

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



Trần Quang Đại

**NGHIÊN CỨU, XÂY DỰNG TRỰC LIÊN THÔNG DỮ LIỆU
PHỤC VỤ LƯU TRỮ, QUẢN LÝ DỮ LIỆU ĐA CẤU TRÚC**

CHUYÊN NGÀNH : HỆ THỐNG THÔNG TIN

MÃ SỐ: 8.48.01.04

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

HÀ NỘI – 2025

Đề án tốt nghiệp được hoàn thành tại:
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. Phan Lý Huỳnh

Phản biện 1:

Phản biện 2:

Đề án tốt nghiệp sẽ được bảo vệ trước Hội đồng chấm đề án tốt nghiệp thạc sĩ
tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: giờ ngày tháng năm

Có thể tìm hiểu đề án tốt nghiệp tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

1. Lý do chọn đề tài

Trong bối cảnh công nghệ thông tin và truyền thông phát triển mạnh mẽ, chuyển đổi số đã và đang trở thành xu hướng tất yếu trong tất cả các lĩnh vực, đặc biệt là trong giáo dục đại học. Chuyển đổi số không chỉ đơn thuần là số hóa các quy trình quản lý, mà còn là quá trình tái cấu trúc toàn diện các hoạt động giảng dạy, nghiên cứu và vận hành theo hướng dựa trên dữ liệu.

Tại Việt Nam, nhiều trường đại học đã triển khai các hệ thống công nghệ thông tin nhằm hỗ trợ quản lý và học tập như hệ thống quản lý đào tạo (LMS), thư viện số, hệ thống quản lý sinh viên,... Tuy nhiên, các hệ thống này thường hoạt động độc lập, dẫn đến tình trạng phân mảnh thông tin, dữ liệu không đồng bộ, gây khó khăn trong khai thác, phân tích và ra quyết định.

Đồng thời, khối lượng dữ liệu được tạo ra tại các trường đại học ngày càng tăng về cả quy mô và hình thái, bao gồm dữ liệu có cấu trúc (dữ liệu bảng điểm, hồ sơ sinh viên), bán cấu trúc (log hệ thống, dữ liệu XML/JSON) và phi cấu trúc (tài liệu PDF, video bài giảng, khảo sát tự do...). Việc tích hợp và quản lý hiệu quả dữ liệu đa cấu trúc là điều kiện tiên quyết để nâng cao hiệu quả vận hành và hoạch định chiến lược trong môi trường giáo dục số.

Do đó, việc nghiên cứu và xây dựng một **trục liên thông dữ liệu** có khả năng **kết nối, tích hợp và chia sẻ dữ liệu đa cấu trúc** giữa các hệ thống trong một trường đại học là yêu cầu cấp thiết. Trục liên thông không chỉ giúp đồng bộ hóa và tối ưu hóa quá trình lưu trữ, mà còn tạo nền tảng cho các ứng dụng phân tích, trực quan hóa và ra quyết định dựa trên dữ liệu.

2. Mục đích nghiên cứu

Mục đích chính của nghiên cứu này là xây dựng một trục liên thông dữ liệu đóng vai trò như một nền tảng trung gian để kết nối, tích hợp và đồng bộ hóa dữ liệu đa cấu trúc từ nhiều hệ thống khác nhau trong một trường đại học. Việc thiết lập một hạ tầng có khả năng xử lý và tích hợp các định dạng dữ liệu không đồng nhất là vô cùng cần

thiết. Trục liên thông dữ liệu sẽ giúp đảm bảo tính tương tác, liên kết và tái sử dụng dữ liệu giữa các hệ thống, đồng thời khắc phục tình trạng phân mảnh thông tin hiện nay.

Thông qua các công cụ phân tích, dữ liệu sẽ được chuyển hóa thành thông tin có giá trị nhằm phục vụ cho các hoạt động quản trị, giảng dạy, nghiên cứu và ra quyết định. Ngoài ra, nghiên cứu cũng chú trọng đến việc trực quan hóa dữ liệu thông qua các biểu đồ, bảng số liệu, đồ thị nhằm tăng tính tương tác và dễ đọc hiểu cho người dùng cuối.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu

Nghiên cứu này tập trung vào ba nhóm đối tượng chính có liên quan mật thiết đến quá trình xây dựng và vận hành một trục liên thông dữ liệu phục vụ lưu trữ, quản lý dữ liệu đa cấu trúc trong một trường đại học:

Thứ nhất, nghiên cứu về dữ liệu đa cấu trúc (multistructured data), bao gồm các loại dữ liệu có cấu trúc, dữ liệu bán cấu trúc và dữ liệu phi cấu trúc. Việc hiểu rõ tính chất, đặc điểm và nhu cầu xử lý của từng loại dữ liệu là cơ sở quan trọng để đề xuất giải pháp lưu trữ, tích hợp và phân tích phù hợp.

Thứ hai, nghiên cứu về trục liên thông dữ liệu X-Road, một trong những nền tảng nổi bật và được áp dụng rộng rãi trong việc kết nối và trao đổi dữ liệu giữa các hệ thống thông tin khác nhau. Nội dung nghiên cứu sẽ làm rõ cấu trúc, cơ chế hoạt động, khả năng tương tác, bảo mật và khả năng mở rộng của X-Road, từ đó đánh giá mức độ phù hợp khi áp dụng trong bối cảnh liên thông dữ liệu tại một trường đại học.

Thứ ba, nghiên cứu về mô hình trực quan hóa dữ liệu và sinh báo cáo thống kê nhằm phục vụ cho mục tiêu phân tích dữ liệu được chia sẻ qua trục liên thông. Nội dung này bao gồm việc tìm hiểu các kỹ thuật trực quan hóa hiện đại, các công cụ hỗ trợ báo cáo dữ liệu động và tương tác, cũng như các tiêu chí đánh giá hiệu quả biểu diễn dữ liệu phục vụ cho quản trị và ra quyết định.

Phạm vi nghiên cứu

Phạm vi của đề tài được xác định rõ ràng để đảm bảo tính khả thi và tập trung trong quá trình triển khai:

Về nội dung, nghiên cứu sẽ tập trung vào quá trình chia sẻ và liên thông dữ liệu đa cấu trúc giữa các hệ thống trong một trường đại học thông qua trục liên thông X-Road. Các loại dữ liệu trong phạm vi nghiên cứu bao gồm: Dữ liệu có cấu trúc, Dữ liệu bán cấu trúc và Dữ liệu phi cấu trúc.

Về thời gian và dữ liệu khảo sát, nghiên cứu thu thập dữ liệu đa cấu trúc thông qua trục liên thông X-Road của một trường đại học trong khoảng thời gian một tháng. Dữ liệu được chia sẻ qua các hệ thống khác nhau sẽ được tổng hợp, chuẩn hóa và lưu trữ theo mô hình hồ dữ liệu (data lake). Từ đó, quá trình phân tích, trích xuất thông tin và xây dựng báo cáo thông kê sẽ được thực hiện để đánh giá hiệu quả hoạt động của mô hình.

4. Phương pháp nghiên cứu

Trong quá trình thực hiện đề tài, để đảm bảo tính khoa học, khách quan và thực tiễn, nhóm nghiên cứu đã áp dụng kết hợp nhiều phương pháp khác nhau, bao gồm cả phương pháp nghiên cứu lý thuyết, thực nghiệm và phân tích đánh giá. Cụ thể như sau:

Phương pháp nghiên cứu lý thuyết

Đây là phương pháp nền tảng được sử dụng trong giai đoạn đầu của nghiên cứu nhằm xây dựng cơ sở lý luận và xác định hướng đi phù hợp cho toàn bộ đề tài. Nội dung bao gồm:

- Khảo sát, tổng hợp và phân tích các công trình nghiên cứu, tài liệu học thuật, báo cáo kỹ thuật trong và ngoài nước liên quan đến lĩnh vực liên thông dữ liệu và dữ liệu đa cấu trúc.

- Nghiên cứu chuyên sâu về các kiến trúc trục liên thông dữ liệu hiện có, đặc biệt là mô hình X-Road, nhằm làm rõ các khía cạnh kỹ thuật như bảo mật, xác thực, mô hình tổ chức, và khả năng tích hợp với các hệ thống hiện hành.

- Tìm hiểu các phương pháp và công cụ được sử dụng trong việc thu thập, lưu trữ và quản lý dữ liệu đa cấu trúc, từ các giải pháp truyền thống như cơ sở dữ liệu quan hệ đến các công nghệ hiện đại như Data Lake, NoSQL, và hệ thống lưu trữ phân tán.

- Khảo sát các kỹ thuật phân tích dữ liệu và trực quan hóa dữ liệu đa cấu trúc, bao gồm cả các công cụ hỗ trợ như Power BI, Tableau, matplotlib, seaborn, cũng như các thuật toán xử lý và thống kê dữ liệu được sử dụng phổ biến hiện nay.

Phương pháp nghiên cứu thực nghiệm

Sau khi đã xây dựng được nền tảng lý luận vững chắc, nhóm tiến hành triển khai mô hình nghiên cứu bằng các bước thực nghiệm cụ thể:

- Đề xuất mô hình trực liên thông dữ liệu phù hợp với bối cảnh thực tế tại Việt Nam, đặc biệt trong môi trường giáo dục. Mô hình được thiết kế dựa trên nguyên lý hoạt động của X-Road, đồng thời điều chỉnh để đáp ứng các yêu cầu kỹ thuật, tổ chức và vận hành phù hợp với điều kiện của các cơ sở giáo dục.

- Triển khai mô hình thực nghiệm trên môi trường thử nghiệm với dữ liệu mô phỏng, nhằm kiểm chứng khả năng tích hợp và chia sẻ dữ liệu đa cấu trúc từ nhiều nguồn khác nhau (bao gồm dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc).

- Xây dựng các kịch bản phân tích và trực quan hóa dữ liệu trên nền tảng dữ liệu thu thập được nhằm kiểm tra khả năng sinh báo cáo thống kê tự động, phục vụ cho nhu cầu quản lý và ra quyết định của các đơn vị sử dụng dữ liệu.

Phương pháp phân tích và đánh giá mô hình

Đây là bước quan trọng nhằm đo lường tính hiệu quả và thực tiễn của mô hình được đề xuất. Các nội dung chính bao gồm:

- Phân tích hiệu năng của mô hình, thông qua các chỉ số như tốc độ xử lý, khả năng đáp ứng truy cập đồng thời, độ trễ trong quá trình truyền tải và chia sẻ dữ liệu.

- Đánh giá tính khả thi và khả năng mở rộng, bao gồm khả năng tích hợp với các hệ thống sẵn có, mức độ linh hoạt trong việc mở rộng quy mô và phạm vi ứng dụng.

- So sánh mô hình đề xuất với các mô hình hiện hữu, nhằm làm nổi bật những điểm mới, điểm mạnh của giải pháp, cũng như chỉ ra các hạn chế cần được cải tiến trong các nghiên cứu tiếp theo.

- Lấy ý kiến chuyên gia và người dùng thử nghiệm để thu thập phản hồi, từ đó hiệu chỉnh giải pháp cho sát với nhu cầu thực tế hơn.

CHƯƠNG 1: TỔNG QUAN VỀ TRỰC LIÊN THÔNG

DỮ LIỆU VÀ DỮ LIỆU ĐA CẤU TRÚC

1.1. Liên thông dữ liệu, trực liên thông dữ liệu cùng các phương pháp đăng ký dịch vụ và chia sẻ dịch vụ thông qua trực liên thông dữ liệu.

Liên thông dữ liệu là khả năng cho phép các hệ thống thông tin khác nhau trao đổi và sử dụng dữ liệu một cách thống nhất. Ba loại liên thông chính bao gồm:

- Liên thông cú pháp: đảm bảo dữ liệu có thể trao đổi thông qua các giao thức, định dạng như XML, JSON, kể cả khi các hệ thống sử dụng nền tảng công nghệ khác nhau.
- Liên thông cấu trúc: yêu cầu các hệ thống hiểu và sử dụng đúng định dạng dữ liệu chung, đảm bảo nội dung truyền tải có cùng cấu trúc logic.
- Liên thông ngữ nghĩa: đảm bảo ý nghĩa của dữ liệu được các hệ thống tiếp nhận hiểu và diễn giải chính xác.

Trực liên thông dữ liệu đóng vai trò trung gian, chịu trách nhiệm kết nối, xác thực, định tuyến, và giám sát việc truyền tải dữ liệu giữa các hệ thống một cách bảo mật, minh bạch và hiệu quả.

1.2. Dữ liệu đa cấu trúc, các loại dữ liệu đa cấu trúc và cách thức lưu trữ, quản lý, phân tích dữ liệu đa cấu trúc.

Dữ liệu đa cấu trúc trong giáo dục bao gồm:

- Dữ liệu có cấu trúc: bảng điểm, hồ sơ sinh viên, thời khóa biểu – dễ truy vấn, quản lý.
- Dữ liệu bán cấu trúc: phản hồi khảo sát dạng JSON, file log, email – có định dạng nhưng không cố định.
- Dữ liệu phi cấu trúc: video bài giảng, tài liệu scan, bài tập viết tay – khó lưu trữ và phân tích hơn.

Các giải pháp lưu trữ phổ biến hiện nay:

- Data Lake: lưu trữ dạng thô, không cần schema cố định.
- Data Warehouse: lưu trữ có cấu trúc, dùng cho phân tích BI.
- Object Storage (MinIO, S3): lưu trữ phi cấu trúc linh hoạt.

1.3. Phương pháp chia sẻ dữ liệu đa cấu trúc thông qua trực liên thông dữ liệu.

Việc chia sẻ dữ liệu đa cấu trúc đòi hỏi một cơ chế tổng hợp, phân tách và phân phối dữ liệu linh hoạt, phù hợp với từng loại dữ liệu cụ thể. Trực liên thông dữ liệu như X-Road thực hiện điều này thông qua các thành phần kỹ thuật chính như sau:

- Định danh và phân loại dữ liệu: Mỗi loại dữ liệu (có cấu trúc, bán cấu trúc, phi cấu trúc) được gán định danh duy nhất và phân loại rõ ràng, nhờ đó có thể áp dụng chính sách xử lý phù hợp.

- Chuẩn hóa dữ liệu đầu vào: Dữ liệu từ các hệ thống nguồn sẽ được chuẩn hóa qua các bước ánh xạ schema, chuyển đổi định dạng (từ XML sang JSON hoặc CSV), loại bỏ dữ liệu dư thừa và phát hiện lỗi logic.

- Thiết lập các luồng trao đổi (data exchange flows): Các kênh chia sẻ dữ liệu được thiết lập thông qua API hoặc dịch vụ web (Web Services) với định nghĩa rõ ràng đầu vào – đầu ra, thời gian xử lý, và cấu trúc dữ liệu đi kèm.

- Bảo mật và xác thực: Dữ liệu được chia sẻ chỉ sau khi trải qua các bước xác thực danh tính, kiểm tra quyền truy cập và mã hóa. X-Road sử dụng kỹ thuật ký số và mã hóa TLS để đảm bảo an toàn dữ liệu.

- Tách riêng dữ liệu theo nhóm xử lý: Các nhóm dữ liệu phi cấu trúc được lưu trữ qua hệ thống object storage, trong khi dữ liệu bán cấu trúc có thể được xử lý thông qua JSON parser hoặc các công cụ như NiFi hoặc Airflow.

- Ghi nhận và theo dõi (logging & auditing): Mọi hoạt động chia sẻ dữ liệu đều được log lại để phục vụ kiểm tra, đánh giá và minh bạch hóa truy xuất.

Phương pháp này giúp đảm bảo dữ liệu từ nhiều nguồn khác nhau có thể được chia sẻ, khai thác và phân tích một cách nhất quán, an toàn và hiệu quả, đồng thời vẫn phù hợp với đặc tính đa dạng và phức tạp của dữ liệu giáo dục hiện nay.

1.4. Phương pháp tổng hợp, thống kê, trực quan hóa dữ liệu.

Trong bối cảnh chuyển đổi số giáo dục đại học, việc xây dựng hệ thống xử lý và phân tích dữ liệu hiệu quả là điều kiện tiên quyết nhằm nâng cao chất lượng quản lý và hỗ trợ ra quyết định. Đề án triển khai quy trình tổng hợp, thống kê và trực quan hóa dữ

liệu thông qua tự động hóa, với mục tiêu đảm bảo tính chính xác, kịp thời và khả năng mở rộng của hệ thống phân tích dữ liệu.

1.4.1. Tổng hợp dữ liệu

Dữ liệu từ các hệ thống phân tán như SIS (quản lý học vụ), LMS (học tập trực tuyến), y tế học đường, phản hồi người học, v.v... được thu thập qua trực liên thông và lưu trữ trong hồ dữ liệu (Data Lake). Việc tổng hợp được thực hiện theo nhiều chiều dữ liệu như thời gian, đối tượng, chuyên ngành hoặc cấp quản lý. Các công cụ ETL tự động sẽ chuẩn hóa, loại bỏ trùng lặp và định dạng lại dữ liệu để phục vụ cho các bước xử lý tiếp theo.

1.4.2. Thống kê dữ liệu

Sau khi dữ liệu được chuẩn hóa, các kỹ thuật thống kê được áp dụng để phân tích các đặc trưng cơ bản như số lượng sinh viên, tỷ lệ tốt nghiệp, mức độ hoàn thành học phần, số lượt sử dụng dịch vụ,... Việc này được thực hiện định kỳ hoặc theo yêu cầu của người dùng cuối. Thống kê được triển khai trên cả dữ liệu có cấu trúc (số liệu học vụ) và bán/phi cấu trúc (kết quả khảo sát, phản hồi dạng văn bản).

1.4.3. Trực quan hóa dữ liệu

Kết quả thống kê được trình bày thông qua các công cụ trực quan như biểu đồ, bản đồ nhiệt, bảng động, nhằm hỗ trợ người dùng không chuyên cũng có thể hiểu và khai thác thông tin. Superset là công cụ chính được đề xuất cho trực quan hóa, nhờ tính linh hoạt, giao diện thân thiện và khả năng tương tác cao. Dashboard có thể thiết kế theo từng nhóm người dùng: lãnh đạo, giảng viên, phòng khảo thí,...

Tóm lại, phương pháp tổng hợp – thống kê – trực quan hóa trong đề án giúp tạo ra một chu trình dữ liệu hoàn chỉnh, từ thu thập đến khai thác, đáp ứng yêu cầu ra quyết định dựa trên dữ liệu trong môi trường giáo dục số

CHƯƠNG 2: ĐỀ XUẤT PHƯƠNG ÁN XÂY DỰNG TRỰC LIÊN THÔNG DỮ LIỆU

2.1. Trục liên thông dữ liệu X-Road phục vụ trao đổi dữ liệu đa cấu trúc giữa các bên

Mô hình tổng thể của hệ thống trục liên thông dữ liệu được thiết kế với mục tiêu tích hợp dữ liệu từ nhiều nguồn khác nhau, xử lý linh hoạt dữ liệu đa cấu trúc, đảm bảo bảo mật và hỗ trợ ra quyết định thông minh. Các thành phần chính bao gồm:

Nguồn dữ liệu (Data Sources):

- Bao gồm hệ thống quản lý đào tạo (LMS, SIS), hệ thống khảo sát, phản hồi sinh viên, dữ liệu khảo cứu từ các biểu mẫu điện tử, và các hệ thống quản lý nội dung học tập (CMS).
- Dữ liệu có thể ở nhiều định dạng khác nhau: CSDL quan hệ (SQL), API REST/GraphQL, tệp (CSV, Excel, PDF), video/audio, log máy chủ.

Thành phần kết nối (Connectors):

- Các kết nối dữ liệu bao gồm API Connector (thu thập dữ liệu qua REST, GraphQL), Database Connector (ODBC, JDBC), File Collector (đọc dữ liệu từ thư mục, SFTP, cloud).
- Các kết nối này sẽ được lập lịch thu thập và định danh dữ liệu đầu vào.

Trục liên thông dữ liệu (Integration Hub - sử dụng X-Road):

- Đảm nhận nhiệm vụ điều phối dữ liệu giữa các hệ thống.
- Bao gồm Central Server, Security Server, và thành phần truy vết.
- Dữ liệu từ các nguồn được xác thực, ký số và mã hóa trước khi luân chuyển.
- Cho phép cấu hình các chính sách phân quyền, xác thực dịch vụ theo mô hình bảo mật đa lớp.

Kho lưu trữ dữ liệu đa cấu trúc (Data Lake - sử dụng MinIO):

- Cho phép lưu trữ dữ liệu thô từ nhiều nguồn với định dạng khác nhau.
- Tổ chức dữ liệu theo lớp raw, cleaned, curated để phục vụ cho ETL và phân tích.

- Hỗ trợ lưu trữ dữ liệu lớn, đảm bảo khả năng mở rộng và truy xuất hiệu quả.

Xử lý và chuyển đổi dữ liệu (ETL - sử dụng Apache Airflow):

- Dữ liệu được nạp từ Data Lake, sau đó thực hiện các thao tác làm sạch, chuyển đổi, chuẩn hóa và trích xuất đặc trưng.
- Sử dụng DAG (Directed Acyclic Graph) để định nghĩa luồng công việc.
- Cho phép theo dõi lịch sử chạy, retry khi lỗi và gửi cảnh báo.

Hệ thống phân tích và hiển thị dữ liệu (Analytics Layer):

- Dữ liệu sau xử lý được đưa vào PostgreSQL hoặc Elasticsearch phục vụ truy vấn phân tích.
- Công cụ trực quan hóa: Superset, Power BI, hoặc chatbot NLP.
- Xây dựng bảng điều khiển theo chức năng: theo dõi tiến độ học tập, đánh giá giảng viên, báo cáo cấp lãnh đạo.

Mô hình này không chỉ đáp ứng nhu cầu tích hợp và chia sẻ dữ liệu hiệu quả mà còn đảm bảo bảo mật, dễ mở rộng, và thân thiện với người dùng. Ngoài ra, nó cho phép mở rộng với các mô-đun học máy để khai thác dữ liệu dự đoán, gợi ý học tập hoặc đánh giá tự động.

2.2 Thu thập biến đổi dữ liệu (Extract Transform Load - ETL).

Để đảm bảo dữ liệu từ các hệ thống khác nhau được xử lý đồng bộ, thống nhất và phục vụ cho các mục tiêu phân tích, hệ thống cần triển khai một quy trình ETL (Extract – Transform – Load) hiệu quả, linh hoạt và có khả năng mở rộng.

Giai đoạn Extract (Trích xuất):

- Dữ liệu được thu thập từ các hệ thống nguồn như hệ quản trị cơ sở dữ liệu (PostgreSQL, MySQL), API (LMS, khảo sát), hoặc tệp lưu trữ (CSV, Excel, JSON, XML).
- Quá trình trích xuất được lập lịch thực hiện định kỳ hoặc theo sự kiện, đảm bảo dữ liệu được cập nhật liên tục và đúng thời điểm.
- Kết nối sử dụng thông qua API Connector, JDBC/ODBC, hoặc giao thức FTP/SFTP cho dữ liệu phi cấu trúc.

Giai đoạn Transform (Chuyển đổi):

- Là bước trung gian quan trọng trong xử lý dữ liệu. Dữ liệu thô được làm sạch, chuẩn hóa, mã hóa định danh, phát hiện và loại bỏ lỗi, khuyết thiếu.
- Các kỹ thuật xử lý bao gồm: đổi kiểu dữ liệu, hợp nhất dữ liệu từ nhiều bảng, ánh xạ giá trị tương đương (mapping), phân loại, chuẩn hóa cấu trúc dữ liệu.
- Có thể bổ sung các bước nâng cao như lọc dữ liệu theo ngữ cảnh, xử lý NLP cho phản hồi tự do, và nhận dạng thực thể (NER).

Giai đoạn Load (Nạp):

- Sau khi xử lý, dữ liệu được nạp vào kho lưu trữ đích như Data Warehouse (PostgreSQL) hoặc hệ thống phân tích (ElasticSearch).
- Với dữ liệu chưa chuẩn hóa, có thể lưu trữ tạm thời trong Data Lake để phân tích sau.
- Nạp dữ liệu hỗ trợ theo hai hình thức: toàn bộ (full load) hoặc chỉ phần thay đổi (incremental load).

Công cụ sử dụng – Apache Airflow:

- Là nền tảng mã nguồn mở hỗ trợ lập lịch và quản lý các luồng công việc ETL thông qua DAG (Directed Acyclic Graph).
- Airflow cho phép người dùng định nghĩa các bước ETL bằng Python, tích hợp chặt chẽ với các hệ thống lưu trữ và dịch vụ API.
- Hệ thống hỗ trợ tự động hóa luồng công việc, thông báo lỗi, giám sát trạng thái và ghi log chi tiết.
- Có thể mở rộng để tích hợp với Spark, Hadoop, hoặc các công cụ phân tích dữ liệu lớn khác.

Tích hợp bảo mật và theo dõi:

- Mỗi luồng ETL đều được theo dõi và ghi log để phục vụ mục đích kiểm toán.
- Cơ chế xác thực truy cập và giới hạn quyền thực thi được thiết lập theo vai trò người dùng (RBAC).

Tổng thể, quy trình ETL đóng vai trò then chốt trong việc đảm bảo tính chính xác, toàn vẹn và giá trị sử dụng của dữ liệu trong hệ thống liên thông. Việc triển khai ETL hiệu quả giúp các tổ chức giáo dục có thể sử dụng dữ liệu để phân tích, báo cáo và ra quyết định một cách linh hoạt, chính xác và kịp thời.

2.3 Hồ dữ liệu tổng hợp và phân tích dữ liệu đa cấu trúc (Datalake).

Hồ dữ liệu (Data Lake) là một thành phần thiết yếu trong hệ thống trực liên thông dữ liệu, đóng vai trò lưu trữ tập trung cho toàn bộ dữ liệu đa cấu trúc từ nhiều nguồn khác nhau. Không giống với kho dữ liệu truyền thống (Data Warehouse), Data Lake cho phép lưu trữ dữ liệu ở định dạng gốc, bao gồm dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc, giúp tối ưu khả năng mở rộng, phân tích và tái sử dụng dữ liệu.

Trong đề án, MinIO được đề xuất làm nền tảng triển khai Data Lake nhờ vào các ưu điểm nổi bật:

- Tương thích S3: MinIO có thể hoạt động như một kho lưu trữ đối tượng S3, dễ dàng tích hợp với các công cụ phân tích và ETL hiện đại.
- Hiệu suất cao và chi phí thấp: Khả năng vận hành trên hạ tầng phần cứng phổ thông giúp giảm chi phí đầu tư.
- Mã nguồn mở và bảo mật tốt: Hỗ trợ mã hóa dữ liệu, kiểm soát truy cập và có khả năng triển khai linh hoạt từ máy chủ vật lý đến môi trường đám mây lai (hybrid cloud).
- Tích hợp với Apache Airflow: Hỗ trợ tự động hóa quy trình ETL từ thu thập, làm sạch, đến lưu trữ dữ liệu.

MinIO được khởi tạo với các thư mục lưu trữ (bucket) theo cấp độ phân tầng như raw-data, clean-data, dashboards, cho phép tổ chức và truy xuất dữ liệu hiệu quả. Dữ liệu được thu thập từ các hệ thống qua trực liên thông sẽ được lưu vào raw-data, xử lý tại các công cụ ETL như Airflow, sau đó được chuyển sang clean-data, sẵn sàng cho trực quan hóa và phân tích.

Kết luận, việc sử dụng MinIO để xây dựng hồ dữ liệu giúp tăng cường khả năng lưu trữ linh hoạt, giảm chi phí và đảm bảo khả năng tích hợp trong hệ sinh thái dữ liệu hiện đại. Điều này đặc biệt phù hợp với bối cảnh giáo dục và chính

phủ điện tử tại Việt Nam, nơi các tổ chức đang chuyển dịch mạnh mẽ sang mô hình quản trị dựa trên dữ liệu

2.4 Trực quan hóa dữ liệu.

Trực quan hóa dữ liệu đóng vai trò then chốt trong việc khai thác giá trị dữ liệu đa cấu trúc, giúp biến dữ liệu thô thành thông tin dễ hiểu và hỗ trợ ra quyết định hiệu quả. Trong đề án, Superset – một công cụ mã nguồn mở – được lựa chọn để triển khai hệ thống trực quan hóa dữ liệu nhờ vào các ưu điểm nổi bật:

- **Miễn phí và cộng đồng phát triển mạnh:** So với các giải pháp thương mại như Tableau hay Power BI, Superset giúp giảm chi phí đầu tư mà vẫn đảm bảo đầy đủ tính năng cần thiết.
- **Giao diện thân thiện và linh hoạt:** Hỗ trợ cả người không chuyên với chức năng kéo-thả và người chuyên sâu với SQL Lab.
- **Hỗ trợ nhiều loại biểu đồ và Dashboard tương tác:** Từ biểu đồ thanh, tròn, đường, bản đồ nhiệt đến biểu đồ phân tán, vùng, cây, giúp người dùng dễ dàng tùy chỉnh và chia sẻ kết quả phân tích.
- **Tích hợp bảo mật và phân quyền:** Cho phép quản lý truy cập theo vai trò người dùng, đảm bảo an toàn dữ liệu khi triển khai trong tổ chức giáo dục hoặc cơ quan nhà nước.
- **Khả năng kết nối với nhiều nguồn dữ liệu:** Có thể truy vấn trực tiếp từ các hệ thống như PostgreSQL, MySQL, MinIO hoặc qua các tập dữ liệu ETL đã xử lý.

Cụ thể trong mô hình đề án, Superset được sử dụng để trực quan hóa dữ liệu sau khi đã được thu thập, làm sạch và lưu trữ trong MinIO thông qua các pipeline ETL của Airflow. Kết quả trực quan hóa gồm các biểu đồ phản ánh xu hướng điểm số, phân tích lớp học, hoặc đánh giá hoạt động học vụ... được trình bày trên các dashboard theo thời gian thực.

Tóm lại, việc tích hợp Superset vào hệ thống trực liên thông dữ liệu không chỉ nâng cao hiệu quả phân tích dữ liệu mà còn tạo điều kiện để người dùng cuối dễ dàng tiếp cận và sử dụng dữ liệu cho các mục tiêu quản trị và học thuật

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

3.1. Chạy thực nghiệm mô hình đề xuất.

Mục tiêu của phần thử nghiệm là đánh giá tính khả thi và hiệu quả của mô hình hạ tầng dữ liệu mã nguồn mở trong môi trường đại học. Việc triển khai mô hình thực nghiệm giúp kiểm chứng khả năng tích hợp, liên thông và xử lý dữ liệu từ nhiều hệ thống khác nhau như SIS, LMS, hệ thống y tế học đường, khảo sát phản hồi và dịch vụ công nội bộ.

3.1.1. Phạm vi thử nghiệm

Thử nghiệm tập trung vào các lĩnh vực dữ liệu thiết yếu trong trường đại học, bao gồm:

- **Dữ liệu người học:** hồ sơ sinh viên, kết quả học tập, tình trạng học vụ.
- **Dữ liệu y tế học đường:** lịch sử tiêm chủng, khám sức khỏe.
- **Dữ liệu phản hồi và khảo sát:** đánh giá học phần, phản ánh sinh viên.
- **Dữ liệu hành chính – dịch vụ công:** đăng ký học phần, xác nhận giấy tờ, thủ tục hành chính.

Mục đích là xử lý các nguồn dữ liệu phân tán này theo hướng chuẩn hóa và thống nhất, giúp tăng cường khả năng khai thác, truy vấn và phân tích.

3.1.2. Cài đặt và cấu hình các thành phần

- **X-Road:** Triển khai một Central Server và hai Security Server để giả lập trao đổi dữ liệu giữa hai hệ thống nội bộ, thử nghiệm chia sẻ dịch vụ XML API (danh sách học sinh, điểm số).
- **Airflow:** Cài đặt qua Docker Compose, cấu hình các DAG ETL để thu thập dữ liệu từ API hoặc tệp JSON/CSV, làm sạch và lưu vào hệ thống MinIO.
- **MinIO:** Thiết lập các bucket như raw-data, clean-data, dashboards, đồng thời cấu hình phân quyền và kiểm tra dữ liệu lưu trữ qua giao diện web.
- **OpenMetadata:** Kết nối với Airflow để tự động ghi nhận dòng dữ liệu, gắn metadata cho bảng dữ liệu, phân quyền truy cập theo nhóm người dùng (giảng viên, lãnh đạo, quản trị viên).

Việc thực nghiệm này tạo tiền đề cho việc triển khai mô hình hạ tầng dữ liệu toàn diện trong các trường đại học, phục vụ cho mục tiêu dài hạn về phân tích học tập, cá nhân hóa học tập và tích hợp trí tuệ nhân tạo trong quản trị giáo dục.

3.2. Các kết quả mà thử nghiệm mang lại.

Quá trình triển khai mô hình trực liên thông dữ liệu trong môi trường giả lập đã mang lại nhiều kết quả tích cực, minh chứng cho tính khả thi và hiệu quả của giải pháp được đề xuất.

3.2.1. Trực liên thông dữ liệu X-Road

Kết nối và chia sẻ dịch vụ thành công giữa các hệ thống giả lập thông qua X-Road.

Thiết lập được danh mục các đơn vị tham gia liên thông dữ liệu, phân loại dịch vụ được chia sẻ.

Cấu hình và kiểm thử thành công dịch vụ API cho phép chia sẻ danh sách học sinh, điểm số, lớp học... giữa các đơn vị.

Kiểm soát quyền truy cập hiệu quả thông qua cơ chế xác thực và chữ ký số, đảm bảo an toàn dữ liệu khi truyền qua trực.

3.2.2. Hệ thống ETL và hồ dữ liệu MinIO

Airflow vận hành ổn định với các DAG ETL tự động:

Thu thập dữ liệu từ nhiều nguồn (API, file CSV/JSON).

Làm sạch dữ liệu và chuẩn hóa định dạng.

Lưu trữ dữ liệu theo cấu trúc phân tầng trong MinIO (raw-data, clean-data, dashboards).

MinIO hoạt động như một Datalake hiệu quả:

Cho phép lưu trữ dữ liệu đa cấu trúc với chi phí thấp.

Dễ tích hợp với công cụ phân tích và trực quan hóa.

3.2.3. Trực quan hóa với Superset

Tạo được các biểu đồ thống kê phản ánh dữ liệu sinh viên như:

Số lượng sinh viên theo khoa/lớp.

Tỷ lệ hoàn thành học phần.

Biểu đồ điểm số trung bình theo thời gian...

Dashboard tương tác được xây dựng phục vụ nhiều nhóm người dùng: giảng viên, lãnh đạo, phòng đào tạo.

3.2.4. Quản lý metadata với OpenMetadata

Ghi nhận và theo dõi dòng dữ liệu (data lineage) tự động.

Gắn metadata cho các bảng dữ liệu, bao gồm mô tả, nguồn gốc, thời gian cập nhật.

Phân quyền truy cập metadata theo nhóm người dùng.

Kết luận, mô hình thử nghiệm cho thấy giải pháp đề xuất hoàn toàn có thể ứng dụng trong môi trường đại học, không chỉ hỗ trợ liên thông và quản trị dữ liệu hiệu quả mà còn tạo tiền đề cho các bước nâng cao như cá nhân hóa học tập và phân tích học vụ dựa trên AI trong tương lai

3.3. Các mục cần cải thiện trong tương lai.

Dựa trên quá trình thực nghiệm và đánh giá mô hình trực liên thông dữ liệu, đề án đã xác định một số hướng cải tiến quan trọng để nâng cao hiệu quả và khả năng triển khai trong thực tế:

3.3.1. Khả năng mở rộng và tối ưu hiệu năng

- Cần nâng cấp hạ tầng để đáp ứng nhu cầu liên thông dữ liệu ở quy mô lớn hơn, như mở rộng giữa các trường đại học hoặc toàn ngành.
- Tích hợp các công nghệ hỗ trợ mở rộng như Docker, Kubernetes, Load Balancer nhằm đảm bảo hệ thống hoạt động ổn định khi có nhiều đơn vị truy cập đồng thời.

3.3.2. Tăng cường bảo mật và kiểm soát truy cập

- Hiện tại mô hình mới áp dụng kiểm soát truy cập theo người dùng/nhóm. Trong tương lai, cần triển khai phân quyền chi tiết hơn, ví dụ: theo thuộc tính hoặc ngưỡng truy cập.
- Bổ sung hệ thống giám sát truy cập, ghi log và cảnh báo bảo mật theo thời gian thực để tăng tính an toàn và minh bạch.

3.3.3. Cải thiện trải nghiệm người dùng

- Tùy biến giao diện Superset, OpenMetadata và dashboard theo nhu cầu sử dụng của từng nhóm người dùng (giảng viên, quản trị viên, lãnh đạo...).

- Tăng tính tương tác và thân thiện của dashboard, đặc biệt với người không chuyên.

3.3.4. Tích hợp các công nghệ mới

- Kết hợp các kỹ thuật trí tuệ nhân tạo (AI) như học máy (machine learning), dự báo học lực, phát hiện sớm nguy cơ bỏ học để nâng cao giá trị khai thác dữ liệu.
- Hỗ trợ truy vấn dữ liệu bằng ngôn ngữ tự nhiên thông qua chatbot hoặc trợ lý ảo.

3.3.5. Chuẩn hóa và mở rộng quy trình chia sẻ dữ liệu

- Cần xây dựng bộ chuẩn chung về từ điển dữ liệu, định dạng API và quy trình liên thông giữa các hệ thống khác nhau để dễ dàng tích hợp và chia sẻ dữ liệu trong quy mô toàn ngành.

Hạn chế của Quá trình Thử nghiệm

Dù hệ thống thử nghiệm đã vận hành ổn định ở mức độ giả lập, một số điểm hạn chế và tiềm năng cải thiện đã được ghi nhận trong quá trình thực nghiệm:

Khả năng mở rộng quy mô (scalability): Hiện tại, hệ thống được triển khai trên môi trường thử nghiệm với dữ liệu và tần suất xử lý còn giới hạn. Khi áp dụng vào thực tế với lượng dữ liệu lớn hơn cần kiểm thử khả năng mở rộng của Airflow, MinIO và X-Road, đồng thời đánh giá khả năng chịu tải và thời gian xử lý thực tế của các module DAG ETL.

Tính sẵn sàng và dự phòng (high availability): Các thành phần hiện đang được triển khai theo cấu hình đơn (single instance), chưa có cơ chế dự phòng (failover) hoặc cân bằng tải (load balancing). Trong môi trường triển khai thực tế, cần thiết lập hệ thống HA cho các thành phần quan trọng như X-Road Security Server, Airflow Scheduler và MinIO.

Tự động hóa và kiểm thử dữ liệu (data validation): Quy trình làm sạch dữ liệu hiện mới ở mức cơ bản (chuyển đổi định dạng, chuẩn hóa tên). Trong tương lai, cần bổ sung các bước kiểm thử chất lượng dữ liệu (data quality checks), phát hiện thiếu giá trị, dữ liệu bất thường hoặc trùng lặp trước khi lưu vào hệ thống.

Quản lý quyền truy cập chi tiết hơn: Dù đã có phân quyền truy cập metadata theo nhóm người dùng trong OpenMetadata, hiện vẫn chưa tích hợp toàn bộ hệ thống với cơ chế xác thực tập trung (ví dụ: LDAP, OAuth2). Việc quản lý phân quyền chi tiết theo vai trò người dùng trong từng lớp học, từng loại dữ liệu sẽ cần hoàn thiện thêm.

Giám sát và cảnh báo: Cần bổ sung các công cụ giám sát (như Prometheus hoặc Grafana) để theo dõi sức khỏe hệ thống, dung lượng lưu trữ, và trạng thái các DAG. Việc tích hợp cảnh báo (alerts) sẽ giúp phát hiện sớm lỗi hệ thống hoặc dữ liệu.

Kết luận Rút ra từ Kết quả Thực nghiệm

Đây là một mô hình hệ thống dữ liệu hiện đại, với kiến trúc được thiết kế chuyên biệt, mang lại những lợi ích đáng kể, cả về mặt kỹ thuật và chiến lược trong quản trị cũng như khai thác dữ liệu trong bối cảnh chuyển đổi số toàn cầu. Trước hết, mô hình này sở hữu khả năng kết nối linh hoạt và mở rộng dễ dàng thông qua việc tích hợp đa dạng các nguồn dữ liệu bằng các đầu nối chuyên dụng (Connectors), cho phép hệ thống nhanh chóng thích nghi với sự xuất hiện của các nguồn dữ liệu mới từ IoT, thiết bị di động, mạng xã hội hay các hệ thống giao dịch phức tạp, đảm bảo thu nạp dữ liệu toàn diện và liên tục. Thứ hai, mô hình giúp đạt được dữ liệu tập trung, đồng bộ và toàn diện nhờ vào trục liên thông dữ liệu và kho lưu trữ trung tâm (Data Lake và Data Warehouse). Điều này loại bỏ hoàn toàn tình trạng "cô lập dữ liệu" (data silo), cho phép người dùng ở mọi cấp độ truy cập và khai thác dữ liệu một cách nhất quán, từ đó giảm thiểu trùng lặp và tối ưu hóa hiệu quả sử dụng tài nguyên dữ liệu. Thứ ba, việc ứng dụng công cụ điều phối dữ liệu (Data Orchestrator) và quy trình ETL tự động hóa mạnh mẽ giúp giảm đáng kể công sức vận hành thủ công và tiết kiệm chi phí dài hạn, đồng thời nâng cao độ chính xác và tính ổn định của quy trình xử lý dữ liệu, giảm thiểu rủi ro do sai sót của con người. Điều này không chỉ tăng cường tính minh bạch trong hoạt động của tổ chức mà còn tạo điều kiện thuận lợi cho nghiên cứu, đổi mới sáng tạo và sự tham gia của cộng đồng vào quản trị công, phù hợp với xu hướng dữ liệu mở đang ngày càng phát triển trên phạm vi toàn cầu. Tóm lại, những lợi ích này khẳng định vai trò then chốt của mô hình hệ thống dữ liệu hiện đại trong việc tối đa hóa giá trị từ dữ liệu, nâng cao năng lực ra quyết định và thúc đẩy sự phát triển bền vững của tổ chức.

KẾT LUẬN

Trong bối cảnh chuyển đổi số đang diễn ra mạnh mẽ trên toàn cầu, đặc biệt trong lĩnh vực giáo dục, việc xây dựng một hạ tầng công nghệ hiệu quả để quản lý và khai thác dữ liệu trở thành một yêu cầu cấp thiết. Thông qua nghiên cứu này, đề án đã làm rõ tầm quan trọng của dữ liệu đa cấu trúc trong đại học số, đồng thời chỉ ra những thách thức trong việc liên thông, lưu trữ và phân tích các loại dữ liệu không đồng nhất đang tồn tại phân tán trên nhiều hệ thống khác nhau.

Việc đề xuất và xây dựng một trục liên thông dữ liệu dựa trên nền tảng như X-Road không chỉ mang lại khả năng kết nối liền mạch giữa các hệ thống thông tin mà còn tạo điều kiện thuận lợi để khai thác tối đa giá trị của dữ liệu. Trục liên thông đóng vai trò quan trọng trong việc đồng bộ hóa, bảo mật và chuẩn hóa dữ liệu, từ đó nâng cao hiệu quả trong công tác quản lý, giảng dạy, nghiên cứu.

Ngoài ra, việc kết hợp khả năng trực quan hóa và phân tích dữ liệu sẽ giúp các trường đại học có cái nhìn toàn diện hơn về hoạt động của mình, từ đó đưa ra các chính sách điều hành hiệu quả, linh hoạt và minh bạch hơn. Kết quả nghiên cứu cũng cho thấy tiềm năng to lớn của việc áp dụng trục liên thông dữ liệu tại các trường đại học. Đây là một đề án khá mới nên những ý kiến và thiết kế trục liên thông dữ liệu của đề án mới chỉ ở mức tương đối. Hơn nữa, do giới hạn về thời gian, điều kiện công tác nên không tránh khỏi những hạn chế và khuyết điểm. Tác giả rất mong nhận được sự đóng góp quý báu của các nhà khoa học, quý thầy cô để đề án được hoàn thiện hơn.