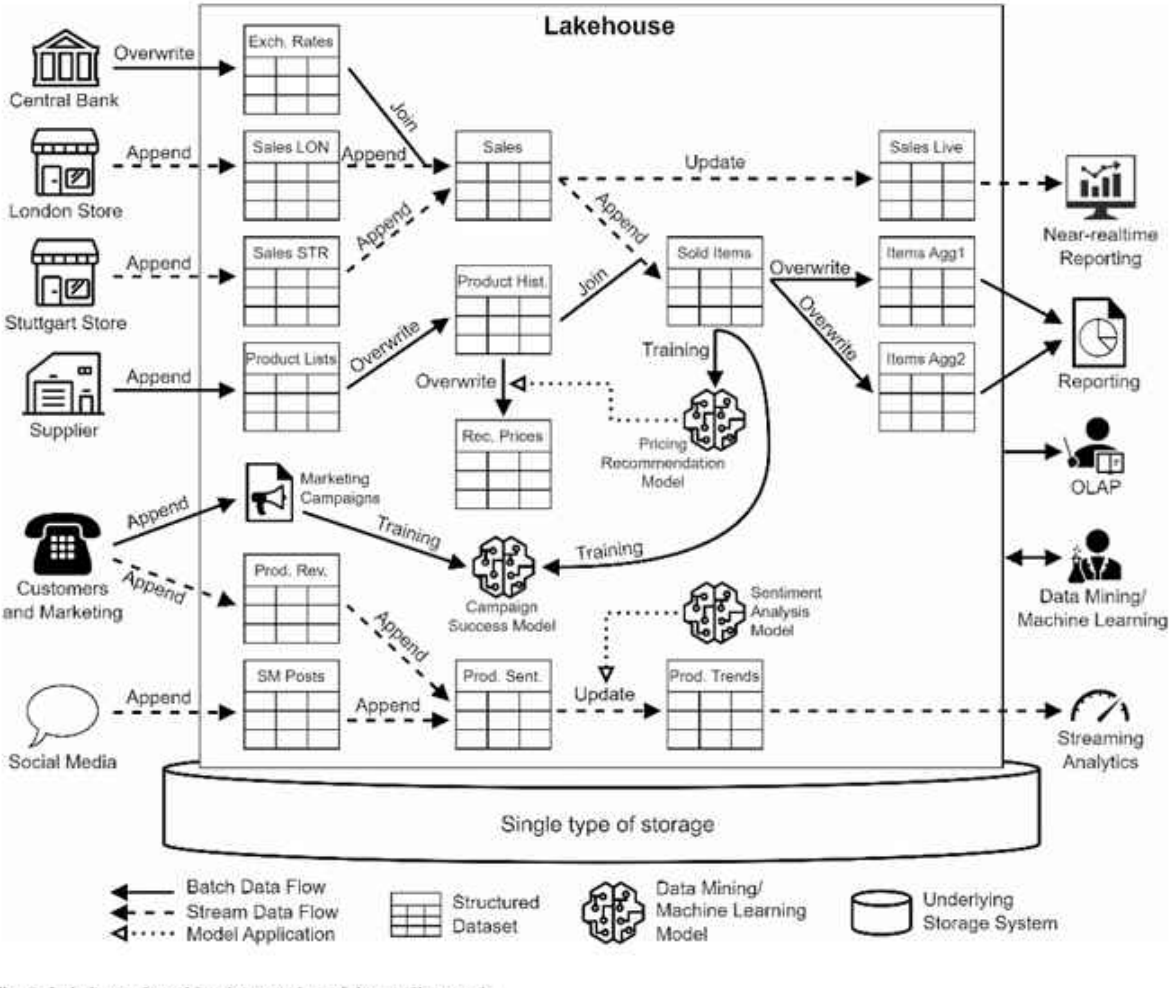
Kiến trúc Data Lakehouse là một mô hình lai tương đối mới, xuất hiện vào khoảng năm 2020–2021 và đã nhanh chóng thu hút sự quan tâm trong cộng đồng nghiên cứu cũng như công nghiệp. Mô hình này được kỳ vọng kết hợp các ưu điểm của Data Lake (lưu trữ linh hoạt, chi phí thấp, hỗ trợ dữ liệu phi cấu trúc) và Data Warehouse (hỗ trợ giao dịch, quản lý dữ liệu chặt chẽ, tối ưu hóa truy vấn) nhằm phục vụ hiệu quả cho các nhu cầu xử lý và khai thác dữ liệu lớn. Nhiều công trình nghiên cứu đã được thực hiện nhằm định nghĩa, mở rộng và triển khai kiến trúc này trong các lĩnh vực cụ thể, bao gồm cả y tế và khoa học dữ liệu.

Dưới đây là tổng quan các nghiên cứu tiêu biểu có liên quan trực tiếp đến đề tài:

- “Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics” – Armbrust et al. (2021) Đây là công trình đặt nền móng cho khái niệm Lakehouse [1]. Tác giả định nghĩa Lakehouse là hệ quản trị dữ liệu được xây dựng trên nền tảng lưu trữ trực tiếp chi phí thấp, đồng thời tích hợp các tính năng quản trị mạnh mẽ của hệ thống cơ sở dữ liệu phân tích (ACID, quản lý phiên bản, chỉ mục, tối ưu hóa truy vấn…). Tác giả cho rằng Lakehouse có thể thay thế hoàn toàn Data Warehouse nhờ vào tính mở, khả năng truy cập linh hoạt và hỗ trợ tốt cho học máy và khoa học dữ liệu.
- “Data Lakehouse - a Novel Step in Analytics Architecture ” – Orescanin & Hlupic (2021) [4] Nghiên cứu này đề xuất một kiến trúc Lakehouse cấp cao gồm ba tầng: Extraction Layer, Data Lake Layer và Data Warehouse Layer. Mỗi tầng có vai trò cụ thể trong việc tiếp nhận, lưu trữ và quản lý dữ liệu phục vụ phân tích.
- “A Lakehouse Architecture for the Management and Analysis of Heterogeneous Data for Biomedical Research and Mega-biobanks” – Oak Ridge National Laboratory (ORNL) [6] Kiến trúc Lakehouse được triển khai trong hệ thống *Knowledge Discovery Infrastructure (KDI)* của ORNL, nhằm hỗ trợ quản lý dữ liệu không đồng nhất trong nghiên cứu y sinh và các ngân hàng sinh học lớn. Hệ thống này đảm bảo tuân thủ các nguyên tắc FAIR, quản lý quyền truy cập phức tạp, và hỗ trợ tiến hóa dữ liệu dẫn xuất.

- “Design of Vessel Data Lakehouse with Big Data and AI Analysis Technology for Vessel Monitoring System” – Lee et al. (2022). Nghiên cứu này xây dựng một kiến trúc Lakehouse cho hệ thống giám sát tàu biển, nhằm xử lý dữ liệu cảm biến và truyền thông từ các tàu theo thời gian thực, đồng thời hỗ trợ phân tích dựa trên công nghệ AI và Big Data.
- “The Evolution from Data Warehouses to Data Lakehouses: A Technical Perspective” Sai Kaushik Ponnekanti (2025) [7]. Nghiên cứu này cung cấp cái nhìn toàn cảnh về sự phát triển của các kiến trúc phân tích dữ liệu, so sánh giữa Data Warehouse, Data Lake và Data Lakehouse, đồng thời thảo luận vai trò của Lakehouse trong việc đáp ứng yêu cầu xử lý dữ liệu lớn hiện nay.
- “Hybrid Cloud Databases for Big Data Analytics: A Review of Architecture, Performance, and Cost Efficiency” của Ashraful Islam (2024) [8]. Trình bày tổng quan các kiến trúc cơ sở dữ liệu đám mây lai trong phân tích dữ liệu lớn. Tác giả phân tích các mô hình triển khai phổ biến, so sánh hiệu năng và chi phí giữa các giải pháp on-premise, cloud và hybrid. Nghiên cứu nhấn mạnh tính linh hoạt và khả năng mở rộng của hệ thống lai trong xử lý dữ liệu lớn. Bài viết cũng đề cập đến các thách thức về bảo mật, đồng bộ hóa và tích hợp hệ thống. Tổng quan này giúp định hướng lựa chọn kiến trúc phù hợp cho các tổ chức có nhu cầu xử lý dữ liệu quy mô lớn.
- “Protecting Sensitive Tabular Data in Hybrid Clouds” của Maya Anderson và các cộng sự (IBM Research) (2023) [10] tập trung vào vấn đề bảo vệ dữ liệu dạng bảng nhạy cảm khi triển khai trong môi trường đám mây lai. Nghiên cứu đề xuất một giải pháp mã hóa dữ liệu linh hoạt, cho phép xử lý truy vấn mà vẫn đảm bảo bảo mật và tuân thủ quy định (ví dụ: GDPR). Hệ thống cho phép phân tách dữ liệu giữa cloud công cộng và private cloud, đảm bảo dữ liệu nhạy cảm luôn được giữ trong khu vực an toàn. Bài viết cũng trình bày mô hình truy vấn kết hợp (hybrid query execution) trên dữ liệu đã mã hóa. Giải pháp được thiết kế đặc biệt cho các tổ chức y tế và tài chính cần bảo vệ dữ liệu cá nhân trong quá trình phân tích.
- “The Lakehouse: State of the Art on Concepts and Technologies” (2024) [12] trình bày tổng quan khái niệm kiến trúc Lakehouse – một mô hình hiện đại kết hợp ưu điểm của Data Warehouse và Data Lake nhằm xử lý hiệu quả cả dữ liệu có cấu trúc và phi cấu trúc trong cùng một nền tảng. Bài viết phân tích các thành phần cốt lõi của Lakehouse như định dạng lưu trữ hỗ trợ ACID (Delta Lake, Apache Hudi, Apache Iceberg), quản lý metadata, phiên bản hóa dữ liệu và cơ chế caching. Ngoài ra, bài báo còn so sánh Lakehouse với các kiến trúc truyền thống, giới thiệu các công cụ hỗ trợ như Apache Spark và Databricks, đồng thời chỉ ra các thách thức như quản lý schema phức tạp, đảm bảo chất lượng dữ liệu và tối ưu chi phí. Đây là tài liệu tổng quan giá trị giúp định hướng nghiên cứu và

ứng dụng kiến trúc Lakehouse trong bối cảnh dữ liệu lớn và trí tuệ nhân tạo phát triển mạnh mẽ.



Hình 1. Mô hình tham khảo áp dụng kiến trúc Lakehouse

- *"From Data Warehouse to Lakehouse: A Comparative Review" (2023)[11]* trình bày cái nhìn tổng quan và so sánh giữa kiến trúc truyền thống Data Warehouse và mô hình Lakehouse hiện đại. Nội dung chính tập trung vào việc phân tích các đặc điểm kỹ thuật, khả năng mở rộng, chi phí lưu trữ, hiệu quả xử lý, và mức độ linh hoạt khi làm việc với dữ liệu có cấu trúc và phi cấu trúc. Bài viết chỉ ra rằng mô hình Lakehouse kết hợp được tính toàn vẹn và hiệu suất cao của Data Warehouse với khả năng lưu trữ linh hoạt, chi phí thấp của Data Lake, nhờ vào các công nghệ như Delta Lake, Apache Iceberg và Hudi. Tác giả cũng thảo luận về những thách thức còn tồn tại trong việc chuyển đổi từ Data Warehouse sang Lakehouse, như quản lý schema, xử lý dữ liệu thời gian thực và tích hợp hệ thống cũ. Bài báo đóng vai trò quan trọng trong việc định hướng các tổ chức khi

lựa chọn mô hình lưu trữ dữ liệu phù hợp trong bối cảnh dữ liệu ngày càng đa dạng và phức tạp.

2.2. Các nghiên cứu liên quan dữ liệu lớn trong lĩnh vực y sinh

Trong lĩnh vực y sinh, Data Lakehouse mang đến những giải pháp đặc thù cho việc xử lý và lưu trữ dữ liệu không cấu trúc (như hình ảnh y tế, hồ sơ bệnh án, và dữ liệu cảm biến IoT). Các nghiên cứu đã chỉ ra rằng, Data Lakehouse có thể giải quyết các vấn đề quan trọng như khả năng xử lý khối lượng lớn dữ liệu không đồng nhất, đồng thời tuân thủ các yêu cầu bảo mật và quyền riêng tư trong lĩnh vực y tế (như HIPAA và GDPR). Mô hình này cũng hỗ trợ tính linh hoạt trong việc kết nối và phân tích dữ liệu từ nhiều nguồn khác nhau, tạo ra một hệ sinh thái dữ liệu mạnh mẽ cho việc nghiên cứu và chăm sóc sức khỏe.

- "Big data analytics in healthcare" – C. Guo & J. Chen (2023) [13] Chương sách này phân tích vai trò của phân tích dữ liệu lớn (Big Data Analytics) trong chăm sóc sức khỏe, đặc biệt trong cải thiện chẩn đoán, điều trị và quản lý bệnh. Tác giả nhấn mạnh tầm quan trọng của việc tích hợp các nguồn dữ liệu y tế khác nhau và khai thác công nghệ trí tuệ nhân tạo để đưa ra quyết định lâm sàng chính xác và kịp thời.
- "A Responsible Framework for Applying Artificial Intelligence on Medical Images and Signals at the Point of Care: The PACS-AI Platform" – Pascal Theriault-Lauzier et al. (2024) [14] Bài nghiên cứu giới thiệu nền tảng PACS-AI, một hệ thống hỗ trợ triển khai AI vào hình ảnh y tế và tín hiệu sinh học tại điểm chăm sóc bệnh nhân. Nền tảng đảm bảo tuân thủ đạo đức, quyền riêng tư và cung cấp công cụ phân tích dữ liệu lớn theo thời gian thực để hỗ trợ chẩn đoán và điều trị.
- "Harnessing Big Data Analytics for Healthcare: A Comprehensive Review of Frameworks, Implications, Applications, and Impacts” [15] Awais Ahmed (2023) Bài tổng quan này hệ thống hóa 180 nghiên cứu trong lĩnh vực Big Data Analytics (BDA) trong chăm sóc sức khỏe. Nội dung đề cập đến các khung kiến trúc phân tích dữ liệu lớn, ứng dụng thực tiễn, thách thức triển khai (như bảo mật, tích hợp dữ liệu), và các tác động tiềm năng đến hệ thống y tế và hiệu quả chăm sóc bệnh nhân.

2.3. Hạn chế của các công trình hiện có

Các công trình nghiên cứu về kiến trúc Data Lakehouse được công bố trong thời gian gần đây chủ yếu tập trung vào việc đề xuất mô hình kiến trúc hoặc phân tích ưu điểm của Data Lakehouse so với Data Warehouse và Data Lake truyền thống. Tuy nhiên, phần lớn các nghiên cứu này chỉ xem xét việc triển khai hệ thống trong một môi trường đơn lẻ, hoặc hoàn toàn trên hạ tầng on-premise, hoặc hoàn toàn trên môi

trường đám mây công cộng. Cách tiếp cận này chưa phản ánh đúng nhu cầu triển khai thực tế trong lĩnh vực y tế, nơi tồn tại đồng thời yêu cầu bảo mật cao đối với dữ liệu nhạy cảm và nhu cầu mở rộng tài nguyên tính toán cho các tác vụ phân tích dữ liệu lớn.

Bên cạnh đó, nhiều nghiên cứu chỉ dừng lại ở mức lý thuyết hoặc thử nghiệm ở quy mô hạn chế, chưa đưa ra các đánh giá định lượng cụ thể về hiệu năng hệ thống. Các chỉ số quan trọng như tốc độ thu thập dữ liệu, độ trễ xử lý, khả năng mở rộng khi tăng tải hoặc tác động của kiến trúc triển khai đến hiệu năng xử lý dữ liệu hầu như chưa được phân tích một cách đầy đủ. Điều này khiến các mô hình được đề xuất khó áp dụng trực tiếp trong các hệ thống vận hành thực tế.

Đặc biệt, rất ít công trình nghiên cứu tập trung vào dữ liệu y tế với các đặc thù riêng như dữ liệu thời gian thực từ thiết bị IoT y tế, dữ liệu hình ảnh y tế dung lượng lớn và dữ liệu định danh bệnh nhân có yêu cầu nghiêm ngặt về bảo mật. Việc thiếu các nghiên cứu thực nghiệm trên dữ liệu y tế thực tế tạo ra khoảng cách lớn giữa nghiên cứu học thuật và nhu cầu triển khai trong các tổ chức chăm sóc sức khỏe.

2.4. Khoảng trống nghiên cứu và định vị đóng góp của luận văn

Xuất phát từ các hạn chế nêu trên, luận văn này hướng đến việc tiếp cận bài toán Data Lakehouse theo hướng thực tiễn hơn, tập trung vào môi trường triển khai và dữ liệu ứng dụng cụ thể trong lĩnh vực y tế. Khác với các nghiên cứu trước chỉ xem xét triển khai on-premise hoặc cloud, luận văn đề xuất và triển khai một mô hình Data Lakehouse trong môi trường đám mây lai (Hybrid Cloud), cho phép kết hợp ưu điểm của cả hai mô hình nhằm đáp ứng đồng thời yêu cầu bảo mật và khả năng mở rộng.

Luận văn không chỉ dừng lại ở việc đề xuất kiến trúc, mà còn xây dựng các pipeline xử lý dữ liệu hoàn chỉnh cho các kịch bản dữ liệu y tế phổ biến, bao gồm xử lý dữ liệu thời gian thực từ thiết bị IoT y tế và xử lý dữ liệu batch từ hệ thống hình ảnh y tế. Toàn bộ quy trình thu thập, lưu trữ, xử lý và chia sẻ dữ liệu được trình bày chi tiết, phản ánh sát với các điều kiện triển khai thực tế tại các cơ sở y tế.

Ngoài ra, luận văn thực hiện các thử nghiệm đánh giá hiệu năng hệ thống dựa trên các số liệu định lượng cụ thể, bao gồm tốc độ thu thập dữ liệu, độ trễ xử lý và khả năng mở rộng khi tăng khối lượng dữ liệu. Việc áp dụng trực tiếp mô hình Data Lakehouse vào dữ liệu y tế và cung cấp các kết quả thực nghiệm giúp làm rõ tính khả thi của kiến trúc đám mây lai, đồng thời cung cấp một hướng dẫn triển khai cụ thể có thể tham khảo và áp dụng trong thực tế.

3. Cơ sở lý thuyết

3.1. Dữ liệu lớn và đặc thù dữ liệu y tế (Big Data in Healthcare)

Trong kỷ nguyên chuyển đổi số, dữ liệu đã trở thành một trong những nguồn tài nguyên quan trọng nhất đối với mọi lĩnh vực kinh tế – xã hội, đặc biệt là lĩnh vực y tế. Sự phát triển mạnh mẽ của các hệ thống thông tin bệnh viện, thiết bị y tế thông minh, công nghệ chẩn đoán hình ảnh và các nền tảng chăm sóc sức khỏe từ xa đã làm cho dữ liệu y tế ngày càng gia tăng nhanh chóng cả về khối lượng, tốc độ phát sinh và mức độ đa dạng. Những đặc điểm này khiến dữ liệu y tế trở thành một trong những dạng dữ liệu điển hình của dữ liệu lớn (Big Data), đồng thời đặt ra nhiều thách thức mới cho việc lưu trữ, xử lý và khai thác dữ liệu.

Dữ liệu lớn thường được mô tả thông qua mô hình 5V, bao gồm Volume (khối lượng), Velocity (tốc độ), Variety (đa dạng), Veracity (độ tin cậy) và Value (giá trị). Mô hình này cung cấp một khung lý thuyết tổng quát để phân tích các đặc trưng của dữ liệu lớn, đồng thời giúp làm rõ các yêu cầu kỹ thuật cần thiết đối với hệ thống xử lý dữ liệu.

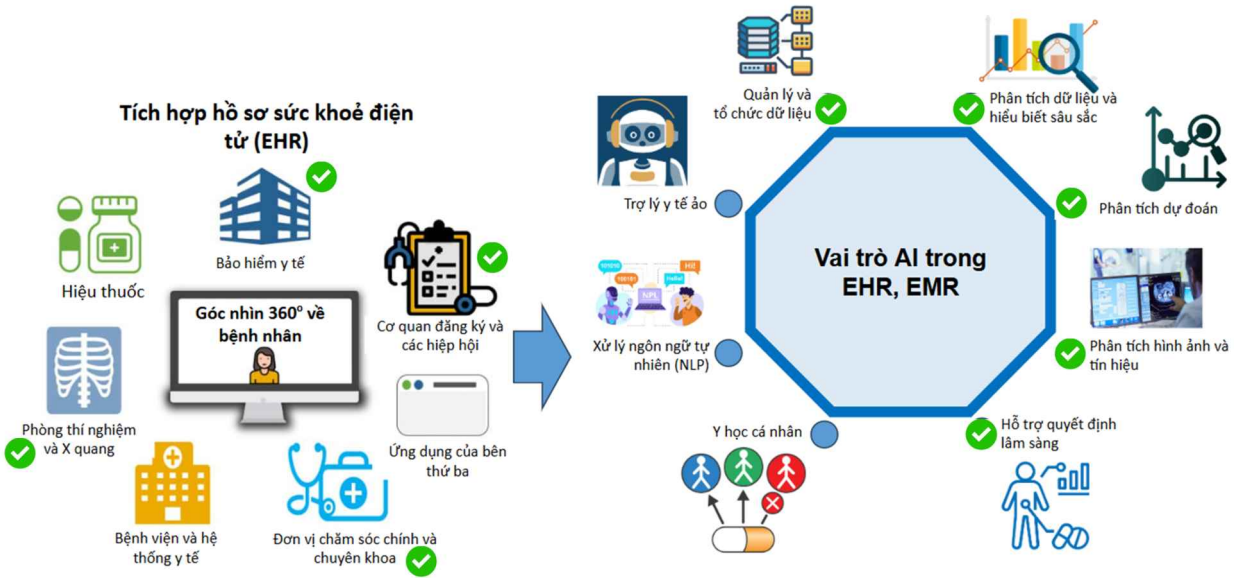
Xét về khối lượng dữ liệu (Volume), lĩnh vực y tế tạo ra một lượng dữ liệu khổng lồ và liên tục tăng theo thời gian. Các bệnh viện và cơ sở y tế hiện đại lưu trữ dữ liệu từ nhiều nguồn khác nhau, bao gồm hồ sơ bệnh án điện tử (Electronic Medical Record – EMR), kết quả xét nghiệm, dữ liệu hình ảnh y tế và dữ liệu sinh hiệu bệnh nhân. Đặc biệt, dữ liệu hình ảnh y tế như CT, MRI và X-ray có dung lượng rất lớn, với mỗi ca chụp có thể lên đến hàng trăm megabyte hoặc hơn. Khi tích lũy trong thời gian dài, khối lượng dữ liệu này nhanh chóng vượt quá khả năng lưu trữ và xử lý của các hệ thống truyền thống.

Về tốc độ phát sinh dữ liệu (Velocity), dữ liệu y tế không chỉ được tạo ra theo từng đợt mà còn phát sinh liên tục với tốc độ cao, nhất là trong môi trường chăm sóc tích cực (ICU). Các thiết bị theo dõi sinh hiệu như điện tâm đồ (ECG), máy đo nhịp tim, máy đo huyết áp hay cảm biến nồng độ oxy trong máu (SpO2) liên tục gửi dữ liệu theo thời gian thực. Dòng dữ liệu này yêu cầu hệ thống phải có khả năng tiếp nhận và xử lý gần như tức thời để phục vụ các ứng dụng giám sát, cảnh báo sớm và hỗ trợ ra quyết định lâm sàng.

Tính đa dạng của dữ liệu (Variety) là một trong những đặc trưng nổi bật nhất của dữ liệu y tế. Dữ liệu y tế bao gồm dữ liệu có cấu trúc như thông tin hành chính bệnh nhân, kết quả xét nghiệm; dữ liệu bán cấu trúc như các file hình ảnh y tế theo chuẩn DICOM; và dữ liệu phi cấu trúc như tín hiệu sinh học, hình ảnh, video, ghi chú lâm sàng và văn bản mô tả của bác sĩ. Sự đa dạng về định dạng và cấu trúc dữ liệu khiến việc

tích hợp và xử lý dữ liệu trở nên phức tạp, đòi hỏi các hệ thống dữ liệu phải có tính linh hoạt cao và hỗ trợ nhiều mô hình xử lý khác nhau.

Độ tin cậy của dữ liệu (Veracity) là yếu tố đặc biệt quan trọng trong lĩnh vực y tế, nơi các quyết định dựa trên dữ liệu có thể ảnh hưởng trực tiếp đến sức khỏe và tính mạng con người. Dữ liệu y tế có thể bị ảnh hưởng bởi nhiều yếu tố như nhiễu tín hiệu từ cảm biến, lỗi thiết bị, gián đoạn kết nối mạng hoặc sai sót trong quá trình nhập liệu. Do đó, việc đảm bảo chất lượng dữ liệu, tính toàn vẹn và khả năng truy vết nguồn gốc dữ liệu là yêu cầu bắt buộc đối với các hệ thống xử lý dữ liệu y tế.



Hình 2. Dữ liệu lớn trong lĩnh vực Y tế

Cuối cùng, giá trị của dữ liệu (Value) thể hiện ở khả năng khai thác dữ liệu để hỗ trợ nâng cao chất lượng chăm sóc sức khỏe, tối ưu hóa quy trình điều trị và thúc đẩy nghiên cứu y sinh. Tuy nhiên, dữ liệu chỉ thực sự mang lại giá trị khi được thu thập đầy đủ, lưu trữ an toàn và xử lý hiệu quả trên một nền tảng hạ tầng dữ liệu phù hợp. Nếu thiếu một kiến trúc dữ liệu hợp lý, dữ liệu y tế lớn có nguy cơ trở thành gánh nặng lưu trữ thay vì nguồn tài nguyên có giá trị.

Từ góc độ xử lý, dữ liệu lớn trong y tế có thể được chia thành hai nhóm chính dựa trên phương thức phát sinh và khai thác. Nhóm thứ nhất là dữ liệu thời gian thực (streaming data), điển hình là dữ liệu sinh hiệu bệnh nhân được thu thập liên tục từ các thiết bị IoT y tế trong ICU. Nhóm dữ liệu này đòi hỏi hệ thống có khả năng xử lý dòng dữ liệu liên tục với độ trễ thấp và độ tin cậy cao. Nhóm thứ hai là dữ liệu xử lý theo lô (batch data), bao gồm dữ liệu hình ảnh y tế, hồ sơ bệnh án và dữ liệu lịch sử điều trị,

thường được xử lý sau khi thu thập và phục vụ cho các bài toán phân tích chuyên sâu hoặc nghiên cứu dài hạn.

Những đặc thù nêu trên cho thấy dữ liệu lớn trong lĩnh vực y tế không chỉ đặt ra thách thức về lưu trữ và xử lý, mà còn liên quan mật thiết đến các yêu cầu về bảo mật, quyền riêng tư và tuân thủ quy định pháp lý. Đây chính là cơ sở lý thuyết quan trọng để nghiên cứu và đề xuất các kiến trúc dữ liệu hiện đại như Data Lakehouse trong môi trường đám mây lai, nhằm đáp ứng đồng thời các yêu cầu về hiệu năng, khả năng mở rộng và an toàn dữ liệu y tế. Nội dung này sẽ được làm rõ hơn trong các mục tiếp theo, bắt đầu với dữ liệu thời gian thực từ điện tâm đồ (ECG) trong môi trường chăm sóc tích cực.

3.2. Dữ liệu thời gian thực trong y tế – ECG 500Hz

Trong môi trường chăm sóc tích cực (Intensive Care Unit – ICU), việc giám sát liên tục các chỉ số sinh tồn của bệnh nhân đóng vai trò then chốt trong hỗ trợ ra quyết định lâm sàng và can thiệp kịp thời. Các hệ thống theo dõi sinh hiệu hiện đại được trang bị nhiều cảm biến khác nhau, trong đó dữ liệu điện tâm đồ (Electrocardiogram – ECG) là một trong những nguồn dữ liệu quan trọng và có tính đặc thù cao. Dữ liệu ECG phản ánh trực tiếp hoạt động điện của tim, cho phép theo dõi nhịp tim, phát hiện sớm các rối loạn nhịp và đánh giá tình trạng tim mạch của bệnh nhân theo thời gian thực.

Dữ liệu ECG trong ICU thường được thu thập với tần số lấy mẫu cao, phổ biến ở mức 250Hz đến 500Hz, thậm chí cao hơn đối với một số thiết bị chuyên dụng. Với tần số 500Hz, mỗi giây hệ thống tạo ra 500 mẫu tín hiệu cho mỗi kênh đo, dẫn đến khối lượng dữ liệu phát sinh rất lớn khi theo dõi liên tục nhiều bệnh nhân trong thời gian dài. Điều này khiến dữ liệu ECG trở thành một dạng dữ liệu thời gian thực điển hình trong lĩnh vực y tế, đồng thời đặt ra các yêu cầu nghiêm ngặt đối với hạ tầng xử lý dữ liệu.

Từ góc độ đặc trưng dữ liệu, dữ liệu ECG mang đầy đủ các đặc điểm của dữ liệu lớn trong y tế. Về khối lượng, dữ liệu được tạo ra liên tục và tích lũy nhanh chóng theo thời gian. Về tốc độ, dữ liệu phát sinh với tần số cao và cần được xử lý gần như tức thời để đảm bảo giá trị sử dụng. Về đa dạng, dữ liệu ECG có thể được kết hợp với các tín hiệu sinh học khác như nhịp thở, huyết áp, SpO2, cũng như dữ liệu bệnh án và thông tin lâm sàng. Về độ tin cậy, tín hiệu ECG có thể bị nhiễu do chuyển động của bệnh nhân, lỗi tiếp xúc điện cực hoặc nhiễu điện từ, đòi hỏi hệ thống phải có khả năng xử lý dữ liệu không hoàn hảo.

Dữ liệu ECG thời gian thực có vai trò quan trọng trong nhiều kịch bản ứng dụng lâm sàng. Trước hết, dữ liệu này được sử dụng để giám sát liên tục tình trạng tim mạch của bệnh nhân, cho phép bác sĩ và nhân viên y tế theo dõi diễn biến sinh lý theo thời gian thực. Thứ hai, dữ liệu ECG là cơ sở để phát hiện sớm các bất thường như loạn

nhịp tim, nhịp tim quá nhanh hoặc quá chậm, rung nhĩ hay nguy cơ ngưng tim. Trong các hệ thống hiện đại, dữ liệu ECG còn được khai thác để xây dựng các mô hình dự báo và cảnh báo sớm, hỗ trợ can thiệp kịp thời và giảm thiểu rủi ro cho bệnh nhân.

Việc xử lý dữ liệu ECG 500Hz đặt ra nhiều thách thức kỹ thuật đối với hệ thống dữ liệu. Trước hết là yêu cầu về độ trễ thấp (low latency). Dữ liệu cần được tiếp nhận, xử lý và phản hồi trong thời gian rất ngắn để đảm bảo giá trị sử dụng trong các tình huống lâm sàng khẩn cấp. Bất kỳ độ trễ lớn nào trong quá trình xử lý cũng có thể làm giảm hiệu quả của hệ thống giám sát và cảnh báo. Do đó, hạ tầng xử lý dữ liệu phải hỗ trợ cơ chế xử lý dòng dữ liệu (stream processing) với hiệu năng cao và khả năng mở rộng linh hoạt.

Bên cạnh đó, hệ thống cần đảm bảo tính sẵn sàng cao và khả năng chịu lỗi. Trong môi trường ICU, dữ liệu ECG cần được thu thập và xử lý liên tục 24/7, không cho phép mất dữ liệu hoặc gián đoạn dịch vụ. Điều này đòi hỏi hệ thống phải có cơ chế lưu trữ tạm thời, sao lưu và phục hồi khi xảy ra sự cố, đồng thời đảm bảo tính toàn vẹn của dữ liệu trong suốt quá trình xử lý.

Một thách thức khác là việc lưu trữ dữ liệu ECG trong dài hạn. Mặc dù dữ liệu thời gian thực thường được sử dụng để giám sát và cảnh báo tức thời, nhưng dữ liệu lịch sử ECG vẫn có giá trị quan trọng cho phân tích hồi cứu, nghiên cứu lâm sàng và đào tạo mô hình học máy. Do đó, hệ thống cần có khả năng lưu trữ dữ liệu ECG thô hoặc dữ liệu đã được xử lý ở các mức độ khác nhau, đồng thời hỗ trợ truy xuất hiệu quả khi cần thiết.

Từ góc độ kiến trúc dữ liệu, dữ liệu ECG 500Hz là ví dụ điển hình cho nhóm dữ liệu streaming trong kiến trúc Data Lakehouse. Dữ liệu được tiếp nhận liên tục từ các thiết bị IoT y tế, đi qua các tầng xử lý trung gian để làm sạch, chuẩn hóa và lưu trữ. Trong mô hình Medallion Architecture, dữ liệu ECG thô có thể được lưu trữ tại tầng Bronze, sau đó được xử lý và làm giàu tại tầng Silver, và cuối cùng được tổng hợp hoặc trích xuất đặc trưng tại tầng Gold để phục vụ phân tích và khai thác.

Ngoài ra, dữ liệu ECG còn đặt ra các yêu cầu nghiêm ngặt về bảo mật và quyền riêng tư. Dữ liệu sinh hiệu bệnh nhân được xem là dữ liệu nhạy cảm, cần được bảo vệ trong suốt quá trình thu thập, truyền tải, lưu trữ và xử lý. Hệ thống xử lý dữ liệu phải hỗ trợ các cơ chế kiểm soát truy cập, phân quyền người dùng và truy vết lịch sử truy cập để đáp ứng các yêu cầu về bảo mật và tuân thủ quy định pháp lý trong lĩnh vực y tế.

Dữ liệu ECG 500Hz đại diện cho một lớp dữ liệu thời gian thực có tính đặc thù cao trong lĩnh vực y tế, với yêu cầu khắt khe về hiệu năng, độ trễ, độ tin cậy và bảo mật. Việc xử lý hiệu quả loại dữ liệu này đòi hỏi một kiến trúc dữ liệu hiện đại, linh hoạt và có khả năng mở rộng. Đây chính là động lực quan trọng thúc đẩy việc nghiên cứu và áp dụng các kiến trúc Data Lakehouse trong môi trường đám mây lai, nhằm đáp ứng đồng thời các yêu cầu của xử lý dữ liệu thời gian thực và phân tích dữ liệu y tế quy mô

lớn. Nội dung tiếp theo sẽ tập trung vào nhóm dữ liệu y tế xử lý theo lô, với trường hợp điển hình là dữ liệu hình ảnh y tế như X-ray, CT và MRI.

3.3. Dữ liệu hình ảnh y tế theo lô – X-ray, CT, MRI

Bên cạnh dữ liệu thời gian thực từ các thiết bị theo dõi sinh hiệu, dữ liệu hình ảnh y tế là một trong những nguồn dữ liệu quan trọng và có khối lượng lớn nhất trong các hệ thống y tế hiện đại. Các kỹ thuật chẩn đoán hình ảnh như X-ray, chụp cắt lớp vi tính (Computed Tomography – CT) và cộng hưởng từ (Magnetic Resonance Imaging – MRI) đóng vai trò thiết yếu trong việc hỗ trợ bác sĩ phát hiện bệnh lý, đánh giá mức độ tổn thương và theo dõi hiệu quả điều trị. Dữ liệu hình ảnh y tế vì vậy không chỉ mang giá trị lâm sàng cao mà còn là nền tảng quan trọng cho các hoạt động phân tích, nghiên cứu và ứng dụng trí tuệ nhân tạo trong y tế.

Khác với dữ liệu ECG được tạo ra liên tục theo thời gian thực, dữ liệu hình ảnh y tế thường được phát sinh theo từng đợt, tương ứng với các ca chẩn đoán hoặc chỉ định lâm sàng cụ thể. Sau khi được thu thập, dữ liệu hình ảnh thường được lưu trữ và xử lý theo cơ chế xử lý theo lô (batch processing). Cách tiếp cận này phù hợp với đặc điểm của dữ liệu hình ảnh, vốn không yêu cầu phản hồi tức thời nhưng lại đòi hỏi khả năng xử lý khối lượng lớn dữ liệu với độ chính xác và độ tin cậy cao.

Dữ liệu hình ảnh y tế thường được lưu trữ dưới định dạng chuẩn DICOM (Digital Imaging and Communications in Medicine). Chuẩn DICOM không chỉ chứa dữ liệu hình ảnh mà còn bao gồm hệ thống metadata phong phú mô tả thông tin bệnh nhân, thông tin thiết bị chụp, thông số kỹ thuật và ngữ cảnh lâm sàng. Nhờ đó, dữ liệu DICOM có cấu trúc phức tạp, vừa mang đặc điểm của dữ liệu bán cấu trúc, vừa gắn chặt với các yêu cầu về quản lý, truy xuất và bảo mật dữ liệu y tế.

Xét về khối lượng dữ liệu, mỗi ca chụp X-ray, CT hoặc MRI có thể tạo ra một hoặc nhiều tập tin với dung lượng từ vài megabyte đến hàng trăm megabyte, thậm chí lên đến vài gigabyte đối với các ca chụp có độ phân giải cao hoặc chụp nhiều lát cắt. Khi được tích lũy trong thời gian dài tại các bệnh viện lớn, dữ liệu hình ảnh y tế nhanh chóng hình thành các kho dữ liệu có dung lượng rất lớn, đặt ra thách thức đáng kể đối với hạ tầng lưu trữ và quản lý dữ liệu.

Về tốc độ xử lý, dữ liệu hình ảnh y tế không yêu cầu xử lý theo thời gian thực như dữ liệu ECG, nhưng lại đòi hỏi khả năng xử lý theo lô với hiệu năng cao. Các tác vụ xử lý thường bao gồm tiền xử lý hình ảnh, trích xuất đặc trưng, chuyển đổi định dạng, cũng như phân tích hình ảnh phục vụ nghiên cứu hoặc đào tạo mô hình học sâu. Những tác vụ này thường tiêu tốn nhiều tài nguyên tính toán và yêu cầu hệ thống có khả năng mở rộng linh hoạt để đáp ứng nhu cầu xử lý trong các khoảng thời gian cao điểm.

Dữ liệu hình ảnh y tế đóng vai trò quan trọng trong nhiều kịch bản ứng dụng. Trước hết, dữ liệu này được sử dụng trực tiếp trong hoạt động chẩn đoán và điều trị, hỗ trợ bác sĩ quan sát và đánh giá tình trạng bệnh lý. Thứ hai, dữ liệu hình ảnh là nguồn dữ liệu đầu vào quan trọng cho các hệ thống hỗ trợ chẩn đoán dựa trên trí tuệ nhân tạo, cho phép phát hiện tự động các bất thường như khối u, xuất huyết, tổn thương mô hoặc dị dạng cấu trúc. Ngoài ra, dữ liệu hình ảnh còn được khai thác trong các nghiên cứu lâm sàng và đào tạo, góp phần nâng cao chất lượng chuyên môn và hiệu quả điều trị.

Việc xử lý dữ liệu hình ảnh y tế theo lô đặt ra nhiều thách thức kỹ thuật. Một trong những thách thức lớn nhất là quản lý và tổ chức dữ liệu ở quy mô lớn, bao gồm cả dữ liệu hình ảnh và metadata đi kèm. Hệ thống cần hỗ trợ lưu trữ lâu dài, đảm bảo tính toàn vẹn của dữ liệu và cho phép truy xuất nhanh chóng khi cần thiết. Đồng thời, dữ liệu hình ảnh y tế phải được bảo vệ nghiêm ngặt để đảm bảo quyền riêng tư của bệnh nhân và tuân thủ các quy định pháp lý liên quan.

Từ góc độ kiến trúc dữ liệu, dữ liệu hình ảnh y tế theo lô là một thành phần quan trọng trong kiến trúc Data Lakehouse. Trong mô hình này, dữ liệu hình ảnh thô có thể được lưu trữ tại tầng Bronze, giữ nguyên định dạng gốc để đảm bảo tính toàn vẹn và khả năng truy vết. Tại tầng Silver, dữ liệu có thể được làm sạch, chuẩn hóa và bổ sung metadata nhằm phục vụ truy vấn và phân tích. Cuối cùng, tại tầng Gold, dữ liệu hình ảnh có thể được tổng hợp, trích xuất đặc trưng hoặc kết hợp với các nguồn dữ liệu khác như dữ liệu sinh hiệu và hồ sơ bệnh án để phục vụ phân tích nâng cao.

So với dữ liệu ECG thời gian thực, dữ liệu hình ảnh y tế theo lô có đặc điểm xử lý khác biệt rõ rệt, nhưng lại bổ sung cho nhau trong hệ thống dữ liệu y tế tổng thể. Sự kết hợp giữa dữ liệu streaming và dữ liệu batch đòi hỏi một kiến trúc dữ liệu thống nhất, có khả năng hỗ trợ đồng thời nhiều mô hình xử lý và nhiều loại dữ liệu khác nhau. Đây chính là tiền đề quan trọng cho việc nghiên cứu các kiến trúc dữ liệu hiện đại như Data Lakehouse, nhằm đáp ứng toàn diện các yêu cầu xử lý dữ liệu lớn trong lĩnh vực y tế.

Tóm lại, dữ liệu hình ảnh y tế theo lô đại diện cho nhóm dữ liệu có khối lượng lớn, cấu trúc phức tạp và giá trị khai thác cao trong các hệ thống y tế hiện đại. Việc lưu trữ và xử lý hiệu quả loại dữ liệu này là một thách thức lớn đối với hạ tầng dữ liệu truyền thống, đồng thời là động lực thúc đẩy việc áp dụng các kiến trúc dữ liệu tiên tiến. Nội dung này cùng với dữ liệu thời gian thực ECG ở mục trước tạo thành nền tảng thực tiễn để đánh giá tính phù hợp của kiến trúc Data Lakehouse trong môi trường đám mây lai cho lĩnh vực y tế.

3.4. Yêu cầu hệ thống xử lý dữ liệu lớn trong y tế

Từ các phân tích ở các mục trước cho thấy dữ liệu y tế hiện đại bao gồm nhiều loại dữ liệu khác nhau, điển hình là dữ liệu thời gian thực từ điện tâm đồ (ECG) với tần số lấy mẫu cao và dữ liệu hình ảnh y tế xử lý theo lô như X-ray, CT và MRI. Mỗi loại dữ liệu có đặc trưng phát sinh, phương thức xử lý và yêu cầu khai thác khác nhau, nhưng cùng tồn tại trong một hệ sinh thái dữ liệu y tế thống nhất. Điều này đặt ra yêu cầu đối với hệ thống xử lý dữ liệu phải có khả năng đáp ứng đồng thời nhiều mục tiêu, từ hiệu năng, độ tin cậy đến bảo mật và khả năng mở rộng.

Một trong những yêu cầu quan trọng nhất đối với hệ thống xử lý dữ liệu y tế là khả năng xử lý đồng thời dữ liệu thời gian thực và dữ liệu theo lô. Dữ liệu ECG 500Hz đòi hỏi cơ chế xử lý dòng dữ liệu liên tục với độ trễ thấp, nhằm đảm bảo khả năng giám sát và cảnh báo kịp thời trong các tình huống lâm sàng. Trong khi đó, dữ liệu hình ảnh y tế lại yêu cầu khả năng xử lý theo lô với hiệu năng cao, phục vụ cho các tác vụ phân tích chuyên sâu và lưu trữ dài hạn. Do đó, hệ thống cần hỗ trợ đa mô hình xử lý (multi-processing paradigm), cho phép tích hợp cả streaming và batch trong cùng một kiến trúc thống nhất.

Yêu cầu về hiệu năng và độ trễ là yếu tố then chốt trong các hệ thống xử lý dữ liệu y tế. Đối với dữ liệu thời gian thực, độ trễ xử lý cần được kiểm soát ở mức thấp để đảm bảo giá trị sử dụng trong các ứng dụng giám sát và cảnh báo. Đối với dữ liệu xử lý theo lô, hệ thống cần có khả năng xử lý khối lượng lớn dữ liệu trong thời gian hợp lý, đồng thời tận dụng hiệu quả tài nguyên tính toán. Điều này đòi hỏi hạ tầng xử lý phải có khả năng mở rộng linh hoạt, cho phép phân bổ tài nguyên động theo nhu cầu thực tế.

Khả năng mở rộng (scalability) là một yêu cầu quan trọng khác của hệ thống xử lý dữ liệu lớn trong y tế. Khối lượng dữ liệu y tế có xu hướng tăng nhanh theo thời gian, cả về số lượng bệnh nhân, số lượng thiết bị theo dõi và độ phân giải của các thiết bị chẩn đoán hình ảnh. Hệ thống dữ liệu cần được thiết kế theo hướng mở rộng ngang, cho phép bổ sung tài nguyên lưu trữ và tính toán mà không làm gián đoạn hoạt động hiện tại. Đồng thời, hệ thống cần hỗ trợ mở rộng linh hoạt giữa môi trường on-premise và đám mây để tận dụng hiệu quả tài nguyên sẵn có.

Bên cạnh hiệu năng và khả năng mở rộng, tính sẵn sàng cao và khả năng chịu lỗi là những yêu cầu không thể thiếu trong hệ thống xử lý dữ liệu y tế. Trong môi trường bệnh viện, dữ liệu cần được thu thập và xử lý liên tục 24/7, không cho phép mất dữ liệu hoặc gián đoạn dịch vụ. Hệ thống cần có các cơ chế dự phòng, sao lưu và phục hồi dữ liệu, cũng như khả năng tự động xử lý khi xảy ra lỗi phần cứng hoặc phần mềm. Việc đảm bảo tính toàn vẹn và nhất quán của dữ liệu trong suốt quá trình xử lý là yếu tố then chốt để duy trì độ tin cậy của hệ thống.

Yêu cầu về quản lý và chất lượng dữ liệu cũng đóng vai trò quan trọng trong hệ thống xử lý dữ liệu y tế. Dữ liệu y tế thường đến từ nhiều nguồn khác nhau, với mức độ hoàn chỉnh và độ tin cậy không đồng đều. Hệ thống cần hỗ trợ các cơ chế làm sạch, chuẩn hóa và kiểm soát chất lượng dữ liệu nhằm đảm bảo dữ liệu đầu vào đáp ứng các tiêu chí cần thiết cho phân tích và khai thác. Đồng thời, khả năng truy vết nguồn gốc dữ liệu (data lineage) và quản lý metadata là yếu tố quan trọng để đảm bảo tính minh bạch và khả năng kiểm toán trong hệ thống dữ liệu y tế.

Bảo mật và quyền riêng tư là những yêu cầu đặc thù và mang tính bắt buộc trong lĩnh vực y tế. Dữ liệu sinh hiệu, hình ảnh y tế và hồ sơ bệnh án đều được xem là dữ liệu nhạy cảm, cần được bảo vệ nghiêm ngặt trong suốt vòng đời dữ liệu. Hệ thống xử lý dữ liệu cần hỗ trợ các cơ chế kiểm soát truy cập, phân quyền chi tiết theo vai trò người dùng, mã hóa dữ liệu và ghi nhận lịch sử truy cập. Những yêu cầu này không chỉ nhằm bảo vệ dữ liệu bệnh nhân mà còn để đáp ứng các quy định pháp lý và tiêu chuẩn trong lĩnh vực y tế.

Ngoài ra, hệ thống xử lý dữ liệu y tế cần đảm bảo khả năng tích hợp và chia sẻ dữ liệu giữa các hệ thống khác nhau. Dữ liệu y tế thường được sử dụng bởi nhiều bên liên quan, bao gồm bác sĩ, nhà nghiên cứu, nhà quản lý và các hệ thống phân tích tự động. Do đó, hệ thống cần hỗ trợ các giao diện truy xuất dữ liệu linh hoạt, cho phép khai thác dữ liệu phục vụ nhiều mục đích khác nhau mà vẫn đảm bảo an toàn và kiểm soát chặt chẽ.

Tổng hợp các yêu cầu trên cho thấy hệ thống xử lý dữ liệu lớn trong y tế cần một kiến trúc dữ liệu hiện đại, linh hoạt và có khả năng mở rộng, đồng thời đảm bảo hiệu năng, độ tin cậy và bảo mật. Các kiến trúc truyền thống như kho dữ liệu (Data Warehouse) hoặc hồ dữ liệu (Data Lake) riêng lẻ khó có thể đáp ứng đầy đủ các yêu cầu này. Đây chính là cơ sở lý thuyết quan trọng để nghiên cứu và đánh giá các kiến trúc dữ liệu mới, trong đó kiến trúc Data Lakehouse được xem là một hướng tiếp cận tiềm năng nhằm giải quyết các thách thức của hệ thống xử lý dữ liệu lớn trong lĩnh vực y tế. Nội dung này sẽ được phân tích chi tiết hơn trong các mục tiếp theo.

3.5. Kho dữ liệu truyền thống (Data Warehouse – DW)

Kho dữ liệu (Data Warehouse – DW) là một trong những kiến trúc lưu trữ và xử lý dữ liệu được sử dụng rộng rãi trong các hệ thống thông tin doanh nghiệp từ nhiều thập kỷ qua. Mục tiêu chính của kho dữ liệu là tích hợp dữ liệu từ nhiều hệ thống nguồn khác nhau, tổ chức dữ liệu theo cấu trúc nhất quán và tối ưu cho các truy vấn phân tích, báo cáo và hỗ trợ ra quyết định. Trong lĩnh vực y tế, kho dữ liệu đã được ứng

dụng trong nhiều hệ thống quản lý bệnh viện, hệ thống báo cáo thống kê và các nền tảng hỗ trợ quản lý hoạt động y tế.

Về mặt kiến trúc, kho dữ liệu truyền thống thường được xây dựng trên các hệ quản trị cơ sở dữ liệu quan hệ (Relational Database Management System – RDBMS) hoặc các hệ quản trị cơ sở dữ liệu phân tích chuyên dụng. Dữ liệu trong kho dữ liệu được tổ chức theo mô hình có cấu trúc chặt chẽ, phổ biến nhất là mô hình lược đồ sao (star schema) hoặc lược đồ bông tuyết (snowflake schema). Trong các mô hình này, dữ liệu được phân chia thành bảng ғᴀᴋᴛ (fact table) và các bảng chiều (dimension table), cho phép tối ưu hóa các truy vấn tổng hợp và phân tích lịch sử.

Kho dữ liệu thường sử dụng phương thức xử lý dữ liệu theo lô, trong đó dữ liệu được trích xuất từ các hệ thống nguồn, biến đổi và nạp vào kho dữ liệu thông qua các quy trình ETL (Extract – Transform – Load). Cách tiếp cận này phù hợp với các hệ thống nghiệp vụ truyền thống, nơi dữ liệu phát sinh theo từng đợt và không yêu cầu xử lý thời gian thực. Trong bối cảnh y tế, kho dữ liệu thường được sử dụng để tổng hợp dữ liệu bệnh nhân, thống kê hoạt động khám chữa bệnh, phân tích chi phí và hỗ trợ các báo cáo quản lý định kỳ.

Một trong những ưu điểm lớn của kho dữ liệu là khả năng cung cấp dữ liệu có cấu trúc, nhất quán và độ tin cậy cao. Dữ liệu trong kho dữ liệu đã được làm sạch, chuẩn hóa và kiểm soát chất lượng trước khi đưa vào sử dụng, giúp đảm bảo tính chính xác và nhất quán cho các hoạt động phân tích. Ngoài ra, kho dữ liệu hỗ trợ tốt cho các công cụ Business Intelligence (BI) và các truy vấn phân tích phức tạp, đáp ứng hiệu quả các nhu cầu báo cáo và phân tích lịch sử.

Tuy nhiên, trong bối cảnh dữ liệu lớn và đặc thù dữ liệu y tế hiện đại, kho dữ liệu truyền thống bộc lộ nhiều hạn chế. Trước hết, kho dữ liệu được thiết kế chủ yếu để xử lý dữ liệu có cấu trúc, trong khi dữ liệu y tế ngày càng đa dạng và bao gồm nhiều dạng dữ liệu bán cấu trúc và phi cấu trúc như tín hiệu sinh học, hình ảnh y tế và văn bản lâm sàng. Việc tích hợp các loại dữ liệu này vào kho dữ liệu truyền thống thường gặp nhiều khó khăn và yêu cầu các bước chuyển đổi phức tạp.

Bên cạnh đó, kho dữ liệu không đáp ứng tốt các yêu cầu xử lý dữ liệu thời gian thực. Các hệ thống kho dữ liệu truyền thống thường có độ trễ cao do phụ thuộc vào quy trình ETL theo lô, không phù hợp với các ứng dụng giám sát và cảnh báo tức thời trong môi trường ICU, nơi dữ liệu ECG và các chỉ số sinh hiệu cần được xử lý gần như ngay lập tức. Điều này hạn chế khả năng ứng dụng kho dữ liệu trong các kịch bản yêu cầu phản hồi nhanh và xử lý dòng dữ liệu liên tục.

Một hạn chế khác của kho dữ liệu là khả năng mở rộng và chi phí đầu tư hạ tầng. Các hệ thống kho dữ liệu truyền thống thường yêu cầu phần cứng chuyên dụng

và chi phí đầu tư lớn để mở rộng dung lượng lưu trữ và năng lực xử lý. Trong bối cảnh dữ liệu y tế tăng trưởng nhanh về khối lượng và độ phức tạp, việc mở rộng kho dữ liệu theo cách truyền thống trở nên kém linh hoạt và tốn kém.

Ngoài ra, kho dữ liệu thường yêu cầu thiết kế lược đồ dữ liệu chặt chẽ ngay từ đầu (schema-on-write), khiến việc thay đổi cấu trúc dữ liệu hoặc tích hợp các nguồn dữ liệu mới trở nên phức tạp. Trong lĩnh vực y tế, nơi các tiêu chuẩn dữ liệu, thiết bị và quy trình nghiệp vụ thường xuyên thay đổi, yêu cầu này làm giảm tính linh hoạt của hệ thống dữ liệu và kéo dài thời gian triển khai.

Từ góc độ tổng thể, kho dữ liệu truyền thống vẫn đóng vai trò quan trọng trong các hệ thống phân tích và báo cáo, đặc biệt đối với dữ liệu có cấu trúc và các nghiệp vụ ổn định. Tuy nhiên, trước các yêu cầu ngày càng cao về xử lý dữ liệu lớn, dữ liệu thời gian thực và dữ liệu phi cấu trúc trong lĩnh vực y tế, kiến trúc kho dữ liệu bộc lộ nhiều hạn chế và không còn đáp ứng đầy đủ nhu cầu của các hệ thống hiện đại. Những hạn chế này là một trong những nguyên nhân thúc đẩy sự ra đời của các kiến trúc dữ liệu mới, trong đó mô hình hồ dữ liệu (Data Lake) và sau đó là kiến trúc Data Lakehouse được nghiên cứu và phát triển nhằm khắc phục những điểm yếu của kho dữ liệu truyền thống.

3.6. Hồ dữ liệu (Data Lake – DL)

Trước những hạn chế của kho dữ liệu truyền thống trong việc xử lý dữ liệu lớn, dữ liệu phi cấu trúc và dữ liệu thời gian thực, mô hình hồ dữ liệu (Data Lake – DL) đã được đề xuất như một hướng tiếp cận mới cho lưu trữ và khai thác dữ liệu ở quy mô lớn. Khác với kho dữ liệu, hồ dữ liệu được thiết kế để lưu trữ toàn bộ dữ liệu thô từ nhiều nguồn khác nhau, cho phép dữ liệu được giữ nguyên ở định dạng gốc và được xử lý khi cần thiết. Cách tiếp cận này đặc biệt phù hợp với bối cảnh dữ liệu y tế hiện đại, nơi dữ liệu phát sinh đa dạng về định dạng và cấu trúc.

Về mặt khái niệm, hồ dữ liệu là một kiến trúc lưu trữ tập trung, cho phép tiếp nhận và lưu trữ dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc với chi phí thấp. Dữ liệu trong hồ dữ liệu thường được lưu trữ trên các hệ thống lưu trữ phân tán hoặc lưu trữ đối tượng (object storage), hỗ trợ khả năng mở rộng linh hoạt và truy cập trực tiếp bởi nhiều hệ thống xử lý khác nhau. Trong lĩnh vực y tế, hồ dữ liệu có thể lưu trữ đồng thời dữ liệu ECG thời gian thực, dữ liệu hình ảnh y tế DICOM, hồ sơ bệnh án và các nguồn dữ liệu liên quan khác.

Một trong những ưu điểm lớn của hồ dữ liệu là khả năng lưu trữ dữ liệu thô theo cơ chế schema-on-read. Điều này cho phép dữ liệu được đưa vào hệ thống mà không cần xác định trước cấu trúc chi tiết, giúp rút ngắn thời gian tích hợp dữ liệu và tăng tính linh hoạt khi bổ sung các nguồn dữ liệu mới. Đối với dữ liệu y tế, nơi các thiết bị, tiêu

chuẩn và định dạng dữ liệu thường xuyên thay đổi, khả năng này giúp hệ thống dễ dàng thích ứng với sự phát triển của công nghệ và nhu cầu nghiệp vụ.

Hồ dữ liệu cũng hỗ trợ tốt cho việc xử lý dữ liệu lớn và dữ liệu phi cấu trúc. Các hệ thống xử lý phân tán có thể truy cập trực tiếp dữ liệu trong hồ dữ liệu để thực hiện các tác vụ phân tích, xử lý theo lô hoặc xử lý thời gian thực. Điều này tạo điều kiện thuận lợi cho các ứng dụng phân tích nâng cao, khoa học dữ liệu và học máy, vốn yêu cầu truy cập trực tiếp vào dữ liệu thô để trích xuất thông tin và xây dựng mô hình.

Trong bối cảnh y tế, hồ dữ liệu cho phép lưu trữ lâu dài dữ liệu sinh hiệu và dữ liệu hình ảnh y tế với chi phí thấp, đồng thời hỗ trợ khai thác dữ liệu cho các nghiên cứu lâm sàng và đào tạo mô hình trí tuệ nhân tạo. Việc lưu trữ dữ liệu ở định dạng gốc cũng giúp đảm bảo tính toàn vẹn của dữ liệu và khả năng truy vết nguồn gốc, phục vụ các yêu cầu kiểm toán và nghiên cứu hồi cứu.

Tuy nhiên, bên cạnh các ưu điểm nổi bật, mô hình hồ dữ liệu cũng tồn tại nhiều hạn chế, đặc biệt khi được áp dụng trong các hệ thống y tế quy mô lớn. Một trong những vấn đề phổ biến nhất là nguy cơ hình thành “data swamp”, tức là hồ dữ liệu trở nên hỗn độn, khó quản lý và khó khai thác do thiếu các cơ chế quản trị dữ liệu chặt chẽ. Khi dữ liệu được lưu trữ ở dạng thô mà không có quy chuẩn rõ ràng về metadata, chất lượng dữ liệu và phân quyền truy cập, việc tìm kiếm và sử dụng dữ liệu trở nên kém hiệu quả.

Hồ dữ liệu truyền thống cũng thiếu các cơ chế đảm bảo tính nhất quán và toàn vẹn của dữ liệu. Không giống như kho dữ liệu, hồ dữ liệu thường không hỗ trợ các giao dịch ACID, dẫn đến nguy cơ dữ liệu không nhất quán khi xảy ra lỗi trong quá trình ghi hoặc cập nhật dữ liệu. Trong lĩnh vực y tế, nơi dữ liệu cần độ tin cậy cao và khả năng truy vết, hạn chế này có thể ảnh hưởng đến chất lượng phân tích và mức độ tin cậy của hệ thống.

Một hạn chế khác của hồ dữ liệu là hiệu năng truy vấn và khả năng tối ưu hóa xử lý. Do dữ liệu được lưu trữ ở dạng thô và phân tán, việc truy vấn dữ liệu trong hồ dữ liệu thường kém hiệu quả hơn so với kho dữ liệu truyền thống, đặc biệt đối với các truy vấn phân tích phức tạp. Điều này gây khó khăn khi hồ dữ liệu được sử dụng làm nền tảng chính cho các ứng dụng phân tích và báo cáo trong lĩnh vực y tế.

Ngoài ra, hồ dữ liệu truyền thống thường thiếu các cơ chế quản lý metadata tập trung và kiểm soát truy cập chi tiết. Trong môi trường y tế, nơi dữ liệu nhạy cảm cần được bảo vệ nghiêm ngặt, việc thiếu các cơ chế quản trị này làm tăng rủi ro về bảo mật và vi phạm quyền riêng tư. Việc tích hợp các công cụ quản lý bảo mật và phân quyền vào hồ dữ liệu thường đòi hỏi nhiều giải pháp bổ sung và làm tăng độ phức tạp của hệ thống.

Tổng hợp các phân tích trên cho thấy hồ dữ liệu là một bước tiến quan trọng so với kho dữ liệu truyền thống, đặc biệt trong việc xử lý dữ liệu lớn và dữ liệu phi cấu trúc. Tuy nhiên, hồ dữ liệu vẫn chưa đáp ứng đầy đủ các yêu cầu về quản trị, hiệu năng và độ tin cậy trong các hệ thống y tế hiện đại. Những hạn chế này đã thúc đẩy sự phát triển của các kiến trúc dữ liệu mới, nhằm kết hợp ưu điểm của kho dữ liệu và hồ dữ liệu, đồng thời khắc phục những điểm yếu của từng mô hình. Kiến trúc Data Lakehouse ra đời trong bối cảnh đó và sẽ được trình bày chi tiết trong mục tiếp theo.

3.7. Kiến trúc Data Lakehouse

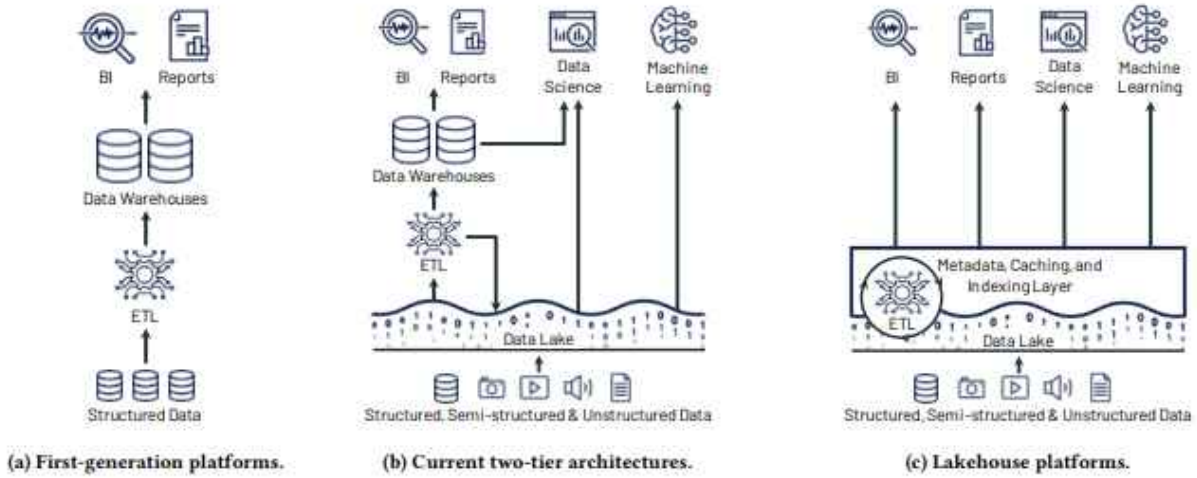
Trước những hạn chế của cả kho dữ liệu truyền thống (Data Warehouse) và hồ dữ liệu (Data Lake) trong việc đáp ứng các yêu cầu xử lý dữ liệu lớn, đa dạng và có độ tin cậy cao, kiến trúc Data Lakehouse đã được đề xuất như một hướng tiếp cận mới nhằm hợp nhất ưu điểm của hai mô hình này. Khái niệm Data Lakehouse được giới thiệu và phát triển dựa trên các nghiên cứu của Armbrust và cộng sự, trong đó kiến trúc Lakehouse được định nghĩa là một nền tảng dữ liệu thống nhất, xây dựng trên lớp lưu trữ chi phí thấp của Data Lake nhưng tích hợp trực tiếp các đặc tính quản trị và tối ưu hóa truyền thống của Data Warehouse.

Về bản chất, Data Lakehouse là một kiến trúc quản lý dữ liệu được thiết kế để hỗ trợ đồng thời nhiều loại dữ liệu và nhiều mô hình xử lý khác nhau trong cùng một hệ thống. Lakehouse cho phép lưu trữ dữ liệu ở định dạng gốc trên các hệ thống lưu trữ phân tán hoặc lưu trữ đối tượng, đồng thời cung cấp các cơ chế quản trị, đảm bảo tính nhất quán và hiệu năng truy vấn tương đương với kho dữ liệu truyền thống. Cách tiếp cận này đặc biệt phù hợp với các hệ thống dữ liệu y tế hiện đại, nơi tồn tại đồng thời dữ liệu thời gian thực, dữ liệu theo lô và dữ liệu phi cấu trúc.

Một trong những đặc điểm cốt lõi của kiến trúc Data Lakehouse là hỗ trợ các giao dịch ACID (Atomicity, Consistency, Isolation, Durability) trên lớp lưu trữ dữ liệu. Nhờ cơ chế này, Lakehouse đảm bảo tính toàn vẹn và nhất quán của dữ liệu ngay cả khi xảy ra lỗi trong quá trình ghi hoặc cập nhật. Điều này khắc phục một trong những hạn chế lớn của Data Lake truyền thống và đặc biệt quan trọng trong lĩnh vực y tế, nơi dữ liệu cần độ tin cậy cao và khả năng kiểm toán chặt chẽ.

Bên cạnh hỗ trợ giao dịch ACID, kiến trúc Data Lakehouse còn cung cấp các cơ chế phiên bản hóa dữ liệu (data versioning) và khả năng truy xuất dữ liệu theo thời điểm (time travel). Các tính năng này cho phép lưu lại lịch sử thay đổi của dữ liệu, hỗ trợ khôi phục dữ liệu về trạng thái trước đó và phục vụ các nhu cầu phân tích hồi cứu. Trong bối cảnh y tế, khả năng truy vết lịch sử dữ liệu có ý nghĩa quan trọng đối với nghiên cứu lâm sàng, kiểm toán và tuân thủ quy định pháp lý.

Một đặc điểm quan trọng khác của Lakehouse là khả năng tối ưu hóa truy vấn và hiệu năng xử lý dữ liệu. Kiến trúc này tích hợp các kỹ thuật như đánh chỉ mục, bộ nhớ đệm và tối ưu hóa truy vấn để nâng cao hiệu năng truy xuất dữ liệu. Nhờ đó, Lakehouse có thể hỗ trợ hiệu quả các khối lượng công việc phân tích phức tạp, từ báo cáo truyền thống đến các tác vụ khoa học dữ liệu và học máy, trên cùng một nền tảng dữ liệu thống nhất.



Hình 3. Kiến trúc tổng quan Data Lakehouse

Kiến trúc Data Lakehouse được xây dựng dựa trên các định dạng dữ liệu mở và tiêu chuẩn bảng mở, cho phép nhiều hệ thống xử lý khác nhau truy cập trực tiếp vào dữ liệu. Các định dạng phổ biến như Apache Parquet và ORC được sử dụng để lưu trữ dữ liệu, giúp tối ưu hóa hiệu năng đọc và giảm chi phí lưu trữ. Trên lớp định dạng này, các tiêu chuẩn bảng mở như Delta Lake, Apache Hudi và Apache Iceberg cung cấp các cơ chế quản trị dữ liệu nâng cao, đóng vai trò nền tảng cho kiến trúc Lakehouse.

So với Data Warehouse, kiến trúc Data Lakehouse mang lại tính linh hoạt cao hơn trong việc xử lý dữ liệu đa dạng và hỗ trợ tốt cho dữ liệu phi cấu trúc. So với Data Lake, Lakehouse khắc phục các vấn đề về quản trị, chất lượng dữ liệu và hiệu năng truy vấn. Nhờ đó, Lakehouse có thể được xem là một nền tảng dữ liệu thống nhất, đáp ứng đồng thời các yêu cầu lưu trữ, xử lý và khai thác dữ liệu trong các hệ thống hiện đại.

Trong bối cảnh dữ liệu y tế, kiến trúc Data Lakehouse thể hiện rõ ưu thế khi hỗ trợ đồng thời xử lý dữ liệu thời gian thực và dữ liệu theo lô. Dữ liệu ECG 500Hz có thể được tiếp nhận và xử lý theo cơ chế streaming, trong khi dữ liệu hình ảnh y tế được xử lý theo lô với hiệu năng cao. Cả hai loại dữ liệu này đều được lưu trữ và quản lý trong

cùng một nền tảng Lakehouse, giúp đơn giản hóa kiến trúc hệ thống và giảm chi phí vận hành.

Ngoài ra, Data Lakehouse tạo điều kiện thuận lợi cho việc tích hợp các công cụ phân tích nâng cao, khoa học dữ liệu và học máy. Việc truy cập trực tiếp vào dữ liệu thô và dữ liệu đã được xử lý trong cùng một nền tảng giúp rút ngắn vòng đời phân tích, từ thu thập dữ liệu đến xây dựng mô hình và triển khai ứng dụng. Điều này đặc biệt có ý nghĩa trong lĩnh vực y tế, nơi nhu cầu nghiên cứu và ứng dụng trí tuệ nhân tạo ngày càng gia tăng.

Một lợi ích quan trọng khác của kiến trúc Data Lakehouse là khả năng hỗ trợ quản trị dữ liệu và bảo mật ở mức cao. Lakehouse cho phép tích hợp các cơ chế quản lý metadata, phân quyền truy cập và truy vết dữ liệu, đáp ứng các yêu cầu nghiêm ngặt về bảo mật và quyền riêng tư trong lĩnh vực y tế. Việc quản trị tập trung giúp đảm bảo dữ liệu được sử dụng đúng mục đích và hạn chế rủi ro truy cập trái phép.

Tóm lại, kiến trúc Data Lakehouse ra đời như một câu trả lời cho những thách thức của các kiến trúc dữ liệu truyền thống trong kỷ nguyên dữ liệu lớn. Bằng cách kết hợp ưu điểm của Data Warehouse và Data Lake, Lakehouse cung cấp một nền tảng dữ liệu thống nhất, linh hoạt và có độ tin cậy cao, đặc biệt phù hợp với các hệ thống dữ liệu y tế hiện đại. Đây là nền tảng lý thuyết quan trọng cho việc thiết kế và triển khai hệ thống Data Lakehouse trong môi trường đám mây lai, sẽ được trình bày chi tiết hơn thông qua kiến trúc Medallion trong mục tiếp theo.

3.8. Kiến trúc Medallion (Bronze – Silver – Gold)

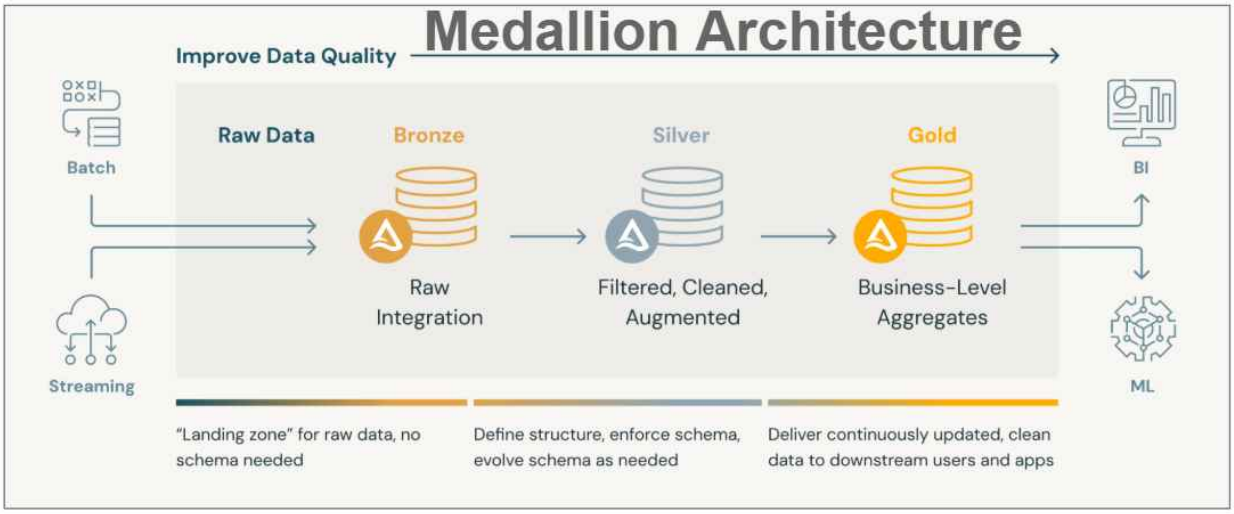
Trong kiến trúc Data Lakehouse, việc tổ chức dữ liệu một cách có hệ thống và theo từng mức độ xử lý đóng vai trò then chốt nhằm đảm bảo khả năng mở rộng, tính nhất quán và hiệu quả khai thác dữ liệu. Kiến trúc Medallion được đề xuất như một phương pháp tổ chức dữ liệu theo nhiều lớp (multi-layer data architecture), trong đó dữ liệu được phân chia thành ba lớp chính: Bronze, Silver và Gold. Mỗi lớp đại diện cho một mức độ xử lý và giá trị gia tăng khác nhau của dữ liệu, giúp hệ thống dữ liệu vận hành ổn định và dễ quản trị.

Về bản chất, kiến trúc Medallion không phải là một công nghệ cụ thể mà là một nguyên tắc thiết kế dữ liệu, được xây dựng trên nền tảng của kiến trúc Data Lakehouse. Phương pháp này cho phép lưu trữ toàn bộ vòng đời của dữ liệu, từ dữ liệu thô ban đầu đến dữ liệu đã được chuẩn hóa và dữ liệu phục vụ phân tích, trong cùng một hệ thống lưu trữ thống nhất.

3.8.1. Lớp Bronze – Dữ liệu thô

Lớp Bronze là lớp dữ liệu đầu tiên trong kiến trúc Medallion, nơi tiếp nhận và lưu trữ dữ liệu ở dạng thô ngay sau khi được thu thập từ các nguồn khác nhau. Dữ liệu tại lớp này được giữ nguyên định dạng ban đầu, chưa trải qua các bước làm sạch, chuẩn hóa hay biến đổi phức tạp. Mục tiêu chính của lớp Bronze là đảm bảo tính đầy đủ và toàn vẹn của dữ liệu gốc, đồng thời đóng vai trò như một bản sao lưu dài hạn cho toàn bộ hệ thống.

Trong bối cảnh dữ liệu y tế, lớp Bronze có thể bao gồm dữ liệu thời gian thực từ các thiết bị IoT y tế như tín hiệu ECG, dữ liệu log hệ thống, cũng như dữ liệu theo lô từ các hệ thống hình ảnh y tế ở định dạng DICOM. Việc lưu trữ dữ liệu thô giúp hệ thống có khả năng tái xử lý dữ liệu khi cần thiết, đặc biệt trong các trường hợp thay đổi yêu cầu phân tích hoặc cập nhật thuật toán xử lý.



Hình 4. Kiến trúc Medallion (Bronze – Silver – Gold)

Một đặc điểm quan trọng của lớp Bronze là khả năng tiếp nhận dữ liệu với tốc độ cao và độ trễ thấp. Do đó, các pipeline xử lý dữ liệu tại lớp này thường được thiết kế đơn giản, tập trung vào việc ghi dữ liệu một cách ổn định và có khả năng mở rộng. Các vấn đề như dữ liệu trùng lặp, dữ liệu thiếu hoặc dữ liệu không đúng định dạng có thể tồn tại ở lớp Bronze và sẽ được xử lý ở các lớp tiếp theo.

3.8.2. Lớp Silver – Dữ liệu đã làm sạch và chuẩn hóa

Lớp Silver là lớp trung gian trong kiến trúc Medallion, nơi dữ liệu từ lớp Bronze được xử lý, làm sạch và chuẩn hóa nhằm nâng cao chất lượng dữ liệu. Tại lớp này, các thao tác như loại bỏ dữ liệu trùng lặp, xử lý giá trị thiếu, chuẩn hóa kiểu dữ liệu và áp dụng các quy tắc kiểm tra chất lượng dữ liệu được thực hiện một cách có hệ thống.

Trong hệ thống dữ liệu y tế, lớp Silver đóng vai trò đặc biệt quan trọng do dữ liệu đầu vào thường có độ phức tạp cao và yêu cầu độ chính xác lớn. Đối với dữ liệu ECG, các bước xử lý tại lớp Silver có thể bao gồm việc đồng bộ hóa thời gian, kiểm tra tần số lấy mẫu, loại bỏ các giá trị nhiễu hoặc bất thường. Đối với dữ liệu hình ảnh y tế, lớp Silver có thể thực hiện việc trích xuất metadata từ các file DICOM, chuẩn hóa thông tin bệnh nhân và phân loại dữ liệu theo từng loại hình ảnh.

Lớp Silver thường được tổ chức theo các bảng dữ liệu có cấu trúc rõ ràng, phù hợp cho việc truy vấn và phân tích cơ bản. Việc chuẩn hóa dữ liệu tại lớp này giúp giảm thiểu rủi ro sai lệch trong quá trình phân tích và tạo nền tảng vững chắc cho các tác vụ khai thác dữ liệu nâng cao ở lớp Gold.

3.8.3. Lớp Gold – Dữ liệu phục vụ phân tích và khai thác

Lớp Gold là lớp dữ liệu cuối cùng trong kiến trúc Medallion, nơi dữ liệu đã được tổng hợp, biến đổi và tối ưu hóa để phục vụ trực tiếp cho các nhu cầu phân tích, báo cáo và nghiên cứu. Dữ liệu tại lớp Gold thường có cấu trúc phù hợp với các mô hình phân tích, chẳng hạn như mô hình dữ liệu dạng sao (star schema) hoặc các bảng tổng hợp chuyên biệt cho từng bài toán.

Trong lĩnh vực y tế, lớp Gold có thể bao gồm các tập dữ liệu tổng hợp phục vụ phân tích xu hướng sức khỏe, nghiên cứu lâm sàng hoặc huấn luyện các mô hình học máy. Ví dụ, dữ liệu ECG sau khi được xử lý và chuẩn hóa ở lớp Silver có thể được tổng hợp theo từng bệnh nhân, từng khoảng thời gian hoặc từng chỉ số y sinh để phục vụ các bài toán phân tích nâng cao.

Một đặc điểm quan trọng của lớp Gold là hiệu năng truy vấn cao và độ ổn định dữ liệu. Do đó, các quy trình xử lý dữ liệu tại lớp này thường được kiểm soát chặt chẽ và chỉ cập nhật khi dữ liệu đã đạt yêu cầu về chất lượng. Điều này giúp đảm bảo rằng các kết quả phân tích và báo cáo được xây dựng trên nền dữ liệu đáng tin cậy.

3.8.4. Vai trò của kiến trúc Medallion trong hệ thống Data Lakehouse y tế

Kiến trúc Medallion đóng vai trò như một khung tổ chức dữ liệu hiệu quả trong kiến trúc Data Lakehouse, giúp phân tách rõ ràng các giai đoạn xử lý dữ liệu và nâng cao khả năng quản trị hệ thống. Việc phân chia dữ liệu theo các lớp Bronze, Silver và Gold giúp hệ thống dễ dàng mở rộng, bảo trì và kiểm soát chất lượng dữ liệu.

Trong bối cảnh hệ thống dữ liệu y tế, kiến trúc Medallion cho phép xử lý linh hoạt cả dữ liệu thời gian thực và dữ liệu theo lô trong cùng một nền tảng. Các pipeline xử lý dữ liệu có thể được thiết kế phù hợp với từng lớp, đảm bảo hiệu năng và độ tin cậy của

hệ thống. Đồng thời, việc lưu trữ dữ liệu theo từng lớp giúp đáp ứng các yêu cầu về kiểm toán, truy vết và tuân thủ quy định pháp lý trong lĩnh vực y tế.

Tóm lại, kiến trúc Medallion là một thành phần quan trọng trong việc hiện thực hóa kiến trúc Data Lakehouse. Bằng cách tổ chức dữ liệu theo các lớp có ý nghĩa rõ ràng, Medallion giúp hệ thống dữ liệu y tế vận hành hiệu quả, linh hoạt và bền vững, tạo nền tảng cho việc triển khai các pipeline xử lý dữ liệu và kiến trúc đám mây lai được trình bày trong các chương tiếp theo của luận văn.

3.9. Mô hình đám mây lai (Hybrid Cloud) cho hệ thống dữ liệu y tế

Trong bối cảnh dữ liệu y tế ngày càng gia tăng cả về quy mô, tốc độ phát sinh và mức độ đa dạng, việc lựa chọn mô hình hạ tầng phù hợp đóng vai trò then chốt đối với hiệu quả vận hành và khả năng mở rộng của hệ thống dữ liệu. Hai hướng tiếp cận phổ biến hiện nay là triển khai hệ thống hoàn toàn tại chỗ (on-premise) hoặc sử dụng nền tảng đám mây công cộng (public cloud). Tuy nhiên, mỗi mô hình đều tồn tại những ưu điểm và hạn chế nhất định khi áp dụng cho lĩnh vực y tế. Do đó, mô hình đám mây lai (Hybrid Cloud) được xem là giải pháp trung hòa, kết hợp lợi thế của cả hai môi trường để đáp ứng các yêu cầu đặc thù của hệ thống dữ liệu y tế.

3.9.1. So sánh mô hình on-premise và mô hình đám mây

Hệ thống on-premise truyền thống cho phép tổ chức y tế toàn quyền kiểm soát hạ tầng, dữ liệu và các chính sách bảo mật. Dữ liệu được lưu trữ và xử lý trong phạm vi nội bộ, giúp giảm thiểu rủi ro rò rỉ thông tin và đáp ứng tốt các yêu cầu pháp lý liên quan đến quyền riêng tư bệnh nhân. Mô hình này đặc biệt phù hợp với các dữ liệu nhạy cảm như hồ sơ bệnh án, thông tin định danh bệnh nhân và dữ liệu lâm sàng thời gian thực.

Tuy nhiên, hệ thống on-premise gặp nhiều hạn chế về khả năng mở rộng và chi phí đầu tư. Việc mở rộng hạ tầng yêu cầu đầu tư phần cứng, bảo trì và nhân sự vận hành, dẫn đến chi phí cao và thời gian triển khai dài. Ngoài ra, năng lực xử lý của hệ thống on-premise thường khó đáp ứng các bài toán phân tích dữ liệu lớn, học máy và trí tuệ nhân tạo ở quy mô lớn.

Ngược lại, nền tảng đám mây công cộng cung cấp tài nguyên tính toán và lưu trữ linh hoạt, cho phép mở rộng gần như không giới hạn theo nhu cầu sử dụng. Các dịch vụ đám mây hỗ trợ tốt cho xử lý dữ liệu lớn, phân tích nâng cao và huấn luyện mô hình học máy. Tuy nhiên, việc đưa toàn bộ dữ liệu y tế lên đám mây công cộng làm gia tăng lo ngại về bảo mật, quyền riêng tư và tuân thủ pháp lý, đặc biệt đối với dữ liệu nhạy cảm và dữ liệu thời gian thực.

3.9.2. Lý do lĩnh vực y tế cần mô hình đám mây lai

Lĩnh vực y tế có những yêu cầu đặc thù mà các mô hình hạ tầng đơn lẻ khó có thể đáp ứng đầy đủ. Thứ nhất, dữ liệu y tế mang tính nhạy cảm cao, đòi hỏi các cơ chế bảo mật nghiêm ngặt và khả năng kiểm soát chặt chẽ. Thứ hai, khối lượng dữ liệu y tế ngày càng lớn, bao gồm cả dữ liệu thời gian thực (như ECG) và dữ liệu theo lô (như hình ảnh y tế), yêu cầu hệ thống có khả năng mở rộng và xử lý linh hoạt. Thứ ba, các ứng dụng phân tích nâng cao và học máy trong y tế đòi hỏi tài nguyên tính toán mạnh mẽ mà hệ thống on-premise truyền thống khó đáp ứng.

Mô hình đám mây lai cho phép kết hợp ưu điểm của cả hai môi trường. Dữ liệu nhạy cảm và các hệ thống lõi có thể được lưu trữ và xử lý tại môi trường on-premise để đảm bảo an toàn và tuân thủ quy định. Trong khi đó, các tác vụ xử lý dữ liệu lớn, phân tích nâng cao và lưu trữ mở rộng có thể được triển khai trên nền tảng đám mây, tận dụng khả năng mở rộng và hiệu năng cao của cloud. Cách tiếp cận này đặc biệt phù hợp với hệ thống dữ liệu y tế hiện đại, nơi yêu cầu vừa bảo mật cao vừa khai thác hiệu quả giá trị của dữ liệu.

3.9.3. Phân vai trò xử lý giữa on-premise và cloud

Trong kiến trúc đám mây lai cho hệ thống dữ liệu y tế, việc phân chia vai trò xử lý giữa môi trường on-premise và cloud là yếu tố then chốt nhằm tối ưu hiệu năng, chi phí và bảo mật. Môi trường on-premise thường đảm nhận các tác vụ xử lý dữ liệu thời gian thực với yêu cầu độ trễ thấp, chẳng hạn như thu thập và xử lý ban đầu dữ liệu ECG từ các thiết bị IoT y tế. Việc xử lý tại chỗ giúp đảm bảo khả năng phản hồi nhanh, phục vụ cảnh báo và hỗ trợ quyết định lâm sàng kịp thời.

Ngoài ra, on-premise cũng đóng vai trò lưu trữ và quản lý các dữ liệu nhạy cảm, bao gồm thông tin định danh bệnh nhân và dữ liệu lâm sàng cốt lõi. Điều này giúp tổ chức y tế duy trì quyền kiểm soát dữ liệu và giảm thiểu rủi ro liên quan đến bảo mật.

Môi trường đám mây, ngược lại, phù hợp cho các tác vụ xử lý theo lô, phân tích dữ liệu lớn và học máy. Các dữ liệu đã được làm sạch và chuẩn hóa từ hệ thống on-premise có thể được đồng bộ lên cloud để phục vụ các bài toán phân tích nâng cao, huấn luyện mô hình trí tuệ nhân tạo và lưu trữ dài hạn. Việc triển khai các lớp xử lý này trên cloud giúp hệ thống tận dụng tài nguyên tính toán linh hoạt và giảm chi phí đầu tư hạ tầng ban đầu.

3.9.4. Vai trò của mô hình đám mây lai trong kiến trúc Data Lakehouse

Mô hình đám mây lai tạo nền tảng hạ tầng phù hợp để triển khai kiến trúc Data Lakehouse cho dữ liệu y tế. Lakehouse đóng vai trò là lớp kiến trúc logic thống nhất, cho phép truy cập và khai thác dữ liệu một cách nhất quán, bất kể dữ liệu được lưu trữ và xử lý ở on-premise hay cloud. Sự kết hợp giữa Hybrid Cloud và Data Lakehouse giúp giảm thiểu sự phân mảnh dữ liệu, nâng cao khả năng quản trị và hỗ trợ đồng thời cả xử lý batch và streaming.

Việc áp dụng mô hình này tạo điều kiện thuận lợi cho việc mở rộng hệ thống trong tương lai, khi nhu cầu lưu trữ và phân tích dữ liệu y tế tiếp tục gia tăng. Đồng thời, kiến trúc đám mây lai cũng cho phép tổ chức y tế từng bước chuyển đổi số, khai thác lợi ích của đám mây mà không cần thay đổi toàn bộ hạ tầng hiện có.

3.9.5. Chuẩn bị logic cho chương thiết kế kiến trúc hệ thống

Từ các phân tích trên, có thể thấy mô hình đám mây lai là lựa chọn phù hợp để triển khai hệ thống Data Lakehouse cho dữ liệu y tế. Việc phân chia vai trò xử lý giữa on-premise và cloud, kết hợp với kiến trúc Lakehouse và Medallion, tạo nên một nền tảng dữ liệu linh hoạt, an toàn và có khả năng mở rộng cao.

Những nguyên lý này là cơ sở để thiết kế kiến trúc hệ thống cụ thể được trình bày trong chương tiếp theo. Chương 4 sẽ tập trung mô tả chi tiết kiến trúc Data Lakehouse trong môi trường đám mây lai, cách tổ chức các thành phần xử lý dữ liệu và sự phối hợp giữa các công nghệ nhằm hiện thực hóa mô hình đề xuất trong bối cảnh thực tế của dữ liệu y tế.

3.10. Công nghệ và thành phần nền tảng của hệ thống Data Lakehouse

Trong kiến trúc Data Lakehouse cho dữ liệu y tế trong môi trường đám mây lai, việc lựa chọn các công nghệ nền tảng đóng vai trò quyết định đến khả năng mở rộng, hiệu năng, độ tin cậy và mức độ bảo mật của hệ thống. Các công nghệ được lựa chọn trong luận văn đều là các nền tảng mã nguồn mở hoặc nền tảng đám mây hiện đại, đã được kiểm chứng rộng rãi trong các hệ thống dữ liệu lớn. Mỗi thành phần đảm nhiệm một vai trò riêng biệt nhưng có sự liên kết chặt chẽ, tạo thành một hệ sinh thái xử lý dữ liệu thống nhất từ thu thập, lưu trữ đến phân tích.

3.10.1. MQTT – Giao thức thu thập dữ liệu thời gian thực

MQTT (Message Queuing Telemetry Transport) là một giao thức truyền thông nhẹ, được thiết kế đặc biệt cho các hệ thống phân tán có tài nguyên hạn chế, chẳng hạn như các thiết bị IoT. Trong bối cảnh y tế hiện đại, nhiều thiết bị giám sát sinh hiệu như máy đo ECG, SpO_2 hoặc huyết áp hoạt động liên tục và tạo ra dòng dữ liệu thời gian thực với tần suất cao.

MQTT hoạt động dựa trên mô hình publish/subscribe, trong đó các thiết bị y tế đóng vai trò là publisher, gửi dữ liệu lên broker trung tâm, và các hệ thống xử lý phía sau đóng vai trò là subscriber. Cơ chế này giúp giảm sự phụ thuộc trực tiếp giữa các thành phần, tăng khả năng mở rộng và đảm bảo hệ thống vẫn hoạt động ổn định khi số lượng thiết bị tăng lên.

Đối với dữ liệu ECG có tần số lấy mẫu lên đến 500Hz, MQTT đáp ứng tốt yêu cầu về độ trễ thấp và khả năng truyền tải liên tục. Ngoài ra, giao thức này hỗ trợ các mức chất lượng dịch vụ (QoS), cho phép cân bằng giữa độ tin cậy và hiệu năng – một yếu tố quan trọng trong môi trường y tế, nơi mất dữ liệu có thể ảnh hưởng đến quyết định lâm sàng.

3.10.2. Apache NiFi – Thu thập và xử lý dữ liệu dòng

Apache NiFi là nền tảng quản lý và điều phối luồng dữ liệu, được thiết kế để xử lý dữ liệu streaming với độ tin cậy cao. Trong hệ thống Data Lakehouse, NiFi đóng vai trò là lớp trung gian giữa tầng thu thập dữ liệu (MQTT) và tầng lưu trữ – xử lý phía sau.

Một trong những ưu điểm lớn của NiFi là khả năng quản lý luồng dữ liệu theo hướng trực quan, cho phép định tuyến, lọc, chuyển đổi và ghi dữ liệu một cách linh hoạt. Điều này đặc biệt quan trọng đối với dữ liệu y tế, vốn có thể đến từ nhiều nguồn khác nhau và có cấu trúc không đồng nhất.

NiFi cung cấp các cơ chế như back-pressure, đảm bảo hệ thống không bị quá tải khi lưu lượng dữ liệu tăng đột biến, cũng như data provenance, cho phép truy vết toàn bộ vòng đời của dữ liệu. Trong lĩnh vực y tế, khả năng truy vết này hỗ trợ kiểm toán dữ liệu và đáp ứng các yêu cầu tuân thủ pháp lý.

3.10.3. Apache Spark – Nền tảng xử lý dữ liệu lớn

Apache Spark là nền tảng xử lý dữ liệu phân tán mã nguồn mở, được thiết kế nhằm khắc phục những hạn chế về hiệu năng và tính linh hoạt của các hệ thống xử lý dữ liệu thế hệ trước. Trong bối cảnh dữ liệu y tế ngày càng gia tăng về quy mô, tốc độ phát sinh và mức độ phức tạp, Spark đóng vai trò trung tâm trong kiến trúc Data

Lakehouse nhờ khả năng xử lý thống nhất cả dữ liệu theo lô (batch) và dữ liệu thời gian thực (streaming).

Nguyên lý xử lý phân tán và mô hình lập trình của Spark:

Spark hoạt động dựa trên mô hình xử lý phân tán, trong đó dữ liệu được chia nhỏ và phân phối trên nhiều nút tính toán trong cụm (cluster). Các tác vụ xử lý được thực thi song song, giúp rút ngắn thời gian xử lý và nâng cao khả năng mở rộng theo chiều ngang. Trái với các hệ thống xử lý dựa hoàn toàn trên đĩa, Spark ưu tiên xử lý dữ liệu trong bộ nhớ (in-memory processing), từ đó cải thiện đáng kể hiệu năng đối với các bài toán phân tích lặp và xử lý dữ liệu lớn.

Mô hình lập trình của Spark dựa trên các tập dữ liệu phân tán bất biến, ban đầu là RDD (Resilient Distributed Dataset) và sau này được mở rộng với DataFrame và Dataset. Các cấu trúc dữ liệu này cho phép biểu diễn dữ liệu y tế ở dạng bảng có schema rõ ràng, thuận lợi cho việc xử lý, tối ưu truy vấn và tích hợp với các công cụ phân tích khác trong kiến trúc Lakehouse.

Spark trong xử lý dữ liệu thời gian thực y tế:

Đối với dữ liệu thời gian thực như tín hiệu ECG có tần số lấy mẫu cao (500Hz), Spark Structured Streaming cung cấp mô hình xử lý dòng dựa trên micro-batch, cho phép xử lý dữ liệu streaming với độ trễ thấp trong khi vẫn giữ được tính nhất quán của mô hình lập trình. Điều này giúp đơn giản hóa việc phát triển các pipeline xử lý dữ liệu thời gian thực, khi cùng một framework có thể xử lý cả dữ liệu streaming và batch.

Khả năng tích hợp với các hệ thống thu thập dữ liệu như MQTT và Apache NiFi giúp Spark dễ dàng tiếp nhận dòng dữ liệu liên tục từ các thiết bị y tế. Trong kiến trúc Data Lakehouse, Spark đóng vai trò xử lý các bước tiền xử lý dữ liệu thời gian thực, bao gồm lọc nhiễu, chuẩn hóa tín hiệu và ghi dữ liệu vào các lớp lưu trữ phù hợp theo kiến trúc Medallion.

Spark trong xử lý dữ liệu theo lô và dữ liệu hình ảnh y tế:

Bên cạnh xử lý streaming, Spark đặc biệt phù hợp cho các tác vụ xử lý dữ liệu theo lô, điển hình là dữ liệu hình ảnh y tế như X-ray, CT và MRI. Các tập dữ liệu hình ảnh này thường có dung lượng lớn, cấu trúc phức tạp và yêu cầu xử lý song song để đảm bảo hiệu năng.

Spark cho phép phân phối việc xử lý các tập dữ liệu hình ảnh trên nhiều nút tính toán, hỗ trợ các thao tác như đọc dữ liệu, chuyển đổi định dạng và trích xuất đặc trưng. Khả năng tích hợp với các thư viện học máy và xử lý dữ liệu nâng cao giúp Spark trở thành nền tảng trung gian hiệu quả giữa tầng lưu trữ dữ liệu và các mô hình phân tích chuyên sâu.

Vai trò của Spark trong kiến trúc Data Lakehouse và Medallion:

Trong kiến trúc Data Lakehouse, Spark đóng vai trò là công cụ xử lý chính để hiện thực hóa các lớp dữ liệu Bronze, Silver và Gold. Ở lớp Bronze, Spark tiếp nhận và ghi dữ liệu thô từ các nguồn khác nhau vào hệ thống lưu trữ. Ở lớp Silver, Spark thực hiện các bước làm sạch, chuẩn hóa và kiểm tra chất lượng dữ liệu. Ở lớp Gold, Spark hỗ trợ các phép tổng hợp, biến đổi dữ liệu để tạo ra các tập dữ liệu sẵn sàng cho phân tích và báo cáo.

Sự kết hợp giữa Spark và Delta Lake giúp đảm bảo tính nhất quán và toàn vẹn của dữ liệu trong toàn bộ vòng đời xử lý. Các tính năng như kiểm soát phiên bản, hỗ trợ giao dịch ACID và tối ưu hóa truy vấn giúp Spark trở thành một thành phần không thể thiếu trong kiến trúc Lakehouse.

Lý do lựa chọn Apache Spark cho hệ thống dữ liệu y tế:

Apache Spark được lựa chọn trong luận văn nhờ các ưu điểm nổi bật: khả năng xử lý dữ liệu lớn với hiệu năng cao, hỗ trợ đồng thời batch và streaming, hệ sinh thái phong phú và khả năng tích hợp tốt với các công nghệ lưu trữ và truy vấn hiện đại. Đặc biệt, Spark phù hợp với mô hình đám mây lai, cho phép triển khai linh hoạt trên cả môi trường on-premise và đám mây.

Trong lĩnh vực y tế, nơi dữ liệu liên tục gia tăng và yêu cầu xử lý ngày càng phức tạp, Spark cung cấp nền tảng xử lý linh hoạt, mở rộng và đáng tin cậy. Vai trò trung tâm của Spark trong kiến trúc Data Lakehouse giúp kết nối các thành phần thu thập, lưu trữ và phân tích dữ liệu, tạo tiền đề cho việc triển khai các ứng dụng phân tích nâng cao và trí tuệ nhân tạo trong chăm sóc sức khỏe.

3.10.4. Apache Airflow – Điều phối và quản lý luồng công việc

Apache Airflow là một nền tảng mã nguồn mở được thiết kế để điều phối, lập lịch và giám sát các luồng công việc (workflow orchestration) trong các hệ thống xử lý dữ liệu phức tạp. Trong bối cảnh hệ thống Data Lakehouse cho dữ liệu y tế, nơi tồn tại đồng thời nhiều pipeline xử lý dữ liệu theo lô (batch) và dữ liệu thời gian thực (streaming), Airflow đóng vai trò trung tâm trong việc quản lý vòng đời xử lý dữ liệu, đảm bảo các tác vụ được thực thi đúng thứ tự, đúng thời điểm và với độ tin cậy cao.

Khác với các công cụ xử lý dữ liệu trực tiếp, Apache Airflow không thực hiện việc xử lý dữ liệu mà tập trung vào việc điều phối các tác vụ xử lý do các hệ thống khác đảm nhiệm, chẳng hạn như Apache Spark, Apache NiFi hoặc các dịch vụ phân tích trên đám mây. Cách tiếp cận này giúp tách biệt rõ ràng giữa lớp điều phối và lớp xử lý, góp phần nâng cao khả năng mở rộng, bảo trì và quản lý hệ thống tổng thể.

Mô hình DAG và quản lý phụ thuộc tác vụ

Apache Airflow sử dụng mô hình DAG (Directed Acyclic Graph) để biểu diễn các luồng công việc. Trong mô hình này, mỗi tác vụ (task) được xem như một đỉnh của đồ

thị, và các cạnh thể hiện mối quan hệ phụ thuộc giữa các tác vụ. Tính chất không chu trình của DAG đảm bảo rằng luồng công việc luôn có điểm bắt đầu và kết thúc rõ ràng, tránh các vòng lặp vô hạn trong quá trình thực thi.

Đối với hệ thống dữ liệu y tế, mô hình DAG cho phép mô tả chính xác các pipeline xử lý phức tạp, ví dụ như: thu thập dữ liệu từ thiết bị IoT y tế, làm sạch dữ liệu, chuẩn hóa schema, lưu trữ vào các tầng Bronze – Silver – Gold và cuối cùng kích hoạt các tác vụ phân tích hoặc huấn luyện mô hình. Việc mô hình hóa pipeline dưới dạng DAG giúp các kỹ sư dữ liệu dễ dàng kiểm soát và điều chỉnh luồng xử lý khi yêu cầu nghiệp vụ thay đổi.

Lập lịch và tự động hóa pipeline dữ liệu

Một trong những chức năng cốt lõi của Apache Airflow là khả năng lập lịch (scheduling) các luồng công việc theo thời gian hoặc theo sự kiện. Airflow hỗ trợ nhiều cơ chế lập lịch linh hoạt, từ chạy định kỳ theo giờ, ngày, tuần cho đến kích hoạt dựa trên điều kiện cụ thể.

Trong môi trường y tế, nơi dữ liệu được tạo ra liên tục và yêu cầu xử lý định kỳ để phục vụ báo cáo, phân tích hoặc lưu trữ lâu dài, khả năng lập lịch của Airflow giúp tự động hóa toàn bộ quy trình ETL. Ví dụ, dữ liệu hình ảnh y tế có thể được xử lý theo lô vào các khung giờ thấp điểm, trong khi dữ liệu tổng hợp từ streaming được cập nhật định kỳ để phục vụ phân tích gần thời gian thực. Việc tự động hóa này giúp giảm sự can thiệp thủ công, hạn chế sai sót và tăng độ ổn định cho hệ thống.

Giám sát, xử lý lỗi và khả năng phục hồi

Apache Airflow cung cấp giao diện quản lý trực quan cho phép theo dõi trạng thái của từng tác vụ và toàn bộ luồng công việc theo thời gian thực. Mỗi tác vụ trong DAG đều có trạng thái rõ ràng như thành công, thất bại, đang chạy hoặc chờ thực thi. Khi xảy ra lỗi, Airflow hỗ trợ cơ chế retry, gửi cảnh báo và ghi log chi tiết để hỗ trợ việc phân tích và khắc phục sự cố.

Đặc điểm này đặc biệt quan trọng đối với các hệ thống dữ liệu y tế, nơi yêu cầu cao về tính liên tục và độ tin cậy. Việc phát hiện sớm và xử lý kịp thời các lỗi trong pipeline giúp đảm bảo dữ liệu được xử lý đầy đủ, nhất quán và không bị gián đoạn, từ đó duy trì chất lượng dữ liệu phục vụ các hoạt động phân tích và nghiên cứu.

Vai trò của Airflow trong kiến trúc Data Lakehouse

Trong kiến trúc Data Lakehouse, Apache Airflow thường đóng vai trò là lớp điều phối trung tâm, kết nối và quản lý các thành phần xử lý dữ liệu khác nhau. Airflow không phụ thuộc vào công nghệ lưu trữ hay xử lý cụ thể, do đó có thể tích hợp linh hoạt với các hệ thống như Delta Lake, MinIO, Hive Metastore và các nền tảng xử lý dữ liệu trên đám mây.

Việc sử dụng Airflow giúp chuẩn hóa cách thức xây dựng và vận hành pipeline dữ liệu trong toàn bộ hệ thống, từ on-premise đến cloud. Điều này đặc biệt phù hợp với mô hình đám mây lai, nơi các tác vụ xử lý được phân bố trên nhiều môi trường khác nhau nhưng vẫn cần được quản lý một cách thống nhất.

Ý nghĩa đối với hệ thống dữ liệu y tế

Trong bối cảnh dữ liệu y tế ngày càng lớn, đa dạng và yêu cầu xử lý liên tục, Apache Airflow cung cấp một nền tảng điều phối mạnh mẽ giúp đảm bảo tính tự động, minh bạch và khả năng kiểm soát của hệ thống Data Lakehouse. Việc áp dụng Airflow không chỉ giúp nâng cao hiệu quả vận hành mà còn tạo nền tảng cho việc mở rộng hệ thống trong tương lai, khi khối lượng dữ liệu và số lượng pipeline tiếp tục gia tăng.

3.10.5. Delta Lake – Lớp lưu trữ dữ liệu Lakehouse

Delta Lake là một lớp lưu trữ dữ liệu mã nguồn mở được xây dựng trên nền tảng lưu trữ dạng file, nhằm bổ sung các đặc tính quản trị và độ tin cậy truyền thống của Data Warehouse cho kiến trúc Data Lake. Trong hệ thống Data Lakehouse cho dữ liệu y tế, Delta Lake đóng vai trò trung tâm trong việc đảm bảo tính nhất quán, toàn vẹn và khả năng quản lý dữ liệu trong suốt vòng đời xử lý, từ dữ liệu thô đến dữ liệu phục vụ phân tích.

Hạn chế của Data Lake truyền thống và động lực ra đời của Delta Lake:

Các kiến trúc Data Lake truyền thống cho phép lưu trữ dữ liệu ở định dạng gốc với chi phí thấp, nhưng thường thiếu các cơ chế kiểm soát chặt chẽ về giao dịch, phiên bản và chất lượng dữ liệu. Điều này dẫn đến các vấn đề như dữ liệu không nhất quán khi có nhiều tiến trình đọc/ghi đồng thời, khó truy vết lịch sử thay đổi và rủi ro mất dữ liệu khi pipeline xử lý gặp lỗi.

Trong lĩnh vực y tế, những hạn chế này trở nên đặc biệt nghiêm trọng do dữ liệu y tế yêu cầu độ chính xác cao, khả năng truy vết và tuân thủ các quy định pháp lý nghiêm ngặt. Delta Lake ra đời nhằm giải quyết các vấn đề này bằng cách đưa các đặc tính của hệ quản trị cơ sở dữ liệu phân tích vào lớp lưu trữ của Data Lake.

Nguyên lý hoạt động của Delta Lake:

Delta Lake sử dụng một cơ chế nhật ký giao dịch (transaction log) để quản lý mọi thay đổi trên dữ liệu. Mỗi thao tác ghi, cập nhật hoặc xóa dữ liệu đều được ghi lại trong transaction log, giúp hệ thống duy trì trạng thái nhất quán của dữ liệu ngay cả khi có nhiều tiến trình truy cập đồng thời.

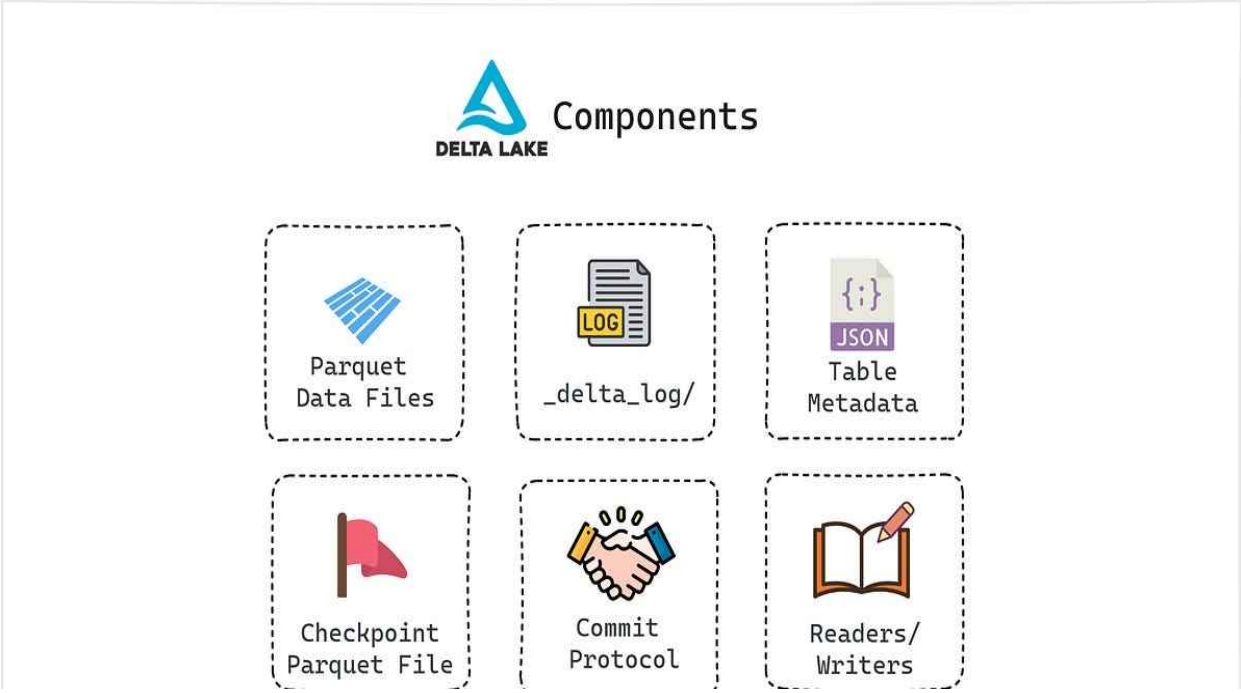
Dữ liệu được lưu trữ dưới dạng các file định dạng mở như Apache Parquet, trong khi metadata và thông tin giao dịch được quản lý riêng biệt. Cách tiếp cận này cho phép Delta Lake kết hợp hiệu quả giữa hiệu năng xử lý và khả năng quản trị dữ

liệu, đồng thời đảm bảo khả năng tương thích với các công cụ xử lý dữ liệu lớn khác trong hệ sinh thái.

Đảm bảo tính toàn vẹn dữ liệu với giao dịch ACID:

Một trong những đặc điểm nổi bật của Delta Lake là hỗ trợ đầy đủ các tính chất ACID (Atomicity, Consistency, Isolation, Durability). Đối với hệ thống dữ liệu y tế, điều này đảm bảo rằng mỗi thao tác ghi dữ liệu đều được thực hiện một cách trọn vẹn hoặc không được thực hiện, tránh tình trạng dữ liệu bị ghi dang dở hoặc không nhất quán.

Tính nhất quán và cô lập của các giao dịch giúp nhiều pipeline xử lý dữ liệu, bao gồm cả xử lý streaming và batch, có thể hoạt động song song mà không gây xung đột. Điều này đặc biệt quan trọng trong kiến trúc Data Lakehouse, nơi dữ liệu thời gian thực và dữ liệu theo lô cùng được ghi vào các lớp lưu trữ chung.



Hình 5. Kiến trúc Medallion (Bronze – Silver – Gold)

Phiên bản hóa và truy vết lịch sử dữ liệu:

Delta Lake hỗ trợ phiên bản hóa dữ liệu thông qua cơ chế time travel, cho phép truy xuất lại trạng thái dữ liệu tại một thời điểm trong quá khứ. Tính năng này có ý nghĩa quan trọng trong lĩnh vực y tế, nơi dữ liệu thường cần được kiểm toán, đối chiếu và truy vết lịch sử thay đổi.

Khả năng truy vết lịch sử giúp các tổ chức y tế dễ dàng kiểm tra lại dữ liệu khi có sai sót, đánh giá lại kết quả phân tích và đáp ứng các yêu cầu kiểm toán nội bộ cũng

như bên ngoài. Ngoài ra, phiên bản hóa dữ liệu còn hỗ trợ việc khôi phục dữ liệu trong trường hợp pipeline xử lý gặp sự cố.

Vai trò của Delta Lake trong kiến trúc Medallion:

Delta Lake là nền tảng lý tưởng để triển khai kiến trúc Medallion với các lớp Bronze, Silver và Gold. Ở lớp Bronze, Delta Lake lưu trữ dữ liệu thô từ các nguồn khác nhau, bao gồm dữ liệu thời gian thực và dữ liệu theo lô, đảm bảo dữ liệu được ghi nhận đầy đủ và có thể truy vết.

Ở lớp Silver, Delta Lake hỗ trợ các bước làm sạch, chuẩn hóa và kiểm tra chất lượng dữ liệu, giúp chuyển đổi dữ liệu từ dạng thô sang dạng có cấu trúc và đáng tin cậy hơn. Ở lớp Gold, Delta Lake cung cấp dữ liệu đã được tổng hợp và tối ưu, sẵn sàng cho các tác vụ phân tích, báo cáo và học máy. Sự nhất quán về định dạng và cơ chế quản lý dữ liệu giữa các lớp giúp giảm độ phức tạp trong thiết kế pipeline và nâng cao hiệu quả vận hành của hệ thống.

Phù hợp với mô hình đám mây lai và dữ liệu y tế:

Delta Lake được thiết kế để hoạt động hiệu quả trên nhiều nền tảng lưu trữ khác nhau, từ hệ thống on-premise như MinIO đến các dịch vụ lưu trữ đám mây. Điều này giúp Delta Lake trở thành lựa chọn phù hợp cho kiến trúc Data Lakehouse trong môi trường đám mây lai.

Trong bối cảnh dữ liệu y tế, Delta Lake cho phép tổ chức y tế tận dụng sức mạnh của đám mây cho phân tích dữ liệu lớn trong khi vẫn duy trì quyền kiểm soát đối với dữ liệu nhạy cảm lưu trữ tại chỗ. Sự linh hoạt này hỗ trợ các chiến lược mở rộng và chuyển đổi số từng bước, phù hợp với yêu cầu thực tế của hệ thống y tế.

Lý do lựa chọn Delta Lake cho hệ thống Data Lakehouse y tế:

Delta Lake được lựa chọn trong luận văn nhờ khả năng kết hợp hiệu quả giữa lưu trữ chi phí thấp và quản trị dữ liệu mạnh mẽ. Các tính năng như hỗ trợ ACID, phiên bản hóa dữ liệu, khả năng tích hợp với các công cụ xử lý dữ liệu lớn và phù hợp với mô hình đám mây lai khiến Delta Lake trở thành thành phần không thể thiếu trong kiến trúc Data Lakehouse.

Trong lĩnh vực y tế, nơi dữ liệu có giá trị cao và yêu cầu độ tin cậy nghiêm ngặt, Delta Lake cung cấp nền tảng lưu trữ đáng tin cậy, hỗ trợ khai thác dữ liệu hiệu quả và tạo tiền đề cho các ứng dụng phân tích nâng cao và trí tuệ nhân tạo trong chăm sóc sức khỏe.

3.10.6. MinIO – Lưu trữ đối tượng trong môi trường on-premise

MinIO là một nền tảng lưu trữ đối tượng mã nguồn mở, được thiết kế theo kiến trúc cloud-native và tương thích hoàn toàn với giao thức Amazon S3. Trong kiến trúc

Data Lakehouse cho dữ liệu y tế, MinIO đóng vai trò là lớp lưu trữ đối tượng chủ đạo trong môi trường on-premise, nơi lưu trữ dữ liệu thô, dữ liệu trung gian và dữ liệu nhạy cảm cần được kiểm soát chặt chẽ.

So sánh MinIO và HDFS trong hệ thống lưu trữ dữ liệu lớn

Trong các hệ thống xử lý dữ liệu lớn, lớp lưu trữ đóng vai trò nền tảng, quyết định khả năng mở rộng, hiệu năng và độ tin cậy của toàn bộ kiến trúc. Hai công nghệ lưu trữ phổ biến thường được xem xét trong các hệ thống Data Lake và Data Lakehouse là HDFS (Hadoop Distributed File System) và MinIO (Object Storage). Mặc dù cùng phục vụ mục tiêu lưu trữ dữ liệu quy mô lớn, hai hệ thống này có những khác biệt căn bản về kiến trúc, mô hình truy cập và phạm vi ứng dụng.

Kiến trúc lưu trữ

HDFS được thiết kế theo mô hình lưu trữ phân tán dựa trên block, trong đó dữ liệu được chia nhỏ thành các block cố định và phân phối trên nhiều node trong cụm. Hệ thống sử dụng NameNode để quản lý metadata và DataNode để lưu trữ dữ liệu thực tế. Kiến trúc này phù hợp với các môi trường xử lý batch truyền thống, nơi dữ liệu được ghi một lần và đọc nhiều lần.

Ngược lại, MinIO được xây dựng theo mô hình lưu trữ đối tượng (object storage), trong đó mỗi đối tượng bao gồm dữ liệu, metadata và định danh duy nhất. MinIO không yêu cầu thành phần quản lý trung tâm như NameNode mà sử dụng cơ chế phân tán metadata, giúp giảm điểm nghẽn và tăng khả năng mở rộng. Kiến trúc này phù hợp với các hệ thống hiện đại sử dụng lưu trữ đối tượng như nền tảng của Data Lakehouse.

Mô hình truy cập dữ liệu

HDFS sử dụng giao thức truy cập đặc thù của Hadoop và yêu cầu các ứng dụng phải tích hợp thông qua API của hệ sinh thái Hadoop. Điều này khiến HDFS gắn chặt với các framework như MapReduce và Spark, nhưng lại hạn chế khả năng tích hợp với các công cụ phân tích hiện đại và dịch vụ đám mây.

MinIO hỗ trợ giao thức S3-compatible, cho phép các ứng dụng truy cập dữ liệu thông qua giao thức HTTP tiêu chuẩn. Nhờ đó, MinIO có thể tích hợp dễ dàng với nhiều công cụ xử lý dữ liệu hiện đại như Spark, Trino, Presto và các nền tảng cloud. Mô hình truy cập này đặc biệt phù hợp với kiến trúc hybrid, nơi dữ liệu cần được chia sẻ giữa on-premise và cloud.

Hiệu năng và khả năng mở rộng

HDFS được tối ưu cho các tác vụ đọc/ghi tuần tự với khối lượng lớn, phù hợp cho xử lý batch truyền thống. Tuy nhiên, việc mở rộng HDFS thường đòi hỏi cấu hình phức tạp và phụ thuộc vào khả năng quản lý cụm Hadoop.

MinIO được thiết kế để mở rộng ngang dễ dàng bằng cách bổ sung node lưu trữ, đồng thời hỗ trợ cơ chế erasure coding giúp tối ưu dung lượng và tăng độ bền dữ liệu. Khả năng mở rộng linh hoạt này giúp MinIO phù hợp với các hệ thống yêu cầu xử lý dữ liệu lớn không đồng đều theo thời gian, như trong lĩnh vực y tế.

Quản lý metadata và schema

Trong HDFS, metadata được quản lý tập trung tại NameNode, dẫn đến nguy cơ trở thành điểm nghẽn và điểm lỗi duy nhất nếu không được cấu hình dự phòng đầy đủ. Ngoài ra, HDFS không hỗ trợ quản lý schema ở mức hệ thống, mà phụ thuộc vào các công cụ bên trên như Hive Metastore.

MinIO lưu trữ metadata cùng với đối tượng, cho phép truy cập linh hoạt và giảm phụ thuộc vào các thành phần quản lý tập trung. Khi kết hợp với các hệ quản trị metadata như Hive Metastore hoặc Glue Catalog, MinIO trở thành nền tảng lưu trữ phù hợp cho các kiến trúc lakehouse hiện đại.

Khả năng tích hợp trong kiến trúc Data Lakehouse

HDFS phù hợp với các kiến trúc Data Lake truyền thống, nơi hệ sinh thái Hadoop đóng vai trò trung tâm. Tuy nhiên, trong các kiến trúc Data Lakehouse hiện đại, nơi lưu trữ đối tượng được ưu tiên, HDFS dần bộc lộ hạn chế về tính linh hoạt và khả năng tích hợp.

MinIO, với vai trò là object storage on-premise, cho phép triển khai kiến trúc Data Lakehouse thống nhất giữa on-premise và cloud. Việc sử dụng MinIO giúp hệ thống dữ liệu y tế tận dụng các công cụ lakehouse như Delta Lake, đồng thời duy trì quyền kiểm soát dữ liệu nhạy cảm trong môi trường nội bộ.

Trong lĩnh vực y tế, dữ liệu thường có dung lượng lớn, đa dạng định dạng và yêu cầu cao về bảo mật. HDFS đáp ứng tốt các bài toán xử lý batch truyền thống nhưng gặp khó khăn khi cần tích hợp với các nền tảng phân tích hiện đại và môi trường đám mây.

MinIO cung cấp giải pháp lưu trữ đối tượng linh hoạt, dễ tích hợp và phù hợp với mô hình đám mây lai. Việc sử dụng MinIO trong hệ thống dữ liệu y tế giúp đảm bảo tính chủ động trong quản lý dữ liệu nhạy cảm, đồng thời sẵn sàng mở rộng lên cloud khi cần thiết.

HDFS và MinIO đại diện cho hai thế hệ lưu trữ dữ liệu lớn với triết lý thiết kế khác nhau. HDFS phù hợp với các hệ thống Hadoop truyền thống và xử lý batch quy mô lớn, trong khi MinIO phù hợp hơn với các kiến trúc Data Lakehouse hiện đại và mô hình đám mây lai. Trong bối cảnh luận văn này, MinIO được lựa chọn làm nền tảng lưu trữ on-premise nhằm đáp ứng yêu cầu linh hoạt, mở rộng và tích hợp của hệ thống dữ liệu y tế lớn.

Vai trò của lưu trữ đối tượng trong hệ thống dữ liệu y tế:

Dữ liệu y tế hiện đại bao gồm nhiều loại dữ liệu có kích thước lớn và cấu trúc đa dạng, như tín hiệu sinh học thời gian thực, hình ảnh y tế chuẩn DICOM và các tập dữ liệu bán cấu trúc. Các hệ thống lưu trữ truyền thống dựa trên block storage hoặc file system gặp nhiều hạn chế khi phải mở rộng quy mô và quản lý dữ liệu phi cấu trúc.

Mô hình lưu trữ đối tượng cho phép lưu trữ dữ liệu dưới dạng các đối tượng độc lập, mỗi đối tượng đi kèm metadata mô tả, giúp hệ thống dễ dàng mở rộng, quản lý và truy xuất dữ liệu. Đối với hệ thống dữ liệu y tế, lưu trữ đối tượng đặc biệt phù hợp cho các tập dữ liệu dung lượng lớn và yêu cầu lưu trữ dài hạn.

Nguyên lý hoạt động và kiến trúc của MinIO:

MinIO được thiết kế theo kiến trúc phân tán, trong đó dữ liệu được phân mảnh và sao chép trên nhiều nút lưu trữ để đảm bảo tính sẵn sàng và độ bền. Hệ thống sử dụng cơ chế erasure coding để tối ưu hóa việc sử dụng dung lượng lưu trữ trong khi vẫn đảm bảo khả năng phục hồi dữ liệu khi có sự cố phần cứng.

Tính tương thích với giao thức S3 cho phép MinIO dễ dàng tích hợp với các công cụ xử lý dữ liệu lớn như Apache Spark, Delta Lake và Trino. Điều này giúp MinIO trở thành lớp lưu trữ linh hoạt, có thể được sử dụng như nền tảng lưu trữ chính cho kiến trúc Data Lakehouse on-premise.

MinIO trong kiến trúc Data Lakehouse và Medallion:

Trong kiến trúc Data Lakehouse, MinIO đóng vai trò là lớp lưu trữ nền tảng cho các lớp dữ liệu Bronze, Silver và Gold khi triển khai trong môi trường on-premise. Ở lớp Bronze, MinIO lưu trữ dữ liệu thô được thu thập trực tiếp từ các nguồn như thiết bị IoT y tế và hệ thống chẩn đoán hình ảnh. Ở lớp Silver và Gold, MinIO tiếp tục cung cấp nền tảng lưu trữ cho dữ liệu đã được xử lý, đảm bảo tính nhất quán và khả năng truy xuất hiệu quả.

Sự kết hợp giữa MinIO và Delta Lake cho phép bổ sung các tính năng quản trị dữ liệu nâng cao trên nền tảng lưu trữ đối tượng, tạo nên một hệ thống lưu trữ vừa linh hoạt vừa đáng tin cậy cho dữ liệu y tế.

Bảo mật và kiểm soát truy cập trong môi trường y tế:

Bảo mật là yếu tố then chốt đối với hệ thống dữ liệu y tế. MinIO hỗ trợ các cơ chế bảo mật như mã hóa dữ liệu khi lưu trữ và khi truyền tải, xác thực và phân quyền truy cập dựa trên chính sách. Các cơ chế này giúp tổ chức y tế kiểm soát chặt chẽ quyền truy cập vào dữ liệu nhạy cảm, giảm thiểu nguy cơ rò rỉ thông tin.

Ngoài ra, việc triển khai MinIO trong môi trường on-premise cho phép tổ chức y tế duy trì quyền kiểm soát vật lý đối với hạ tầng lưu trữ, đáp ứng tốt các yêu cầu về tuân thủ pháp lý và chính sách nội bộ.

MinIO trong mô hình đám mây lai:

Trong kiến trúc đám mây lai, MinIO đóng vai trò là cầu nối giữa hệ thống lưu trữ on-premise và các nền tảng đám mây. Nhờ khả năng tương thích S3, dữ liệu lưu trữ trên MinIO có thể được đồng bộ hoặc truy cập bởi các dịch vụ xử lý dữ liệu trên cloud, chẳng hạn như Databricks.

Cách tiếp cận này cho phép tổ chức y tế tận dụng sức mạnh của đám mây cho phân tích dữ liệu lớn trong khi vẫn giữ dữ liệu nhạy cảm tại chỗ. Việc phân tách vai trò lưu trữ và xử lý giúp hệ thống đạt được sự cân bằng giữa bảo mật và khả năng mở rộng.

Lý do lựa chọn MinIO cho hệ thống Data Lakehouse y tế:

MinIO được lựa chọn trong luận văn nhờ các ưu điểm nổi bật: hiệu năng cao, khả năng mở rộng theo chiều ngang, tương thích với hệ sinh thái S3 và khả năng triển khai linh hoạt trong môi trường on-premise. Đặc biệt, MinIO phù hợp với yêu cầu lưu trữ dữ liệu y tế nhạy cảm, cho phép tổ chức y tế duy trì quyền kiểm soát dữ liệu trong khi vẫn xây dựng được kiến trúc Data Lakehouse hiện đại.

Với vai trò là lớp lưu trữ đối tượng on-premise, MinIO tạo nền tảng vững chắc cho toàn bộ hệ thống dữ liệu y tế trong mô hình đám mây lai, đồng thời hỗ trợ hiệu quả cho các pipeline xử lý dữ liệu và các ứng dụng phân tích nâng cao trong các chương tiếp theo của luận văn.

3.10.7. Hive Metastore – Quản lý metadata và schema

Hive Metastore là thành phần quản lý metadata trung tâm trong hệ sinh thái dữ liệu lớn, chịu trách nhiệm lưu trữ và quản lý thông tin mô tả về cấu trúc dữ liệu, schema, bảng, phân vùng và vị trí lưu trữ dữ liệu vật lý. Trong kiến trúc Data Lakehouse cho dữ liệu y tế, Hive Metastore đóng vai trò then chốt trong việc tổ chức, quản trị và khai thác dữ liệu một cách nhất quán và có kiểm soát.

Vai trò của metadata trong hệ thống dữ liệu y tế:

Metadata được xem là “dữ liệu của dữ liệu”, cung cấp thông tin mô tả giúp hệ thống hiểu và xử lý dữ liệu một cách chính xác. Trong lĩnh vực y tế, nơi dữ liệu có cấu trúc phức tạp, đa nguồn và thay đổi theo thời gian, metadata đóng vai trò đặc biệt quan trọng. Các thông tin như schema của bảng dữ liệu, ý nghĩa của từng trường, nguồn gốc dữ liệu và thời điểm thu thập đều cần được quản lý rõ ràng để đảm bảo tính chính xác và khả năng truy vết.

Việc thiếu quản lý metadata có thể dẫn đến sự nhầm lẫn trong khai thác dữ liệu, gia tăng rủi ro sai lệch kết quả phân tích và gây khó khăn trong việc tuân thủ các quy định về kiểm toán và bảo mật dữ liệu y tế. Do đó, một hệ thống quản lý metadata tập trung là thành phần không thể thiếu trong kiến trúc Data Lakehouse.

Nguyên lý hoạt động của Hive Metastore:

Hive Metastore hoạt động như một kho lưu trữ metadata tập trung, thường được triển khai dưới dạng một dịch vụ độc lập, sử dụng cơ sở dữ liệu quan hệ để lưu trữ thông tin metadata. Các công cụ xử lý và truy vấn dữ liệu như Apache Spark, Trino và các hệ thống BI truy cập Hive Metastore để lấy thông tin về schema, cấu trúc bảng và vị trí dữ liệu trước khi thực hiện các tác vụ xử lý.

Cách tiếp cận này giúp tách biệt rõ ràng giữa dữ liệu vật lý và metadata, cho phép nhiều công cụ khác nhau truy cập và khai thác dữ liệu trên cùng một nền tảng lưu trữ. Trong kiến trúc Data Lakehouse, Hive Metastore đóng vai trò như “bộ não” điều phối thông tin cấu trúc dữ liệu, đảm bảo các thành phần trong hệ thống có chung một cách hiểu về dữ liệu.

Quản lý schema và hỗ trợ tiến hóa dữ liệu:

Một thách thức lớn trong hệ thống dữ liệu y tế là sự thay đổi liên tục của schema, do các thiết bị, hệ thống và tiêu chuẩn y tế không ngừng được cập nhật. Hive Metastore hỗ trợ quản lý schema một cách tập trung, cho phép theo dõi và kiểm soát các thay đổi cấu trúc dữ liệu theo thời gian.

Khả năng quản lý schema tập trung giúp hệ thống hạn chế các lỗi phát sinh khi schema thay đổi, đồng thời hỗ trợ các pipeline xử lý dữ liệu thích ứng linh hoạt với các thay đổi này. Trong kiến trúc Data Lakehouse, sự kết hợp giữa Hive Metastore và các định dạng lưu trữ như Delta Lake giúp đảm bảo quá trình tiến hóa schema diễn ra một cách có kiểm soát và an toàn.

Hive Metastore trong kiến trúc Medallion:

Trong kiến trúc Medallion (Bronze – Silver – Gold), Hive Metastore đóng vai trò quản lý metadata cho từng lớp dữ liệu, giúp phân biệt rõ ràng trạng thái và mục đích sử dụng của dữ liệu. Ở lớp Bronze, metadata giúp mô tả dữ liệu thô, nguồn gốc và thời điểm thu thập. Ở lớp Silver, metadata phản ánh các bước làm sạch và chuẩn hóa dữ liệu. Ở lớp Gold, metadata cung cấp thông tin về các tập dữ liệu đã được tổng hợp và tối ưu cho phân tích.

Cách tổ chức này giúp người dùng và các công cụ phân tích dễ dàng hiểu được bối cảnh và mức độ tin cậy của dữ liệu, từ đó khai thác dữ liệu một cách hiệu quả và an toàn hơn.

Hỗ trợ quản trị dữ liệu và tuân thủ quy định:

Trong hệ thống dữ liệu y tế, quản trị dữ liệu và tuân thủ các quy định pháp lý là yêu cầu bắt buộc. Hive Metastore hỗ trợ các chức năng quản trị dữ liệu thông qua việc quản lý metadata tập trung, tạo nền tảng cho việc áp dụng các chính sách phân quyền và kiểm soát truy cập.

Thông qua việc kết hợp Hive Metastore với các hệ thống quản lý quyền truy cập, tổ chức y tế có thể xác định rõ ai được phép truy cập vào dữ liệu nào, ở mức độ nào. Điều này giúp giảm thiểu rủi ro truy cập trái phép và đảm bảo tuân thủ các yêu cầu về bảo mật và quyền riêng tư dữ liệu y tế.

Phù hợp với mô hình đám mây lai và hệ sinh thái Lakehouse:

Hive Metastore được thiết kế để hoạt động hiệu quả trong các môi trường phân tán, bao gồm cả on-premise và cloud. Khả năng tích hợp với nhiều công cụ xử lý và truy vấn dữ liệu giúp Hive Metastore trở thành thành phần lý tưởng cho kiến trúc Data Lakehouse trong môi trường đám mây lai.

Trong mô hình này, Hive Metastore cung cấp một lớp metadata thống nhất, cho phép các thành phần xử lý dữ liệu trên on-premise và cloud truy cập và khai thác dữ liệu một cách nhất quán. Điều này giúp giảm thiểu sự phức tạp trong quản lý hệ thống và nâng cao khả năng mở rộng trong tương lai.

Lý do lựa chọn Hive Metastore cho hệ thống Data Lakehouse y tế:

Hive Metastore được lựa chọn trong luận văn nhờ khả năng quản lý metadata tập trung, hỗ trợ tiến hóa schema và tích hợp tốt với các công cụ xử lý dữ liệu lớn. Đối với hệ thống dữ liệu y tế, nơi yêu cầu cao về tính nhất quán, khả năng truy vết và quản trị dữ liệu, Hive Metastore cung cấp nền tảng quản lý metadata đáng tin cậy và linh hoạt.

Với vai trò là lớp quản lý metadata trung tâm, Hive Metastore góp phần quan trọng vào việc xây dựng một hệ thống Data Lakehouse hoàn chỉnh, hỗ trợ hiệu quả cho các pipeline xử lý dữ liệu và các ứng dụng phân tích nâng cao trong lĩnh vực y tế.

3.10.8. Trino – Công cụ truy vấn phân tán

Trino là công cụ truy vấn SQL phân tán, cho phép truy vấn dữ liệu từ nhiều nguồn khác nhau một cách thống nhất. Trong hệ thống Data Lakehouse, Trino đóng vai trò cung cấp khả năng truy vấn hiệu quả trên dữ liệu được lưu trữ trong Delta Lake và các hệ thống lưu trữ khác.

Khả năng truy vấn dữ liệu mà không cần sao chép giúp Trino giảm độ trễ và chi phí xử lý, đồng thời hỗ trợ các nhà phân tích và nhà khoa học dữ liệu khai thác dữ liệu y tế một cách linh hoạt. Trino đặc biệt phù hợp cho các tác vụ phân tích tương tác và truy vấn dữ liệu ở quy mô lớn.

3.10.9. Databricks – Nền tảng xử lý dữ liệu trên đám mây

Databricks là một nền tảng xử lý và phân tích dữ liệu lớn dựa trên đám mây, được thiết kế nhằm hỗ trợ toàn bộ vòng đời dữ liệu từ thu thập, xử lý, lưu trữ đến phân tích nâng cao. Trong kiến trúc Data Lakehouse cho dữ liệu y tế trong môi trường đám mây lai, Databricks đóng vai trò là lớp xử lý trung tâm trên cloud, bổ sung năng lực tính toán linh hoạt và khả năng mở rộng cao cho hệ thống on-premise.

Nguồn gốc và định hướng kiến trúc của Databricks:

Databricks được phát triển bởi các nhà sáng lập của Apache Spark với mục tiêu cung cấp một nền tảng hợp nhất cho xử lý dữ liệu lớn và phân tích nâng cao. Khác với các hệ thống xử lý truyền thống, Databricks được xây dựng theo định hướng lakehouse-native, trong đó dữ liệu được lưu trữ trên các hệ thống lưu trữ đối tượng và được xử lý trực tiếp mà không cần sao chép sang các kho dữ liệu riêng biệt.

Định hướng này giúp Databricks trở thành nền tảng phù hợp cho các tổ chức cần xử lý khối lượng dữ liệu lớn, đa dạng và thay đổi liên tục, như trong lĩnh vực y tế. Việc tích hợp sâu với Delta Lake giúp Databricks kết hợp được ưu điểm của Data Lake và Data Warehouse trong một kiến trúc thống nhất.

Databricks như lớp xử lý cloud trong mô hình Hybrid Cloud:

Trong mô hình đám mây lai, Databricks thường được triển khai trên các nền tảng đám mây công cộng để đảm nhận các tác vụ xử lý yêu cầu tài nguyên tính toán lớn, chẳng hạn như xử lý batch khối lượng dữ liệu hình ảnh y tế hoặc phân tích dữ liệu lịch sử dài hạn. Trong khi đó, hệ thống on-premise tập trung vào thu thập dữ liệu thời gian thực và lưu trữ dữ liệu nhạy cảm.

Sự phân tách này cho phép hệ thống tận dụng ưu điểm của cả hai môi trường: khả năng kiểm soát và bảo mật của on-premise, cùng với tính linh hoạt và khả năng mở rộng của cloud. Databricks đóng vai trò cầu nối, giúp mở rộng năng lực xử lý dữ liệu mà không làm gián đoạn các pipeline đang hoạt động trên hạ tầng nội bộ.

Hỗ trợ xử lý dữ liệu batch và streaming:

Databricks cung cấp môi trường xử lý thống nhất cho cả dữ liệu batch và dữ liệu thời gian thực thông qua Apache Spark. Điều này đặc biệt quan trọng trong hệ thống dữ liệu y tế, nơi dữ liệu ECG thời gian thực và dữ liệu hình ảnh y tế theo lô cần được xử lý song song trong cùng một kiến trúc.

Khả năng xử lý hợp nhất giúp giảm độ phức tạp trong thiết kế hệ thống, đồng thời đảm bảo tính nhất quán trong cách tiếp cận dữ liệu. Các pipeline xử lý dữ liệu có thể được triển khai và quản lý tập trung, từ đó nâng cao hiệu quả vận hành và khả năng mở rộng của hệ thống.

Vai trò trong việc xây dựng các lớp dữ liệu Medallion:

Databricks hỗ trợ trực tiếp mô hình Medallion thông qua việc xử lý và chuyển đổi dữ liệu giữa các lớp Bronze, Silver và Gold. Trong hệ thống Data Lakehouse y tế, Databricks đảm nhiệm các tác vụ làm sạch, chuẩn hóa và tổng hợp dữ liệu, giúp nâng cao chất lượng dữ liệu trước khi phục vụ cho phân tích và nghiên cứu.

Việc triển khai Medallion trên Databricks giúp hệ thống tận dụng các cơ chế quản lý dữ liệu tiên tiến như kiểm soát phiên bản và đảm bảo tính nhất quán của dữ liệu. Điều này đặc biệt quan trọng trong bối cảnh dữ liệu y tế yêu cầu độ tin cậy cao và khả năng truy vết lịch sử thay đổi.

Khả năng mở rộng và tối ưu tài nguyên:

Một trong những lợi thế nổi bật của Databricks là khả năng mở rộng tài nguyên tính toán theo nhu cầu xử lý. Trong các giai đoạn cao điểm, hệ thống có thể tăng cường tài nguyên để xử lý khối lượng dữ liệu lớn, sau đó giảm tài nguyên khi nhu cầu giảm xuống.

Cách tiếp cận này giúp tối ưu chi phí vận hành, đặc biệt phù hợp với các hệ thống dữ liệu y tế có khối lượng xử lý không đồng đều theo thời gian. Việc tách biệt lưu trữ và tính toán cũng giúp hệ thống linh hoạt hơn trong việc mở rộng và tối ưu hiệu năng.

Hỗ trợ phân tích nâng cao và nghiên cứu dữ liệu y tế:

Databricks không chỉ hỗ trợ xử lý dữ liệu truyền thống mà còn cung cấp môi trường thuận lợi cho các hoạt động phân tích nâng cao. Trong bối cảnh dữ liệu y tế, nền tảng này có thể được sử dụng để khai thác dữ liệu đã được chuẩn hóa nhằm phục vụ nghiên cứu, thống kê và hỗ trợ ra quyết định.

Khả năng tích hợp với các thư viện phân tích và công cụ khoa học dữ liệu giúp Databricks trở thành nền tảng phù hợp cho các nghiên cứu liên ngành, nơi dữ liệu y tế cần được xử lý và phân tích trên quy mô lớn.

Đảm bảo quản trị và bảo mật dữ liệu:

Trong môi trường đám mây, bảo mật và quản trị dữ liệu là những yếu tố then chốt. Databricks hỗ trợ các cơ chế quản lý truy cập và phân quyền, giúp đảm bảo rằng chỉ những đối tượng được ủy quyền mới có thể truy cập và xử lý dữ liệu nhạy cảm.

Sự kết hợp giữa Databricks và các lớp quản trị metadata giúp hệ thống duy trì tính nhất quán và tuân thủ các yêu cầu về bảo mật dữ liệu y tế. Điều này tạo nền tảng cho việc triển khai các hệ thống phân tích dữ liệu y tế trên cloud một cách an toàn và có kiểm soát.

Lý do lựa chọn Databricks trong luận văn:

Databricks được lựa chọn trong luận văn nhờ khả năng hỗ trợ toàn diện cho kiến trúc Data Lakehouse, tích hợp chặt chẽ với Delta Lake và phù hợp với mô hình đám mây lai. Nền tảng này cho phép mở rộng năng lực xử lý dữ liệu y tế mà không làm mất đi tính kiểm soát và bảo mật của hệ thống on-premise.

Việc sử dụng Databricks trong mô hình đề xuất giúp minh họa rõ ràng tính khả thi và hiệu quả của kiến trúc Data Lakehouse trong môi trường hybrid, đồng thời tạo tiền đề cho các nghiên cứu và triển khai thực tế trong lĩnh vực dữ liệu y tế lớn.

3.10.10. Docker – Công nghệ container hóa trong môi trường on-premise

Trong hệ thống Data Lakehouse được đề xuất trong luận văn, Docker được sử dụng như một công nghệ container hóa **chỉ áp dụng cho môi trường on-premise**, nhằm phục vụ việc triển khai, vận hành và quản lý các thành phần xử lý dữ liệu nội bộ. Việc giới hạn Docker trong phạm vi on-premise xuất phát từ yêu cầu kiểm soát hạ tầng, bảo mật dữ liệu y tế nhạy cảm và tính chủ động trong vận hành hệ thống.

Docker cho phép đóng gói các dịch vụ xử lý dữ liệu cùng với toàn bộ thư viện phụ thuộc, cấu hình và môi trường thực thi vào các container độc lập. Nhờ đó, các thành phần của hệ thống Data Lakehouse on-premise có thể được triển khai một cách nhất quán, giảm thiểu sự phụ thuộc vào cấu hình phần cứng và hệ điều hành cụ thể.

Vai trò của Docker trong kiến trúc on-premise

Trong môi trường on-premise, hệ thống dữ liệu y tế thường bao gồm nhiều thành phần mã nguồn mở khác nhau như Apache NiFi, Apache Airflow, Hive Metastore, Trino và các dịch vụ hỗ trợ lưu trữ, giám sát. Docker đóng vai trò như một lớp trừu tượng hóa hạ tầng, giúp các thành phần này được triển khai và vận hành độc lập nhưng vẫn có thể kết nối và phối hợp với nhau trong cùng một hệ thống.

Việc sử dụng Docker giúp đơn giản hóa quá trình cài đặt và cấu hình các dịch vụ phức tạp, đặc biệt trong bối cảnh hệ thống phải duy trì hoạt động liên tục 24/7. Các container có thể được khởi động, dừng hoặc cập nhật mà không ảnh hưởng đến toàn bộ hệ thống, góp phần nâng cao tính ổn định và khả năng bảo trì của kiến trúc on-premise.

Đảm bảo tính nhất quán và khả năng tái lập môi trường

Một trong những lợi ích quan trọng của Docker trong môi trường on-premise là khả năng đảm bảo tính nhất quán giữa các môi trường triển khai. Trong các hệ thống dữ liệu y tế, sự khác biệt nhỏ về phiên bản thư viện hoặc cấu hình hệ thống cũng có thể dẫn đến lỗi nghiêm trọng trong quá trình xử lý dữ liệu.

Docker cho phép định nghĩa môi trường thực thi một cách tường minh thông qua các image, giúp tái lập chính xác môi trường xử lý khi cần mở rộng hệ thống, khôi phục

sau sự cố hoặc triển khai trên phần cứng mới. Điều này đặc biệt quan trọng đối với các hệ thống lưu trữ và xử lý dữ liệu y tế yêu cầu độ tin cậy cao.

Hỗ trợ kiến trúc phân tán trong nội bộ hệ thống

Mặc dù Docker trong luận văn này không được sử dụng để triển khai các hệ thống cloud-native hay orchestration quy mô lớn, công nghệ container vẫn đóng vai trò quan trọng trong việc xây dựng kiến trúc phân tán ở mức nội bộ. Mỗi thành phần chức năng của hệ thống Data Lakehouse on-premise có thể được triển khai như một container riêng biệt, giúp tách biệt trách nhiệm và giảm sự phụ thuộc lẫn nhau giữa các dịch vụ.

Cách tiếp cận này giúp hệ thống dễ dàng mở rộng theo chiều ngang trong phạm vi on-premise bằng cách tăng số lượng container cho các dịch vụ có tải cao, chẳng hạn như các pipeline xử lý dữ liệu streaming hoặc các dịch vụ truy vấn phân tích.

Lý do không sử dụng Docker cho lớp cloud

Trong kiến trúc được đề xuất, lớp cloud được triển khai trên nền tảng Databricks – một dịch vụ quản lý hoàn chỉnh, nơi các vấn đề về container hóa, phân bổ tài nguyên và vận hành hạ tầng đã được trừu tượng hóa. Do đó, Docker không được sử dụng trực tiếp ở lớp cloud để tránh trùng lặp chức năng và tăng độ phức tạp không cần thiết.

Việc tách biệt rõ vai trò của Docker ở on-premise và nền tảng quản lý ở cloud giúp kiến trúc tổng thể trở nên rõ ràng, dễ quản lý và phù hợp với mô hình đám mây lai trong lĩnh vực y tế.

Ý nghĩa đối với hệ thống dữ liệu y tế

Việc sử dụng Docker trong môi trường on-premise mang lại nhiều lợi ích cho hệ thống dữ liệu y tế, bao gồm tăng tính ổn định, khả năng tái lập và chủ động trong vận hành. Công nghệ container hóa giúp hệ thống dễ dàng thích ứng với sự thay đổi về khối lượng dữ liệu và yêu cầu xử lý, đồng thời duy trì mức độ kiểm soát cao đối với dữ liệu nhạy cảm.

Tóm lại, Docker trong luận văn này được xem là công nghệ hỗ trợ triển khai và vận hành hệ thống Data Lakehouse on-premise một cách hiệu quả, đóng vai trò nền tảng cho sự ổn định và mở rộng của toàn bộ kiến trúc dữ liệu y tế trong môi trường đám mây lai.

4. NỘI DUNG VÀ PHƯƠNG PHÁP THỰC HIỆN

4.1. Phát biểu bài toán

Trong thực tế vận hành tại các bệnh viện quy mô trung bình ở Việt Nam, phòng hồi sức tích cực (ICU) thường có quy mô khoảng 20 giường bệnh và đóng vai trò trung tâm trong việc theo dõi, điều trị các ca bệnh nặng. Mỗi giường bệnh được trang bị các thiết bị theo dõi sinh hiệu, trong đó điện tâm đồ (ECG) là một trong những nguồn dữ liệu quan trọng nhất, yêu cầu thu thập và xử lý liên tục theo thời gian thực.

Trong kịch bản nghiên cứu của luận văn, hệ thống ICU gồm 20 máy ECG hoạt động độc lập, mỗi máy kết nối qua Bluetooth tới một thiết bị trung gian (transmitter). Các transmitter này có nhiệm vụ thu thập dữ liệu ECG thời gian thực và gửi dữ liệu về máy chủ MQTT thông qua mạng nội bộ của bệnh viện. Để đảm bảo tính liên tục của hệ thống, dữ liệu tại transmitter chỉ được lưu trữ tạm thời với dung lượng giới hạn và sẽ bị xóa khi bộ nhớ đầy. Do đó, yêu cầu bắt buộc đặt ra là mọi dữ liệu ECG phải được sao chép và lưu trữ đầy đủ tại hệ thống trung tâm nhằm tránh mất dữ liệu và đảm bảo khả năng truy xuất lịch sử khi cần thiết.

Dữ liệu ECG trong ICU có đặc tính là dữ liệu thời gian thực, tốc độ cao, phát sinh liên tục 24/7, với yêu cầu nghiêm ngặt về độ trễ ingest, độ ổn định của pipeline và khả năng mở rộng khi số lượng thiết bị tăng. Việc chỉ lưu trữ dữ liệu ở mức cục bộ hoặc dựa vào các cơ chế lưu trữ tạm thời tại thiết bị không đáp ứng được các yêu cầu về an toàn dữ liệu, truy vết nguồn gốc và phân tích dài hạn phục vụ nghiên cứu và cải tiến chất lượng điều trị.

Bên cạnh dữ liệu thời gian thực, hệ thống bệnh viện còn phát sinh lượng lớn dữ liệu hình ảnh y tế, đặc biệt là dữ liệu MRI. Trong thực tế hiện nay, dữ liệu MRI thường được lưu trữ tại các ổ đĩa chia sẻ nội bộ (Windows shared drive) để phục vụ nhu cầu truy cập nhanh trong nội bộ bệnh viện. Tuy nhiên, mô hình lưu trữ này tồn tại nhiều hạn chế: thiếu khả năng quản lý metadata, khó mở rộng khi dung lượng tăng nhanh, không hỗ trợ truy vết dữ liệu, không phù hợp cho phân tích quy mô lớn và tiềm ẩn rủi ro mất dữ liệu khi xảy ra sự cố phần cứng.

Do đó, dữ liệu MRI cần được thu thập theo cơ chế batch, định kỳ (ví dụ theo chu kỳ hàng giờ), và chuyển từ hệ thống shared drive nội bộ vào một hệ thống lưu trữ tập trung có khả năng mở rộng, quản lý dữ liệu có cấu trúc và hỗ trợ phân tích nâng cao. Việc chuyển đổi từ mô hình lưu trữ file truyền thống sang kiến trúc Data Lakehouse giúp chuẩn hóa quy trình ingest, đảm bảo tính toàn vẹn dữ liệu, hỗ trợ quản lý phiên bản và tạo nền tảng cho các tác vụ phân tích và huấn luyện mô hình trí tuệ nhân tạo trong tương lai.

Một yêu cầu quan trọng khác trong bối cảnh y tế tại Việt Nam là tuân thủ các quy định pháp luật liên quan đến bảo mật và chủ quyền dữ liệu. Toàn bộ dữ liệu y tế đầy đủ, bao gồm dữ liệu định danh bệnh nhân, phải được lưu trữ và kiểm soát trong hạ tầng nội bộ của bệnh viện. Hệ thống bên ngoài không được phép truy cập trực tiếp vào mạng nội bộ. Tuy nhiên, nhu cầu chia sẻ dữ liệu không nhạy cảm (đã được ẩn danh hoặc tổng hợp) lên môi trường đám mây là cần thiết để phục vụ các hoạt động nghiên cứu trí tuệ nhân tạo và hợp tác khoa học giữa các cơ sở nghiên cứu tại nhiều địa điểm khác nhau như TP. Hồ Chí Minh, Hà Nội, Vương quốc Anh và Hoa Kỳ.

Trong bối cảnh đó, bài toán đặt ra là làm thế nào để xây dựng một kiến trúc Data Lakehouse trên nền tảng đám mây lai, vừa đảm bảo lưu trữ đầy đủ và an toàn dữ liệu y tế trong môi trường nội bộ, vừa cho phép chia sẻ dữ liệu đã xử lý lên đám mây phục vụ nghiên cứu, đồng thời đáp ứng các yêu cầu về truy vết nguồn gốc dữ liệu, hiệu năng xử lý, độ tin cậy và khả năng mở rộng.

Luận văn này tập trung giải quyết bài toán trên thông qua việc thiết kế và triển khai một mô hình Data Lakehouse lai cho phòng ICU 20 giường bệnh, tích hợp cả pipeline xử lý dữ liệu ECG thời gian thực và pipeline xử lý dữ liệu MRI batch, nhằm đảm bảo dữ liệu được thu thập đầy đủ, chính xác, có thể kiểm soát và sẵn sàng cho các bài toán phân tích và nghiên cứu y sinh trong thực tế.

4.2. Mô tả dữ liệu sử dụng trong luận văn

Trong luận văn này, hai bộ dữ liệu y tế công khai được sử dụng nhằm mô phỏng các kịch bản xử lý dữ liệu đặc trưng trong hệ thống Data Lakehouse cho lĩnh vực chăm sóc sức khỏe, bao gồm dữ liệu thời gian thực (real-time) và dữ liệu batch dung lượng lớn. Việc lựa chọn các bộ dữ liệu này nhằm đảm bảo tính đa dạng về cấu trúc, kích thước, tần suất sinh dữ liệu cũng như phản ánh sát với các bài toán thực tế trong môi trường y tế.

4.2.1. Bộ dữ liệu ECG-ID (Dữ liệu điện tâm đồ thời gian thực)

Bộ dữ liệu ECG-ID được cung cấp bởi PhysioNet, là một tập dữ liệu điện tâm đồ phục vụ nghiên cứu nhận dạng sinh trắc học dựa trên tín hiệu ECG. Bộ dữ liệu bao gồm tổng cộng 310 bản ghi ECG thu thập từ 90 tình nguyện viên, với độ tuổi từ 13 đến 75, bao gồm cả nam và nữ.

Mỗi bản ghi ECG có các đặc điểm kỹ thuật như sau:

- Tín hiệu ECG đạo trình I, được ghi trong khoảng thời gian 20 giây.
- Tần số lấy mẫu 500 Hz, độ phân giải 12-bit trong dải ±10 mV.

- Mỗi bản ghi bao gồm hai loại tín hiệu: tín hiệu thô (raw signal) và tín hiệu đã được lọc (filtered signal).
- Mỗi bản ghi có kèm theo các annotation tự động cho các đỉnh sóng R và T, cùng với thông tin mô tả như tuổi, giới tính và thời điểm ghi nhận dữ liệu.

Đặc điểm nổi bật của bộ dữ liệu ECG-ID là dữ liệu có nhiễu, mang tính liên tục theo thời gian và có thể phát sinh với tần suất cao, tương tự như dữ liệu thu thập từ các thiết bị y tế IoT trong thực tế (máy đo điện tim, thiết bị theo dõi bệnh nhân). Do đó, bộ dữ liệu này được sử dụng trong luận văn để mô phỏng kịch bản ingest và xử lý dữ liệu thời gian thực trong kiến trúc Data Lakehouse, tập trung đánh giá các yếu tố như độ trễ (latency), thông lượng (throughput) và khả năng mở rộng của hệ thống.

4.2.2. Bộ dữ liệu MRI cột sống thắt lưng (Dữ liệu ảnh y tế batch)

Bên cạnh dữ liệu thời gian thực, luận văn sử dụng bộ dữ liệu Lumbar Spine MRI được công bố trên nền tảng Mendeley Data để đại diện cho kịch bản xử lý dữ liệu batch dung lượng lớn trong lĩnh vực y tế.

Bộ dữ liệu này bao gồm dữ liệu MRI đã được ẩn danh của 515 bệnh nhân có triệu chứng đau lưng, với tổng cộng 48.345 lát cắt ảnh MRI. Mỗi bệnh nhân có thể có một hoặc nhiều nghiên cứu MRI khác nhau, được chụp tại các thời điểm khác nhau.

Các đặc điểm chính của bộ dữ liệu bao gồm:

- Ảnh MRI ở hai mặt phẳng sagittal và axial.
- Độ phân giải phổ biến là 320×320 pixel, với độ sâu màu 12-bit cho mỗi pixel.
- Dữ liệu được lưu trữ dưới định dạng DICOM, là định dạng tiêu chuẩn trong lưu trữ và trao đổi ảnh y tế.
- Tổng dung lượng dữ liệu xấp xỉ 5.8 GB.

Bộ dữ liệu MRI này mang đặc trưng của dữ liệu y tế phi cấu trúc, dung lượng lớn, thời gian xử lý dài và yêu cầu cao về lưu trữ cũng như khả năng mở rộng. Trong luận văn, dữ liệu MRI được sử dụng để đánh giá hiệu năng của kiến trúc Data Lakehouse trong các tác vụ xử lý batch, bao gồm ingest dữ liệu lớn, tổ chức lưu trữ theo mô hình Medallion (Bronze–Silver–Gold) và tối ưu truy vấn phân tích.

4.2.3. Ý nghĩa của việc lựa chọn dữ liệu

Việc kết hợp hai bộ dữ liệu ECG-ID và MRI cho phép luận văn mô phỏng đầy đủ hai kịch bản phổ biến trong hệ thống dữ liệu y tế hiện đại:

- Xử lý dữ liệu thời gian thực từ các thiết bị y tế IoT.

- Xử lý dữ liệu ảnh y tế dung lượng lớn theo mô hình batch.

Qua đó, luận văn không chỉ đánh giá kiến trúc Data Lakehouse trên khía cạnh lý thuyết mà còn kiểm chứng khả năng áp dụng thực tế của mô hình trong môi trường đám mây lai, đáp ứng các yêu cầu đặc thù của lĩnh vực y tế như hiệu năng, khả năng mở rộng và quản lý dữ liệu phức tạp.

4.3 Kiến trúc hệ thống

4.3.1. Nguyên tắc thiết kế kiến trúc

Kiến trúc hệ thống Data Lakehouse trong nghiên cứu này được thiết kế dựa trên các nguyên tắc cốt lõi nhằm đảm bảo tính an toàn, khả thi và phù hợp với đặc thù dữ liệu y tế trong môi trường bệnh viện.

Thứ nhất, hệ thống ưu tiên tuyệt đối vấn đề bảo mật và chủ quyền dữ liệu y tế. Toàn bộ dữ liệu gốc, dữ liệu đầy đủ và dữ liệu định danh bệnh nhân được lưu trữ, xử lý và kiểm soát trong hạ tầng nội bộ của bệnh viện. Các hệ thống bên ngoài không có quyền truy cập trực tiếp vào mạng nội bộ, nhằm tuân thủ các quy định pháp luật hiện hành về bảo vệ dữ liệu y tế tại Việt Nam.

Thứ hai, kiến trúc phải hỗ trợ đồng thời cả hai mô hình xử lý dữ liệu là dữ liệu thời gian thực (streaming) và dữ liệu theo lô (batch). Điều này cho phép hệ thống tiếp nhận và xử lý liên tục dữ liệu IoT từ các thiết bị y tế, đồng thời xử lý định kỳ các tập dữ liệu lớn như hình ảnh MRI, đáp ứng nhu cầu vận hành thực tế của bệnh viện.

Thứ ba, hệ thống được thiết kế với khả năng mở rộng theo chiều ngang, cho phép tăng số lượng thiết bị thu thập dữ liệu và dung lượng lưu trữ theo thời gian mà không cần thay đổi kiến trúc tổng thể. Nguyên tắc này đặc biệt quan trọng trong bối cảnh dữ liệu y tế tăng nhanh cả về khối lượng lẫn đa dạng định dạng.

Thứ tư, kiến trúc phải hỗ trợ truy vết nguồn gốc dữ liệu (data lineage) và kiểm soát chất lượng dữ liệu trong toàn bộ vòng đời xử lý. Việc ghi nhận đầy đủ metadata, lịch sử xử lý và luồng di chuyển của dữ liệu giúp đảm bảo tính minh bạch, khả năng kiểm tra và độ tin cậy của dữ liệu y tế.

Thứ năm, chỉ những tập dữ liệu đã được xử lý, chuẩn hóa và ẩn danh mới được phép chia sẻ lên môi trường đám mây nhằm phục vụ nghiên cứu và phân tích nâng cao. Nguyên tắc này giúp cân bằng giữa yêu cầu bảo mật nội bộ và nhu cầu khai thác dữ liệu cho các mục đích khoa học và trí tuệ nhân tạo.

Thứ sáu, kiến trúc hướng đến hạn chế sự phụ thuộc vào nhà cung cấp (vendor lock-in) bằng cách ưu tiên sử dụng các công nghệ mã nguồn mở và chuẩn giao tiếp phổ biến. Điều này giúp hệ thống có khả năng chuyển đổi, mở rộng hoặc tích hợp với các nền tảng khác trong tương lai khi cần thiết.

Cuối cùng, hệ thống được xây dựng theo hướng dễ tương thích, thuận tiện cho vận hành và bảo trì. Kiến trúc microservice cho phép các thành phần hoạt động độc lập, giúp việc giám sát, khắc phục sự cố và mở rộng từng thành phần trở nên linh hoạt và hiệu quả hơn trong môi trường vận hành 24/7 của bệnh viện.

4.3.2. Thiết kế kiến trúc tổng thể hệ thống

Kiến trúc hệ thống được đề xuất trong luận văn được xây dựng theo mô hình Data Lakehouse trên nền tảng đám mây lai (Hybrid Cloud–On-Premise), nhằm đáp ứng đồng thời các yêu cầu về bảo mật dữ liệu y tế, xử lý dữ liệu thời gian thực, xử lý dữ liệu theo lô và khả năng mở rộng trong tương lai. Kiến trúc này kết hợp chặt chẽ giữa hạ tầng nội bộ của bệnh viện và môi trường đám mây phục vụ nghiên cứu, trong đó dữ liệu y tế gốc được kiểm soát hoàn toàn trong hệ thống on-premise, còn môi trường đám mây chỉ tham gia vào các giai đoạn phân tích và khai thác dữ liệu đã được xử lý.

Toàn bộ hệ thống được thiết kế tuân theo Medallion Architecture, bao gồm các lớp chính là Ingestion Layer, Storage Layer (Bronze, Silver, Gold) và Processing & Orchestration Layer. Cách tiếp cận này giúp tách biệt rõ ràng các giai đoạn thu thập, lưu trữ và xử lý dữ liệu, đồng thời hỗ trợ hiệu quả cho việc kiểm soát chất lượng, truy vết nguồn gốc dữ liệu và mở rộng hệ thống khi khối lượng dữ liệu tăng lên.

Về tổng thể, hạ tầng on-premise của bệnh viện đóng vai trò trung tâm trong việc thu thập, lưu trữ và xử lý dữ liệu y tế gốc. Môi trường đám mây không được phép truy cập trực tiếp vào mạng nội bộ mà chỉ tiếp nhận các tập dữ liệu đã được xử lý, tổng hợp và ẩn danh nhằm phục vụ các hoạt động phân tích và nghiên cứu trí tuệ nhân tạo.

Lớp Ingestion Layer chịu trách nhiệm tiếp nhận dữ liệu từ các nguồn phát sinh trong bệnh viện, bao gồm cả dữ liệu thời gian thực và dữ liệu theo cơ chế batch. Đối với dữ liệu ECG thời gian thực trong phòng ICU, mỗi máy ECG được kết nối qua Bluetooth tới một thiết bị transmitter. Các transmitter này thu thập tín hiệu ECG liên tục và truyền dữ liệu về máy chủ MQTT (Mosquitto) thông qua mạng nội bộ của bệnh viện. Giao thức MQTT được lựa chọn do có đặc tính nhẹ, độ trễ thấp và phù hợp với các hệ thống IoT y tế yêu cầu vận hành liên tục 24/7. Apache NiFi được triển khai như một consumer, subscribe vào các topic MQTT để tiếp nhận dữ liệu ECG, đồng thời thực hiện các bước xử lý ban đầu như chuẩn hóa định dạng, gắn metadata và kiểm soát luồng dữ liệu.

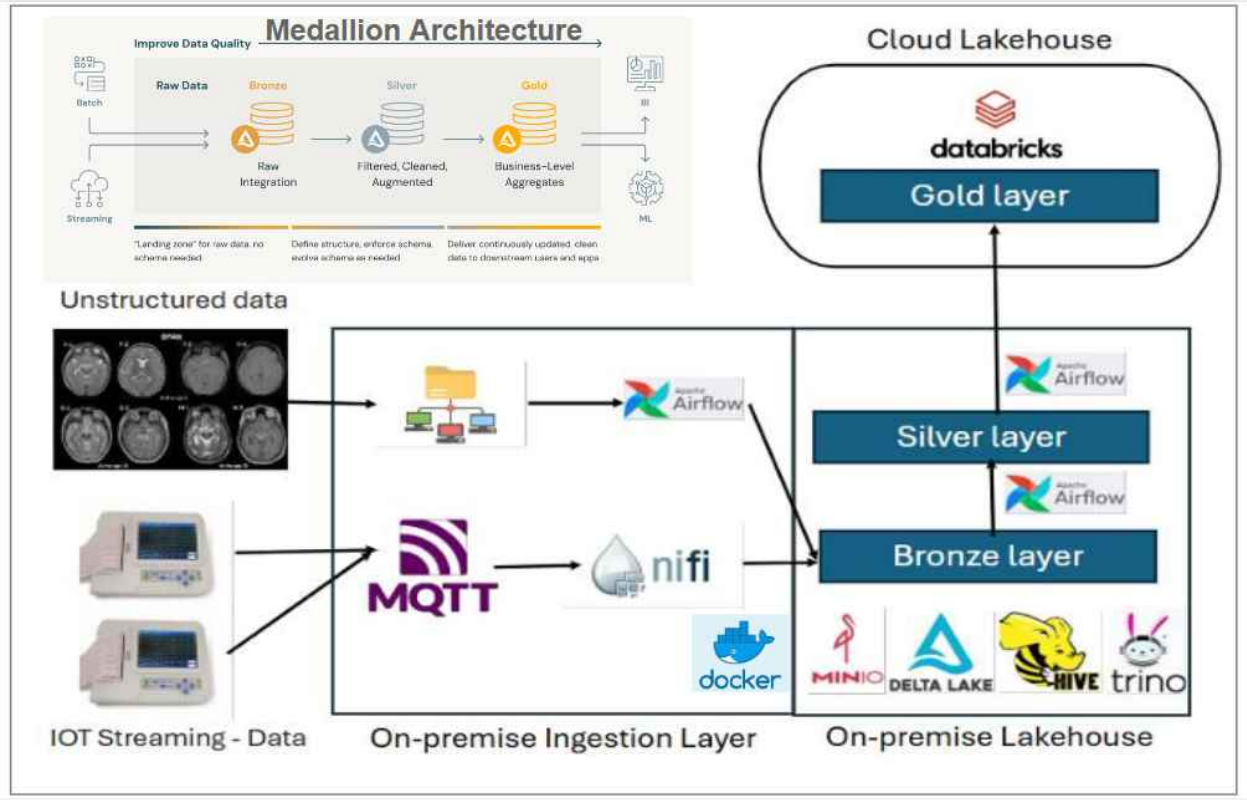
Đối với dữ liệu hình ảnh MRI, nguồn dữ liệu là các file DICOM được lưu trữ tại hệ thống Windows shared drive nội bộ. Trong kiến trúc đề xuất, Apache NiFi được sử dụng để thực hiện cơ chế ingest theo batch, định kỳ theo chu kỳ hàng giờ. Việc batch ingest giúp giảm tải cho hệ thống lưu trữ nội bộ, đồng thời đảm bảo dữ liệu MRI được đồng bộ đầy đủ và có kiểm soát vào hệ thống lưu trữ tập trung.

Sau khi được thu thập, dữ liệu được lưu trữ tại Storage Layer trong môi trường on-premise, sử dụng hệ thống lưu trữ đối tượng MinIO, đóng vai trò tương đương với Amazon S3 trong hạ tầng nội bộ. Tại Bronze Layer, dữ liệu được lưu trữ ở trạng thái thô, bao gồm dữ liệu ECG chưa qua làm sạch và các file MRI gốc. Lớp Bronze được thiết kế theo nguyên tắc lưu trữ đầy đủ dữ liệu gốc nhằm phục vụ các mục đích truy vết, kiểm tra và phục hồi dữ liệu khi cần thiết.

Tiếp theo, dữ liệu từ Bronze Layer được xử lý bằng Apache Spark kết hợp với Delta Lake để tạo thành Silver Layer. Tại lớp này, dữ liệu được làm sạch, chuẩn hóa schema, loại bỏ các bản ghi lỗi, bổ sung metadata và tổ chức phân vùng theo các tiêu chí phù hợp như thời gian, thiết bị hoặc bệnh nhân (đã được ẩn danh). Việc sử dụng Delta Lake giúp đảm bảo các đặc tính ACID, hỗ trợ quản lý phiên bản dữ liệu, schema evolution và time travel, những yếu tố đặc biệt quan trọng đối với dữ liệu y tế có yêu cầu cao về tính toàn vẹn và khả năng kiểm toán. Hive Metastore được tích hợp để quản lý metadata, cho phép các công cụ xử lý và truy vấn truy cập dữ liệu một cách nhất quán.

Processing & Orchestration Layer đóng vai trò điều phối và thực thi các pipeline xử lý dữ liệu trong toàn hệ thống. Apache Spark (Delta Spark) được sử dụng làm engine xử lý chính cho cả dữ liệu streaming và batch, trong đó Spark Structured Streaming đảm nhiệm xử lý luồng dữ liệu ECG thời gian thực, còn các Spark batch job được sử dụng để xử lý dữ liệu MRI. Apache Airflow được triển khai như một công cụ điều phối trung tâm, chịu trách nhiệm lập lịch, giám sát và tự động hóa toàn bộ pipeline từ ingest, xử lý, kiểm tra chất lượng đến đồng bộ dữ liệu. Việc sử dụng Airflow giúp hệ thống đáp ứng yêu cầu vận hành ổn định và liên tục 24/7, đồng thời hỗ trợ khôi phục khi xảy ra sự cố. Ngoài ra, Trino được triển khai như một lớp truy vấn phân tích, cho phép truy vấn trực tiếp trên Silver Layer mà không cần sao chép dữ liệu, phục vụ các nhu cầu phân tích nhanh trong nội bộ bệnh viện.

Gold Layer được triển khai trên nền tảng Databricks trong môi trường đám mây và chỉ tiếp nhận các tập dữ liệu đã được xử lý và ẩn danh từ Silver Layer nội bộ. Việc truyền dữ liệu từ on-premise lên đám mây được thực hiện thông qua các cơ chế bảo mật như JDBC hoặc REST API kết hợp với token xác thực, đảm bảo môi trường đám mây không thể truy cập trực tiếp vào mạng nội bộ của bệnh viện. Tại Gold Layer, dữ liệu được tổng hợp và tối ưu cho các tác vụ phân tích, báo cáo và huấn luyện mô hình trí tuệ nhân tạo. Mô hình này cho phép các nhóm nghiên cứu tại nhiều địa điểm khác nhau như TP. Hồ Chí Minh, Hà Nội, Vương quốc Anh và Hoa Kỳ khai thác dữ liệu một cách an toàn, đồng thời tuân thủ các quy định pháp luật về bảo mật và chủ quyền dữ liệu y tế tại Việt Nam.



Hình 6. Kiến trúc Data Lake House Hybrid

4.4. Luồng xử lý dữ liệu chi tiết cho dữ liệu ECG và MRI

Phần này trình bày chi tiết luồng dữ liệu từ lúc phát sinh tại thiết bị y tế cho đến khi dữ liệu được lưu trữ, xử lý và khai thác trong kiến trúc Data Lakehouse lai. Hai luồng dữ liệu chính được xem xét gồm: (1) dữ liệu ECG thời gian thực trong phòng ICU và (2) dữ liệu hình ảnh MRI được xử lý theo cơ chế batch.

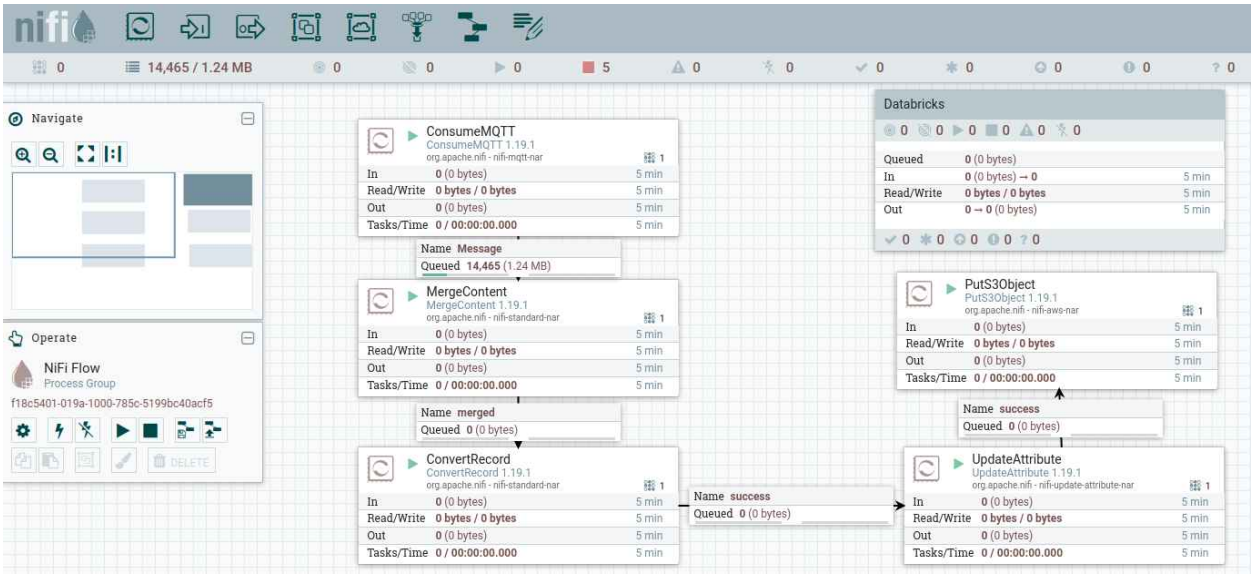
4.4.1 Luồng dữ liệu ECG thời gian thực

Dữ liệu ECG được phát sinh liên tục từ 20 máy ECG được lắp đặt tại phòng hồi sức tích cực (ICU), tương ứng với 20 giường bệnh. Mỗi máy ECG ghi nhận tín hiệu điện tim của bệnh nhân theo tần suất cao và tạo ra các bản ghi dữ liệu dạng chuỗi thời gian. Mỗi bản ghi ECG bao gồm các thông tin cơ bản như thời điểm đo, định danh thiết bị, giá trị biên độ tín hiệu ECG và thông tin chú thích nếu có.

Một bản ghi ECG tại nguồn có cấu trúc như sau: timestamp là thời điểm ghi nhận tín hiệu, device_name là định danh thiết bị hoặc giường bệnh, ecg là giá trị biên độ tín

hiệu điện tim và annotation là thông tin đánh dấu bổ sung, thường để trống trong điều kiện vận hành bình thường.

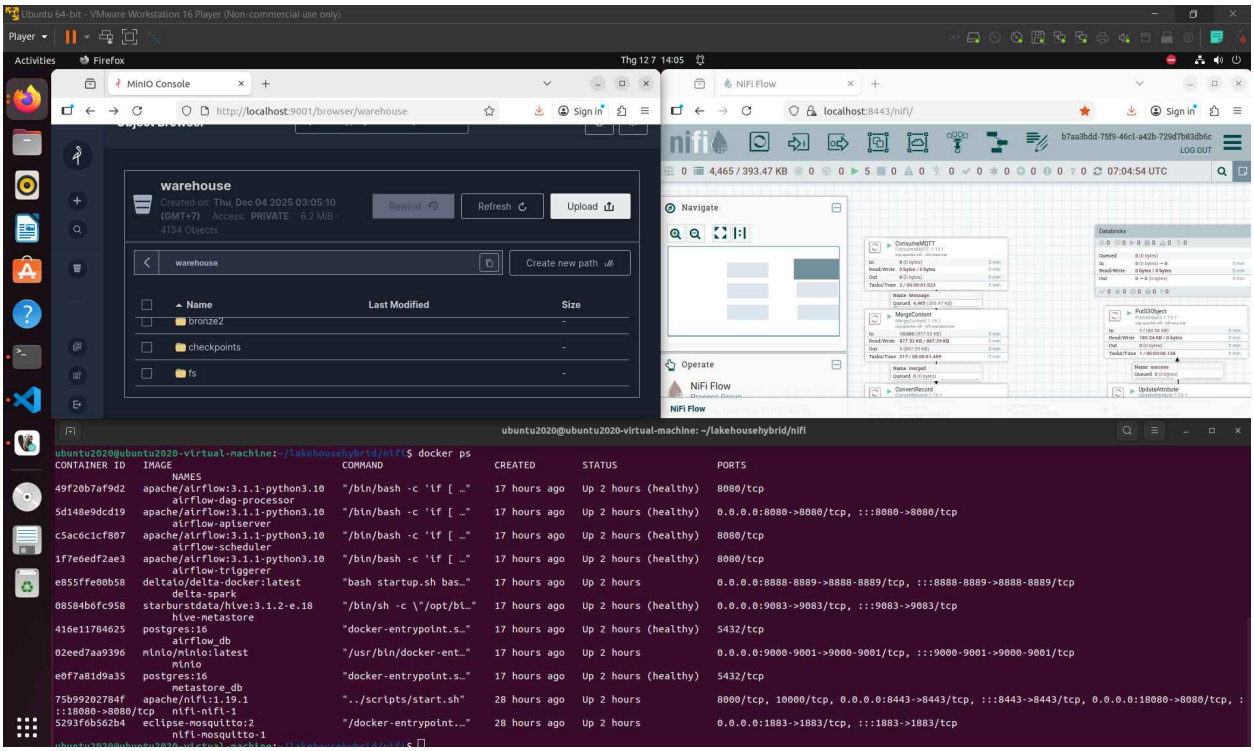
Tại tầng thiết bị, tín hiệu ECG được truyền từ máy đo đến các thiết bị transmitter thông qua kết nối Bluetooth. Các transmitter này đóng vai trò là lớp biên (edge layer), có nhiệm vụ thu thập dữ liệu từ thiết bị đo, tạm lưu dữ liệu trong bộ nhớ cục bộ để đảm bảo tính liên tục khi xảy ra gián đoạn mạng và gửi dữ liệu về hệ thống trung tâm ngay khi kết nối được khôi phục.



Hình 7. Luồng xử lý dữ liệu ECG bằng NIFI

Từ transmitter, dữ liệu ECG được truyền về máy chủ MQTT (Mosquitto) thông qua mạng nội bộ của bệnh viện. MQTT được cấu hình theo mô hình publish–subscribe, trong đó mỗi transmitter phát dữ liệu ECG lên các topic tương ứng với thiết bị hoặc giường bệnh. Cách tổ chức này giúp giảm độ trễ trong quá trình truyền dữ liệu và cho phép hệ thống dễ dàng mở rộng khi số lượng thiết bị ECG tăng trong tương lai.

Apache NiFi được triển khai trong hạ tầng nội bộ của bệnh viện như một thành phần tiếp nhận dữ liệu thời gian thực. NiFi subscribe vào các topic MQTT để thu thập dữ liệu ECG ngay khi dữ liệu được phát sinh. Tại NiFi, dữ liệu ECG được thực hiện các bước xử lý ban đầu nhằm đảm bảo tính nhất quán và khả năng truy vết, bao gồm việc chuẩn hóa định dạng dữ liệu, gắn thêm metadata như thời gian ingest và nguồn dữ liệu, đồng thời kiểm tra các lỗi cơ bản về cấu trúc dữ liệu.



Hình 8. Dữ liệu ECG được đặt trong Lakehouse On-premise Minio

Sau khi xử lý sơ bộ, dữ liệu ECG được ghi vào Bronze Layer trên hệ thống lưu trữ đối tượng MinIO. Tại Bronze Layer, dữ liệu được lưu trữ dưới dạng file Parquet và được phân vùng theo thời gian ingest. Bronze Layer lưu trữ dữ liệu ở trạng thái gần với dữ liệu gốc nhất, bao gồm cả các bản ghi trùng lặp hoặc chưa được làm sạch, nhằm đảm bảo khả năng truy vết nguồn gốc dữ liệu và phục hồi khi xảy ra lỗi ở các bước xử lý tiếp theo.

Tiếp theo, Apache Spark Structured Streaming đọc dữ liệu ECG từ Bronze Layer để xây dựng Silver Layer. Trong giai đoạn này, Spark thực hiện các bước xử lý chính bao gồm loại bỏ các bản ghi trùng lặp dựa trên timestamp và device_name, chuẩn hóa schema dữ liệu, kiểm tra và loại bỏ các bản ghi không hợp lệ, đồng thời chuẩn hóa trường thời gian theo chuẩn event time. Dữ liệu sau khi được làm sạch và chuẩn hóa được ghi vào bảng Delta Lake tại Silver Layer với các đặc tính ACID.

Dữ liệu ECG tại Silver Layer đã có cấu trúc rõ ràng, chất lượng ổn định và sẵn sàng cho các tác vụ truy vấn và phân tích. Từ Silver Layer, các tác vụ xử lý tiếp theo được thực hiện để xây dựng Gold Layer trên nền tảng Databricks Cloud. Trong giai đoạn này, dữ liệu ECG được ẩn danh hóa thông tin thiết bị, tổng hợp theo các cửa sổ thời gian phù hợp và trích xuất các đặc trưng cần thiết cho phân tích và huấn luyện mô hình trí tuệ nhân tạo.

Kết quả cuối cùng được lưu trữ trong bảng Delta Lake ecg_gold trên Databricks. Bảng ecg_gold đóng vai trò là bảng dữ liệu phân tích và nghiên cứu, chỉ chứa dữ liệu đã được xử lý và không bao gồm thông tin định danh bệnh nhân. Cách tiếp cận này đảm bảo tuân thủ các yêu cầu về bảo mật và chủ quyền dữ liệu y tế, đồng thời tạo nền tảng dữ liệu sẵn sàng cho các bài toán phân tích và nghiên cứu y sinh trong môi trường đám mây.

4.4.2 Luồng dữ liệu MRI theo cơ chế batch

Dữ liệu MRI được tạo ra từ hệ thống chụp cộng hưởng từ và ban đầu được lưu trữ tại các ổ đĩa chia sẻ nội bộ (Windows shared drive) dưới dạng file DICOM. Đây là mô hình lưu trữ phổ biến trong các bệnh viện do đáp ứng tốt yêu cầu truy cập nhanh, tương thích với hệ thống PACS hiện hữu và đảm bảo dữ liệu được kiểm soát trong mạng nội bộ.

Tuy nhiên, hình thức lưu trữ trên shared drive tồn tại nhiều hạn chế khi dữ liệu MRI tăng nhanh về dung lượng và số lượng nghiên cứu, bao gồm khó mở rộng, thiếu khả năng truy vết lịch sử xử lý, không hỗ trợ phân tích dữ liệu quy mô lớn và tiềm ẩn rủi ro mất dữ liệu khi xảy ra sự cố phần cứng. Do đó, để phục vụ lưu trữ dài hạn và các bài toán phân tích, dữ liệu MRI cần được chuyển vào hệ thống Data Lakehouse.

Apache Airflow được sử dụng để điều phối luồng ingest dữ liệu MRI theo cơ chế batch. Trong nghiên cứu này, Airflow được cấu hình thực hiện ingest theo chu kỳ định kỳ hàng giờ. Một DAG Airflow chịu trách nhiệm giám sát các thư mục trên Windows shared drive, phát hiện các file DICOM mới được tạo hoặc thay đổi kể từ lần chạy trước và thực hiện sao chép dữ liệu vào hệ thống lưu trữ đối tượng MinIO trong môi trường nội bộ.

Quá trình ingest do Airflow điều phối bao gồm các bước sau:

- Quét thư mục MRI trên shared drive dựa trên timestamp hoặc checksum để xác định dữ liệu mới.
- Sao chép các file DICOM nguyên gốc vào Bronze Layer trên MinIO, giữ nguyên cấu trúc và định dạng ban đầu.
- Ghi nhận metadata ingest như thời gian xử lý, nguồn dữ liệu, loại thiết bị chụp và mã nghiên cứu.

- Đảm bảo quá trình sao chép không làm gián đoạn hoạt động truy cập của các hệ thống lâm sàng đang sử dụng shared drive.

Sau khi dữ liệu MRI được lưu trữ tại Bronze Layer, Airflow tiếp tục điều phối các job xử lý batch bằng Apache Spark để xây dựng Silver Layer. Trong giai đoạn này, Spark thực hiện:

- Đọc và phân tích file DICOM để trích xuất các metadata cần thiết cho phân tích.
- Kiểm tra tính toàn vẹn của file và phát hiện dữ liệu lỗi hoặc không đầy đủ.
- Chuẩn hóa cấu trúc dữ liệu và tổ chức lưu trữ theo thời gian, loại nghiên cứu hoặc bộ phận lâm sàng.
- Lưu kết quả xử lý dưới định dạng Delta Lake nhằm hỗ trợ ACID, quản lý phiên bản và truy vết lịch sử thay đổi dữ liệu.

Cuối cùng, các tập dữ liệu MRI đã được xử lý, ẩn danh hoặc trích xuất đặc trưng có thể được đồng bộ lên Gold Layer trên nền tảng Databricks Cloud. Gold Layer phục vụ các bài toán phân tích hình ảnh y tế, huấn luyện mô hình trí tuệ nhân tạo và nghiên cứu đa trung tâm, trong khi vẫn đảm bảo dữ liệu định danh và dữ liệu gốc đầy đủ được lưu trữ an toàn trong hệ thống nội bộ, tuân thủ các quy định pháp luật về bảo vệ dữ liệu y tế tại Việt Nam.

5. PHÂN TÍCH CHIẾN LƯỢC BẢO MẬT, TUÂN THỦ DỮ LIỆU Y TẾ VÀ ĐÁNH GIÁ HỆ THỐNG

5.1. Chiến lược bảo mật và tuân thủ dữ liệu y tế

Trong lĩnh vực y tế, dữ liệu bệnh nhân được xem là loại dữ liệu đặc biệt nhạy cảm, yêu cầu mức độ bảo mật cao và tuân thủ nghiêm ngặt các quy định pháp luật về quyền riêng tư và chủ quyền dữ liệu. Do đó, trong kiến trúc Data Lakehouse đám mây lai được đề xuất, chiến lược bảo mật không được xem là một thành phần bổ trợ mà được tích hợp xuyên suốt trong toàn bộ quá trình thiết kế và vận hành hệ thống.

Nguyên tắc cốt lõi của chiến lược bảo mật trong nghiên cứu này là phân tách rõ ràng giữa dữ liệu y tế đầy đủ, dữ liệu định danh và dữ liệu phục vụ phân tích. Toàn bộ dữ liệu gốc, bao gồm dữ liệu sinh hiệu thời gian thực và dữ liệu hình ảnh y tế chứa thông tin định danh bệnh nhân, được lưu trữ và kiểm soát hoàn toàn trong hạ tầng nội bộ của bệnh viện. Môi trường đám mây không có quyền truy cập trực tiếp vào mạng nội bộ, cũng như không được phép truy vấn dữ liệu gốc chưa qua xử lý.

Về bảo mật dữ liệu trong quá trình truyền, tất cả các luồng dữ liệu từ thiết bị y tế đến hệ thống trung tâm đều được truyền trong mạng nội bộ của bệnh viện, hạn chế tối đa việc phơi lộ dữ liệu ra bên ngoài. Đối với các luồng đồng bộ dữ liệu từ hệ thống on-premise lên môi trường đám mây, việc truyền dữ liệu được thực hiện thông qua các kênh bảo mật như HTTPS hoặc JDBC có xác thực bằng token, đảm bảo dữ liệu không bị nghe lén hoặc can thiệp trong quá trình truyền.

Về bảo mật dữ liệu khi lưu trữ, hệ thống sử dụng cơ chế phân vùng lưu trữ theo mô hình Medallion. Bronze Layer lưu trữ dữ liệu thô và dữ liệu gốc, chỉ cho phép các thành phần xử lý nội bộ truy cập. Silver Layer chứa dữ liệu đã được làm sạch và chuẩn hóa, nhưng vẫn nằm trong hạ tầng nội bộ. Gold Layer trên môi trường đám mây chỉ tiếp nhận dữ liệu đã được ẩn danh, tổng hợp hoặc trích xuất đặc trưng, không còn khả năng truy ngược trực tiếp về bệnh nhân cụ thể.

Chiến lược phân quyền truy cập được triển khai theo nguyên tắc phân quyền tối thiểu (least privilege). Mỗi nhóm người dùng, bao gồm nhân viên kỹ thuật, nhà phân tích dữ liệu và nhóm nghiên cứu, chỉ được cấp quyền truy cập đúng với vai trò và phạm vi công việc của mình. Việc quản lý metadata và schema thông qua Hive Metastore giúp kiểm soát quyền truy cập ở mức bảng và cột dữ liệu, hạn chế nguy cơ lộ thông tin nhạy cảm.

Ngoài ra, hệ thống hỗ trợ truy vết nguồn gốc dữ liệu thông qua việc ghi nhận metadata, lịch sử xử lý và phiên bản dữ liệu trong Delta Lake. Cơ chế này cho phép

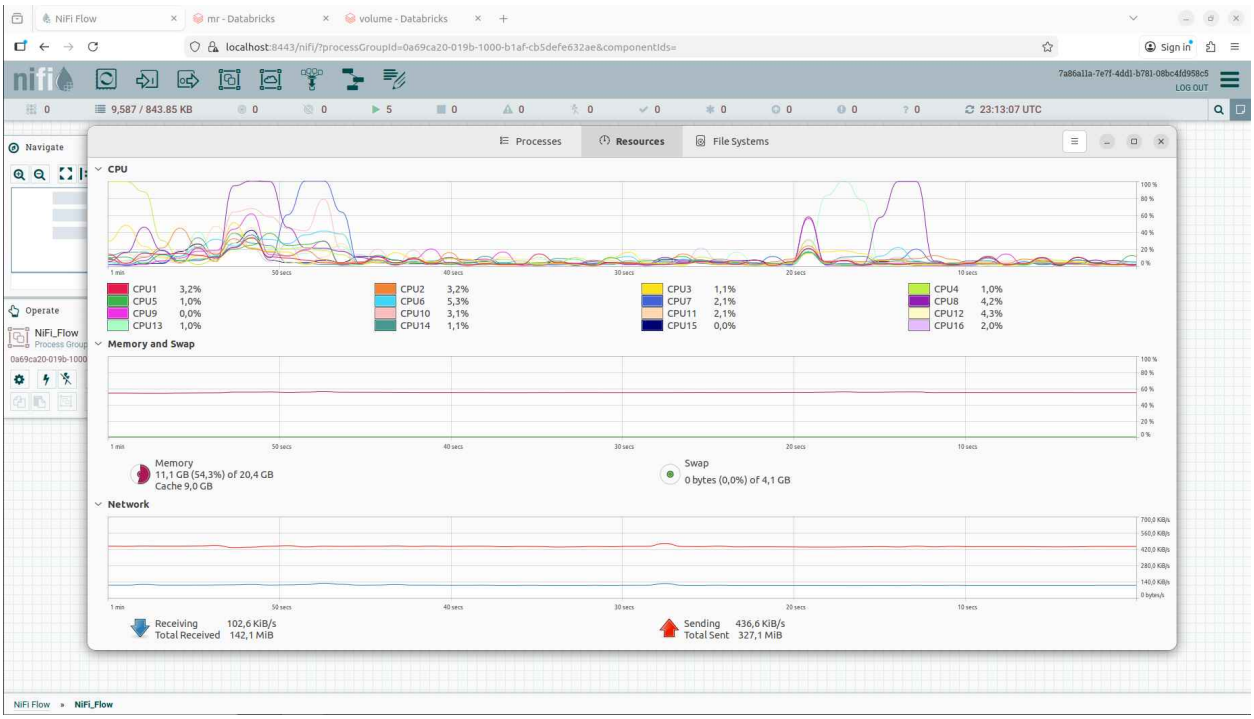
truy xuất lại quá trình hình thành của mỗi tập dữ liệu, phục vụ công tác kiểm toán, điều tra sự cố và đảm bảo tính minh bạch trong quản lý dữ liệu y tế.

Xét trên phương diện kiến trúc, mô hình Data Lakehouse đám mây lai được đề xuất đáp ứng các nguyên tắc cốt lõi của các tiêu chuẩn bảo vệ dữ liệu phổ biến như HIPAA và GDPR ở mức thiết kế hệ thống, bao gồm kiểm soát truy cập, phân tách dữ liệu định danh, bảo vệ dữ liệu khi truyền và khi lưu trữ, cũng như đảm bảo quyền kiểm soát dữ liệu thuộc về tổ chức sở hữu dữ liệu.

5.2. Mục tiêu và phương pháp đánh giá hệ thống

Mục tiêu của phần đánh giá là kiểm chứng khả năng áp dụng thực tế của kiến trúc Data Lakehouse đám mây lai trong bối cảnh dữ liệu y tế, thông qua các tiêu chí định lượng và định tính. Việc đánh giá tập trung vào ba khía cạnh chính: hiệu năng xử lý dữ liệu, độ ổn định của hệ thống và khả năng đáp ứng các yêu cầu bảo mật và tuân thủ dữ liệu.

Hai kịch bản đánh giá chính được thực hiện tương ứng với hai loại dữ liệu trong luận văn, bao gồm xử lý dữ liệu ECG thời gian thực và xử lý dữ liệu MRI theo cơ chế batch. Các thử nghiệm được triển khai trên môi trường mô phỏng hạ tầng bệnh viện quy mô trung bình, phù hợp với kịch bản ICU 20 giường bệnh.



Hình 9. Đánh giá khả năng chịu tải của hệ thống

Các chỉ số đánh giá chính bao gồm độ trễ ingest (ingestion latency), thông lượng xử lý (throughput), độ ổn định của pipeline trong quá trình vận hành liên tục và khả năng mở rộng khi tăng khối lượng dữ liệu.

Cấu hình máy thử nghiệm:

- Bộ xử lý (CPU): Intel(R) Core(TM) i7-14700HX, xung nhịp 2.10 GHz, 16 processors
- Bộ nhớ (RAM): 20 GB
- Ổ cứng (Storage): 100 GB

Kịch bản đánh giá	Số luồng ECG đồng thời	đồng thờiMức sử dụng RAM trung bình (%)	Mức sử dụng CPU trung bình (%)	Nhận xét
Kịch bản 1	1 luồng	~60%	~50%	Hệ thống hoạt động ổn định, còn dư tài nguyên

Kịch bản 2	20 luồng	~80%	~95%	CPU đạt ngưỡng bão hòa, hệ thống bắt đầu chịu tải cao

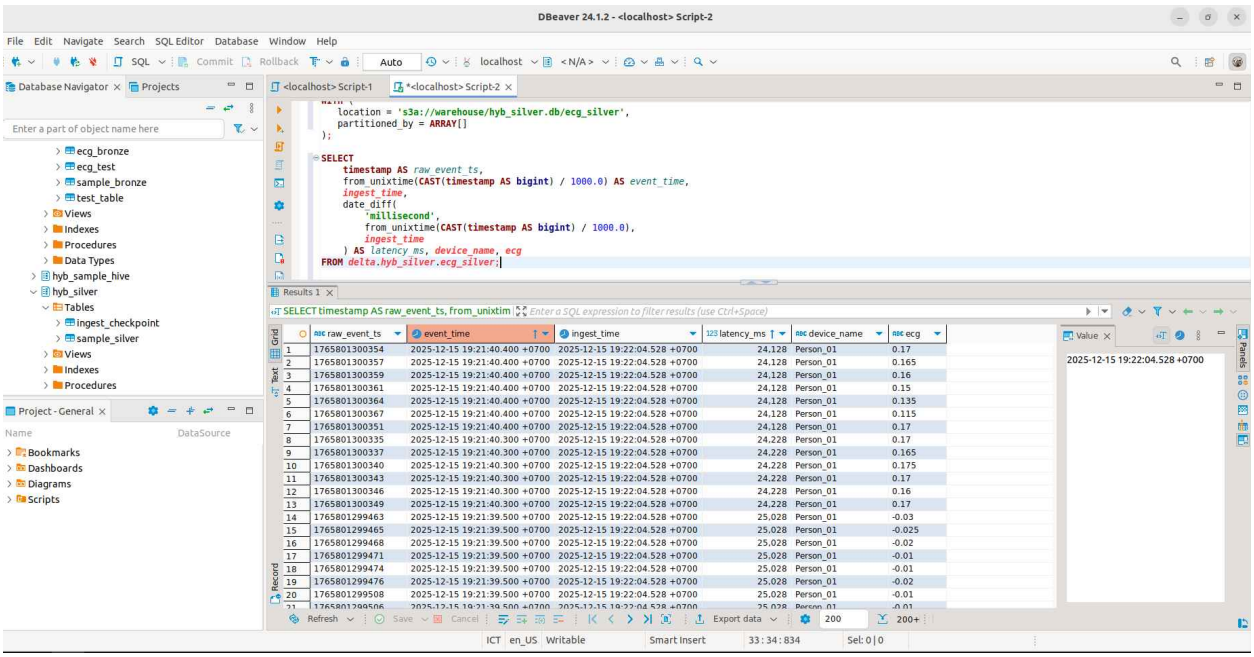
Bảng 1. Đánh giá khả năng chịu tải của hệ thống

Bảng số liệu trình bày kết quả đo đạc mức sử dụng tài nguyên hệ thống trong quá trình xử lý dữ liệu ECG thời gian thực với các kịch bản số luồng khác nhau. Khi số luồng ECG tăng từ 10 lên 20, mức sử dụng RAM tăng từ khoảng 60% lên 80%, trong khi mức sử dụng CPU tăng mạnh từ khoảng 50% lên gần 95%. Điều này cho thấy pipeline xử lý ECG thời gian thực có khả năng mở rộng tốt ở mức tải trung bình, tuy nhiên khi số luồng tăng cao, CPU trở thành tài nguyên giới hạn chính của hệ thống. Kết quả này phù hợp với đặc thù của các tác vụ xử lý streaming yêu cầu tính toán liên tục và độ trễ thấp.

5.3. Đánh giá pipeline xử lý dữ liệu ECG thời gian thực

Đối với dữ liệu ECG, hệ thống được đánh giá trong kịch bản 20 thiết bị ECG phát sinh dữ liệu liên tục với tần suất cao. Dữ liệu được publish từ các transmitter lên MQTT và được Apache NiFi tiếp nhận theo thời gian thực.

Kết quả thử nghiệm cho thấy độ trễ trung bình từ thời điểm dữ liệu ECG được phát sinh tại transmitter đến khi dữ liệu được ghi vào Bronze Layer dao động trong khoảng vài giây, đáp ứng yêu cầu xử lý gần thời gian thực trong môi trường ICU. Spark Structured Streaming xử lý dữ liệu ổn định trong điều kiện luồng dữ liệu liên tục, không ghi nhận tình trạng backlog kéo dài hay mất dữ liệu trong suốt quá trình thử nghiệm.



Hình 10. Kết quả triển khai dữ liệu ECG

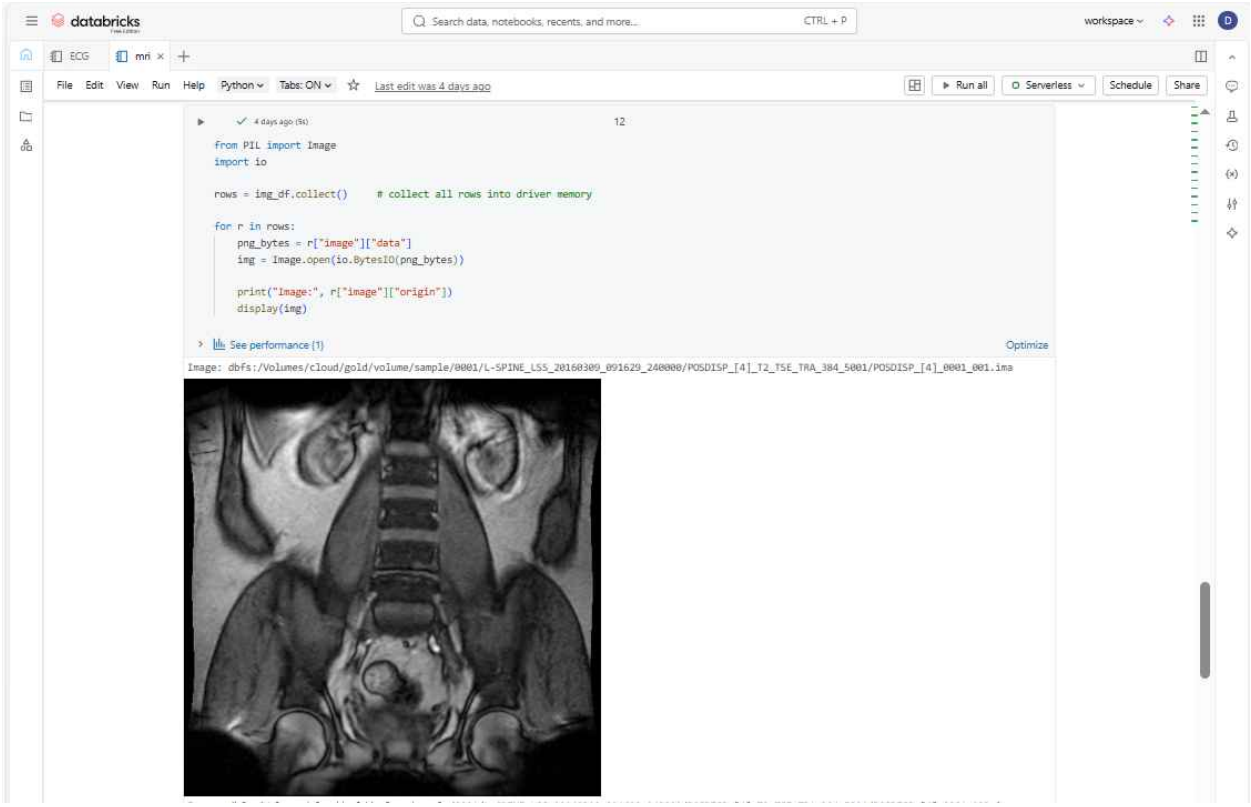
Thông lượng xử lý của pipeline cho phép tiếp nhận đồng thời dữ liệu từ nhiều thiết bị ECG mà không cần thay đổi kiến trúc. Việc phân vùng dữ liệu theo thời gian và thiết bị giúp hệ thống duy trì hiệu năng truy vấn ổn định tại Silver Layer. Gold Layer trên Databricks cho phép thực hiện các truy vấn phân tích tổng hợp với thời gian phản hồi phù hợp cho các tác vụ nghiên cứu và phân tích.

Kết quả cho thấy kiến trúc đề xuất đáp ứng tốt các yêu cầu về độ trễ, tính liên tục và khả năng mở rộng đối với dữ liệu ECG thời gian thực trong môi trường ICU.

5.4. Đánh giá pipeline xử lý dữ liệu MRI theo cơ chế batch

Đối với dữ liệu MRI, pipeline batch được đánh giá thông qua việc ingest toàn bộ tập dữ liệu DICOM với dung lượng xấp xỉ 5.8 GB theo chu kỳ định kỳ. Quá trình ingest do Apache Airflow điều phối cho thấy khả năng phát hiện và sao chép dữ liệu mới một cách ổn định, không ảnh hưởng đến hoạt động truy cập của hệ thống shared drive nội bộ.

Thời gian ingest batch hàng giờ phù hợp với đặc thù dữ liệu MRI, vốn không yêu cầu xử lý thời gian thực. Apache Spark xử lý batch dữ liệu MRI ổn định, cho phép trích xuất metadata, kiểm tra tính toàn vẹn file và tổ chức dữ liệu vào Silver Layer dưới định dạng Delta Lake.



Hình 11. Kết quả triển khai dữ liệu hình ảnh MRI trên Databricks

Việc lưu trữ dữ liệu MRI trong Data Lakehouse giúp cải thiện đáng kể khả năng truy vấn metadata và hỗ trợ các tác vụ phân tích quy mô lớn so với mô hình lưu trữ truyền thống trên shared drive. Gold Layer trên Databricks cho phép các nhóm nghiên cứu truy cập dữ liệu MRI đã được xử lý và ẩn danh một cách thuận tiện, đồng thời đảm bảo dữ liệu gốc vẫn được bảo vệ trong hạ tầng nội bộ.

5.5. Đánh giá tổng hợp và thảo luận

Kết quả đánh giá cho thấy kiến trúc Data Lakehouse đám mây lai được đề xuất có khả năng đáp ứng đồng thời các yêu cầu khắt khe về hiệu năng, bảo mật và khả năng mở rộng trong lĩnh vực y tế. Hệ thống xử lý hiệu quả cả dữ liệu thời gian thực và dữ liệu batch dung lượng lớn, đồng thời đảm bảo dữ liệu y tế được kiểm soát chặt chẽ trong hạ tầng nội bộ.

So với các mô hình chỉ sử dụng on-premise hoặc cloud thuần túy, kiến trúc lai mang lại sự cân bằng giữa bảo mật, hiệu năng và khả năng hợp tác nghiên cứu. Điều này cho thấy tính khả thi và tiềm năng ứng dụng thực tế của mô hình Data Lakehouse đám mây lai trong các bệnh viện quy mô trung bình tại Việt Nam.

6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1. Kết luận

Luận văn đã tập trung nghiên cứu và đề xuất một mô hình Data Lakehouse trên nền tảng đám mây lai nhằm phục vụ lưu trữ và xử lý dữ liệu y tế lớn trong bối cảnh các hệ thống bệnh viện hiện đại ngày càng phải đối mặt với khối lượng dữ liệu tăng nhanh, đa dạng về định dạng và yêu cầu nghiêm ngặt về bảo mật. Kết quả nghiên cứu được công khai mã nguồn mở tại trang Github: https://github.com/PhongDinhCS/lakehousehybrid.

Trên cơ sở phân tích các đặc thù của dữ liệu y tế, bao gồm dữ liệu sinh hiệu thời gian thực, dữ liệu hình ảnh y tế dung lượng lớn và dữ liệu hồ sơ bệnh nhân có cấu trúc, luận văn đã chỉ ra những hạn chế của các mô hình lưu trữ và xử lý truyền thống, cũng như các kiến trúc chỉ thuần on-premise hoặc thuần cloud. Từ đó, mô hình Data Lakehouse đám mây lai được đề xuất như một hướng tiếp cận phù hợp nhằm cân bằng giữa yêu cầu bảo mật, hiệu năng xử lý và khả năng mở rộng.

Luận văn đã xây dựng một kiến trúc tổng thể cho hệ thống Data Lakehouse đám mây lai, trong đó hạ tầng on-premise đóng vai trò trung tâm trong việc thu thập, lưu trữ và xử lý dữ liệu y tế gốc, còn môi trường đám mây được sử dụng để mở rộng khả năng phân tích, khai thác và nghiên cứu trên dữ liệu đã được xử lý và ẩn danh. Việc phân tách rõ ràng giữa hai tầng này giúp đảm bảo chủ quyền dữ liệu y tế, đồng thời tận dụng được sức mạnh tính toán và hệ sinh thái phân tích của nền tảng đám mây.

Về mặt triển khai, luận văn đã thiết kế và mô phỏng thành công các pipeline xử lý dữ liệu theo mô hình Medallion Architecture, bao gồm Bronze, Silver và Gold Layer. Các pipeline này cho phép xử lý hiệu quả cả dữ liệu thời gian thực thông qua cơ chế streaming và dữ liệu dung lượng lớn thông qua cơ chế batch. Việc ứng dụng các công nghệ như MQTT, Apache NiFi, Apache Spark Structured Streaming, Delta Lake và Databricks đã chứng minh tính khả thi của mô hình trong môi trường thực tế.

Kết quả đánh giá cho thấy hệ thống đáp ứng tốt các yêu cầu về độ trễ trong xử lý dữ liệu ECG thời gian thực, đảm bảo tính liên tục và ổn định trong quá trình vận hành. Đồng thời, pipeline batch xử lý dữ liệu hình ảnh y tế cho thấy khả năng quản lý và khai thác hiệu quả các tập dữ liệu dung lượng lớn mà không ảnh hưởng đến hoạt động thường nhật của hệ thống nội bộ. Chiến lược bảo mật và phân quyền được tích hợp xuyên suốt kiến trúc giúp đảm bảo an toàn dữ liệu và phù hợp với các nguyên tắc cơ bản của các tiêu chuẩn bảo vệ dữ liệu y tế hiện hành.

Từ các kết quả đạt được, có thể khẳng định rằng mô hình Data Lakehouse đám mây lai được đề xuất trong luận văn là một giải pháp khả thi và có giá trị thực tiễn cao cho bài toán lưu trữ và xử lý dữ liệu y tế lớn tại các cơ sở y tế, đặc biệt là các bệnh viện có yêu cầu cao về bảo mật nhưng vẫn mong muốn tận dụng sức mạnh của công nghệ đám mây cho các hoạt động phân tích và nghiên cứu.

6.2. Đóng góp khoa học của luận văn

Luận văn này đóng góp một số kết quả có ý nghĩa khoa học và thực tiễn trong lĩnh vực lưu trữ và xử lý dữ liệu y tế lớn, cụ thể như sau:

Thứ nhất, luận văn đề xuất một mô hình kiến trúc Data Lakehouse trên nền tảng đám mây lai, được thiết kế và đặc thù hóa cho bài toán dữ liệu y tế. Kiến trúc này phân tách rõ ràng vai trò giữa hạ tầng nội bộ của bệnh viện và môi trường đám mây, trong đó toàn bộ dữ liệu y tế gốc và dữ liệu định danh được lưu trữ, xử lý trong hệ thống on-premise, còn môi trường đám mây chỉ tiếp nhận các tập dữ liệu đã được xử lý và ẩn danh. Cách tiếp cận này góp phần chuẩn hóa mô hình kiến trúc Data Lakehouse phù hợp với các yêu cầu về bảo mật, chủ quyền dữ liệu và tuân thủ pháp lý trong lĩnh vực y tế.

Thứ hai, luận văn đề xuất một khung phân loại dữ liệu y tế dựa trên đặc tính phát sinh và phương thức xử lý, bao gồm dữ liệu thời gian thực từ các thiết bị y tế IoT, dữ liệu hình ảnh y tế dung lượng lớn xử lý theo cơ chế batch và dữ liệu hồ sơ bệnh nhân có tính định danh. Trên cơ sở đó, luận văn xây dựng các chiến lược ingest, lưu trữ và xử lý tương ứng cho từng loại dữ liệu trong cùng một kiến trúc Data Lakehouse thống nhất, góp phần làm rõ cách áp dụng mô hình Medallion Architecture cho các kịch bản dữ liệu y tế đa dạng.

Thứ ba, luận văn cung cấp minh chứng thực nghiệm cho khả năng tích hợp và vận hành đồng thời các pipeline xử lý dữ liệu thời gian thực và dữ liệu theo lô trong kiến trúc Data Lakehouse. Thông qua việc triển khai pipeline xử lý dữ liệu ECG thời gian thực và pipeline xử lý dữ liệu MRI theo cơ chế batch, luận văn chứng minh rằng mô hình Lakehouse có thể đáp ứng các yêu cầu về độ trễ, tính toàn vẹn dữ liệu và khả năng mở rộng trong bối cảnh dữ liệu y tế đa nguồn, đa định dạng.

Thứ tư, luận văn góp phần vào nghiên cứu về quản trị dữ liệu y tế thông qua việc áp dụng các nguyên tắc quản lý vòng đời dữ liệu, truy vết nguồn gốc dữ liệu và kiểm soát chất lượng dữ liệu trong mô hình Bronze–Silver–Gold. Việc sử dụng Delta Lake và hệ thống quản lý metadata giúp đảm bảo tính nhất quán, khả năng kiểm toán và độ tin cậy của dữ liệu y tế trong toàn bộ quá trình xử lý.

Cuối cùng, về mặt phương pháp nghiên cứu, luận văn kết hợp giữa thiết kế kiến trúc, triển khai hệ thống và đánh giá thực nghiệm trên các bộ dữ liệu y tế công khai.

Cách tiếp cận này góp phần cung cấp một mô hình tham chiếu cho việc nghiên cứu và triển khai các hệ thống Data Lakehouse trong lĩnh vực y tế, đặc biệt trong bối cảnh đám mây lai tại các bệnh viện quy mô vừa và lớn ở Việt Nam.

6.3. Hạn chế của nghiên cứu

Mặc dù đã đạt được các mục tiêu đề ra, luận văn vẫn còn tồn tại một số hạn chế nhất định. Thứ nhất, các thử nghiệm được thực hiện trên môi trường mô phỏng với quy mô giới hạn, chưa phản ánh đầy đủ các kịch bản phức tạp trong môi trường bệnh viện lớn với hàng trăm hoặc hàng nghìn thiết bị y tế hoạt động đồng thời. Thứ hai, các đánh giá về hiệu năng chủ yếu tập trung vào các chỉ số cơ bản như độ trễ và thông lượng, chưa đi sâu vào phân tích chi phí vận hành dài hạn hoặc tối ưu hóa tài nguyên ở quy mô lớn. Ngoài ra, các khía cạnh liên quan đến quản lý vòng đời dữ liệu và tích hợp sâu với các hệ thống thông tin bệnh viện hiện hữu vẫn chưa được triển khai đầy đủ trong phạm vi nghiên cứu.

6.4. Hướng phát triển trong tương lai

Trong thời gian tới, mô hình Data Lakehouse đám mây lai được đề xuất có thể tiếp tục được mở rộng và hoàn thiện theo nhiều hướng khác nhau. Một trong những hướng phát triển quan trọng là mở rộng quy mô hệ thống để đánh giá khả năng xử lý trong môi trường bệnh viện lớn, đồng thời nghiên cứu các cơ chế tự động mở rộng tài nguyên nhằm tối ưu chi phí vận hành.

Bên cạnh đó, việc tích hợp các mô hình học máy và trí tuệ nhân tạo vào tầng phân tích, đặc biệt là các mô hình phát hiện bất thường trên dữ liệu ECG hoặc phân tích hình ảnh y tế, sẽ góp phần nâng cao giá trị ứng dụng của hệ thống. Ngoài ra, các cơ chế quản lý metadata nâng cao, kiểm soát chất lượng dữ liệu tự động và giám sát pipeline theo thời gian thực cũng là những hướng nghiên cứu tiềm năng.

Cuối cùng, việc nghiên cứu sâu hơn về các tiêu chuẩn và quy định pháp lý liên quan đến dữ liệu y tế tại Việt Nam, cũng như khả năng tích hợp mô hình với các hệ thống y tế quốc gia, sẽ giúp tăng tính ứng dụng thực tế và khả năng triển khai rộng rãi của mô hình Data Lakehouse đám mây lai trong tương lai.

7. TÀI LIỆU THAM KHẢO

[1]. Michael Armbrust, Ali Ghodsi, Reynold Xin, Matei Zaharia. (2021). Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. CIDR ’21, Jan. 2021, Online

[2]. Cherradi, Mohamed & El Haddadi, Anass. (2024). Data Lakehouse: Next Generation Information System. Seminars in Medical Writing and Education. 3. 10.56294/mw202467.

[3]. Aziz, Farooq & Business, C. (2024). Next-Generation Healthcare Analytics: The Open Lakehouse Framework. INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS. 11. 94-105. 10.2139/ssrn.5065660.

[4]. Orescanin, Drazen & Hlupic, Tomislav. (2021). Data Lakehouse - a Novel Step in Analytics Architecture. 1242-1246. 10.23919/MIPRO52101.2021.9597091.

[5]. Chaudhari, Akash Vijayrao & Charate, Pallavi. (2025). Optimizing Data Lakehouse Architectures for Scalable Real-Time Analytics. International Journal of Scientific Research in Science Engineering and Technology. 12. 809-822. 10.32628/IJSRSET25122198.

[6]. Begoli, Edmon & Goethert, Ian & Knight, Kathryn. (2021). A Lakehouse Architecture for the Management and Analysis of Heterogeneous Data for Biomedical Research and Mega-biobanks. 4643-4651. 10.1109/BigData52589.2021.9671534.

[7]. Ponnekanti, Sai. (2025). The Evolution from Data Warehouses to Data Lakehouses: A Technical Perspective. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 11. 2248-2263. 10.32628/CSEIT25112711.

[8]. Islam, Ashraful. (2024). HYBRID CLOUD DATABASES FOR BIG DATA ANALYTICS: A REVIEW OF ARCHITECTURE, PERFORMANCE, AND COST EFFICIENCY. 96-114.

[9]. Bello, Sadis & Brown, James. (2024). HYBRID CLOUD STRATEGIES: BALANCING CONTROL AND FLEXIBILITY IN SENSITIVE INDUSTRIES.

[10]. M Anderson, G Gershinsky, E Salant, S Garcia (2023) Protecting Sensitive Tabular Data in Hybrid Clouds. arXiv preprint arXiv:2312.01354

[11]. Harby, Ahmed & Zulkernine, Farhana. (2022). From Data Warehouse to Lakehouse: A Comparative Review. 10.1109/BigData55660.2022.10020719.

[12]. Schneider, Jan & Gröger, Christoph & Lutsch, Arnold & Schwarz, Holger & Mitschang, Bernhard. (2024). The Lakehouse: State of the Art on Concepts and Technologies. SN Computer Science. 5. 10.1007/s42979-024-02737-0.

[13]. Guo, Chonghui & Chen, Jingfeng. (2023). Big Data Analytics in Healthcare. 10.1007/978-981-99-1075-5_2.

[14]. Theriault-Lauzier, Pascal & Cobin, Denis & Tastet, Olivier & Langlais, Elodie & Taji, Bahareh & Kang, Guson & Chong, Aun-Yeong & So, Derek & Tang, An & Gichoya, Judy & Chandar, Sarath & Déziel, Pierre-Luc & Hussin, Julie & Kadoury, Samuel & Avram, Robert. (2024). A Responsible Framework for Applying Artificial Intelligence on Medical Images and Signals at the Point of Care: The PACS-AI Platform. Canadian Journal of Cardiology. 40. 10.1016/j.cjca.2024.05.025.

[15]. Ahmed, Awais & Xi, Rui & Hou, Mengshu & Shah, Syed & Hameed, Sufian. (2023). Harnessing Big Data Analytics for Healthcare: A Comprehensive Review of Frameworks, Implications, Applications, and Impacts. IEEE Access. PP. 1-1. 10.1109/ACCESS.2023.3323574.