

BÁO CÁO LUẬN VĂN THẠC SĨ

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

GVHD1: PGS.TS Thoại Nam

GVHD2: TS. Nguyễn Lê Duy Lai

Học viên thực hiện:

Đinh Thanh Phong - 2270243

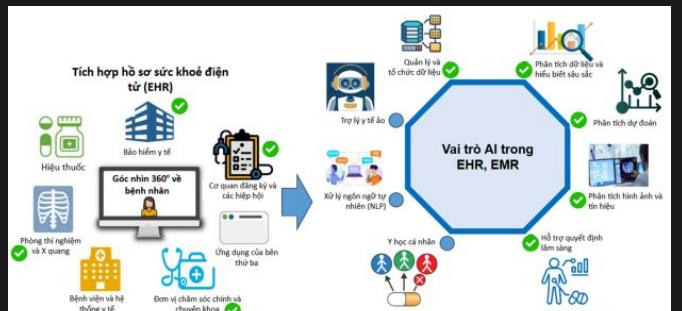
Nội dung trình bày



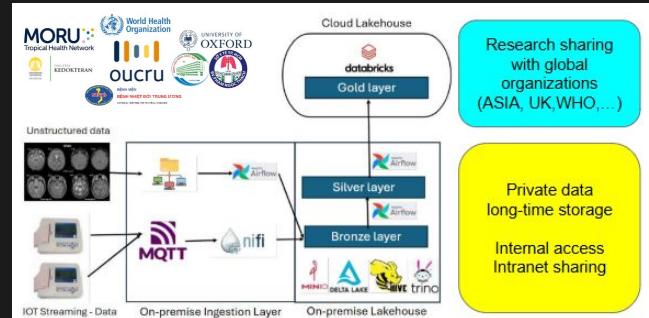
1. TỔNG QUAN MỤC TIÊU VÀ KẾT QUẢ CỦA LUẬN VĂN
2. BỐI CẢNH & ĐỘNG LỰC NGHIÊN CỨU
3. PHÁT BIỂU BÀI TOÁN
4. CÁC NGHIÊN CỨU LIÊN QUAN
5. TẬP DỮ LIỆU SỬ DỤNG (ECG - MRI)
6. KIẾN TRÚC MÔ HÌNH HYBRID CLOUD AND ON-PREMISE DATA LAKEHOUSE
7. TRIỂN KHAI LUỒNG DỮ LIỆU REALTIME (ECG)
8. TRIỂN KHAI LUỒNG DỮ LIỆU BATCH (MRI)
9. KẾT LUẬN & ĐÁNH GIÁ
10. TÀI LIỆU THAM KHẢO

- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
- 2.Bối Cảnh & Động Lực
Nghiên Cứu
- 3.Phát Biểu Bài Toán
- 4.Các Nghiên Cứu Liên
Quan
- 5.Tập Dữ Liệu Sử Dụng
(Ecg - Mri)
- 6.Kiến Trúc Mô Hình
- 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
- 8.Triển Khai Luồng Dữ
Liệu Batch (Mri)
- 9.Kết Luận & Đánh Giá
- 10.Tài Liệu Tham Khảo

1. TỔNG QUAN MỤC TIÊU VÀ KẾT QUẢ CỦA LUẬN VĂN



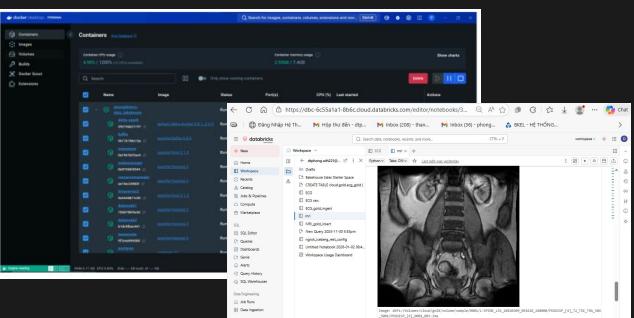
Nhu cầu lưu trữ dữ liệu lớn tại các bệnh viện theo tiêu chuẩn VN và chia sẻ dữ liệu giữa các viện nghiên cứu quốc tế



Đánh giá các lựa chọn và triển khai kiến trúc Data Lakehouse Hybrid Cloud - On premise



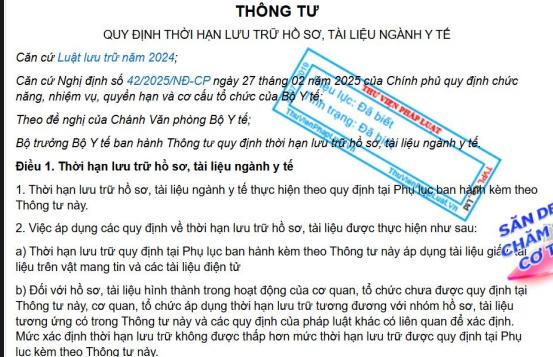
Bài toán lưu trữ và xử lý đa dạng loại dữ liệu y tế:
Dữ liệu Realtime - Dữ liệu phi cấu trúc



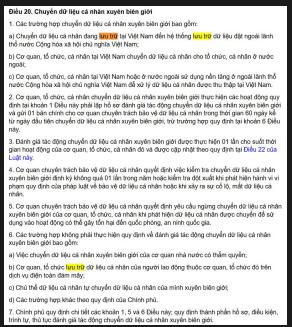
Xây dựng luồng pipeline thực nghiệm xử lý dữ liệu ECG - MRI
Đưa ra đánh giá hiệu năng và đánh giá bảo mật an toàn thông tin

- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
 - 2.Bối Cảnh & Động Lực
Nghiên Cứu
 - 3.Phát Biểu Bài Toán
 - 4.Các Nghiên Cứu Liên
Quan
 - 5.Tập Dữ Liệu Sử Dụng
(Ecg - Mri)
 - 6.Kiến Trúc Mô Hình
 - 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
 - 8.Triển Khai Luồng Dữ
Liệu Batch (Mri)
 - 9.Kết Luận & Đánh Giá
 - 10.Tài Liệu Tham Khảo

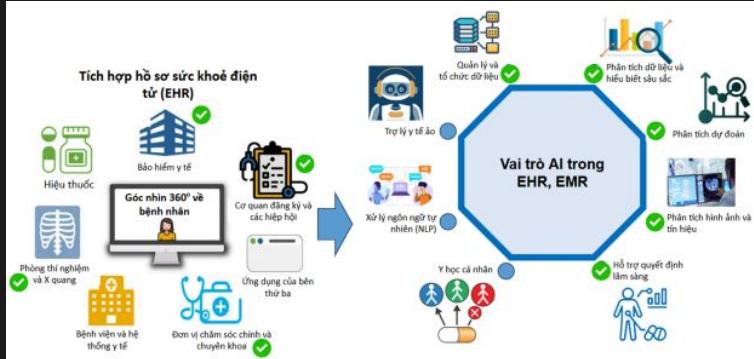
2. BỐI CẢNH & ĐỘNG LỰC NGHIÊN CỨU



Luật Việt Nam: yêu cầu lưu trữ dữ liệu y tế từ 10 năm đến vĩnh viễn



Luật yêu cầu phải xin phép theo quy định chính phủ để được đưa dữ liệu cá nhân ra nước ngoài



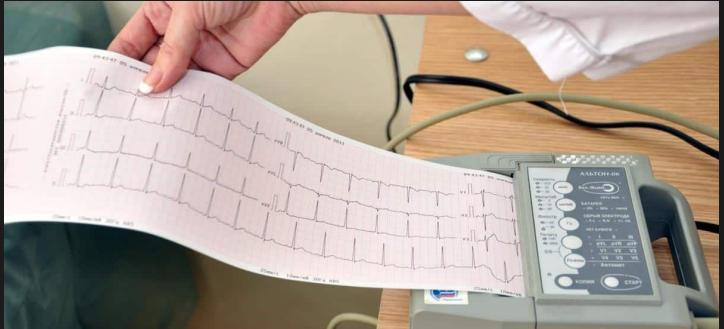
Nhu cầu số hóa tài liệu, chuyển đổi số, chia sẻ liên thông dữ liệu y tế giữa các cơ quan và dữ liệu tập trung quốc gia



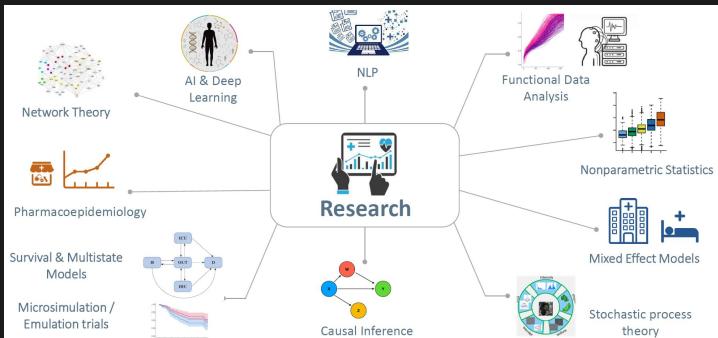
Nhu cầu chia sẻ dữ liệu nghiên cứu lâm sàng giữa các tổ chức y tế trong nước và quốc tế

- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
- 2.Bối Cảnh & Động Lực
Nghiên Cứu
- 3.Phát Biểu Bài Toán
- 4.Các Nghiên Cứu Liên
Quan
- 5.Tập Dữ Liệu Sử Dụng
(Ecg - Mri)
- 6.Kiến Trúc Mô Hình
- 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
- 8.Triển Khai Luồng Dữ
Liệu Batch (Mri)
- 9.Kết Luận & Đánh Giá
- 10.Tài Liệu Tham Khảo

2. BỐI CẢNH & ĐỘNG LỰC NGHIÊN CỨU



Thiết bị y tế giới hạn lưu trữ và sẽ ghi đè dữ liệu cũ khi đầy bộ nhớ: cần phương án tự động lưu trữ tập trung trước khi xoá



Nhu cầu sử dụng Healthcare Machine Learning tăng nhanh, cần môi trường sẵn sàng dữ liệu phục vụ nghiên cứu học máy



Định dạng dữ liệu đa dạng tùy thuộc vào công nghệ và thương hiệu nhà sản xuất: cần tập trung lưu trữ và xử lý chuẩn hóa



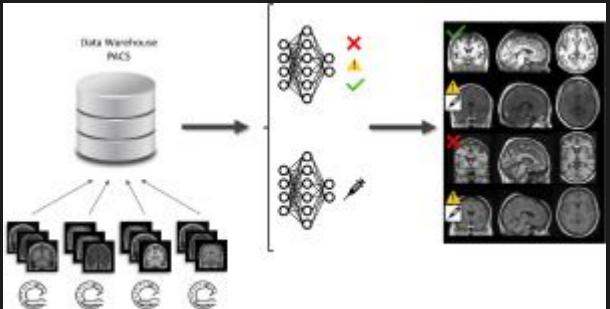
Nguồn lực bảo trì cơ sở dữ liệu cần được tối ưu và chức năng phân quyền truy cập hệ thống cần được đảm bảo an ninh thông tin

- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
- 2.Bối Cảnh & Động Lực
Nghiên Cứu
- 3.Phát Biểu Bài Toán
- 4.Các Nghiên Cứu Liên
Quan
- 5.Tập Dữ Liệu Sử Dụng
(Ecg - Mri)
- 6.Kiến Trúc Mô Hình
- 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
- 8.Triển Khai Luồng Dữ
Liệu Batch (Mri)
- 9.Kết Luận & Đánh Giá
- 10.Tài Liệu Tham Khảo

3. PHÁT BIỂU BÀI TOÁN



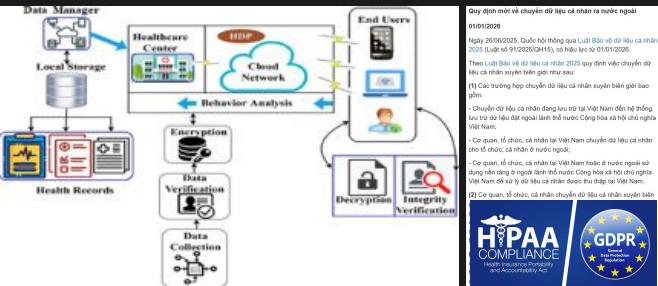
Hệ thống ICU 20 giường với các thiết bị ECG tạo ra dòng dữ liệu sinh hiệu thời gian thực, yêu cầu thu thập và xử lý liên tục 24/7.



Dữ liệu MRI được thu thập tập trung theo cơ chế batch và chuyển đổi định dạng nhằm sẵn sàng cho lưu trữ và chia sẻ nghiên cứu.



Dữ liệu ECG được tập trung từ nhiều thiết bị cầm tay về hệ thống trung tâm trước khi bị ghi đè, đồng thời được chuyển đổi sang dạng bảng để phục vụ phân tích và nghiên cứu.



Dữ liệu gốc được lưu trữ đầy đủ và nguyên bản trong hạ tầng nội bộ, trong khi dữ liệu lâm sàng không nhạy cảm được chia sẻ có kiểm soát cho các tổ chức nghiên cứu quốc tế.

- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
 - 2.Bối Cảnh & Động Lực
Nghiên Cứu
 - 3.Phát Biểu Bài Toán
 - 4.Các Nghiên Cứu Liên
Quan
 - 5.Tập Dữ Liệu Sử Dụng
(Ecg - Mri)
 - 6.Kiến Trúc Mô Hình
 - 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
 - 8.Triển Khai Luồng Dữ
Liệu Batch (Mri)
 - 9.Kết Luận & Đánh Giá
 - 10.Tài Liệu Tham Khảo

4. Các nghiên cứu liên quan

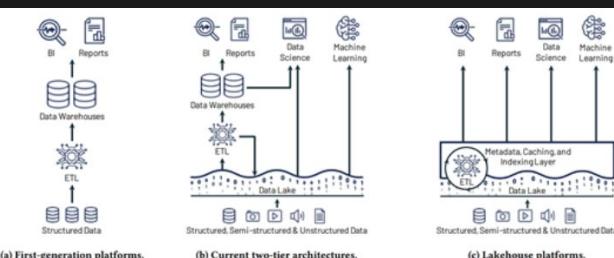


“Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics” – Armbrust et al. (2021) [1]: Giới thiệu khái niệm Lakehouse với khả năng kết hợp ưu điểm của Data Warehouse và Data Lake, hỗ trợ tốt cho phân tích và học máy.

“From Data Warehouse to Lakehouse: A Comparative Review” (2023) [11]: So sánh đặc điểm kỹ thuật, hiệu năng và chi phí giữa kiến trúc truyền thống và Lakehouse, nêu bật vai trò của Delta Lake, Iceberg, Hudi.

“The Evolution from Data Warehouses to Data Lakehouses: A Technical Perspective” – Sai Kaushik Ponnekanti (2025) [7]: Trình bày sự phát triển từ Data Warehouse đến Lakehouse, phân tích vai trò Lakehouse trong xử lý dữ liệu lớn hiện đại.

	Data Warehouse	Data Lake	Lakehouse
Data	Rational data from transactional systems, operational databases & business applications	All data including structured, semi-structured, and unstructured.	Query Every kind of data Image, audio, video, and others.
Data Access	SQL Only	Open API, SQL, Python	Open API, SQL, Python
Data Format	Proprietary Format	Open format	Open Format
Governance	Fine-grained Security	Weak Governance and security	Fine-grained Security
Reliability	High with ACID Transactions	Low Quality- Data Swamps	High with ACID Transactions.
Performance	Fast query results using local language	Faster query results, decoupling of computing and storage	Faster and deeper insights without data movement.
Scalability	Scaling becomes expensive	Low cost of scaling Regardless of the data-type	Low cost of scaling regardless of the data-type



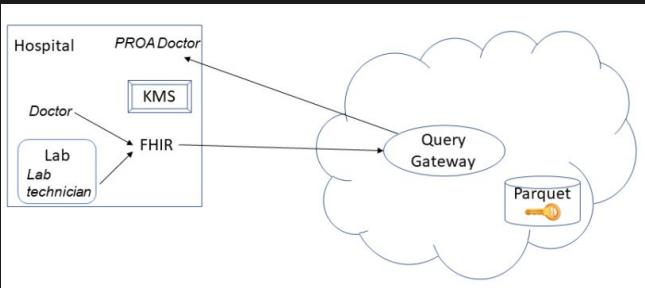
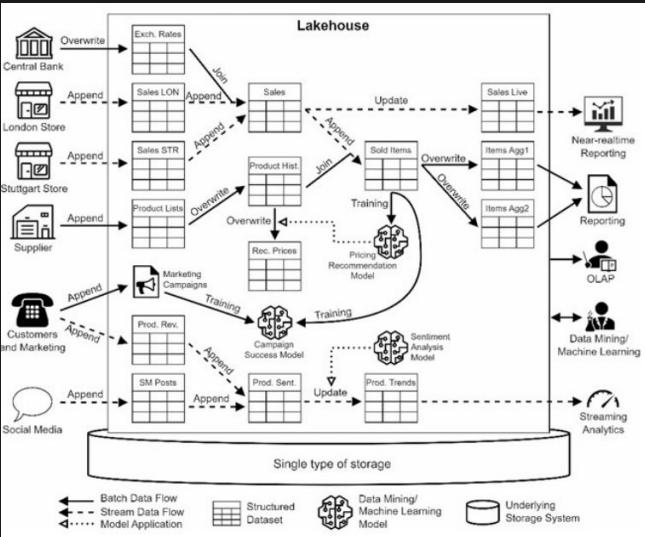
- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
- 2.Bối Cảnh & Động Lực
Nghiên Cứu
- 3.Phát Biểu Bài Toán
- 4.Các Nghiên Cứu Liên
Quan
- 5.Tập Dữ Liệu Sử Dụng
(Ecg - Mri)
- 6.Kiến Trúc Mô Hình
- 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
- 8.Triển Khai Luồng Dữ
Liệu Batch (Mri)
- 9.Kết Luận & Đánh Giá
- 10.Tài Liệu Tham Khảo

4. Các nghiên cứu liên quan

“The Lakehouse: State of the Art on Concepts and Technologies” (2024) [12]: Tổng quan kiến trúc Lakehouse, phân tích các thành phần như định dạng hỗ trợ ACID, quản lý metadata, caching và công cụ như Apache Spark, Databricks.

“Design of Vessel Data Lakehouse with Big Data and AI Analysis Technology for Vessel Monitoring System” – Lee et al. (2022): Đề xuất kiến trúc Lakehouse cho giám sát tàu biển theo thời gian thực, tích hợp xử lý dữ liệu cảm biến và AI.

“Protecting Sensitive Tabular Data in Hybrid Clouds” – Maya Anderson et al. (2023) [10]: Đề xuất giải pháp mã hóa dữ liệu dạng bảng trong đám mây lai, hỗ trợ truy vấn an toàn và tuân thủ GDPR, phù hợp với lĩnh vực y tế và tài chính.



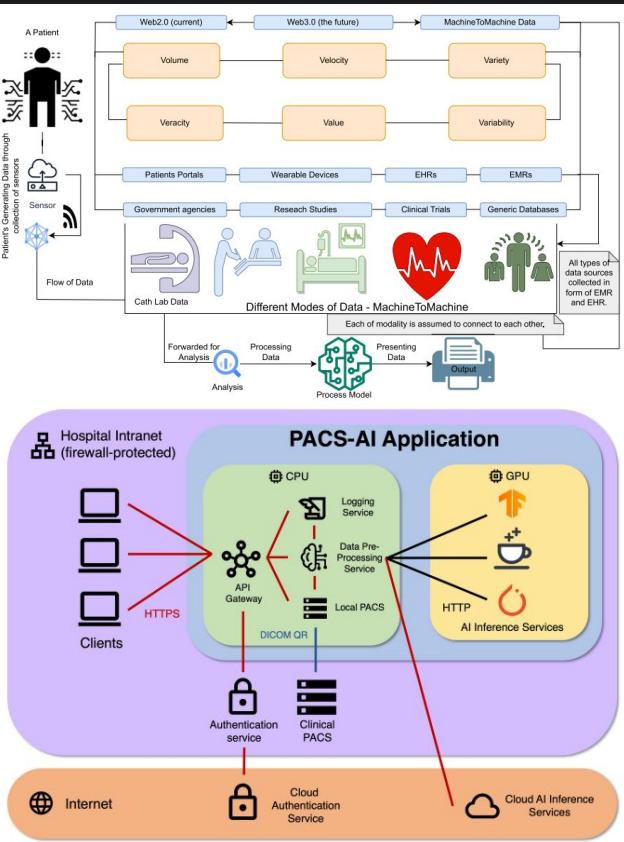
- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
- 2.Bối Cảnh & Động Lực
Nghiên Cứu
- 3.Phát Biểu Bài Toán
- 4.Các Nghiên Cứu Liên
Quan
- 5.Tập Dữ Liệu Sử Dụng
(Ecg - Mri)
- 6.Kiến Trúc Mô Hình
- 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
- 8.Triển Khai Luồng Dữ
Liệu Batch (Mri)
- 9.Kết Luận & Đánh Giá
- 10.Tài Liệu Tham Khảo

4. Các nghiên cứu liên quan

“Big Data Analytics in Healthcare” – C. Guo & J. Chen (2023) [13]: Phân tích vai trò của Big Data Analytics trong chẩn đoán, điều trị và quản lý bệnh bằng cách tích hợp dữ liệu và ứng dụng AI trong chăm sóc sức khỏe.

“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]: Giới thiệu nền tảng PACS-AI hỗ trợ triển khai AI cho hình ảnh y tế và tín hiệu sinh học theo thời gian thực, đảm bảo tuân thủ đạo đức và quyền riêng tư.

“Harnessing Big Data Analytics for Healthcare: A Comprehensive Review of Frameworks, Implications, Applications, and Impacts” – Awais Ahmed (2023) [15]: Tổng quan 180 nghiên cứu về BDA trong y tế, trình bày các khung kiến trúc, ứng dụng, thách thức và tác động đến hiệu quả chăm sóc sức khỏe.



- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
- 2.Bối Cảnh & Động Lực
Nghiên Cứu
- 3.Phát Biểu Bài Toán
- 4.Các Nghiên Cứu Liên
Quan
- 5.Tập Dữ Liệu Sử Dụng
(Ecg - MRI)
- 6.Kiến Trúc Mô Hình
- 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
- 8.Triển Khai Luồng Dữ
Liệu Batch (MRI)
- 9.Kết Luận & Đánh Giá
- 10.Tài Liệu Tham Khảo

5. Tập dữ liệu sử dụng

5.1 ECG

Published: March 6, 2014, Version: 1.0.0

Biometric Human Identification based on ECG (March 6, 2014, 1 p.m.)

The ECG-ID Database is a set of 310 ECGs from 90 volunteers, created and contributed to PhysioBank by Tatiana Lugovaya, who used the ECGs in her master's thesis. An excellent summary of this thesis, with a discussion of the challenges in using ECGs as biometrics, and a comparison of the author's methods and results with those of three previous studies, is also available.

When using this resource, please cite the original publication:
Lugovaya T.S. Biometric human identification based on electrocardiogram. [Master's thesis] Faculty of Computing Technologies and Informatics, Electrotechnical University "LETI", Saint-Petersburg, Russian Federation; June 2005.

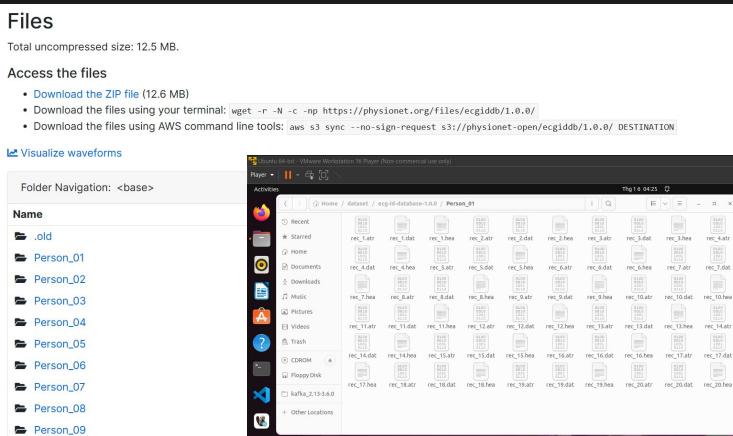
Please include the standard citation for PhysioNet: (show more options)

Golberger A., Amaral L., Glass L., Hausdorff J., Ivanov P. C., Mark R., & Stanley H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* [Online]. 101 (23), pp. e215-e220. RRID:SCR_007345.

<https://physionet.org/content/ecgiddb/1.0.0/>

Bộ dữ liệu ECG-ID do PhysioNet cung cấp bao gồm các bản ghi điện tâm đồđạo trình I, thu thập từ 90 tình nguyện viên với độ tuổi đa dạng, phục vụ nghiên cứu sinh trắc học và phân tích tín hiệu ECG.

Mỗi bản ghi ECG có thời lượng khoảng 20 giây, được lấy mẫu ở tần số 500Hz với độ phân giải 12-bit, phản ánh đặc tính dữ liệu thời gian thực có độ chi tiết cao.



Dữ liệu ECG-ID bao gồm cả tín hiệu thô và tín hiệu đã được lọc, kèm theo annotation.

Mỗi bản ghi ECG trong bộ dữ liệu được lưu dưới ba tệp chính, trong đó tệp .dat chứa tín hiệu ECG thô dạng nhị phân, tệp .hea lưu trữ metadata và thông số kỹ thuật của bản ghi (tần số lấy mẫu, số kênh, độ phân giải, thời lượng), và tệp .atr chứa các annotation tự động như vị trí đỉnh sóng R và T phục vụ phân tích tín hiệu.

- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
- 2.Bối Cảnh & Động Lực
Nghiên Cứu
- 3.Phát Biểu Bài Toán
- 4.Các Nghiên Cứu Liên
Quan
- 5.Tập Dữ Liệu Sử Dụng
(Ecg - MRI)
- 6.Kiến Trúc Mô Hình
- 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
- 8.Triển Khai Luồng Dữ
Liệu Batch (MRI)
- 9.Kết Luận & Đánh Giá
- 10.Tài Liệu Tham Khảo

5. Tập dữ liệu sử dụng

5.2 MRI



Lumbar Spine MRI Dataset

Published: 3 April 2019 | Version 2 | DOI: 10.17632/k57fr854j2.2

Contributors: Sud Sidiman, Al Al Kafri, Friska Natalia, Hira Meidha, Nunik Afriiana, Wasfi Al-Rashdan, Mohammad Bashtawi, Mohammad Al-Jumaily

Description

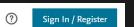
This data set contains anonymised clinical MRI study, or a set of scans, of 515 patients with symptomatic back pains. Each patient data can have one or more MRI studies associated with it. Each study contains slices, i.e., individual images taken from either sagittal or axial view, of the lowest three vertebrae and the lowest three IVDs. The axial view slices are mainly taken from the last three IVDs – including the one between the last vertebrae and the sacrum. The orientation of the slices of the last IVD are made to follow the spine curve whereas those of the other IVDs are usually made in blocks – i.e., parallel to each other. There are between four to five slices per IVD and they begin from the top of the IVD towards its bottom. Many of the top and bottom slices cut through the vertebrae leaving between one to three slices that cut the IVD cleanly and show purely the image of that IVD. In most cases, the total number of slices in axial view ranges from 12 to 15. However, in some cases, there may be up to 20 slices since the study contains slices of more than just three vertebrae. The scans in sagittal view also vary but all contain at least the last seven vertebrae and the sacrum. While the number of vertebrae varies, each scan always includes the first two sacral slices.

There are a total 48,345 MRI slices in our dataset. The majority of the slices have an image resolution of 320x320 pixels; however, there are slices from three studies with 320x310 pixel resolution. The pixels in all slices have 12-bit per pixel precision which is higher than the standard 8-bit grayscale images. Specifically for all axial-view slices, the slice thickness are uniformly 4mm with centre-to-centre distance between adjacent slices to be 4.4mm. The horizontal and vertical pixel spacing is 0.6875 mm uniformly across all axial-view slices.

The majority of the MRI studies were taken with the patient in Head-First-Supine position with the rests were taken with the patient in Feet-First-Supine position. Each study can last between 15 to 45 minutes and a patient may have one or more study associated with them taken at a different time or a few days apart.

<https://data.mendeley.com/datasets/k57fr854j2/2>

Bộ dữ liệu Lumbar Spine MRI bao gồm các nghiên cứu MRI đã được ấn danh của 515 bệnh nhân có triệu chứng đau lưng, với mỗi bệnh nhân có thể có một hoặc nhiều lần chụp khác nhau. Mỗi nghiên cứu chứa các lát cắt MRI ở hai mặt phẳng sagittal và axial, tập trung vào ba đốt sống thắt lưng cuối cùng, các đĩa đệm liên đốt (IVD) tương ứng và xương cùng, phản ánh đầy đủ cấu trúc giải phẫu vùng thắt lưng.



Dataset metrics

Citations 11

Usage

Views: 25506

Downloads: 4809



[View details >](#)

Latest version

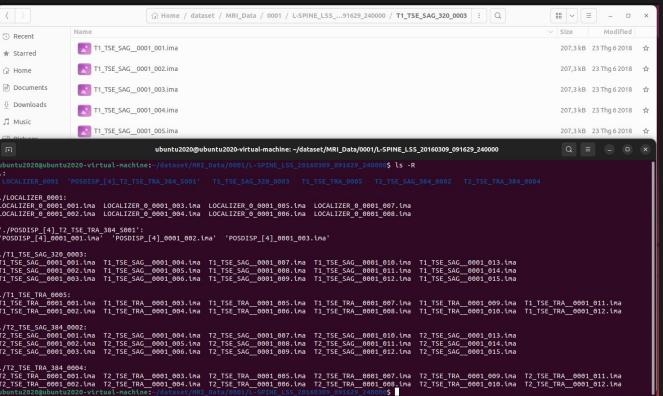
Version 2

Published: 3 Apr 2019

DOI: 10.17632/k57fr854j2.2

Cite this dataset

Sudiman, Sud; Al Kafri, Alz; Natalia, Friska; Meidha, Hira; Afriiana, Nunik; Al-Rashdan, Wasfi; Bashtawi, Mohammad; Al-Jumaily, Mohammad (2019), "Lumbar Spine MRI Dataset", Mendeley Data, V2, doi: 10.17632/k57fr854j2.2



Tổng cộng bộ dữ liệu gồm 48.345 lát cắt MRI với độ phân giải chủ yếu 320×320 pixel, độ sâu màu 12-bit và thông số chụp đồng nhất ở mặt phẳng axial.

Mỗi nghiên cứu MRI được tổ chức theo cấu trúc thư mục dạng series, trong đó các thư mục con (LOCALIZER, T1_TSE, T2_TSE ở các mặt phẳng sagittal và axial) đại diện cho từng chuỗi chụp khác nhau, và mỗi tệp .ima là một lát cắt DICOM riêng lẻ, phản ánh cách dữ liệu ảnh y tế được lưu trữ phân tán theo series trong thực tế bệnh viện.

- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
- 2.Bối Cảnh & Động Lực
Nghiên Cứu
- 3.Phát Biểu Bài Toán
- 4.Các Nghiên Cứu Liên
Quan

- 5.Tập Dữ Liệu Sử Dụng
(Ecg - Mri)
- 6.Kiến Trúc Mô Hình
- 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
- 8.Triển Khai Luồng Dữ
Liệu Batch (Mri)
- 9.Kết Luận & Đánh Giá
- 10.Tài Liệu Tham Khảo

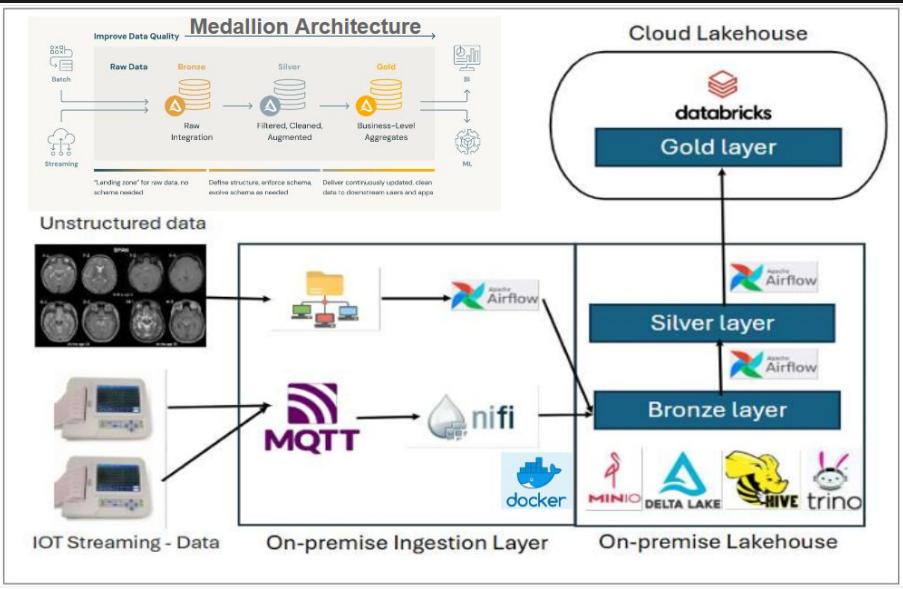
6. Kiến trúc mô hình

Kiến trúc Data Lakehouse được thiết kế theo mô hình đám mây lai (Hybrid Cloud–On-Premise), ưu tiên tuyệt đối bảo mật và chủ quyền dữ liệu y tế khi toàn bộ dữ liệu gốc được lưu trữ và xử lý trong hạ tầng nội bộ bệnh viện.

Ưu tiên các lựa chọn mã nguồn mở và ít phụ thuộc vào vendor lock.

Hệ thống hỗ trợ đồng thời xử lý dữ liệu thời gian thực (ECG) và dữ liệu batch (MRI) theo Medallion Architecture, đảm bảo khả năng mở rộng, truy vết nguồn gốc và kiểm soát chất lượng dữ liệu.

Chỉ các tập dữ liệu đã được làm sạch, chuẩn hóa và ẩn danh mới được chia sẻ lên môi trường đám mây để phục vụ phân tích và nghiên cứu trí tuệ nhân tạo.



- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
- 2.Bối Cảnh & Động Lực
Nghiên Cứu
- 3.Phát Biểu Bài Toán
- 4.Các Nghiên Cứu Liên
Quan
- 5.Tập Dữ Liệu Sử Dụng
(Ecg - MRI)
- 6.Kiến Trúc Mô Hình
- 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
- 8.Triển Khai Luồng Dữ
Liệu Batch (MRI)
- 9.Kết Luận & Đánh Giá
- 10.Tài Liệu Tham Khảo

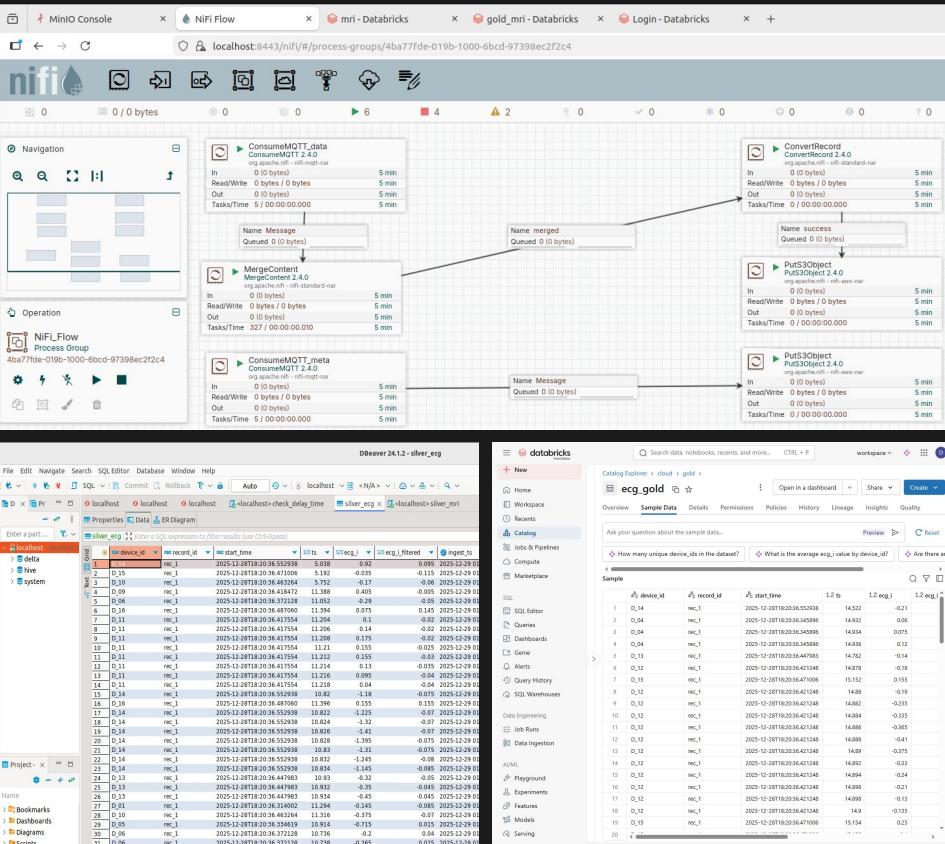
7. Triển khai luồng dữ liệu Real-Time ECG

Dữ liệu ECG thời gian thực được phát từ các thiết bị ICU lên MQTT theo mô hình publish–subscribe và được Apache NiFi subscribe để thu thập.

NiFi chuẩn hóa schema, gắn metadata và ghi dữ liệu dưới định dạng Avro vào MinIO tại Bronze Layer.

Apache Spark Structured Streaming đọc dữ liệu Avro từ Bronze, xử lý làm sạch và ghi vào bảng Delta Lake tại Silver Layer.

Từ Silver Layer, Spark tiếp tục loại bỏ hoặc ẩn danh các trường dữ liệu cá nhân và đầy dữ liệu phân tích thông qua API lên Databricks Cloud để xây dựng Gold Layer phục vụ phân tích và nghiên cứu.



- 1.Tổng Quan Mục Tiêu
- Và Kết Quả Của Luận Văn
- 2.Bối Cảnh & Động Lực
- Nghiên Cứu
- 3.Phát Biểu Bài Toán
- 4.Các Nghiên Cứu Liên Quan
- 5.Tập Dữ Liệu Sử Dụng (Ecg - Mri)
- 6.Kiến Trúc Mô Hình
- 7.Triển Khai Luồng Dữ Liệu Realtime (Ecg)
- 8.Triển Khai Luồng Dữ Liệu Batch (Mri)
- 9.Kết Luận & Đánh Giá
- 10.Tài Liệu Tham Khảo

8. Triển khai luồng dữ liệu Batch (MRI)

Dữ liệu MRI được tạo ra từ hệ thống chụp và lưu trữ ban đầu dưới dạng file DICOM trên Windows shared drive trong mạng nội bộ bệnh viện.

Apache Airflow điều phối luồng ingest theo cơ chế batch định kỳ, giám sát shared drive, phát hiện dữ liệu mới và sao chép nguyên bản file DICOM vào Bronze Layer trên MinIO, kèm metadata ingest để đảm bảo khả năng truy vết.

Airflow tiếp tục kích hoạt các job Apache Spark batch đọc dữ liệu DICOM từ Bronze, trích xuất metadata và lưu trữ dưới dạng bảng theo binary column bên trong silver layer.

Từ Silver Layer, dùng API Databricks để chuyển dữ liệu dạng decode base64 lên Gold Layer trên Databricks Cloud để phục vụ phân tích hình ảnh y tế nghiên cứu, trong khi dữ liệu gốc vẫn được lưu trữ an toàn on-premise, tuân thủ quy định bảo mật y tế.

- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
 - 2.Bối Cảnh & Động Lực
Nghiên Cứu
 - 3.Phát Biểu Bài Toán
 - 4.Các Nghiên Cứu Liên
Quan
 - 5.Tập Dữ Liệu Sử Dụng
(Ecg - Mri)
 - 6.Kiến Trúc Mô Hình
 - 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
 - 8.Triển Khai Luồng Dữ
Liệu Batch (Mri)
 - 9.Kết Luận & Đánh Giá
 - 10.Tài Liệu Tham Khảo

9. Kết luận và đánh giá

Dữ liệu ECG được kiểm tra đầy đủ 100%, không ghi nhận mất mát dữ liệu trong quá trình ingest; tổng cộng khoảng 200.000 bản ghi được thu thập và truyền thành công từ MQTT → NiFi → Bronze → Silver → Cloud, đảm bảo toàn vẹn dữ liệu trong suốt quá trình truyền tải và xử lý.

Kích bản 1: Độ trễ xử lý được đo ở mức trung bình khoảng 30 giây cho một thiết bị ECG

Kích bản 2: gồm 20 thiết bị ECG hoạt động đồng thời, dữ liệu được ghi đầy đủ vào Silver Layer sau khoảng 3 phút, không phát sinh thất thoát hay backlog kéo dài.

Dữ liệu MRI theo cơ chế batch được ingest đầy đủ 100%, với 300 file DICOM được xử lý và đồng bộ thành công từ Bronze → Silver → Gold; tổng thời gian xử lý và đưa dữ liệu lên Gold Layer là khoảng 30 phút, không ghi nhận mất mát dữ liệu trong quá trình batch processing.

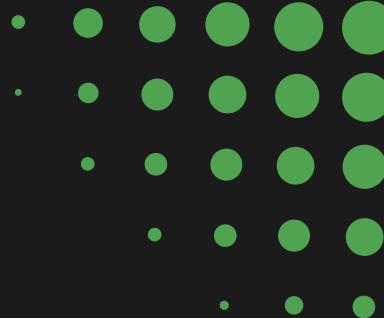
The screenshot shows a Databricks workspace interface. On the left, there's a sidebar with various navigation options like Home, Workspace, Catalog, Jobs & Pipelines, Compute, Marketplace, SQL, SQL Editor, Queries, Dashboards, Genie, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, AI/ML, Playground, Experiments, Features, Models, and Seving. The main area has a table titled "SELECT device_id, MIN(to_unixtime(at_timezone)) FROM device_id GROUP BY device_id" with 20 rows of data. Below the table is a Python notebook titled "mri" with code for processing pixel arrays and displaying MRI slices. A preview image of an MRI scan is shown on the right.

device_id	device_id	min_delay_to_minio_sec	avg_delay_to_minio_sec	max_delay_to_minio_sec	min_delay_to_silver_sec	avg_delay_to_silver_sec	max_delay_to_silver_sec
1	D_01	41.286000134	127.839601087	196.702000111	48.8500001431	136.7230841208	207.720000287
2	D_02	41.3570001125	131.4265999943	198.620999984	48.9430000782	142.3817063061	208.369999958
3	D_03	41.2860001323	127.449999998	196.620000128	48.8500001431	136.7230841208	207.720000287
4	D_04	41.2890000962	127.292999714	195.980000248	48.7900000248	138.261550582	207.5670001507
5	D_05	41.349999627	130.2863998894	197.887999773	48.8539998511	141.1350026002	208.263999993
6	D_06	41.289999523	129.169800021	196.763999939	48.8500001431	139.9755824322	207.8180000782
7	D_07	41.2380000545	131.1320000529	196.8699999373	44.0000000629	142.029344000	207.9240002632
8	D_08	41.2380000595	127.4438999332	195.853999945	48.773999844	138.84647152	207.8920002392
9	D_09	41.3020000458	129.02520000172	197.7100000381	48.8120000362	139.9357029272	207.9020001888
10	D_10	41.202998302	126.987999940	196.626999855	48.753000021	137.8924769504	207.611000061
11	D_11	41.2170000706	127.8788999447	196.714999142	48.8669998837	138.7966639573	207.7170000076
12	D_12	41.2380000752	129.4930000529	197.873000030	48.809000042	140.8110000109	207.8710000160
13	D_13	41.2530000023	130.2089999733	197.5880000866	44.0359997749	141.1274787872	207.9530000687
14	D_14	41.1800000668	130.4094999542	197.4600000381	48.7699999809	141.335943446	207.7980000973
15	D_15	41.21500001526	128.2584000439	195.9370000362	48.7699999054	139.1220764554	207.5190001049
16	D_16	41.1110000782	127.1110000529	196.100000103	48.8239999542	138.1144928236	207.7109999457
17	D_17	41.2379999352	129.5363999892	197.5399999873	48.8990000405	140.8466639505	207.8970000390
18	D_18	41.0110001564	129.638000061	196.6970000267	43.7920000553	140.5525504752	207.6930000782
19	D_19	41.0339999199	129.5595999561	197.379999876	43.80699999218	140.533662668	207.7199999352
20	D_20	41.131000042	128.1122000553	196.5829999447	48.6830000877	138.9930322662	207.5910000801

- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
- 2.Bối Cảnh & Động Lực
Nghiên Cứu
- 3.Phát Biểu Bài Toán
- 4.Các Nghiên Cứu Liên
Quan
- 5.Tập Dữ Liệu Sử Dụng
(Ecg - MRI)
- 6.Kiến Trúc Mô Hình
- 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
- 8.Triển Khai Luồng Dữ
Liệu Batch (Mri)
- 9.Kết Luận & Đánh Giá
- 10.Tài Liệu Tham Khảo

9. Kết luận và đánh giá

- Luận văn đề xuất mô hình kiến trúc Data Lakehouse đam mê lai được đặc thù hóa cho dữ liệu y tế, phân tách rõ vai trò on-premise và cloud, góp phần đảm bảo bảo mật, chủ quyền dữ liệu và tuân thủ pháp lý.
- Xây dựng khung phân loại dữ liệu y tế (ECG thời gian thực, MRI batch, dữ liệu định danh) và thiết kế các pipeline xử lý tương ứng theo mô hình Bronze–Silver–Gold trong một kiến trúc thống nhất.
- Cung cấp minh chứng thực nghiệm cho khả năng vận hành đồng thời pipeline streaming và batch, đáp ứng yêu cầu về độ trễ, tính toàn vẹn dữ liệu và khả năng mở rộng.
- Áp dụng các nguyên tắc quản trị dữ liệu y tế như quản lý vòng đời, truy vết nguồn gốc và kiểm soát chất lượng dữ liệu thông qua Delta Lake và metadata.
- Đóng góp chia sẻ mã nguồn kiến trúc tại Github: [GitHub - PhongDinhCS/lakehousehybrid](https://github.com/PhongDinhCS/lakehousehybrid)
- Hạn chế chính của nghiên cứu là thử nghiệm ở quy mô mô phỏng, chưa đánh giá sâu chi phí vận hành và tích hợp đầy đủ với hệ thống bệnh viện lớn.
- Hướng phát triển trong tương lai gồm mở rộng quy mô hệ thống, tích hợp AI cho phân tích ECG và MRI, tăng cường quản lý metadata, giám sát pipeline thời gian thực và nghiên cứu sâu hơn các quy định pháp lý dữ liệu y tế tại Việt Nam.



- 1.Tổng Quan Mục Tiêu
Và Kết Quả Của Luận
Văn
- 2.Bối Cảnh & Động Lực
Nghiên Cứu
- 3.Phát Biểu Bài Toán
- 4.Các Nghiên Cứu Liên
Quan
- 5.Tập Dữ Liệu Sử Dụng
(Ecg - Mri)
- 6.Kiến Trúc Mô Hình
- 7.Triển Khai Luồng Dữ
Liệu Realtime (Ecg)
- 8.Triển Khai Luồng Dữ
Liệu Batch (Mri)
- 9.Kết Luận & Đánh Giá
- 10.Tài Liệu Tham Khảo

10. 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.

10. TÀI LIỆU THAM KHẢO

- 
- [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.

Cảm ơn

Em xin chân thành cảm ơn quý Thầy, Cô đã theo dõi phần trình bày.