HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



TRẦN ĐỨC LONG

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI - 2025

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



TRẦN ĐỨC LONG

TÍCH HỢP DỮ LIỆU CHO HỆ THỐNG QUẢN LÝ CÁN BỘ, VIÊN CHỨC CỦA VIỆN NĂNG LƯỢNG NGUYÊN TỬ

CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN

MÃ SỐ: 8.48.01.04 (Hệ thống thông tin)

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. Trần Thiện Chính

HÀ NỘI - 2025

LỜI CAM ĐOAN

Tôi xin cam đoan rằng, đây là công trình nghiên cứu của tôi trong đó có sự giúp đỡ rất lớn của thầy hướng dẫn là TS. Trần Thiện Chính. Các nội dung nghiên cứu và kết quả trong đề tài này là hoàn toàn trung thực.

Trong luận văn, tôi có tham khảo đến một số tài liệu, một số bài báo và công bố khoa học của một số tác giả đã được liệt kê tại phần Tài liệu tham khảo ở cuối luận văn này.

Hà Nội, ngày tháng 12 năm 2025

Tác giả

Trần Đức Long

LỜI CẢM ƠN

Đầu tiên tác giả xin trân trọng cảm ơn tất cả các thầy cô giảng viên các Giáo sư, Phó Giáo sư, Tiến Sĩ cùng toàn thể các thầy, cô giáo, giảng viên Học viện Công nghệ Bưu chính Viễn thông đã hết mình giảng dạy, tạo điều kiện thuận lợi cho tác giả trong suốt quá trình học tập tại Học viện.

Tác giả xin trân trọng cảm ơn TS. Trần Thiện Chính và các thầy cô trong hội đồng khoa học đã giành thời gian hướng dẫn, chỉ bảo giúp luận văn này được hoàn thiện.

Tác giả xin trân trọng cảm ơn Ban giám hiệu Học viện, Khoa Sau Đại học, Khoa Công nghệ thông tin đã quan tâm, giúp đỡ tạo điều kiện trong quá trình nghiên cứu và học tập tại Học viện.

Tác giả xin chân thành cảm ơn gia đình, bạn bè, đồng nghiệp đã luôn đọc viên và tiếp thêm nghị lực để tác giả có thể hoàn thành khóa học và luận văn này.

Cuối cùng, xin gửi lời tri ân đến vợ, người bạn đồng hành luôn sát cánh, ủng hộ tôi.

Mặc dù đã cố gắng, tuy nhiên do nhân bất thập toàn nên việc sai sót là không thể tránh khỏi. Tác giả rất mong được sự nhận xét, góp ý, từ các nhà khoa học, các thành viên Hội đồng và từ các thầy cô giảng viên.

Xin trân trọng cảm ơn!

Hà Nội, tháng 12 năm 2025

Tác giả

Trần Đức Long

MỤC LỤC

LỜI CAM ĐOAN i
LỜI CẢM ƠN ii
MỤC LỤC iii
DANH MỤC VIẾT TẮT v
DANH MỤC BẢNG, BIỂU, HÌNH MINH HOẠ vi
MỞ ĐẦU 1
CHƯƠNG 1: TỔNG QUAN VỀ VẤN ĐỀ NGHIÊN CỨU 6
1.1. Cách tiếp cận vấn đề quản lý và khai thác dữ liệu 6
1.1.1. Tích hợp dữ liệu (Data Integration) 7
1.1.2. Tổng hợp dữ liệu (Data Aggregation) 7
1.1.3. Hợp nhất dữ liệu (Data Consolidation) 8
1.1.4. So sánh giữa Tích hợp dữ liệu (Data Integration), Tổng hợp dữ liệu (Data Aggregation) và Hợp nhất dữ liệu (Data Consolidation) 8
1.2. Giới thiệu về Viện Năng lượng nguyên tử Việt Nam 10
1.2.1. Công tác quản lý cán bộ tại Viện 12
1.2.2. Hệ thống quản lý công chức, viên chức trong cơ quan nhà nước Việt Nam 14
1.3. Các hệ thống quản lý cán bộ, công chức, viên chức đã được triển khai tại Viện Năng lượng nguyên tử 17
1.3.1. Hệ thống quản lý cán bộ, công chức, viên chức của Bộ KH&CN xây dựng năm 2020 17
1.3.2. Hệ thống quản lý cán bộ, công chức, viên chức của Bộ KH&CN năm 2024 19
1.3.3. Hệ thống quản lý cán bộ, công chức, viên chức của Bộ KH&CN (sau sắp xếp) năm 2025 21
CHƯƠNG 2: CÁC KỸ THUẬT TÍCH HỢP DỮ LIỆU 25
2.1. Lý thuyết về Tích hợp dữ liệu - Data Intergration 25
2.1.1. Tích hợp dữ liệu trên thế giới 25

2.1.2. Minh họa sự cần thiết của việc Tích hợp dữ liệu tại Viện Năng lượng nguyên tử 27
2.1.3. Tóm lược về tích hợp dữ liệu 29
2.2. Các kỹ thuật tích hợp dữ liệu 30
2.2.1. Kỹ thuật Hợp nhất dữ liệu ETL/ELT (Extract, Transform, Load/Etract, Load, Tranform – Trích xuất, Chuyển đổi, Tải/Trích xuất, Tải, Chuyển đổi) 30
2.2.2. Kỹ thuật Data Virtualization 34
2.3. Khảo sát công nghệ để giải quyết bài toán Tích hợp dữ liệu 37
2.3.1. Kỹ thuật Tích hợp lựa chọn 37
2.3.2. Công cụ Tích hợp lựa chọn 38
CHƯƠNG 3: XÂY DỰNG CÔNG CỤ PHẦN MỀM VÀ THỬ NGHIỆM ÁP DỤNG, ĐÁNH GIÁ 39
3.1. Xây dựng và cấu hình công cụ phần mềm 40
3.2. Kết quả thử nghiệm 48
3.3. Đánh giá kết quả đã đạt được của luận văn 51
3.3.1. Về ưu điểm và hạn chế 51
3.3.2. Đánh giá về mặt định tính 51
3.3.3. Đánh giá về mặt định lượng 52
3.4. Định hướng phát triển 53
KẾT LUẬN 55
DANH MỤC TÀI LIỆU THAM KHẢO 57

DANH MỤC VIẾT TẮT

Từ viết tắt	Từ đầy đủ
NLNT	Năng lượng nguyên tử
NLNTVN	Năng lượng nguyên tử Việt Nam
KH&CN	Khoa học và Công nghệ
ETL	Extract, Transform, Load – Trích xuất, Chuyển đổi, Tải
ELT	Extract, Load, Transform – Trích xuất, Tải, Chuyển đổi
QLCBCCVC	Quản lý cán bộ, công chức, viên chức
TCCB	Tổ chức cán bộ
CSDL	Cơ sở dữ liệu
CSDLQG	Cơ sở dữ liệu quốc gia
CCCD	Căn cước công dân

DANH MỤC BẢNG, BIỂU, HÌNH MINH HOẠ

Bảng 1. So sánh các cách giải quyết vấn đề tích hợp dữ liệu 9

Hình 1.1 Tổng doanh thu từ các loại hình dịch vụ của Viện NLNTVN qua các năm 11
Hình 1.2 Sơ đồ cơ cấu tổ chức Viện Năng lượng nguyên tử Việt Nam 12
Hình 1.3 Giao diện hệ thống QLCBCCVC của Bộ KH&CN 2020..... 18
Hình 1.4 Giao diện Hệ thống quản lý cán bộ của Bộ KH & CN (2024) 20
Hình 1.5 Giao diện Hệ thống quản lý cán bộ của Bộ KH & CN (2025) 22
Hình 1.6 Lỗi lặp hồ sơ trong quá trình đồng bộ sau sáp nhập Bộ 23

Hình 2.1 Sự quan tâm đến "Tích hợp dữ liệu" qua công cụ tìm kiếm Google 26
Hình 2.2 Sơ đồ minh họa quy trình tích hợp dữ liệu 27
Hình 2.3 Sơ đồ mô tả quy trình ETL 31
Hình 2.4 So sánh giữa kỹ thuật ETL và ELT..... 33
Hình 2.5 Minh họa về Ảo hóa dữ liệu giữa các nguồn dữ liệu và người dùng 35

Hình 3.1 Cấu hình hệ thống 40
Hình 3.2 Trích xuất dữ liệu từ các file lưu trữ dữ liệu..... 41
Hình 3.3 Dữ liệu về cán bộ hiện tại được trích xuất ra từ hệ thống mới 202542
Hình 3.4 Định nghĩa chuẩn hoá CCCD..... 43
Hình 3.5 Định nghĩa tích hợp theo các dữ liệu số CCCD..... 43
Hình 3.6 Quy trình Chuyển đổi dữ liệu, làm sạch dữ liệu và chuẩn hoá về định dạng. 44
Hình 3.7 Quy trình xử lý loại bỏ email cán bộ bị trùng..... 45
Hình 3.8 Chức năng tải dữ liệu của hệ thống..... 46
Hình 3.9 Chức năng chuyển định dạng dữ liệu sang .csv..... 47
Hình 3.10 File dữ liệu cán bộ thứ 1 49
Hình 3.11 File dữ liệu cán bộ thứ 2 49
Hình 3.12 File dữ liệu cán bộ thứ 3 50
Hình 3.13 Dữ liệu cán bộ được tích hợp từ 3 file dữ liệu có sẵn..... 50

MỞ ĐẦU

1. Lý do chọn đề tài

Hiện nay, trước bối cảnh cuộc Cách mạng công nghiệp lần thứ Tư và xu thế chuyển đổi số đang diễn ra mạnh mẽ, rộng khắp trong mọi mặt của nền kinh tế - xã hội tại Việt Nam nói riêng và trên cả thế giới nói chung. Để tiến hành triển khai nhiệm vụ quan trọng này, Ngày 25/6/2020 Thủ tướng chính phủ đã ra Quyết định 893/QĐ-TTg về việc Phê duyệt Đề án xây dựng cơ sở dữ liệu quốc gia về cán bộ, công chức, viên chức trong các cơ quan nhà nước [1]. Tại Quyết định này nêu rõ quan điểm, mục tiêu, và yêu cầu đối với hệ thống quản lý cơ sở dữ liệu quốc gia do Bộ Nội Vụ khai thác, quản lý và sử dụng.

Viện Năng lượng nguyên tử Việt Nam (tên tiếng Anh: Vietnam Atomic Energy Institute, tên viết tắt là VINATOM là tổ chức sự nghiệp khoa học hạng đặc biệt trực thuộc Bộ Khoa học và Công nghệ (KH&CN). Cơ cấu tổ chức của Viện hiện nay gồm 03 đơn vị chức năng, 03 viện nghiên cứu, 05 trung tâm nghiên cứu tổng cộng hơn 800 cán bộ đang làm việc. Đặc thù công việc yêu cầu Viện phải thường xuyên triển khai rà soát, tổng hợp, thống kê dữ liệu về công chức, viên chức tại đơn vị để có thể tạo được các báo cáo, văn bản liên quan đến yếu tố con người tại Viện. Do vậy có rất nhiều dữ liệu liên quan đến cán bộ, công chức, viên chức, người lao động (hồ sơ cán bộ, tuyển dụng, lương, chế độ chính sách, tuyển dụng, bổ nhiệm, đào tạo, vv...) cần phải rà soát và tổng hợp. Tuy nhiên, khi tiến hành đưa ra các văn bản, báo cáo này, Viện thường xuyên phải cập nhật lại các dữ liệu từ các nguồn khác nhau, sau đó tiến hành tinh chỉnh lại dữ liệu, đưa về cùng một định dạng thống nhất, sau đó xử lý thông tin để có thể sử dụng vào từng văn bản cụ thể. Việc này không chỉ rất mất thời gian và công sức, mà việc sai sót do yếu tố con người và dữ liệu không nhất quán là hoàn toàn không thể tránh khỏi.

Theo Kế hoạch 141/KH-BCĐTKNQ18 ngày 06/12/2024 của Ban Chỉ đạo về tổng kết việc thực hiện Nghị quyết số 18-NQ/TW về Kế hoạch định hướng sắp xếp, tinh gọn tổ chức bộ máy của Chính phủ có nêu về định hướng cơ cấu, sắp xếp và hợp

nhất 14 bộ, cơ quan ngang Bộ. Cụ thể theo Kế hoạch, "Hợp nhất Bộ Thông tin và Truyền thông và Bộ Khoa học và Công nghệ, thực hiện chức năng quản lý nhà nước về các ngành, lĩnh vực hiện đang giao cho Bộ Thông tin và Truyền thông và Bộ Khoa học và Công nghệ. Sau hợp nhất Bộ mới có tên là Bộ Khoa học và Công nghệ". Chính vì thế nên số lượng cơ sở dữ liệu của cả hai Bộ ngành sau sát nhập bị nhân lên, chồng chéo và khiến việc truy xuất và quản lý dữ liệu phức tạp hơn.

Trên thực tế thì Cơ sở dữ liệu về viên chức và người lao động tại Viện Năng lượng nguyên tử hiện đang rất chồng chéo và rời rạc nguyên do từ việc sáp nhập Bộ, ngành nên vô hình chung phát sinh nhiều cơ sở dữ liệu về cán bộ, dẫn đến dữ liệu không chính xác và phức tạp trong quá trình cập nhật và khai thác. Chính vì lý do thực tế cấp thiết này, học viên đề xuất đưa ra giải pháp Tích hợp dữ liệu – Data Integration để kết hợp các dữ liệu không đồng nhất từ nhiều nguồn khác nhau thành một tập hợp dữ liệu thống nhất. Điều này sẽ cho phép công tác khai thác dữ liệu từ nhiều nguồn hiện hữu một cách dễ dàng và hiệu quả, chính xác.

Xuất phát từ những vấn đề nêu trên học viên đã chọn đề tài **"Tích hợp dữ liệu cho hệ thống quản lý cán bộ, viên chức của Viện Năng lượng nguyên tử"**

Bố cục của luận văn này kết cấu gồm 3 chương, đi từ việc trình bày lý thuyết, đặc thù công việc đến thực tiễn cách tiếp cận vấn đề và triển khai công cụ phần mềm, cụ thể như sau:

- Phần mở đầu
- Phần nội dung luận văn gồm:

Chương 1. Tổng quan về vấn đề nghiên cứu

- Giới thiệu về Viện Năng lượng nguyên tử Việt Nam
- Giới thiệu về Hệ thống quản lý công chức, viên chức trong cơ quan nhà nước Việt Nam
- Giới thiệu về các hệ thống quản lý cán bộ đã được triển khai tại đơn vị

Chương 2. Các kỹ thuật tích hợp dữ liệu

- Cách tiếp cận vấn đề xử lý dữ liệu

- Lý thuyết về Tích hợp dữ liệu
- Các kỹ thuật tích hợp dữ liệu
- Đề xuất công nghệ áp dụng

Chương 3. Triển khai áp dụng, thử nghiệm, đánh giá

2. Mục đích nghiên cứu:

Mục đích chính của luận văn là nghiên cứu về cách thức quản lý cơ sở dữ liệu cán bộ, viên chức và người lao động tại Viện Năng lượng nguyên tử Việt Nam trong bối cảnh đất nước đổi mới, đẩy mạnh ứng dụng phát triển ứng dụng công nghệ thông tin và chuyển đổi số. Học viên muốn thông qua luân văn này trình bày cụ thể tình huống thực tế về bài toán quản lý và khai thác cơ sở dữ liệu về cán bộ và người lao động tại đơn vị trước và sau sáp nhập cơ quan chủ quản là Bộ Khoa học và Công nghệ.

Mục tiêu tiếp theo của Luận văn sẽ tìm hiểu lý thuyết và các công trình nghiên cứu về Data Integration và Data Warehouse nhằm giải quyết bài toán được đặt ra cho hệ thống dữ liệu về quản lý cán bộ, viên chức và người lao động tại Viện NLNTVN. Từ đó tiến tới giới thiệu về các phương pháp tích hợp dữ liệu phổ biến hiện hành.

Cuối cùng, mục tiêu chủ đạo của nghiên cứu là đưa ra một mô hình tích hợp dữ liệu phù hợp để áp dụng vào hệ thống quản lý dữ liệu về viên chức và người lao động hiện hành nhằm thoả mãn và giải quyết các khó khăn và vướng mắc hiện có. Ngoài ra nghiên cứu cũng đưa ra các ví dụ trực quan như hình ảnh, biểu đồ và mô phỏng các thao tác, hướng dẫn, dễ đọc và dễ hiểu cho cán bộ thực hiện công tác chuyên môn.

Nghiên cứu này hướng tới có thể góp phần hỗ trợ xây dựng hệ thống quản lý cán bộ, công chức, viên chức, người lao động hiện hành và được triển khai thử nghiệm thực tế tại Viện Năng lượng nguyên tử Việt Nam, tiến tới trong Bộ KH&CN.

3. Đối tượng và phạm vi nghiên cứu:

- Luận văn tập trung nghiên cứu về lý thuyết Tích hợp dữ liệu, các cách thức triển khai tích hợp dữ liệu và nhu cầu thực tế về việc áp dụng kỹ thuật Data

Intergration trong hệ thống dữ liệu về quản lý cán bộ, viên chức và người lao động tại Viện Năng lượng nguyên tử Việt Nam.

- Giới hạn phạm vi nghiên cứu của đề tài: Để đảm bảo tính khả thi và thực tế của luận văn, tác giả tập trung vào các phương pháp và cách triển khai tích hợp dữ liệu ngắn gọn, phù hợp với kinh phí và thân thiện với người dùng nhất đồng thời bám sát vào các nhu cầu thực tế tại đơn vị áp dụng là Viện Năng lượng nguyên tử.

4. Phương pháp nghiên cứu, cách tiếp cận:

- *Nghiên cứu lý thuyết:* Luận văn lấy lý thuyết Tích hợp dữ liệu làm nền tảng và được định hình xuyên suốt trong toàn bộ nghiên cứu. Tiến tới tiến hành việc nghiên cứu hệ thống văn bản quy phạm pháp luật và các văn bản liên quan đến dữ liệu quản lý cán bộ, công chức, viên chức, người lao động trong cơ quan. Nghiên cứu tài liệu, bài báo về Data Integration và Data Warehouse. Các phương pháp Tích hợp dữ liệu, các mô hình triển khai Tích hợp dữ liệu.

- *Nghiên cứu thực tiễn:* Từ nền tảng lý thuyết được nghiên cứu, Luận văn tiếp tục nghiên cứu các cách thức đưa phương pháp Tích hợp dữ liệu và thử nghiệm trên một số cơ sở dữ liệu được xây dựng nhằm mô phỏng lại hệ thống quản lý cán bộ, công chức, viên chức và người lao động tại Viện Năng lượng nguyên tử Việt Nam nhằm giải quyết các khó khăn, rào cản hiện hữu.

- *Cách tiếp cận nghiên cứu:* Dựa trên cơ sở các kiến thức đã được trang bị trong quá trình học tập, nghiên cứu, đề xuất xây dựng, thiết kế tích hợp dữ liệu quản lý cán bộ, công chức, viên chức, người lao động của Viện Năng lượng nguyên tử Việt Nam phù hợp, thích ứng với yêu cầu thực tiễn và quy định của pháp luật có liên quan. Từ đó tiến tới áp dụng vào thực tế với cơ sở dữ liệu về công chức, viên chức và người lao động tại Viện Năng lượng nguyên tử Việt Nam. Song song với đó là học hỏi, khảo sát và phản hồi lại ý kiến của các cán bộ, đồng nghiệp có liên quan đến công tác khai thác và sử dụng hệ thống quản lý cán bộ nhằm cải thiện và hiệu chỉnh mô hình Tích hợp dữ liệu sao cho thân thiện, hiệu quả, kinh tế và có tính ứng dụng cao.

Kết quả dự kiến thu được:

- Tóm tắt các nghiên cứu và lý thuyết liên quan đến Tích hợp dữ liệu và các mô hình triển khai Tích hợp dữ liệu hiện hành.

- Xây dựng phương án tích hợp dữ liệu quản lý cán bộ, công chức, viên chức, người lao động tại Viện Năng lượng nguyên tử Việt Nam, có khả năng tổng hợp dữ liệu từ nhiều dạng, nhiều nguồn, lưu trữ dữ liệu tổ chức cán bộ một cách chính xác, hiệu quả, bảo mật quyền riêng tư, hiển thị dữ liệu đúng với yêu cầu truy xuất, dễ dàng sử dụng thông tin. Đảm bảo đáp ứng yêu cầu dữ liệu "đúng, đủ, sạch, sống, thống nhất, dùng chung".

- Khả năng và địa chỉ ứng dụng: Có khả năng ứng dụng trong Viện Năng lượng nguyên tử Việt Nam, giúp giảm thiểu rất nhiều các công tác nhập, xuất dữ liệu đối với từng cá nhân cán bộ, viên chức và người lao động tại đơn vị. Đặt nền móng cho bài toán tích hợp dữ liệu để ứng dụng vào hệ thống cơ sở dữ liệu của Bộ Khoa học và Công nghệ về công chức, viên chức và người lao động

- Dự kiến hiệu quả mang lại: Giúp nâng cao hiệu quả của công tác tổ chức cán bộ, có khả năng kết nối với trục liên thông kết nối, chia sẻ dữ liệu cấp Bộ (của Bộ KH&CN) đảm bảo yêu cầu về bảo mật và an toàn thông tin, giúp tiết kiệm thời gian tra cứu và giảm bớt yêu cầu về lưu trữ.

CHƯƠNG 1: TỔNG QUAN VỀ VẤN ĐỀ NGHIÊN CỨU

Tích hợp dữ liệu hay còn được biết đến là Data Integration là một khái niệm quan trọng trong việc giải quyết bài toán quản lý và khai thác dữ liệu trong thời đại 4.0. Nó đóng vai trò trung tâm và nền móng không thể thiếu trong việc giải quyết các thách thức trong việc quản lý, khai thác và phân tích Dữ liệu lớn (Big Data). Quá trình này rất quan trọng vì Big Data thường đến từ các nguồn dữ liệu đa dạng, có khối lượng lớn và đòi hỏi tốc độ cao. Chính vì vậy thách thức đặt ra trong việc xử lý dữ liệu Big Data đòi hỏi các cơ sở dữ liệu phải được gom lại để xử lý hiệu quả và Tích hợp dữ liệu là kỹ thuật hoàn hảo để giải quyết vấn đề đó. Hiện tại có ba hướng giải pháp chính cho bài toán quản lý và khai thác dữ liệu hiệu quả hiện nay thường được sử dụng là Tích hợp dữ liệu (Data Integration), Hợp nhất dữ liệu (Data Consolidation) và Tổng hợp dữ liệu (Data Aggregation) tuỳ vào nhu cầu sử dụng và trường hợp sử dụng mà người dùng có thể chọn phương án thích hợp. Tuy nhiên, để có thể thực hiện việc Data Consolidation hay Data Aggregation thì điều kiện tiên quyết là dữ liệu cần phải được Tích hợp, nên có thể coi việc Tích hợp dữ liệu là nền móng căn bản và quan trọng nhất trong bài toán quản lý dữ liệu. Từ đó học viên tiến hành nghiên cứu về bài toán Tích hợp dữ liệu. "Tích hợp dữ liệu là vấn đề kết hợp dữ liệu tại nhiều nguồn và cho phép người dùng có một cách nhìn hợp nhất với những dữ liệu đó" [3].

Nghiên cứu và tìm hiểu lý thuyết về Tích hợp dữ liệu – Data Intergration, các kỹ thuật tích hợp dữ liệu, các công cụ liên quan để thực hiện việc tích hợp dữ liệu để qua đó nghiên cứu đề xuất phương án giải quyết bài toán "Tích hợp dữ liệu cho hệ thống quản lý cán bộ, viên chức của Viện Năng lượng nguyên tử".

Nghiên cứu xây dựng và đề xuất phương án Tích hợp dữ liệu nhằm giải quyết bài toán về khai thác và sử dụng cơ sở dữ liệu cán bộ, viên chức, người lao động tại Viện Năng lượng nguyên tử Việt Nam sau sáp nhập.

1.1. Cách tiếp cận vấn đề quản lý và khai thác dữ liệu

Trong bối cảnh các cơ sở dữ liệu ngày một lớn và phân hóa một cách dàn trải trên thế giới nói chung và tại Việt Nam nói riêng ở trên thì việc quản lý và khai thác

dữ liệu là yêu cầu vô cùng cấp thiết. Cùng với nhu cầu phát triển của con người, việc công nghệ phát triển để phục vụ nhân loại là tất yếu. Tuy nhiên, các cơ sở dữ liệu đi kèm theo nó liên tục được tạo mới, liên tục thay đổi và được hình thành từ nhiều nguồn luôn luôn là một mớ hỗn độn khiến các nhà khoa học, các nhà nghiên cứu luôn đau đầu trong việc quản lý và khai thác chúng một cách hiệu quả. Để có thể khai thác, sử dụng và quản lý các cơ sở dữ liệu trên một cách khoa học và dễ hiểu nhất, các nhà khoa học đã nghiên cứu không ngừng và đề xuất nhiều giải pháp nhằm giải bài toán này. Trong đó ba hướng kỹ thuật, giải pháp chính cho bài toán quản lý và khai thác dữ liệu hiệu quả hiện nay thường được sử dụng là Tích hợp dữ liệu (Data Integration), Hợp nhất dữ liệu (Data Consolidation) và Tổng hợp dữ liệu (Data Aggregation) với mục tiêu cuối cùng là đưa ra một nguồn sự thật duy nhất (Single Source of Truth). Tuy nhiên trong ngữ cảnh rộng hơn thì thuật ngữ "Tích hợp dữ liệu" thường được dùng để chỉ toàn bộ các kỹ thuật trên và trong kỹ thuật quản trị dữ liệu nó ám chỉ quy trình bao trùm nhiều bước nhỏ bao gồm cả các bài toán về hợp nhất và tổng hợp dữ liệu.

1.1.1. Tích hợp dữ liệu (Data Integration)

Tích hợp dữ liệu hay Data Integration thường được sử dụng để đề cập đến quá trình kết hợp dữ liệu từ nhiều nguồn, nhiều cơ sở dữ liệu, databasse khác nhau và rời rạc trở thành một thể thống nhất (thường là một cơ sở dữ liệu) nhằm đảm bảo tính toàn vẹn của dữ liệu đồng thời cung cấp khả năng truy xuất dữ liệu đồng bộ [3]. Cũng theo tác giả Lenzerini thì đây là bước nền tảng mang tính quyết định cho phép người dùng có thể khai thác thông tin từ nhiều nguồn khác nhau, đây là một lĩnh vực nghiên cứu có tính chất quan trọng trong việc quản trị cơ sở dữ liệu. Ứng dụng điển hình của quá trình này là giúp giảm thiểu phân mảnh dữ liệu, hỗ trợ truy vấn và khai thác dữ liệu cụ thể trong các nền tảng Dữ liệu lớn (Big Data) như Kho dữ liệu (Data warehouse), Hồ dữ liệu (Data lake) hoặc các hệ thống Internet vạn vật (Internet of Things)...

1.1.2. Tổng hợp dữ liệu (Data Aggregation)

Tổng hợp dữ liệu hay Data Aggregation là thuật ngữ nhằm chỉ quá trình thu thập dữ liệu từ nhiều nguồn khác nhau hoặc nhiều hệ thống (thường là các cửa hàng, các công ty con, các đơn vị trực thuộc, các điểm có cùng một tiêu chuẩn và dạng dữ liệu) sau đó tiến hành gộp lại thành một dữ liệu mang tính tóm tắt tổng quan, thống nhất (thường là một báo cáo tổng hợp trung bình số liệu, tổng số liệu, giá trị thấp nhất và giá trị cao nhất trong các cơ sở dữ liệu) để giúp người dùng phân tích sâu hơn và hỗ trợ việc quản trị và sử dụng dữ liệu. Quá trình này giúp đơn hóa công tác thống kê, cung cấp một cái nhìn tổng thể cho người quản lý dữ liệu. Cách tiếp cận việc xử lý dữ liệu này giúp giảm thiểu đáng kể khối lượng dữ liệu cũng như thời gian phân tích dữ liệu. Theo các tác giả Karthick & Kalrani nhận định trong nghiên cứu của họ năm 2014, "Data Aggregation không tập trung vào tính toàn vẹn tuyệt đối của dữ liệu nguồn mà nó tập trung vào việc tổng hợp, phân tích và tóm tắt dữ liệu một cách tổng quan nhất". Ứng dụng cụ thể của quá trình này là giúp việc quản trị doanh nghiệp có thể dễ dàng so sánh doanh số giữa các mặt hàng, so sánh doanh số và tiêu thụ tại mỗi điểm bán hàng, tổng quan các báo cáo thống kê, kết quả kinh doanh nhằm đưa ra phương án quản trị, tìm ra xu hướng thị trường, tổng hợp dữ liệu sử dụng các thiết bị IoT…

1.1.3. Hợp nhất dữ liệu (Data Consolidation)

Hợp nhất dữ liệu hay Data Consolidation là thuật ngữ thường được dùng để nhắc đến quá trình thu thập và kết hợp dữ liệu từ nhiều nguồn dữ liệu khác nhau thành một nguồn dữ liệu duy nhất. Khác với Tích hợp dữ liệu, quá trình này thường được sử dụng với các nguồn dữ liệu sau khi đã được tích hợp với mục đích chính là loại bỏ các dữ liệu bị trùng lặp, có mâu thuẫn với nhau nhằm đảm bảo dữ liệu cuối. Tác giả Batini et al [12] đề cập và ca ngợi đến quá trình này là một "single source of truth" (nguồn sự thật duy nhất) trong quản trị dữ liệu. Mục đích chính của Hợp nhất dữ liệu là nó giúp đơn giản hóa việc quản lý dữ liệu, giảm thiểu chi phí vận hành và tăng cường tính chính xác trong việc phân tích dữ liệu cho người dùng trong Data Warehouse hoặc trong hệ thống các máy chủ…

1.1.4. So sánh giữa Tích hợp dữ liệu (Data Integration), Tổng hợp dữ liệu (Data Aggregation) và Hợp nhất dữ liệu (Data Consolidation)

Bảng 1. So sánh các cách giải quyết vấn đề tích hợp dữ liệu

	Tích hợp dữ liệu	Tổng hợp dữ liệu	Hợp nhất dữ liệu
Mục đích	Kết hợp dữ liệu từ nhiều nguồn, nhiều cơ sở dữ liệu khác nhau thành một cơ ở dữ liệu thống nhất	Tổng hợp số liệu, tóm tắt dữ liệu và tính toán dữ liệu cơ bản để phục vụ phân tích	Chuẩn hóa, hợp nhất dữ liệu vào một hệ thống trung tâm, loại bỏ trùng lặp và dữ liệu xấu
Đặc thù	Đặc điểm chính là khả năng kết nối và tương thích giữa các hệ thống cơ sở dữ liệu khác nhau	Đặc điểm chính là khả năng tính toán, rút gọn và phân nhóm các dữ liệu	Đặc điểm chính là độ chính xác, tính toàn vẹn và đưa ra một nguồn duy nhất
Kết quả	Tạo ra một hệ thống cơ sở dữ liệu đa nguồn có khả năng truy xuất và khai thác đồng bộ	Tạo ra báo cáo cơ sở dữ liệu dễ đọc và hình dung, phục vụ mục đích phân tích	Tạo ra cơ sở dữ liệu duy nhất, được chuẩn hóa và có tính ổn định lâu dài
Ứng dụng	ETL, IoT, Big Data	Báo cáo tài chính, phân tích kết quả kinh doanh	Data warehouse
Vai trò	Bước nền móng quan trọng trong khối liên thông dữ liệu	Tạo ra thông tin dễ hiểu và tập trung, giúp người dùng nhận định thông tin	Đảm bảo cơ sở dữ liệu đáng tin cậy, dài hạn

Từ đó có thể thể thấy mối quan hệ giữa ba quá trình này mang tính tuần tự và bổ trợ lẫn nhau trong một quy trình tổng thể về quản trị liên thông dữ liệu: Tích hợp dữ liệu => Tổng hợp dữ liệu => Hợp nhất dữ liệu. Tuy nhiên tùy thuộc vào mục đích

sử dụng và yêu cầu về dữ liệu đầu ra và nguồn kinh phí mà người dùng có thể chọn phương án quản lý dữ liệu hợp lý nhất.

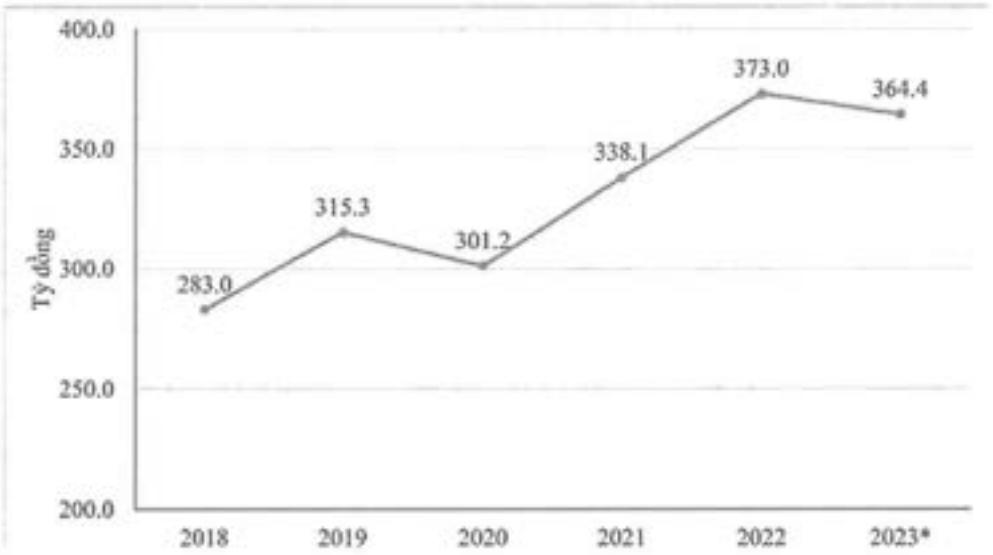
1.2. Giới thiệu về Viện Năng lượng nguyên tử Việt Nam

Viện Năng lượng nguyên tử Việt Nam tiền thân là Viện Nghiên cứu hạt nhân Đà Lạt chính thức tiếp nhận lò phản ứng hạt nhân Đà Lạt vào cuối năm 1975 sau khi thống nhất đất nước từ Việt Nam Cộng hoà. Vào năm 1979, Viện được chính thức thành lập theo quyết định số 59-CP ngày 23/2/1979 của Hội đồng Chính phủ, đặt dưới sự chỉ đạo trực tiếp của Thủ tướng Chính phủ.

Viện Năng lượng nguyên tử Việt Nam là tổ chức sự nghiệp khoa học và công nghệ hạng đặc biệt trực thuộc Bộ Khoa học và Công nghệ có chức năng giúp Bộ trưởng nghiên cứu cơ bản, nghiên cứu ứng dụng và triển khai các hoạt động ứng dụng kết quả nghiên cứu trong lĩnh vực năng lượng nguyên tử; hỗ trợ kỹ thuật phục vụ công tác quản lý nhà nước về năng lượng nguyên tử, an toàn bức xạ và hạt nhân; tổ chức hoạt động đào tạo, dịch vụ trong lĩnh vực năng lượng nguyên tử theo quy định của pháp luật.

Viện Năng lượng nguyên tử Việt Nam có tên giao dịch quốc tế bằng tiếng Anh là: ***Vietnam Atomic Energy Institute*** (viết tắt là **VINATOM**). Viện có tư cách pháp nhân, là đơn vị dự toán theo quy định của Luật Ngân sách, có con dấu riêng, được mở tài khoản tại Kho bạc nhà nước và ngân hàng để hoạt động theo quy định của pháp luật. Viện có trụ sở chính đặt tại 59 Lý Thường Kiệt, Quận Hoàn Kiếm, Thành phố Hà Nội. Điện thoại: 024.39427580; Fax: 024.39424133;

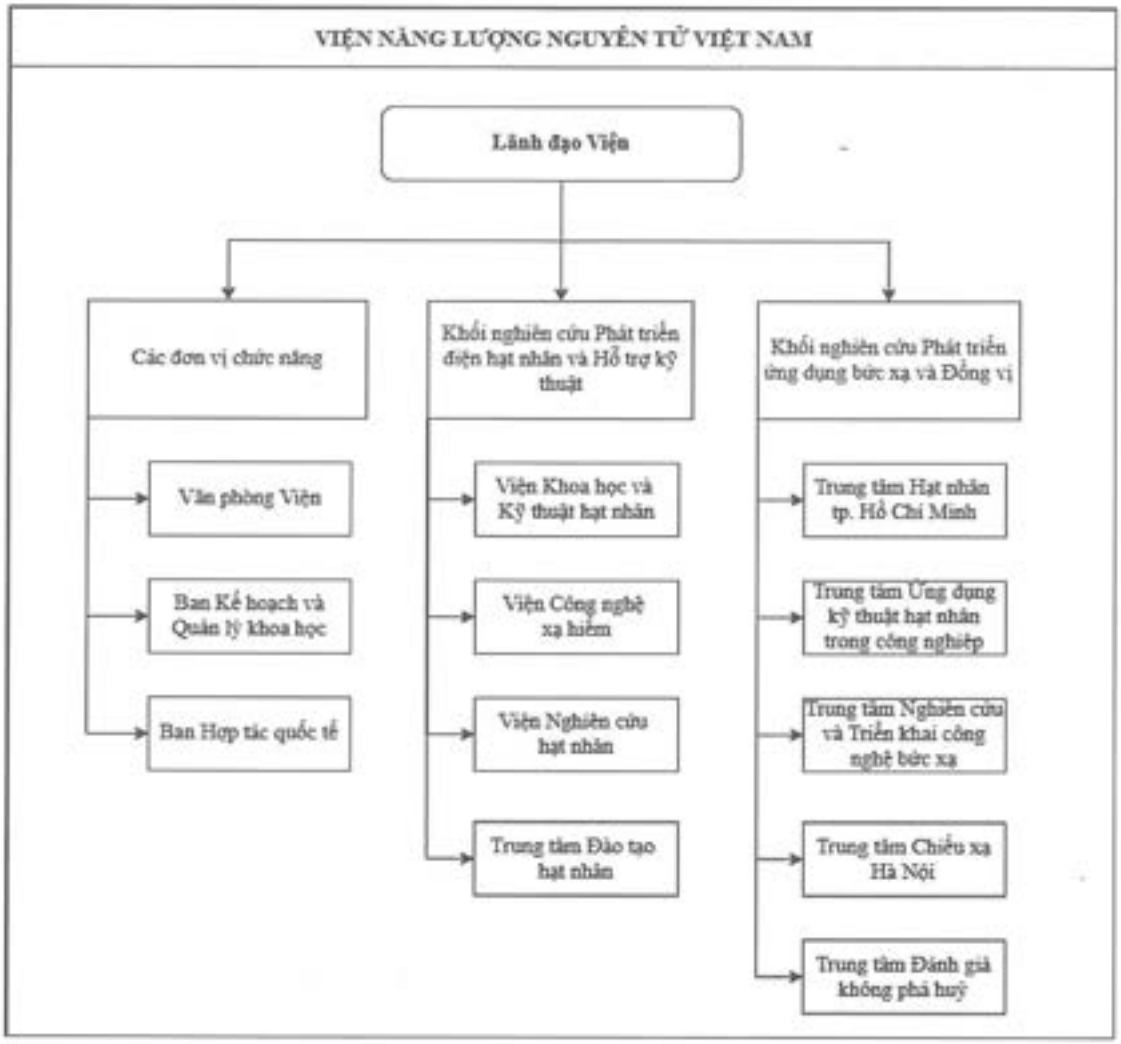
Tính đến năm 2023, mặc dù nguồn lực đầu tư của nhà nước còn khiêm tốn, tình hình kinh tế, chính trị thế giới nhiều biến động phức tạp, trong nước còn nhiều khó khăn, các hoạt động triển khai dịch vụ của các đơn vị trực thuộc Viện vẫn được duy trì tương đối ổn định. Tổng doanh thu năm 2023 (tính đến tháng 12) của toàn Viện đạt 364,367 tỷ đồng.



Hình 1.1 Tổng doanh thu từ các loại hình dịch vụ của Viện NLNTVN qua các năm

Viện có quy mô hoạt động đặc thù, bao gồm cả nghiên cứu lý thuyết và nghiên cứu ứng dụng, đào tạo về ngành năng lượng hạt nhân, dịch vụ chiếu xạ, dịch vụ y tế liên quan đến xạ trị, hoá trị, nghiên cứu và khai thác đất hiếm vv… Viện có nhiều Trung tâm và Viện nghiên cứu trực thuộc trải dài trên toàn quốc.

Cơ cấu tổ chức của Viện Năng lượng nguyên tử Việt Nam được thể hiện ở Hình 1.2, mô tả quan hệ giữa Lãnh đạo Viện trực tiếp chỉ đạo các đơn vị chức năng, các Viện nghiên nghiên cứu và các Trung tâm nghiên cứu trực thuộc



Hình 1.2 Sơ đồ cơ cấu tổ chức Viện Năng lượng nguyên tử Việt Nam

1.2.1. Công tác quản lý cán bộ tại Viện

Một trong những nhiệm vụ chính của Văn phòng Viện là giúp Viện trưởng quản lý công tác tổ chức, cán bộ, công chức, viên chức và người lao động trong của Viện. Ngoài ra, Văn phòng cũng đảm nhiệm việc quản lý hồ sơ, cơ sở dữ liệu về công chức, viên chức, tài liệu của đơn vị theo phân cấp của Viện Năng lượng nguyên tử Việt Nam và các quy định của pháp luật.

Do đặc thù công tác và nhiệm vụ được giao của Viện Năng lượng nguyên tử trong sự nghiệp phát triển đất nước, Viện hoạt động ở nhiều mảng, ngành liên quan đến nghiên cứu và ứng dụng kết quả nghiên cứu trong lĩnh vực năng lượng nguyên tử; chiếu xạ; đất hiếm và tổ chức các hoạt động đào tạo, dịch vụ trong lĩnh vực năng

lượng nguyên tử theo quy định của pháp luật. Để có thể đáp ứng được nhiệm vụ được giao, Viện đã thành lập nhiều Trung tâm, Công ty và Viện nghiên cứu trực thuộc Viện trong lịch sử phát triển của mình. Hiện tại Viện có 12 đơn vị trực thuộc, gồm 3 đơn vị chức năng và 9 đơn vị nghiên cứu, triển khai trên toàn quốc. Vị trí địa lý của từng đơn vị trực thuộc cũng tuỳ thuộc vào nhu cầu công tác, vị trí địa lý, mật độ dân cư và yếu tố an ninh năng lượng, an ninh quốc gia mà được thành lập nên dẫn đến việc các đơn vị rải rác ở các miền trên tổ quốc (miền bắc có 8 đơn vị tại Hà Nội, miền trung có 2 đơn vị tại Đà Lạt và miền nam có 2 đơn vị tại thành phố Hồ Chí Minh). Chính vì Viện được giao và đảm nhận nhiều công tác chuyên môn đặc thù và có nhiều cơ sở trực thuộc nên Viện hiện tại đang sở hữu một nguồn nhân lực tương đối lớn và dàn trải khắp mọi miền trên tổ quốc. Chính vì thế nên yêu cầu có một hệ thống quản lý nhân sự chung là cấp thiết do việc quản lý hồ sơ cán bộ thủ công qua hồ sơ giấy là hoàn toàn không khả thi và sẽ khó có cái nhìn toàn cảnh về chất lượng cán bộ, viên chức và người lao động trên toàn Viện.

Công tác tổ chức cán bộ là một phần quan trọng trong công tác quản lý nhân sự của mọi tổ chức, cơ quan, viện nghiên cứu, doanh nghiệp hay các tổ chức chính trị, xã hội và tại Viện NLNTVN cũng không phải là ngoại lệ. Điều này không chỉ ảnh hưởng đến việc tuyển dụng, chế độ chính sách, đào tạo và phát triển đội ngũ nhân lực mà còn bao gồm việc tham mưu và triển khai các chính sách nhân sự, giúp đỡ Lãnh đạo Viện về các công tác quy hoạch, bổ nhiệm, điều chuyển một cách phù hợp, đảm bảo sự phát triển bền vững và hiệu quả của tổ chức. Bên cạnh đó, công tác tổ chức cán bộ cũng bao gồm cả việc thi đua, khen thưởng, kỷ luật, lịch sử công tác, thống kê các thông tin và các đặc thù của con người tại Viện. Trong các Viện nghiên cứu như Viện Năng lượng nguyên tử, công tác tổ chức cán bộ đóng vai trò thiết yếu trong việc bảo đảm sự vận hành ổn định, hiệu quả, đồng thời tạo ra các điều kiện để đội ngũ cán bộ, viên chức có thể phát huy tối đa năng lực, đóng góp cho sự phát triển chung của tổ chức.

Chính vì tầm quan trọng của công tác này trong mọi cơ quan, tổ chức cùng với lượng dữ liệu liên quan đến nhân sự và các công việc triển khai ngày một lớn và phức tạp. Điều này khiến cho việc có một hệ thống quản lý cơ sở dữ liệu về cán bộ trong

mọi cơ quan, tổ chức là cần thiết để khiến cho người quản lý có được một cái nhìn toàn diện về người làm việc. Yêu cầu cấp thiết với công tác cán bộ tại Viện là có một hệ thống dữ liệu tổng hợp về thông tin chuyên ngành; độ tuổi; sở trường công tác; địa chỉ; chức vụ của toàn bộ cán bộ thuộc Viện trên khắp toàn quốc tập trung tại cùng trên một hệ thống là vô cùng cấp thiết. Hệ thống này sẽ giúp người quản trị con người có cái nhìn tổng quan về nguồn lực nhân sự tại đơn vị, giúp phân bổ, điều chuyển công tác phù hợp với chuyên môn nhân sự và đặc thù công việc yêu cầu, góp phần thực hiện các nhiệm vụ quan trọng của Viện trong thời đại mới của đất nước.

1.2.2. Hệ thống quản lý công chức, viên chức trong cơ quan nhà nước Việt Nam

Một hệ thống quản lý cán bộ công chức, viên chức phải là giải pháp tổng thể được thiết kế và xây dựng để áp dụng cho từng đặc thù riêng biệt của mỗi cơ quan, tổ chức sử dụng nó. Qua đó nó giúp cho các cơ quan, tổ chức này có thể dễ dàng theo dõi thông tin, dữ liệu của mỗi cán bộ, công chức, viên chức trong quá trình công tác từ khi bắt đầu vào cơ quan cho đến nghỉ hưu. Đồng thời hệ thống cũng cần có khả năng lưu trữ những dữ liệu này theo tiêu chuẩn, có khả năng chỉnh sửa dữ liệu cán bộ trong quá trình công tác tại đơn vị. Hệ thống giúp các đơn vị có thẩm quyền quản lý cán bộ công chức, viên chức quản lý hồ sơ nhân sự hiệu quả, giảm chi phí, thời gian, công sức, thuận tiện trong việc tra cứu tìm kiếm hồ sơ cán bộ công chức, viên chức. Hệ thống này về phần mềm thường sẽ có các chức năng phân quyền đến từng đơn vị, cho phép Cán bộ quản lý được phân quyền cập nhật, chỉnh sửa và thay đổi thông tin hồ sơ cán bộ, công chức, viên chức của đơn vị mình mà vẫn đảm bảo tính bảo mật của hồ sơ giữa các đơn vị, không để thông tin bị lộ, lọt.

Hệ thống quản lý công chức, viên chức và người lao động tại Việt Nam có nhiệm vụ chính là giúp lãnh đạo quản lý, cơ quan chủ quản và bộ phận chức năng có cái nhìn toàn cảnh về bức tranh nhân sự tại đơn vị, lĩnh vực mà mình chịu trách nhiệm quản lý để có thể cố vấn, tham mưu hay trực tiếp đưa ra những quyết định theo thời gian thực liên quan đến nhân sự, tránh lãng phí tài năng và nguồn lực, giúp giải quyết công việc hiệu quả và triệt để. Công tác này yêu cầu cần được thực hiện nhiều tầng, nhiều lớp tuỳ theo cấp độ phân cấp quản lý, mỗi cấp tương ứng với một đơn vị, một

cấp quản lý nên yêu cầu phải có hệ thống riêng biệt. Ví dụ cụ thể tại mô hình hoạt động của Viện Năng lượng nguyên tử thì một đơn vị trực thuộc sẽ có một hệ thống quản lý nhân sự tại đơn vị mình và do trưởng đơn vị phụ trách quản lý, tất cả các đơn vị trực thuộc Viện Năng lượng nguyên tử sẽ có hệ thống quản lý các đơn vị trực thuộc Viện và do lãnh đạo Viện quản lý, tương ứng với Viện và các cơ quan thuộc Bộ KH&CN sẽ có hệ thống quản lý cán bộ thuộc Bộ và trên nữa là hệ thống quản lý cán bộ nhà nước do Bộ Nội Vụ quản lý, khai thác và sử dụng. Mỗi một hệ thống này sẽ có một cơ sở dữ liệu riêng biệt, phục vụ các công tác chuyên biệt. Để phục vụ công tác nhân sự tại Viện nói riêng và trên toàn quốc nói chung thì yêu cầu xuất dữ liệu và nạp dữ liệu giữa các hệ thống là vô cùng cần thiết.

Tính đến thời điểm năm 2024, bộ máy cơ quan hành chính Nhà nước nước Việt Nam ta hiện nay có tổng cộng 18 Bộ (Bộ Quốc phòng; Bộ Công an; Bộ Ngoại giao; Bộ Xây dựng; Bộ Tư pháp; Bộ Tài chính; Bộ Công thương; Bộ Giao thông Vận tải; Bộ Lao động – Thương binh và Xã hội; Bộ Thông tin và Truyền thông; Bộ Văn hóa, Thể thao và Du lịch; Bộ Giáo dục và Đào tạo; Bộ Nông nghiệp và Phát triển nông thôn; Bộ Kế hoạch và Đầu tư; Bộ Y tế; Bộ Nội vụ; Bộ Khoa học và Công nghệ; Bộ Tài nguyên và Môi trường) 4 cơ quan ngang Bộ (Ủy ban Dân tộc; Thanh tra Chính phủ; Ngân hàng Nhà nước; Văn phòng Chính phủ) và 9 cơ quan thuộc Chính phủ không phải cơ quan hành chính (Viện Hàn lâm Khoa học Xã hội Việt Nam; Viện Hàn lâm KH&CN Việt Nam; Thông tấn xã Việt Nam; Đài Tiếng nói Việt Nam; Đài Truyền hình Việt Nam; Bảo hiểm Xã hội Việt Nam; Ban Quản lý Lăng Chủ tịch Hồ Chí Minh; Ủy ban Quản lý vốn Nhà nước tại doanh nghiệp; Ủy ban Giám sát Tài chính Quốc gia). Với mỗi Bộ, cơ quan ngang Bộ lại có nhiều cơ quan trực thuộc Bộ và với mỗi cơ quan, đơn vị. Theo Bộ Nội vụ thông tin tại Diễn đàn quản trị đất nước tốt (Hội nghị ACCSM21) vào ngày 4 tháng 8 thì nước ta có 1.761.045 viên chức và 233.219 công chức trên cả nước. Để có thể quản lý, sử dụng và khai thác thông tin từ lượng dữ liệu cán bộ, công chức, viên chức khổng lồ này, việc xây dựng hệ thống quản lý dữ liệu công chức, viên chức là cấp thiết.

Để giải quyết yêu cầu cấp thiết trên, theo đề nghị của Bộ trưởng Bộ Thông tin và Truyền thông, Thủ tướng Chính phủ đã ra Quyết định số 1819/QĐ-TTg ngày 26 tháng 10 năm 2015 về việc Phê duyệt chương trình quốc gia về ứng dụng công nghệ thông tin trong hoạt động của cơ quan nhà nước giai đoạn 2016 – 2020. Để thực hiện, triển khai Quyết định này, Bộ Nội vụ đã ban hành Quyết định số 4223/QĐ-BNV ngày 30 tháng 11 năm 2016 về việc Ban hành Danh mục chuẩn thông tin và Quy định kỹ thuật về dữ liệu dùng cho phần mềm, cơ sở dữ liệu cán bộ, công chức, viên chức và cán bộ, công chức. Quyết định này được ban hành nhằm mục đích công bố danh mục chuẩn thông tin gồm các trường thông tin được áp dụng thống nhất trong việc cập nhật dữ liệu, lưu trữ, tra cứu, tìm kiếm, thống kê và tích hợp dữ liệu về cán bộ, công chức, viên chức và cán bộ, công chức cấp xa.

Theo đề nghị của Bộ Nội Vụ, vào ngày 25 tháng 6 năm 2020, Thủ tướng Chính phủ đã ban hành Quyết định số 893/QĐ-TTg về việc phê duyệt đề án xây dựng cơ sở dữ liệu quốc gia về cán bộ, công chức, viên chức trong các cơ quan nhà nước. Quyết định này được đưa ra với quan điểm "Xây dựng và phát triển cơ sở dữ liệu quốc gia về cán bộ, công chức, viên chức trong các cơ quan nhà nước phải phù hợp, thống nhất, đồng bộ với quá trình xây dựng Chính phủ điện tử/Chính quyền điện tử; đảm bảo thống nhất về quản lý cán bộ, công chức, viên chức và tổ chức bộ máy, biên chế; thông tin cơ bản về cán bộ, công chức và tổ chức bộ máy, biên chế của các cơ quan nhà nước được quản lý tập trung, thống nhất trong cơ sở dữ liệu". Hệ thống này sẽ do Bộ Nội Vụ xây dựng, lưu trữ tại Bộ Nội Vụ và do Bộ Nội vụ quản lý và vận hành. Tuy nhiên Cơ sở dữ liệu về cán bộ, công chức, viên chức thành phần của từng bộ, ngành và địa phương sẽ do các bộ, ngành và địa phương đó xây dựng, vận hành theo đặc thù riêng của từng bộ, ngành và địa phương.

Tuy nhiên theo Theo Kế hoạch 141/KH-BCĐTKNQ18 năm 2024 định hướng sắp xếp, tinh gọn tổ chức bộ máy của Chính phủ có nêu về định hướng cơ cấu, sắp xếp và hợp nhất 14 bộ, cơ quan ngang Bộ. Hiện tại, cơ cấu tổ chức của bộ máy Chính phủ tại nước ta sau sắp xếp gồm 14 Bộ và 3 Cơ quan ngang Bộ. Chính vì thế nên gây ra tình trạng cơ sở dữ liệu về nhân sự vừa thừa, vừa thiếu, lại chồng chéo về mặt dữ

liệu và hệ thống quản lý cán bộ, viên chức và người lao động tại Viện Năng lượng nguyên tử Việt Nam cũng không phải là ngoại lệ.

1.3. Các hệ thống quản lý cán bộ, công chức, viên chức đã được triển khai tại Viện Năng lượng nguyên tử

Do đặc thù của Viện Năng lượng nguyên tử Việt Nam là đơn vị hành chính sự nghiệp cấp 3, được phân cấp và quản lý bởi Bộ KH&CN nên cán bộ thuộc Viện đều thuộc sự quản lý của Bộ. Chính vì thế nên Viện Năng lượng nguyên tử Việt Nam sử dụng và khai thác các phần mềm quản lý cán bộ do Bộ KH&CN triển khai. Cụ thể, trong thời gian qua Viện đã tham gia xây dựng, triển khai và khai thác từ những hệ thống quản lý cán bộ sau:

1.3.1. Hệ thống quản lý cán bộ, công chức, viên chức của Bộ KH&CN xây dựng năm 2020

Để có thể xây dựng được cơ sở dữ liệu đầy đủ, chính xác về đội ngũ cán bộ, công chức, viên chức; sử dụng thông tin cán bộ, công chức, viên chức để phục vụ cho việc tra cứu, xây dựng các báo cáo, thống kê, kiểm soát chặt chẽ số lượng cán bộ, công chức, viên chức, Bộ KH&CN đã ban hành Quyết định 1057/QĐ-BKHCN vào ngày 23 tháng 4 năm 2020 về việc công bố kế hoạch triển khai phần mềm quản lý cán bộ công chức, viên chức và người lao động tại Bộ KH&CN năm 2020, trong đó có Viện Năng lượng năng lượng nguyên tử Việt Nam là Viện nghiên cứu trực thuộc. Hệ thống này được truy cập tại địa chỉ: "https://qlcb.most.gov.vn/Pages/default.aspx". Hệ thống QLCBCCVC được sử dụng là hệ thống do Trung tâm Công nghệ thông tin triển khai và do Vụ TCCB quản lý, cho phép Bộ KH&CN quản lý và khai thác thông tin hồ sơ cán bộ của các đơn vị trực thuộc. Hệ thống được tạo trên nền tảng Microsoft ASP.NET Framework, viết bằng JavaScript và chạy trên máy chủ tại Bộ KH&CN ở địa chỉ 113 Trần Duy Hưng, Hà Nội. Bên cạnh đó, Vụ TCCB cũng phân quyền quản lý đến từng đơn vị trực thuộc Bộ nhằm thực hiện việc cập nhật, sử dụng và khai thác hồ sơ cán bộ tại đơn vị qua Công văn số 203/TCCB ngày 23 tháng 4 năm 2020 và Viện Năng lượng nguyên tử Việt Nam cũng không phải là ngoại lệ.



Hình 1.3 Giao diện hệ thống QLCBCCVC của Bộ KH&CN 2020

Hệ thống này quản lý bằng cách phân quyền tài khoản Quản lý Hệ thống cho từng đơn vị (mỗi đơn vị sẽ có 1 tài khoản lý). Tài khoản quản lý này cho phép cán bộ phụ trách quản lý cán bộ có khả năng khởi tạo tài khoản theo từng cán bộ, công chức, viên chức đang làm việc tại đơn vị qua email cá nhân. Cán bộ làm việc tại đơn vị sẽ tiến hành đăng ký tài khoản với cán bộ phụ trách quản lý, sau đó cán bộ quản lý cán bộ sẽ tiến hành khởi tại tài khoản và hướng dẫn các cán bộ tại đơn vị mình khai hồ sơ vào Form Sơ yếu lý lịch cán bộ, công chức mẫu 2C-BNV/2008 gồm các thông tin cán bộ như họ tên; ngày tháng năm sinh; thông tin tuyển dụng; thông tin chức vụ; thông tin chức danh; diễn biến lương; quá trình đào tạo; trình độ lý luận chính trị; quản lý nhà nước; trình độ tin học, ngoại ngữ; khen thưởng, kỷ luật; quan hệ gia đình và các thông tin liên quan đến quá trình quản lý cán bộ khác.

Dữ liệu cán bộ được khởi tạo qua tài khoản được cấp quyền bởi cơ quan chủ quản và cá nhân cán bộ, công chức, viên chức và người lao động trực thuộc tự kê khai thông tin cá nhân liên quan đến công tác nhân sự dưới sự hướng dẫn và hỗ trợ của cán bộ quản lý nhân sự. Sau đó cán bộ quản lý sẽ căn cứ và đối chiếu và xác thực các thông tin được khai báo với những thông tin trên hồ sơ, bằng cấp hiện hữu và tiến hành phê duyệt. Từ đó hệ thống sẽ giúp cán bộ quản lý lưu trữ và có cái nhìn tổng thể về các thông tin liên quan đến nhân sự tại đơn vị mình. Hệ thống này cũng có chức năng giúp cán bộ quản lý cán bộ có thể trích xuất thông tin, báo cáo theo từng mục

đích cụ thể nhằm phục vụ các báo cáo, thống kê, danh sách với các mục đích khác nhau tại đơn vị mình quản lý ra file excel hoặc file pdf theo yêu cầu. Hệ thống cũng có thể dựa trên các thông tin như ngày sinh và đưa ra lộ trình hưu trí, chế độ chính sách và nâng lương theo đợt, tránh việc quên lãng, sai sót do việc quản trị thủ công truyền thống mắc phải.

1.3.2. Hệ thống quản lý cán bộ, công chức, viên chức của Bộ KH&CN năm 2024

Thực hiện Quyết định số 893/QĐ-TTg ngày 25/6/2020 của Thủ tướng Chính phủ, Bộ Nội Vụ đã ban hành Công văn Số 4659/BNV-VP ngày 06 tháng 8 năm 2024 về việc bổ sung và làm sạch dữ liệu CBCCVC tới Bộ trưởng Bộ KH&CN với mục tiêu rà soát, cập nhật đầy đủ tiêu chí trong hồ sơ nhằm đồng bộ với CSDLQG về CBCCVC trước ngày 31/10/2024 để kịp thời đối khớp với CSDLQG về dân cư làm sạch dữ liệu để đưa vào sử dụng, khai thác phục vụ công tác quản lý. Lý do có yêu cầu này tới Bộ KH&CN là do Hệ thống QLCBCCVC được Bộ KH&CN triển khai và lưu trữ dữ liệu tại Trung tâm Công nghệ thông tin cán bộ dưới định dạng NCHAR, NVARCHAR và NTEXT gây xung đột dữ với dữ liệu tại CSDLQG về dân cư. Một ví dụ cụ thể là CSDLQG về dân cư được đồng bộ với các cơ sở dữ liệu tại các Bộ, Ban, Ngành thông qua số CCCD 12 số gắn chíp được Bộ Công an triển khai từ ngày 01/01/2021. Tuy nhiên trường dữ liệu CCCD tại Hệ thống QLCBCCVC của Bộ KH&CN lại là dạng chuỗi kí tự nên không thể đồng bộ được với CSDLQG về dân cư. Ngoài trường CCCD này ra, hầu như toàn bộ các trường dữ liệu khác đều bị xung đột về kiểu dữ liệu khiến cho việc đồng bộ không thể thực hiện được.

Thực hiện chỉ đạo của Bộ trưởng Bộ KH&CN ngày 8/8/2024 tại Công văn số 4659/BNV-VP ngày 06/8/2024 của Bộ Nội vụ về việc bổ sung và làm sạch dữ liệu về CBCCVC; Trung tâm Công nghệ thông tin Bộ đã tiến hành triển khai phần mềm mới đồng thời phối hợp với Vụ Tổ chức cán bộ hướng dẫn các cán bộ được phân quyền phụ trách quản lý hệ thống QLCBCCVC hiện có. Qua đó, các đơn vị trực thuộc Bộ, trong đó có Viện Năng lượng nguyên tử Việt Nam, tiến hành rà soát, cập nhật đầy đủ các trường

thông tin theo yêu cầu của Bộ Nội Vụ và đồng bộ với CSDL quốc gia về CBCCVC. Địa chỉ phần mềm mới tại https://qlccvc.most.gov.vn/Pages/default.aspx đối với cán bộ được phân quyền quản lý và https://ccvc.most.gov.vn/Pages/default.aspx đối với cán bộ tự khai. Qua đó toàn bộ các cán bộ thuộc Bộ tiến hành khai lại những thông tin của mình giống như đã triển khai tại hệ thống cũ trước đó.



Hình 1.4 Giao diện Hệ thống quản lý cán bộ của Bộ KH & CN (2024)

Dưới góc độ kỹ thuật thì hệ thống mới là một nhà kho dữ liệu – data warehouse - có giao diện tương đối tương đồng với hệ thống cũ và vẫn được triển khai trên cùng Framework ASP.Net của Microsoft. Hệ thống này cũng được xây dựng dựa trên quy trình tích hợp data warehouse kiểu mẫu, từ nhiều nguồn dữ liệu được lưu trữ tại trung tâm, dữ liệu được tải và truy vấn theo yêu cầu sử dụng và hiển thị tại dashboard trung tâm. Điểm khác biệt chính là hệ thống này sử dụng một hệ thống cơ sở dữ liệu nguồn hoàn toàn khác, không kế thừa được từ các hệ thống cơ sở dữ liệu cũ do khác định dạng dữ liệu dẫn đến không thể đồng bộ được dữ liệu theo yêu cầu của Bộ Nội Vụ. Chính vì hệ thống gặp vấn đề quá nghiêm trọng và không thể khắc phục từ lúc khởi tạo nên Bộ KH&CN qua sự tham mưu của Trung tâm Công nghệ thông tin và Vụ Tổ chức cán bộ đã tiến hành triển khai tới các đơn vị trực thuộc hệ thống quản trị nhân sự mới và Viện Năng lượng nguyên tử là một trong số những đơn vị thuộc bộ cũng bị ảnh hưởng.

Do đặc thù Viện đông cán bộ, rải rác nên việc triển khai tới từng cán bộ, vận động cán bộ và người lao động khai và nhập lại dữ liệu là tương đối phức tạp và thiếu hiệu quả. Như đã được trình bày, việc khởi tạo và khai báo dữ liệu cán bộ được thực hiện một cách thủ công tới từng cán bộ hiện hữu. Ngày 01/10/2024, Viện đã hoàn thành việc triển khai và cập nhật dữ liệu của từng cán bộ thuộc Viện tại thời điểm trên để đáp ứng yêu cầu sử dụng và khai thác của các cấp trên.

1.3.3. Hệ thống quản lý cán bộ, công chức, viên chức của Bộ KH&CN (sau sắp xếp) năm 2025

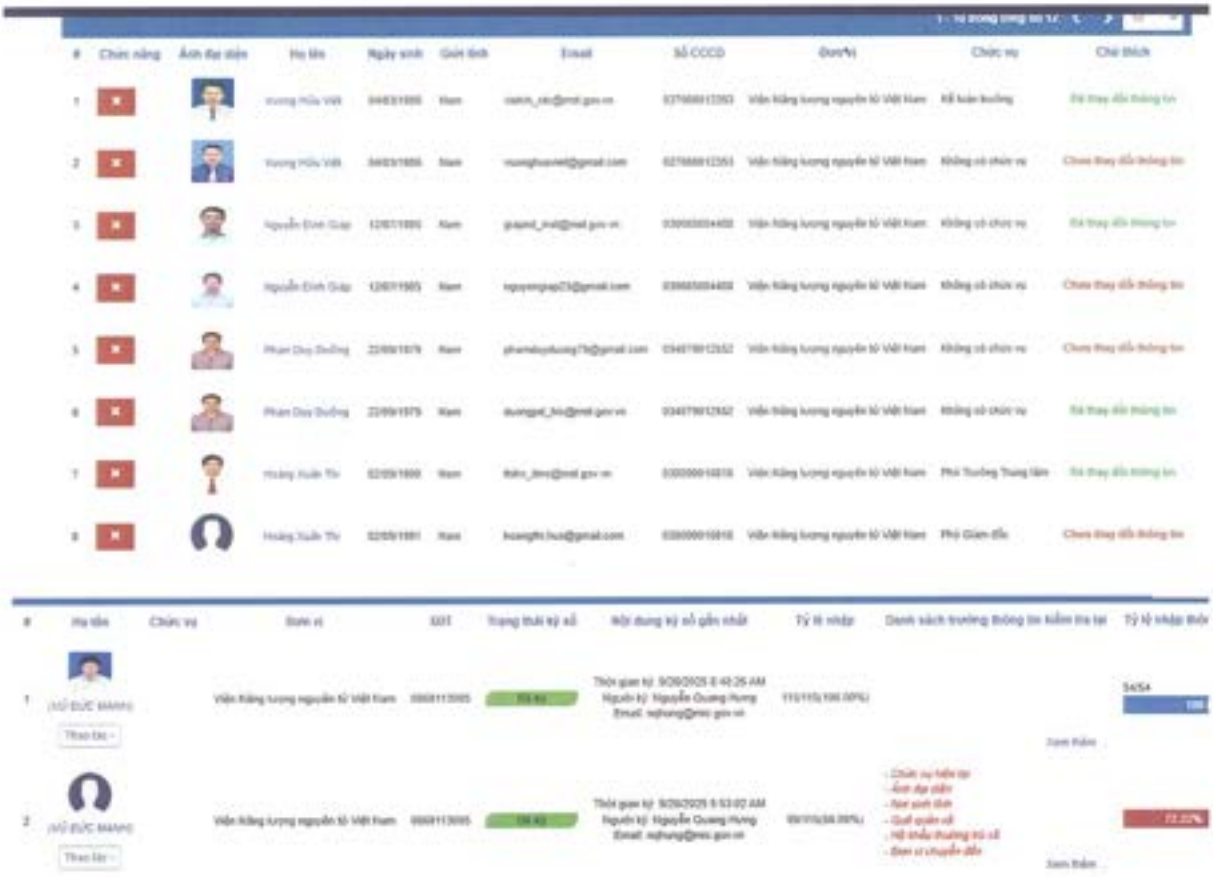
Như đã đề cập trước đó, theo Kế hoạch 141/KH-BCĐTKNQ18 năm 2024 định hướng sắp xếp, tinh gọn tổ chức bộ máy của Chính phủ có nêu về định hướng cơ cấu, sắp xếp và hợp nhất 14 bộ, cơ quan ngang Bộ. Cụ thể theo Kế hoạch, "Hợp nhất Bộ Thông tin và Truyền thông và Bộ KH&CN" và điều này trực tiếp tác động tới Viện Năng lượng nguyên tử bởi việc thay đổi cơ quan chủ quản. Bộ Thông tin và Truyền thông (trước sáp nhập) đã tiến hành triển khai, khai thác và vận hành hệ thống quản lý công chức, viên chức và người lao động theo yêu cầu của Bộ Nội Vụ từ trước đó, đồng thời Bộ Khoa học công nghệ (trước sáp nhập) cũng đã có hệ thống như đề cập ở trên. Thực tế phát sinh ra vấn đề về nhiều cơ sở dữ liệu rời rạc thuộc hai Bộ ngành trước sáp nhập và sau sáp nhập trở thành một cần thiết phải tập hợp, làm sạch và chuẩn hoá dữ liệu từ các kho dữ liệu của cả các đơn vị trực thuộc hai Bộ thành một thể thống nhất để có thể dễ dàng phục vụ mục đích khai thác, quản trị phân tích và báo cáo.

Trong quá trình triển khai, cơ quan chủ quản mới là Bộ KH&CN (sau sáp nhập) kết luận hệ thống về cơ sở dữ liệu của hai Bộ được xây dựng khác nhau và không có định hướng kết hợp từ đầu nên xảy ra xung đột dữ liệu trong giai đoạn đồng bộ giữa hai hệ thống. Chính vì thế nên cơ sở dữ liệu về viên chức và người lao động tại Viện Năng lượng nguyên tử Việt Nam cũng gặp nhiều khó khăn trong quá trình khai thác, sử dụng và báo cáo. Viện được sự hướng dẫn và hỗ trợ của Vụ Tổ chức cán bộ và Trung tâm Thông tin Bộ đã tiến hành triển khai lại việc kê khai lại hồ sơ cán bộ theo như yêu cầu.



Hình 1.5 Giao diện Hệ thống quản lý cán bộ của Bộ KH & CN (2025)

Như đã được giới thiệu tại mục 1, Viện Năng lượng nguyên tử Việt Nam gồm 12 đơn vị trực thuộc, hơn 800 cán bộ, người lao động đang công tác trong đó hầu hết cán bộ đều thuộc diện phải kê khai, khởi tạo lại hồ sơ do việc đồng bộ hệ thống giữa cơ quan chủ quản gặp lỗi. Thực tế hiện nay Viện Năng lượng nguyên tử đã và đang triển khai, hướng dẫn đến từng cán bộ, khởi tạo lại tài khoản, tiến hành kê khai lại thông tin, quá trình công tác như đã từng làm tại Hệ thống QLCBCCVC trước đó. Toàn bộ công việc này đang được Viện tiến hành một cách thủ công, điều đó dẫn đến những bất cập, khó khăn trong việc triển khai.



Hình 1.6 Lỗi lặp hồ sơ trong quá trình đồng bộ sau sáp nhập Bộ

Hình 1.6 đưa ra một vài ví dụ về lỗi đồng bộ hồ sơ cán bộ tại hệ thống quản lý viên chức và người lao động tại Viện Năng lượng nguyên tử. Có thể nhìn thấy theo Hình 1.6, nhiều cán bộ bị lặp hồ sơ do hệ thống không tích hợp được dữ liệu và thông tin giữa các cơ sở dữ liệu nên các hồ sơ bị trùng lặp và chồng kéo. Hiện tại cách duy nhất để khắc phục là đối chiếu từng hồ sơ cán bộ để kiểm tra xem hồ sơ bị trùng lặp dữ liệu hay là cán bộ trùng tên. Quy trình xử lý này rất thủ công và thiếu hiệu quả.

Qua quá trình nghiên cứu, thảo luận cùng với sự giúp đỡ, hỗ trợ nhiệt tình từ phía các cán bộ chuyên môn ở cả Bộ KH&CN trước và sau sáp nhập, học viên đã phần nào xác định được vấn đề kỹ thuật khiến phát sinh các vấn đề trên trong quy trình đồng bộ dữ liệu đã được triển khai này. Nguyên nhân cụ thể và đề xuất kỹ thuật về hướng giải quyết sẽ được trình bày và giải quyết trong chương 3 của luận văn này.

Xuất phát từ những bất cập cụ thể trong công tác của học viên trong việc quản lý, vận hành và khai thác hệ thống QLCBCCVC của Viện Năng lượng nguyên tử Việt Nam cũng như của cơ quan chủ quản là Bộ KH&CN (sau sáp nhập), bài toán được đặt ra là làm thế nào để có thể tích hợp được dữ liệu từ các cơ sở dữ liệu khác nhau về một nguồn dữ liệu được xác định trước, nhanh chóng và đơn giản với hệ thống dữ liệu được chuẩn hóa, chính xác và thực tế. Hệ thống QLCBCCVC này từ việc có được một nguồn dữ liệu cán bộ, viên chức và người lao động tại Viện Năng lượng nguyên tử Việt Nam duy nhất sẽ có khả năng trích xuất thông tin cán bộ từ kho dữ liệu, giúp cho công tác quản trị con người, báo cáo, thống kê dữ liệu về cán bộ đang làm việc một cách thống nhất giúp cho lãnh đạo quản lý và các cán bộ thực hiện công tác nhân sự có cái nhìn tổng thể và số liệu chính xác, có thể kịp thời đề xuất, tham mưu và đưa ra những quyết định, chính sách về nhân sự kịp thời. Từ đó tác giả đề xuất giải pháp **"Tích hợp dữ liệu cho hệ thống quản lý cán bộ, viên chức của Viện Năng lượng nguyên tử"**

CHƯƠNG 2: CÁC KỸ THUẬT TÍCH HỢP DỮ LIỆU

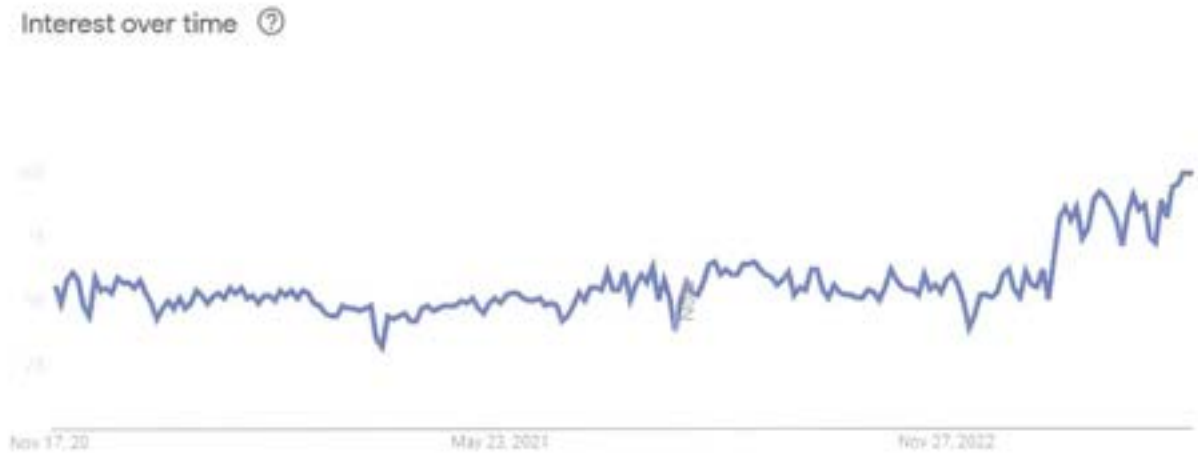
2.1. Lý thuyết về Tích hợp dữ liệu - Data Intergration

2.1.1. Tích hợp dữ liệu trên thế giới

Trong bối cảnh xã hội hiện đại ở thế kỷ 21, trên thế giới nói chung và tại Việt Nam nói riêng đã quá quen thuộc với việc tìm kiếm thông tin qua Google hoặc tài liệu chỉ bằng một vài dòng từ khóa trên điện thoại cũng như máy vi tính. Chính vì thế nên dữ liệu và các cơ sở dữ liệu đang dần dần trở thành một kho báu vô cùng quý giá với người sử dụng cũng như các công ty, tổ chức, các đơn vị cung cấp. Tuy nhiên ít người có thể hình dung được lượng dữ liệu khủng lồ mà các hệ thống trên Internet xử lý và truyền tải đến người dùng phức tạp như thế nào vì dữ liệu thường được lưu trữ ở nhiều nguồn và nhiều dạng khác nhau. Vấn đề tích hợp dữ liệu ngày càng được quan tâm trong cả giới nghiên cứu kỹ thuật và cả giới kinh doanh bởi sự bùng nổ về khối lượng dữ liệu và nhu cầu chia sẻ, khai thác những dữ liệu sẵn có. Để có thể cung cấp các dịch vụ, dữ liệu và khai thác thông tin một cách chính xác và tối ưu nhất đến người sử dụng thì việc ứng dụng tích hợp dữ liệu là cần thiết hơn bao giờ hết.

Vấn đề để kết nối và tổng hợp các nguồn dữ liệu không đồng nhất trong một giao diện duy nhất để phục vụ mục đích truy vấn, khai thác thông tin đã tồn tại một thời gian khá dài. Cụ thể là từ những năm 1980, các nhà khoa học, nhà nghiên cứu trong lĩnh vực công nghệ thông tin và khoa học dữ liệu đã bắt đầu tìm cách để thiết kế các hệ thống có khả năng tương tác với các dữ liệu lưu trữ tại các database khác nhau. Cụ thể là vào năm 1981, John Miles Smith và các đồng tác giả đã đưa ra một giải pháp gọi là Multibase (tạm dịch là Đa cơ sở) [4] . Multibase ngày đó theo như nhóm tác giả đề xuất là một phần mềm hệ thống cho phép truy cập cùng lúc nhiều cơ sở dữ liệu có sẵn, bất nhất và riêng rẽ. Thông qua bài báo được công bố vào ngày 4 tháng 5 năm 1981 tại hội nghị máy tính quốc gia, nhóm tác giả đã mô tả cấu trúc cơ bản của Multibase và xác định rõ định hướng phát triển và thực hiện trong những nghiên cứu sau đó.

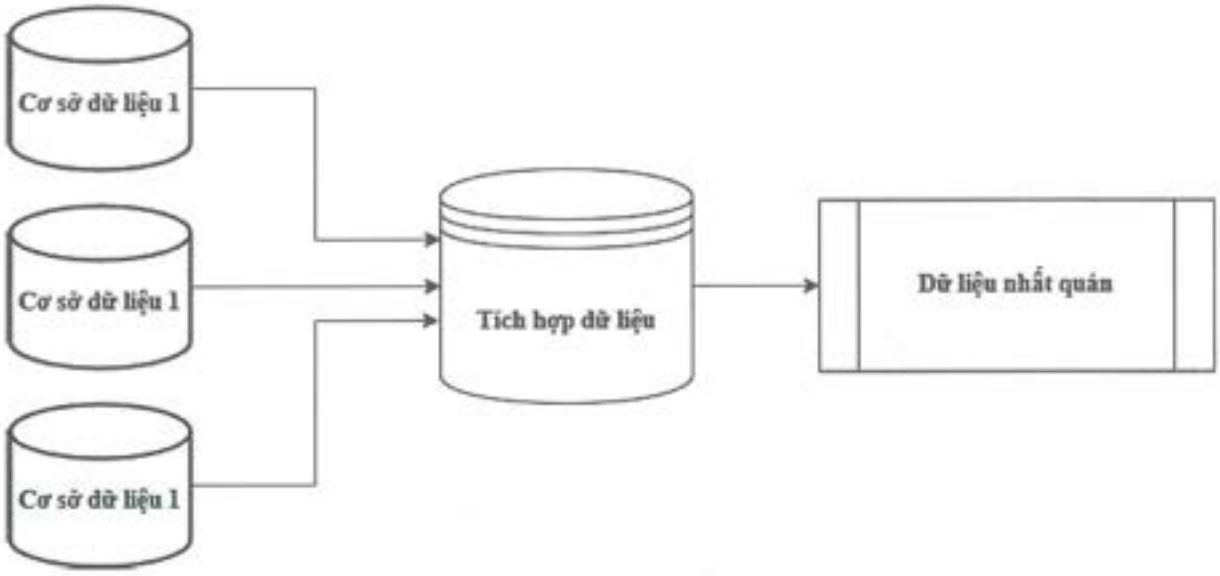
Từ khi được đề cập đến nay, mức độ quan tâm trên thế giới đến việc "tích hợp dữ liệu" không những không suy giảm mà còn liên tục gia tăng đột biến, cụ thể ở Hình 2.1 sau đây thể hiện mối quan tâm đến chủ đề này qua Google Trend trong 5 năm đổ lại.



Hình 2.1 Sự quan tâm đến "Tích hợp dữ liệu" qua công cụ tìm kiếm Google

Lý thuyết về Tích hợp dữ liệu luôn là một chủ đề khiến các nhà nghiên cứu, các tác giả trong và ngoài nước quan tâm sâu sắc và luôn là một đề tài gây tranh cãi. Chính vì vậy, mỗi bài nghiên cứu được đăng tải, một tạp chí khoa học lại đưa ra một định nghĩa khác nhau cho Tích hợp dữ liệu theo mục đích sử dụng và từ nhiều góc nhìn khác nhau của mỗi tác giả. Ví dụ như theo Maurizio Lenzerini định nghĩa tại bài báo của mình thì "Tích hợp dữ liệu là vấn đề tổng hợp dữ liệu tại các nguồn khác nhau và đem đến cho người dùng một cái nhìn thống nhất về những dữ liệu đó". Nhưng theo như Brenton Louie định nghĩa tại tạp chí ScienceDirect thì "Bản chất của việc tích hợp dữ liệu là khả năng truy vấn từ các nguồn dữ liệu khác nhau. Các nguồn dữ liệu này bao gồm nhưng không chỉ là những cơ sở dữ liệu có liên quan với nhau hoặc có thể là các cơ sở dữ liệu tương đồng trong một mạng"[18]. Ở một góc nhìn khác từ một công ty, một tập đoàn công nghệ thông tin thì IBM định nghĩa "Cụm từ Tích hợp dữ liệu được dùng để ám chỉ quy trình tổng hợp và thống nhất dữ liệu từ nhiều nguồn lại thành một định dạng thống nhất và mạch lạc nhằm có thể dùng cho các công tác phân tích, đề xuất và đưa ra quyết định".

Mặc dù thuật ngữ kỹ thuật "Tích hợp dữ liệu" hay "Data Integration" có nhiều khái niệm và cách định nghĩa khác nhau nhưng ở một góc độ rộng hơn, dựa trên những nghiên cứu tiền nhiệm, thì ta có thể thấy thuật ngữ này thường được dùng để mô tả quy trình liên quan đến việc tổng hợp, kết hợp dữ liệu có thể khai được tại các nguồn dữ liệu khác nhau để cung cấp cho người quản trị dữ liệu có được một cái nhìn thống nhất về chúng. Nhu cầu về Tích hợp dữ liệu luôn tăng cùng với nhu cầu về xử lý, chia sẻ và khai thác Dữ liệu lớn bùng nổ trong bối cảnh hiện tại.



Hình 2.2 Sơ đồ minh họa quy trình tích hợp dữ liệu

2.1.2. Minh họa sự cần thiết của việc Tích hợp dữ liệu tại Viện Năng lượng nguyên tử

Viện Năng lượng nguyên tử Việt Nam là một tổ chức nghiên cứu khoa học đồng thời vừa đảm nhiệm sản xuất, kinh doanh các sản phẩm, tổ chức hội nghị và đào tạo tiến sĩ trong lĩnh vực nguyên tử. Viện Năng lượng nguyên tử gồm nhiều đơn vị trực thuộc, phụ trách những mảng công việc cụ thể và có trụ sở tại khắp mọi miền tổ quốc từ Bắc vào Nam. Nếu thống kê chi tiết về từng đầu mục công việc tại từng cơ sở nghiên cứu, viện khoa học, trung tâm triển khai thì một đơn vị như Viện dễ dàng cần đến cả trăm cơ sở dữ liệu khác nhau.

Bộ phận đảm nhiệm công tác Quản lý cán bộ tại Viện Năng lượng nguyên tử có nhiều các cơ sở dữ liệu lưu trữ các thông tin liên quan đến từng cán bộ, công chức, viên chức và người lao động tại đơn vị ở cả dạng điện tử và dạng giấy. Công tác này cũng đòi hỏi phải có thông tin và hệ thống dữ liệu khác về lương gồm các thông tin về quy hoạch, bổ nhiệm, đào tạo, lương, hệ số, bậc lương, phụ cấp, thuế thu nhập, ngày làm thêm giờ để tính được lương nhận của từng cán bộ một cách công bằng và thỏa đáng. Đội ngũ quản lý mảng đào tạo tại Viện cũng phải có cơ sở dữ liệu riêng để đảm bảo việc quản lý các khóa học được tổ chức và chia khóa học sao cho đúng cán bộ vào đúng khóa nhằm nâng cao chất lượng chuyên môn người lao động và phân bố nguồn kinh phí hợp lý. Đồng thời Bộ phận phụ trách đề tài và sản phẩm chiếu xạ cũng phải có từng cơ sở dữ liệu của riêng họ để quản lý được các đề tài nghiên cứu khoa học hay số lượng sản phẩm chiếu xạ tại đơn vị.

Để có thể có cái nhìn thống nhất và khách quan nhất về hơn 800 cán bộ đang công tác và làm việc tại Viện, việc Tích hợp dữ liệu cán bộ giữa các đơn vị trực thuộc trên toàn quốc là nhu cầu cấp thiết. Chỉ khi người quản trị có cái nhìn thống nhất về cán bộ, nghiên cứu viên đang công tác tại đơn vị, họ mới có thể tham mưu, cố vấn lãnh đạo đơn vị ra các quyết định liên quan đến nhân sự như đào tạo, bồi dưỡng, quy hoạch, bổ nhiệm hay luân chuyển, điều động, biệt phái một cách hiệu quả. Công tác này là công tác quan trọng bậc nhất trong các nhiệm vụ của bài toán quản trị con người tại một tập thể vì nó giống như trái tim có nhiệm vụ đưa máu đến các cơ quan trong cơ thể. Kỹ thuật Tích hợp dữ liệu trong thời đại dữ liệu ngày càng nhiều và càng phân tán giúp giải quyết triệt để vấn đề trùng lặp dữ liệu trong công tác quản lý nhân sự.

Bằng việc lấy một ví dụ hết sức tối giản, chúng ta có thể thấy được có vô vàn lý do cho sự cần thiết và nguồn gốc của rất nhiều các cơ sở dữ liệu trải khắp các phòng ban, các đơn vị trong một tổ chức, tập thể nói chung và tại Viện Năng lượng nguyên tử Việt Nam. Nhưng càng ngày phạm vi hoạt động của Viện càng rộng, càng ngày càng có nhiều cán bộ, công chức, viên chức và người lao động khiến cho nhu cầu đào tạo càng ngày càng tăng dẫn đến Viện thành lập hẳn một đơn vị chuyên trách về tổ chức thực hiện đào tạo về an toàn bức xạ và chuyên môn hạt nhân.

Chính vì những lý do trên nên số lượng nguồn dữ liệu liên quan đến nhân sự tại một tập thể, công ty đặc biệt là tại cơ quan quản lý nhà nước là rất lớn và không thống nhất. Mặc dù vậy nhưng đơn vị phụ trách công tác cán bộ luôn cần đến khả năng có thể truy cập thông tin liên quan đến tất cả cán bộ từ tất cả các cơ sở dữ liệu này. Hiện tại để truy cập một thông tin khi cần thiết, cán bộ tổ chức phải liên hệ tới từng đầu mối quản lý từng cơ sở dữ liệu riêng biệt. Việc tổng hợp dữ liệu từ các cơ sở dữ liệu lại là một vấn đề cấp thiết với công tác quản lý nói chung và công tác cán bộ nói riêng.

Đặc biệt trong bối cảnh hiện tại, sau khi Chính phủ Việt Nam ra quyết định về sắp xếp và tinh gọn bộ máy, nguồn cơ sở dữ liệu về nhân sự tại Viện vì thế mà tăng lên do thay đổi cơ quan chủ quản. Nguồn dữ liệu tăng lên khiến cho các công tác nhân sự cùng với việc tổng hợp và báo cáo số liệu trở nên phức tạp hơn và thiếu chính xác.

2.1.3. Tóm lược về tích hợp dữ liệu

Tổng hợp từ những nghiên cứu tiền nhiệm, chúng ta có thể hiểu khái quát tích hợp dữ liệu là các kỹ thuật, công cụ, phương thức hay một quá trình tổng hợp, kết hợp các dữ liệu không đồng nhất từ các nguồn dữ liệu khác nhau thành một tập hợp dữ liệu thống nhất như đã được định trước. Đặc điểm của một hệ thống tích hợp dữ liệu có tính linh hoạt về nguồn dữ liệu. Nguồn dữ liệu có thể sử dụng bao gồm cả các cơ sở dữ liệu trong các hệ thống khác nhau, cũng có thể là thông tin con người từ các tổ chức khác nhau, cũng có thể là hoá đơn bán hàng từ các cửa hàng khác nhau. Các nguồn dữ liệu có thể sử dụng các ngôn ngữ khác nhau, các định dạng khác nhau. Tích hợp dữ liệu giúp người dùng có có một cái nhìn nhất quán và tổng thể đầy đủ dữ liệu của họ để từ đó có thể khai thác, hiểu biết các thông tin chuyên sâu hơn.

Mục đích then chốt của việc Tích hợp dữ liệu là nó cho phép khả năng truy cập đồng bộ với một nhóm những nguồn dữ liệu không đồng nhất. Hiện nay có rất nhiều giải pháp cho việc tích hợp dữ liệu. Tuỳ theo nhu cầu, mục đích sử dụng thực tế, nguồn kinh phí cho phép, loại Database và cấu trúc của Database mà áp dụng các mô hình tích hợp dữ liệu khác nhau. Chính vì vậy mà không chỉ các công ty, tập đoàn

lớn xử lý Dữ liệu lớn mới cần thiết phải xử dụng Tích hợp dữ liệu mà kể cả trong những tổ chức nhỏ hơn cũng nên áp dụng công nghệ này để việc vận hành trở nên thuận lợi hơn.

2.2. Các kỹ thuật tích hợp dữ liệu

Nhu cầu về tích hợp dữ liệu luôn tồn tại từ khi những cơ sở dữ liệu đầu tiên được ra đời, từ lúc dữ liệu còn được lưu trữ dưới dạng vật lý như thông tin được viết trên giấy hay trên các phiếu đục lỗ. Việc truy xuất dữ liệu lúc này thường xuyên phải được thực hiện thủ công bằng cách tìm kiếm bằng tay sau đó đọc và lọc trên giấy. Cho đến ngày nay, việc tích hợp dữ liệu thủ công vẫn là một lựa chọn đối với các cơ sở dữ liệu có quy mô nhỏ do đặc tính dễ triển khai của nó. Tuy nhiên việc người dùng thu thập và tổng hợp dữ liệu từ các nguồn một cách thủ công sẽ dễ dẫn đến sai sót, tốn nhiều thời gian và không hiệu quả đối với khối lượng dữ liệu lớn. Cùng với sự phát triển của công nghệ cũng như nhu cầu quản lý dữ liệu, ngày nay các nhà quản trị dữ liệu có nhiều cách để tiếp cận với kỹ thuật này. Sau đây tác giả sẽ đưa ra một vài kỹ thuật tích hợp thông dụng

2.2.1. Kỹ thuật Hợp nhất dữ liệu ETL/ELT (Extract, Transform, Load/Etract, Load, Tranform – Trích xuất, Chuyển đổi, Tải/Trích xuất, Tải, Chuyển đổi)

Trong tất cả các kỹ thuật dùng để triển khai quá trình Tích hợp dữ liệu hiện nay thì kỹ thuật ETL là một trong số những kỹ thuật thông dụng nhất. Đối với các nhà khoa học máy tính hay các chuyên gia quản trị dữ liệu thì ELT luôn là lựa chọn kỹ thuật đầu tiên được cân nhắc khi giải quyết bài toán Tích hợp dữ liệu. Nó cũng là kỹ thuật được sử dụng nhiều nhất và phổ biến nhất trong việc xử lý Big Data và trong Data Warehouse. Từ khi ra đời đến nay có thể coi kỹ thuật ETL là một kỹ thuật cốt lõi trong quá trình Tích hợp dữ liệu. Trong bối cảnh thời đại mới khi mà nhu cầu và yêu cầu về chuyển đổi số càng ngày được đề cao cả ở trong nước và quốc tế thì ETL đóng vai trò như một nền móng căn bản để có thể tiến hành xây dựng và triển khai nhiệm vụ chuyển đổi số, nhằm bắt kịp và hội nhập với thế giới.

Kỹ thuật ETL – Tên tiếng anh đầy đủ là Extract, Transform, Load – hay còn được gọi là quy trình Trích xuất, Chuyển đổi, Tải là một kỹ thuật quan trọng trong bài toán quản lý dữ liệu và tích hợp dữ liệu. ETL là cụm từ chỉ kỹ thuật, quy trình xử lý tính toán dữ liệu gồm 3 giai đoạn chính như tên gọi của nó theo thứ tự là Trích xuất, Chuyển đổi và Tải. Ở giai đoạn đầu dữ liệu được trích xuất từ nguồn, sau đó được chuyển đổi và tải lên một kho chứa dữ liệu. Đặc điểm chính của kỹ thuật này là dữ liệu có thể được thu thập từ nhiều nguồn và sau quá trình chuyển đổi cũng có thể xuất ra một hay nhiều đích. Theo tác giả Ralph Kimball đề cập đến trong cuốn sách The Data Warehouse ETL Toolkit thì "Một hệ thống ETL được thiết kế chuẩn mực và hợp lý sẽ tách, trích xuất dữ liệu từ các hệ thống nguồn, chuẩn hoá các các tiêu chuẩn về dữ liệu như kiểu dữ liệu và tính trọn vẹn của dữ liệu, đồng thời đảm bảo được dữ liệu đầu ra tuân thủ về mặt cấu trúc với các yêu cầu được đặt ra. Hệ thống ETL này cũng có thể cung cấp dữ liệu ở định dạng sẵn sàng để các nhà phát triển và quản trị viên có thể xây dựng cơ sở dữ liệu dựa trên quyết định của người dùng cuối". [13]



Hình 2.3 Sơ đồ mô tả quy trình ETL

Ở giai đoạn đầu của quy trình, kỹ thuật ETL bắt đầu từ việc trích xuất dữ liệu từ các nguồn dữ liệu hoặc các nguồn hệ thống được cung cấp. Bước này thường được các nhà nghiên cứu khoa học máy tính cho là khía cạnh quan trọng bậc nhất trong cả quy trình của kỹ thuật ELT. Trong hầu hết phần lớn các trường hợp áp dụng, việc

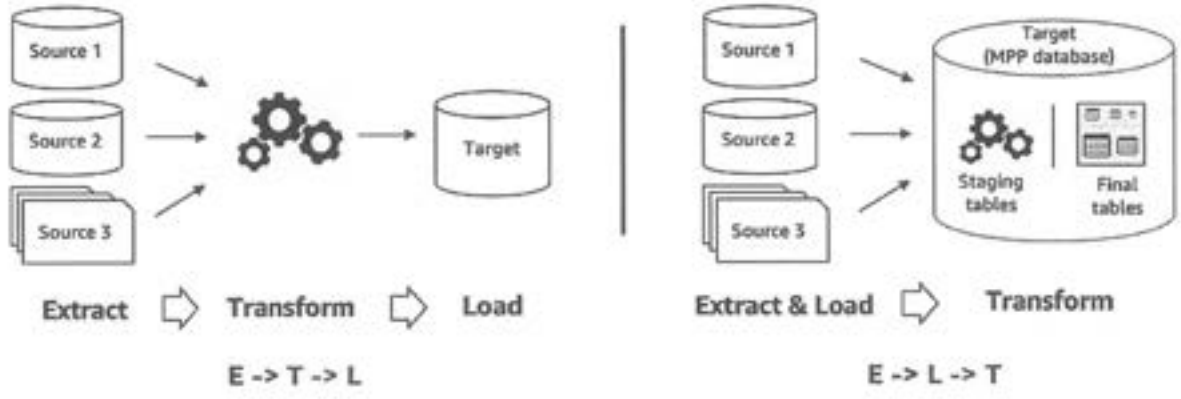
trích xuất dữ liệu chuẩn xác đóng vai trò nền móng cho sự thành công của các bước kế tiếp. Từ xưa đến nay thì hầu hết các Kho dữ liệu được lập ra thường tổng hợp dữ liệu từ nhiều nguồn dữ liệu, cơ sở dữ liệu và hệ thống dữ liệu khác nhau. Các nguồn dữ liệu này thường được cấu thành bởi nhiều kiểu định dạng cơ sở dữ liệu và cấu trúc dữ liệu khác nhau như csv, tsv, xml hay json… Các dữ liệu thô tại các nguồn này cần phải được trích xuất để có thể sử dụng cho các bước tiếp theo và giai đoạn Trích xuất hay Extract sẽ đảm nhiệm phần công tác này. Một phần trách nhiệm của quá trình này là xác thực xem dữ liệu được lấy ra từ các nguồn được cung cấp có đúng giá trị mong đợi hay không và sẽ tiến hành xác thực để tiếp tục quy trình.

Giai đoạn thứ hai của quy trình ETL, sau khi đã hoàn thiện việc trích xuất dữ liệu cần thiết theo yêu cầu thì dữ liệu được đưa vào giai đoạn chuyển đổi. Đối với một quy trình ETL thì mặc dù giai đoạn Trích xuất mang tính nền tảng cho quá trình tuy nhiên giai đoạn Chuyển đổi – Transform lại là giai đoạn phức tạp và quan trọng nhất. Dữ liệu thô sau khi được trích xuất từ nguồn tại giai đoạn một sẽ được làm sạch và biến đổi tại đây theo các yêu cầu của người dùng. Trong các kỹ thuật khoa học máy tính, thuật ngữ Chuyển đổi dữ liệu hay Data Transformation thường được dùng chỉ quá trình chuyển đổi dữ liệu từ một định dạng hoặc cấu trúc này sang một định dạng hoặc cấu trúc khác để chuẩn bị nó trước khi được nạp vào đích cuối tuỳ vào mục đích sử dụng, phân tích. Trong ETL, một chức năng quan trọng của Data Transformation là làm sạch dữ liệu với đối tượng dữ liệu gốc và loại bỏ những dữ liệu bị trùng lặp. Ví dụ cụ thể với dữ liệu quản lý nhân sự: Khi ta tiến hành trích xuất dữ liệu liên quan về 1 nhân sự cụ thể từ nhiều database khác nhau như database về lương, database đơn vị chủ quản và database đào tạo thì cả 3 bảng thông tin về nhân sự này thường sẽ trùng lặp các cột thông tin về họ tên, năm sinh, số căn cước, và quá trình Chuyển đổi sẽ xoá các cột dữ liệu trùng ở kết quả cuối.

Giai đoạn cuối cùng của quy trình ETL là quy trình Tải dữ liệu hay còn gọi là Load. Đây là bước cuối cùng và cũng là bước có tính trực quan nhất của kỹ thuật ETL. Dữ liệu thô sau khi được trích xuất từ các nguồn dữ liệu đầu vào, qua quá trình

chuyển đổi và làm sạch theo những quy tắc đã được quy định từ trước, sẽ được tải vào một đích đến được chỉ định từ trước (có thể là bất kể dạng lưu trữ dữ liệu nào, từ các mô hình cơ sở dữ liệu database căn bản đến các mô hình datawarehouse hay datalake phức tạp). Tuỳ theo nhu cầu sử dụng và độ phức tạp của mô hình dữ liệu đích mà quy trình này thực hiện các hình thức tải khác nhau.

Song hành với kỹ thuật ETL là kỹ thuật ELT hay còn được biết đến với tên gọi đầy đủ là Extract, Load, Transform (Trích xuất, Tải, Chuyển đổi). Cụ thể về các bước của ELT về cơ bản có nhiều nét tương đồng với ETL về mặt kỹ thuật thực hiện đều gồm ba giai đoạn chính như đã được trình bày. Tuy nhiên, ELT khác biệt với ETL ở thứ tự và quy trình ba bước của nó. Với ETL, người dùng tiến hành trích xuất dữ liệu thô từ các nguồn khác nhau, sau đó tiến hành chuyển đổi dữ liệu và cuối cùng là tải tài nguyên dữ liệu sau khi được chuyển đổi vào đích đến. Còn với ELT, người dùng sau khi tiến hành trích xuất dữ liệu thô từ các nguồn sẽ tiến hành tải luôn các dữ liệu đó ở trạng thái nguyên bản sau khi được trích xuất vào hệ quản trị dữ liệu đích sau đó mới tiến hành chuyển đổi dữ liệu nếu cần ngay tại hệ thống đích.



Hình 2.4 So sánh giữa kỹ thuật ETL và ELT

Quy trình kỹ thuật ETL và ELT mặc dù có nhiều điểm tương đồng ở từng giai đoạn riêng biệt nhưng chúng có những điểm khác biệt cụ thể, và các ưu nhược riêng, các nhà nghiên cứu khoa học máy tính thường tuỳ vào từng trường hợp cụ thể để lựa chọn kỹ thuật sử dụng. Điểm khác biệt lớn nhất giữa hai kỹ thuật này là vị trí thực

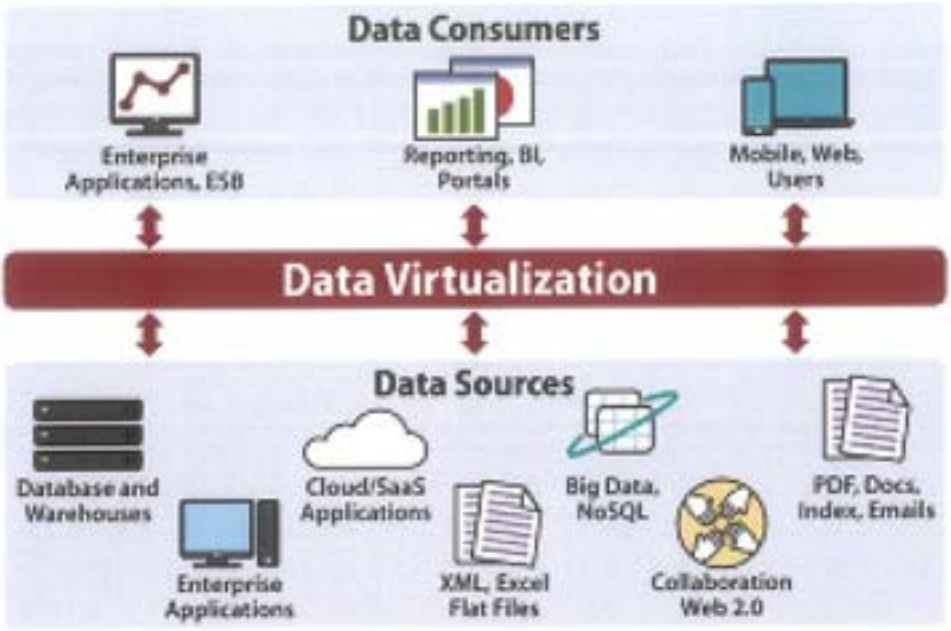
hiện giai đoạn chuyển đổi và tải. Chính từ điểm khác biệt này dẫn đến sự chênh lệch về tốc độ của hai quy trình này. ELT nhanh hơn ETL do ETL có thêm một bước trung gian là chuyển đổi dữ liệu trước khi tải vào mục tiêu cuối trong khi ELT thực hiện thao tác này song song, dữ liệu được chuyển đổi theo nhu cầu tại thời gian thực. Tuy vậy, ETL đã tồn tại và được nghiên cứu rất lâu trong ngành khoa học máy tính nên việc triển khai và tìm kiếm tài liệu và được hỗ trợ rộng rãi. ELT có tuổi đời ít hơn so với ETL và là một khái niệm khá mới trong ngành quản trị dữ liệu nên việc triển khai và hỗ trợ còn phức tạp và ít nền tảng hỗ trợ. Trên thị trường hiện nay, phần lớn các công cụ Tích hợp dữ liệu thường có thiên hướng sử dụng quy trình ETL, trong khi đó thì ELT lại phổ biến và được áp dụng nhiều hơn trong các ứng dụng quản trị dữ liệu, nhà kho dữ liệu và hồ dữ liệu.

Một số công cụ đáng chú ý cung cấp giải pháp cho quy trình ETL trong thời đại này là: Informatica PowerCenter – Công cụ tích hợp dữ liệu thường được dùng kèm với trung tâm tích hợp dữ liệu của công ty Informatica; Apache Airflow – Nền tảng mã nguồn mở cho các quy trình ETL do AirBnB phát triển; Pentaho Data Integration (Cũng được biết đến với tên gọi Pentaho Kettle) – Công cụ mã nguồn mở do tập đoàn Pentaho phát triển; Talend Open Studio – Công cụ ETL với mã nguồn mở miễn phí do công ty Talend phát triển. Thư viện Pandas Python – Thư viện mã nguồn mở trong Python.

2.2.2. Kỹ thuật Data Virtualization

Ảo hoá dữ liệu hay còn được gọi là Data Virtualization là một trong những phương pháp tích hợp dữ liệu khá mới so với ETL hay ELT. Trong khoa học máy tính, ảo hoá dữ liệu là công nghệ cung cấp một lớp lưu trữ ảo từ kết quả truy vấn các dữ liệu phân tán với các nguồn dữ liệu khác nhau. Từ đó, các công cụ truy vấn và hiển thị sẽ dựa trên các lớp lưu trữ ảo này mà cung cấp các chức năng truy cập, khai thác và thao tác với dữ liệu trên thời gian thật tại nơi lưu trữ. Theo tác giả Tejasvi Addagada đề cập trong bài nghiên cứu của mình thì "Ảo hoá dữ liệu là toàn bộ quá trình giới thiệu một lớp logic giữa người sử dụng dữ liệu và nền tảng cung cấp dữ

liệu. Lớp logic này có thể chạy theo nhu cầu người dùng tại thời gian thực, nó có thể thao tác trích dữ liệu, chuyển đổi dữ liệu và kết hợp dữ liệu trong môi trường không gian ảo để đưa những dữ liệu cần thiết đến với người dùng" [14].



Hình 2.5 Minh họa về Ảo hóa dữ liệu giữa các nguồn dữ liệu và người dùng

Điểm mạnh vượt trội của công nghệ Ảo hoá dữ liệu so với kỹ thuật ETL phổ thông được đề cập trước đó tại luận văn này là nó phép người dùng và các ứng dụng quản trị dữ liệu truy cập và thao tác dữ liệu theo thời gian thực, tuỳ vào tình hình thực tế và nhu cầu hiện tại mà có thể truy vấn dữ liệu hay dạng dữ liệu mong muốn một cách nhanh chóng mà không cần tải dữ liệu vào đích đến là cơ sở dữ liệu hay kho dữ liệu truyền thống. Lớp logic ảo hoá của công nghệ này đóng vai trò như một phiên dịch viên thời gian thực, là cầu nối giữa người dùng dữ liệu và các cơ sở dữ liệu nguồn.

Về cách thức hoạt động của công nghệ này cũng có nét tương đồng với kỹ thuật ETL ở điểm nó cũng gồm 3 bước chính, tuy nhiên cụ thể cách thức hoạt động, mục tiêu cụ thể và nhiệm vụ cần hoàn thành ở từng bước là khác nhau. Đầu tiên hệ thống Ảo hóa sử dụng các giao thức kết nối để kết nối trực tiếp với các nguồn dữ liệu, các nguồn dữ liệu này có thể có cấu trúc khác nhau và từ những nền tảng dữ liệu cũng

như cơ sở dữ liệu, nhà kho dữ liệu lớn hay các hồ dữ liệu khác nhau. Các giao thức này không trích xuất dữ liệu hay sao chép dữ liệu mà mục đích chính là thiết lập quyền truy cập vào các nguồn dữ liệu. Sau đó hệ thống sẽ tạo một lớp ảo hay còn được biết đến là lớp logic ảo hóa. Lớp ảo này đóng vai trò cầu nối giữa các nguồn dữ liệu được liên kết ở bước đầu với người dùng, tạo lập sơ đồ kết nối. Tại đây, lớp ảo này sẽ mô tả cấu trúc logic dữ liệu, mối quan hệ giữa các cơ sở dữ liệu, các chỉ tiêu và quy tắc ánh xạ. Sau đó lớp này cũng có nhiệm vụ hiển thị cho người dùng một mô hình dữ liệu thống nhất và đơn giản hơn sau khi đã được dịch từ dữ liệu nguồn, che đi sự phức tạp và hỗn loạn của dữ liệu nguồn theo yêu cầu tại thời gian thực. Cuối cùng dữ liệu được tổng hợp và cung cấp tới một bộ dữ liệu thống nhất tới người dùng tại một điểm truy cập duy nhất.

Một kỹ thuật tích hợp dữ liệu quan trọng trong môi trường tích hợp dữ liệu cùng với các kỹ thuật trên mà luận văn này mong muốn trình bày là Data Federation – hay còn gọi là Liên kết dữ liệu. Data Federation là một kỹ thuật tích hợp dữ liệu cho phép người dùng truy cập và truy vấn dữ liệu từ nhiều nguồn dữ liệu khác nhau cùng lúc mà không cần phải trích xuất, sao chép về một nơi duy nhất.

Cũng tương tự như kỹ thuật Data Virtualization được đề cập ở trên thì Data Federation cũng tạo ra các cơ sở dữ liệu ảo nhằm cho phép người dùng tương tác và xử lý dữ liệu trực tiếp mà không cần thông qua quy trình ETL truyền thống. Tuy nhiên điểm khác biệt chính là Data Federation không tạo ra một lớp ảo hóa chuyên biệt mà chỉ tập trung vào việc truy vấn dữ liệu qua một giao diện. Đây là cấp độ tích hợp căn bản so với Data Virtualization vì nó không có các lớp logic dữ liệu hay tối ưu hóa như công nghệ Ảo hóa. Kỹ thuật này đóng vai trò nền tảng cho kỹ thuật Ảo hóa dữ liệu.

Dưới góc độ ngành quản trị dữ liệu lớn, công nghệ Ảo hóa dữ liệu là một phát triển mang tính đột phá. Thay vì mô hình mô hình tích hợp ETL truyền thống thực hiện trích xuất, sao chép và lưu trữ dữ liệu thì công nghệ này cho phép người dùng tương tác một cách trực quan tới các nhóm dữ liệu trên một lớp ảo tại thời gian thực, giúp tiết kiệm rất nhiều thời gian và chi phí cho hạ tầng lưu trữ. Tuy nhiên do công nghệ này

còn mới nên các nghiên cứu cũng như tài liệu hỗ trợ cũng như cộng đồng chưa thực sự phát triển. Một số công cụ Ảo hóa dữ liệu đáng chú hiện nay gồm có: Microsft Hyper-V do Microsoft phát triển và duy trì, tích hợp sâu với hệ sinh thái Microsoft và Windows; Xen – Nền tảng ảo hóa mã nguồn mở do Linux Foundation phát triển, được sử dụng rộng rãi với tính linh hoạt cao; Vmware vSphere do Vmware cung cấp dịch vụ và triển khai, có tính bảo mật cao và được sử dụng tại các doanh nghiệp lớn.

2.3. Khảo sát công nghệ để giải quyết bài toán Tích hợp dữ liệu

Nhiệm vụ chính của luận văn này là giải quyết bài toán "Tích hợp dữ liệu cho hệ thống quản lý cán bộ, viên chức của Viện Năng lượng nguyên tử". Các kỹ thuật và công cụ được nghiên cứu, so sánh để đưa ra lựa chọn đều dựa trên nhu cầu và thực tế tại đơn vị áp dụng. Kỹ thuật và công cụ được chọn phải giải quyết được vấn đề đã nêu ra về xung đột dữ liệu trong bối cảnh sáp nhập Bộ, Ngành. Kỹ thuật được được ra cần phải dễ dàng áp dụng, có khả năng tùy chỉnh dễ dàng sử dụng, dễ dàng tùy chỉnh theo nhu cầu và mục đích người dùng và tối ưu hóa về chi phí đầu tư.

2.3.1. Kỹ thuật Tích hợp lựa chọn

Dựa trên các yếu tố đề ra và từ nhu cầu thực tế, nghiên cứu nhận định áp dụng kỹ thuật tích hợp dữ liệu theo quy trình ETL là hợp lý nhất với bối cảnh thực trạng tại Viện Năng lượng nguyên tử Việt Nam. Toàn bộ các kỹ thuật tích hợp dữ liệu phổ biến được khảo sát ở mục trước đều có ưu điểm và nhược điểm đặc thù. Tổng quan về góc độ kỹ thuật thì các công nghệ này khác nhau chủ yếu ở giai đoạn thực hiện chuyển đổi dữ liệu và cách tương tác với các dữ liệu nguồn.

Đặc điểm của kỹ thuật ETL là trích xuất dữ liệu đến vùng trung gian, thực hiện các thuật toán logic về chuyển đổi dữ liệu tại đây, sau đó tải về nguồn dữ liệu đích. Đặc tính của ETL giúp cho dữ liệu được đảm bảo về chất lượng, phù hợp với các data warehouse truyền thông, giúp giảm tải cho hệ thống đích vì dữ liệu được làm sạch kỹ lưỡng trước khi tải lên. Tuy nhiên việc tích hợp qua ETL thường chỉ được thực hiện định kỳ do quy trình sẽ lặp lại từ đầu với mỗi yêu cầu tích hợp, điều này khiến tải

trọng nặng lên vùng trung gian với khối lượng dữ liệu lớn. Chính vì những đặc tính đó nên ETL rất phù hợp cho việc giải quyết bài toán tích hợp hệ thống quản lý nhân sự nói chung do quy trình chuyển đổi không cần linh hoạt mà chỉ cần áp dụng theo chuẩn với dữ liệu nguồn.

Khác với ETL thì ELT di chuyển dữ liệu đến đích trước khi chuyển đổi. Đặc thù trong quy trình này khiến cho ELT linh hoạt hơn trong môi trường dữ liệu lớn. Tuy nhiên, chính điều này khiến cho tài nguyên hệ thống đích chịu áp lực nặng nề và chất lượng dữ liệu ban đầu không được đảm bảo do không được qua quá trình làm sạch trước khi tải. Những đặc điểm này phù hợp cho mục đích tích hợp tại các dữ liệu lớn, nhu cầu chuyển đổi dữ liệu linh hoạt.

Song song với ELT và ETL thì kỹ thuật tích hợp dựa trên lớp ảo hóa cũng được cân nhắc. Tuy nhiên, kinh phí phát sinh cho các dự án này thường cao hơn các quy trình ETL truyền thống do khả năng truy xuất và chuyển đổi dữ liệu ở thời gian thực. Tuy nhiên đặc tính chính của việc ảo hóa này không phù hợp đối với bài toán tích hợp dữ liệu trong quản lý cán bộ tại Viện Năng lượng nguyên tử đã đề ra.

2.3.2. Công cụ Tích hợp lựa chọn

Quy trình ETL được chọn là định hướng kỹ thuật để giải quyết bài toán tích hợp dữ liệu trong quản lý cán bộ tại Viện Năng lượng nguyên tử. Trong số các công cụ khảo sát tại mục trên, việc sử dụng thư viện pandas để trích xuất và thao tác dữ liệu từ nguồn nổi bật lên như một giải pháp tối ưu nhất trong bối cảnh cụ thể tại Viện.

Pandas là công cụ thư viện hoàn toàn miễn phí cho ngôn ngữ python được nhà phát triển Wes McKinney xây dựng từ năm 2008. Thư viện này sử dụng một cấu trúc dữ liệu riêng là Dataframe để phục vụ các công tác chuyển đổi dữ liệu. Chính vì cấu trúc này nó đem lại cho người dùng nhiều chức năng xử lý tùy theo nhu cầu và bối cảnh thực tế rất linh hoạt mà ít công cụ đem lại được.

Pandas được phân phối hoàn toàn miễn phí, linh hoạt, dễ dàng sử dụng và có sự hỗ trợ cộng đồng lớn, không bị phụ thuộc vào công ty hay tập đoàn nào.

CHƯƠNG 3: XÂY DỰNG CÔNG CỤ PHẦN MỀM VÀ THỬ NGHIỆM ÁP DỤNG, ĐÁNH GIÁ

Như đã được học viên trình bày ở trên, mục tiêu chính của nghiên cứu này là đưa ra một mô hình tích hợp dữ liệu nhằm giải quyết tình trạng chồng chéo cơ sở dữ liệu về cán bộ, viên chức và người lao động tại Viện Năng lượng nguyên tử Việt Nam sau sáp nhập Bộ, Ban, Ngành. Nghiên cứu này sẽ giúp người dùng có thể trích xuất, chuyển đổi dữ liệu liên quan đến cán bộ đang công tác và làm việc tại Viện từ nhiều nguồn cơ sở dữ liệu cũ và mới để gộp lại thành một nguồn dữ liệu duy nhất, tải lên hệ thống phần mềm quản lý cán bộ, công chức, viên chức và người lao động của Bộ KH&CN, tiến tới đồng bộ tới hệ thống cán bộ quốc gia của Bộ Nội Vụ.

Căn cứ vào hiện trạng và nhu cầu cho việc giải bài toán Tích hợp dữ liệu cán bộ, viên chức và người lao động tại Viện Năng lượng nguyên tử Việt Nam, học viên tiến hành nghiên cứu và xây dựng mô hình tích hợp dữ liệu theo quy trình ETL tại chính trụ sở Viện. Mô hình này được xây dựng và phát triển để phục vụ cho công tác quản lý cán bộ, viên chức và người lao động; giải quyết được các vấn đề và nhu cầu phát sinh do ảnh hưởng của việc sáp nhập Cơ quan, Bộ, Ban, Ngành. Qua quá trình tích hợp theo quy trình ETL, hệ thống này sẽ đạt được những kết quả sau:

- Trích xuất được dữ liệu từ nhiều nguồn;
- Trích xuất được dữ liệu cụ thể từ các nguồn trên;
- Trích xuất dữ liệu được từ nhiều hệ thống lưu trữ dữ liệu cùng lúc;
- Làm sạch dữ liệu tại các trường thông tin;
- Loại bỏ các trường, cột dữ liệu trùng lặp;
- Chuẩn hoá yêu cầu về trường dữ liệu;
- Chuẩn hoá định dạng, lỗi chính tả viết hoa về các thông tin định danh;
- Chuẩn hoá thông tin về mã số định danh (yêu cầu mã số định danh điện tử theo quy định mới, chuẩn 12 số);

- Thay thế thông tin hòm thư điện tử cá nhân (gmail, yahoo, hotmail) và hòm thư điện tử thuộc Bộ KH&CN cũ (most.gov) thành tài khoản điện tử mới do Bộ KH&CN mới cung cấp và quản lý (mst.gov.vn);

- Tải và xuất dữ liệu sau khi được trích xuất, tổng hợp, chuyển đổi thành một nguồn dữ liệu duy nhất;

- Tuỳ chỉnh môi trường, định dạng, hệ thống đích cho dữ liệu đầu ra, phục vụ công tác đồng bộ với kho dữ liệu của hệ thống phần mềm quản lý cán bộ, công chức, viên chức và người lao động của Bộ KH&CN.

3.1. Xây dựng và cấu hình công cụ phần mềm

Qua khảo sát các kỹ thuật tích hợp dữ liệu hiện hành tại chương 2, để phục vụ cho công tác thực hiện những nhiệm vụ được đề ra ở trên, học viên tiến hành nghiên cứu và xây dựng quy trình ETL với sự hỗ trợ của thư viện Pandas trong môi trường Python. Công cụ sử dụng để thực hiện là Pycharm 251.25410.159; Python phiên bản 3.13 và thư viện Pandas qua cú pháp cài đặt Pip [pip install --pre --extra-index https://pypi.anaconda.org/scientific-python-nightly-wheels/simple pandas].

Hệ thống ETL được đề xuất có cấu hình như sau:

```
import os
import pandas as pd

# Cấu hình hệ thống
INPUT_FOLDER = r"E:\Pycharm\python1_DataIntergration\targets"  # Đường dẫn địa chỉ thư mục chứa nguồn dữ liệu
OUTPUT_FILE = r"E:\Pycharm\python1_DataIntergration\output\result.xlsx"  # Đường dẫn địa chỉ thư mục chứa dữ liệu xuất
```

Hình 3.1 Cấu hình hệ thống

Hệ thống Tích hợp dữ liệu được đề xuất sử dụng hai thư viện os và pandas là nền móng để xây dựng và phát triển. Hàm reduce trong pandas được sử dụng để tương tác kết hợp dữ liệu, tổng hợp dữ liệu trong nhiều DataFrame.

Cấu hình input_folder và output_file nhằm chỉ địa điểm thư mục tổng hợp dữ liệu nguồn được sử dụng và địa chỉ thư mục tải dữ liệu đầu ra của hệ thống. Người dùng hệ thống hoàn toàn có thể thay đổi đường dẫn các thư mục tuỳ thuộc vào mục đích sử dụng.

Để phục vụ nhiệm vụ trích xuất dữ liệu từ nhiều nguồn đầu vào tại giai đoạn đầu tiên của ETL, chương trình sẽ thực hiện theo quy trình được mô tả tại Hình 3.2 (Theo như cấu hình được đề xuất, việc trích xuất sẽ được tiến hành trên các cơ sở dữ liệu có mặt tại thư mục input_folder được cài đặt theo tuỳ chỉnh của người dùng):

```
#Định nghĩa logic tích hợp trên dataframe
def combine_files_on_cccd_prioritize_email(input_folder, output_file):  1 usage
    all_dataframes = []

    # Lặp lại với toàn bộ file trong thư mục
    for file in os.listdir(input_folder):
        file_path = os.path.join(input_folder, file)

        if file.endswith(".csv"):
            df = pd.read_csv(file_path, dtype=str)
            all_dataframes.append(df)

        elif file.endswith(".xlsx"):
            xls = pd.ExcelFile(file_path)
            for sheet in xls.sheet_names:
                df = pd.read_excel(file_path, sheet_name=sheet, dtype=str)
                all_dataframes.append(df)
```

Hình 3.2 Trích xuất dữ liệu từ các file lưu trữ dữ liệu

Các dữ liệu được trích xuất từ các file dữ liệu sẽ được gán vào các dataframes với mỗi file dữ liệu hệ thống sẽ gắn vào một dataframe để phục vụ các nhiệm vụ chuyển đổi dữ liệu phía sau.

Trong trường hợp cụ thể này (nhằm phục vụ công tác tích hợp hệ thống quản lý cán bộ, công chức, viên chức và người lao động tại Bộ KH&CN), hệ thống được cấu hình để chỉ sử dụng dữ liệu có định dạng .xlsx và .csv. (Trong các trường hợp cụ thể khác, người dùng hoàn toàn có thể thiết lập các tuỳ chỉnh để có thể thu thập thông tin từ các kiểu dữ liệu và các nguồn dữ liệu khác mà pandas có hỗ trợ) Sau khi hệ thống này xác định được các nguồn dữ liệu đúng định dạng được thiết lập trước đó, hệ thống sẽ thử đọc dữ liệu tại các nguồn này xem có thể sử dụng được không. Trong trường hợp dữ liệu đọc được bởi pandas, hệ thống sẽ trích xuất dữ liệu thô và append dữ liệu từ file nguồn vào tất cả các dataframes.

Để có thể xử lý được bài toán đề ra, nghiên cứu cần thiết phải làm rõ hơn về nguyên nhân phát sinh bài toán 'tích hợp dữ liệu cho hệ thống quản lý cán bộ, viên chức của Viện Năng lượng nguyên tử". Học viên đã tiến hành khảo sát cụ thể về quy trình quản lý và khai thác hồ sơ cán bộ của Vụ Tổ chức cán bộ, Bộ KH&CN sau sáp nhập.

Quy trình đồng bộ sau sáp nhập giữa các hệ thống dữ liệu tại Bộ KH&CN trước và sau sáp nhập được áp dụng và triển khai đối với các đơn vị cơ sở trực thuộc nói chung và tại Viện Năng lượng nguyên tử nói riêng được thực hiện theo quy trình: Dữ liệu cán bộ được trích xuất từ các nguồn thông tin tại nhà kho dữ liệu là hệ thống quản lý quản lý cán bộ của Bộ KH&CN (2024) dưới dạng dữ liệu thô và sau đó được tải lại vào hệ thống phần mềm mới của Bộ KH&CN (2025) mà không được chuyển đổi hay làm sạch. Quy trình này được thực hiện một cách nhanh chóng trong bối cảnh sắp xếp và tinh gọn bộ máy hành chính nhà nước. Chính vì bỏ qua giai đoạn chuyển đổi dữ liệu mà dữ liệu không được làm sạch cũng như loại bỏ trùng lặp dẫn đến nhiều lỗi. Hình 3.3 là ví dụ về một số dữ liệu về cán bộ, viên chức và người lao động tại Viện Năng lượng nguyên tử Việt Nam được học viên trích xuất ra từ hệ thống mới tại thời điểm hiện tại.

A	B	C	D
9	Nguyễn Thị Thu Hà	034179010151	hantt_itrre@mst.gov.vn
10	Nguyễn Thị Thu Hà	001178038310	hantt_inst@mst.gov.vn
11	Nguyễn Thị Thu Hà	001176047943	nguyenthithuha55@mst.gov.vn
12	Nguyễn Tiến Phong	024078000210	phongnt_nde@mst.gov.vn
13	NGUYỄN TIẾN PHONG	001087000895	qlkh.nde@gmail.com
14	Nguyễn Trung Kiên	035083011272	kiennt_nde@mst.gov.vn
15	Nguyễn Trung Kiên	001089029911	trungkien211013@gmail.com
16	Phạm Bảo Ngọc	068194000848	ngocpb_nri@mst.gov.vn
17	Phạm Bảo Ngọc	001189019141	ngocpb_inst@mst.gov.vn
18	PHẠM THỊ THU TRANG	036189001823	trangptt@mst.gov.vn
19	PHẠM THỊ THU TRANG	NULL	cherrypham.hanoi@gmail.com
20	Phạm Văn Dũng	001067017076	dungpv_nri@mst.gov.vn
21	PHẠM VĂN DŨNG	001067067076	phamdungatbx@gmail.com
22	VŨ ĐỨC MẠNH	034096017423	vuducmanh05021996@gmail.com
23	VŨ ĐỨC MẠNH	034096017424	manhvd_inst@mst.gov.vn
24	VŨ ĐỨC MẠNH	34096017424	manhvd_inst@mst.gov.vn

Hình 3.3 Dữ liệu về cán bộ hiện tại được trích xuất ra từ hệ thống mới 2025

Bảng thống kê được thể hiện tại Hình 3.3 là một minh chứng cụ thể cho sự cấp thiết của đề tài "Tích hợp dữ liệu cho hệ thống quản lý cán bộ, viên chức của

Viện Năng lượng nguyên tử". Có thể thấy rõ nhiều cán bộ bị trùng lặp dữ liệu dẫn đến việc khi tạo ra các báo cáo thống kê nhân sự, số liệu và thông tin bị sai lệch, không có giá trị sử dụng.

Trước khi thực hiện các tác vụ khác có liên quan, giá trị "CCCD" được định nghĩa như sau

```
# Định nghĩa chuẩn hoá CCCD
def normalize_cccd(cccd_value):  1 usage
    """
    Chuẩn hoá số định danh cá nhân điện tử (CCCD) 12 số được hiểu như sau:
    - Nếu như dữ liệu chứa ít hơn 12 ký tự, thêm ký tự "0" ở đầu dãy cho đến khi đủ 12;
    - Nếu dữ liệu đã đủ 12 ký tự, giữ nguyên.
    """
    try:
        cccd_str = str(cccd_value).strip()
        if cccd_str.isdigit():
            if len(cccd_str) < 12:
                return cccd_str.zfill(12)
            else:
                return cccd_str
        else:
            return cccd_str
    except Exception:
        return str(cccd_value)
```

Hình 3.4 Định nghĩa chuẩn hoá CCCD

```
# Chia nhóm dữ liệu dựa trên CCCD, đặt CCCD làm key data
merged_df = combined_df.groupby( by: "CCCD", as_index=False).apply(merge_rows)
```

Hình 3.5 Định nghĩa tích hợp theo các dữ liệu số CCCD

Giá trị tham chiếu xuyên suốt giữa các hệ thống lưu trữ dữ liệu cơ sở của cán bộ, viên chức và người lao động là số định danh cá nhân (căn cước điện tử 12 số) do bộ Công An cấp. Do việc phát sinh nhiều nguồn dữ liệu liên quan đến công tác quản trị con người nên việc dữ liệu trùng lặp giữa các database là không tránh khỏi. Ví dụ ở nguồn dữ liệu nào thì số định danh cá nhân (CCCD) cũng đi kèm với họ tên cán bộ, ngày tháng năm sinh, đơn vị công tác, tuy nhiên ở các hệ dữ liệu khác nhau sẽ có những thông tin khác nhau liên quan đến cán bộ đó như thông tin lương, ngạch, hay quá trình

đào tạo, bồi dưỡng, quy hoạch, bổ nhiệm. Dữ liệu liên quan đến các thông tin này của cán bộ sau khi được trích xuất và làm sạch sẽ được xử lý qua quá trình khử trùng lặp dữ liệu. Hệ thống được đề xuất tiến hành khử trùng lặp bằng cách đối chiếu các dữ liệu giữa các cột và sẽ tiến hành giữ lại giá trị non-empty cuối cùng nhận được.

Tại hệ thống ETL được đề xuất để ứng dụng trong việc quản lý cán bộ, viên chức và người lao động tại Viện năng lượng nguyên tử Việt Nam, các dữ liệu sau khi được trích xuất từ các dữ liệu nguồn được qua các chu trình xử lý sau:

```
# Kết nối liên kết các dataframe
combined_df = pd.concat(all_dataframes, ignore_index=True)

# Chuẩn hoá CCCD
if "CCCD" in combined_df.columns:
    combined_df["CCCD"] = combined_df["CCCD"].apply(normalize_cccd)

# Chuẩn hoá Họ Tên
if "Họ Tên" in combined_df.columns:
    combined_df["Họ Tên"] = combined_df["Họ Tên"].astype(str).str.title().str.strip()

# Loại bỏ dữ liệu trung lặp
combined_df.drop_duplicates(inplace=True)
```

Hình 3.6 Quy trình Chuyển đổi dữ liệu, làm sạch dữ liệu và chuẩn hoá về định dạng.

Một vấn đáng chú ý của việc dùng trực tiếp dữ liệu thô từ hệ thống để đồng bộ sang là các giá trị "Họ và Tên" cán bộ không được chuẩn hóa về mặt định dạng. Các giá trị Họ và Tên tại Hệ thống Quản lý nhân sự của Bộ KH&CN trước sáp nhập được lưu dưới định dạng toàn bộ ký tự đều in hoa trong khi đó, các giá trị này tại hệ thống mới chỉ được viết hoa chữ cái đầu; một số khác do lỗi người dùng khi khai báo nhập thông tin địa chỉ và họ tên đều ở dạng chữ cái thông thường. Trong giai đoạn Transform, Hệ thống được đề xuất có nhiệm vụ chuẩn hoá các dữ liệu Họ và Tên cùng địa chỉ được viết hoa chữ cái đầu theo tiêu chuẩn.

Giải pháp để loại bỏ trùng lặp dữ liệu dựa trên số CCCD điện tử cũng như chuẩn hóa và làm sạch dữ liệu về số CCCD đã được đề xuất trước đó. Hệ thống tích

hợp dữ liệu được đề xuất sẽ tiến hành gộp các hồ sơ cán bộ có cùng số CCCD thành một bản lưu trữ.

Vấn đề phát sinh trong trường hợp này là một cá nhân sẽ có nhiều hòm thư điện tử (Hòm thư điện tử cá nhân dùng để đăng nhập tài khoản tại hệ thống quản lý nhân sự tại Bộ KH&CN trước sáp nhập). Một cách tiếp cận là tự động loại bỏ tất cả những hòm thư điện tử cá nhân không được Bộ KH&CN (sau sáp nhập) cấp như @gmail; @yahoo; @hotmail; @mst.gov; @most.gov… và chỉ giữ lại các hòm thư do Bộ (sau sáp nhập) cung cấp có địa chỉ tại @mst.gov.vn. Tuy nhiên, hướng giải quyết này là không khả thi do nhiều cán bộ thuộc diện chuyên gia độc lập được ký hợp đồng lao động; hợp đồng chuyên môn hay các cán bộ thuộc diện lao động phổ thông như lái xe; tạp vụ; bảo vệ… đều không được cung cấp tài khoản thư điện tử @mst.gov.vn. Để giải quyết vấn đề này, nhiệm vụ của hệ thống tích hợp được đề xuất để giải quyết các vấn đề dữ liệu liên quan đến địa chỉ hòm thư điện tử được định nghĩa như sau:

```
# Hàm kết hợp giá trị email
def merge_rows(group):
    # Ưu tiên sử dụng mst.gov.vn email
    mst_rows = group[group['Email'].str.endswith('@mst.gov.vn', na=False)]
    if not mst_rows.empty:
        base = mst_rows.iloc[0].copy()
    else:
        base = group.iloc[0].copy()

    # Sử dụng giá trị non-empty đầu tiên trong tập hợp dữ liệu
    for col in group.columns:
        if pd.isna(base[col]) or base[col] == "":
            non_empty = group[col].dropna()
            if not non_empty.empty:
                base[col] = non_empty.iloc[0]
    return base
```

Hình 3.7 Quy trình xử lý loại bỏ email cán bộ bị trùng

Vẫn trong giai đoạn Chuyển đổi dữ liệu của quy trình ETL được đề xuất, hệ thống sẽ tiến hành ưu tiên hòm thư điện tử @mst.gov.vn nếu phát sinh trường hợp một cán bộ có nhiều địa chỉ thư điện điện tử. Trong trường hợp cán bộ không có tài khoản @mst.gov.vn như ví dụ này, hệ thống sẽ sử dụng giá trị non-empty mới nhất. Sau khi đã xử lý xong tình trạng một cá nhân tồn tại nhiều địa chỉ thư điện tử, hệ thống sẽ tiến thành gộp các bản ghi nhân sự theo số CCCD định danh điện tử 12 số và gán địa chỉ email tương ứng.

Hệ thống sau khi hoàn thành việc trích xuất dữ liệu và chuyển đổi dữ liệu sẽ tiến hành tải dữ liệu vào tệp tin lưu trữ dữ liệu đầu ra theo đường dẫn thư mục output_file được cấu hình theo tuỳ chọn ở mục trước.

```
        # Tải dữ liệu vào file .xlsx mong muốn
        merged_df.to_excel(output_file, index=False, engine="openpyxl")

▷   if __name__ == "__main__":
        combine_files_on_cccd_prioritize_email(INPUT_FOLDER, OUTPUT_FILE)
        print(f"Combined file saved as {OUTPUT_FILE}")
```

Hình 3.8 Chức năng tải dữ liệu của hệ thống

Hệ thống ETL được đề xuất có nhiệm vụ chính là giải quyết bài toán "tích hợp dữ liệu cho hệ thống quản lý cán bộ, viên chức của Viện Năng lượng nguyên tử" nên dữ liệu sau khi được tích hợp sẽ được tải lên nguồn dữ liệu mới dưới định dạng .xlsx.

Định dạng này được lựa chọn làm SSO (Single source of Truth) cuối để tải dữ liệu do hai yếu tố chính phát sinh từ nhu cầu thực tế. Đầu tiên, cán bộ quản trị có thể dễ dàng tương tác với thông tin dữ liệu được tải lên sau khi tích hợp, giúp người dùng có thể sử dụng luôn những dữ liệu này trong công tác quản trị con người, phục vụ các công tác chuyên môn cũng như báo cáo. Người dùng hoàn toàn có thể kiểm tra lại thông tin đầu ra một lần cuối trước khi được sử dụng làm nguồn dữ liệu cho việc đồng bộ. Lý do thứ hai mà hệ thống lưu trữ dữ liệu này được lựa chọn vì theo đề xuất của cán bộ đầu mối chuyên trách đảm nhận nhiệm vụ quản trị dữ liệu và đồng bộ hệ thống

sau sắp xếp khuyến khích lưu trữ dữ liệu có định dạng .xlsx do nhà kho dữ liệu trước đó có chức năng xuất dữ liệu dưới định dạng này và hệ thống mới cũng đã có khả năng hỗ trợ trích xuất dữ liệu từ các tệp dữ liệu định dạng này.

Tuy nhiên, để hệ thống tích hợp dữ liệu được đề xuất hoàn thiện hơn, học viên muốn trình bày và đưa vào đề xuất khả năng convert các dạng file lưu trữ dữ liệu khác nhau với sự hỗ trợ của thư viện pandas trong môi trường python. Do thư viện này hỗ trợ ghi dữ liệu trên hầu hết các hệ thống quản lý dữ liệu phổ biến như MySQL, SQLite, SQL Server, Oracle,… Nên nó cũng có khả năng hỗ trợ nhiều định dạng hệ thống dữ liệu khác nhau. Hình 3.9 mô tả về quy trình chuyển đổi dữ liệu sang định dạng .csv tùy theo nhu cầu sử dụng của người dùng.

```
combineExcel.py    convertExcelToCSV.py
import pandas as pd
import os

# Đường dẫn file Excel
INPUT_FILE = r"E:\Pycharm\python1_DataIntergration\output\result.xlsx"
OUTPUT_FILE = r"E:\Pycharm\python1_DataIntergration\output\result.csv"  # Tự động đặt tên nếu không có đường dẫn
SHEET_NAME = 0  # Có thể là tên sheet hoặc số thứ tự

def excel_to_csv(input_file, output_file=None, sheet_name=0):
    # Tự động đặt tên file CSV nếu không được cung cấp
    if not output_file:
        base = os.path.splitext(input_file)[0]
        output_file = base + '.csv'

    # Đọc file Excel
    xl = pd.ExcelFile(input_file)

    # Trích xuất dữ liệu từ sheet
    df = xl.parse(sheet_name)

    # Ghi ra file CSV
    df.to_csv(output_file, index=False, encoding="utf-8")

    print(f"Converted '{input_file}' → '{output_file}' successfully.")

# Gọi chuyển đổi
excel_to_csv(INPUT_FILE, OUTPUT_FILE, SHEET_NAME)
```

Hình 3.9 Chức năng chuyển định dạng dữ liệu sang .csv

Hệ thống được đề xuất cũng có khả năng chuyển đổi dữ liệu sau khi tích hợp qua quy trình trước đó sang định dạng .csv theo nhu cầu người dùng. Đầu tiên, người dùng sẽ tiến hành cấu hình đường dẫn cho thư mục đầu ra và đầu vào dữ liệu tương tự với quy trình ETL được đề xuất trước đó. Hàm sheet_name = 0 cho phép người

dùng tuỳ chỉnh sheet dữ liệu trên hệ thống lưu trữ với định dạng .xlsx. Trong trường hợp sử dụng giá trị mặc định, hệ thống sẽ chuyển đổi sheet đầu tiên được đọc tại tệp dữ liệu nguồn.

3.2. Kết quả thử nghiệm

Để phục vụ công tác nghiên cứu xây dựng thử nghiệm vận hành hệ thống tích hợp dữ liệu được đề xuất, học viên tiến hành khởi tạo 3 tệp chứa dữ liệu về cán bộ, người lao động dựa trên các yếu tố và hiện trạng tại các bản ghi hồ sơ nhân sự có sẵn tại hệ thống quản lý cán bộ, viên chức, công chức và người lao động của Bộ KH&CN (sau sáp nhập) nhằm mô phỏng lại hiện trạng thực tế về bài toán tích hợp dữ liệu tại Viện Năng lượng nguyên tử.

Tệp dữ liệu thứ 1 được khởi tạo là tệp dữ liệu về thông tin cán bộ tại Viện được trích xuất ra từ Hệ thống quản lý CCVCNLĐ của Bộ KH&CN (trước sáp nhập). Tệp dữ liệu thứ 2 được khởi tạo là tệp dữ liệu được khởi tạo bởi Bộ KH&CN (mới) nhằm phục công tác đồng bộ. Tệp dữ liệu thứ 3 là tệp dữ liệu cán bộ được yêu cầu bổ sung thông tin, gồm các trường thông tin bổ sung mà tại hệ thống quản lý CCVCCNLĐ của Bộ KH&CN (cũ) nhưng cần thiết phải cập nhật tại hệ thống phần mềm quản lý CCVCNLĐ tại Bộ&CN (mới) nhằm phục vụ các công tác cán bộ.

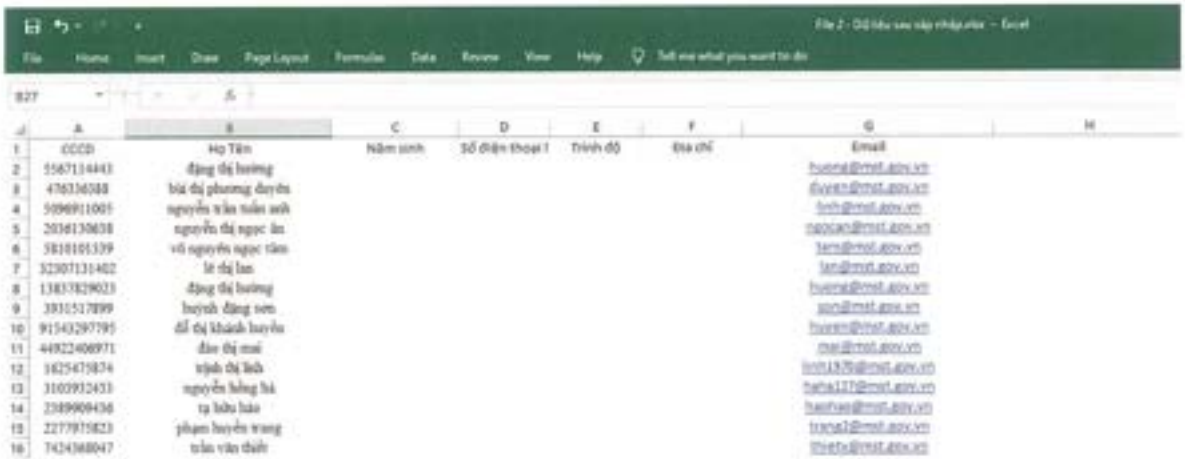
Tệp dữ liệu thứ 1 được cấu hình với các đặc điểm sau để mô phỏng dữ liệu được trích xuất từ hệ thống quản lý nhân sự được sử dụng trước khi sáp nhập bộ ngành của Bộ KH&CN (năm 2024): số CCCD bị mất số 0, Họ Tên viết in Hoa, Email cá nhân tại các địa chỉ rời rạc, một số sử dụng Email được Bộ KH&CN (trước sáp nhập) cung cấp tại địa chỉ @most.gov.vn và một số trường dữ liệu chứa thông tin liên quan.



CCCD	Họ Tên	Năm sinh	Số điện thoại liên hệ	Trình độ	Địa chỉ	Email
5567114441	ĐẶNG THỊ HƯỜNG	1971	091 870 03 31	Tiến sỹ	Phường Điện Biên, Thành phố Thanh Hóa, Thanh Hóa	huong@mst.gov.vn
476036088	BÙI THỊ PHƯƠNG DUYÊN	1978	0304 613 019	Cử nhân	286 Đường Nguyễn Tất Lực, Phường 8, Đà Lạt, Lâm Đồng	duyen78@mst.gov.vn
5096911001	NGUYỄN TRẦN TUẤN ANH	1990	0825 855 173	Thạc sỹ	Phường Mỹ An, Quận Ngũ Hành Sơn, Đà Nẵng	trtuananh117@gmail.com
2036130658	NGUYỄN THỊ NGỌC ÂN	1999	096 952 25 93	Kỹ sư	157 Minh Phụng, Phường 9, Quận 11, Thành phố Hồ Chí Minh	anh88@yahoo.com
5810101339	VÕ NGUYỄN NGỌC TÂM	1994	0360 875 468	Cử nhân	9a Vạn Hạnh, Phường 10, Quận 10, Thành phố Hồ Chí Minh	tamngan@mst.gov.vn
32307131402	LÊ THỊ LAN	1983	0829 867 002	Cử nhân	Cộng Hòa, Phường 4, Tân Bình, Thành phố Hồ Chí Minh	lhlan8888@gmail.com
13837829023	ĐẶNG THỊ HƯỜNG	1980	0810 001 687	Cử nhân	115 Đông Hội, Đông Anh, Hà Nội	huongdang@gmail.com
3931517899	HUỲNH ĐẶNG SƠN	1989	0369 339 099	Thạc sỹ	Phường Mỹ An, Quận Ngũ Hành Sơn, Đà Nẵng	sonhh@gmail.com
91543297795	ĐỖ THỊ KHÁNH HUYỀN	1996	0362 674 940	Cử nhân	211 P. Thanh Trang, Triều Quí, Gia Lâm, Hà Nội	lamer@yahoo.com
44922400971	ĐÀO THỊ MAI	1992	0869 685 904	Cử nhân	435 Đ. Cổ Linh, Long Biên, Hà Nội	maidao@gmail.com
1625475874	TRỊNH THỊ LỊCH	1978	0878 874 581	Tiến sỹ	809 Phạm Hùng, Mễ Trì, Nam Từ Liêm, Hà Nội	lichtt78@gmail.com
3103932433	NGUYỄN HỒNG HÀ	1978	0848 501 017	Tiến sỹ	111 P. Nghĩa Dũng, Phúc xá, Ba Đình, Hà Nội	haha127@gmail.com
2389909436	TA HỮU HẢO	1981	091 706 33 33	Cử nhân	3 Tháng 4, Phường 3, Đà Lạt, Lâm Đồng	haohau@mst.gov.vn
2277975823	PHẠM HUYỀN TRANG	2000	091 336 14 88	Cử nhân	114 Đường Lý Tự Trọng, Phường 1, Đà Lạt, Lâm Đồng	trang2k@gmail.com
7424368047	TRẦN VĂN THIẾT	1994	091 325 50 19	Thạc sỹ	67 Đường Nguyễn Tất Lực, Phường 8, Đà Lạt, Lâm Đồng	thuanvietlam@gmail.com

Hình 3.10 File dữ liệu cán bộ thứ 1

Tệp dữ liệu thứ 2 được cấu hình với các đặc điểm sau để mô phỏng dữ liệu được trích xuất từ hệ thống quản lý nhân sự được xây dựng sau khi sáp nhập bộ ngành của Bộ KH&CN (năm 2025): Họ và Tên sử dụng định dạng khác, Email được cung cấp và khởi tạo bởi Bộ KH&CN sau sắp xếp bộ máy nhà nước thống nhất là @mst.gov.vn. Các trường thông tin khác trắng vì hệ thống mới được xây dựng và khởi tạo năm 2025, mới chỉ có thông tin căn bản về sổ định danh điện tử và hòm thư được cấp cho cán bộ, viên chức thuộc Bộ KH&CN quản lý.

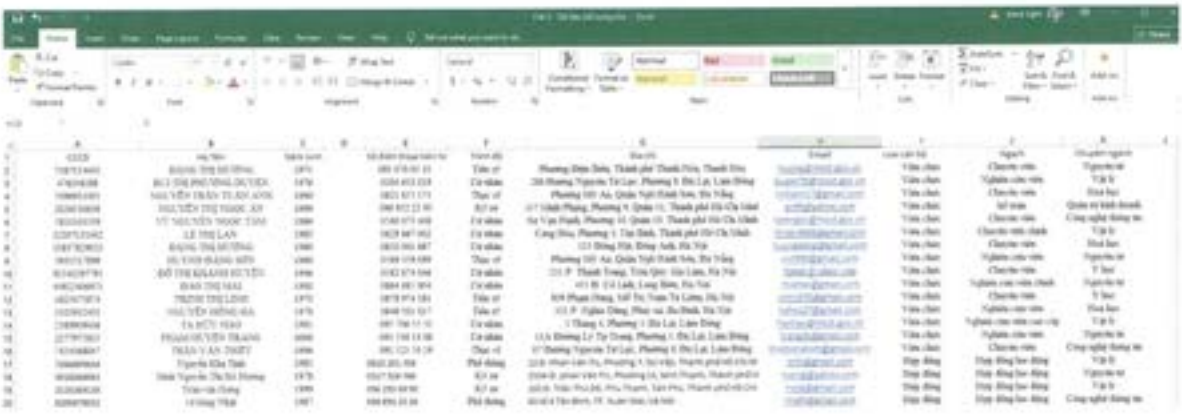


CCCD	Họ Tên	Năm sinh	Số điện thoại	Trình độ	Địa chỉ	Email
5567114441	đặng thị hường					huong@mst.gov.vn
476036088	bùi thị phương duyên					duyen@mst.gov.vn
5096911001	nguyễn trần tuấn anh					tuvh@mst.gov.vn
2036130658	nguyễn thị ngọc ân					ngocan@mst.gov.vn
5810101339	võ nguyễn ngọc tâm					tam@mst.gov.vn
32307131402	lê thị lan					lan@mst.gov.vn
13837829023	đặng thị hường					huong@mst.gov.vn
3931517899	huỳnh đặng sơn					son@mst.gov.vn
91543297795	đỗ thị khánh huyền					huyen@mst.gov.vn
44922400971	đào thị mai					mai@mst.gov.vn
1625475874	trịnh thị lịch					lichtt78@mst.gov.vn
3103932433	nguyễn hồng hà					haha127@mst.gov.vn
2389909436	ta hữu hảo					haohau@mst.gov.vn
2277975823	phạm huyền trang					trang2@mst.gov.vn
7424368047	trần văn thiết					thiettv@mst.gov.vn

Hình 3.11 File dữ liệu cán bộ thứ 2

Tệp dữ liệu thứ 3 được cấu hình với các đặc điểm sau để mô phỏng dữ liệu được Bộ KH&CN yêu cầu cung cấp bổ sung sau khi rà soát và nhận thấy dữ liệu cán bộ tại cả 2 hệ thống đều chưa đủ các thông tin: bổ sung thêm cán bộ thuộc diện hợp

đồng lao động không được cấp tài khoản thư @mst.gov.vn; bổ sung các trường thông tin khác; lỗi chính tả về định dạng phát sinh trong quá trình khai báo.



Hình 3.12 File dữ liệu cán bộ thứ 3

Với 3 File dữ liệu cán bộ như giả định được đưa vào quy trình tích hợp dữ liệu, hệ thống được đề xuất cho người dùng kết quả là dữ liệu được ghi vào định dạng tuỳ chọn cụ thể như hình ảnh kết quả được thể hiện tại Hình 3.13.



Hình 3.13 Dữ liệu cán bộ được tích hợp từ 3 file dữ liệu có sẵn

Sau tích hợp, dữ liệu đã được tổng hợp ở cả 3 File và toàn bộ các dữ liệu trùng lặp đã bị loại bỏ. Nhân sự được tích hợp ở kết quả cuối bao gồm cả 15 nhân sự là viên chức thuộc biên chế nhà nước, có tài khoản @mst.gov.vn ban đầu tại File 1 và 2 cộng với 4 nhân sự thuộc diện ký hợp đồng lao động với đơn vị tại File 3 thông tin bổ sung. Tất cả các số CCCD được chuẩn hoá thành dạng 12 số, không bị mất số 0 ở đầu dãy. Họ và Tên cán bộ tại đơn vị sau tích hợp cũng được chuẩn hoá về cùng một định dạng

viết hoa các chữ cái đầu tiên. Cột Email đã được chuẩn hoá theo tình trạng thực tế, ưu tiên địa chỉ thư điện thử @mst.gov.vn với các cán bộ có tài khoản là viên chức thuộc biên chế nhà nước. Với những cán bộ còn lại thuộc diện lao động hợp đồng, sử dụng dữ liệu email cá nhân. Các trường thông tin được yêu cầu bổ sung tại File 3 cũng được tổng hợp vào dữ liệu nhân sự cuối.

Bài luận văn này không sử dụng hoàn toàn các dữ liệu nhân sự sẵn có tại Viện Năng lượng nguyên tử Việt Nam. Dữ liệu và thông tin chi tiết về cán bộ là thông tin nhân sự thuộc quyền sở hữu của nhà nước. Để đảm bảo tính bảo mật về thông tin cán bộ nên luận văn này chỉ sử dụng một phần để mô phỏng giả định lại thực trạng đã được chứng minh trước đó. Mọi trùng lặp về họ tên, địa chỉ, mã số định danh và các thông tin khác chỉ là sự trùng hợp.

3.3. Đánh giá kết quả đã đạt được của luận văn

3.3.1. Về ưu điểm và hạn chế

1) Ưu điểm

- Công cụ tích hợp dữ liệu về quản lý cán bộ, viên chức: đầy đủ các thông tin, dữ liệu kết hợp được từ nhiều nguồn khác nhau.
- Dữ liệu được quản lý tập trung, thống nhất, giảm thiểu tối đa tình trạng trùng lặp hoặc thất lạc hồ sơ. (Tỷ lệ trùng lặp hồ sơ sau khi áp dụng hệ thống tích hợp dữ liệu tại Viện Năng lượng nguyên tử giảm từ 76% xuống còn 5%)
- Dữ liệu về cán bộ tại đơn vị được chuẩn hóa và làm sạch.
- Phục vụ trực tiếp cho công tác quản lý nhân sự tại đơn vị

2) Hạn chế

- Một số nghiệp vụ đặc thù của đơn vị chưa được tùy biến sâu.
- Thiếu phân quyền truy cập để kiểm soát dữ liệu sau tích hợp.

3.3.2. Đánh giá về mặt định tính

1) Đánh giá về hiệu quả sử dụng

- Giảm đáng kể thời gian tổng hợp, thống kê và lập báo cáo về dữ liệu cán bộ tại đơn vị so với các hình thức thực hiện thủ công truyền thống.

- Nâng cao hiệu quả công tác quản lý cán bộ, hỗ trợ lãnh đạo trong việc theo dõi, ra quyết định.

- Tăng tính minh bạch, chính xác trong quản lý thông tin cán bộ, viên chức.

2) Đánh giá về kỹ thuật và công nghệ

- Giao diện thân thiện, dễ sử dụng, phù hợp với người dùng không chuyên về CNTT.

- Hệ thống hoạt động ổn định, tốc độ xử lý đáp ứng yêu cầu khai thác thường xuyên.

3) Đánh giá về khả năng mở rộng và tích hợp

- Phần mềm có khả năng mở rộng, nâng cấp trong tương lai khi thay đổi quy định hoặc quy mô quản lý.

- Có thể tích hợp với các hệ thống khác (quản lý văn bản, chấm công, tiền lương...) nếu được đầu tư bổ sung.

4) Đánh giá mức độ hài lòng của người dùng

- Đa số người sử dụng đánh giá phần mềm đáp ứng tốt nhu cầu công việc.

- Giao diện và chức năng dễ tiếp cận, thời gian làm quen ngắn.

3.3.3. Đánh giá về mặt định lượng

Trên cơ sở kết quả đã thực hiện của luận văn, mà cụ thể là công cụ phần mềm "Tích hợp dữ liệu cho hệ thống quản lý cán bộ, viên chức của Viện Năng lượng nguyên tử" đã được triển khai áp dụng thực tế tại Viện Năng lượng nguyên tử Việt Nam, có thể đánh giá định lượng dựa trên các tiêu chí sau:

Với những Kết quả thu được từ thử nghiệm tại mục 3.2 và các hình 3.10, 3.11, 3.12, 3.13 của quá trình tích hợp dữ liệu cán bộ, viên chức và người lao động tại Viện Năng lượng nguyên tử Việt Nam sau khi áp dụng hệ thống được đánh giá theo các quy định hiện hành về dữ liệu như sau:

- Tiêu chí về tính chính xác của dữ liệu (Dữ liệu nhập, xuất phải chính xác, không có sai lệch, thay đổi về nội dung): Kết quả đánh giá là Đạt.

- Tiêu chí về tính Toàn vẹn dữ liệu (Dữ liệu có ràng buộc, không bị trùng lặp hồ sơ): Kết quả đánh giá là Đạt.

- Tiêu chí về giao diện (Hệ thống có tính thân thiện, dễ thao tác): Kết quả đạt được là Chấp nhận được.

- Tiêu chí về Tốc độ làm quen (Người dùng sử dụng thành thạo một cách nhanh chóng): Kết quả đánh giá là Đạt.

- Tiêu chí về thời gian phản hồi (Đáp ứng và phản hồi nhanh chóng trong thời gian thao tác và sử dụng): Kết quả đánh giá là Đạt.

- Tiêu chí về tính ổn định (Lỗi xảy ra trong quá trình là ít hoặc không có, hệ thống không bị gián đoạn trong quá trình làm việc): Kết quả đánh giá là Đạt.

Đánh giá chung: Phần mềm đáp ứng đầy đủ các nghiệp vụ cốt lõi theo quy định hiện hành

Phần mềm đáp ứng yêu cầu quản lý cán bộ, viên chức trong giai đoạn hiện nay.

3.4. Định hướng phát triển

Về mặc logic, hệ thống được đề xuất đã vận hành ổn định tại môi trường giả định với dữ liệu cán bộ được mô phỏng lại và có tiềm năng ứng dụng thực tế để giải quyết bài toán về hệ thống quản lý nhân sự tại Viện. Trong trường hợp thực tế thì hệ thống cần phải hoàn thiện hơn về định hướng mở rộng. Tại môi trường giả lập, hệ thống được xác định một số chức năng chuyển đổi dữ liệu dựa trên các vấn đề hiện hữu tại cơ sở dữ liệu của Viện. Để có thể hoàn thiện hơn cần đưa hệ thống vào thử nghiệm thực tế để có thể có cái nhìn chi tiết hơn về các vấn đề cụ thể.

Về tính khả thi, các vấn đề được đưa ra để giải quyết chỉ là những vấn đề hiện hữu trong quá trình quản trị dữ liệu cán bộ tại Viện Năng lượng nguyên tử. Hệ thống cần được vận hành một thời gian với vai trò là cầu nối giữa các hệ thống dữ liệu của Viện và của Bộ để có thể hiệu chỉnh cụ thể hơn về các vấn đề chưa hiện hữu.

Về mặt người dùng, hệ thống cần có một giao diện cụ thể để người dùng có thể tương tác và hiệu chỉnh một cách dễ dàng hơn. Hiện tại các hiệu chỉnh vẫn được triển khai một cách thủ công bằng cách tác động trực tiếp vào code nguồn, điều này dễ gây ra sai sót khi thao tác xử lý khối công việc lớn hơn.

Để có thể tiếp tục hỗ trợ và ứng dụng cho việc giải quyết bài toán tích hợp dữ liệu cho hệ thống quản lý cán bộ, viên chức và người lao động tại Viện NLNTVN và hướng tới áp dụng cho hệ thống của Bộ KH&CN, học viên cần phải nghiên cứu chuyên sâu thêm một cách cụ thể về hệ thống đang được triển khai và sử dụng. Từ đó, tiến tới mở rộng nghiên cứu bằng cách tham khảo các công nghệ khác để có thể áp dụng và triển khai sát với nhu cầu thực tế.

KẾT LUẬN

Trong bối cảnh đất nước đẩy mạnh công cuộc chuyển đổi số quốc gia trên cả 3 trụ cột chính là: Chính phủ số, kinh tế số và xã hội số, việc triển khai và hoàn thiện số hóa các hồ sơ cán bộ là vô cùng cần thiết. Với chủ đề của năm 2025 là "Chuyển đổi số toàn dân, toàn diện, toàn trình để tăng tốc, bứt phá phát triển kinh tế số", công tác chuyển đổi số và khai thác thông tin, dữ liệu nhân sự là vô cùng cấp thiết với toàn bộ các cơ quan, công ty, tập thể quản lý nhân sự. Công tác này đóng vai trò là nền tảng để thực hiện tất cả các công tác quản trị con người trên nền tảng số hiện nay.

Khởi điểm về bài toán của nghiên cứu này dựa trên nhu cầu tích hợp dữ liệu về cán bộ, công chức, viên chức và người lao động trên toàn bộ hệ thống quản lý cán bộ quốc gia. Nhiệm vụ này phát sinh từ công tác sáp nhập Tỉnh, Thành cũng như các cơ quan Bộ, Ngành và nó có tính chất nền tảng, dễ dàng hiệu chỉnh cũng như nhân rộng, có khả năng áp dụng rộng rãi tại thời điểm này. Nghiên cứu này đóng vai trò quan trọng trong công tác đồng bộ, chuyển đổi, chuẩn hóa và làm sạch dữ liệu từ các nguồn hệ thống dữ liệu khác nhau.

Đề tài đã đặt ra vấn đề thực tế về quản lý dữ liệu cán bộ, công chức, viên chức tại Viện Năng lượng nguyên tử Việt Nam trong bối cảnh toàn thể các cơ quan bộ ngành trên cả nước tập trung chuyển đổi số trong quản lý cán bộ, công chức, viên chức theo quyết định 893/QĐ-TTg và việc sáp nhập cơ quan chủ quản dẫn đến chồng chéo, trùng lặp và xung đột dữ liệu về cán bộ, công chức, viên chức tại đơn vị.

Luận văn tập trung nghiên cứu tích hợp dữ liệu cán bộ, công chức, viên chứ của Việt Năng lượng nguyên tử Việt Nam là một đề tài có tính cấp thiết, xuất phát từ nhu cầu thực tế, với dữ liệu và hệ thống hiện hữu, không chỉ thuần túy về lý thuyết. Hơn nữa, cơ quan chủ quản là Bộ Khoa học và Công nghệ được sáp nhập với Bộ Thông tin và Truyền thông trong chương trình cải cách của quốc gia, cho nên đã xảy ra việc chồng chéo về cơ sở dữ liệu, làm cho các công tác quản trị nhân lực trở nên phức tạp, không thể đáp ứng được yêu cầu trong thời xu thế chuyển đổi số diễn ra mạnh mẽ hiện nay. Kết quả của đề tài luận văn đã đưa ra giải pháp tích hợp dữ liệu

nhằm làm sạch, chuẩn hóa và hợp nhất dữ liệu cán bộ, công chức, viên chức phù hợp với yêu cầu quản lý hiện nay của Viện Năng lượng nguyên tử Việt nam nói riêng và có khả năng hoàn thiện mở rộng tiến tới áp dụng trong Bộ Khoa học và Công nghệ nói chung.

Về mặt lý thuyết, luận văn đã trình bày và giới thiệu các nền tảng lý thuyết về Tích hợp dữ liệu thông qua các công trình nghiên cứu kinh điển và hiện đại (Lenzerini, Batini, IBM).

Về mặt thực tiễn, luận văn nghiên cứu về các cách tiếp cận vấn đề tích hợp dữ liệu trong thực tiễn hiện nay, qua đó đề xuất một hệ thống tích hợp dữ liệu phù hợp với nhu cầu thực tế tại Viện Năng lượng nguyên tử Việt Nam. Kết quả của luận văn đã được triển khai và xây dựng được quy trình tích hợp dữ liệu có khả năng: Kết hợp dữu liệu từ nhiều nguồn khác nhau; Chuẩn hóa và làm sạch dữ liệu về cán bộ; Loại bỏ hồ sơ trùng lặp và có giá trị phục vụ trực tiếp cho công tác quản lý nhân sự tại Viện Năng lượng nguyên tử Việt Nam. Luận văn thể hiện rõ tính ứng dụng thực tiễn, bám sát với thực tế và có tính ứng dụng cao nhằm giải quyết vấn đề về quản lý dữ liệu cán bộ, công chức, viên chức tại Viện Năng lượng nguyên tử Việt Nam.

Công cụ phần mềm hệ thống quản lý cán bộ, công chức, viên chức đã đáp ứng cơ bản và hiệu quả các yêu cầu quản lý hiện nay, góp phần hiện đại hóa công tác quản lý nhân sự, nâng cao hiệu suất làm việc và chất lượng quản lý của Viện Năng lượng nguyên tử Việt nam. Tuy nhiên, để phát huy tối đa hiệu quả, cần tiếp tục hoàn thiện một số chức năng, nâng cao khả năng tùy biến và tích hợp với các hệ thống liên quan trong thời gian tới.

Đây là một nghiên cứu mang tính chất ứng dụng với nhu cầu phát sinh từ định hướng phát triển quốc gia, do còn nhiều tiềm năng chưa được khai phá, các trường hợp cụ thể chưa được tìm ra dẫn đến các ý tưởng chuyển đổi dữ liệu mới chỉ ở mức tương đối. Do giới hạn về thời gian, kinh phí và điều kiện công tác nên không thể tránh khỏi hạn chế. Học viên mong muốn nhận được sự góp ý, nhận xét và phản hồi của các quý thầy cô, nhà khoa học, đồng nghiệp để luận văn được hoàn thiện hơn.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Thủ Tướng Chính phủ, "Quyết định số: 893/QĐ-TTg về việc Phê duyệt Đề án xây dựng cơ sở dữ liệu quốc gia về cán bộ, công chức, viên chức trong các cơ quan nhà nước," ngày 25/6/2020.

[2] Thông tư 06/2023/TT-BNV ngày 04/5/2023 về việc ban hành quy chế cập nhật, sử dụng, khai thác dữ liệu, thông tin của cơ sở dữ liệu quốc gia về cán bộ, công chức, viên chức trong các cơ quan nhà nước.

[3] Maurizio Lenzerini (6/2002). Data Integration: A Theoretical Perspective. [Online] Từ: https://www.researchgate.net/publication/220266329_Data_Integration_A_Theoretical_Perspective và https://doi.org/10.1145/543613.543644. Truy cập ngày 03/09/2025

[4] John Miles Smith; và đồng nghiệp (1982). *"Multibase: integrating heterogeneous distributed database systems"*. [Online] Từ: https://dl.acm.org/doi/10.1145/1500412.1500483. Truy cập ngày 03/09/2025

[5] Wei Li. (21/01/2020). *Towards Integrating Heterogeneous Data: A Spatial DBMS Solution from a CRC-LCL Project in Australia*. [Online] Từ: https://www.mdpi.com/2220-9964/9/2/63. Truy cập ngày 05/09/2025.

[6] Giuseppe Fusco và Lerina Aversano. (02/3/2020). *An approach for semantic integration of heterogeneous data sources.* [Online] Từ: http://www.dit.unitn.it/~pavel/OM/articles/peerj-cs-254.pdf. Truy cập ngày 07/09/2025.

[7] Keith D. Foote. (19/7/2023). *The Fundamentals of Data Integration*. [Online] Từ: https://www.dataversity.net/the-fundamentals-of-data-integration/. Truy cập ngày 11/09/2025.

[8] W.H. Inmon, *Building the Data Warehouse*, Nhà xuất bản Wiley Computer Publishing, 2002.

[9] V.Poe, Prentice. (1996). *Building a Data Warehouse for Decision Support,* Nhà xuất bản Prentice Hall, 1997

[10] Ralph Kimball và Margy Ross. *The Data Warehouse Toolkit– The Complete Guide to Dimensional Modeling*. Nhà xuất bản Wiley, 2002

[11] Karthick, N., & Kalrani, X. A. (2014). *A Survey on Data Aggregation in Big Data and Cloud Computing*. IJCTT, 2014

[12] Batini, C., Lenzerini M., & Navathe, S. B. *Data and Information Quality*. Springer, 2011

[13] Ralph Kimball (2004). *The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data.* Nhà xuất bản Wiley, Indianapolis, IN, 2004

[14] Tejasvi Addagada (18/10/2019). *Data Virtualization: An Agile Approach to Data Integration.* Công bố online tại nền tảng Research gate qua địa chỉ https://www.researchgate.net/publication/368607696_Data_Virtualization_An_Agile_Approach_to_Data_Integration. Truy cập ngày 20/09/2025.

[15] Steve Ruggles, J. David Hacker và Matthew Sobek (1995). *Order out of Chaos: The Integrated Public Use Microdata Series. Historical Methods, Volume 28.* [Online] từ: https://www.tandfonline.com/toc/vhim20/28/1. Truy cập ngày 20/09/2025.

[16] Surani, Ibrahim (2020). *Data Integration for Business Intelligence: Best Practices.* [Online] Từ: https://web.archive.org/web/20201130072535/https://www.dataversity.net/data-integration-for-business-intelligence-best-practices/ Truy cập ngày 20/09/2025.

[17] Shubhra S. Ray; et al (2009). *Combining Multi-Source Information through Functional Annotation based Weighting: Gene Function Prediction in Yeast.*

[Online] Từ http://shubhrasankar.tripod.com/cgi-bin/combiningMultisourceIEEE.pdf. Truy cập ngày 21/09/2025.

[18] Brenton Louie, Peter Mork, Fernando Martin-Sanchez, Alon Halevy, Peter Tarczy-Hornoch (2007), *Data Integration and genomic medicine* , Tạp chí Science Direct, [Online] từ: https://www.sciencedirect.com/science/article/pii/S1532046406000244. Truy cập ngày 10/10/2025

KiểmTraTàiLiệu

BÁO CÁO KIỂM TRA TRÙNG LẶP

Thông tin tài liệu

Tên tài liệu:	B23CHIS016 - Luận văn ThS Trần Đức Long - Data Integration
Tác giả:	Trần Đức Long
Điểm trùng lặp:	12
Thời gian tải lên:	16:37 05/02/2026
Thời gian sinh báo cáo:	16:39 05/02/2026
Các trang kiểm tra:	67/67 trang

Kết quả kiểm tra trùng lặp



Nguồn trùng lặp tiêu biểu

tailieu.vn 123docz.net vi.wikipedia.org

Học viên

Trần Đức Long

BẢN CAM ĐOAN

Tôi cam đoan đã thực hiện việc kiểm tra mức độ tương đồng nội dung đề án tốt nghiệp qua phần mềm Kiểm tra tài liệu một cách trung thực và đạt kết quả mức độ tương đồng 12% toàn bộ nội dung đề án tốt nghiệp. Bản đề án tốt nghiệp kiểm tra qua phần mềm là bản cứng đề án tốt nghiệp đã nộp để bảo vệ trước hội đồng. Nếu sai tôi xin chịu các hình thức kỷ luật theo quy định hiện hành của Học viện.

Học viên

Trần Đức Long

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

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

BÁO CÁO GIẢI TRÌNH
SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Họ và tên học viên: Trần Đức Long

Chuyên ngành: Hệ thống thông tin

Khóa: 2023 đợt 1

Tên đề tài: Tích hợp dữ liệu cho hệ thống quản lý cán bộ, viên chức của Viện Năng lượng nguyên tử Việt Nam

Người hướng dẫn khoa học: TS. Trần Thiện Chính

Ngày bảo vệ: 24/01/2026

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

TT	Ý kiến hội đồng	Sửa chữa của học viên
1	Làm rõ bài toán và đóng góp của đề tài	Đã bổ sung làm rõ bài toán và đóng góp của đề tài bắt đầu từ dòng thứ 16 trang 55 đến hết dòng 2 trang 54 và dòng 9 trang 54 đến hết dòng 22 trang 56
2	Làm rõ cách đánh giá kết quả của đề tài	Đã bổ sung phần đánh giá kết quả của hệ thống tại mục 3.3 tại trang 51, trang 52, trang 53.
3	Kiểm tra lại các tài liệu tham khảo	Đã kiểm tra lại và chỉnh sửa các đường link lỗi các tài liệu tham khảo trực tuyến. Chỉnh sửa lại format và Bổ sung thêm ngày truy cập tài liệu trực tuyến.

Hà Nội, ngày 04 tháng 02 năm 2026

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN	THƯ KÝ HỘI ĐỒNG	NGƯỜI HƯỚNG DẪN KHOA HỌC	HỌC VIÊN
PGS.TSKH. Hoàng Đăng Hải	TS. Phạm Hoàng Duy	TS. Trần Thiện Chính	Trần Đức Long

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

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

BIÊN BẢN
HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 3299/QĐ-HV ngày 26 tháng 12 năm 2025 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi ..10... giờ 30.... phút, ngày 24 tháng 01 năm 2025 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Trần Đức Long**

Tên đề án tốt nghiệp: **Tích hợp dữ liệu cho hệ thống quản lý cán bộ, viên chức của Viện năng lượng nguyên tử**

Ngành: **Hệ thống thông tin** Mã số: **8.48.01.04**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: ..05../ 05

TT	HỌ VÀ TÊN	TRÁCH NHIỆM TRONG HĐ	GHI CHÚ
1	**PGS.TSKH. Hoàng Đăng Hải**	Chủ tịch	
2	**TS. Phạm Hoàng Duy**	Thư ký	
3	**PGS.TS. Bùi Thu Lâm**	Phản biện 1	
4	**TS. Trịnh Bảo Ngọc**	Phản biện 2	
5	**TS. Trần Tiến Công**	Uỷ viên	

Các nội dung thực hiện:

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên (Có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

Thầy Lâm: Làm rõ hơn về đóng góp của đề tài? Tài liệu tham khảo kiểm tra lại?
Cách đánh giá kết quả của đề tài? Bổ sung yêu cầu an toàn hoạt động?
Thầy Ngọc: Giải thích sự biến động của đề tài? Làm rõ hơn đề tài?

7. Trả lời của học viên:

Học viên trả lời các câu hỏi của hội đồng

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên *(có văn bản kèm theo).*

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:
  1. Trưởng Ban kiểm phiếu: TS. Trần Tiến Công
  2. Uỷ viên Ban kiểm phiếu: TS. Trịnh Bảo Ngọc
  3. Uỷ viên Ban kiểm phiếu: TS. Phạm Hoàng Duy
- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.
- Ban kiểm phiếu làm việc:
- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu *(có Biên bản họp Ban kiểm phiếu kèm theo)*
- **Điểm trung bình của đề án tốt nghiệp:** 7,5

Kết luận:

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:

Làm rõ hơn về tính cấp thiết của đề tài

Làm rõ cách đánh giá kết quả của đề tài

2. Đề nghị Học viện công nhận (hoặc không) và cấp bằng (hoặc không) thạc sĩ cho học viên: Trần Đức Long

3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho NCS.: Không

Buổi làm việc kết thúc vào 12h40 cùng ngày.

Chủ tịch	**Thư ký**
PGS.TSKH. Hoàng Đăng Hải	**TS. Phạm Hoàng Duy**

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

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: Tích hợp dữ liệu cho hệ thống quản lý cán bộ, viên chức của Viện Năng lượng nguyên tử
Chuyên ngành: Hệ thống thông tin
Mã chuyên ngành: 8.48.01.04
Họ và tên học viên: Trần Đức Long
Họ và tên người nhận xét: Trịnh Bảo Ngọc
Học hàm, học vị: Tiến sĩ
Chuyên ngành: Kỹ thuật phần mềm
Cơ quan công tác: Trường đại học Hà Nội
Số điện thoại: 0987 355 799 E-mail: ngoctb@hanu.edu.vn

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Về cơ sở khoa học: Đề tài được xây dựng trên nền tảng lý thuyết vững chắc về Tích hợp dữ liệu (Data Integration), có phân biệt rõ ràng với Data Aggregation và Data Consolidation. Tác giả đã tham chiếu các công trình nghiên cứu kinh điển và hiện đại (Lenzerini, Batini, IBM, các mô hình ETL/ELT, Data Virtualization), cho thấy việc lựa chọn hướng tiếp cận là có cơ sở khoa học, phù hợp với chuyên ngành Hệ thống thông tin.

Về cơ sở thực tiễn: Đề tài bám sát bài toán thực tế tại Viện Năng lượng nguyên tử Việt Nam (chưa thấy có minh chứng cụ thể nào về việc này, tuy nhiên nếu đúng vậy), và trong bối cảnh:

- Chuyển đổi số trong quản lý CBCCVC theo Quyết định 893/QĐ-TTg;
- Thay đổi, sáp nhập cơ quan chủ quản dẫn đến chồng chéo, trùng lặp, xung đột dữ liệu giữa các hệ thống quản lý nhân sự của Bộ KH&CN trước và sau sáp nhập;
- Quy trình cập nhật dữ liệu hiện nay còn thủ công, tốn thời gian, dễ sai sót.

Tính cấp thiết: Đề tài có tính cấp thiết cao, xuất phát từ nhu cầu thật, dữ liệu thật, hệ thống thật, không mang tính lý thuyết thuần túy. Việc đề xuất giải pháp tích hợp dữ liệu nhằm làm sạch, chuẩn hóa và hợp nhất dữ liệu CBCCVC là hoàn toàn phù hợp với yêu cầu quản lý hiện nay của Viện và có khả năng mở rộng áp dụng ở cấp Bộ.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

Kết cấu đề án: Tổng số: khoảng 53 trang (không kể phụ lục), gồm 03 chương, ngoài ra có phần Mở đầu, Kết luận và Tài liệu tham khảo.

Tóm tắt nội dung các chương:

- Chương 1(18 trang): Tổng quan vấn đề nghiên cứu, trình bày cơ sở lý luận về tích hợp dữ liệu; phân tích thực trạng công tác quản lý CBCCVC tại Viện Năng lượng nguyên tử; khảo sát và đánh giá các hệ thống quản lý cán bộ đã triển khai (2020, 2024, 2025), chỉ ra rõ các bất cập về dữ liệu và nguyên nhân phát sinh lỗi.
- Chương 2 (14 trang): Các kỹ thuật tích hợp dữ liệu, hệ thống hóa lý thuyết Data Integration; trình bày các kỹ thuật ETL/ELT, Data Virtualization; phân tích ưu nhược điểm và lựa chọn kỹ thuật phù hợp với bài toán quản lý nhân sự tại Viện.
- Chương 3 (13 trang): Xây dựng công cụ phần mềm, thử nghiệm và đánh giá Trình bày quy trình xây dựng, cấu hình công cụ tích hợp dữ liệu; mô tả các bước trích xuất – làm sạch – chuẩn hóa – tích hợp dữ liệu (theo CCCD); đánh giá kết quả thử nghiệm và đề xuất hướng phát triển tiếp theo.

Kết quả đạt được:

- Xây dựng được mô hình và quy trình tích hợp dữ liệu khả thi cho dữ liệu CBCCVC;
- Thử nghiệm thành công trên nhiều tập dữ liệu thực tế, cho kết quả làm sạch và loại bỏ trùng lặp;

III/ Những vấn đề cần giải thích thêm:

Về hình thức trình bày:

- Còn một số lỗi nhỏ về chính tả, khoảng trắng, ngắt dòng (ví dụ: "Intergration" → "Integration");
- Một số hình/bảng chưa được căn chỉnh đồng đều hoặc chú thích chưa thật sự ngắn gọn; Nhiều hình vẽ còn ở chất lượng thấp cần thay thế, ví dụ: Hình 2.5 khá mờ khó nhìn rõ
- Cũng cần nhất quán cách vẽ các hình cùng với mục đích, ví dụ như mô tả cấu trúc hệ thống, đã dùng đen trắng là đen trắng tất cả, không phải chỗ đen trắng, chỗ dùng màu, ví dụ như Hình 2.2, Hình 2.3
- Cần thống nhất cách viết thuật ngữ tiếng Anh (in nghiêng hoặc trong ngoặc).

Về nội dung:

- Chương 2 có thể rút gọn phần tổng quan thế giới để dành dung lượng phân tích sâu hơn vào so sánh kỹ thuật lựa chọn;
- Chương 3 nên bổ sung thêm tiêu chí đánh giá định lượng (thời gian xử lý, tỷ lệ bản ghi trùng, độ chính xác sau tích hợp);
- Phần định hướng phát triển có thể làm rõ hơn khả năng tích hợp với CSDLQG về CBCCVC.
- Cần trau chuốt lại cách viết, tránh phải các lỗi diễn đạt không đáng có, cần thực hiện xuyên suốt. Ví dụ 1 lỗi như: "Mặc dù thuật ngữ kỹ thuật "Tích hợp dữ liệu" hay "Data Integration" có nhiều khái niệm và cách định nghĩa khác nhau". Đây là 1 thuật ngữ diễn đạt theo 2 ngôn ngữ khác nhau, viết như thế này thì thành ra lại đang mô tả 2 thuật ngữ?
- Việc phân tích tổng quan các kỹ thuật tích hợp dữ liệu còn thiếu, ví dụ nên liệt kê và phân tích thêm về: Semantic Data Integration, AI assist, Metadata driven, nên tìm hiểu sâu hơn và bổ sung để hoàn thiện phần này, từ đó làm tiền đề cho việc phân tích lý do tại sao chọn kỹ thuật như trong bài

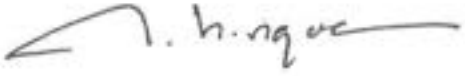
- Đồng thời, đề tài hoàn toàn chỉ tập trung vào phân tích lý thuyết, còn cụ thể nghiên cứu chính của đề tài là ứng dụng Tích hợp dữ liệu vào 1 hệ thống thật thì hoàn toàn không thấy việc dẫn giải từng bước ở bất cứ đâu, nếu chỉ dùng các thư viện có sẵn (vd như pandas) và code thêm 1 chút nữa và gọi đây là tích hợp dữ liệu thì làm luận văn tốt nghiệp có phải dễ quá không?

Câu hỏi:

1. Đánh giá một cách khách quan, kỹ thuật ETL/ELT không có gì đặc sắc, bản chất đó là khái niệm hóa một quy trình phổ thông, thông thường trong việc tích hợp dữ liệu. Thực tế là người làm không cần nghe tên kỹ thuật này và biết nó là gì, với tư duy logic cơ bản cũng sẽ làm theo các bước này. Vậy tại sao nhóm tác giả lại chọn phương pháp này, lý do thông dụng được nêu ra (tại 2.2.1) không hợp lý lắm, đúng ra phải là phương pháp tốt nhất chứ?
2. So sánh ETL và Data Virtualization trong bối cảnh hệ thống quản lý CBCCVC: trong trường hợp nào nên ưu tiên ETL, và khi nào Data Virtualization sẽ phù hợp hơn?

IV/ Kết luận:
Đồng ý.

Ngày 24 tháng 01 năm 2026
NGƯỜI NHẬN XÉT



TS Trịnh Bảo Ngọc

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: Tích hợp dữ liệu cho hệ thống quản lý cán bộ, viên chức của Viện Năng lượng nguyên tử Việt Nam

Chuyên ngành: Hệ thống thông tin
Mã chuyên ngành: 8.48.01.04

Họ và tên học viên: Trần Đức Long

Họ và tên người nhận xét: PGS TS Bùi Thu Lâm
Học hàm, học vị: PGS TS
Chuyên ngành: KHMT
Cơ quan công tác: Học viện Kỹ thuật Mật mã
Số điện thoại: 0968699975 E-mail: lambt.actvn@gmail.com

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài

Đề tài được lựa chọn xuất phát từ nhu cầu thực tiễn và có tính thời sự cao, gắn trực tiếp với quá trình chuyển đổi số trong cơ quan nhà nước và đặc biệt là bối cảnh sắp xếp, sáp nhập các bộ ngành, dẫn tới tình trạng phân mảnh, chồng chéo và thiếu nhất quán dữ liệu cán bộ, công chức, viên chức.

Học viên đã trình bày tương đối đầy đủ bối cảnh pháp lý (Quyết định 893/QĐ-TTg, các kế hoạch sắp xếp bộ máy), bối cảnh tổ chức (đặc thù của Viện Năng lượng nguyên tử Việt Nam), qua đó làm rõ tính cấp thiết và ý nghĩa thực tiễn của bài toán tích hợp dữ liệu trong quản lý nhân sự.

Về cơ sở khoa học, đề tài dựa trên các khái niệm và kỹ thuật phổ biến trong lĩnh vực Data Integration, ETL/ELT, Data Consolidation, phù hợp với chuyên ngành Hệ thống thông tin và định hướng ứng dụng của chương trình đào tạo thạc sĩ.

Tuy nhiên, phần cơ sở khoa học mới dừng ở mức tổng quan mô tả, chưa làm rõ được khoảng trống nghiên cứu (research gap) hay sự khác biệt cụ thể giữa giải pháp của học viên so với các công trình/giải pháp đã có.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được

Luận văn được bố cục hợp lý, gồm 3 chương, bám sát logic từ bối cảnh – lý thuyết – triển khai thử nghiệm, phù hợp với một đề án thạc sĩ theo định hướng ứng dụng.

- Chương 1 trình bày tổng quan về bài toán tích hợp dữ liệu, phân biệt rõ các khái niệm Data Integration, Data Aggregation và Data Consolidation; đồng thời mô

tả chi tiết thực trạng hệ thống quản lý cán bộ tại Viện Năng lượng nguyên tử Việt Nam.

- Chương 2 giới thiệu các kỹ thuật tích hợp dữ liệu, tập trung chủ yếu vào ETL/ELT và Data Virtualization, kèm theo phân tích lý do lựa chọn công nghệ phù hợp với điều kiện thực tế của đơn vị.
- Chương 3 trình bày quá trình xây dựng công cụ, cấu hình hệ thống, xử lý dữ liệu thực (chuẩn hóa CCCD, loại bỏ trùng lặp, chuyển đổi định dạng), và minh họa kết quả tích hợp thông qua các tập dữ liệu cụ thể.

Kết quả đạt được cho thấy học viên đã triển khai một quy trình tích hợp dữ liệu ở mức cơ bản, có khả năng:

- Kết hợp dữ liệu từ nhiều nguồn khác nhau;
- Chuẩn hóa và làm sạch dữ liệu cán bộ;
- Giảm thiểu trùng lặp hồ sơ;
- Phục vụ trực tiếp cho công tác quản lý nhân sự tại đơn vị.

Đề tài thể hiện rõ tính ứng dụng thực tiễn, phù hợp với đối tượng là cán bộ kỹ thuật/QLCNTT trong cơ quan nhà nước.

III/ Những vấn đề cần giải thích thêm (các điểm yếu, thiếu sót)

Đây là phần còn nhiều hạn chế, học viên cần giải thích và làm rõ thêm trước khi bảo vệ:

1. Tài liệu tham khảo chưa trích dẫn nghiêm túc, danh sách tài liệu cần kiểm chứng (link hiện đang sai) và format chuẩn chỉnh hơn.
2. Thiếu mô hình kiến trúc tổng thể (overall architecture)
    - Luận văn chưa trình bày rõ kiến trúc tổng thể của hệ thống tích hợp dữ liệu (ví dụ: kiến trúc 3 tầng, pipeline ETL đầy đủ, vị trí của công cụ tích hợp trong hệ thống CNTT của Viện).
    - Các hình minh họa chủ yếu mang tính thao tác kỹ thuật, chưa thể hiện được bức tranh kiến trúc hệ thống ở mức thiết kế.
3. Chưa có tiêu chí đánh giá định lượng hiệu quả tích hợp
    - Kết quả thử nghiệm chủ yếu mang tính mô tả (ảnh chụp dữ liệu trước/sau).
    - Chưa có các chỉ số đánh giá cụ thể như: tỷ lệ trùng lặp trước/sau tích hợp, thời gian xử lý, mức độ giảm sai sót, khả năng mở rộng dữ liệu.
    - Điều này làm giảm tính thuyết phục về hiệu quả kỹ thuật của giải pháp.
4. Phần bảo mật và an toàn dữ liệu còn rất mờ nhạt
    - Dữ liệu cán bộ, viên chức là dữ liệu nhạy cảm, nhưng luận văn hầu như chưa phân tích:
        - Phân quyền truy cập;
        - Kiểm soát truy cập dữ liệu tích hợp;

- Rủi ro rò rỉ dữ liệu trong quá trình ETL;
- Tuân thủ các quy định về bảo vệ dữ liệu cá nhân.

- Đây là một thiếu sót nghiêm trọng nếu xét trong bối cảnh ứng dụng thực tế tại cơ quan nhà nước.

5. Chưa làm rõ tính mới (novelty) của đề tài
   - Giải pháp tích hợp dữ liệu được đề xuất chủ yếu dựa trên các kỹ thuật phổ biến (ETL, chuẩn hóa dữ liệu).
   - Luận văn chưa chỉ ra rõ:
     - Điểm cải tiến so với các hệ thống hiện có;
     - Hay đóng góp mới ở mức phương pháp, quy trình hay tổ chức dữ liệu.
6. So sánh các phương án kỹ thuật còn đơn giản
   - Việc lựa chọn ETL thay vì Data Virtualization hoặc các kiến trúc hiện đại hơn (Data Lake, Microservices Data Integration) chưa được phân tích sâu về ưu – nhược điểm trong bối cảnh dài hạn.

IV/ Kết luận

Luận văn là một đề án ứng dụng, có tính thực tiễn cao, phù hợp với chuyên ngành Hệ thống thông tin và đối tượng đào tạo thạc sĩ theo định hướng ứng dụng. Học viên đã cơ bản thể hiện được khả năng tiếp cận bài toán thực tế, triển khai kỹ thuật tích hợp dữ liệu và xây dựng công cụ thử nghiệm.

Tuy nhiên, luận văn còn hạn chế về chiều sâu học thuật, đánh giá định lượng, phân tích bảo mật và kiến trúc hệ thống, cần được học viên giải thích rõ ràng trong buổi bảo vệ.

Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp, với điều kiện học viên cần làm rõ các vấn đề nêu trên trong phần trả lời hội đồng.

Ngày 24 tháng 1 năm 2026

NGƯỜI NHẬN XÉT

Bùi Thu Lâm