

QUẢN LÝ DỮ LIỆU LỚN

(OVERVIEW OF BIG DATA MANAGEMENT)

Hoàng Thị Hà

Khoa Công nghệ Thông tin – Học viện Nông nghiệp Việt Nam

Tóm tắt:

Quản lý dữ liệu lớn đóng vai trò then chốt trong việc khai thác giá trị từ dữ liệu, tối ưu hóa hoạt động và ra quyết định chiến lược của tổ chức.

Bài viết trình bày tổng quan về dữ liệu lớn và quản lý dữ liệu lớn, bao gồm: khái niệm dữ liệu lớn, các đặc trưng của dữ liệu lớn, ứng dụng của dữ liệu lớn, đồng thời phân tích các cơ hội, thách thức trong quản lý dữ liệu lớn, giới thiệu các công cụ quản lý dữ liệu lớn và luồng quy trình quản lý dữ liệu lớn.

Từ khóa: Big Data, Big Data Management, Dữ liệu lớn, quản lý dữ liệu lớn, phân tích dữ liệu lớn

1. ĐẶT VẤN ĐỀ

Trong thời đại công nghệ số phát triển mạnh mẽ, dữ liệu trở thành nguồn tài nguyên chiến lược và có giá trị lớn đối với mọi lĩnh vực. Tuy nhiên, cùng với sự bùng nổ về số lượng và tốc độ tạo ra dữ liệu, việc lưu trữ, quản lý và khai thác hiệu quả khối lượng dữ liệu khổng lồ này đặt ra nhiều thách thức. Do đó, việc nghiên cứu và triển khai quản lý dữ liệu lớn là cấp thiết, đặc biệt trong bối cảnh chuyển đổi số đang diễn ra mạnh mẽ tại Việt Nam và trên toàn cầu.

2. TỔNG QUAN VỀ DỮ LIỆU LỚN

2.1. Khái niệm về dữ liệu lớn

Dữ liệu lớn (Big Data) là một thuật ngữ dùng để chỉ tập hợp dữ liệu có khối lượng lớn, tốc độ xử lý cao, đa dạng về định dạng và đòi hỏi các phương pháp, công cụ đặc biệt để lưu trữ, quản lý, phân tích để nâng cao sự hiểu biết về dữ liệu, hỗ trợ ra quyết định [1]

2.2. Các đặc điểm của dữ liệu lớn

Dữ liệu lớn có khá nhiều các đặc điểm, dưới đây là các đặc điểm chính [2]:

- **Volume (Khối lượng):** Lượng dữ liệu lớn được tạo ra hàng ngày, tính bằng terabyte (TB), petabyte (PB), thậm chí exabyte (EB).
- **Velocity (Tốc độ):** Dữ liệu được sinh ra và cập nhật liên tục theo thời gian thực. Ví dụ: dữ liệu giao dịch chứng khoán, dữ liệu từ cảm biến giao thông hoặc dữ liệu livestream.
- **Variety (Đa dạng):** Dữ liệu có thể ở nhiều dạng như: văn bản, hình ảnh, âm thanh, video, dữ liệu cảm biến, dữ liệu log...
- **Veracity (Độ tin cậy):** Dữ liệu có thể không hoàn chỉnh, nhiễu hoặc không nhất quán. Việc đảm bảo tính chính xác và đáng tin cậy là một thách thức lớn.
- **Value (Giá trị):** Không phải mọi dữ liệu đều có giá trị. Quá trình khai thác dữ liệu lớn nhằm tìm ra thông tin hữu ích phục vụ ra quyết định.

2.3. Phân loại dữ liệu lớn

Dữ liệu lớn được chia thành ba loại chính:

- **Dữ liệu có cấu trúc:** Dữ liệu có định dạng rõ ràng và có thể lưu trữ trong cơ sở dữ liệu quan hệ như MySQL, SQL Server hoặc Oracle.
- **Dữ liệu bán cấu trúc:** Có cấu trúc một phần nhưng không thể lưu trong bảng truyền thống. Ví dụ: tệp XML, tài liệu JSON, email.
- **Dữ liệu phi cấu trúc:** Không có cấu trúc và không thể lưu dạng bảng. Ví dụ: video, âm thanh, văn bản, dữ liệu máy móc tạo ra.

2.4. Sự khác biệt giữa dữ liệu nhỏ và dữ liệu lớn

Có một số điểm khác biệt giữa dữ liệu nhỏ và dữ liệu lớn. Những điểm khác biệt này bao gồm khối lượng, tốc độ, tính đa dạng, độ chính xác, giá trị, thời gian [3]

Bảng 1. So sánh dữ liệu lớn và dữ liệu nhỏ

Đặc điểm	Dữ liệu nhỏ(Small data)	Dữ liệu lớn(Big data)
Khối lượng	< 1 TB	> 1 TB
Tốc độ	Dòng chảy ổn định	Dữ liệu không lờ đến liên tục
Đa dạng	Có cấu trúc	Rất đa dạng: có cấu trúc, bán cấu trúc, phi cấu trúc: văn bản, hình ảnh, âm thanh, v.v.

Đặc điểm	Dữ liệu nhỏ(Small data)	Dữ liệu lớn(Big data)
Độ tin cậy	Dữ liệu chất lượng cao	Thường thiếu độ chính xác và nhiều nhiễu
Giá trị	Phân tích, báo cáo kinh doanh	Khai phá mẫu, dự đoán, phát hiện bất thường
Thời gian	Dữ liệu thể hiện giá trị kinh doanh theo thời gian cụ thể	Đôi khi, dữ liệu lịch sử trở nên không liên quan đến việc phân tích thông tin kinh doanh

2.5. Tại sao “dữ liệu lớn” lại quan trọng?

Khi dữ liệu lớn được thu thập, xử lý và phân tích đúng cách, các doanh nghiệp có thể hiểu rõ hơn về hoạt động kinh doanh, khách hàng, sản phẩm và đối thủ. Những hiểu biết này giúp:

- ✓ Tăng hiệu suất làm việc
- ✓ Tăng doanh số
- ✓ Giảm chi phí
- ✓ Cải thiện dịch vụ khách hàng
- ✓ Phát triển sản phẩm tốt hơn

Ví dụ: cảm biến trong sản phẩm sản xuất cung cấp dữ liệu theo thời gian thực (telemetry). Doanh nghiệp có thể phân tích dữ liệu mạng xã hội và blog để hiểu lý do tại sao khách hàng không mua hàng. Phân tích dữ liệu từ trung tâm hỗ trợ khách hàng có thể nâng cao sự hài lòng.

2.6. Ứng dụng của dữ liệu lớn

Dữ liệu lớn đã và đang được ứng dụng rộng rãi trong nhiều lĩnh vực, mang lại giá trị thực tiễn và khả năng ra quyết định tốt hơn.

a.. Trong lĩnh vực tài chính – ngân hàng

Big Data giúp phân tích hành vi khách hàng, phát hiện gian lận giao dịch, tối ưu hóa danh mục đầu tư và cá nhân hóa dịch vụ. Nhiều tổ chức tài chính sử dụng mô hình AI kết hợp dữ liệu lớn để chấm điểm tín dụng, phát hiện bất thường trong giao dịch.

b.. Trong y tế

Dữ liệu lớn hỗ trợ phân tích hồ sơ bệnh án điện tử, hình ảnh y tế, dữ liệu gen và cảm biến sinh học để cá nhân hóa điều trị, dự đoán dịch bệnh và tối ưu hóa quy trình chăm sóc bệnh nhân.

c.. Trong thương mại điện tử và marketing

Amazon, Alibaba, Shopee sử dụng Big Data để phân tích hành vi mua sắm, đưa ra gợi ý sản phẩm, tối ưu hóa quảng cáo và quản lý chuỗi cung ứng. Các thuật toán đề xuất dựa trên hành vi người dùng được xây dựng từ dữ liệu lớn.

d.. Trong nông nghiệp thông minh

Dữ liệu cảm biến từ thiết bị IoT có thể giúp nông dân theo dõi độ ẩm đất, nhiệt độ, ánh sáng... từ đó điều chỉnh phương pháp canh tác. Các mô hình dự đoán năng suất cây trồng sử dụng dữ liệu vệ tinh và dữ liệu khí tượng lịch sử.

e. Trong giao thông và đô thị thông minh

Dữ liệu lớn giúp tối ưu hóa quản lý giao thông, phân tích luồng người, dự báo nhu cầu đi lại, từ đó đưa ra phương án điều phối giao thông thông minh. Hệ thống thu phí không dừng (ETC), camera giao thông AI đều dựa trên xử lý dữ liệu lớn.

f. Trong giáo dục và nghiên cứu

Big Data được sử dụng để phân tích kết quả học tập, đánh giá hiệu quả giảng dạy và cá nhân hóa quá trình học tập. Trong nghiên cứu khoa học, các mô hình phân tích dữ liệu lớn giúp tìm kiếm tri thức mới, đặc biệt trong lĩnh vực vật lý, thiên văn học, sinh học phân tử...

3. QUẢN LÝ DỮ LIỆU LỚN

3.1. Quản lý dữ liệu lớn là gì

Quản lý dữ liệu lớn (Big Data Management) là một quá trình phức tạp, là tập hợp các kỹ thuật, công cụ và nền tảng (platforms) nhằm lưu trữ, tiền xử lý, bảo mật dữ liệu lớn một cách hiệu quả [4]. Không giống như quản lý dữ liệu truyền thống vốn chỉ dựa trên các hệ quản trị quan hệ (RDBMS), quản lý dữ liệu lớn đòi hỏi khả năng xử lý phân tán, làm việc với dữ liệu phi cấu trúc, đồng thời đảm bảo hiệu suất và khả năng mở rộng theo thời gian.

Quản lý dữ liệu lớn không chỉ là bài toán kỹ thuật, mà còn là một thách thức về chiến lược tổ chức, mô hình vận hành, bảo mật thông tin. Việc quản lý dữ liệu lớn hiệu quả có thể giúp doanh nghiệp:

- ✓ Tối ưu hóa chuỗi cung ứng và hoạt động sản xuất
- ✓ Hiểu rõ hành vi và nhu cầu khách hàng
- ✓ Dự đoán xu hướng thị trường và ra quyết định chính xác
- ✓ Cải thiện trải nghiệm người dùng và dịch vụ khách hàng

3.2. Cơ hội và thách thức trong quản lý dữ liệu lớn

3.2.1. Cơ hội

a. Tăng cường khả năng ra quyết định

Phân tích dữ liệu lớn cho phép doanh nghiệp đưa ra các quyết định dựa trên bằng chứng thực tiễn, thay vì chỉ dựa vào kinh nghiệm chủ quan. Ví dụ, ngành bán lẻ có thể dự đoán xu hướng tiêu dùng dựa trên dữ liệu mua sắm và tương tác trên mạng xã hội.

b. Tối ưu hóa hoạt động vận hành

Phân tích dữ liệu lớn giúp phát hiện những điểm yếu, điểm nghẽn và cơ hội tối ưu hóa, từ đó hợp lý hóa hoạt động và tiết kiệm chi phí. Dữ liệu từ các thiết bị cảm biến, phản hồi của khách hàng,... có thể giúp doanh nghiệp xác định điểm nghẽn trong quy trình và điều chỉnh kịp thời để tăng hiệu suất.

c. Cải thiện tương tác khách hàng

Big Data hỗ trợ doanh nghiệp phân khúc khách hàng, phân tích hành vi và xây dựng mô hình cá nhân hóa, từ đó cải thiện chiến dịch marketing và giữ chân người dùng.

d. Phát hiện gian lận và đảm bảo an toàn

Ngành tài chính sử dụng mô hình phân tích dữ liệu lớn để phát hiện giao dịch bất thường, giúp giảm thiểu rủi ro tín dụng và gian lận tài chính.

e. Thúc đẩy đổi mới sáng tạo

Dữ liệu lớn tạo nền tảng cho sự đổi mới, cho phép các tổ chức phát triển sản phẩm, dịch vụ và mô hình kinh doanh mới để đáp ứng nhu cầu thị trường luôn biến đổi.

3.2.2. Thách thức

Mặc dù dữ liệu lớn có thể mang lại những cơ hội đáng kể cho các doanh nghiệp và tổ chức, nhưng việc áp dụng nó không hề đơn giản và đi kèm với nhiều thách thức. Các thách thức này có thể được phân loại thành bốn nhóm chính: kỹ thuật, phân tích, tổ chức và đạo đức/pháp lý [5]:

a. Thách thức kỹ thuật: Cơ sở hạ tầng hiện tại có thể không đủ khả năng xử lý khối lượng dữ liệu lớn, tốc độ tạo dữ liệu cao và đa dạng định dạng. Các hệ thống lưu trữ và xử lý phân tán như Hadoop và Spark là rất cần thiết, nhưng việc tích hợp và bảo trì chúng khá phức tạp và tốn kém. Hơn nữa, việc đảm bảo tính bảo mật, riêng tư và chất lượng của dữ liệu là một vấn đề lớn trong bối cảnh dữ liệu phân tán và dữ liệu đến từ nhiều nguồn không đồng nhất.

b. Thách thức phân tích: Để thu được hiểu biết có giá trị từ big data, cần có các phương pháp phân tích nâng cao, như học máy, khai phá dữ liệu và phân tích dự đoán... Tuy nhiên, những phương pháp này đòi hỏi chuyên môn cao và năng lực tính toán mạnh. Ngoài ra, việc xử lý dữ liệu không có cấu trúc – chẳng hạn như văn bản, hình ảnh và video – vẫn là một thách thức lớn. Một vấn đề khác là sự lệch (bias) trong dữ liệu, có thể dẫn đến kết quả phân tích sai lệch và ảnh hưởng không tốt đến quyết định.

c. Thách thức tổ chức: Nhiều tổ chức vẫn chưa phát triển được văn hóa dựa trên dữ liệu và gặp khó khăn trong việc tích hợp “big data” vào quy trình ra quyết định. Việc thiếu kỹ năng và chuyên môn, sự không thiện chí thay đổi từ nhân viên và vấn đề về quản trị dữ liệu là những rào cản phổ biến. Ngoài ra, việc liên kết dữ liệu lớn với các mục tiêu kinh doanh rõ ràng và đo lường được tác động của nó cũng là những thách thức lớn.

d. Thách thức đạo đức và pháp lý: Việc thu thập, lưu trữ và sử dụng dữ liệu lớn làm dấy lên nhiều câu hỏi đạo đức và pháp lý, đặc biệt là về quyền riêng tư và sự đồng thuận của người dùng.

3.3. Các công cụ quản lý dữ liệu lớn

Sự gia tăng nhanh chóng về khối lượng, tốc độ và đa dạng của dữ liệu đã dẫn đến sự phát triển của nhiều công cụ và công nghệ để lưu trữ, xử lý và phân tích dữ liệu lớn. Dưới đây là một số công cụ tiêu biểu được chia theo nhóm chức năng:

3.3.1. Hệ thống lưu trữ và xử lý phân tán

a. Apache Hadoop: Một framework nguồn mở dùng để lưu trữ và xử lý hiệu quả các bộ dữ liệu lớn có kích thước từ hàng gigabyte đến hàng petabyte dữ liệu. Thay vì sử dụng một máy tính lớn để lưu trữ và xử lý dữ liệu, Hadoop cho phép tạo nhóm nhiều máy tính để phân tích các bộ dữ liệu lớn đồng thời một cách nhanh chóng hơn.

b. Apache Spark: Công cụ cung cấp khả năng xử lý dữ liệu trong bộ nhớ (in-memory), nhanh hơn đáng kể so với Hadoop trong nhiều trường hợp, và hỗ trợ nhiều ngôn ngữ lập trình như Scala, Python và Java.

3.3.2. Hệ quản trị dữ liệu phi quan hệ (NoSQL)

Các hệ quản trị cơ sở dữ liệu NoSQL như MongoDB, Cassandra, HBase và Redis được thiết kế để lưu trữ và truy xuất khối lượng lớn dữ liệu phi cấu trúc và bán cấu trúc. Chúng cung cấp khả năng mở rộng cao, tính linh hoạt trong mô hình dữ liệu và hiệu suất truy vấn tốt cho những khối dữ liệu lớn.

a. MongoDB: Hệ quản trị cơ sở dữ liệu hướng tài liệu (document database) dùng để quản lý các dữ liệu lưu trữ phi cấu trúc, không có tính ràng buộc, toàn vẹn nên tính sẵn sàng cao, hiệu suất lớn và dễ dàng mở rộng lưu trữ. MongoDB sử dụng lưu trữ dữ liệu dưới dạng Document JSON nên mỗi một collection sẽ có các kích cỡ và các document khác nhau.

b. Apache Cassandra: Hệ cơ sở dữ liệu phân tán, tối ưu cho khối lượng lớn dữ liệu ghi và khả năng mở rộng theo chiều ngang.

c. HBase: Cơ sở dữ liệu dạng cột, được xây dựng trên cơ sở Apache Hadoop. Công nghệ được thiết kế để lưu trữ và xử lý lượng lớn dữ liệu cấu trúc, bán cấu trúc trên các cụm máy chủ phân tán. HBase có khả năng mở rộng linh hoạt để đáp ứng nhu cầu về gia tăng dữ liệu trong những trường hợp cần thiết.

3.3.3. Công cụ thu thập và tích hợp dữ liệu

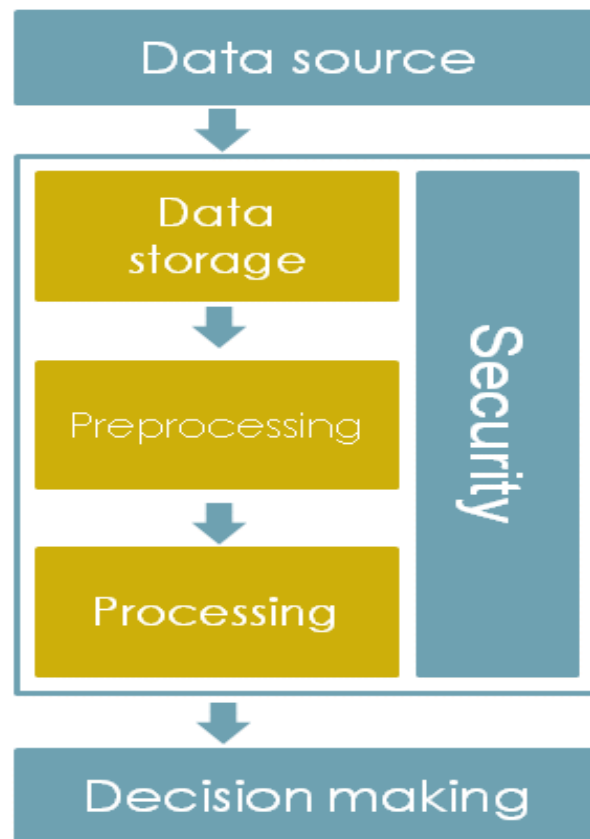
a. Apache Flume và Apache Sqoop: là các công cụ được sử dụng để thu thập và chuyển đổi dữ liệu từ các nguồn khác nhau đến một kho lưu trữ tập trung như HDFS (Hadoop Distributed File System)

b. Apache nifi được sử dụng để tự động hóa và kiểm soát các luồng dữ liệu giữa các hệ thống. Nó cung cấp cho chúng ta một giao diện trên nền web mà có thể thu thập, xử lý, phân tích dữ liệu. NiFi được biết đến với khả năng xây dựng luồng chuyển dữ liệu tự động giữa các hệ thống. Đặc biệt là hỗ trợ rất nhiều kiểu nguồn và đích khác nhau như: hệ thống file, cơ sở dữ liệu quan hệ, phi quan hệ,... Ngoài ra NiFi sẽ hỗ trợ các thao tác với dữ liệu như lọc chỉnh sửa, thêm bớt nội dung.

3.3.4. Công cụ trực quan hóa và phân tích dữ liệu. Để phân tích dữ liệu lớn, có rất nhiều công cụ như Apache Hive (truy vấn dữ liệu lớn bằng ngôn ngữ giống SQL), Apache Pig (ngôn ngữ lập trình kịch bản phân tích dữ liệu), và các thư viện học máy như Apache Mahout hoặc MLlib (Spark). Ngoài ra, các công cụ trực quan hóa dữ liệu như Tableau, Power BI, hoặc các thư viện Python như Matplotlib, Seaborn và Plotly cũng được sử dụng để trình bày kết quả phân tích theo cách dễ hiểu.

3.4. Quy trình quản lý dữ liệu lớn

Quản lý dữ liệu lớn là một quy trình gồm nhiều bước, mỗi bước đều cần giám sát và bảo mật. Quy trình quản lý dữ liệu lớn được trình bày như Hình 1 [6]



Hình 1. Luồng quy trình quản lý dữ liệu lớn

✓ **Bước 1:** Lưu trữ dữ liệu (*Data Storage*)

Một trong những mối quan tâm cơ bản của việc quản lý Dữ liệu lớn là cách xây dựng các hệ thống có khả năng lưu trữ dữ liệu ở quy mô petabyte. Các hệ thống quản lý lưu trữ như vậy cần có khả năng lưu giữ dữ liệu đủ dữ liệu, cải thiện khả năng truy xuất và tối ưu hóa tính khả dụng của dữ liệu để xử lý

✓ **Bước 2:** Tiền xử lý dữ liệu (*Data Pre-processing*)

Để thực hiện phân tích dữ liệu một cách hiệu quả, dữ liệu trước tiên phải đáp ứng một số yêu cầu về chất lượng. Các thao tác tiền xử lý nhằm mục đích chính là chuyển đổi dữ liệu thô không nhất quán hoặc không đầy đủ thành định dạng dễ hiểu, sẵn sàng cho các bước phân tích tiếp theo. Do đó, chuẩn bị dữ liệu để xử lý tiếp theo bằng cách cấu trúc chúng theo một định dạng chuẩn để thu được dữ liệu tích hợp

✓ **Bước 3: Xử lý dữ liệu (Data Processing)**

Bước then chốt trong quy trình quản lý Dữ liệu lớn là xử lý dữ liệu.

Trong giai đoạn này, người ta thu được những hiểu biết sâu sắc về mối quan hệ giữa các đặc trưng, cho phép phát triển các phân tích hiệu quả.

Giai đoạn này bao gồm nhiều phương pháp phân tích khác nhau như phân tích thống kê, khai phá dữ liệu và học máy.

✓ **Bước 4: Bảo mật (Security)**

Bảo mật ngày càng trở thành một vấn đề nghiêm trọng. Bảo mật liên quan đến nhiều khía cạnh như quyền riêng tư, tính toàn vẹn, tính bảo mật và tính sẵn sàng của dữ liệu và được thực hiện ở tất cả các khâu trên.

4. KẾT LUẬN

Dữ liệu lớn đang đóng vai trò trung tâm trong quá trình chuyển đổi số của nền kinh tế toàn cầu. Từ phân tích hành vi khách hàng trong thương mại điện tử đến dự đoán dịch bệnh trong y tế, từ quản lý chuỗi cung ứng đến điều hành đô thị thông minh, từ giáo dục đến kinh doanh, sản xuất...dữ liệu lớn không chỉ là tài nguyên công nghệ mà còn là lợi thế cạnh tranh cốt lõi của các tổ chức.

Tuy nhiên, quản lý dữ liệu lớn cũng đặt ra nhiều thách thức về kỹ thuật, hạ tầng, nhân lực, bảo mật và đạo đức. Việc lựa chọn công cụ phù hợp, xây dựng kiến trúc linh hoạt và thực thi quy trình quản trị chặt chẽ là điều kiện tiên quyết để khai thác thành công giá trị từ dữ liệu. Trong bối cảnh Việt Nam đang thúc đẩy chiến lược chuyển đổi số quốc gia, quản lý dữ liệu lớn cần được coi là ưu tiên hàng đầu trong mọi lĩnh vực.

Để tiến xa hơn, Việt Nam cần đầu tư bài bản vào hạ tầng dữ liệu, xây dựng đội ngũ chuyên gia dữ liệu, hoàn thiện khung pháp lý bảo vệ dữ liệu cá nhân, và đẩy mạnh hợp tác công – tư trong nghiên cứu, triển khai và chuyển giao công nghệ Big Data. Chỉ khi làm chủ được dữ liệu lớn, các tổ chức và quốc gia mới có thể làm chủ được tương lai số.

TÀI LIỆU THAM KHẢO

1. TechAmerica Foundation's Federal Big Data Commission. (2012). Demystifying big data: A practical guide to transforming the business of Government.
2. G'eczy, P.: Big data characteristics. *Macro Theme Rev.* 3(6), 94–104 (2014)
3. Kitchin, R., Lauriault, T.P.: Small data in the era of big data. *GeoJournal* 80(4), 463–475 (2015)
4. Siddiqua, A., Hashem, I. A. T., Yaqoob, I., Marjani, M., Shamshirband, S., Gani, A., & Nasaruddin, F. (2016). A survey of big data management: Taxonomy and state-of-the-art. *Journal of Network and Computer Applications*, 71, 151-166.
5. Oguntimilehin, A., & Ademola, E. O. (2014). A review of big data management, benefits and challenges. *A Review of Big Data Management, Benefits and Challenges*, 5(6), 1-7.
6. Aisha Siddiqua; Ibrahim Abaker Targio Hashem; Ibrar Yaqoob; Mohsen Marjani; Shahabuddin Shamshirband; Abdullah Gani; Fariza Nasarudin. A survey of big data management (2016): Taxonomy and state-of-the-art. *Journal of Network and Computer Applications*
7. Bộ Thông tin và Truyền thông (2023). *Báo cáo Chuyển đổi số Quốc gia*. <https://dx.gov.vn>
8. Apache Hadoop Project. <https://hadoop.apache.org>
9. Apache Spark Project. <https://spark.apache.org>
10. MongoDB Documentation. <https://www.mongodb.com>