Dữ liệu lớn và làm chủ công nghệ dữ liệu lớn tại Việt Nam

TS. Trần Việt Trung
Viện Công nghệ thông tin và truyền thông, Đại học Bách khoa, Hà Nội

Ngày nay, dữ liệu được thu thập và sinh ra với tốc độ tăng theo cấp số nhân. Theo một nghiên cứu đưa ra bởi viện McKinsey Global 2011[40], 40 zettabytes tức khoảng 43 nghìn tỉ gigabytes dữ liệu sẽ được tạo ra vào năm 2020. Con số này tăng 300 lần so với số liệu thống kê được vào năm 2005. Tác nhân chính gây ra bùng nổ dữ liệu có thể kể đến sự phổ biến của điện thoại thông minh, xu hướng mạng xã hội chia sẻ, và mạng lưới vạn vật kết nối internet (internet of things)[41]. Dữ liệu lớn (Big data) là thuật ngữ được dùng để tính chất hoá sự bùng nổ của dữ liệu và vai trò của dữ liệu trong mọi mặt của đời sống xã hội và sản xuất. Dữ liệu lớn được đặc trưng bởi 1 trong 5 tính chất: (1) Dung lượng lớn (Big volume); (2) Tốc độ lớn (Big Velocity); (3) Đa dạng lớn (Big Variety); (4) Độ tin cậy (Big Veracity); và (5) Giá trị lớn (Big Value).

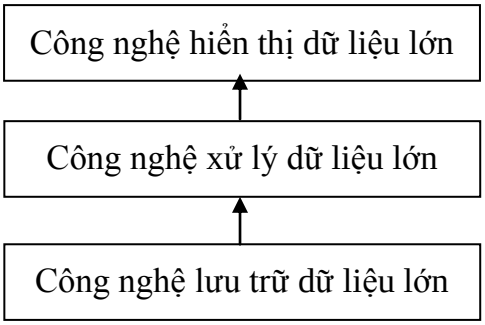
Dữ liệu lớn có tầm quan trọng được ví như nguồn tài nguyên dầu mỏ mới của thế kỷ 21[42]. Ứng dụng khai thác dữ liệu lớn có thể tạo ra nguồn lợi khổng lồ trong mọi lĩnh vực. Trong lĩnh vực chăm sóc sức khoẻ y tế, khai thác dữ liệu lớn trong nghiên cứu phương thuốc và phương pháp điều trị cho bệnh nhân là ngành công nghiệp tạo ra 300 tỉ USD lợi nhuận mỗi năm tại Mĩ. Cũng theo nghiên cứu này, châu Âu tiết kiệm 250 tỉ EUR mỗi năm nhờ ứng dụng dữ liệu lớn trong lĩnh vực hành chính công. Trong lĩnh vực bán lẻ, khai phá dữ liệu lớn được kỳ vọng giúp tăng doanh thu lên tới 60%. Trong lĩnh vực sản xuất, dữ liệu lớn giúp cắt giảm chi phí sản xuất lên tới 50%.

1. Công nghệ cho dữ liệu lớn

Song hành với nguồn lợi khổng lồ mà dữ liệu lớn mang lại là những thách thức không nhỏ về mặt công nghệ, đòi hỏi các mô hình lưu trữ, xử lý và phân tích mới. Để đáp ứng với các tính chất của dữ liệu lớn, các mô hình này được cài đặt trên môi trường phân tán, kết tập năng lực lưu trữ và xử lý của hàng ngàn máy chủ. Công nghệ cho dữ liệu lớn được phân thành 3 tầng chính, như Hình 1:

[40] *McKinsey Global, Big data: The next frontier for innovation, competition, and productivity, 2011*
[41] *https://en.wikipedia.org/wiki/Internet_of_Things*
[42] *McKinsey Global, Big data: The next frontier for innovation, competition, and productivity, 2011*

Hình 1: Kiến trúc phân tầng công nghệ dành cho dữ liệu lớn



1.1. Công nghệ lưu trữ dữ liệu lớn

Công nghệ lưu trữ dữ liệu lớn đáp ứng bài toán Big Volume và Big Velocity, tức giải quyết bài toán lượng dữ liệu khổng lồ và tốc độ xử lý cao. Hai bài toán này được giải quyết bằng cách phân mảnh dữ liệu và phân tán trên nhiều server lưu trữ. Khi truy xuất dữ liệu thì cho phép truy xuất đồng thời nhiều server lưu trữ cùng một lúc để tăng thông lượng.

- Hệ thống tập tin phân tán: HDFS

HDFS[43] (Hadoop Distributed File System) là hệ thống quản lý tập tin được thiết kế để tối ưu cho bài toán lưu trữ các tập tin có kích thước lớn hàng GB, thậm chí TB. Để giải quyết bài toán này, dữ liệu của các tập tin lớn sẽ được chia nhỏ thành các khối lớn (ví dụ 64MB) và phân tán trên các nút lưu trữ. So với các hệ thống tập tin khác, HDFS không tối ưu cho bài toán lưu trữ hàng tỉ tập tin nhỏ với kích thước mỗi tập tin chỉ vài KB. Ưu điểm của thiết kế tập tin lớn là giảm tải cho hệ thống quản lý không gian tập tin, giảm thời gian thao tác trên các thư mục hay tìm kiếm tập tin.

- Cơ sở dữ liệu không quan hệ: NoSQL[44]

Mô hình cơ sở dữ liệu truyền thống dược thiết kế và tối ưu cho lưu trữ và xử lý dữ liệu nghiệp vụ doanh nghiệp không còn phù hợp với sự đa dạng của dữ liệu lớn (Big Variety). Xu thế cơ sở dữ liệu mới, gọi tên NoSQL được đưa ra và sử dụng rộng rãi tại các công ty internet lớn, như Google, Yahoo, Facebook, Amazon. Cơ sở dữ liệu NoSQL đáp ứng lưu trữ dữ liệu với lược đồ mô hình dữ liệu linh hoạt, đáp ứng đa dạng định dạng dữ liệu, tính khả mở, giao tiếp ứng dụng API đơn giản, loại bỏ các tính chất không thực sự cần thiết của cơ sở dữ liệu quan hệ truyền thống (đảm bảo ACID, ngôn ngữ truy vấn SQL).

[43] *http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html*
[44] *https://en.wikipedia.org/wiki/NoSQL*

NoSQL phân làm 4 nhóm chính: Cơ sở dữ liệu khoá – giá trị (key/value store), cơ sở dữ liệu văn bản (document store), cơ sở dữ liệu dạng đồ thị (graph store), cơ sở dữ liệu siêu cột (column familly store).

1.2. Công nghệ xử lý dữ liệu lớn

Để xử lý và phân tích dữ liệu lớn cần mô hình phương thức tính toán khác biệt so với các mô hình truyền thống. Phương pháp xử lý dữ liệu lớn là kết tập, phối hợp năng lực xử lý của nhiều máy tính vào giải quyết một bài toán chung.

Công nghệ xử lý dữ liệu lớn phổ biến là mô hình tính toán MapReduce[45], được Google đưa ra vào năm 2004. Với mô hình tính toán này, các máy tính này sẽ hoạt động song song nhưng độc lập với nhau, mục đích là làm rút ngắn thời gian xử lý toàn bộ dữ liệu. Tính toán MapReduce được phân tán trên các nút lưu trữ. So với các mô hình tính toán khác mà dữ liệu được sao chép đến các nút tính toán và thực hiện trên các nút đó, mô hình tính toán MapReduce khác biệt ở chỗ mã chương trình được sao chép tới các nút lưu trữ để thực thi. Đây là một trong những điểm mấu chốt tiên tiến của MapReduce vì quan điểm di chuyển mã chương trình thì tiết kiệm và hiệu quả hơn di chuyển dữ liệu mà có thể lên tới hàng TB. Hơn nữa, nền tảng tính toán MapReduce được thiết kế để thực thi với các máy chủ phổ thông, không cần năng lực tính toán và lưu trữ lớn như mô hình tính toán song song MPI. Điều này đạt được nhờ vào thiết kế chịu lỗi cao.

1.3. Công nghệ hiển thị dữ liệu lớn

Hiển thị trực quan lượng dữ liệu khổng lồ và các tri thức khai thác được từ dữ liệu là đòi hỏi cần thiết khi làm việc với dữ liệu lớn. Việc hiển thị dữ liệu dưới dạng trực quan giúp người khai thác có cái nhìn toàn cảnh về dữ liệu và tri thức mang lại từ dữ liệu. Các công cụ cho phép hiển thị và tương tác trực quan với dữ liệu lớn hiện nay phổ biến là các công cụ như Tableau, Pentahoo, SAS, vv...

2. Thách thức ứng dụng lưu trữ, xử lý dữ liệu lớn tại Việt Nam

2.1. Thách thức về dữ liệu

Tại Việt Nam, nguồn dữ liệu lưu trữ trong các cơ quan, tổ chức thường phân tán và khó tiếp cận. Mỗi chi nhánh, đơn vị tổ chức một cơ sở dữ liệu riêng, và chỉ chia sẻ, lưu trữ tập trung một phần khối lượng dữ liệu. Hơn nữa, định dạng dữ liệu này đa phần không được chuẩn hoá. Mỗi công ty, tổ chức thường lưu trữ thông tin theo cấu trúc riêng, tồn tại dưới nhiều dạng như tệp tin văn bản, excel, trong cơ sở dữ liệu, vv... Các thông tin đầy đủ về mặt nội dung đa phần tồn tại dưới dạng lưu trữ vật

[45] *Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters" Communications of the ACM 51.1 (2008): 107-113*

lý trên giấy tờ. Thông tin được số hoá chỉ mang tính chất tóm tắt, không phản ánh chính xác và có nhiễu.

2.2. Thách thức về công nghệ

Xét tới thời điểm hiện tại, hệ sinh thái Hadoop[46] là nền tảng công nghệ cho dữ liệu lớn mã nguồn mở phổ biến nhất. Hadoop bao gồm các thành phần cơ bản như hệ thống tệp tin phân tán HDFS, cơ sở dữ liệu bán cấu trúc Hbase, mô hình tính toán MapReduce, và bộ xử lý truy vấn Hive. Hadoop được phát triển từ 2006 cho phép lưu trữ và xử lý dữ liệu phân tán. Tuy nhiên, tốc độ truy xuất dữ liệu lớn được lưu trữ phân tán trên hàng ngàn máy chủ hiện vẫn là một lĩnh vực thu hút đông đảo cộng đồng nghiên cứu và phát triển nguồn mở. Truy vấn dữ liệu lớn trong thời gian tương tác (nhỏ hơn vài chục giây) là đòi hỏi thực tiễn trong các bài toán phân tích, giám sát và dự báo. Mô hình tính toán MapReduce được thiết kế xử lý dữ liệu phân tán với thời gian thực thi mỗi công việc từ vài phút đến hàng giờ. Bộ xử lý truy vấn Hive cho phép truy xuất dữ liệu lớn qua bộ tập lệnh tương tự như SQL, với tốc độ xử lý truy vấn chậm. Hive không đáp ứng trong thời gian tương tác (nhỏ hơn vài chục giây), do mỗi câu lệnh truy vấn đề được ánh xạ và thực hiện bởi một chuỗi các công việc sử dụng MapReduce.

2.3. Thách thức về con người

Việc sử dụng Hadoop đòi hỏi các kỹ năng vận hành hệ thống, phát triển phần mềm, khai phá dữ liệu chuyên biệt. Hadoop không phù hợp với với đại đa số người sử dụng truyền thống vốn quen làm việc với dữ liệu nhỏ, lưu trữ trên hệ quản trị cơ sở dữ liệu quan hệ và sử dụng truy vấn SQL trong khai thác dữ liệu. Tại Việt Nam, hiện chưa có chương trình đào tạo đưa công nghệ lưu trữ và xử lý dữ liệu lớn vào nội dung đào tạo một cách chính thức. Các kỹ sư, chuyên gia về dữ liệu lớn số lượng không đáng kể so với tiềm năng dữ liệu lớn tại Việt Nam. Họ chủ yếu được đào tạo tại nước ngoài, hoặc tự đào tạo trong các công ty, tổ chức lớn tiên phong về khai phá dữ liệu lớn.

Bên cạnh đó, việc vận hành công nghệ dữ liệu lớn như Hadoop đòi hỏi năng lực quản trị, tinh chỉnh, tối ưu hệ thống phân tán gồm nhiều tầng như tầng thiết bị lưu trữ, tầng mạng, tầng máy chủ, vvv. Việc lựa chọn công nghệ, các công cụ, các giải thuật cho các bài toán dữ liệu lớn là sự lựa chọn đòi hỏi nhiều kinh nghiệm chuyên gia. Hình bên dưới (Hình 2) thể hiện bức tranh toàn cảnh các công nghệ, công cụ dành cho dữ liệu lớn ở thời điểm hiện tại.

[46] *https://hadoop.apache.org/*

Hình 2: Toàn cảnh công nghệ dữ liệu lớn

2.4. Thách thức về hạ tầng

Việc lưu trữ và khai thác dữ liệu lớn đòi hỏi đầu tư về hạ tầng tính toán rất lớn vì cần rất nhiều năng lực lực lưu trữ và tính toán, đa phần đòi hỏi cụm máy tính có thể lên tới hàng chục ngàn máy chủ. Đây cũng chính là lý do chính mà tiên phong trong dữ liệu lớn là các công ty internet toàn cầu như Google, Amazon, Facebook,... Các công ty nhỏ và vừa, với nguồn vốn hạn hẹp sẽ không có đủ vốn để đầu tư xây dựng hạ tầng tính toán đủ mạnh cho việc khai phá dữ liệu lớn. Tuy nhiên với sự phát triển của điện toán đám mây trong thời gian gần đây sẽ giảm chi phí đầu tư hạ tầng khi các công ty có thể thuê cụm máy chủ trong một khoảng thời gian theo nhu cầu.

Tóm lại, dữ liệu lớn là nguồn tài nguyên quan trọng được ví như dầu mỏ. Tại Việt Nam, dữ liệu lớn vẫn là một lĩnh vực mới mẻ. Khai phá dữ liệu lớn hiện tồn tại dưới dạng thử nghiệm tại một số tập đoàn lớn như Viettel, FPT, các công ty dịch vụ internet như VCCORP, VNG. Trong lĩnh vực hành chính công chưa có báo cáo áp dụng được công bố chính thức. Triển khai khai phá dữ liệu lớn tại Việt Nam gặp rất nhiều rào cản như chi phí đầu tư hạ tầng máy chủ, tính thuyết phục khi áp dụng khai phá dữ liệu lớn, tính sẵn sàng của công nghệ và đặc biệt là nguồn nhân lực có kĩ năng về làm việc và khai thác dữ liệu lớn. Trong lĩnh vực thống kê nhà nước, dữ liệu đã được lưu trữ là rất lớn nhưng chưa được khai thác đúng nghĩa để mang lại nhiều giá trị tri thức quan trọng từ đó đưa ra các dự báo, phân tích chính sác nguyên nhân, kết quả từ dữ liệu. Vì vậy cần phải đẩy mạnh nghiên cứu xây dựng hạ tầng và khảo sát áp dụng khai thác dữ liệu lớn trong thống kê nhà nước. Dữ liệu lớn đã có, việc cần làm là tổ chức và khai thác hiệu quả nguồn dữ liệu quan trọng ấy./.