

ĐẠI HỌC BÁCH KHOA HÀ NỘI
TRƯỜNG ĐIỆN – ĐIỆN TỬ



ĐỒ ÁN TỐT NGHIỆP

Đề tài:

**BÁO CÁO, DỰ ĐOÁN TÌNH HÌNH COVID-19 SỬ DỤNG
DỊCH VỤ CLOUD MICROSOFT AZURE CHO DATA
ENGINEER**

Sinh viên thực hiện: DƯƠNG CÔNG KIÊN

MSSV: 20182614

ĐVTN.T04 – K63

Giảng viên hướng dẫn: PGS. TS. HOÀNG MẠNH THẮNG

Hà Nội, 6/2022

ĐẠI HỌC BÁCH KHOA HÀ NỘI
TRƯỜNG ĐIỆN – ĐIỆN TỬ



ĐỒ ÁN TỐT NGHIỆP

Đề tài:

**BÁO CÁO, DỰ ĐOÁN TÌNH HÌNH COVID-19 SỬ DỤNG
DỊCH VỤ CLOUD MICROSOFT AZURE CHO DATA
ENGINEER**

Sinh viên thực hiện: **DƯƠNG CÔNG KIÊN**

MSSV: 20182614

ĐVTN: 04 – K63

Giảng viên hướng dẫn: **PGS. TS. HOÀNG MẠNH THẮNG**

Cán bộ phản biện:

Hà Nội, 6/2022

ĐÁNH GIÁ QUYỀN ĐỒ ÁN TỐT NGHIỆP

(Dùng cho giảng viên hướng dẫn)

Giảng viên đánh giá: PGS. TS. Hoàng Mạnh Thắng.....

Họ và tên sinh viên: Dương Công Kiên MSSV: 20182614.....

Tên đồ án: Báo cáo, dự đoán tình hình Covid-19 sử dụng dịch vụ Cloud Microsoft Azure cho Data Engineer.....

Chọn các mức điểm phù hợp cho sinh viên trình bày theo các tiêu chí dưới đây:

Rất kém (1); Kém (2); Đạt (3); Giới (4); Xuất sắc (5)

Có sự kết hợp giữa lý thuyết và thực hành (20)					
	Nêu rõ tính cấp thiết và quan trọng của đề tài, các vấn đề và các giả thuyết (bao gồm mục đích và tính phù hợp) cũng như phạm vi ứng dụng của đồ án	1	2	3	4
1	Cập nhật kết quả nghiên cứu gần đây nhất (trong nước/quốc tế)	1	2	3	4
3	Nêu rõ và chi tiết phương pháp nghiên cứu/giải quyết vấn đề	1	2	3	4
4	Có kết quả mô phỏng/thực nghiệm và trình bày rõ ràng kết quả đạt được	1	2	3	4

Có khả năng phân tích và đánh giá kết quả (15)						
5	Kế hoạch làm việc rõ ràng bao gồm mục tiêu và phương pháp thực hiện dựa trên kết quả nghiên cứu lý thuyết một cách có hệ thống	1	2	3	4	5
6	Kết quả được trình bày một cách logic và dễ hiểu, tất cả kết quả đều được phân tích và đánh giá thỏa đáng.	1	2	3	4	5
7	Trong phần kết luận, tác giả chỉ rõ sự khác biệt (nếu có) giữa kết quả đạt được và mục tiêu ban đầu để ra đồng thời cung cấp lập luận để đề xuất hướng giải quyết có thể thực hiện trong tương lai.	1	2	3	4	5

Kỹ năng viết quyển đồ án (10)						
	Đồ án trình bày đúng mẫu quy định với cấu trúc các chương logic và đẹp mắt (bảng biểu, hình ảnh rõ ràng, có tiêu đề, được đánh số thứ tự và được giải thích hay đề cập đến trong đồ án, có căn lề, dấu cách sau dấu chấm, dấu phẩy v.v), có mở đầu chương và kết luận chương, có liệt kê tài liệu tham khảo và có trích dẫn đúng quy định	1	2	3	4	5
8	Kỹ năng viết xuất sắc (cấu trúc câu chuẩn, văn phong khoa học, lập luận logic và có cơ sở, từ vựng sử dụng phù hợp v.v.)	1	2	3	4	5
9						

Thành tựu nghiên cứu khoa học (5) (chọn 1 trong 3 trường hợp)		
10a	Có bài báo khoa học được đăng hoặc chấp nhận đăng/đạt giải SVNC khoa học giải 3 cấp Viện trở lên/các giải thưởng khoa học (quốc tế/trong nước) từ giải 3 trở lên/ Có đăng ký bằng phát minh sáng chế	5
10b	Được báo cáo tại hội đồng cấp Viện trong hội nghị sinh viên nghiên cứu khoa học nhưng không đạt giải từ giải 3 trở lên/Đạt giải khuyến khích trong các kỳ thi quốc gia và quốc tế khác về chuyên ngành như TI contest.	2
10c	Không có thành tích về nghiên cứu khoa học	0
Điểm tổng		/50

Điểm tổng quy đổi về thang 10

Nhận xét khác của cán bộ phản biện

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

Ngày: ... / ... / 2022

Người nhận xét
(Ký và ghi rõ họ tên)

ĐÁNH GIÁ QUYỀN ĐỒ ÁN TỐT NGHIỆP

(Dùng cho cán bộ phản biện)

Giảng viên đánh giá:.....

Ho và tên sinh viên: Dương Công Kiên MSSV: 20182614.....

Tên đồ án: Báo cáo, dự đoán tình hình Covid-19 sử dụng dịch vụ Cloud Microsoft Azure cho Data Engineer.....

Chọn các mức điểm phù hợp cho sinh viên trình bày theo các tiêu chí dưới đây:

Rất kém (1); Kém (2); Đạt (3); Giỏi (4); Xuất sắc (5)

Có sự kết hợp giữa lý thuyết và thực hành (20)					
	Nêu rõ tính cấp thiết và quan trọng của đề tài, các vấn đề và các giả thuyết (bao gồm mục đích và tính phù hợp) cũng như phạm vi ứng dụng của đồ án	1	2	3	4
1	Cập nhật kết quả nghiên cứu gần đây nhất (trong nước/quốc tế)	1	2	3	4
3	Nêu rõ và chi tiết phương pháp nghiên cứu/giải quyết vấn đề	1	2	3	4
4	Có kết quả mô phỏng/thực nghiệm và trình bày rõ ràng kết quả đạt được	1	2	3	4

Có khả năng phân tích và đánh giá kết quả (15)						
5	Kế hoạch làm việc rõ ràng bao gồm mục tiêu và phương pháp thực hiện dựa trên kết quả nghiên cứu lý thuyết một cách có hệ thống	1	2	3	4	5
6	Kết quả được trình bày một cách logic và dễ hiểu, tất cả kết quả đều được phân tích và đánh giá thỏa đáng.	1	2	3	4	5
7	Trong phần kết luận, tác giả chỉ rõ sự khác biệt (nếu có) giữa kết quả đạt được và mục tiêu ban đầu để ra đồng thời cung cấp lập luận để đề xuất hướng giải quyết có thể thực hiện trong tương lai.	1	2	3	4	5

Kỹ năng viết quyển đồ án (10)						
	Đồ án trình bày đúng mẫu quy định với cấu trúc các chương logic và đẹp mắt (bảng biểu, hình ảnh rõ ràng, có tiêu đề, được đánh số thứ tự và được giải thích hay đề cập đến trong đồ án, có cẩn lè, dấu cách sau dấu chấm, dấu phẩy v.v), có mở đầu chương và kết luận chương, có liệt kê tài liệu tham khảo và có trích dẫn đúng quy định	1	2	3	4	5
8	Đồ án trình bày đúng mẫu quy định với cấu trúc các chương logic và đẹp mắt (bảng biểu, hình ảnh rõ ràng, có tiêu đề, được đánh số thứ tự và được giải thích hay đề cập đến trong đồ án, có cẩn lè, dấu cách sau dấu chấm, dấu phẩy v.v), có mở đầu chương và kết luận chương, có liệt kê tài liệu tham khảo và có trích dẫn đúng quy định	1	2	3	4	5
9	Kỹ năng viết xuất sắc (cấu trúc câu chuẩn, văn phong khoa học, lập luận logic và có cơ sở, từ vựng sử dụng phù hợp v.v.)	1	2	3	4	5

Thành tựu nghiên cứu khoa học (5) (chọn 1 trong 3 trường hợp)		
10a	Có bài báo khoa học được đăng hoặc chấp nhận đăng/đạt giải SVNC khoa học giải 3 cấp Viện trở lên/các giải thưởng khoa học (quốc tế/trong nước) từ giải 3 trở lên/ Có đăng ký bằng phát minh sáng chế	5
10b	Được báo cáo tại hội đồng cấp Viện trong hội nghị sinh viên nghiên cứu khoa học nhưng không đạt giải từ giải 3 trở lên/Đạt giải khuyến khích trong các kỳ thi quốc gia và quốc tế khác về chuyên ngành như TI contest.	2
10c	Không có thành tích về nghiên cứu khoa học	0
Điểm tổng		/50

Điểm tổng quy đổi về thang 10

Nhận xét khác của cán bộ phản biện

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

Ngày: ... / ... / 2022

Người nhận xét
(Ký và ghi rõ họ tên)

LỜI NÓI ĐẦU

Cách mạng công nghệ lần thứ tư hay còn được biết đến với tên gọi cách mạng công nghiệp 4.0 là cuộc cách mạng tập trung vào những công nghệ hiện đại dựa trên nền tảng kĩ thuật số. Trong đó, vai trò của dữ liệu lớn trong các xu hướng công nghệ là rất quan trọng. Nó được xem là một trong các công nghệ lõi trong Công nghiệp 4.0, là một trong bốn thành phần chính của Internet kết nối vạn vật (IoT). Dữ liệu lớn được mô tả như loại hàng hóa mới cho nền kinh tế thế kỷ 21.

Việc tận dụng và khai thác dữ liệu để phục vụ các mục đích cải thiện mức độ hiệu quả, hiệu suất của một hoạt động nhất định ngày càng trở nên quan trọng và đem lại lợi ích cực kỳ to lớn. Dữ liệu được xem là tài sản cực kỳ chủ lực không thuộc lĩnh vực tài chính và nhân lực, tài nguyên này cần phải được quản lý và sử dụng đúng cách, triệt để tận dụng được mọi tiềm năng mà nó đem lại. Nếu một doanh nghiệp bỏ qua hoặc không thành thạo để khai thác và sử dụng dữ liệu thì đó là một “thất bại” nặng nề mà các doanh nghiệp sẽ phải hứng chịu. Với tốc độ bùng nổ dữ liệu như hiện nay, việc bắt kịp xu hướng và làm chủ dữ liệu sẽ đem lại lợi ích to lớn cho doanh nghiệp.

Một khi Big Data được cân nhắc về tính cấp bách, quan trọng thì những công cụ hỗ trợ khai thác giá trị bên trong Big Data sẽ ngày càng được chú ý hơn. Ngoài những phần mềm hỗ trợ cho việc phân tích, thì việc áp dụng các hệ thống lưu trữ, quản lý Big Data như điện toán đám mây (cloud computing) cũng cực kỳ cần thiết. Trước tình hình đó, tôi quyết định nghiên cứu đề tài “Báo cáo, dự đoán tình hình Covid-19 sử dụng dịch vụ Cloud Microsoft Azure cho Data Engineer” để hiểu và tận dụng được tầm quan trọng của Cloud Computing đối với Big Data vào thực tế.

Cuối cùng, tôi xin được gửi lời cảm ơn chân thành đến PGS. TS. Hoàng Mạnh Thắng, giảng viên Trường Điện – Điện tử, Đại học Bách Khoa Hà Nội đã tận tình hướng dẫn và tạo điều kiện giúp tôi hoàn thành được đề tài.

LỜI CAM ĐOAN

Tôi là Dương Công Kiên, mã số sinh viên 20182614, sinh viên lớp Điện tử 04 – K63. Người hướng dẫn là PGS. TS. Hoàng Mạnh Thắng. Tôi xin cam đoan toàn bộ nội dung được trình bày trong đồ án “Báo cáo, dự đoán tình hình Covid-19 sử dụng dịch vụ Cloud Microsoft Azure cho Data Engineer” là kết quả của quá trình tìm kiếm và nghiên cứu của tôi. Các dữ liệu được nêu trong đồ án hoàn toàn trung thực, phản ánh đúng kết quả mà tôi đã làm được. Mọi thông tin trích dẫn đều tuân thủ các quy định về sở hữu trí tuệ; các tài liệu tham khảo được liệt kê rõ ràng. Tôi xin chịu hoàn toàn trách nhiệm với những nội dung được viết trong đồ án này.

Hà Nội, ngày 28, tháng 06, năm 2022

Người cam đoan

Dương Công Kiên

MỤC LỤC

DANH MỤC KÝ HIỆU VÀ CHỮ VIẾT TẮT	i
DANH MỤC HÌNH VẼ	ii
DANH MỤC BẢNG BIỂU.....	iv
TÓM TẮT ĐỒ ÁN	v
ABSTRACT.....	vi
PHẦN MỞ ĐẦU.....	vii
CHƯƠNG 1. ĐẶT VẤN ĐỀ.....	1
<i>1.1 Đặt vấn đề.....</i>	<i>1</i>
<i>1.2 Mục tiêu nghiên cứu, đối tượng và phạm vi đề tài</i>	<i>2</i>
<i>1.3 Ý nghĩa của đề tài.....</i>	<i>2</i>
<i>1.4 Những nghiên cứu gần đây.....</i>	<i>2</i>
<i>1.5 Kết quả dự kiến</i>	<i>3</i>
<i>1.6 Kế hoạch thực hiện đề tài</i>	<i>3</i>
<i>1.7 Kết luận chương.....</i>	<i>4</i>
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT.....	5
<i>2.1 Tổng quan về Big Data và Microsoft Azure</i>	<i>5</i>
2.1.1 Dữ liệu lớn và phân tích dữ liệu	5
2.1.2 Các công nghệ đặc biệt hỗ trợ Big Data	7
2.1.3 Các dịch vụ của Microsoft Azure ứng dụng cho Big Data.....	13
<i>2.2 Ứng dụng dữ liệu lớn vào thực thế</i>	<i>16</i>
2.2.1 Ứng dụng trong ngành Ngân hàng	16
2.2.2 Ứng dụng trong ngành Y tế	17
2.2.3 Ứng dụng trong ngành Thương mại điện tử.....	17
2.2.4 Ứng dụng trong ngành Digital Marketing hoặc Social Media	18
2.2.5 Một số ứng dụng khác	19
<i>2.3 Kết luận chương.....</i>	<i>19</i>

CHƯƠNG 3. XÂY DỰNG BÁO CÁO COVID-19 SỬ DỤNG CÁC DỊCH VỤ CỦA MICROSOFT AZURE.....	20
<i> 3.1 Tổng quan bài toán.....</i>	<i> 20</i>
3.1.1 Những dịch vụ Microsoft Azure được sử dụng.....	20
3.1.2 Các nguồn dữ liệu đầu vào.....	23
<i> 3.2 Xây dựng giải pháp thực hiện bài toán.....</i>	<i> 24</i>
3.2.1 Giải pháp thiết kế (Solution Architecture)	24
3.2.2 Cài đặt môi trường	25
<i> 3.3 Triển khai bài toán.....</i>	<i> 26</i>
3.3.1 Thu thập dữ liệu (Ingestion).....	26
3.3.2 Chuyển đổi dữ liệu (Transformation)	29
3.3.3 Trực quan hóa dữ liệu (Exploitation)	43
3.3.4 Quản lý luồng dữ liệu (Monitoring).....	45
<i> 3.4 Kết luận chương.....</i>	<i> 46</i>
CHƯƠNG 4. KẾT QUẢ VÀ THẢO LUẬN.....	46
<i> 4.1 Kết quả thu được.....</i>	<i> 46</i>
4.1.1 Kết quả xây dựng hệ thống trên Azure Data Factory	47
4.1.2 Kết quả trực quan hóa dữ liệu	48
4.1.3 Kết quả giám sát quá trình vận hành của các đường ống dữ liệu (data pipeline) ..	50
<i> 4.2 Đánh giá kết quả.....</i>	<i> 51</i>
<i> 4.3 Kết luận chương.....</i>	<i> 51</i>
KẾT LUẬN	52
<i> Kết luận chung.....</i>	<i> 52</i>
<i> Hướng phát triển</i>	<i> 52</i>
TÀI LIỆU THAM KHẢO.....	53

DANH MỤC KÝ HIỆU VÀ CHỮ VIẾT TẮT

STT	Chữ viết tắt	Chữ viết đầy đủ
1	ADF	Azure Data Factory
2	AI	Artificial Intelligence
3	CPU	Central Processing Unit
4	ETL	Extract - Transform - Load
5	HDFS	Hadoop Distributed File System
6	IMDB	In-memory Databases
7	IoT	Internet of Things
8	ML	Machine Learning
9	OLAP	On-line analytical processing
10	OLTP	On-line transactional processing
11	SQL	Structured Query Language
12	SSIS	SQL Server Integration Services
13	SSMS	SQL Server Management Studio

DANH MỤC HÌNH VẼ

Hình 2.1 Hệ thống Hadoop Distributed File System	8
Hình 2.2 Mô hình YARN	9
Hình 2.3 Mô hình MapReduce.....	10
Hình 2.4 Xây dựng Spark trên Hadoop.....	11
Hình 2.5 Apache Spark Core.....	12
Hình 3.1 Những dịch vụ của Azure Data Factory	21
Hình 3.2 Sơ đồ giải pháp thiết kế.....	24
Hình 3.3 Sơ đồ khói hệ thống	25
Hình 3.4 Đăng ký Microsoft Azure.....	25
Hình 3.5 Các dịch vụ cần sử dụng trên Azure.....	26
Hình 3.6 Lấy dữ liệu dân số	27
Hình 3.7 Pipeline lấy dữ liệu dân số	27
Hình 3.8 Lấy dữ liệu dịch bệnh từ trang web ECDC.....	28
Hình 3.9 Kết quả thu thập dữ liệu ECDC	28
Hình 3.10 Kết quả khói Thu thập dữ liệu.....	29
Hình 3.11 Data Flow chuyển đổi dữ liệu ca nhiễm và tử vong.....	30
Hình 3.12 Bộ dữ liệu sau khi hoàn thành Data Flow (1)	30
Hình 3.13 Chuyển đổi bộ dữ liệu raw	31
Hình 3.14 Data Flow (2)	32
Hình 3.15 Tạo xác thực quyền truy cập cho HD Insight	33
Hình 3.16 Add role assignment	33
Hình 3.17 Tạo cụm HD Insight – Hadoop	34
Hình 3.18 Giao diện chính của HD Insight.....	35
Hình 3.19 Các node trong cụm HD Insight.....	35
Hình 3.20 Hive trên Hadoop - HD Insight.....	36
Hình 3.21 Kết quả vận hành pipeline chạy file (.hql) trên HD Insight.....	36
Hình 3.22 Kết quả thu được sau quá trình chuyển đổi qua HD Insight.....	37

Hình 3.23 Spark Cluster trên Databricks.....	37
Hình 3.24 Mount dữ liệu với Python.....	38
Hình 3.25 Bộ dữ liệu dân số gốc	38
Hình 3.26 Chuyển đổi bộ dữ liệu dân số theo độ tuổi.....	39
Hình 3.27 Pipeline chuyển đổi dữ liệu dân số với Databricks	39
Hình 3.28 Python Notebook chạy trên Cluster Databricks	40
Hình 3.29 Kết quả quá trình chuyển đổi.....	40
Hình 3.30 Cơ sở dữ liệu SQL trên Azure	41
Hình 3.31 SQL Query Editor.....	41
Hình 3.32 Kết quả tạo các bảng tạm trên cơ sở dữ liệu.....	42
Hình 3.33 Kết quả sao chép dữ liệu vào SQL Database.....	42
Hình 3.34 Kết quả quá trình sao chép dữ liệu vào Database.....	43
Hình 3.35 Giao diện Power BI	43
Hình 3.36 Kết nối Power BI với SQL Database.....	44
Hình 3.37 Dữ liệu được đã được lấy vào Power BI	44
Hình 3.38 Trực quan hóa dữ liệu với Power BI	45
Hình 3.39 Giám sát các luồng dữ liệu với chức năng “Monitor”	46
Hình 4.1 Những pipeline đã xây dựng được	47
Hình 4.2 Những bộ dataset trước và sau khi xử lý dữ liệu.....	48
Hình 4.3 Những Data Flow đã xây dựng.....	48
Hình 4.4 Số ca nhiễm và tử vong ở EU/EEA & UK từ 2/3/2020 đến 25/10/2020.....	49
Hình 4.5 Số ca nhiễm và tử vong ở UK, France & Germany từ 24/8/2020 đến 25/10/2020	49
Hình 4.6 Tỉ lệ tiêm Vaccine ở khu vực EU/EEA & UK	50
Hình 4.7 Biểu đồ quản lý vận hành luồng dữ liệu.....	51

DANH MỤC BẢNG BIỂU

Bảng 1.1 Kế hoạch thực hiện đồ án 4

TÓM TẮT ĐỒ ÁN

Xử lý dữ liệu vẫn luôn là bài toán quan trọng bởi những ứng dụng thực tiễn mà dữ liệu mang lại trong cuộc sống như kinh tế, khoa học công nghệ,... Xử lý dữ liệu được coi là bài toán cơ bản, nền tảng cho những bài toán phân tích dữ liệu, học máy sử dụng nguồn dữ liệu đó. Vì vậy, xử lý và quản lý luồng dữ liệu trở thành vấn đề được nhiều cá nhân, doanh nghiệp quan tâm, nhằm tận dụng tối đa tiềm năng dữ liệu mang lại và tiết kiệm chi phí nhất có thể. Với sự phát triển của điện toán đám mây, lưu trữ đám mây dần thay thế cho các dạng lưu trữ trên máy chủ cục bộ truyền thống, trở thành công cụ hỗ trợ mạnh mẽ cho bài toán lưu trữ, xử lý và phân tích dữ liệu. Một trong số những dịch vụ điện toán đám mây nổi tiếng có thể kể đến chính là Microsoft Azure Data Factory. Vì vậy, việc đưa ra ứng dụng của Azure Data Factory vào bài toán dữ liệu là cần thiết.

Trong những năm gần đây, Azure Data Factory được coi như một giải pháp điện toán đám mây mạnh mẽ, giải quyết được những nhược điểm còn tồn đọng của phương pháp lưu trữ truyền thống. Với ưu điểm có thể quản lý dữ liệu không cần máy chủ và khả năng làm việc với lượng dữ liệu lớn. Nhờ đó, Azure giúp giảm thiểu đáng kể chi phí cho việc xây dựng máy chủ, nhân lực và tận dụng tối đa được nguồn lợi từ dữ liệu. Với những đặc điểm nổi bật của Azure Data Factory, đồ án đưa ra giải pháp xây dựng đường đi, quản lý các luồng dữ liệu sử dụng các dịch vụ mà Azure cung cấp. Kết quả thực nghiệm cho thấy rằng giải pháp tôi đưa ra đã giải quyết được bài toán xử lý, quản lý và phân tích dữ liệu. Ngoài ra, dựa vào phản hồi của các luồng dữ liệu có thể thấy dữ liệu được quản lý triệt để theo thời gian thực.

ABSTRACT

Data processing is still an important problem because of the practical applications that data brings in life such as economics, science and technology,... Data processing is considered as a fundamental problem, the foundation of the problem. foundation for problem analysis data, machine learning uses that source data. Therefore, the process of processing and managing data flows becomes a matter of interest to many individuals and businesses, making the most of the data brought and saving costs as much as possible. With the development of the cloud, the storage cloud replaces the form storage on the server communication system, becoming a powerful support tool for the problem of storing, processing and analyzing data instead. Because. Some famous cloud computing services can be said to be Microsoft Azure Data Factory. Therefore, the Azure Data Factory application to input the problem data is necessary.

In recent years, Azure Data Factory is considered as a powerful cloud computing solution that solves the remaining shortcomings of traditional storage methods. With the advantage of being able to manage data without a server and the ability to work with large amounts of data. As a result, Azure significantly reduces the cost of building servers, human resources, and makes the most of data. With the outstanding features of Azure Data Factory, the project offers a solution to build paths and manage data flows using the services that Azure provides. Experimental results show that my solution has solved the problem of data processing, management and analysis. In addition, based on the feedback of the data streams, it is possible to see that the data is thoroughly managed in real time.

PHẦN MỞ ĐẦU

Trong những năm gần đây, dữ liệu cho thấy sự bùng nổ đáng kinh ngạc trong mọi lĩnh vực như y tế, tài chính, khoa học công nghệ,... Nhiều công nghệ xây dựng máy chủ lưu trữ cục bộ đã được đề xuất, xây dựng nhằm giải quyết bài toán dữ liệu. Tuy vậy, việc xây dựng các hệ thống máy chủ này không hề đơn giản cũng như vô cùng tốn kém đối với lượng dữ liệu lớn. Đối mặt với vấn đề này, các giải pháp điện toán đám mây đã nổi lên như một cuộc cách mạng mới. Azure Data Factory được biết đến như một công cụ đám mây mạnh mẽ để giải quyết các bài toán về dữ liệu lớn. Do đó, tôi đưa ra đồ án này để tìm hiểu và ứng dụng Azure Data Factory vào bài toán dữ liệu thực tế với bộ dữ liệu Covid-19. Để trình bày cụ thể được nội dung và kết quả thu được, đồ án được bố trí gồm 4 chương như sau:

Chương 1: Trình bày cụ thể về vấn đề đang phải đối mặt và các phương án giải quyết.

Chương 2: Cơ sở lý thuyết về dữ liệu lớn và Microsoft Azure.

Chương 3: Triển khai xây dựng báo cáo Covid-19 sử dụng dịch vụ của Microsoft Azure.

Chương 4: Kết quả thu được và những bàn luận về các kết quả thu được.

CHƯƠNG 1. ĐẶT VẤN ĐỀ

Chương này trình bày các vấn đề về dữ liệu lớn và ứng dụng của điện toán đám mây vào làm việc với dữ liệu từ đó dẫn dắt đến việc lựa chọn đề tài của nhóm. Bên cạnh đó, mục tiêu, phạm vi, kết quả dự kiến của đề tài cùng với kế hoạch thực hiện đồ án cũng được nêu ra trong chương này.

1.1 Đặt vấn đề

Phương pháp quản lý dữ liệu đã phát triển qua nhiều năm với nhiều lần lặp lại như Xử lý giao dịch trực tuyến (On-line transactional processing - OLTP), Kho dữ liệu và Marts dữ liệu, Xử lý phân tích trực tuyến (On-line analytical processing - OLAP), Hồ dữ liệu và cuối cùng là khái niệm Data Lakehouse đang trở nên phổ biến trong kỉ nguyên đám mây. Mặc dù có nhiều loại nền tảng và sản phẩm quản lý dữ liệu khác nhau và các phương pháp tiếp cận đa dạng như nhau để điều chỉnh các công cụ và công nghệ này nhằm quản lý dữ liệu, nhưng một câu tạo trung tâm của tất cả những điều này là các đường ống dẫn dữ liệu thông minh. Khi dữ liệu phát triển vượt ra ngoài giới hạn của các mô hình dữ liệu có cấu trúc và quan hệ, nhu cầu ngày càng tăng đối với các đường ống dữ liệu theo hướng siêu dữ liệu sẽ đặc biệt hữu ích để xử lý sự đa dạng và khối lượng dữ liệu thường được tìm thấy trên các hồ dữ liệu trên đám mây.

Một thách thức chính với các đường ống dữ liệu biểu hiện theo thời gian là việc phát triển đường ống dữ liệu bắt đầu ở mức độ khiêm tốn với kết nối điểm-điểm từ nguồn đến đích. Khi quy mô dữ liệu phát triển và lược đồ của các đối tượng dữ liệu thay đổi theo thời gian, việc phù hợp với tốc độ phát triển các đường ống dữ liệu điểm-điểm mới cũng như duy trì các đường ống dữ liệu hiện có ngày càng trở nên khó khăn và kém hiệu quả. Ít nhất, các đường ống dữ liệu chỉ xử lý việc nhập dữ liệu từ các đối tượng dữ liệu mới hơn có thể được tạo theo hướng siêu dữ liệu như một điểm khởi đầu để giảm số lượng đường ống dữ liệu giống hệt nhau chỉ khác nhau về nguồn và đích. Đám mây Azure hỗ trợ nhiều sản phẩm và nền tảng quản lý dữ liệu khác nhau, có thể là nguồn hoặc đích từ quan điểm đường ống dữ liệu. Azure Data Factory (ADF) là dịch vụ chính từ Azure để xây dựng đường ống dữ liệu và chúng ta sẽ xem cách tạo đường ống theo hướng siêu dữ liệu bằng cách sử dụng nó. Trong phần tiếp theo của báo cáo sẽ đưa ra mục tiêu, phạm vi nghiên cứu cụ thể của đề tài.

1.2 Mục tiêu nghiên cứu, đối tượng và phạm vi đề tài

Như đã nêu trong các phần trước, chi phí và tài nguyên đáp ứng cho sự bùng nổ dữ liệu là một vấn đề lớn đối với một số doanh nghiệp. Trong đề tài này, tôi xin đưa ra một giải pháp cho vấn đề nêu trên. Đó là ứng dụng dịch vụ đám mây Microsoft Azure với nhiều ưu điểm về mặt chi phí, tốc độ, đáp ứng được khối lượng dữ liệu lớn, đa dạng và sản sinh nhanh. Hiện nay, điện toán đám mây đang là một lĩnh vực nhận được sự quan tâm của nhiều doanh nghiệp và được các doanh nghiệp đầu tư rất mạnh để giải quyết bài toán dữ liệu của doanh nghiệp.

Mục tiêu chính của đồ án lần này là hướng đến việc xây dựng đường ống dữ liệu và quản lý luồng dữ liệu trên Azure Data Factory. Bên cạnh đó, tôi thực hiện xây dựng một báo cáo trực quan hóa những dữ liệu được lấy trực tiếp từ cơ sở dữ liệu đã tạo được qua đường ống dữ liệu ở phần trên. Cụ thể hơn, bài toán mà đề tài hướng đến là bài toán thu thập, xử lý, trực quan hóa và quản lý dữ liệu.

1.3 Ý nghĩa của đề tài

Đề tài này được đưa ra như một giải pháp cho vấn đề thu thập, xử lý và phân tích dữ liệu. Các luồng dữ liệu sẽ được hoạt động theo một pipeline đã định nghĩa từ trước một cách hoàn toàn tự động thông qua Azure Data Factory. Điều này sẽ giúp tiết kiệm được thời gian, nhân lực cũng như chi phí cho bài toán dữ liệu. Mặc dù các pipeline với chỉ thực sự tối ưu với các định dạng dữ liệu đầu vào nhất định (.csv), tuy vậy, tôi sẽ cố gắng để làm cho pipeline hoạt động được với nhiều định dạng dữ liệu đầu vào hơn. Về cơ bản, đồ án này sẽ hữu ích trong thời điểm hiện tại do việc ứng dụng điện toán đám mây vào xử lý bài toán dữ liệu ngày càng được quan tâm. Tuy nhiên, trong tương lai, khi bùng nổ dữ liệu cũng như tốc độ phát triển của công nghệ ngày càng nhanh thì hệ thống cũng cần phải phát triển thêm để có thể bắt kịp được công nghệ.

1.4 Những nghiên cứu gần đây

Gần đây, Azure đang nhận được nhiều sự quan tâm từ các cá nhân, doanh nghiệp bởi những điểm nổi bật của nó, vì vậy, Azure Data Factory đang trở thành một giải pháp tốt cho các bài toán dữ liệu hiện nay. Do đó, nhiều bài báo, hướng dẫn được đưa ra.

Microsoft trực tiếp cung cấp một số bài báo khoa học (Whitepapers) để mọi người có thể hiểu sâu hơn về Azure. Bài báo [1] mô tả cách Azure Data Factory có thể cho phép bạn xây dựng một kho dữ liệu hiện đại, cho phép phân tích nâng cao để thúc đẩy các ứng dụng SaaS thông minh và nâng các gói dịch vụ tích hợp máy chủ Structured Query Language (SQL) của bạn trên Azure. Bài báo [2] đề cập đến sự phức tạp của việc di chuyển hàng chục TB dữ liệu từ kho dữ liệu quan hệ tại chỗ hiện có (ví dụ: Netezza, Oracle, Teradata, máy chủ SQL) sang Azure (ví dụ: Blob Storage hoặc Azure Data Lake Storage) bằng Azure Data Factory. Những thách thức và thực tiễn tốt nhất được minh họa xung quanh khả năng phục hồi, hiệu suất, khả năng mở rộng, quản lý và bảo mật cho hành trình nhập dữ liệu lớn vào Azure của Azure Data Factory. Bài báo [3] trình bày một số phương pháp hay nhất về tích hợp và triển khai liên tục của Azure Data Factory. Bài báo [4] tóm tắt sự hỗ trợ hiện tại của Azure Data Factory về tích hợp dữ liệu SAP, bao gồm kịch bản mục tiêu, các tùy chọn kết nối SAP và so sánh cho các yêu cầu khác nhau và giới thiệu về từng đầu nối SAP trong Data Factory. Bài báo [5] đề cập đến lý do tại sao bạn muốn di chuyển khối lượng công việc SSIS (SQL Server Integration Services) hiện có của mình sang Nhà máy Dữ liệu Azure và giải quyết các vấn đề và cân nhắc phổ biến. Sau đó, chúng tôi sẽ hướng dẫn bạn qua các chi tiết kỹ thuật của việc tạo Azure-SSIS IR và sau đó hướng dẫn bạn cách tải lên, thực thi và giám sát các gói của bạn thông qua Azure Data Factory bằng cách sử dụng các công cụ mà bạn có thể quen thuộc như SQL Server Management Studio (SSMS).

1.5 Kết quả dự kiến

Sau khi hoàn tất quá trình xây dựng hệ thống, dữ liệu sẽ được trực quan hóa dưới dạng một bản báo cáo bao gồm các biểu đồ thể hiện các số liệu cụ thể. Các luồng dữ liệu cũng được quản lý dựa trên các thông số phản hồi của hệ thống. Bên cạnh đó, chúng tôi hi vọng hệ thống sẽ đáp ứng được các yêu cầu về tốc độ, khối lượng và độ đa dạng của dữ liệu đầu vào cũng như đầu ra.

1.6 Kế hoạch thực hiện đề tài

Như trên bìa của đồ án, người thực hiện đồ án là tôi, Dương Công Kiên, dưới sự hướng dẫn của thầy Hoàng Mạnh Thắng. Sự phân chia nội dung, kế hoạch nghiên cứu được trình bày trong Bảng 1.1.

Bảng 1.1 Kế hoạch thực hiện đồ án

STT	Nội dung nghiên cứu	Người phụ trách
1	Tìm hiểu về Azure Data Factory	Dương Công Kiên
2	Tìm hiểu về Azure Data Lake Storage Gen 2	Dương Công Kiên
3	Tìm hiểu về Hadoop, Spark. Pipeline	Dương Công Kiên
4	Tìm hiểu về bộ dữ liệu Covid-19	Dương Công Kiên
5	Áp dụng phương pháp đã tìm hiểu vào xây dựng luồng dữ liệu	Dương Công Kiên
6	Xây dựng báo cáo dựa trên dữ liệu đã chuyển đổi	Dương Công Kiên

1.7 Kết luận chương

Trong chương đầu tiên này, đồ án đã trình bày vấn đề về dữ liệu và giải pháp điện toán đám mây cùng với đó là mục tiêu nghiên cứu, đối tượng và phạm vi đề tài. Ngoài ra, ý nghĩa của đề tài và nội dung nghiên cứu cũng được đưa ra trong chương này. Chương tiếp theo sẽ trình bày về cơ sở lý thuyết, các công nghệ và ứng dụng vào thực tế.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

Trong chương này, các cơ sở lý thuyết được sử dụng cho đề tài sẽ được trình bày một cách ngắn gọn. Nội dung sẽ tập trung vào các công nghệ hỗ trợ cho bài toán dữ liệu. Đôi với mỗi công nghệ được đề cập, tôi sẽ đưa ra vai trò và ứng dụng của nó vào bài toán thực tế.

2.1 Tổng quan về Big Data và Microsoft Azure

2.1.1 Dữ liệu lớn và phân tích dữ liệu

- Dữ liệu lớn:

Dữ liệu lớn là dữ liệu có chứa nhiều loại dữ liệu, khối lượng và tốc độ dữ liệu ngày càng lớn. Hay dữ liệu lớn là những tập dữ liệu lớn hơn, phức tạp hơn, đặc biệt là từ các nguồn dữ liệu mới. Các tập dữ liệu này quá lớn đến nỗi phần mềm xử lý dữ liệu truyền thống không thể quản lý chúng. Nhưng khối lượng dữ liệu khổng lồ này có thể được sử dụng để giải quyết các vấn đề kinh doanh mà trước đây bạn không thể giải quyết. Ba đặc điểm này của dữ liệu lớn còn được gọi là “three Vs” hay “3 V”. Gồm có:

- **Volume** (khối lượng): là số lượng dữ liệu quan trọng. Với dữ liệu lớn, bạn sẽ phải xử lý khối lượng lớn dữ liệu phi cấu trúc, mật độ thấp. Đây có thể là dữ liệu có giá trị không xác định, chẳng hạn như nguồn cấp dữ liệu Facebook, dòng nhập chuột trên trang web hoặc ứng dụng dành cho thiết bị di động hoặc thiết bị hỗ trợ cảm biến. Đôi với một số tổ chức, đây có thể là hàng chục terabyte dữ liệu. Đôi với những tổ chức khác, nó có thể là hàng trăm petabyte.
- **Velocity** (tốc độ): là tốc độ nhanh chóng mà dữ liệu được nhận và (có thể) được thực hiện. Thông thường, tốc độ cao nhất của luồng dữ liệu trực tiếp vào bộ nhớ so với được ghi vào đĩa. Một số sản phẩm thông minh hỗ trợ internet hoạt động trong thời gian thực hoặc gần thời gian thực và sẽ yêu cầu đánh giá và hành động theo thời gian thực.
- **Variety** (dự đa dạng): Sự đa dạng đề cập đến nhiều loại dữ liệu có sẵn. Các kiểu dữ liệu truyền thống được cấu trúc và nằm gọn trong cơ sở dữ liệu quan hệ. Với sự gia tăng của dữ liệu lớn, dữ liệu xuất hiện trong các kiểu dữ liệu phi cấu trúc mới. Các loại dữ liệu không có cấu trúc và bán cấu trúc, chẳng hạn như văn

bản, âm thanh và video, yêu cầu xử lý trước bổ sung để tìm ra ý nghĩa và hỗ trợ siêu dữ liệu.

Ngoài “3 V” nói trên, gần đây người ta còn quan tâm đến 2 đặc tính khác của dữ liệu lớn đó là: (giá trị) và tính xác thực. Dữ liệu có giá trị nội tại. Nhưng nó không có ích lợi gì cho đến khi giá trị đó được phát hiện. Điều quan trọng không kém: Dữ liệu của bạn trung thực đến mức nào — và bạn có thể dựa vào nó ở mức độ nào?

- **Value** (giá trị): Dữ liệu có giá trị nội tại. Nhưng nó không có ích lợi gì cho đến khi giá trị đó được phát hiện.
- **Veracity** (tính xác thực): Điều quan trọng không kém chính là: Dữ liệu của bạn trung thực đến mức nào và bạn có thể dựa vào nó ở mức độ nào.

Ngày nay, dữ liệu lớn đã trở thành tài sản. Hãy nghĩ về một số công ty công nghệ lớn nhất thế giới, một phần lớn giá trị mà họ cung cấp đến từ dữ liệu của họ. Dữ liệu này được họ liên tục phân tích để tạo ra hiệu quả hơn và phát triển các sản phẩm mới. Những đột phá công nghệ gần đây đã làm giảm chi phí lưu trữ và tính toán dữ liệu theo cấp số nhân, khiến việc lưu trữ nhiều dữ liệu trở nên dễ dàng và ít tốn kém hơn bao giờ hết. Với khối lượng dữ liệu lớn ngày càng tăng, giá thành rẻ hơn và dễ tiếp cận hơn, bạn có thể đưa ra các quyết định kinh doanh chính xác và chính xác hơn. Tìm kiếm giá trị trong dữ liệu lớn không chỉ là phân tích nó (đó là một lợi ích hoàn toàn khác). Đó là toàn bộ quá trình khám phá yêu cầu các nhà phân tích sâu sắc, người dùng doanh nghiệp và giám đốc điều hành đặt câu hỏi phù hợp, nhận ra các mẫu, đưa ra các giả định sáng suốt và dự đoán hành vi.

- Phân tích dữ liệu:

Điều thực sự mang lại giá trị từ tất cả các tổ chức dữ liệu lớn đang thu thập là phân tích dữ liệu (Analytics). Nếu không phân tích, nó chỉ là một bó dữ liệu với việc sử dụng hạn chế trong kinh doanh. Bằng cách áp dụng phân tích vào dữ liệu lớn, các công ty có thể nhận thấy những lợi ích như tăng doanh thu, dịch vụ khách hàng được cải thiện, hiệu quả cao hơn và tăng khả năng cạnh tranh. Phân tích dữ liệu liên quan đến việc kiểm tra bộ dữ liệu để thu thập thông tin chi tiết hoặc rút ra kết luận về những gì bao gồm trong đó, chẳng hạn các xu hướng và dự đoán về hoạt động trong tương lai.

Bằng cách phân tích dữ liệu, các tổ chức có thể đưa ra những quyết định kinh doanh tốt hơn như thời gian và địa điểm nên chạy chiến dịch tiếp thị hoặc giới thiệu sản phẩm hoặc dịch vụ mới. Việc phân tích có thể tham khảo các ứng dụng kinh doanh thông minh hay tiên tiến hơn, phân tích dự đoán như ứng dụng được các tổ chức khoa học sử dụng. Loại phân tích dữ liệu cao cấp nhất là data mining, nơi các nhà phân tích đánh giá bộ dữ liệu lớn để xác định mối quan hệ, mô hình và xu hướng. Phân tích dữ liệu có thể bao gồm phân tích dữ liệu thăm dò (để xác định các mẫu và mối quan hệ trong dữ liệu) và phân tích dữ liệu xác nhận (áp dụng các kỹ thuật thống kê để tìm ra giả thiết về bộ dữ liệu đó có đúng hay không). Một mảng khác là phân tích dữ liệu định lượng (hoặc phân tích dữ liệu số có các biến có thể so sánh theo thống kê) so với phân tích dữ liệu định tính (tập trung vào các dữ liệu không phải dữ liệu cá nhân như video, hình ảnh và văn bản).

2.1.2 Các công nghệ đặc biệt hỗ trợ Big Data

Để làm việc được với dữ liệu lớn, các doanh nghiệp cần phải có cơ sở hạ tầng để thu thập và lưu trữ dữ liệu. Đồng thời họ cần phải cung cấp quyền truy cập và đảm bảo thông tin khi vận chuyển, phân tích dữ liệu. Ở cấp độ cao, hệ thống dữ liệu lớn gồm có hệ thống lưu trữ, các máy chủ, phần mềm quản lý và phân tích dữ liệu, phần mềm kinh doanh thông minh (BI), các ứng dụng dữ liệu lớn. Phần lớn cơ sở hạ tầng sẽ được các công ty đầu tư trung tâm dữ liệu xây dựng cục bộ. Tuy nhiên, khi dữ liệu ngày càng lớn thì càng nhiều tổ chức chuyển dần sang các dịch vụ điện toán đám mây để xử lý bài toán dữ liệu lớn của họ.

Ngoài những cơ sở hạ tầng công nghệ thông tin cơ bản được sử dụng cho dữ liệu nói chung, có một số công nghệ đặc biệt dành cho dữ liệu lớn mà cơ sở hạ tầng công nghệ thông tin nên triển khai. Có thể kể tên các hệ sinh thái dưới đây:

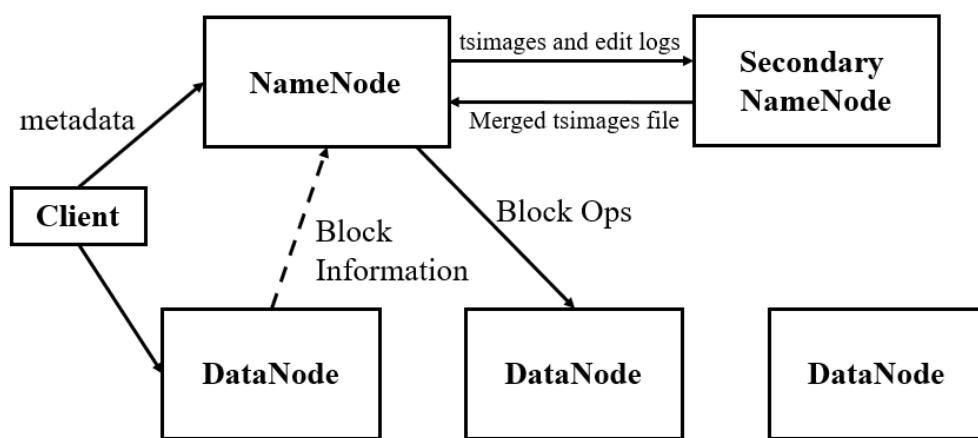
- **Hệ sinh thái Hadoop (Hadoop Ecosystem):**

Hadoop là một trong những công nghệ liên quan chặt chẽ với dữ liệu lớn. Apache Hadoop phát triển phần mềm mã nguồn mở cho máy tính có khả năng mở rộng và phân phối dữ liệu.

Hadoop là một framework cho phép xử lý phân phối các bộ dữ liệu lớn trên cụm máy tính sử dụng mô hình lập trình đơn giản. Nó được thiết kế để mở rộng từ

một máy chủ duy nhất sang hàng ngàn máy con, mỗi máy cung cấp tính toán và lưu trữ cục bộ. Hệ sinh thái Hadoop đặc trưng bởi các yếu tố sau đây:

- Hadoop Distributed File System (HDFS – Hệ thống lưu trữ dữ liệu phân tán): là nền tảng của Hadoop, do đó là một thành phần rất quan trọng trong hệ sinh thái Hadoop. Đây là một phần mềm Java cung cấp nhiều tính năng như khả năng mở rộng, tính sẵn sàng cao, khả năng chịu lỗi, hiệu quả về chi phí,.v.v. Nó cung cấp khả năng lưu trữ dữ liệu phân tán mạnh mẽ cho Hadoop. Các thành phần chính của HDFS gồm có: DataNode, NameNode và Secondary NameNode.

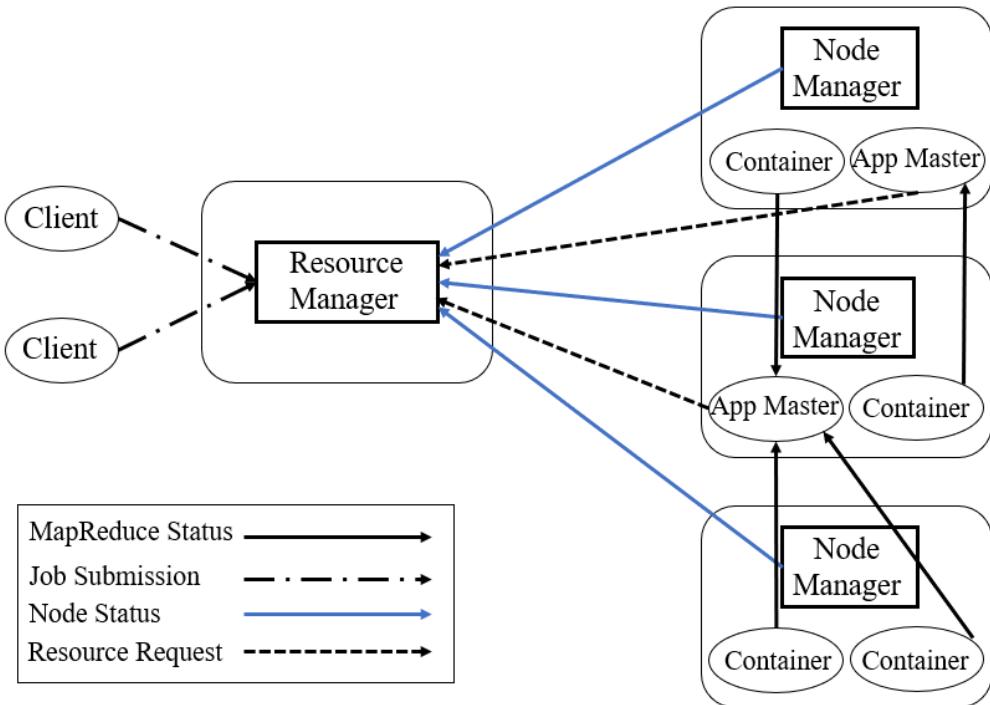


Hình 2.1 Hệ thống Hadoop Distributed File System

Trong Hình 2.1 mô tả quá trình hoạt động của kiến trúc Hadoop Distributed File System.

- + DataNode là các nút lưu trữ dữ liệu thực tế. HDFS lưu trữ dữ liệu theo cách phân tán. Dữ liệu được chia thành các khối.
- + NameNode chịu trách nhiệm quản lý không gian hệ thống file, kiểm soát quyền truy cập của client. Ngoài ra, nó còn thực hiện các tác vụ như mở, đóng và đặt tên cho các tệp và thư mục. NameNode có hai tệp chính: FSImage và Edits log.
- + Secondary NameNode: Nếu NameNode không khởi động lại trong một thời gian dài, kích thước của Edits log sẽ tăng lên. Điều này sẽ làm tăng thời gian chết của cụm khi khởi động lại NameNode. Trong trường hợp này, Secondary NameNode sẽ nhận Edits log theo chu kỳ và cập nhật FSImage mới trên NameNode.

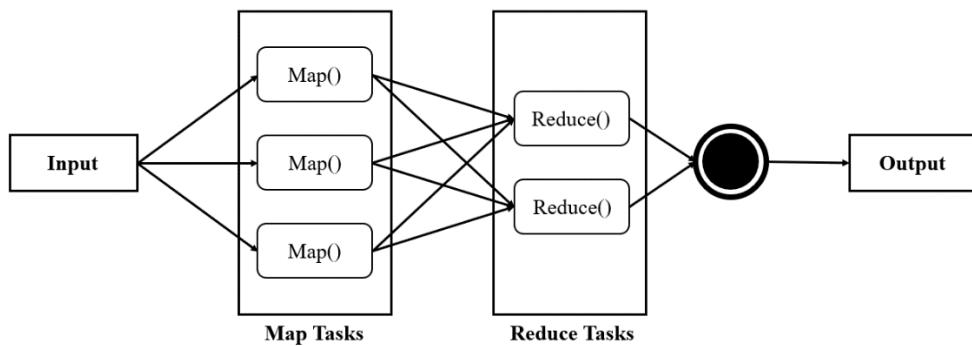
- Hadoop YARN: một framework cho việc thiết lập lịch làm việc và quản lý tài nguyên gồm có các thành phần: Node Manager và Resource Manager.



Hình 2.2 Mô hình YARN

Hình 2.2 Mô tả quá trình hoạt động của YARN

- + Node Manager: Quản lí các nút riêng lẻ trong một cụm Hadoop. Nó giám sát việc sử dụng tài nguyên như CPU, bộ nhớ,... của nút cục bộ và tương tự với Resource Manager.
- + Resource Manager: chịu trách nhiệm theo dõi các tài nguyên trong cụm và lên lịch các tác vụ như công việc Map-Reduce.
- Hadoop MapReduce: thành phần xử lý dữ liệu của Hadoop. Nó áp dụng tính toán trên các tập dữ liệu đồng thời do đó cải thiện hiệu suất. MapReduce hoạt động theo hai giai đoạn:
 - + Map: Giai đoạn này nhận đầu vào là các cặp key-value và tạo ra đầu ra dưới dạng các cặp key-value. Giai đoạn này xử lý dữ liệu để cung cấp cho giai đoạn tiếp theo.
 - + Reduce: Sắp xếp các cặp key-value sau đó đưa vào xử lý, tính toán tổng hợp.



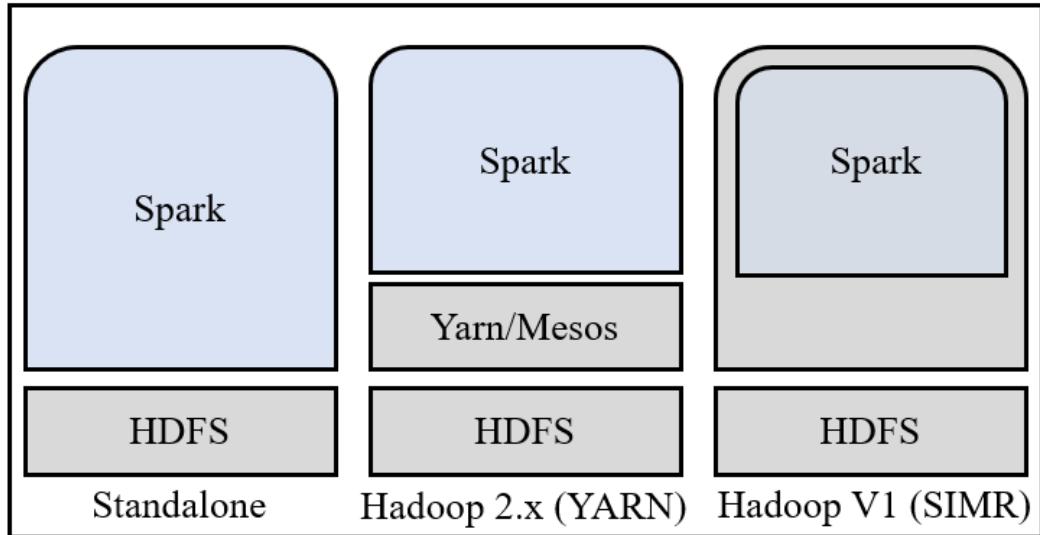
Hình 2.3 Mô hình MapReduce

Hình 2.3 mô tả quá trình hoạt động của mô hình Map-Reduce.

- **Apache Spark:**

Apache Spark là công nghệ điện toán cụm (cluster-computing) nhanh như chớp, được thiết kế để tính toán nhanh chóng. Nó dựa trên Hadoop MapReduce và nó mở rộng mô hình MapReduce để sử dụng hiệu quả nó cho nhiều loại tính toán hơn, bao gồm các truy vấn tương tác và xử lý luồng. Tính năng chính của Spark là tính toán cụm trong bộ nhớ giúp tăng tốc độ xử lý của ứng dụng. Spark được thiết kế để bao gồm một loạt các khôi lượng công việc như ứng dụng hàng loạt, thuật toán lặp lại, truy vấn tương tác và phát trực tuyến. Ngoài việc hỗ trợ tất cả khôi lượng công việc này trong một hệ thống tương ứng, nó làm giảm gánh nặng quản lý trong việc duy trì các công cụ riêng biệt.

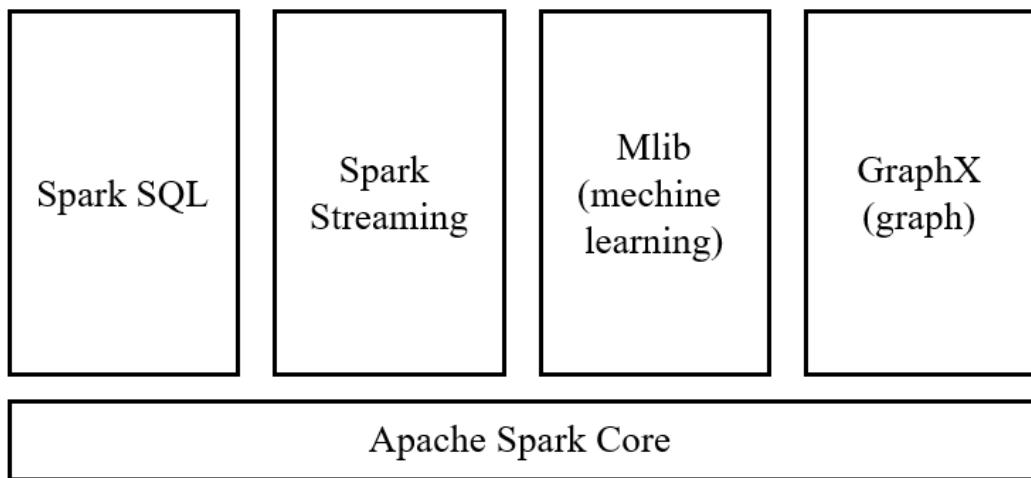
- Xây dựng Spark trên Hadoop: Các hình thức triển khai được mô tả trong Hình 2.4.



Hình 2.4 Xây dựng Spark trên Hadoop

- **Độc lập (Standalone):** Spark Triển khai độc lập có nghĩa là Spark chiếm vị trí trên đầu HDFS (Hệ thống tệp phân tán Hadoop) và không gian được phân bổ cho HDFS một cách rõ ràng. Ở đây, Spark và MapReduce sẽ chạy đồng thời với nhau để bao gồm tất cả các công việc tia lửa trên cụm.
- **Hadoop Yarn:** Việc triển khai Hadoop Yarn có nghĩa là đơn giản, spark chạy trên Yarn mà không cần cài đặt trước hoặc truy cập root. Nó giúp tích hợp Spark vào hệ sinh thái Hadoop hoặc ngăn xếp Hadoop. Nó cho phép các thành phần khác chạy trên đầu ngăn xếp.
- **Spark in MapReduce (SIMR) -** Spark trong MapReduce được sử dụng để khởi chạy công việc spark ngoài việc triển khai độc lập. Với SIMR, người dùng có thể khởi động Spark và sử dụng trình bao của nó mà không cần bất kỳ quyền truy cập quản trị nào.
 - Spark có những đặc tính chính sau:
- **Tốc độ (speed):** Spark giúp chạy ứng dụng trong cụm Hadoop, nhanh hơn tới 100 lần trong bộ nhớ và nhanh hơn 10 lần khi chạy trên đĩa. Điều này có thể thực hiện được bằng cách giảm số lượng các thao tác đọc / ghi vào đĩa. Nó lưu trữ dữ liệu xử lý trung gian trong bộ nhớ.
- **Hỗ trợ nhiều ngôn ngữ:** Spark cung cấp các API tích hợp sẵn bằng Java, Scala hoặc Python. Do đó, bạn có thể viết ứng dụng bằng các ngôn ngữ khác nhau. Spark đưa ra 80 toán tử cấp cao để truy vấn tương tác.

- **Phân tích nâng cao** (Advance Analytics): Spark không chỉ hỗ trợ “Map” và “reduce”. Nó cũng hỗ trợ các truy vấn SQL, dữ liệu truyền trực tuyến, máy học (ML) và các thuật toán đồ thị.



Hình 2.5 Apache Spark Core

- Hình 2.5 thể hiện những thư viện sẵn có trong hệ sinh thái Apache Spark Core:
 - + Apache Spark Core: Spark Core là công cụ thực thi chung cơ bản cho nền tảng tia lửa mà tất cả các chức năng khác đều được xây dựng dựa trên. Nó cung cấp tính toán trong bộ nhớ và tham chiếu bộ dữ liệu trong hệ thống lưu trữ bên ngoài.
 - + Spark SQL: Spark SQL là một thành phần trên Spark Core giới thiệu một phần trùu tượng hóa dữ liệu mới được gọi là Schema RDD, cung cấp hỗ trợ cho dữ liệu có cấu trúc và bán cấu trúc.
 - + Spark Streaming: Spark Streaming tận dụng khả năng lập lịch nhanh chóng của Spark Core để thực hiện phân tích luồng. Nó nhập dữ liệu trong các lô nhỏ và thực hiện các phép biến đổi RDD (Tập dữ liệu phân tán có khả năng phục hồi) trên các lô dữ liệu nhỏ đó.
 - + MLlib (Thư viện học máy) MLlib là một khung công tác học máy phân tán trên Spark vì kiến trúc Spark dựa trên bộ nhớ phân tán. Theo điểm chuẩn, nó được thực hiện bởi các nhà phát triển MLlib dựa trên việc triển khai Phương diện Ít nhất Luân phiên (ALS). Spark MLlib nhanh gấp 9 lần so với phiên bản dựa trên đĩa Hadoop của Apache Mahout (trước khi Mahout có được giao diện Spark).
 - + GraphX: GraphX là một khung xử lý đồ thị phân tán trên Spark. Nó cung cấp một API để thể hiện tính toán đồ thị có thể lập mô hình các đồ thị do người dùng xác

định bằng cách sử dụng API trùu tượng Pregel. Nó cũng cung cấp thời gian chạy được tối ưu hóa cho sự trùu tượng này.

- **Data lakes:** Data lakes là các kho lưu trữ chứa khối lượng dữ liệu thô rất lớn ở định dạng gốc của nó cho đến khi những người dùng doanh nghiệp cần dữ liệu. Các yếu tố giúp tăng trưởng data lakes là những phong trào kỹ thuật số và sự phát triển của IoT. Các data lakes được thiết kế để giúp người dùng dễ dàng truy cập vào một lượng lớn dữ liệu khi có nhu cầu.
- **Cơ sở dữ liệu NoSQL:** Các cơ sở dữ liệu SQL thông thường được thiết kế cho các transaction đáng tin cậy và các truy vấn ngẫu nhiên. Nhưng chúng có những hạn chế như giản đồ cứng nhắc làm cho chúng không phù hợp với một số loại ứng dụng. Cơ sở dữ liệu NoSQL nêu ra những hạn chế, và lưu trữ và quản lý dữ liệu theo những cách cho phép tốc độ hoạt động cao và sự linh hoạt tuyệt vời. Nhiều cơ sở dữ liệu đã được phát triển bởi các công ty để tìm cách tốt hơn để lưu trữ nội dung hoặc xử lý dữ liệu cho các trang web lớn. Không giống như các cơ sở dữ liệu SQL, nhiều cơ sở dữ liệu NoSQL có thể được mở rộng theo chiều ngang trên hàng trăm hoặc hàng ngàn máy chủ.
- **Cơ sở dữ liệu trong bộ nhớ (In-memory Databases):** Cơ sở dữ liệu trong bộ nhớ (IMDB) là một hệ thống quản lý cơ sở dữ liệu chủ yếu dựa vào bộ nhớ chính (Random Access Memory - RAM), thay vì ổ cứng để lưu trữ dữ liệu. Cơ sở dữ liệu trong bộ nhớ nhanh hơn các cơ sở dữ liệu được tối ưu hóa trong đĩa, một điểm quan trọng để sử dụng phân tích big data và tạo ra các kho dữ liệu và các siêu dữ liệu.

2.1.3 Các dịch vụ của Microsoft Azure ứng dụng cho Big Data

- **Azure Blob Storage [6]:** là giải pháp lưu trữ đối tượng của Microsoft cho đám mây. Bộ lưu trữ Blob được tối ưu hóa để lưu trữ một lượng lớn dữ liệu phi cấu trúc. Dữ liệu phi cấu trúc là dữ liệu không tuân theo một mô hình hoặc định nghĩa dữ liệu cụ thể, chẳng hạn như dữ liệu văn bản hoặc dữ liệu nhị phân.
 - Azure Blob Storage được thiết kế cho:
 - + Cung cấp hình ảnh hoặc tài liệu trực tiếp đến trình duyệt.
 - + Lưu trữ tệp để truy cập phân tán.
 - + Truyền phát video và âm thanh.
 - + Ghi vào tệp nhật ký.

- + Lưu trữ dữ liệu để sao lưu và khôi phục, khôi phục sau thảm họa và lưu trữ.
 - + Lưu trữ dữ liệu để phân tích bằng dịch vụ tại chỗ hoặc dịch vụ lưu trữ trên Azure.
- **Azure Data Lake Storage Gen 2 [7]:** là một tập hợp các khả năng dành riêng cho phân tích dữ liệu lớn, được xây dựng trên Azure Blob Storage. Data Lake Storage Gen2 hội tụ các khả năng của Azure Data Lake Storage Gen1 với Azure Blob Storage. Ví dụ: Data Lake Storage Gen2 cung cấp ngữ nghĩa hệ thống tệp, bảo mật cấp tệp và quy mô. Vì những khả năng này được xây dựng trên bộ lưu trữ Blob, bạn cũng sẽ nhận được bộ nhớ theo cấp, chi phí thấp, với tính khả dụng cao và khả năng khôi phục sau thảm họa.
 - Những lợi ích của Azure Data Lake Storage Gen 2: Dễ mở rộng; chi phí rẻ; một dịch vụ có nhiều mảng; hỗ trợ Blob Storage; open source;...
 - + Truy cập tương thích với Hadoop: Data Lake Storage Gen2 cho phép bạn quản lý và truy cập dữ liệu giống như cách bạn làm với Hệ thống tệp phân tán Hadoop (HDFS). Trình điều khiển ABFS mới (được sử dụng để truy cập dữ liệu) có sẵn trong tất cả các môi trường Apache Hadoop. Các môi trường này bao gồm Azure HDInsight, Azure Databricks và Azure Synapse Analytics.
 - + Một tập hợp lớn các quyền POSIX: Mô hình bảo mật cho Data Lake Gen2 hỗ trợ các quyền ACL và POSIX cùng với một số chi tiết bổ sung dành riêng cho Data Lake Storage Gen2. Cài đặt có thể được định cấu hình thông qua Storage Explorer hoặc thông qua các khung như Hive và Spark.
 - + Tiết kiệm chi phí: Data Lake Storage Gen2 cung cấp dung lượng lưu trữ và giao dịch với chi phí thấp. Các tính năng như vòng đồi của Azure Blob Storage tối ưu hóa chi phí khi chuyển đổi dữ liệu qua vòng đồi của nó.
 - + Trình điều khiển được tối ưu hóa: Trình điều khiển ABFS được tối ưu hóa đặc biệt cho phân tích dữ liệu lớn. Các API REST tương ứng được hiển thị thông qua điểm cuối `dfs.core.windows.net`.
- **Azure SQL Databases [8]:** là một nền tảng được quản lý hoàn toàn dưới dạng công cụ cơ sở dữ liệu dịch vụ (PaaS) xử lý hầu hết các chức năng quản lý cơ sở dữ liệu như nâng cấp, vá lỗi, sao lưu và giám sát mà không cần sự tham gia của người dùng. Cơ sở dữ liệu Azure SQL luôn chạy trên phiên bản ổn định mới nhất của công cụ cơ sở dữ liệu SQL Server và hệ điều hành được vá với tính khả dụng 99,99%. Các khả năng PaaS được tích hợp trong Cơ sở dữ liệu Azure

SQL cho phép bạn tập trung vào các hoạt động quản trị và tối ưu hóa cơ sở dữ liệu theo miền cụ thể, rất quan trọng đối với doanh nghiệp của bạn.

- Với Cơ sở dữ liệu Azure SQL, bạn có thể tạo một lớp lưu trữ dữ liệu hiệu suất cao và có sẵn cho các ứng dụng và giải pháp trong Azure. Cơ sở dữ liệu SQL có thể là lựa chọn phù hợp cho nhiều ứng dụng đám mây hiện đại vì nó cho phép bạn xử lý cả dữ liệu quan hệ và cấu trúc phi quan hệ, chẳng hạn như đồ thị, JSON, không gian và XML.
 - Cơ sở dữ liệu Azure SQL dựa trên phiên bản ổn định mới nhất của công cụ cơ sở dữ liệu Microsoft SQL Server. Bạn có thể sử dụng các tính năng xử lý truy vấn nâng cao, chẳng hạn như công nghệ trong bộ nhớ hiệu suất cao và xử lý truy vấn thông minh. Trên thực tế, các khả năng mới nhất của SQL Server được phát hành trước tiên cho Cơ sở dữ liệu SQL, sau đó đến chính SQL Server. Bạn nhận được các khả năng mới nhất của SQL Server mà không phải trả phí để và hoặc nâng cấp, đã được thử nghiệm trên hàng triệu cơ sở dữ liệu.
 - Cơ sở dữ liệu SQL cho phép bạn dễ dàng xác định và mở rộng quy mô hiệu suất trong hai mô hình truy vấn dữ liệu khác nhau: mô hình truy vấn dựa trên vCore và mô hình truy vấn dựa trên DTU. Cơ sở dữ liệu SQL là một dịch vụ được quản lý hoàn toàn có sẵn tính sẵn sàng cao, các bản sao lưu và các hoạt động bảo trì thông thường khác được tích hợp sẵn. Microsoft xử lý tất cả các bản vá và cập nhật mã SQL và hệ điều hành. Bạn không phải quản lý cơ sở hạ tầng cơ bản.
- **Azure Databricks** [9]: Azure Databricks là một nền tảng phân tích dữ liệu được tối ưu hóa cho nền tảng dịch vụ đám mây Microsoft Azure. Azure Databricks cung cấp ba môi trường để phát triển các ứng dụng chuyên sâu về dữ liệu: Databricks SQL, Databricks Data Science & Engineering, và Databricks Machine Learning.
 - Databricks SQL cung cấp một nền tảng dễ sử dụng cho các nhà phân tích muốn chạy truy vấn SQL trên hồ dữ liệu của họ, tạo nhiều kiểu trực quan hóa để khám phá kết quả truy vấn từ các quan điểm khác nhau, đồng thời xây dựng và chia sẻ trang tổng quan.
 - Databricks Data Science & Engineering cung cấp một không gian làm việc tương tác cho phép cộng tác giữa các kỹ sư dữ liệu, nhà khoa học dữ liệu và kỹ sư máy học. Đối với đường ống dữ liệu lớn, dữ liệu (thô hoặc có cấu trúc) được

nhập vào Azure thông qua Azure Data Factory theo lô hoặc được truyền trực tuyến gần thời gian thực bằng Apache Kafka, Event Hub hoặc IoT Hub. Dữ liệu này được đưa vào một hồ dữ liệu để lưu trữ lâu dài liên tục, trong Azure Blob Storage hoặc Azure Data Lake Storage. Là một phần của quy trình phân tích của bạn, hãy sử dụng Azure Databricks để đọc dữ liệu từ nhiều nguồn dữ liệu và biến nó thành thông tin chi tiết đột phá bằng Spark.

- Databricks Machine Learning là một môi trường học máy tích hợp từ đầu đến cuối kết hợp các dịch vụ được quản lý để theo dõi thử nghiệm, đào tạo mô hình, phát triển và quản lý tính năng cũng như cung cấp tính năng và mô hình.
- **Azure HDInsight [10]:** Azure HDInsight là một dịch vụ phân tích mã nguồn mở, toàn phổ, được quản lý trên đám mây dành cho doanh nghiệp. Với HDInsight, bạn có thể sử dụng các khuôn khổ nguồn mở như Hadoop, Apache Spark, Apache Hive, LLAP, Apache Kafka, Apache Storm, R, v.v. trong môi trường Azure của bạn.
- Azure HDInsight là một phân phối đám mây của các thành phần Hadoop. Azure HDInsight giúp xử lý một lượng lớn dữ liệu dễ dàng, nhanh chóng và tiết kiệm chi phí trong một môi trường có thể tùy chỉnh. Bạn có thể sử dụng các khuôn khổ nguồn mở phổ biến nhất như Hadoop, Spark, Hive, LLAP, Kafka, Storm, R, v.v. Với các khuôn khổ này, bạn có thể kích hoạt một loạt các kịch bản như trích xuất, chuyển đổi và tải (ETL), lưu trữ dữ liệu, học máy và IoT.

2.2 Ứng dụng dữ liệu lớn vào thực thế

2.2.1 Ứng dụng trong ngành Ngân hàng

Trong hệ thống ngân hàng, Big Data đã và đang được ứng dụng hiệu quả thể hiện vai trò quan trọng của mình trong mọi hoạt động của ngân hàng: từ thu tiền mặt đến quản lý tài chính. Ngân hàng ứng dụng Big Data vào những bài toán:

- Sử dụng các kỹ thuật phân cụm giúp đưa ra quyết định quan trọng. Hệ thống phân tích có thể xác định các địa điểm chi nhánh nơi tập trung nhiều nhu cầu của khách hàng tiềm năng, để đề xuất lập chi nhánh mới.
- Kết hợp nhiều quy tắc được áp dụng trong các lĩnh vực ngân hàng để dự đoán lượng tiền mặt cần thiết sẵn sàng cung ứng ở một chi nhánh tại thời điểm cụ thể hàng năm.

- Khoa học dữ liệu hiện đang là nền tảng của hệ thống ngân hàng kĩ thuật số.
- Machine learning và AI đang được nhiều ngân hàng sử dụng để phát hiện các hoạt động gian lận và báo cáo cho các chuyên viên liên quan.
- Khoa học dữ liệu hỗ trợ xử lý, lưu trữ và phân tích lượng dữ liệu khổng lồ từ các hoạt động hàng ngày và giúp đảm bảo an ninh cho ngân hàng.

2.2.2 *Ứng dụng trong ngành Y tế*

Khoa học dữ liệu đang dần khẳng định vai trò khá quan trọng trong việc cải thiện sức khỏe con người ngày nay. Big Data không chỉ được ứng dụng để xác định phương hướng điều trị mà giúp cải thiện quá trình chăm sóc sức khỏe. Big Data từ lúc được ứng dụng vào lĩnh vực chăm sóc sức khỏe, đã tạo nên nhiều tác động lớn trong việc giảm lãng phí tiền bạc và thời gian. Ở một số quốc gia, chính phủ đã tài trợ các dự án ứng dụng Big Data để phát triển cơ sở hạ tầng mới và các dịch vụ y tế khẩn cấp. Ngành Y tế ứng dụng Big Data vào các bài toán sau:

- Cho phép người quản lý ca dự đoán các bác sĩ cần thiết vào những thời điểm cụ thể
- Theo dõi tình trạng bệnh nhân bằng việc theo dõi hồ sơ sức khỏe điện tử.
- Sử dụng các thiết bị kỹ thuật số có thể đeo, hệ thống Big Data có thể theo dõi bệnh nhân và gửi báo cáo cho các bác sĩ liên quan.
- Big Data có thể đánh giá các triệu chứng và xác định nhiều bệnh ở giai đoạn đầu.
- Có thể lưu giữ các hồ sơ nhạy cảm được bảo mật và lưu trữ lượng dữ liệu khổng lồ một cách hiệu quả.
- Các ứng dụng Big Data cũng có thể báo trước khu vực có nguy cơ bùng phát dịch như: sốt xuất huyết hoặc sốt rét.

2.2.3 *Ứng dụng trong ngành Thương mại điện tử*

Thương mại điện tử không chỉ tận hưởng những lợi ích của việc điền hành trực tuyến mà còn phải đối mặt với nhiều thách thức để đạt được các mục tiêu kinh doanh. Lý do là bởi các doanh nghiệp dù là nhỏ hay lớn, khi đã tham gia vào thị trường này đều cần đầu tư mạnh để cải tiến công nghệ. Big Data có thể tạo lợi thế cạnh tranh cho doanh nghiệp bằng cách cung cấp thông tin chuyên sâu và các bản báo cáo phân tích xu hướng tiêu dùng. Thương mại điện tử ứng dụng Big Data vào các bài toán sau:

- Có thể thu thập dữ liệu và yêu cầu của khách hàng ngay cả trước khi khách thực sự bắt đầu giao dịch.
- Tạo ra một mô hình tiếp thị hiệu suất cao.
- Nhà quản lý trang thương mại điện tử có thể xác định các sản phẩm được xem nhiều nhất và tối ưu thời gian hiển thị của các trang sản phẩm này.
- Đánh giá hành vi của khách hàng và đề xuất các sản phẩm tương tự. Điều này làm tăng khả năng bán hàng, từ đó tạo ra doanh thu cao hơn.
- Nếu bất kỳ sản phẩm nào được thêm vào giỏ hàng nhưng cuối cùng không được khách hàng mua, Big Data có thể tự động gửi code khuyến mãi cho khách hàng cụ thể đó.
- Các ứng dụng Big Data còn có thể tạo một báo cáo tùy chỉnh theo các tiêu chí: độ tuổi, giới tính, địa điểm của khách truy cập, v.v.
- Xác định các yêu cầu của khách hàng, những gì họ muốn và tập trung vào việc cung cấp dịch vụ tốt nhất để thực hiện nhu cầu của họ.
- Phân tích hành vi, sự quan tâm của khách hàng và theo xu hướng của họ để tạo ra các sản phẩm hướng đến khách hàng.
- Cung cấp các sản phẩm tốt hơn với chi phí thấp hơn.
- Có thể thu thập nhiều dữ liệu về hành vi khách hàng để thiết kế mô hình tiếp thị tối ưu dành được tùy biến theo đối tượng hoặc nhóm đối tượng, tăng khả năng bán hàng.
- Tìm ra sự tương đồng giữa khách hàng và nhu cầu của họ. Từ đó, việc nhắm mục tiêu các chiến dịch quảng cáo có thể được tiến hành dễ dàng hơn dựa trên những phân tích đã có trước đó.

2.2.4 Ứng dụng trong ngành Digital Marketing hoặc Social Media

Digital Marketing là chìa khóa để cánh cửa thành công cho bất kỳ doanh nghiệp nào. Giờ đây, không chỉ các công ty lớn có thể điều hành các hoạt động quảng cáo tiếp thị mà cả các doanh nhân nhỏ cũng có thể chạy các chiến dịch quảng cáo thành công trên các nền tảng truyền thông xã hội và quảng bá sản phẩm của họ. Big Data đã tiếp sức cho Digital Marketing phát triển thực sự mạnh mẽ, và nó đã trở thành một phần không thể thiếu của bất kỳ doanh nghiệp nào. Digital Marketing ứng dụng Big Data như sau:

- Phân tích thị trường, đối thủ cạnh tranh và đánh giá mục tiêu kinh doanh. Điều này giúp cho doanh nghiệp xác định rõ hơn, đâu là cơ hội tốt để tiếp tục tiến hành các kế hoạch kinh doanh tiếp theo.
- Có thể xác định người dùng trên các phương tiện truyền thông xã hội và nhắm mục tiêu cho họ dựa trên nhân khẩu học, giới tính, thu nhập, tuổi tác và sở thích.
- Tạo báo cáo sau mỗi chiến dịch quảng cáo bao gồm hiệu suất, sự tham gia của khán giả và những gì có thể được thực hiện để tạo kết quả tốt hơn.
- Khoa học dữ liệu được sử dụng cho các khách hàng nhằm mục tiêu và nuôi dưỡng chu trình khách hàng.
- Tập trung vào các chủ đề được tìm kiếm cao và tư vấn cho các chủ doanh nghiệp thực hiện chúng trên chiến lược nội dung để xếp hạng trang web doanh nghiệp trên cao hơn trên google (SEO).
- Có thể tạo đối tượng tương tự bằng cách sử dụng cơ sở dữ liệu đối tượng hiện có để nhắm mục tiêu các khách hàng tương tự và kiểm được lợi nhuận.

2.2.5 Một số ứng dụng khác

Còn rất nhiều ngành đang áp dụng rất mạnh mẽ Bigdata như nông nghiệp, giáo dục...cho phép chúng ta có insight ngày càng tốt để ra quyết định nhanh chóng và chính xác.

2.3 Kết luận chương

Trong chương này đã tìm hiểu về dữ liệu lớn, các công nghệ liên quan và các dịch vụ của Microsoft Azure phục vụ cho dữ liệu lớn. Ngoài ra, trong chương này còn thể hiện tổng quát về phân tích dữ liệu cũng như ứng dụng, tầm quan trọng của nó trong cuộc sống thực tiễn hiện nay. Trong quá trình tìm hiểu, tôi thấy được ở Việt Nam hiện đang có rất nhiều doanh nghiệp đầu tư mạnh trong lĩnh vực điện toán đám mây. Điển hình có thể kể đến CMC, đây là một công ty triển khai hệ thống lên Cloud Amazon hoặc Google cho các dự án lớn. Trong chương tiếp theo, tôi đưa ra phương pháp thực hiện ứng dụng Microsoft Azure vào bài toán dữ liệu với bộ dữ liệu thực tế.

CHƯƠNG 3. XÂY DỰNG BÁO CÁO COIVD-19 SỬ DỤNG CÁC DỊCH VỤ CỦA MICROSOFT AZURE

Dựa trên những lý thuyết về Big Data và các công nghệ liên quan, chương này trình bày phương pháp thu thập, chuyển đổi và phân tích dữ liệu thực tế từ bộ dữ liệu Covid-19 được cung cấp bởi ECDC và bộ dữ liệu dân số được cung cấp bởi EUROSTAT. Tuy nhiên, trong đồ án lần này, tôi chỉ sử dụng số liệu đến đầu năm 2021 vì một số lí do sau: những dữ liệu này thực sự có giá trị cao vào thời điểm đó để phục vụ cho việc kiểm soát và điều trị dịch bệnh khi chưa có vac-xin; bên cạnh đó, hiện tại, ECDC không còn cung cấp công khai bộ dữ liệu của họ dưới dạng (.csv) như trước. Tuy nhiên, điều này vẫn đáp ứng được mục tiêu chính của đồ án là tìm hiểu về ứng dụng của Microsoft Azure vào thu thập, chuyển đổi và phân tích dữ liệu tự động.

3.1 Tổng quan bài toán

3.1.1 *Những dịch vụ Microsoft Azure được sử dụng*

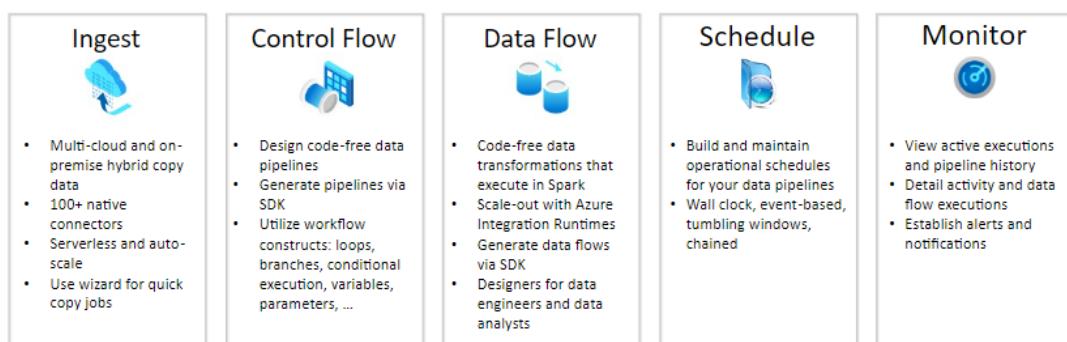
Trong thế giới dữ liệu lớn, dữ liệu thô, chưa được tổ chức thường được lưu trữ trong các hệ thống lưu trữ quan hệ, không quan hệ và các hệ thống lưu trữ khác. Tuy nhiên, về bản chất, dữ liệu thô không có ngữ cảnh hoặc ý nghĩa thích hợp để cung cấp thông tin chi tiết có ý nghĩa cho các nhà phân tích, nhà khoa học dữ liệu hoặc những người ra quyết định kinh doanh. Dữ liệu lớn yêu cầu một dịch vụ có thể sắp xếp và vận hành các quy trình để tinh chỉnh các kho dữ liệu thô khổng lồ này thành những thông tin chi tiết về kinh doanh có thể hành động được. Azure Data Factory là một dịch vụ đám mây được quản lý được xây dựng cho các dự án kết hợp trích xuất-chuyển đổi-tải (ETL), trích xuất-tải-chuyển đổi (ELT) và tích hợp dữ liệu phức tạp này.

Ví dụ: hãy tưởng tượng một công ty trò chơi thu thập hàng petabyte nhật ký trò chơi được tạo ra bởi các trò chơi trên đám mây. Công ty muốn phân tích các nhật ký này để hiểu rõ hơn về sở thích, nhân khẩu học và hành vi sử dụng của khách hàng. Nó cũng muốn xác định các cơ hội bán thêm và bán kèm, phát triển các tính năng mới hấp dẫn, thúc đẩy tăng trưởng kinh doanh và cung cấp trải nghiệm tốt hơn cho khách hàng của mình. Để phân tích các nhật ký này, công ty cần sử dụng dữ liệu tham khảo như thông tin khách hàng, thông tin trò chơi và thông tin chiến dịch tiếp thị có trong kho

dữ liệu tại chỗ. Công ty muốn sử dụng dữ liệu này từ kho dữ liệu tại chỗ, kết hợp dữ liệu này với dữ liệu nhật ký bô sung mà công ty có trong kho dữ liệu đám mây. Để trích xuất thông tin chi tiết, nó hi vọng sẽ xử lý dữ liệu được kết hợp bằng cách sử dụng cụm Spark trong đám mây (Azure HDInsight) và xuất bản dữ liệu đã chuyển đổi vào kho dữ liệu đám mây như Azure Synapse Analytics để dễ dàng tạo báo cáo trên đó. Họ muốn tự động hóa quy trình làm việc này, đồng thời theo dõi và quản lý nó theo lịch trình hàng ngày. Họ cũng muốn thực thi nó khi các tệp đến vùng chứa lưu trữ blob.

Azure Data Factory là nền tảng giải quyết các tình huống dữ liệu như vậy. Đây là dịch vụ tích hợp dữ liệu và ETL dựa trên đám mây cho phép bạn tạo quy trình làm việc theo hướng dữ liệu để điều phối việc di chuyển dữ liệu và chuyển đổi dữ liệu trên quy mô lớn. Sử dụng Azure Data Factory, bạn có thể tạo và lập lịch các quy trình làm việc theo hướng dữ liệu (được gọi là đường ống) có thể nhập dữ liệu từ các kho dữ liệu khác nhau. Bạn có thể xây dựng các quy trình ETL phức tạp biến đổi dữ liệu một cách trực quan với các luồng dữ liệu hoặc bằng cách sử dụng các dịch vụ tính toán như Azure HDInsight Hadoop, Azure Databricks và Azure SQL Database. Ngoài ra, bạn có thể xuất bản dữ liệu đã chuyển đổi của mình lên các cửa hàng dữ liệu như Azure Synapse Analytics để sử dụng các ứng dụng thông minh kinh doanh (BI). Cuối cùng, thông qua Azure Data Factory, dữ liệu thô có thể được sắp xếp thành các kho dữ liệu và hồ dữ liệu có ý nghĩa để đưa ra các quyết định kinh doanh tốt hơn. Azure Data Factory chứa một loạt các hệ thống được kết nối với nhau cung cấp một nền tảng end-to-end hoàn chỉnh cho các kỹ sư dữ liệu.

Code-Free ETL as a service



Hình 3.1 Những dịch vụ của Azure Data Factory

Hình 3.1 đưa ra những dịch vụ làm việc với dữ liệu được cung cấp bởi Azure:

- **Kết nối và lấy dữ liệu:** Với Data Factory, bạn có thể sử dụng Hoạt động sao chép trong đường dẫn dữ liệu để di chuyển dữ liệu từ cả kho dữ liệu tại chỗ và nguồn dữ liệu trên đám mây sang kho dữ liệu tập trung trên đám mây để phân tích thêm. Ví dụ: bạn có thể thu thập dữ liệu trong Azure Data Lake Storage và chuyển đổi dữ liệu sau này bằng cách sử dụng dịch vụ tính toán Azure Data Lake Analytics. Bạn cũng có thể thu thập dữ liệu trong bộ lưu trữ Azure Blob và chuyển đổi dữ liệu đó sau này bằng cách sử dụng cụm Azure HDInsight Hadoop.
- **Biến đổi dữ liệu:** Sau khi dữ liệu có trong kho dữ liệu tập trung trên đám mây, hãy xử lý hoặc chuyển đổi dữ liệu đã thu thập bằng cách sử dụng luồng dữ liệu ánh xạ ADF. Luồng dữ liệu cho phép các kỹ sư dữ liệu xây dựng và duy trì các biểu đồ chuyển đổi dữ liệu thực thi trên Spark mà không cần hiểu các cụm Spark hoặc lập trình Spark. Nếu bạn thích chuyển đổi mã bằng tay, ADF hỗ trợ các hoạt động bên ngoài để thực hiện các chuyển đổi của bạn trên các dịch vụ tính toán như HDInsight Hadoop, Spark, Data Lake Analytics và Machine Learning.
- **CI/CD và xuất bản dữ liệu:** Data Factory cung cấp hỗ trợ đầy đủ cho CI / CD của đường ống dữ liệu của bạn bằng cách sử dụng Azure DevOps và GitHub. Điều này cho phép bạn từng bước phát triển và cung cấp các quy trình ETL của mình trước khi xuất bản thành phẩm. Sau khi dữ liệu thô đã được tinh chỉnh thành dạng tiêu dùng sẵn sàng cho doanh nghiệp, hãy tải dữ liệu vào Kho dữ liệu Azure, Cơ sở dữ liệu Azure SQL, Azure Cosmos DB hoặc bất kỳ công cụ phân tích nào mà người dùng doanh nghiệp của bạn có thể truy cập từ các công cụ thông minh kinh doanh của họ.
- **Quản lý luồng dữ liệu:** Sau khi bạn đã xây dựng và triển khai thành công đường ống tích hợp dữ liệu của mình, cung cấp giá trị kinh doanh từ dữ liệu đã được tinh chế, hãy theo dõi các hoạt động và đường ống đã lên lịch để biết tỷ lệ thành công và thất bại. Azure Data Factory có hỗ trợ tích hợp cho việc giám sát đường ống thông qua Azure Monitor, API, PowerShell, nhật ký Azure Monitor và bảng tình trạng trên cổng Azure.

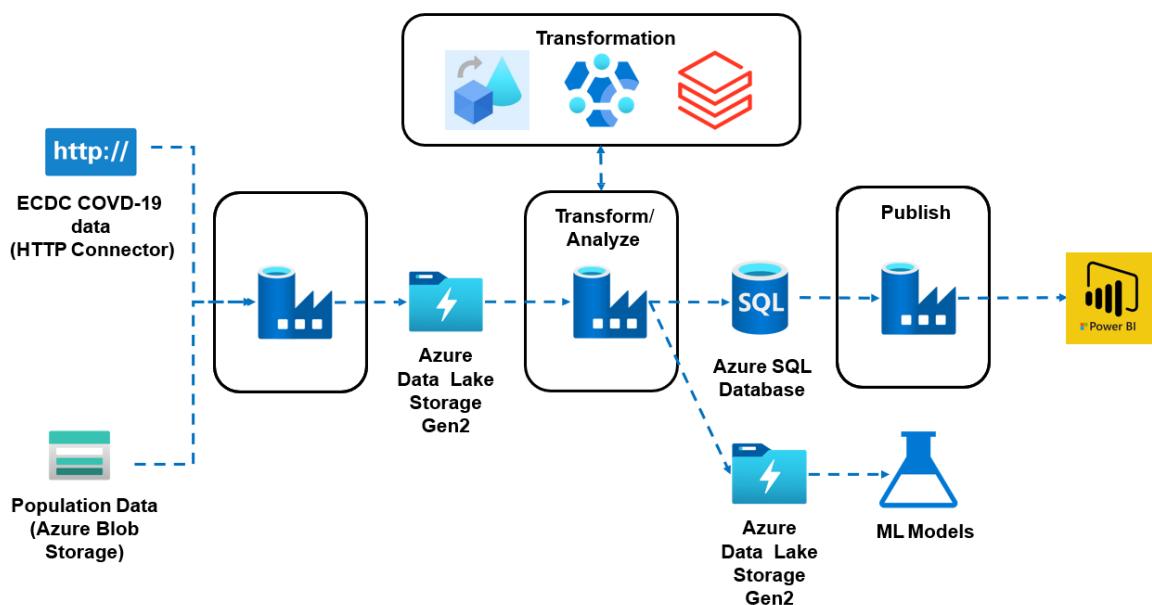
3.1.2 Các nguồn dữ liệu đầu vào

Vấn đề mà tôi sẽ phát triển trong đồ án này dựa trên báo cáo và dự đoán của Covid-19 lây lan. Thứ nhất, tôi muốn tạo một nền tảng dữ liệu mà từ đó tôi có thể chạy mô hình học máy để dự đoán sự lây lan của vi-rút và tìm hiểu các thông tin chi tiết khác từ dữ liệu. Thứ hai, tôi muốn tạo một nền tảng dữ liệu mà từ đó tôi có thể dễ dàng báo cáo về xu hướng Covid-19 sử dụng công cụ báo cáo. Đó là hai mục tiêu chính trong dự án của tôi, giải pháp mà tôi đang xây dựng sẽ có giới hạn để báo cáo về dữ liệu chỉ liên quan đến các nước EU và Vương quốc Anh. Tôi làm điều này vì tôi đã từng sử dụng bộ dữ liệu này để làm một số công việc phân tích nên tương đối quen thuộc. Bộ dữ liệu đầu vào hoàn toàn có thể thay đổi, lấy từ Công thông tin hoặc các nguồn của Tổ chức Y tế Thế giới. Dữ liệu mà tôi xây dựng sẽ cung cấp thông tin chi tiết về các trường hợp Covid-19 đã được xác nhận hàng ngày trên các nền tảng. Những ca tử vong đáng tiếc do hậu quả của covid hàng ngày, nhập viện và số liệu điều trị tích cực (ICU) của các trường hợp, họ sẽ có cả số mới mỗi tuần cũng như số người phải nhập viện vào cuối ngày. Xét nghiệm chi tiết như các bài xét nghiệm được thực hiện mỗi tuần và bất kỳ trường hợp nhiễm mới nào từ bài xét nghiệm. Và cuối cùng, tôi cũng sẽ nhận được số liệu thống kê về dân số ở mọi quốc gia theo nhóm tuổi. Dữ liệu này sau đó có thể được tôi sử dụng để dự đoán sự lây lan của vi-rút, những vấn đề như việc tạo ra các mô hình học máy sẽ nằm ngoài phạm vi của đồ án này. Mục tiêu chính của chúng tôi ở đây là tạo ra một nền tảng dữ liệu mà từ đó tôi có thể tạo ra máy mô hình học tập, xây dựng và điền vào kho dữ liệu với một tập hợp con dữ liệu để nó có thể được sử dụng để báo cáo về các xu hướng. Kho dữ liệu sẽ bao gồm thông tin chi tiết về các trường hợp được xác nhận, tỉ lệ tử vong đáng tiếc, nhập viện và các trường hợp ICU từ số lượng hàng tuần của chúng tôi trong dữ liệu như. Cũng như các con số thử nghiệm sau đó sẽ là một báo cáo từ dữ liệu này. Tôi sẽ sử dụng trang web Eurostat cho dữ liệu dân số theo độ tuổi. Như ở đầu chương đã nêu, những nguồn dữ liệu này đã bị giới hạn nên tôi chỉ sử dụng những số liệu đã lấy được trước đó và tập trung vào việc xây dựng các đường đi của dữ liệu để đúng trọng tâm của đề tài đã đưa ra.

3.2 Xây dựng giải pháp thực hiện bài toán

3.2.1 Giải pháp thiết kế (Solution Architecture)

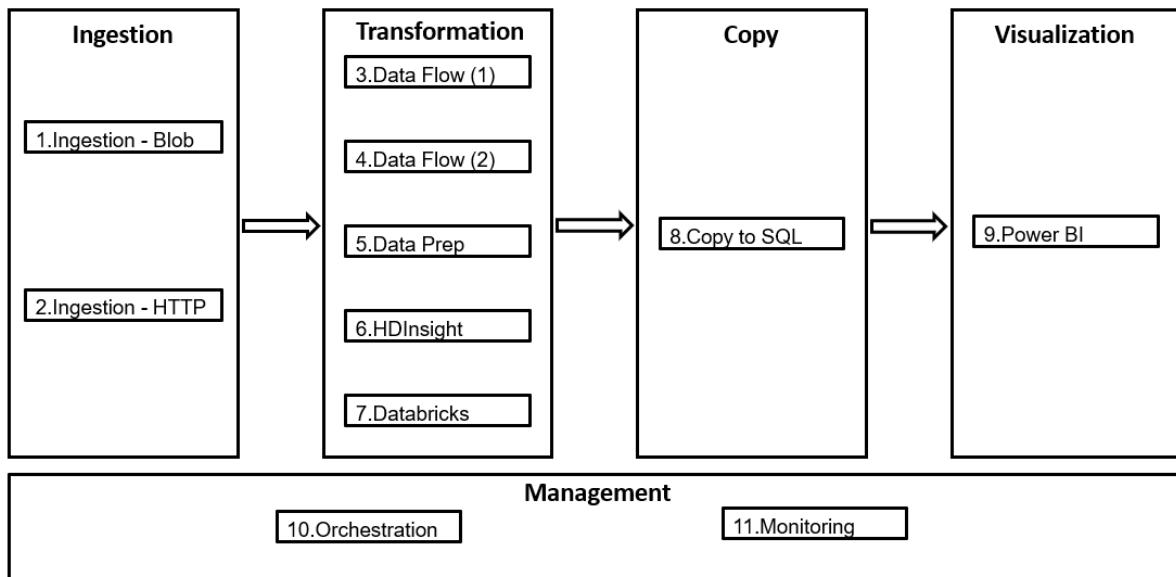
Để giải quyết các yêu cầu của bài toán, tôi đưa ra giải pháp thiết kế như Hình 3.2:



Hình 3.2 Sơ đồ giải pháp thiết kế

Dữ liệu đầu vào được lấy từ hai nguồn là Website ECDC và Website Euro STAT. Sau đó, dữ liệu sẽ được đưa vào Azure Data Factory để thực hiện biến đổi. Cuối cùng, dữ liệu sẽ được sao chép vào SQL Database và dùng cho việc phân tích với Power BI. Cụ thể bài toán được biểu diễn theo các khối như được mô tả trong Hình 3.3. Hệ thống bao gồm các khối chính:

- Ingestion: thu thập dữ liệu
- Transformation - copy: chuyển đổi dữ liệu và sao chép dữ liệu
- Visualization: trực quan hóa dữ liệu
- Orchestration: triển khai
- Monitoring: Giám sát hệ thống

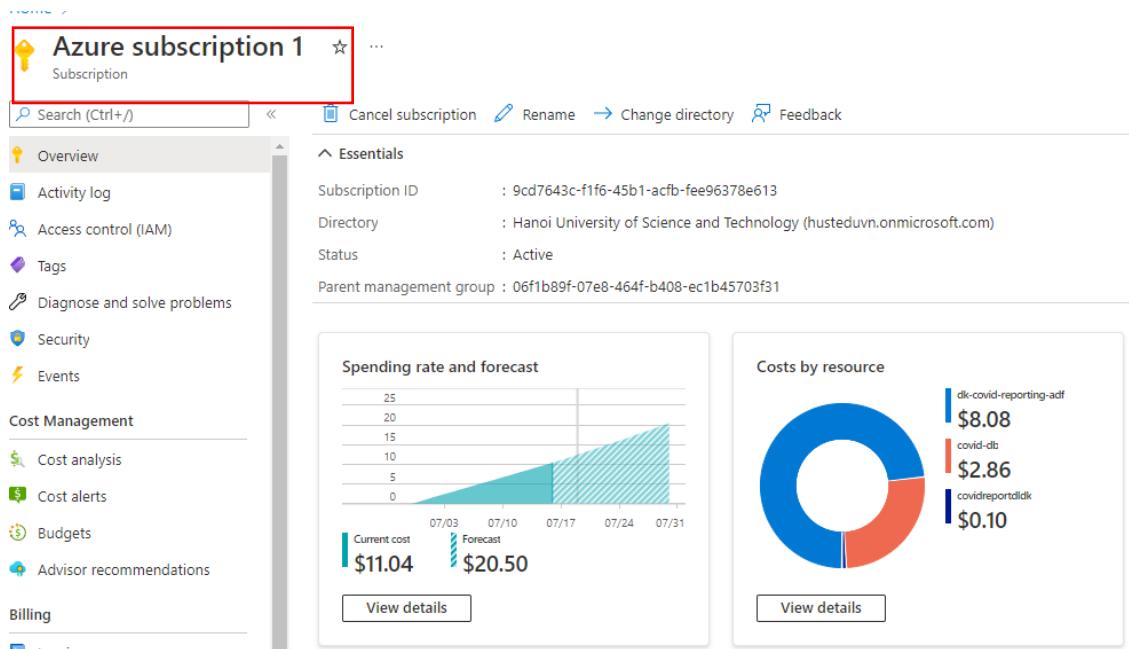


Hình 3.3 Sơ đồ khái niệm hệ thống

Chi tiết quá trình thực hiện sẽ được trình bày trong các phần tiếp theo của chương.

3.2.2 Cài đặt môi trường

Trước hết, để có thể thao tác và làm việc với dữ liệu trên Azure, chúng ta cần thông qua bước cài đặt môi trường. Hay chúng ta cần đăng ký tài khoản và thiết lập các dịch vụ cần thiết. Trước hết, ta cần phải đăng ký một tài khoản Microsoft Azure.



Hình 3.4 Đăng ký Microsoft Azure

Hình 3.4 thể hiện tài khoản Azure đã được đăng ký như trong vùng khoanh đỏ và những khoản phí đã trả cho các dịch vụ đã sử dụng trên Azure. Sau khi đã có tài khoản Azure, ta bắt đầu tạo cài đặt các dịch vụ cần thiết như trong Hình 3.5 theo các hướng dẫn của Microsoft sau:

- Azure Data Factory [1]
- Azure Storage Account và Azure Blob Storage [6]
- Azure Data Lake Storage Gen 2 [7]
- Azure SQL Database [8]
- Azure HD Insight [10]
- Azure Databricks [9]

Name	Type	Last Viewed
Azure subscription 1	Subscription	5 minutes ago
covid-db	SQL database	23 hours ago
covidreportldk	Storage account	4 weeks ago
dk-covid-reporting-adf	Data factory (V2)	4 weeks ago
covid-rpt-databrick	Azure Databricks Service	4 weeks ago
covid-reporting-rg	Resource group	4 weeks ago
covid-hdins-identity	Managed Identity	4 weeks ago
covidreportsadk	Storage account	3 months ago

Hình 3.5 Các dịch vụ cần sử dụng trên Azure

Ngoài các dịch vụ trên Azure, ta còn cần phải chuẩn bị Visual Studio để lập trình Python.

3.3 Triển khai bài toán

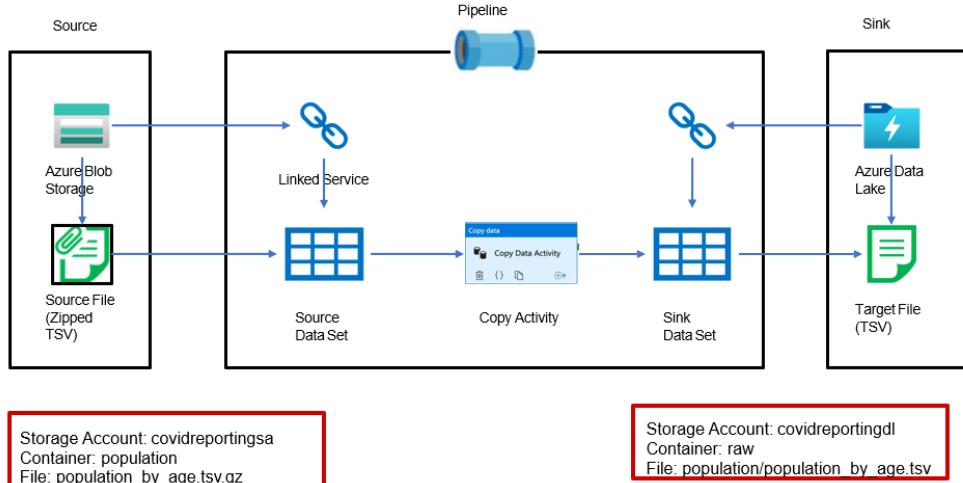
Sau khi đã xây dựng được sơ đồ giải pháp thiết kế cũng như sơ đồ khói của hệ thống và cài đặt hoàn tất các môi trường cần sử dụng, ta bắt đầu đi vào triển khai thực hiện xây dựng hệ thống.

3.3.1 Thu thập dữ liệu (Ingestion)

Ở bước thu thập dữ liệu, tôi sẽ xây dựng các đường ống dữ liệu trên Azure Data Factory để lấy được dữ liệu từ trang web và đưa vào kho lưu trữ của Azure. Bài toán đặt ra bao gồm hai nguồn dữ liệu chính:

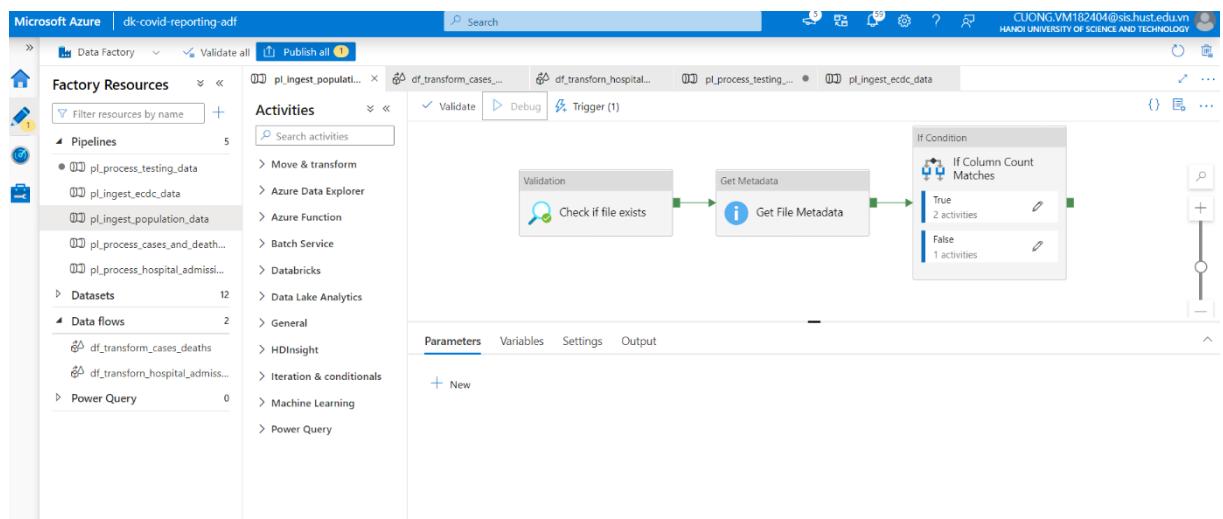
- Trang web ECDC: Cung cấp các dữ liệu liên quan đến dịch bệnh
- Trang web EURO Stat: Cung cấp các dữ liệu liên quan đến dân số, độ tuổi.

- Đầu tiên, tôi sẽ đi vào phần lấy dữ liệu từ trang web EURO Stat. Dữ liệu sẽ đi theo sơ đồ Hình 3.6: dữ liệu được lấy từ trang web và đưa vào Azure Blob Storage, đi qua một đường ống dữ liệu, sau đó sẽ thu được một tệp dữ liệu nén dưới dạng (.tsv).



Hình 3.6 Lấy dữ liệu dân số

Đường ống dữ liệu (Pipeline) sẽ được tôi xây dựng trên Azure Data Factory. Kết quả của đường ống dữ liệu được mô tả trong Hình 3.7.

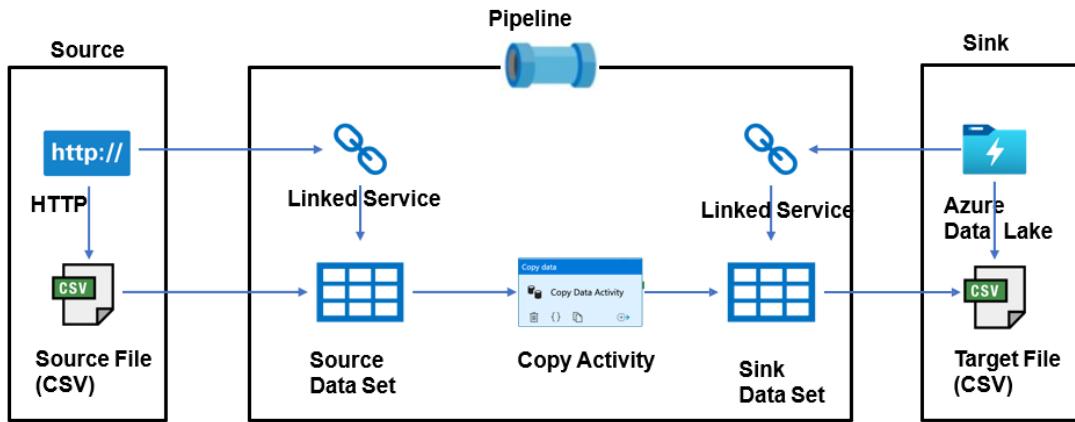


Hình 3.7 Pipeline lấy dữ liệu dân số

Các bước thực thi:

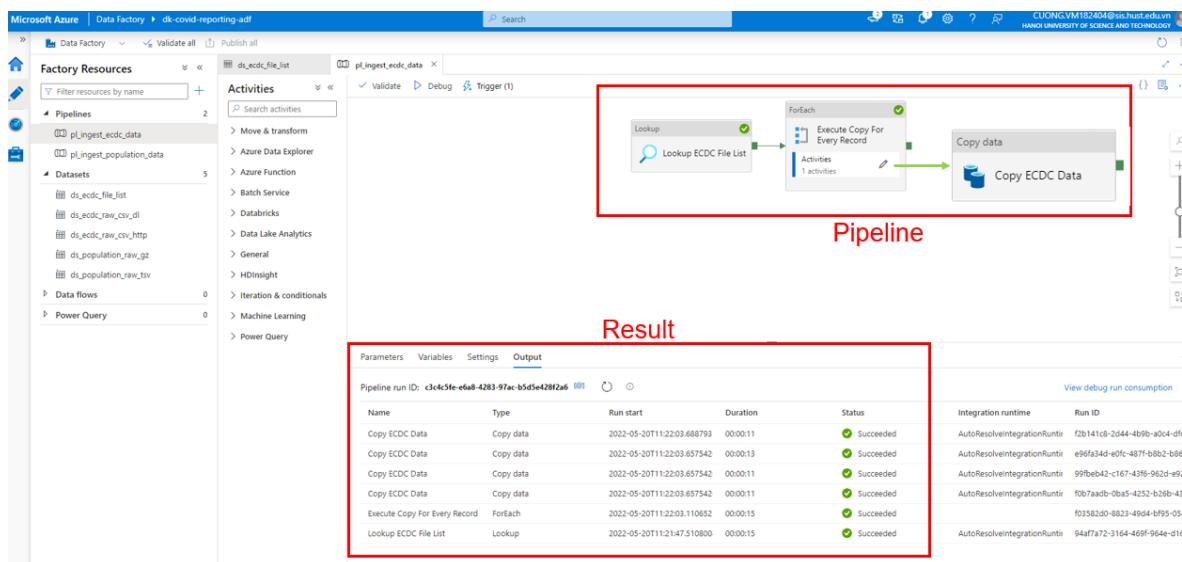
- Kiểm tra xem trong hệ thống đã có file dữ liệu chưa.
- Nếu chưa có thì sẽ lấy file về từ kho dữ liệu và đưa vào Data Lake Storage để giải nén.

- Lấy dữ liệu từ ECDC Website:
- Các bước thực hiện được mô tả như Hình 3.8. Dữ liệu từ nguồn http:// đưa qua đường ống dữ liệu, sau đó sẽ thu được các bộ dữ liệu lưu trong Azure Data Lake dưới dạng (.csv).



Hình 3.8 Lấy dữ liệu dịch bệnh từ trang web ECDC

Đường ống dữ liệu (Pipeline) được thực thi với kết quả (Result) như Hình 3.9



Hình 3.9 Kết quả thu thập dữ liệu ECDC

Sau khi thu thập dữ liệu từ ECDC ta sẽ được các bộ dữ liệu sau:

- Bộ dữ liệu ca nhiễm và tử vong: ds_raw_cases_and_deaths
 - Bộ dữ liệu quản lý của bệnh viện: ds_raw_hospital_admissions
 - Bộ dữ liệu xét nghiệm: ds_testing
- ⇒ Sau khi đi qua khối Thu thập dữ liệu (Ingestion) ta được kết quả là các bộ dataset như Hình 3.10.



6

Hình 3.10 Kết quả khôi Thu thập dữ liệu

Chính những dữ liệu này sẽ được đưa vào khôi Transformation để xử lý, chuyển đổi để thu được dữ liệu như mong muốn cho quá trình phân tích.

3.3.2 Chuyển đổi dữ liệu (Transformation)

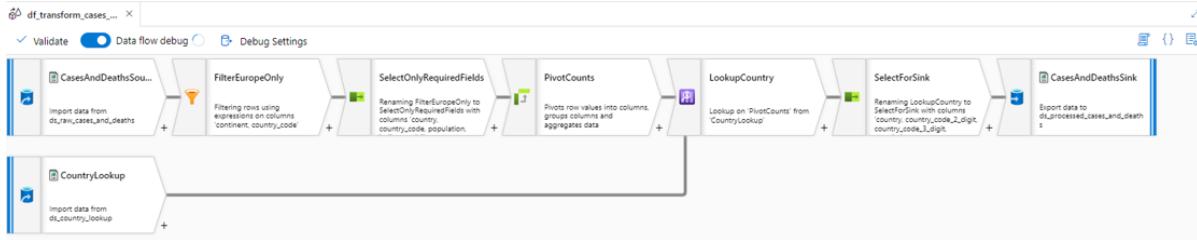
Ở phần này, tôi sẽ xây dựng các đường ống dữ liệu (pipeline) để biến đổi dữ liệu (transform). Kết quả sẽ thu được những bộ dữ liệu đảm bảo yêu cầu phân tích và ứng dụng vào các bài toán học máy sau này. Như tôi đã trình bày ở phần Sơ đồ khái niệm 3.3, khôi Transformation sẽ gồm 5 phần nhỏ: Data Flow (1), Data Flow (2), Data Prepare, HD Insight và Databricks.

3.3.2.1 Data Flow (1): Chuyển đổi dữ liệu số ca nhiễm và ca tử vong.

Ở phần này, tôi sẽ sử dụng Azure Data Factory để xây dựng Pipeline nhằm thực thi các yêu cầu sau:

- Lọc dữ liệu (chỉ lấy những dữ liệu khu vực EU)
- Chọn những cột cần thiết và bỏ đi các cột còn lại.
- Pivot dữ liệu
- Look up dữ liệu

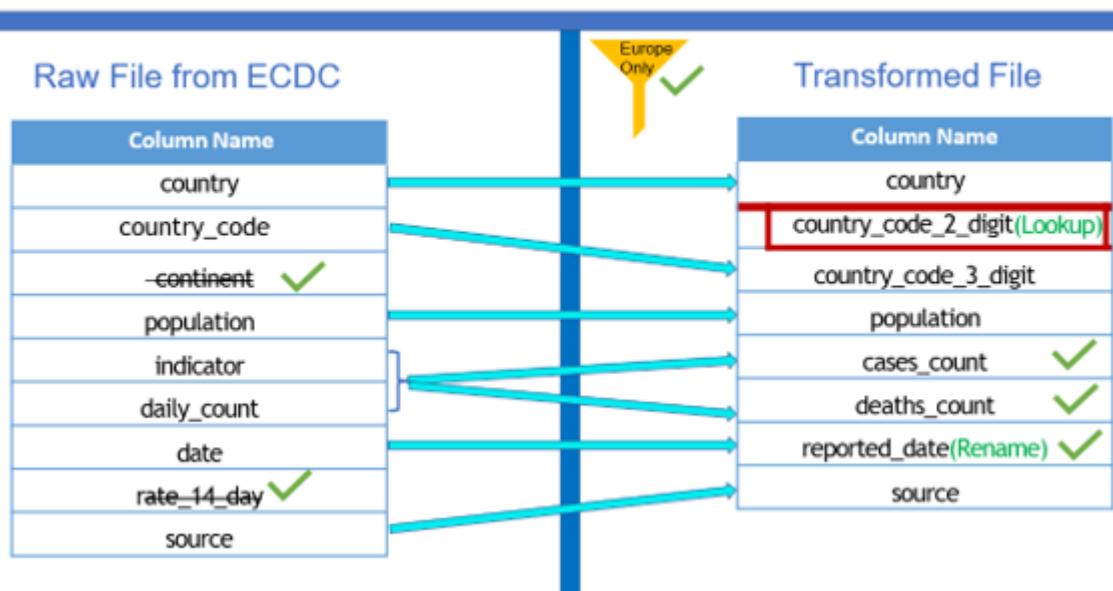
Pipeline được thể hiện ở Hình 3.11, thứ tự đường đi từ trái sang phải: Import; Filter; Select; Pivot; Lookup; Select For Sink; Export.



Hình 3.11 Data Flow chuyển đổi dữ liệu ca nhiễm và tử vong

Sau quá trình chuyển đổi dữ liệu Data Flow (1), ta thu được dataset có dạng như cột Transformed File trong Hình 3.12.

Transform Cases & Deaths Data



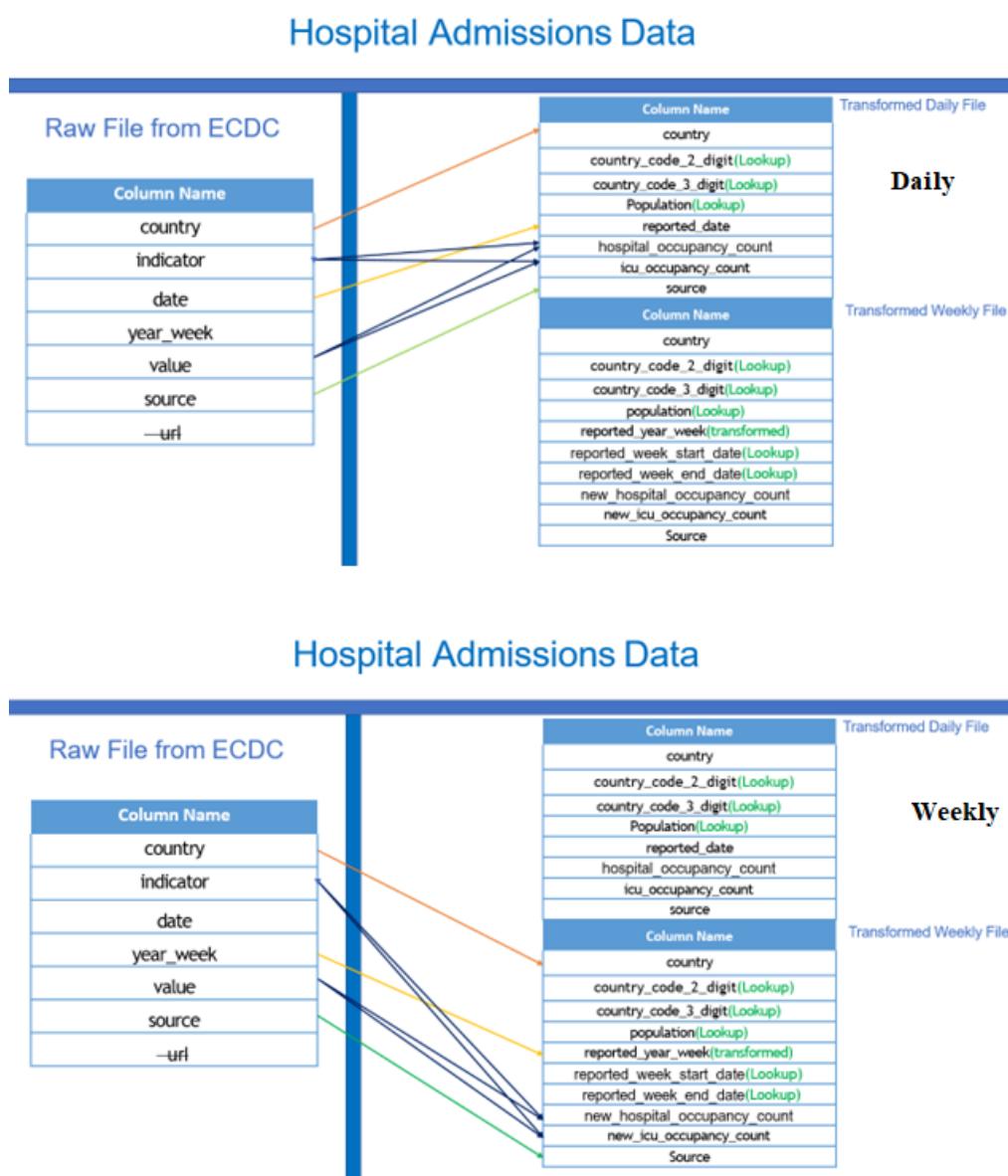
Hình 3.12 Bộ dữ liệu sau khi hoàn thành Data Flow (1)

- Như trong Hình 3.12 mô tả, bộ dữ liệu đã được thực hiện các chuyển đổi như sau:
 - + Bỏ các trường “continent”, “rate_14_day”
 - + Lookup để có trường “country_code_2_digit”
 - + Trường “country_code” được chuyển sang “country_code_3_digit”
 - + Trường “date” được đổi thành “reported_date”
 - + Từ các trường “indicator” và “daily_count” để tạo ra các trường “cases_count” và “deaths_count”
 - + Các trường còn lại được giữ nguyên

Các thao tác xây dựng pipe thực hiện trên Azure Data Factory với UX/UI đơn giản, chỉ cần thực hiện các cài đặt chi tiết cho từng khối mà mình mong muốn. Điều này cũng là một trong những lợi ích lớn mà Azure cung cấp cho người dùng.

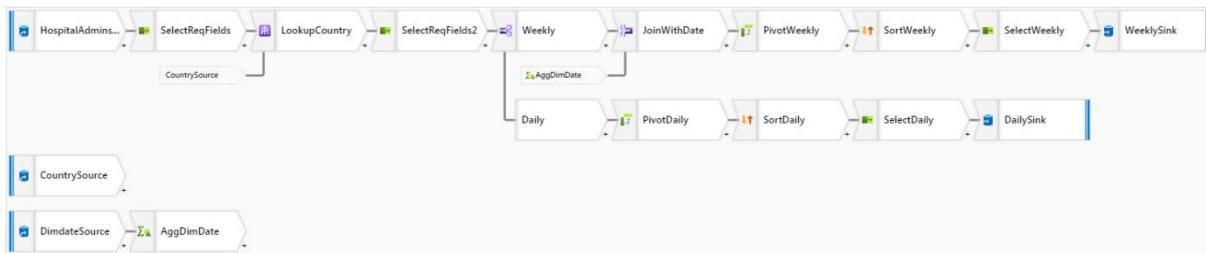
3.3.2.2 Data Flow (2): Chuyển đổi dữ liệu bệnh viện

Tương tự như phần Data Flow (1), phần này tôi cũng xây dựng pipeline để thực hiện chuyển đổi dữ liệu từ file raw của ECDC, bỏ đi những dữ liệu thừa và giữ lại các trường dữ liệu mà mình mong muốn. Việc triển khai pipeline sẽ chia bộ dữ liệu Hospital Admissions thành các dữ liệu theo từng ngày và từng tuần theo các bước biến đổi như Hình 3.13.



Hình 3.13 Chuyển đổi bộ dữ liệu raw

Để thực hiện được yêu cầu trên, tôi đã cấu hình một pipeline như Hình 3.14



Hình 3.14 Data Flow (2)

- Data Flow gồm có các bước chính sau:
 - Select
 - Lookup
 - Join
 - Sort
 - Aggregate
 - Pivot
- ⇒ Các thao tác này có thể thực hiện thủ công trên Excel hoặc Power BI đối với những người đã quen làm việc với dữ liệu hoặc phân tích dữ liệu. Tuy nhiên, với Azure, việc này hoàn toàn được thực hiện trên Cloud Azure, chỉ cần cấu hình theo đúng ý muốn của mình. Điều này sẽ giúp tiết kiệm được thời gian cũng như công sức trong quá trình chuyển đổi dữ liệu.
- ⇒ Sau khi hoàn thành quá trình chuyển đổi, ta sẽ thu được dataset ca nhiễm và ca tử vong theo ý muốn và có thể sử dụng cho các khôi sau.

3.3.2.3 HD Insight

Ở phần này tôi sẽ chuyển đổi dữ liệu xét nghiệm (testing file) được lấy từ EDCD và đưa vào Azure Data Lake thông qua khôi Ingestion. Việc chuyển đổi này sẽ được thực hiện với Hadoop thông qua dịch vụ Azure HD Insight.

- Tạo cụm HD Insight:

Hadoop sẽ làm việc với hệ thống dữ liệu phân tán (HDFS), vì vậy việc đầu tiên chúng ta cần làm là tạo cụm (cluster) HD Insight để xử lý dữ liệu.

- Trước khi tạo cụm HD Insight, để Hadoop có thể làm việc được với các thư mục dữ liệu đã tạo trong Azure Data Lake Storage Gen 2, ta cần phải tạo “User

Assigned Managed Identity” như trong Hình 3.15. Điều này sẽ giúp cho Hadoop có quyền truy cập vào các bộ dữ liệu trong Azure Data Lake Storage Gen 2.

Hình 3.15 Tạo xác thực quyền truy cập cho HD Insight

- Sau khi hoàn tất tạo “User Assigned Managed Identity”, để Hadoop có thể truy cập vào Data Lake, ta cần thiết lập Access Control bằng cách “Add role assignment” như trong Hình 3.16:

Hình 3.16 Add role assignment

- Khi đã có quyền truy cập vào Data Lake, ta bắt đầu tạo cụm HD Insight như trong Hình 3.17.

Create HDInsight cluster

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription *

Resource group *

Cluster details

Name your cluster, pick a region, and choose a cluster type and version. [Learn more](#)

Cluster name *

Region *

Availability zone

Cluster type *

Version *

Cluster credentials

Enter new credentials that will be used to administer or access the cluster.

Cluster login username *

Cluster login password *

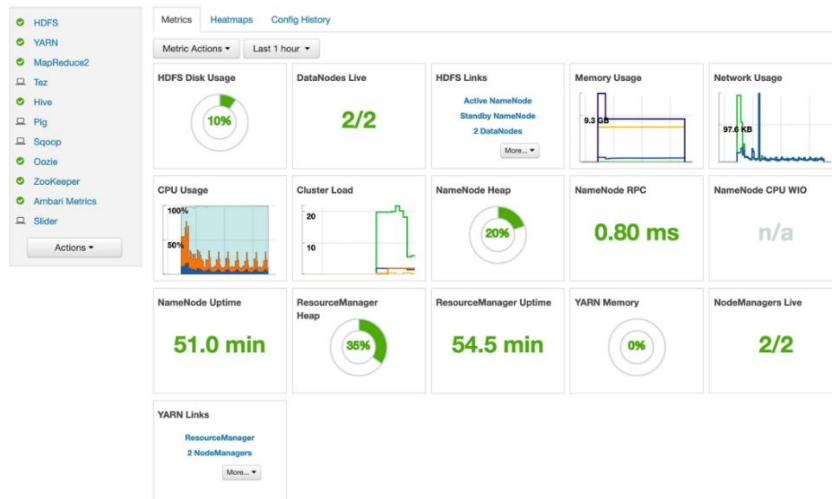
Confirm cluster login password *

Secure Shell (SSH) username *

Use cluster login password for SSH

Hình 3.17 Tạo cụm HD Insight – Hadoop

- Ở bước này, chúng ta cần chú ý chọn các thông số phù hợp với nhu cầu, không quá dư thừa để đảm bảo tiết kiệm nhất chi phí. Các thông số cần chú ý:
 - + Số nhân (core)
 - + Dung lượng bộ nhớ
 - + Đời GPU/CPU
- Quản lý và vận hành HD Insight:



Hình 3.18 Giao diện chính của HD Insight

- Như Hình 3.18 ta có thể thấy các thông số của HDFS, các Node,... đều được thể hiện rất cụ thể thông qua các biểu đồ. Ngoài ra, ta còn có thể theo dõi các node đang chạy để đảm bảo quá trình vận hành ổn định như trong Hình 3.19.

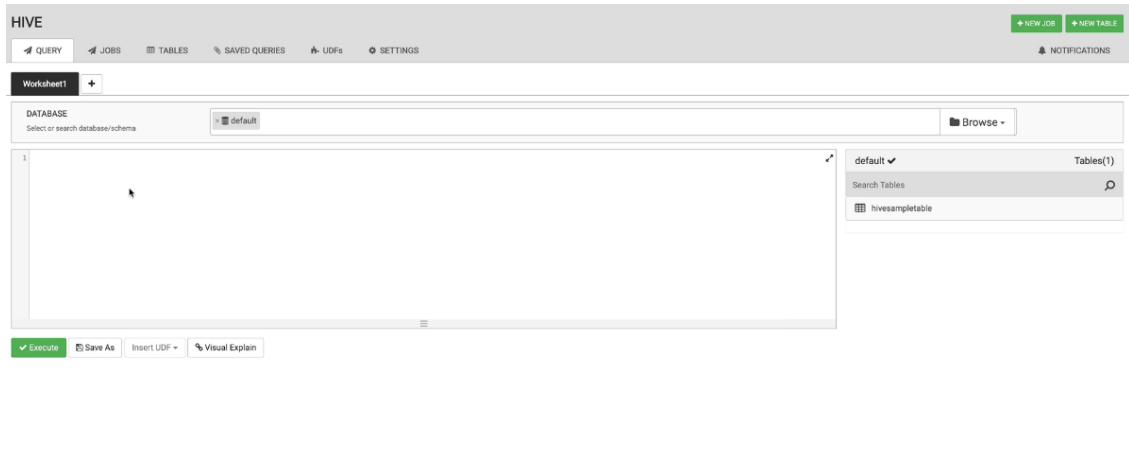
The screenshot shows the HD Insight Node list page with the following details:

<input type="checkbox"/>	Name	IP Address	Rack	Cores	RAM	Disk Usage	Load Avg	Versions	Components
<input type="checkbox"/>	hn0-covid.ams5bzg1rkxevcy2d...	10.0.0.17	/default-rack	2 (2)	13.67GB	<div style="width: 100%;"> </div>	18.06	HDP-2.6.5.3026-7	22 Components
<input type="checkbox"/>	hn1-covid.ams5bzg1rkxevcy2d...	10.0.0.16	/default-rack	2 (2)	13.67GB	<div style="width: 100%;"> </div>	5.67	HDP-2.6.5.3026-7	17 Components
<input type="checkbox"/>	wn1-covid.ams5bzg1rkxevcy2d...	10.0.0.10	/default-rack	2 (2)	13.67GB	<div style="width: 100%;"> </div>	0.58	HDP-2.6.5.3026-7	7 Components
<input type="checkbox"/>	wn2-covid.ams5bzg1rkxevcy2d...	10.0.0.8	/default-rack	2 (2)	13.67GB	<div style="width: 100%;"> </div>	0.33	HDP-2.6.5.3026-7	7 Components
<input type="checkbox"/>	zk0-covid.ams5bzg1rkxevcy2d...	10.0.0.7	/default-rack	1 (1)	1.88GB	<div style="width: 100%;"> </div>	1.14	HDP-2.6.5.3026-7	4 Components
<input type="checkbox"/>	zk1-covid.ams5bzg1rkxevcy2d...	10.0.0.6	/default-rack	1 (1)	1.88GB	<div style="width: 100%;"> </div>	0.32	HDP-2.6.5.3026-7	4 Components
<input type="checkbox"/>	zk2-covid.ams5bzg1rkxevcy2d...	10.0.0.4	/default-rack	1 (1)	1.88GB	<div style="width: 100%;"> </div>	0.59	HDP-2.6.5.3026-7	4 Components

Show: 10 | 1 - 7 of 7 | ← →

Hình 3.19 Các node trong cụm HD Insight

- Chạy Query Hive trên cụm HD Insight để chuyển đổi dữ liệu testing.



Hình 3.20 Hive trên Hadoop - HD Insight

Hình 3.20 mô tả không gian làm việc với Hive trên HD Insight. Ta có thể thực hiện các thao tác như: Query, xem Jobs, Tables,...

- Upload file code “[covid_transform_testing.hql](#)” để chạy các dòng lệnh chuyển đổi file testing.
- Kết quả chạy lệnh được mô tả như Hình 3.21: ta có thể thấy được các thông số như thời gian thực thi, tài nguyên tiêu tốn,...

```

20/11/12 15:54:07 [main]: WARN parse.RowResolver: Duplicate column info for c.country_code_3_digit was overwritten in RowResolver map: _col2: string by _col2: string
20/11/12 15:54:07 [main]: WARN parse.RowResolver: Duplicate column info for t.year_week was overwritten in RowResolver map: _col3: string by _col3: string
20/11/12 15:54:07 [main]: WARN parse.RowResolver: Duplicate column info for t.new_cases was overwritten in RowResolver map: _col6: bigint by _col6: bigint
20/11/12 15:54:07 [main]: WARN parse.RowResolver: Duplicate column info for t.tests_done was overwritten in RowResolver map: _col7: bigint by _col7: bigint
20/11/12 15:54:07 [main]: WARN parse.RowResolver: Duplicate column info for t.population was overwritten in RowResolver map: _col8: bigint by _col8: bigint
20/11/12 15:54:07 [main]: WARN parse.RowResolver: Duplicate column info for t.testing_rate was overwritten in RowResolver map: _col9: double by _col9: double
20/11/12 15:54:07 [main]: WARN parse.RowResolver: Duplicate column info for t.positivity_rate was overwritten in RowResolver map: _col10: double by _col10: double
20/11/12 15:54:07 [main]: WARN parse.RowResolver: Duplicate column info for t.testing_data_source was overwritten in RowResolver map: _col11: string by _col11: string
20/11/12 15:54:08 [main]: ERROR calcite.RelOptHiveTable: No Stats for covid_reporting_lookup@dim_date, Columns: the_year, week_of_year
20/11/12 15:54:08 [main]: ERRUN parse.CalcitePlanner: CBO failed due to missing column stats (see previous errors), skipping CBO
20/11/12 15:54:08 [main]: WARN parse.RowResolver: Duplicate column info for t.country was overwritten in RowResolver map: _col0: string by _col0: string
20/11/12 15:54:08 [main]: WARN parse.RowResolver: Duplicate column info for c.country_code_2_digit was overwritten in RowResolver map: _col1: string by _col1: string
20/11/12 15:54:08 [main]: WARN parse.RowResolver: Duplicate column info for c.country_code_3_digit was overwritten in RowResolver map: _col2: string by _col2: string
20/11/12 15:54:08 [main]: WARN parse.RowResolver: Duplicate column info for t.year_week was overwritten in RowResolver map: _col3: string by _col3: string
20/11/12 15:54:08 [main]: WARN parse.RowResolver: Duplicate column info for t.new_cases was overwritten in RowResolver map: _col6: bigint by _col6: bigint
20/11/12 15:54:08 [main]: WARN parse.RowResolver: Duplicate column info for t.tests_done was overwritten in RowResolver map: _col7: bigint by _col7: bigint
20/11/12 15:54:08 [main]: WARN parse.RowResolver: Duplicate column info for t.population was overwritten in RowResolver map: _col8: bigint by _col8: bigint
20/11/12 15:54:08 [main]: WARN parse.RowResolver: Duplicate column info for t.testing_rate was overwritten in RowResolver map: _col9: double by _col9: double
20/11/12 15:54:08 [main]: WARN parse.RowResolver: Duplicate column info for t.positivity_rate was overwritten in RowResolver map: _col10: double by _col10: double
20/11/12 15:54:08 [main]: WARN parse.RowResolver: Duplicate column info for t.testing_data_source was overwritten in RowResolver map: _col11: string by _col11: string
Query ID = yarn_20201112155403_9103a8b0-12f3-476f-8f41-017cfca59c89e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1605184425458_0004)
Status: DAG finished successfully in 34.76 seconds
Query Execution Summary
OPERATION DURATION
Compile Query 8.11s
Prepare Plan 1.67s
Submit Plan 2.31s
Start DAG 1.51s
Run DAG 34.77s
Task Execution Summary
VERTICES TOTAL_TASKS FAILED_ATTEMPTS KILLED_TASKS DURATION(ms) CPU_TIME(ms) GC_TIME(ms) INPUT_RECORDS OUTPUT_RECORDS
Map 1 1 0 0 16857.00 9,608 507 1,108 1,108
Map 2 1 0 0 21148.00 13,900 711 2,411 1,108
Map 4 1 0 0 17480.00 9,860 573 208 208
Reducer 3 1 0 0 1918.00 2,710 288 1,108 0
Loading data to table covid_reporting_processed.testing
Table covid_reporting_processed.testing stats: [numFiles=1, numRows=1108, totalSize=116837, rawDataSize=115729]
OK
Time taken: 50.808 seconds

```

Hình 3.21 Kết quả vận hành pipeline chạy file (.hql) trên HD Insight

- Sau khi hoàn thành quá trình chuyển đổi sẽ thu được file testing mới có thể sử dụng cho các khối sau như trong Hình 3.22.

```

File Edit Selection View Go Run Terminal Help
E: > Documents > DATN > covid19-main > testing > 000000_0
1 Austria,AT,AUT,2020-W15,2020-04-05,2020-04-11,2841,12339,8858775,139.285623576623,16.5410487073507,Manual webscraping
2 Austria,AT,AUT,2020-W16,2020-04-12,2020-04-18,855,58488,8858775,666.226724349586,1.4613832581042,Manual webscraping
3 Austria,AT,AUT,2020-W17,2020-04-19,2020-04-25,472,33443,8858775,377.512692217603,1.4113563666537,Manual webscraping
4 Austria,AT,AUT,2020-W18,2020-04-26,2020-05-02,336,26598,8858775,300.244672655879,1.2632528761561,Country website
5 Austria,AT,AUT,2020-W19,2020-05-03,2020-05-09,307,42153,8858775,475.833283947273,0.728299290679193,Country website
6 Austria,AT,AUT,2020-W20,2020-05-10,2020-05-16,363,46061,8858775,519.27044889865,0.789113288146884,Country website
7 Austria,AT,AUT,2020-W21,2020-05-17,2020-05-23,267,39348,8858775,444.169763878189,0.678568536749009,Country website
8 Austria,AT,AUT,2020-W22,2020-05-24,2020-05-30,231,46677,8858775,526.901292785966,0.494890417121923,Country website
9 Austria,AT,AUT,2020-W23,2020-05-31,2020-06-06,184,41063,8858775,463.529899678003,0.448091956262329,Country website
10 Austria,AT,AUT,2020-W24,2020-06-07,2020-06-13,192,35243,8858775,397.831528625572,0.544789036128648,Country website
11 Austria,AT,AUT,2020-W25,2020-06-14,2020-06-20,233,15775,8858775,178.072824638945,1.4770286022187,Country website
12 Austria,AT,AUT,2020-W26,2020-06-21,2020-06-27,315,61905,8858775,608.798648797379,0.508844196753089,Country website
13 Austria,AT,AUT,2020-W27,2020-06-28,2020-07-04,634,45284,8858775,511.176771054689,1.40005299885169,Country website
14 Austria,AT,AUT,2020-W28,2020-07-05,2020-07-11,599,48936,8858775,552.491432477967,1.22404773581821,Country website
15 Austria,AT,AUT,2020-W29,2020-07-12,2020-07-18,713,51929,8858775,586.187142127439,1.37382855822373,Country website
16 Austria,AT,AUT,2020-W30,2020-07-19,2020-07-25,841,99229,8858775,1120.121089959505,0.847534496925932,Country website
17 Austria,AT,AUT,2020-W31,2020-07-26,2020-08-01,875,57416,8858775,648.125728444396,1.52396545172708,Country website
18 Austria,AT,AUT,2020-W32,2020-08-02,2020-08-08,771,56554,8858775,638.395263453469,1.257205502709538,Country website
19 Austria,AT,AUT,2020-W33,2020-08-09,2020-08-15,1276,56622,8858775,639.162863939902,2.25354102645615,Country website
20 Austria,AT,AUT,2020-W34,2020-08-16,2020-08-22,1888,76497,8858775,863.516682611309,2.4680706432932,Country website
21 Austria,AT,AUT,2020-W35,2020-08-23,2020-08-29,1838,77105,8858775,870.379934029223,2.38376240191946,Country website
22 Austria,AT,AUT,2020-W36,2020-08-30,2020-09-05,2037,83733,8858775,945.198404971342,2.43273261438143,Country website
23 Austria,AT,AUT,2020-W37,2020-09-06,2020-09-12,3977,86241,8858775,973.509317033111,4.61149569230413,Country website
24 Austria,AT,AUT,2020-W38,2020-09-13,2020-09-19,4997,102617,8858775,1158.36557537583,4.86956352261321,Country website
25 Austria,AT,AUT,2020-W39,2020-09-20,2020-09-26,4992,118816,8858775,1250.91787521412,4.50476465492348,Country website
26 Austria,AT,AUT,2020-W40,2020-09-27,2020-09-03,5079,139874,8858775,1477.33744225359,3.8808194522976,Country website
27 Austria,AT,AUT,2020-W41,2020-10-04,2020-10-10,6666,124663,8858775,1407.22616840365,5.3721609459102,TESSy
28 Austria,AT,AUT,2020-W42,2020-10-11,2020-10-17,9810,129647,8858775,1463.48676876882,7.56670034786767,TESSy
29 Austria,AT,AUT,2020-W43,2020-10-18,2020-10-24,15275,158997,8858775,1794.79668464319,9.60789958502211,TESSy
30 Austria,AT,AUT,2020-W44,2020-10-25,2020-10-31,26814,167926,8858775,1895.58940146917,15.967747698391,TESSy
31 Austria,AT,AUT,2020-W45,2020-11-01,2020-11-07,39918,199567,8858775,2252.7668687939,28.802304998384,TESSy
32 Belgium,BE,BEL,2020-W09,2020-02-23,2020-02-29,18,82,11455519,0.715812177518976,21.9512195121951,Country website
33 Belgium,BE,BEL,2020-W10,2020-03-01,2020-03-07,324,4406,11455519,38.4618104164464,7.35360871538811,Country website
34 Belgium,BE,BEL,2020-W11,2020-03-08,2020-03-14,1198,9924,11455519,86.6307323133941,12.0717452640865,Country website
35 Belgium,BE,BEL,2020-W12,2020-03-15,2020-03-21,3401,17066,11455519,14R.976227091937,19.928512R125325,Country website

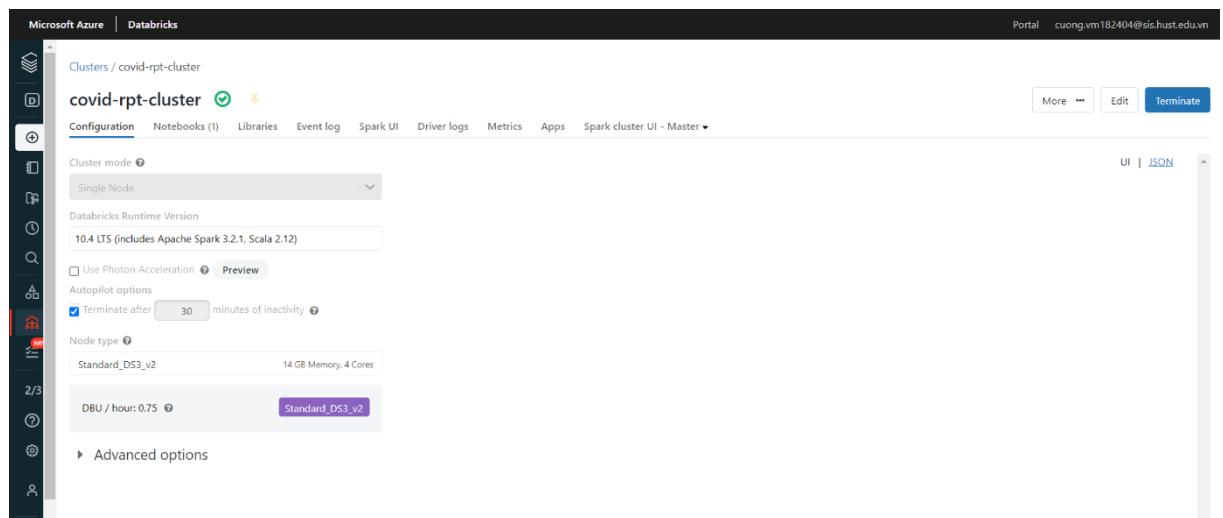
```

Ln 1 Col 1 Spaces: 4 UTF-8 LF Plain Text R

Hình 3.22 Kết quả thu được sau quá trình chuyển đổi qua HD Insight

3.3.2.4 Databricks

Ở phần này, tôi sẽ thực hiện chuyển đổi file dữ liệu dân số theo độ tuổi để thu được bộ dữ liệu như mong muốn cho quá trình phân tích, trực quan hóa dữ liệu. Trước khi đi vào chuyển đổi. Tôi sử dụng Databricks để ứng dụng Spark vào việc liên kết dữ liệu. Các Spark Cluster (cụm Spark) sẽ giúp tôi thực hiện điều này một cách dễ dàng. Để làm được điều đó, trước tiên tôi sẽ tạo một Spark Cluster trên Databricks như trong Hình 3.23. Cần phải chú ý chọn đúng các thông số để đảm bảo tính tối ưu cho hệ thống và tiết kiệm chi phí nhất có thể.



Hình 3.23 Spark Cluster trên Databricks

- Sau khi đã tạo được Spark Cluster, tôi sẽ sử dụng ngôn ngữ Python để thực hiện quá trình liên kết các dữ liệu với nhau (mount) như trong Hình 3.24: Chạy các dòng lệnh Python.

The screenshot shows the Microsoft Azure Databricks interface. At the top, it says "Microsoft Azure | Databricks". Below that, there's a sidebar with icons for file operations like copy, move, and delete. The main area has tabs for "mount_storage" and "Python". A command bar at the top of the code editor says "Command took 24.82 seconds -- by cuong.vm182404@sis.hust.edu.vn at 6/22/2022, 11:59:27 PM on covid-rpt-cluster".

Cmd 6:

```
Mount the processed container
Update the storage account name before executing
```

Cmd 7:

```
1 dbutils.fs.mount(
2   source = "abfss://processed@covidreportdlk.dfs.core.windows.net/",
3   mount_point = "/mnt/covidreportdlk/processed",
4   extra_configs = configs)
```

Out[3]: True

Cmd 8:

```
Mount the lookup container
Update the storage account name before executing
```

Hình 3.24 Mount dữ liệu với Python

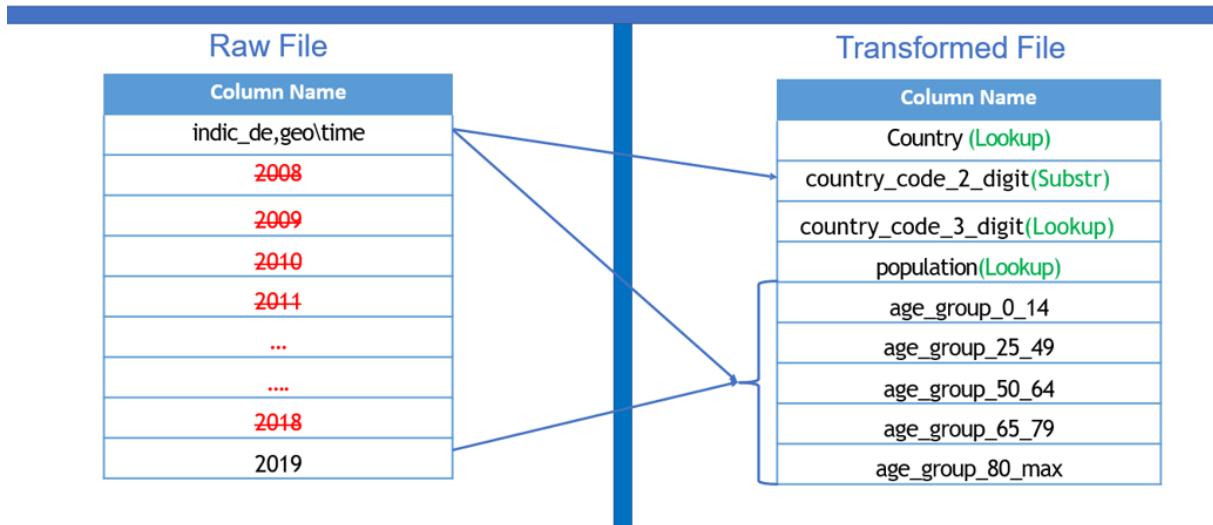
- Sau khi đã thực hiện liên kết dữ liệu với Spark Cluster trên Azure Databricks, ta bắt đầu đi vào quá trình biến đổi dữ liệu. Trước khi chuyển đổi, ta cần xem xét bộ dữ liệu gốc và quyết định thao tác với nó để đưa ra được bộ dữ liệu đầu ra đúng như mong muốn.

indic_de,geo\time	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
PC_Y0_14,UK	17.7	17.7	17.6	17.6	17.6	17.6	17.6	17.7	17.7	17.8	17.9	17.9
PC_Y15_24,UK	13.2	13.1	13.1	13.1	13	12.9	12.7	12.6	12.4	12.1	11.9	11.8
PC_Y25_49,UK	35.2	35.1	34.9	34.7	34.4	34.2	33.9	33.6	33.4	33.2	33	32.8
PC_Y50_64,UK	18	18	18.1	18.2	18.2	18.1	18.2	18.4	18.5	18.7	18.9	19.1
PC_Y65_79,UK	11.5	11.6	11.7	11.9	12.1	12.5	12.8	13	13.1	13.2	13.3	13.4
PC_Y80_MAX,UK	4.4	4.5	4.5	4.6	4.6	4.7	4.7	4.8	4.8	4.9	4.9	5

Hình 3.25 Bộ dữ liệu dân số gốc

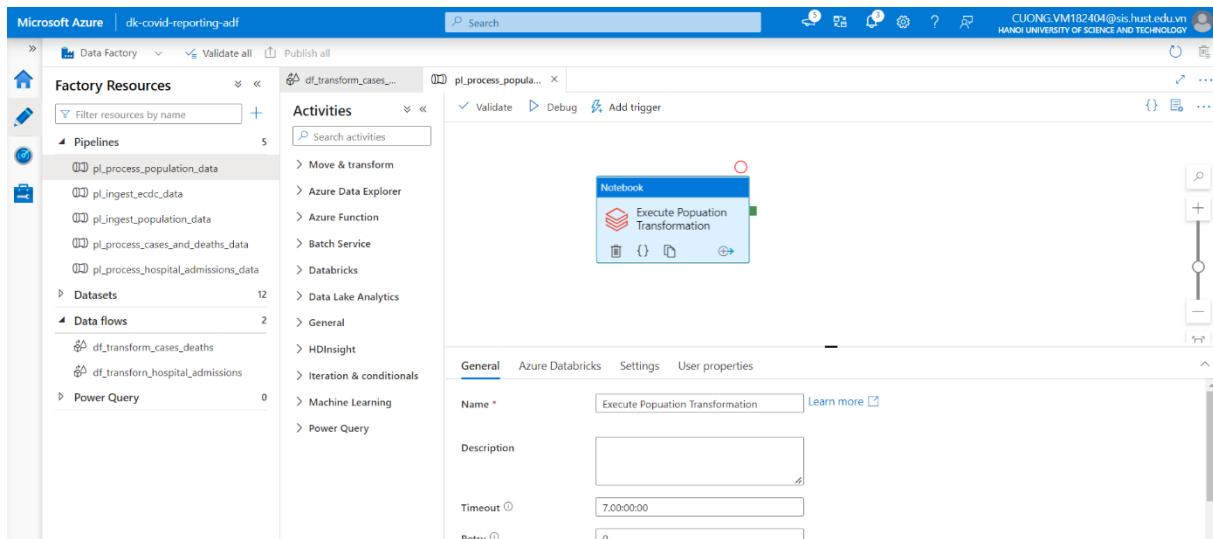
- Như trong Hình 3.25 ta có thể thấy, trong file trên có thể bỏ đi các cột từ 2008 đến 2018, chỉ để lại cột cuối để tính toán. Sau đó, ta có thể chia lại cột đầu tiên thành các cột nhỏ tương ứng với các nhóm tuổi và mã quốc gia. Đầu vào và đầu ra của quá trình chuyển đổi được thể hiện như trong Hình 3.26.

Transform Population By Age Data



Hình 3.26 Chuyển đổi bộ dữ liệu dân số theo độ tuổi

- Để thực hiện quá trình chuyển đổi, tôi sử dụng đường ống dữ liệu để Databricks chạy cluster có chứa Python Notebook. Thông qua đó, các script sẽ chuyển đổi dữ liệu để được đầu ra mong muốn. Pipeline như Hình 3.27



Hình 3.27 Pipeline chuyển đổi dữ liệu dân số với Databricks

- File Python chuyển đổi dữ liệu khi chạy trên Databricks như trong Hình 3.28.

Hình 3.28 Python Notebook chạy trên Cluster Databricks

- Kết quả thu được là các file như trong Hình 3.29.

Hình 3.29 Kết quả quá trình chuyển đổi

- ⇒ Như vậy, tôi đã hoàn tất quá trình chuyển đổi dữ liệu dân số theo độ tuổi với Spark Cluster trên Azure Databricks. Bộ dữ liệu đầu ra sẽ được sử dụng ở các phần tiếp theo.

3.3.2.5 Sao chép dữ liệu sang Database

Sau khi hoàn tất quá trình chuyển đổi dữ liệu theo thứ tự: Data Flow (1); Data Flow (2); HD Insight; Databricks, ta đã thu được tất cả các bộ dữ liệu như mong muốn. Tiếp theo đó, ta cần sao chép bộ dữ liệu trên từ Azure Data Lake sang Azure SQL Database để có thể sử dụng cho quá trình phân tích dữ liệu trong Power BI.

- Để thực hiện điều đó, trước tiên tôi cần tạo một Cơ sở dữ liệu mới với Azure SQL Database với các thông tin như trong Hình 3.30.

Hình 3.30 Cơ sở dữ liệu SQL trên Azure

- Sau khi đã tạo được một cơ sở dữ liệu để lưu dữ liệu, tôi cần tạo các bảng tạm chưa có dữ liệu với các trường tương tự như các bộ dữ liệu có trong Azure Data Lake Storage Gen 2 để chuẩn bị cho quá trình sao chép. Điều này được thực hiện bằng cách chạy cách dòng lệnh query trên trình Query Editor của Azure như trong Hình 3.31.

```

1 CREATE SCHEMA covid_reporting
2 GO
3
4 CREATE TABLE covid_reporting.cases_and_deaths
5 (
6     country           VARCHAR(100),
7     country_code_2_digit  VARCHAR(2),
8     country_code_3_digit  VARCHAR(3),
9     population        BIGINT,
10    cases_count       BIGINT,
11    deaths_count      BIGINT,
12    reported_date    DATE,
13    source            VARCHAR(500)
14 )
15 GO
16
17 CREATE TABLE covid_reporting.hospital_admissions_daily
18 (
19     country           VARCHAR(100),
20     country_code_2_digit  VARCHAR(2),

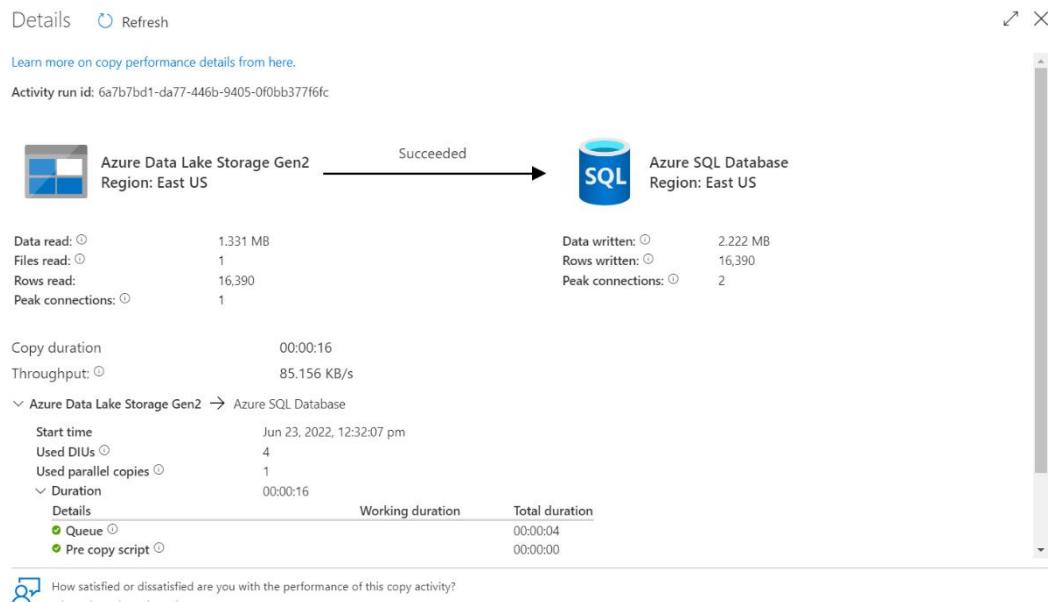
```

Hình 3.31 SQL Query Editor

- Kết quả thu được là các bảng chưa có dữ liệu như Hình 3.32

Hình 3.32 Kết quả tạo các bảng tạm trên cơ sở dữ liệu

- Sau khi đã có các bảng tạm để chứa dữ liệu. Tôi sẽ xây dựng một pipeline để sao chép dữ liệu từ Azure Data Lake Storage Gen 2 sang Azure SQL Database. Pipeline được xây dựng có kết quả như Hình 3.33: dữ liệu được chuyển từ Azure Data Lake Storage Gen 2 sang Azure SQL Database.



Hình 3.33 Kết quả sao chép dữ liệu vào SQL Database

- Tiếp theo, quay trở lại Query Editor để kiểm tra xem dữ liệu đã được sao chép thành công hay chưa. Kết quả được thể hiện trong Hình 3.34.

Query 1 × Query 2 × **Query 3** ×

Run Cancel query Save query Export data as Show only Editor

1 SELECT TOP (1000) * FROM [covid_reporting].[hospital_admissions_daily]

2

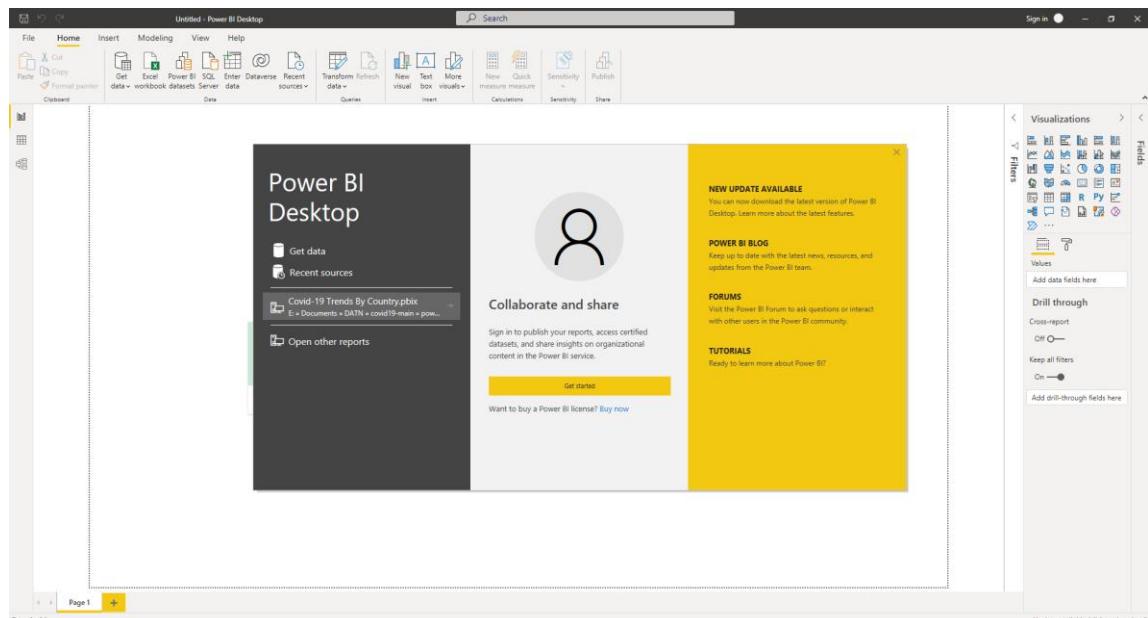
country	country_code_2_digit	country_code_3_digit	population	reported_date	hospital,occupancy_co...	icu_occupancy_count	source
Austria	AT	AUT	8858775	2020-10-25T00:00:00...	1225	174	Country_Website
Belgium	BE	BEL	11455519	2020-10-25T00:00:00...	4825	756	Country_Website
Bulgaria	BG	BGR	7000039	2020-10-25T00:00:00...	1976	138	External_Github
Czechia	CZ	CZE	10649800	2020-10-25T00:00:00...	5613	828	Country_Website
Denmark	DK	DNK	5806081	2020-10-25T00:00:00...	127	18	Country_Website
Estonia	EE	EST	1324820	2020-10-25T00:00:00...	29	4	Country_API
France	FR	FRA	67012883	2020-10-25T00:00:00...	16454	2575	Country_Website
Hungary	HU	HUN	9772756	2020-10-25T00:00:00...	2449	221	JRC

Hình 3.34 Kết quả quá trình sao chép dữ liệu vào Database

- Như trong Hình 3.34 có thể thấy các bảng đã được tạo vào trong các bảng đều đã có đầy đủ dữ liệu. Những dữ liệu này sẽ được Power BI lấy về và thực hiện trực quan hóa ở phần tiếp theo.

3.3.3 Trực quan hóa dữ liệu (Exploitation)

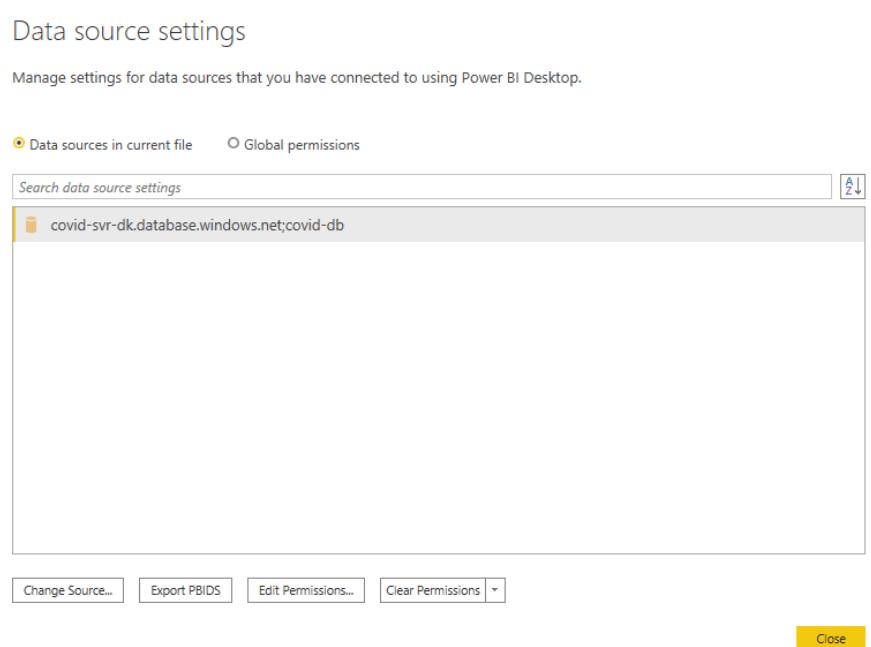
Từ những con số, những bộ dữ liệu đã được chuyển đổi thông qua điện toán đám mây Azure, để có thể biết được những con số đó thể hiện điều gì, ta cần phải trực quan hóa dữ liệu bằng các bảng biểu, biểu đồ. Điều đó được thực hiện dễ dàng với Microsoft Power BI – một công cụ phân tích dữ liệu vô cùng mạnh mẽ.



Hình 3.35 Giao diện Power BI

Hình 3.35 mô tả giao diện ban đầu của Power BI.

- Trước khi bắt đầu phân tích dữ liệu với Power BI, tôi cần phải liên kết Power BI tới Azure SQL Database để lấy dữ liệu như trong Hình 3.36. Cần kết nối đến server “covid-svr-dk.database.windows.net”, database “covid-db”.



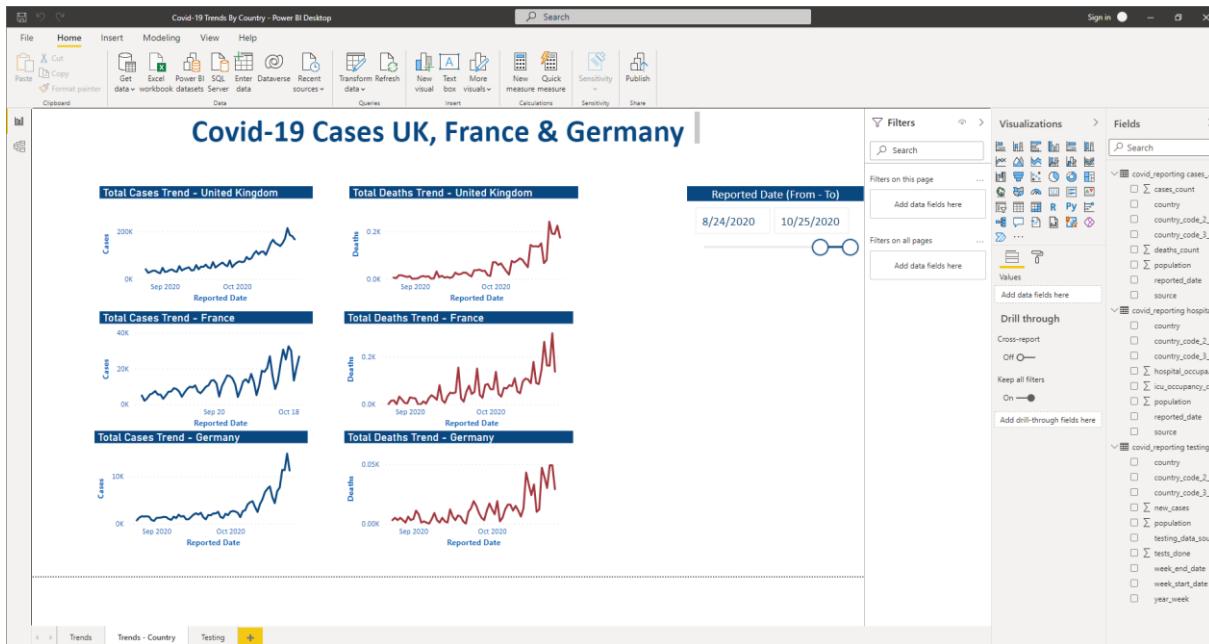
Hình 3.36 Kết nối Power BI với SQL Database

- Sau khi kết nối thành công, Power BI sẽ nhận được các bảng chứa dữ liệu từ SQL Database như Hình 3.37.

country	country_code_2_digit	country_code_3_digit	population	cases_count	deaths_count	reported_date	iso
Russia	RU	RUS	145934460	8849	127	5/22/2020	Epic
Germany	DE	DEU	83019213	2089	11	9/29/2020	Epic
Greece	EL	GRC	10724599	56	4	4/11/2020	Epic
Croatia	HR	HRV	4076246	91	1	3/28/2020	Epic
Malta	MT	MILT	495559	6	0	4/14/2020	Epic
Andorra	AD	AND	76177	0	0	2/5/2020	Epic
Bosnia and Herzegovina	BA	BIH	3280815	58	0	6/19/2020	Epic
Belarus	BY	BLR	9449321	0	0	2/25/2020	Epic
Iceland	IS	ISL	356991	0	0	5/20/2020	Epic
Latvia	LV	LVA	1919668	12	1	3/2/2020	Epic
Poland	PL	POL	37972812	371	15	7/3/2020	Epic
Belgium	BE	BEL	1145519	654	3	9/1/2020	Epic
Czechia	CZ	CZE	10649800	0	0	2/4/2020	Epic
Malta	MT	MILT	495559	68	2	10/13/2020	Epic
Denmark	DK	DNK	5806081	435	0	9/29/2020	Epic
Latvia	LV	LVA	1919968	6	1	4/8/2020	Epic
Austria	AT	AUT	8858775	143	0	3/14/2020	Epic
Malta	MT	MILT	495559	13	0	4/17/2020	Epic
Luxembourg	LU	LUX	613894	109	0	10/14/2020	Epic
Portugal	PT	PRT	10276617	0	0	1/29/2020	Epic
Albania	AL	ALB	2862427	23	0	5/31/2020	Epic
Andorra	AD	AND	76177	0	0	5/28/2020	Epic
Azerbaijan	AZ	AZE	10393175	43	0	4/4/2020	Epic
Montenegro	ME	MNE	622182	296	8	9/23/2020	Epic
Serbia	RS	SRB	6963764	83	1	9/25/2020	Epic
Armenia	AM	ARM	2963234	39	0	4/1/2020	Epic
North Macedonia	MK	MKD	2077132	30	3	5/18/2020	Epic
Slovakia	SK	SVK	5450421	41	0	3/22/2020	Epic
Lithuania	LT	LTU	2794184	58	3	4/17/2020	Epic

Hình 3.37 Dữ liệu đã được lấy vào Power BI

- Khi đã có được dữ liệu, tôi bắt đầu thực hiện quá trình trực quan hóa dữ liệu bằng cách xây dựng các bảng biểu, biểu đồ. Những biểu đồ này sẽ đưa ra những thông tin rõ ràng thay vì những con số trong bộ dữ liệu.



Hình 3.38 Trực quan hóa dữ liệu với Power BI

Hình 3.38 mô tả các biểu đồ được xây dựng trên Power BI từ dữ liệu trên Database.

- Sau khi hoàn thành quá trình trực quan hóa dữ liệu, tôi sẽ thực hiện giám sát các luồng dữ liệu đã được xây dựng. Việc này sẽ được tôi trình bày rõ hơn ở phần tiếp theo.

3.3.4 Quản lý luồng dữ liệu (Monitoring)

Ngay khi xây dựng được các đường ống dữ liệu hoàn chỉnh, tôi đều thêm các tín hiệu kích hoạt (trigger) với thời gian lặp cố định (thường là một lần/ngày). Điều này sẽ giúp cho dữ liệu được cập nhật liên tục và đảm bảo đường ống hoạt động trơn chu. Để thuận tiện cho quá trình bảo trì, nâng cấp hệ thống cũng như theo dõi các luồng dữ liệu, Azure Data Factory cung cấp cho người dùng công cụ quản lý tiện lợi “Monitor”.

The screenshot shows the Microsoft Azure Data Factory interface. On the left, a sidebar menu is open with several options: Dashboards, Runs, Pipeline runs (which is selected and highlighted with a red box), Trigger runs, Runtimes & sessions, Integration runtimes, Data flow debug, Notifications, and Alerts & metrics. The main area is titled 'Pipeline runs' and displays a table of recent runs. The table has columns for Pipeline name, Run start, Run end, Duration, Triggered by, Status, Error, Run, Parameters, Annotations, and Run ID. Five rows are listed, all showing a status of 'Succeeded'. The table is paginated with 'Showing 1 - 5 items' at the top.

Pipeline name	Run start	Run end	Duration	Triggered by	Status	Error	Run	Parameters	Annotations	Run ID
pl_sql_cases_and_deaths...	Jul 29, 2022, 7:08:40 am	Jul 29, 2022, 7:09:01 am	00:00:21	tr_sql_case_and_deat...	Succeeded		Original			c37fee2e-1456-4d51-4
pl_sql_hospital_admissions...	Jul 29, 2022, 7:06:24 am	Jul 29, 2022, 7:06:38 am	00:00:13	tr_sql_hospital_admi...	Succeeded		Original			c13ad17f-dd6d-40b6-
pl_process_cases_and_deaths...	Jul 29, 2022, 7:02:07 am	Jul 29, 2022, 7:08:29 am	00:06:21	tr_process_cases_an...	Succeeded		Original			49f06a44-4939-4675-
pl_process_hospital_admissio...	Jul 29, 2022, 7:02:01 am	Jul 29, 2022, 7:06:13 am	00:04:11	tr_process_hospital_...	Succeeded		Original			a1ab45cf-c092-4744-4
pl_ingest_ecdc_data	Jul 29, 2022, 7:01:16 am	Jul 29, 2022, 7:01:44 am	00:01:27	tr_ingest_ecdc_data	Succeeded		Original			40290866-1b45-4176

Hình 3.39 Giám sát các luồng dữ liệu với chức năng “Monitor”

- Các thông tin này sẽ giúp cho người dùng biết được đường ống dữ liệu của mình có hoạt động tốt hay không, nếu xảy ra lỗi thì sẽ báo cụ thể lỗi nằm ở đâu để người dùng có thể khắc phục kịp thời. Điều này sẽ đảm bảo cho hệ thống luôn hoạt động một cách tốt nhất.

Ngoài ra, Azure còn cung cấp cho người dùng các công cụ quản lý tài nguyên, chi phí đã trả cho các dịch vụ của Azure.

3.4 Kết luận chương

Chương 3 đã trình bày cụ thể về phương pháp thu thập, chuyển đổi và phân tích dữ liệu sử dụng các dịch vụ của Azure như: Azure Data Factory, Azure Data Lake Storage Gen 2, Azure SQL Database, Azure HD Insight, Azure Databricks,... Bên cạnh đó còn sử dụng Power BI để trực quan hóa dữ liệu là quản lý các luồn dữ liệu với Azure Data Factory. Chương này đã cho thấy sự tối ưu khi sử dụng điện toán đám mây để giải quyết các bài toán về dữ liệu. Trong chương tới, tôi sẽ tổng kết lại kết quả thu được và thảo luận thêm về đồ án này.

CHƯƠNG 4. KẾT QUẢ VÀ THẢO LUẬN

Chương 4 sẽ tập trung đưa ra kết quả của bài toán thông qua các số liệu trên các bảng biểu của báo cáo cũng như các số liệu phản hồi của các pipeline. Bên cạnh đó, dựa trên kết quả thu được, tôi đưa ra một số đánh giá và nhận xét về các yếu tố ảnh hưởng đến bài toán này.

4.1 Kết quả thu được

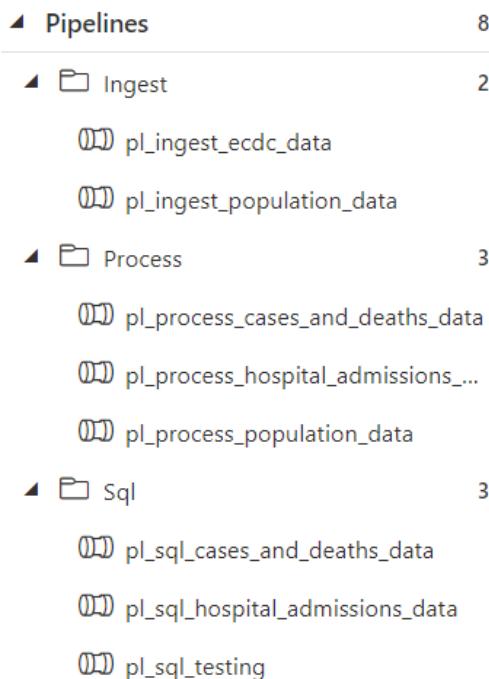
Kết quả của đề tài được thể hiện qua các phần sau:

- Các bộ dataset thu được
- Các pipeline, data flow đã xây dựng được
- Kết quả trực quan hóa được thể hiện qua các biểu đồ
- Kết quả giám sát quá trình vận hành của các đường ống dữ liệu (data pipeline)

4.1.1 Kết quả xây dựng hệ thống trên Azure Data Factory

Sau khi hoàn tất xây dựng và triển khai hệ thống, ta sẽ thu được những dataset, pipeline và data flow như các hình dưới đây:

- Pipeline:



Hình 4.1 Những pipeline đã xây dựng được

- Dataset:

Datasets		16
└ Lookup	3	
└ ds_country_lookup		
└ ds_dim_date_lookup		
└ ds_ecdc_file_list		
└ Process	4	
└ ds_process_testing		
└ ds_processed_cases_and_deaths		
└ ds_processed_hospital_admissio...		
└ ds_processed_hospital_admissio...		
└ Raw	6	
└ ds_ecdc_raw_csv_dl		
└ ds_ecdc_raw_csv_http		
└ ds_population_raw_gz		
└ ds_population_raw_tsv		
└ ds_raw_cases_and_deaths		
└ ds_raw_hospital_admissions		
└ Sql	3	
└ ds_sql_cases_and_deaths		
└ ds_sql_hospital_admissions		
└ ds_sql_testing		

Hình 4.2 Những bộ dataset trước và sau khi xử lý dữ liệu

- Data Flow:

Data flows		2
└ df_transform_cases_deaths		
└ df_transform_hospital_admissions		

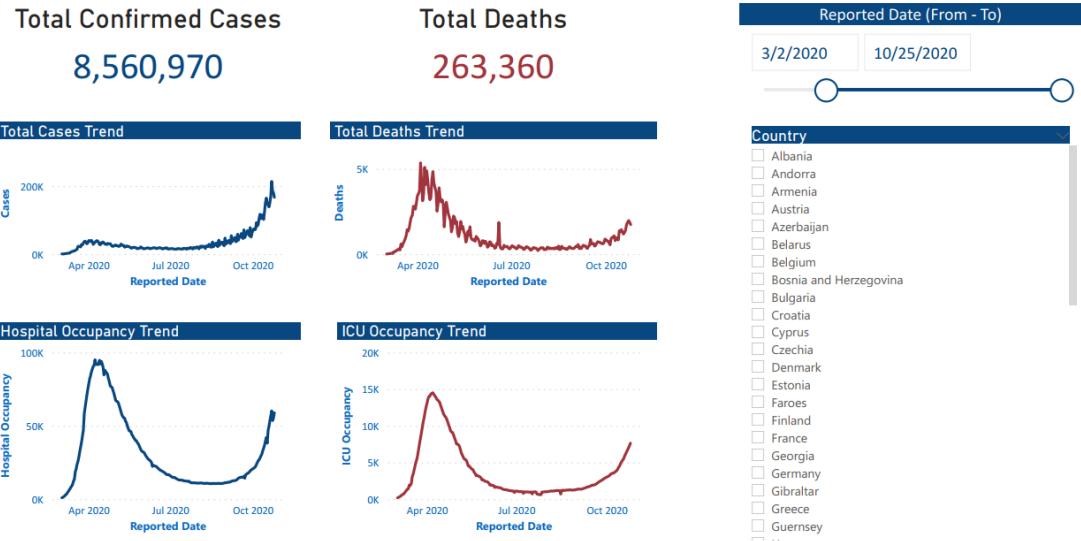
Hình 4.3 Những Data Flow đã xây dựng

4.1.2 Kết quả trực quan hóa dữ liệu

Những bộ dữ liệu chỉ toàn những con số từ ở nhiều nguồn khác nhau, sau khi đi qua các khối như trong phần giải pháp thiết kế đã đưa ra, ta sẽ thu được những bản báo

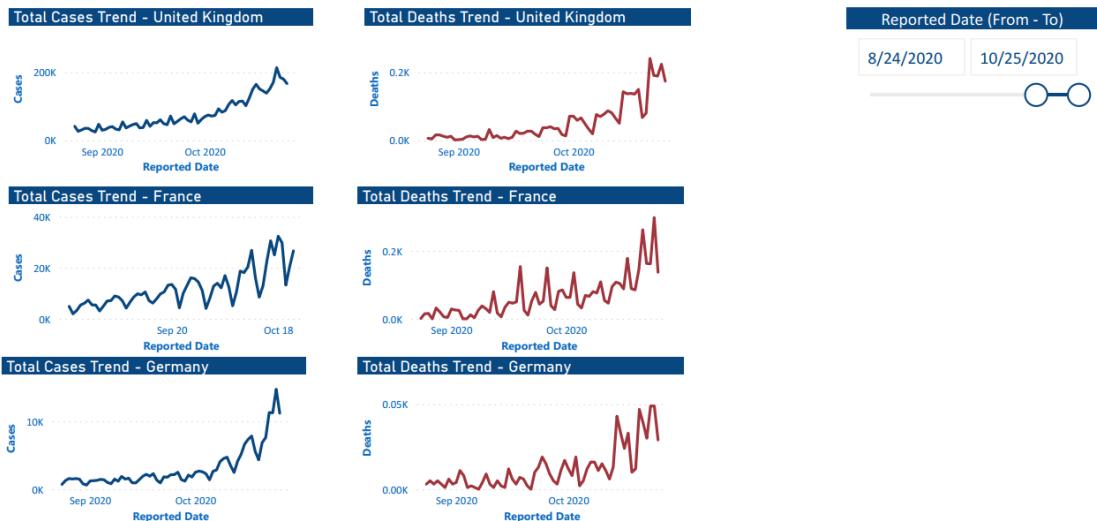
cáo trực quan nhất, dễ hiểu nhất để phân tích tình hình của dịch bệnh. Kết quả được thể hiện trong các hình dưới đây:

Covid-19 Cases EU/EEA & UK



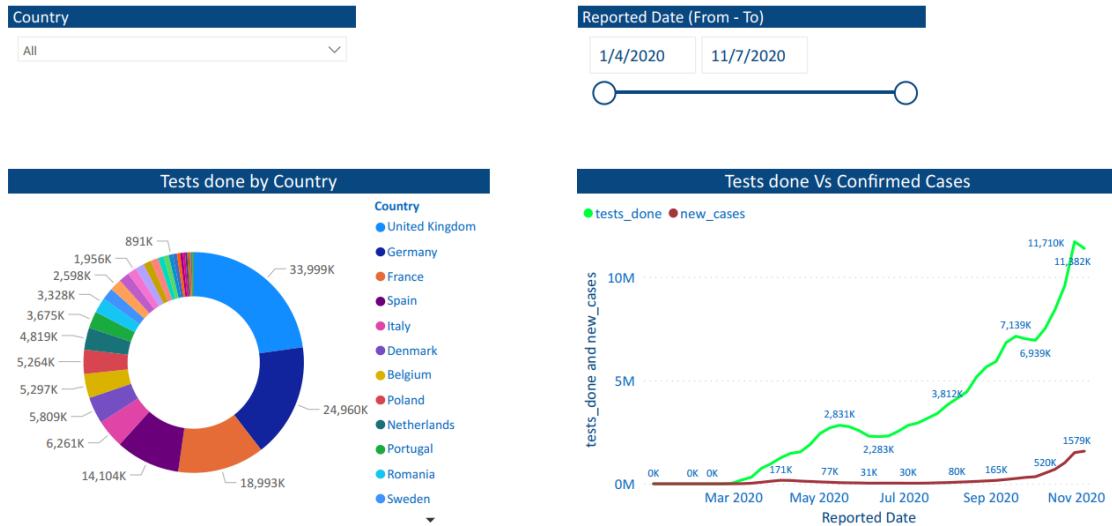
Hình 4.4 Số ca nhiễm và tử vong ở EU/EEA & UK từ 2/3/2020 đến 25/10/2020

Covid-19 Cases UK, France & Germany



Hình 4.5 Số ca nhiễm và tử vong ở UK, France & Germany từ 24/8/2020 đến 25/10/2020

Covid-19 Testing EU/EEA & UK

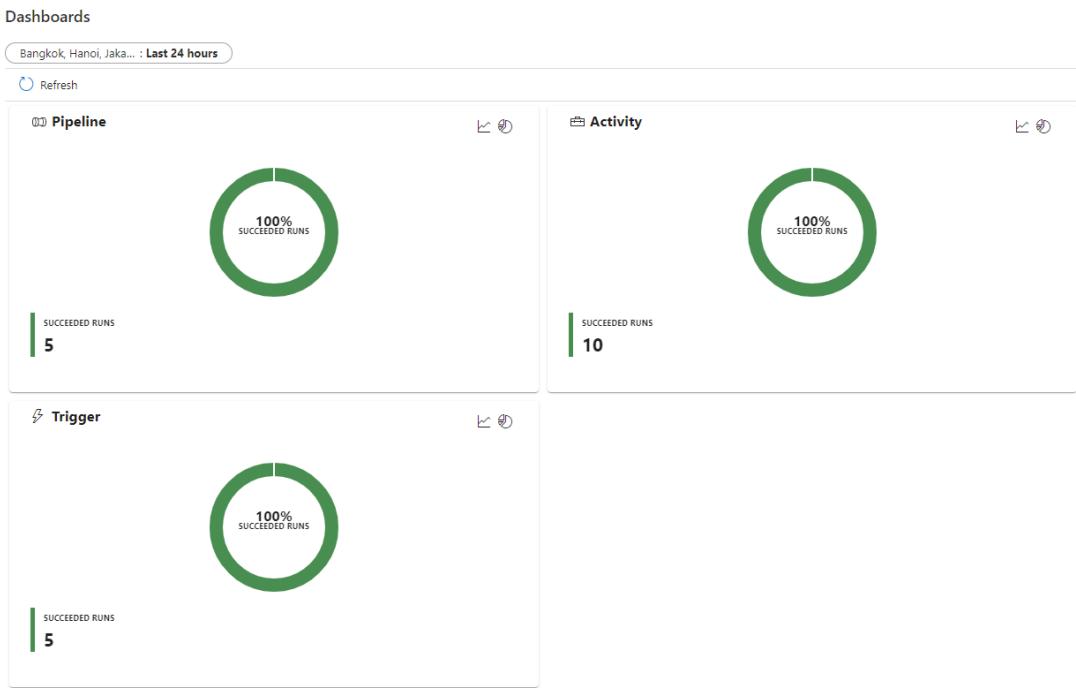


Hình 4.6 Tỉ lệ tiêm Vaccine ở khu vực EU/EEA & UK

⇒ Từ những biểu đồ trên ta có thể đọc được những thông tin mà dữ liệu đã đưa ra theo từng khu vực, từng mốc thời gian khác nhau. Điều này sẽ giúp cho người nhận được những báo cáo dễ hình dung hơn về tình hình dịch bệnh tại thời điểm đó hay vì đọc những con số.

4.1.3 Kết quả giám sát quá trình vận hành của các luồng dữ liệu (data pipeline)

Azure Data Factory cung cấp công cụ cho người dùng có thể quản lý các luồng dữ liệu của mình một cách đơn giản với các biểu đồ trực quan. Kết quả quản lý luồng dữ liệu được thể hiện như Hình 4.7. Khi có một luồng dữ liệu vận hành không đúng như lúc cấu hình, hệ thống sẽ gửi thông báo về Email để người dùng kịp thời phát hiện và bảo trì hệ thống. Điều này sẽ đảm bảo cho việc các luồng dữ liệu luôn hoạt động tốt và đưa ra kết quả đúng như mong muốn.



Hình 4.7 Biểu đồ quản lý vận hành luồng dữ liệu

Như trên Hình 4.7 ta có thể thấy tất cả các đường ống dữ liệu (pipeline) cũng như các tín hiệu kích hoạt (trigger) đều hoạt động hiệu quả, không xảy ra bất kỳ lỗi nào.

4.2 Đánh giá kết quả

Từ kết quả thu được có thể thấy dữ liệu từ nhiều nguồn khác nhau đã được lấy về, chuyển đổi và phân tích trực quan để đưa ra các góc nhìn cụ thể hơn về dữ liệu. Bên cạnh đó, kết quả quá trình làm việc với dữ liệu trên Azure cũng cho ta thấy được những ưu điểm của việc ứng dụng giải pháp đám mây Microsoft Azure vào việc xử lý các bài toán dữ liệu.

Bên cạnh những ưu điểm về mặt kỹ thuật, chi phí, thông qua các bước thực hiện được đề cập đến trong báo cáo ta có thể thấy việc thực hiện xây dựng hệ thống trên Azure là rất dễ dàng, chỉ cần bám sát các tài liệu được cung cấp bởi Microsoft và các kỹ thuật cốt lõi của dữ liệu lớn là có thể xây dựng được các đường ống dữ liệu.

4.3 Kết luận chương

Như vậy, Chương 4 đã đưa ra kết quả thực nghiệm thu được sau khi triển khai hệ thống thu thập, chuyển đổi và phân tích dữ liệu tự động với Solution Architecture và bộ dữ liệu nêu trong Chương 3. Bên cạnh đó, chương này đưa ra những đánh giá về kết quả thu được.

KẾT LUẬN

Kết luận chung

Như vậy, đồ án đã đưa ra phương pháp giải quyết các bài toán về dữ liệu như: lấy dữ liệu, chuyển đổi dữ liệu và trực quan hóa dữ liệu. Trong báo cáo trình bày chi tiết về lý thuyết cơ bản của dữ liệu lớn, các công cụ hỗ trợ, giới thiệu về dịch vụ đám mây Microsoft Azure cũng như cách triển khai hệ thống làm việc với dữ liệu trên Azure Data Factory và phân tích dữ liệu với Power BI. Tôi đã thực hiện trực tiếp trên Azure Portal và Power BI với bộ dữ liệu Covid-19 và dân số của khu vực UK.

Kết quả thực nghiệm thu được sau khi xây dựng và triển khai hệ thống cho thấy rằng phương pháp sử dụng các dịch vụ đám mây vào xử lý các bài toán dữ liệu là phương pháp tối ưu và đem lại nhiều lợi ích cho người dùng cả về mặt kỹ thuật và mặt kinh tế. Không những vậy, những đánh giá còn cho thấy rằng việc làm việc với Azure rất dễ dàng và tiện lợi. Như vậy, có thể thấy việc ứng dụng giải pháp đám mây (cụ thể là Microsoft Azure) vào giải quyết bài toán dữ liệu (ở đây là xây dựng báo cáo Covid-19) sẽ là giải pháp đáng được quan tâm bởi các cá nhân, doanh nghiệp lớn.

Hướng phát triển

Mặc dù đã có được những kết quả nhất định nhưng do thời gian còn hạn chế nên tôi chưa mở rộng được hệ thống với các bộ dữ liệu lớn hơn. Vì vậy trong tương lai, tôi hi vọng có thể tiến hành áp dụng phương pháp đề xuất trên vào nhiều mô hình dữ liệu phức tạp. Bên cạnh đó, sau khi đã có được các bộ dữ liệu, tôi sẽ tiến hành xây dựng các thuật toán ML-AI để dự đoán các số liệu trong tương lai. Đây hiện đang là xu hướng của công nghệ, vì vậy tôi hi vọng có thể sớm triển khai được vấn đề này trong tương lai.

TÀI LIỆU THAM KHẢO

- [1] Microsoft. (2019). *Azure Data Factory: Data Ingtegration in the Cloud* [Online]. Available: https://azure.microsoft.com/mediahandler/files/resourcefiles/azure-data-factory-data-integration-in-the-cloud/Azure_Data_Factory_Data_Integration_in_the_Cloud.pdf.
- [2] Y. Xu, S. G. Sugiyama and L. Lovosevic. (2018). *Data Migration from on-premise relational Data Warehouse to Azure using Azure Data Factory* [Online]. Available: https://azure.microsoft.com/mediahandler/files/resourcefiles/data-migration-from-on-premise-relational-data-warehouse-to-azure-data-lake-using-azure-data-factory/Data_migration_from_on-prem_RDW_to_ADLS_using_ADF.pdf. [Accessed 5 July 2022].
- [3] B. John and I. BA. (2019). *Azure Data Factory With Azure DevOps* [Online]. Available: <https://azure.microsoft.com/mediahandler/files/resourcefiles/whitepaper-adf-on-azuredevops/Azure%20data%20Factory-Whitepaper-DevOps.pdf>. [Accessed 5 July 2022].
- [4] B. John, I. BA, Y. Xu and Gaurav. (2019). *Azure Data Factory-Passing Parameters* [Online]. Available: <https://azure.microsoft.com/mediahandler/files/resourcefiles/azure-data-factory-passing-parameters/Azure%20data%20Factory-Whitepaper-PassingParameters.pdf>. [Accessed 5 July 2022].
- [5] S. Winarko, M. Kromer and Microsoft. (2018). *Azure Data Factory: SSIS in the Cloud* [Online]. Available: https://azure.microsoft.com/mediahandler/files/resourcefiles/azure-data-factory-ssis-in-the-cloud/Azure_Data_Factory_SSIS_in_the_Cloud.pdf. [Accessed 5 July 2022].
- [6] Microsoft. "Introduction to Azure Blob storage," [Online]. Available: <https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blobs-introduction>. [Accessed 7 2022].
- [7] Microsoft, "Azure Data Lake Storage Gen 2 Introduction," [Online]. Available: <https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage>

introduction. [Accessed 7 2022].

- [8] Microsoft, "What is the Azure SQL Databases service?," [Online]. Available: <https://docs.microsoft.com/en-us/azure/azure-sql/database/sql-database-paas-overview?view=azuresql>. [Accessed 7 2022].
- [9] Microsoft, "Azure Databricks documentation," [Online]. Available: <https://docs.microsoft.com/en-us/azure/databricks/>. [Accessed 7 2022].
- [10] Microsoft, "Azure HDInsight documentation," [Online]. Available: <https://docs.microsoft.com/en-us/azure/hdinsight/>. [Accessed 7 2022].
- [11] I. Bill, *Data Lake Architecture - Designing the Data Lake and Avoiding the Garbage Dump*. Technics Publications, 2016.

