1. MỞ ĐẦU

"Big Data" được biết như là một giải pháp lý tưởng để xử lý một tập dữ liệu lớn có cấu trúc, bán cấu trúc hoặc là phi cấu trúc như dữ liệu từ các weblogs, mạng xã hội, e-mail, cảm biến và các bức ảnh mà có thể được khai thác nhằm tìm ra những thông tin hữu ích. Big Data trên thực tế đang được ứng dụng vào rất nhiều lĩnh vực của xã hội, tạo những chuyển biến ấn tượng, giúp tăng hiệu quả và năng suất của doanh nghiệp. Trong lĩnh vực ngân hàng, Big Data được sử dụng bởi các kỹ thuật phân cụm dữ liệu giúp đưa ra những quyết định quan trọng. Trong lĩnh vực thương mại Big Data cùng với các kỹ thuật phân tích dữ liệu để nhanh chóng xác định những địa điểm, chi nhánh nơi tập trung nhiều nhu cầu của khách hàng tiềm năng và đề xuất thành lập chi nhánh mới. Đối với lĩnh vực tài chính, Big Data có thể kết hợp với nhiều quy tắc được áp dụng trong 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. Trong lĩnh vực y tế, Big Data không chỉ được ứng dụng để xác định phương hướng điều trị mà giúp cải thiện và hỗ trợ quá trình chăm sóc sức khỏe của bệnh nhân.

Năm 2014, công ty phân tích dữ liệu Gartner [3] đưa ra khái niệm mới có thể chấp nhận được về mô hình "5Vs" của Big Data. Đó là, 5 đặc trưng của Big Data: tăng về lượng (volume), tăng về vận tốc (velocity) , tăng về chủng loại (variety), tăng về độ chính xác (veracity) và tăng về giá trị thông tin (value). Hiểu một cách đơn giản, đó chính là sự phát triển không ngừng của khối lượng dữ liệu cần lưu trữ, cách thức để xử lý dữ liệu với tốc độ cao, tính đa dạng dữ liệu (variety): Theo IBM [1], chỉ có 20% dữ liệu thu được là có cấu trúc nhưng thực tế là 80% dữ liệu trên thế giới đều là dạng phi cấu trúc hoặc bán cấu trúc. Ngoài ra còn phải quản lý được các dữ liệu mới được tạo ra và các dữ liệu được cập nhật, độ chính xác của xử lý và giá trị thông tin được lưu trữ.

Tất cả các quan điểm trên đều hướng tới việc trả lời cho câu hỏi: Big Data là vấn đề gì và tại sao chúng ta cần phải nghiên cứu và tìm hiểu nó. Vấn đề này được các nhà cung cấp dịch vụ, các trung tâm tích hợp dữ liệu nghiên cứu, tìm hiểu phương pháp tốt nhất để lưu trữ loại dữ liệu với 5 yếu tố trên. Nhìn chung, có bốn lợi ích chính mà Big Data có thể mang lại đó là: cắt giảm chi phí, giảm thời gian tìm kiếm thông tin, tăng thời gian phát triển và tối ưu hóa sản phẩm, đồng thời hỗ trợ con người đưa ra những quyết định đúng đắn và hợp lý. Trong phạm vi bài báo này chúng tôi đề xuất giải pháp xử lý của Big Data trong lĩnh vực y tế, nơi sản sinh một tập dữ liệu khổng lồ từ các xét nghiệm, điều trị SARS-CoV-2 trong hơn 2 năm qua.

2. BIG DATA VÀ CÔNG CỤ XỬ LÝ

Kỹ thuật xử lý dữ liệu trong Big Data chủ yếu là NoSQL (cơ sở dữ liệu theo cột, cặp khóa-giá trị) [4], do mô hình dữ liệu quan hệ không thể đáp ứng được các

loại dữ liệu bán cấu trúc và không cấu trúc. Trong thực tế nhiều công ty lớn cũng đã đưa ra các công cụ khác nhau để xử lý Big Data.

2.1 Giải pháp Hadoop/MapReduce

Năm 2004, Google công bố kiến trúc của hệ thống file phân tán GFS (Google File System) và công cụ MapReduce. Từ đó Hadoop, một Framework, cùng với GFS và MapReduce được ra đời bởi Dough Cutting để xử lý các Big Data.

2.1.1 Hadoop

Apache Hadoop [5] là một framework dùng để chạy những ứng dụng trên một cluster lớn được xây dựng trên những phần cứng thông thường. Thư viện phần mềm Hadoop là một khuôn mẫu cho phép xử lý phân tán các bộ dữ liệu lớn trên các nhóm máy tính sử dụng các mô hình lập trình đơn giản. Nó được thiết kế để mở rộng từ một máy chủ duy nhất sang hàng ngàn máy khác, mỗi máy cung cấp tính toán và lưu trữ cục bộ. Hadoop thực hiện mô hình MapReduce, đây là mô hình phân tán song song, ứng dụng sẽ được chia nhỏ ra thành nhiều phân đoạn dữ liệu khác nhau (phân tán), và các phần này sẽ được thực hiện trên đồng thời trên nhiều node khác nhau (song song). Bên cạnh đó, Hadoop cung cấp một hệ thống file phân tán (HDFS) cho phép lưu trữ dữ liệu lên trên nhiều node. Cả Map/Reduce và HDFS đều được thiết kế sao cho framework sẽ tự động quản lý được các lỗi, các hư hỏng về phần cứng của các node.

2.1.2 MapReduce

Có thể hiểu một cách đơn giản, MapReduce phân chia các công việc xử lý thành nhiều khối công việc nhỏ, phân tán khắp các nút tính toán (giai đoạn Map), rồi thu hồi các kết quả (giai đoạn Reduce). MapReduce có thể chạy trên các phần cứng thông thường, không đòi hỏi các server chạy MapReduce phải là các máy tính có cấu hình cao với khả năng tính toán, lưu trữ và truy xuất mạnh mẽ. Do đó, chi phí triển khai MapReduce sẽ rẻ hơn. MapReduce làm đơn giản hoá các giải thuật tính toán phân tán bằng cách chỉ cần cung cấp hai hàm Map và Reduce cùng với một số thành phần xử lý dữ liệu đầu vào.

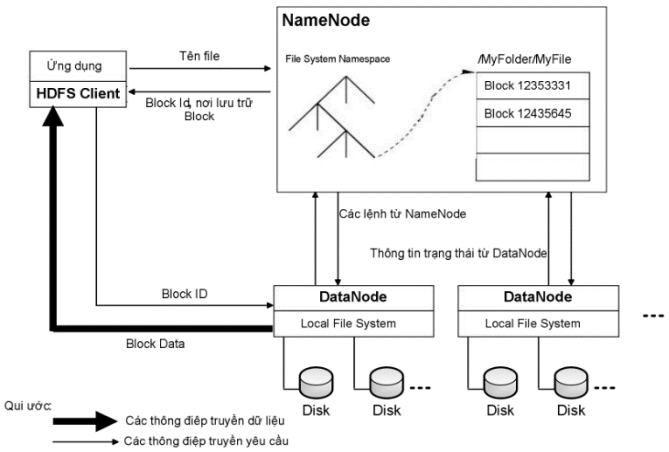
Hàm Map: Người dùng đưa một cặp dữ liệu (key, value) làm input cho hàm Map, và tùy vào mục đích của người dùng mà hàm Map sẽ trả ra danh sách các cặp dữ liệu (intermediate key, value).

Hàm Reduce: Hệ thống sẽ gom nhóm tất cả value theo intermediate key từ các output của hàm map, để tạo thành tập các cặp dự liệu với cấu trúc là (key, tập các value cùng key). Dữ liệu input của hàm Reduce là từng cặp dữ liệu được gom nhóm ở trên và sau khi thực hiện xử lý nó sẽ trả ra cặp dữ liệu (key, value) output cuối cùng cho người dùng. Cho đến nay, Hadoop đã trở thành giải pháp nguồn mở hàng đầu hỗ trợ mô hình MapReduce. Hadoop được viết bằng Java, tuy nhiên hỗ

trợ phát triển MapReduce trên nhiều ngôn ngữ khác ngoài Java như C++, Pearl, Python, …

2.2 Kiến trúc Hadoop File System (HDFS)

Giống như các hệ thống file khác, HDFS duy trì một cấu trúc cây phân cấp các file [6], thư mục mà các file sẽ đóng vai trò là các node lá. Trong HDFS, mỗi file sẽ được chia ra làm một hay nhiều block và mỗi block này sẽ có một block ID để nhận diện. Mỗi block của file sẽ được lưu trữ thành ra nhiều bản sao khác nhau vì mục đích an toàn dữ liệu.



Hình 1. Kiến trúc của HDFS

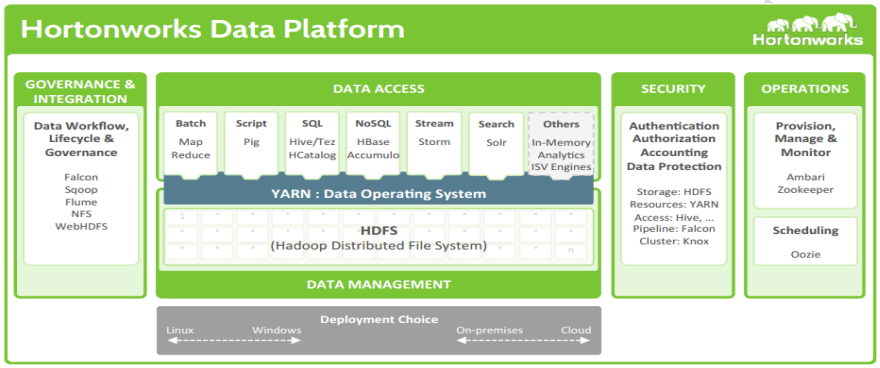
2.3 Xử lý Big Data với Hornworks Sanbox

2.3.1 Giới thiệu Hortonworks Data Platform

Hortonworks Data Platform (HDP) là một nền tảng phát triển và xây dựng hoàn toàn mở, HDP được thiết kế để đáp ứng nhu cầu xử lý dữ liệu lớn của doanh nghiệp. HDP là linh hoạt, cung cấp khả năng mở rộng tuyến tính, mở rộng lưu trữ và tính toán trên một loạt các phương pháp truy cập, batch và real-time, search và streaming. Nó bao gồm một tập hợp toàn diện các khả năng xử lý dữ liệu cho doanh nghiệp như: governance, integration, security và operation. HDP cho phép triển khai Hadoop bất cứ nơi nào, từ cloud cho đến hệ thống tại chỗ, trên cả Linux và Windows.

Sự xuất hiện và bùng nổ của các loại dữ liệu mới trong những năm gần đây đã gây áp lực không nhỏ về kiến trúc dữ liệu cho các tổ chức. Để đối phó, nhiều người đã chuyển sang Apache Hadoop để quản lý sự bùng nổ của dữ liệu.

Apache Hadoop có nguồn gốc là để quản lý và truy cập dữ liệu, và chỉ bao gồm 2 thành phần là: Hadoop Distributed File System (HDFS) và MapReduce, một khuôn khổ xử lý cho dữ liệu lưu trữ trong HDFS. Theo thời gian, nền tảng Hadoop mở rộng kết hợp với một loạt các dự án khác để thành một nền tảng hoàn chỉnh. Nền tảng này chia thành 5 loại sau: data access, data management, security, operations và governance.



Hình 2. Cấu trúc của HDP

2.3.2 Các thành phần của Hortonworks Data Platform

- Data Management: Quản lý lưu trữ và xử lý tất cả các dữ liệu, bao gồm các thành phần nền tảng của HDP là Apache Hadoop YARN và Hadoop Distributed File System (HDFS).
- Data Access: HDP cho phép nhiều công cụ xử lý dữ liệu bằng ngôn ngữ truy vấn có cấu trúc như SQL hoặc truy cập dữ liệu với độ trễ thấp như NoSQL. Truyền phát thời gian thực đến khoa học dữ liệu và xử lý hàng loạt để sử dụng dữ liệu được lưu trữ trong một nền tảng duy nhất.
- Governance and Integration: HDP mở rộng data access và management với các công cụ mạnh mẽ để quản lý và tích hợp
- Security: HDP cung cấp một cách tiếp cận tập trung để quản lý, cho phép triển khai chính sách bảo mật một cách nhất quán trong suốt nền tảng dữ liệu với các yêu cầu để xác thực, cấp phép, kiểm tra và bảo vệ dữ liệu.

2.3.3 Cài đặt Hortonworks Sandbox (HWS)

- Cài đặt HWS, phiên bản 2.6.5 trên hệ điều hành Windows 10.
- Cấu hình máy tính: CPU: 64 bit (x64-based processor); Hệ điều hành: Windows 64 bit; Bộ nhớ chính Ram: > 8 GB;

Tải các phần mềm:

- Oracle VM VirtualBox mới nhất dành cho Windows tại link: https://www.virtualbox.org/wiki/Downloads
- Hortonworks Data Platform (HDP) on Hortonworks Sandbox, tại link: https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html

3. ỨNG DỤNG BIGDATA BÀI TOÁN TRONG LĨNH VỰC Y TẾ

3.1 Giới thiệu

Big Data trong y tế được sử dụng để mô tả và thống kê khối lượng thông tin khổng lồ được tạo ra từ việc áp dụng công nghệ kỹ thuật số để thu thập hồ sơ của bệnh nhân và giúp quản lý hoạt động của bệnh viện hoặc phục vụ cho việc

nghiên cứu về một sự kiện y tế. Công việc này là quá lớn và phức tạp đối với các công nghệ truyền thống.

Việc áp dụng phân tích dữ liệu lớn trong y tế mang lại rất nhiều kết quả tích cực và cũng là cứu sống con người. Về bản chất, Big Data đề cập đến lượng thông tin khổng lồ được tạo ra bởi quá trình số hóa mọi thứ, được tổng hợp và phân tích bằng các công nghệ cụ thể. Được áp dụng cho y tế, nó sẽ sử dụng dữ liệu sức khỏe cụ thể của một dân số (hoặc của một cá nhân cụ thể) và có khả năng giúp ngăn ngừa dịch bệnh, chữa bệnh, cắt giảm chi phí, ...

Trong nhiều năm qua, việc thu thập một lượng lớn dữ liệu cho mục đích y tế rất tốn kém và mất thời gian. Với công nghệ luôn cải tiến ngày nay, việc thu thập dữ liệu đó trở nên dễ dàng hơn, tạo báo cáo chăm sóc sức khỏe toàn diện và chuyển đổi chúng thành những thông tin chi tiết quan trọng có liên quan, sau đó có thể sử dụng để cung cấp dịch vụ chăm sóc tốt hơn.

3.2 Mô tả bài toán

3.2.1 Dữ liệu COVID-19

COVID-19 là một bệnh truyền nhiễm do vi-rút SARS-CoV-2 gây ra. Hầu hết người mắc bệnh COVID-19 sẽ gặp các triệu chứng từ nhẹ đến trung bình và hồi phục mà không cần phải điều trị đặc biệt. Tuy nhiên, một số người sẽ chuyển bệnh nghiêm trọng và cần được hỗ trợ y tế. Vi-rút này có thể lây từ miệng hoặc mũi của người bị nhiễm bệnh dưới dạng các giọt nhỏ khi họ ho, hắt hơi, nói chuyện, hát hoặc thở. Chúng ta có thể bị nhiễm bệnh khi hít phải vi-rút nếu đang ở gần người nhiễm COVID-19 hoặc chạm vào bề mặt có vi-rút rồi lại chạm tay vào mắt, mũi hoặc miệng. Vi-rút dễ lây lan hơn trong nhà và ở những nơi đông đúc.

- Các triệu chứng thường gặp nhất: sốt, ho, mệt mỏi, mất vị giác hoặc khứu giác,
- Các triệu chứng ít gặp hơn: đau họng, đau đầu, đau nhức, tiêu chảy, da nổi mẩn hay ngón tay hoặc ngón chân bị tấy đỏ hoặc tím tái, mắt đỏ hoặc ngứa.

Dữ liệu COVID-19 của mỗi cá nhân về quá trình tiêm chủng và điều trị rất khó tìm kiếm và sử dụng hợp pháp vì liên quan đến bảo mật dữ liệu cá nhân. Chúng tôi đã sử dụng dữ liệu công khai của Jonhn Hoppkin University và dữ liệu tiêm chủng của các nước ở châu Âu trong giai đoạn 2020-2022 để lưu trữ và xử lý trên môi trường HDP. Dữ liệu này có thể được download và lưu trữ dưới dạng file XLSx, CSV, JSON, XML.

Chúng ta sẽ thử nghiệm HWS với tập tin dữ liệu chuẩn "Data", kích thước 4,3 MB, gồm 160.782 mẫu tin với 14 fields, chứa tất cả dữ liệu Covid trong 2 năm 2020, 2021 của châu Âu. Tập tin này được download từ link: https://www.ecdc.europa.eu/en/publications-data/data-covid-19-vaccination-eu-eea

Bảng 2. Diễn giải các trường của tập tin dữ liệu chuẩn

YearWeekISO	Thời gian (tuần) nhận vac xin
ReportingCountry	Mã quốc gia theo chuẩn ISO 3166-1-alpha-2
Denominator	Tổng số dân số cho từng nhóm đối tượng được liệt kê trong TargetGroup
NumberDosesReceived	Số liều vac xin được nhận trong thời gian (tuần) báo cáo
FirstDose	Số liều vắc xin đầu tiên đã tiêm cho các cá nhân trong tuần báo cáo.
FirstDoseRefused	Số cá nhân từ chối tiêm
SecondDose	Số vắc xin liều thứ hai được tiêm cho các cá nhân trong tuần báo cáo
UnknownDose	Số liều được sử dụng trong tuần báo cáo mà loại liều không được chỉ định (tức là số lượng không biết đó là liều thứ nhất hay thứ hai).
Region	Region = country code
TargetGroup	Nhóm đối tượng tiêm phòng, bao **gồm**: ALL: người trên 18; <18; ..., HCW: người chăm sóc y tế; LTCF: người được chăm sóc dài hạn; ...
Vaccine	Tên vắc xin. Các vắc xin bổ sung sẽ được bổ sung khi được phê duyệt hoặc theo yêu cầu.
Population	Dân số theo độ tuổi cụ thể cho mỗi quốc gia

Xét truy vấn:

```
SELECT ReportingCountry, CONCAT(ROUND((sum(FirstDose) / max(Denominator) * 100),2),'%') as UptakePercentage
    FROM ecdc_covid19_vaccine_tracker
    WHERE ReportingCountry in ('DE','ES','DK') AND TargetGroup = 'ALL'
    GROUP BY ReportingCountry
```

Nếu sử dụng SQL hoặc Hortonworks Sandbox (HWS) thì chúng ta có được kết quả như sau:

ReportingCountry	UptakePercentage
DE	75.7%
DK	87.99%
ES	86.66%

Mặc dù từ một câu truy vấn cả SQL và HWS đều cho cùng một kết quả nhưng vấn đề đặt ra ở đây là chúng ta cần biết khả năng xử lý và lưu trữ của mỗi phương pháp với các ưu nhược điểm của nó như thế nào.

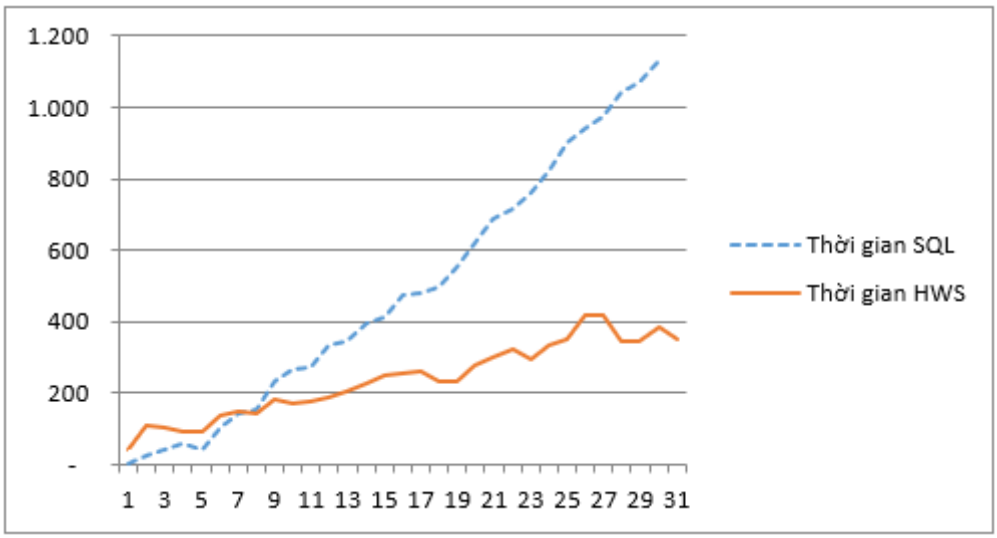
Thực hiện 31 lần test với tập tin dữ liệu chuẩn và các tập tin được nhân bản từ tập tin dữ liệu chuẩn chúng tôi được quả như sau:

Bảng 2. Số liệu thực nghiệm từ SQL và HWS

Lần xử lý	Số mẫu tin	SQL Server		Hortonworks Sandbox	
		Dung lượng lưu trữ	Thời gian cho kết quả	Dung lượng lưu trữ	Thời gian cho kết quả
1	160.782	47 MB	1s	27 MB	27s 672ms
2	3.067.600	900 MB	23s	543 MB	1m 48s 567ms
3	6.135.200	1.800 MB	43s	1.087 MB	1m 45s 923ms
4	9.202.800	2.701 MB	57s	1.631 MB	1m 30s 217ms
5	12.270.400	3.601 MB	40s	2.175 MB	1m 32s 716ms
6	15.338.000	4.502 MB	1m 45s	2.719 MB	2m 19s 243ms
7	18.405.600	5.402 MB	2m 21s	3.263 MB	2m 26s 7ms
8	21.473.200	6.303 MB	2m 33s	3.807 MB	2m 23s 479ms
9	24.540.800	7.203 MB	3m 51s	4.351 MB	3m 2s 899ms
10	27.608.400	8.103 MB	4m 25s	4.895 MB	2m 52s 676ms
11	30.676.000	9.004 MB	4m 32s	5.439 MB	2m 59s 401ms
12	33.743.600	9.904 MB	5m 31s	5.983 MB	3m 7s 723ms
13	36.811.200	10.805 MB	5m 45s	6.527 MB	3m 25s 444ms
14	39.878.800	11.705 MB	6m 37s	7.071 MB	3m 49s 642ms
15	42.946.400	12.606 MB	6m 50s	7.615 MB	4m 11s 205ms
16	46.014.000	13.506 MB	7m 53s	8.159 MB	4m 17s 28ms
17	49.081.600	14.407 MB	8m 00s	8.703 MB	4m 18s 868ms
18	52.149.200	15.307 MB	8m 14s	9.247 MB	3m 55s 830ms
19	55.216.800	16.207 MB	9m 11s	9.791 MB	3m 53s 1ms
20	58.284.400	17.108 MB	10m 22s	10.334 MB	4m 37s 736ms
21	61.352.000	18.008 MB	11m 30s	10.878 MB	5m 1s 430ms
22	64.419.600	18.909 MB	11m 58s	11.422 MB	5m 21s 614ms
23	67.487.200	19.809 MB	12m 43s	11.966 MB	4m 53s 553ms
24	70.554.800	20.710 MB	13m 44s	12.510 MB	5m 31s 161ms
25	73.622.400	21.610 MB	15m 03s	13.054 MB	5m 50s 815ms
26	76.690.000	22.510 MB	15m 41s	13.598 MB	6m 577ms
27	79.757.600	23.411 MB	16m 13s	14.142 MB	6m 575ms

28	82.825.200	24.311 MB	17m 19s	14.686 MB	5m 43s 969ms
29	85.892.800	25.212 MB	17m 48s	15.230 MB	5m 46s 741ms
30	88.960.400	26.112 MB	18m 50s	15.774 MB	6m 25s 625ms
31	92.028.000	27.013 MB	Báo lỗi	16.318 MB	5m 51s 52ms

Biểu đồ dưới đây so sánh thời gian xử lý của HWS và SQL với 31 lần test:



Hình 3. So sánh thời gian xử lý của HWS và SQL

Từ thực nghiệm chúng tôi rút ra được một số kết quả sau:

a. Cùng một tập tin với số mẫu tin xác định nhưng dung lượng lưu trữ của nó trong HWS luôn luôn có kích cỡ nhỏ hơn dung lượng lưu trữ trong SQL.
b. Thời gian xử lý của SQL tốt hơn HWS đối với những tập tin có kích thước nhỏ nhưng với tập tin có kích thước lớn thì HWS lại có thời gian xử lý nhanh hơn.
c. HWS có khả năng xử lý dữ liệu bán cấu trúc như dữ liệu XML hoặc dữ liệu không có cấu trúc như dữ liệu từ các mạng xã hội, weblog, hình ảnh giao thông, ...
d. SQL chỉ xử lý dữ liệu có cấu trúc, không có khả năng xử lý các tập tin có kích cỡ khá lớn (thường là báo lỗi hoặc thời gian xử lý rất lâu)

4. KẾT LUẬN

Mục đích bài báo này là giới thiệu về Big Data, giải pháp xử lý và lưu trữ đối với các loại dữ liệu có kích thước lớn, cụ thể ở đây là dữ liệu Covid được sinh trong quá trình tiêm chủng và điều trị Covid trong thời gian qua. Bài báo cũng bàn về cơ chế xử lý cho Big Data trong môi trường Hadoop, đó là HDFS và MapReduce, để người đọc hiểu được nguyên tắc xử lý các Big Data trên nền tảng phân tán Hadoop. Việc tìm hiểu các công cụ này sẽ mang lại nhiều lựa chọn giải

pháp xử lý và lưu trữ và cho việc phát triển ứng dụng phân tán với nhiều mục đích khác nhau.