TẬP SAN TIN HỌC QUẢN LÝ
Tập 03, số 1&2, 2014, 53-73.

BIG DATA: BỨC TRANH TOÀN CẢNH

Lê Thị Quỳnh Nga[1]
Nguyễn Mạnh Tuấn[1]

Abstract: *Given the advantages of Big Data and the significant impact that Big Data and its related applications have had on achieving competitive edge, Big Data has been considered as new capability for driving and achieving business value. However, to gain the full potential of Big Data and to successfully implement a Big Data project, it is necessary that Big Data related problems are defined. This paper provides an overview about Big Data, including Big Data definition, characteristics and what we need to know about Big Data in business and technology perspectives. Then, this paper provides some Big Data problems in research and in practice.*

Tóm tắt: *Với những ưu điểm và tác động mạnh mẽ của Dữ liệu lớn (Big Data) và các ứng dụng liên quan, Big Data đang được xem như một yếu tố quyết định đến việc phát triển cũng như mang lại lợi thế cạnh tranh của các tổ chức. Tuy nhiên, để đạt được sự thành công trong việc xây dựng và thực hiện các dự án Big Data, những vấn đề có liên quan cần được xác định, từ đó tìm ra phương hướng để giải quyết. Bài báo này cung cấp cái nhìn tổng quan về Big Data, các ứng dụng của Big Data và khía cạnh kỹ thuật của Big Data. Bài báo cũng nêu một số khó khăn mà các nhà nghiên cứu và các doanh nghiệp cần quan tâm.*

Từ khóa:*Big Data, các ứng dụng, Hadoop, dữ liệu.*

PHẦN 1 GIỚI THIỆU

Ngày nay, sự phát triển của Internet đã làm thay đổi mạnh mẽ cách thức hoạt động của các tổ chức. Các ứng dụng Web 2.0, mạng xã hội, điện toán đám mây đã một phần mang lại cho các tổ chức phương thức kinh doanh mới [1]. Trong kỷ nguyên của IoT[2], các cảm biến được nhúng vào trong các thiết bị di động như điện thoại di động, ô tô, và máy móc công nghiệp góp phần vào việc tạo và chuyển dữ liệu, dẫn đến sự bùng nổ của dữ liệu có thể thu thập được [2]. Theo một báo cáo của IDC, năm 2011, lượng dữ liệu được tạo ra trên thế giới là 1.8ZB[3], tăng gần 9 lần chỉ trong 5 năm [3]. Dưới sự bùng nổ này, thuật ngữ Big Data được sử dụng để chỉ những bộ dữ liệu khổng lồ, chủ yếu không có cấu trúc[4], được thu thập từ nhiều nguồn khác nhau.

Với những tác động trong việc khám phá giá trị tiềm ẩn to lớn, Big Data đang được xem là một yếu tố mới quan trọng mang lại lợi ích cho các tổ chức trong nhiều lĩnh vực khác nhau [4, 5]. Trong một khảo sát của tổ chức Oracle Corp và

---

[1] Khoa Hệ Thống Thông Tin Kinh Doanh – ĐH Kinh Tế HCM

[2] IoT (Internet of Things) chỉ các cảm biến được cảm biến được nhúng trong các thiết bị, được liên kết với nhau bởi các mạng máy tính (Trích từ nguồn http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_internet_of_things)

[3] 1 zettabyte (ZB) = 2^{40} gigabytes (GB) ~ 1,000,000,000,000GB

[4] Dữ liệu không cấu trúc chỉ các loại dữ liệu không theo một định dạng cụ thể như: hình ảnh, âm thanh, video, văn bản…

Accenture PLC, 57% chuyên gia tài chính đánh giá đầu tư vào Big Data sẽ là yếu tố then chốt để đạt được lợi thế cạnh tranh [6]. Chính vì những lợi ích to lớn mà Big Data có thể mang lại, nhiều tổ chức đã đầu tư mạnh vào việc nghiên cứu và ứng dụng vào Big Data. Theo một báo cáo từ Gartner, năm 2014, 73% tổ chức được khảo sát đã mua hoặc có ý định đầu tư vào các dự án Big Data, con số này năm 2013 là 64%. [7]

Mục tiêu của nghiên cứu này nhằm đưa cái nhìn toàn cảnh về Big Data, những ứng dụng của Big Data trong các lĩnh vực, các yếu tố kỹ thuật liên quan đến Big Data. Đồng thời, một số vấn đề cũng cần xem xét khi thực hiện các dự án Big Data nhằm đảm bảo dự án thành công.

Cấu trúc của nghiên cứu như sau: Phần 2. Định nghĩa, các đặc trưng của Big Data. Phần 3. 4. Các ứng dụng của Big Data trong một số lĩnh vực, trong đó trình bày rõ một số trường hợp cụ thể về ứng dụng của Big Data trong hoạt động tài chính, bảo hiểm và thương mại, Phần 5. Khía cạnh kỹ thuật của Big Data. Phần 6. Những thách thức cần giải quyết. Phần 7. Kết luận.

PHẦN 2 BIG DATA: ĐỊNH NGHĨA VÀ CÁC ĐẶC TRƯNG

Dữ liệu lớn (Big data) là thuật ngữ dùng để mô tả các bộ dữ liệu có kích thước rất lớn, khả năng phát triển nhanh, và rất khó thu thập, lưu trữ, quản lý và phân tích với các công cụ thống kê hay ứng dụng cơ sở dữ liệu truyền thống [5]. Một số đặc trưng của Big Data bao gồm Dung lượng (volume), Tốc độ (velocity), Tính đa dạng (variety), và Giá trị (value)

(1) *Dung lượng (Volume)*: Dung lượng của Big Data đang tăng lên mạnh mẽ từng ngày. Theo tài liệu của Intel vào tháng 9/2013, cứ mỗi 11 giây, 1 PB[1] dữ liệu được tạo ra trên toàn thế giới, tương đương với một đoạn video HD dài 13 năm [8]. Facebook phải xử lý khoảng 500 TB[2] dữ liệu mỗi ngày [9]. Lợi ích thu được từ việc xử lý một khối lượng lớn dữ liệu chính là điểm thu hút chủ yếu của Big Data, tuy nhiên cũng đặt ra nhiều khó khăn trong việc tìm ra những phương pháp, kỹ thuật để xử lý khối lượng dữ liệu này.

(2) *Tốc độ (velocity)*: với sự ra đời của các kỹ thuật, công cụ, ứng dụng lưu trữ, nguồn dữ liệu liên tục được bổ sung với tốc độ nhanh chóng. Tổ chức McKinsey Global ước tính lượng dữ liệu đang tăng trưởng với tốc độ 40%/năm, và sẽ tăng 44 lần từ năm 2009 đến 2020. [6].

(3) *Tính đa dạng (variety)*: Dữ liệu được thu thập từ nhiều nguồn khác nhau, từ các thiết bị cảm biến, thiết bị di động, qua mạng xã hội .v.v… [4]. Các kiểu dữ liệu có cấu trúc, bán cấu trúc và không có cấu trúc tồn tại dưới nhiều hình thức bao gồm hình ảnh, âm thanh, video, văn bản, v.v…

[1] 1 petabyte (PB)= 2^{20} gigabytes (GB) ~ 1,000,000 GB

[2] 1 terabyte (TB) = 2^{10} gigabytes (GB) ~ 1000 GB

(4) *Giá trị (value)*: đây là đặc trưng quan trọng nhất của Big Data, đề cập đến quá trình trích xuất các giá trị to lớn đang tiềm ẩn trong các bộ dữ liệu khổng lồ.

PHẦN 3 CÁC ỨNG DỤNG BIG DATA

Big Data và các ứng dụng có liên quan đang ngày càng được sử dụng rộng rãi trong các tổ chức, trong các lĩnh vực khác nhau, nhằm giảm thiểu các rủi ro, hỗ trợ tổ chức trong việc quản lý các hoạt động hằng ngày cũng như ra quyết định [7]. Các cơ quan chính phủ tìm cách phân tích dữ liệu nhằm tìm ra những cách thức thu thuế một cách khéo léo, dự đoán được tỷ lệ thất nghiệp, xu hướng nghề nghiệp trong tương lai [8], các doanh nghiệp trong lĩnh vực y tế cũng đang chủ động hơn trong việc quản lý và theo dõi sức khỏe khách hàng, thiết kế các gói sản phẩm hợp lý nhằm giảm chi phí chăm sóc sức khỏe. Ngành khách sạn và du lịch sử dụng dữ liệu từ nhiều nguồn như mạng xã hội và tạo ra những gói kỳ nghỉ cá nhân cho các khách hàng. Các doanh nghiệp phân tích dữ liệu nhằm tìm hiểu hành vi khách hàng và tư vấn cho họ về danh mục sản phẩm, thời gian và địa điểm mua có những chính sách giá hấp dẫn.

Nhiều nghiên cứu đã tìm hiểu về các ứng dụng của Big Data và các lĩnh vực trong đó Big Data có thể được áp dụng. Chẳng hạn, Hsinchun, Chiang [10] phân tích một số ứng dụng của Big Data bao gồm thương mại điện tử, chính phủ điện tử, (3) khoa học và công nghệ, chăm sóc sức khỏe, và an ninh và an toàn công cộng. O'Leary [4] mô tả một số ưu điểm cũng như trở ngại của Big Data và các ứng dụng nền tảng cảm biến trên thiết bị di động trong quản lý cơ sở hạ tầng đường bộ. McKinsey & Company thực hiện nghiên cứu về những giá trị dữ liệu mang lại đối với y tế, quản lý công, bán lẻ, sản xuất ở Mỹ. Báo cáo nêu rõ nếu Big Data được sử dụng một cách sáng tạo và hiệu quả để cải tiến năng suất và chất lượng công việc, các doanh nghiệp bán lẻ Mỹ có thể tăng lợi nhuận trên 60%, chi tiêu cho công nghiệp y tế Mỹ có thể giảm trên 8%, các nền kinh tế phát triển ở châu Âu cũng có thể tiết kiệm được 149 triệu Euro nhờ việc cải tiến hiệu suất hoạt động [5].

Bài viết này tổng hợp một số lĩnh vực mà Big Data được áp dụng, cũng như các kỹ thuật được sử dụng qua Bảng 1 và một số ví dụ cụ thể về ứng dụng của Big Data qua Bảng 2. Các khía cạnh kỹ thuật liên quan đến Big Data sẽ được trình bày cụ thể hơn ở Phần 5

Lĩnh vực	Ứng dụng	Kỹ thuật	Trích dẫn
Thương mại	• Phân khúc thị trường và khách hàng • Phân tích hành vi khách hàng tại cửa hàng • Tiếp thị trên nền tảng định vị • Phân tích tiếp thị chéo kênh, tiếp thị đa kênh • Quản lý các chiến dịch tiếp thị và khách hàng thân thiết • So sánh giá • Phân tích và quản lý chuỗi cung ứng	• Phân tích cấu trúc dữ liệu • Phân tích mạng xã hội • Phân tích văn bản và web • Phân tích tâm lý • Phát hiện yếu tố bất thường	[1, 5, 10]

Tài chính	• Phân tích thông tin khách hàng trong thời gian thực • Quản lý mối quan hệ khách hàng • Quản lý và phân tích rủi ro • Phân tích và phát hiện gian lận • Phân tích, xếp hạng rủi ro tín dụng	• Phân tích đa phương tiện • Phân tích mạng • Phân tích mạng xã hội • Phân tích văn bản và web • Phân tích tâm lý	[1]
Chính trị, chính phủ điện tử	• Hệ thống chính phủ điện tử. • Hệ thống bầu cử, bỏ phiếu • Phân tích quy định và việc tuân thủ quy định. • Phân tích, giám sát, theo dõi và phát hiện gian lận, mối đe dọa, an ninh mạng, phát hiện xâm nhập. • Quản lý tiêu thụ năng lượng và khí thải carbon	• Phân tích nội dung và văn bản • Phân tích thông tin ngữ nghĩa • Phân tích và giám sát phương tiện truyền thông xã hội • Phân tích mạng xã hội • Phân tích tâm lý	[1, 10]
An ninh và an toàn công cộng	• Phân tích tội phạm • Tội phạm công nghệ cao • Khủng bố • An ninh mạng	• Luật kết hợp • Phân cụm và phân lớp dữ liệu • Phân tích mạng lưới tội phạm • Phân tích quan điểm • Phân tích tấn công mạng	[10]
Viễn thông	• Phân tích hành vi, thói quen người tiêu dùng • Phân tích định vị người dùng di động • Ghi nhận chi tiết cuộc gọi trong thời gian thực • Tối ưu hóa hệ thống	• Xử lý chi tiết cuộc gọi • Mô hình dự báo • Phân tích tâm lý • Phân tích di động	[1, 11]
Y tế và chăm sóc sức khỏe	• Phân tích dữ liệu chẩn đoán trong thời gian thực • Phân tích chất lượng dịch vụ chăm sóc bệnh nhân • Phát hiện gian lận bảo hiểm y tế • Tối ưu hóa các gói sức khỏe.	• Phân tích văn bản • Phân tích Web • Phân tích đa phương tiện • Phân tích mạng	[1, 10, 12]
Giao thông vận tải	• Phân tích dữ liệu thời tiết và giao thông trong thời gian thực • Tối ưu hóa tuyến đường vận chuyển • Giảm thiểu tình trạng ùn tắc giao thông	• Phân tích văn bản • Phân tích Web • Phân tích đa phương tiện • Phân tích di động	[4, 13]

Bảng 1: Một số lĩnh vực khai thác ứng dụng Big Data

Lĩnh vực	**Công ty/ Sự kiện**	**Mục tiêu**	**Kỹ thuật áp dụng**	**Kết quả**	**Trích dẫn**

Tài chính	China Merchants Bank (CMB)	Quản lý mối quan hệ với khách hàng	• Hoạt động: Tích điểm, đổi điểm tích lũy • Xây dựng hệ thống cảnh báo khách hàng ngưng sử dụng dịch vụ	• Bán được các sản phẩm tín dụng lãi suất cao cho 20% khách hàng có khả năng ngưng sử dụng dịch vụ của ngân hàng • Tỷ lệ khách ngưng sử dụng thẻ Gold Cards giảm 15%, thẻ Sunflower Cards giảm 7%.	[13]
Thương mại	Amazon	Quản lý mối quan hệ với khách hàng, tăng doanh thu bán hàng	Xây dựng hệ tư vấn, sử dụng thuật toán item-to-item collaborative filtering match[1]	Doanh thu bán hàng của công ty tăng 29% từ USD 9.9 tỷ đô la (quý 2, 2011) lên $12.83 tỷ (quý 2, 2012)	[14]
Giao thông vận tải, vận chuyển	UPS[2] sử dụng ứng dụng Big Data và IoT	Tối ưu hóa các tuyến đường vận chuyển, thiết kế lại việc bốc và dỡ hàng của tài xế trong thời gian thực	• Sử dụng hệ thống định vị toàn cầu (GPS), thiết bị cảm biến để theo dõi vị trí các xe tải • Các kỹ thuật tối ưu hóa để tìm tuyến đường tối ưu nhất	Năm 2011, công ty tiết kiệm được 42.28 triệu km vận chuyển	[13]
Y tế	Google và dịch cúm A/H1N1 năm 2009	Phân tích dữ liệu dịch cúm trong thời gian thực, nhằm dự báo khả năng lan tỏa dịch cúm	Phân tích các truy vấn tìm kiếm về dịch cúm	Dự báo chính xác mức độ hiện tại của dịch cúm tại mỗi khu vực của Hoa Kỳ, thời gian, địa bàn lan tỏa dịch cúm.	[15]
Y tế	Phụ sản Quốc tế Sài Gòn (SIH)	Quản lý mối quan hệ với khách hàng, cung cấp dịch vụ tốt nhất cho bệnh nhân	• Sử dụng dịch vụ phần cứng và phần mềm của IBM để xây dựng mô hình dự đoán nhu cầu bệnh nhân, lên lịch và điều phối khám bệnh	Các thông tin về bệnh nhân được cung cấp kịp thời giúp nâng cao hoạt động của bệnh viện và bác sĩ	[16]

[1] Thuật toán item-to-item collaborative filtering match: thuật toán này xây dựng một ma trận các sản phẩm tương đồng bằng cách tìm kiếm những sản phẩm thường được mua cùng với nhau để tư vấn cho người dùng những sản phẩm đi kèm phù hợp nhất đối với sản phẩm họ lựa chọn

[2] UPS (United Parcel Service of North America, Inc.) là công ty vận tải lớn nhất thế giới.

Quản lý công	Obama thắng chiến dịch bầu cử	Tìm những hình thức vận động cử tri thích hợp	• Thu thập và tạo ra một cơ sở dữ liệu chứa thông tin của các cử tri tiềm năng bao gồm tiểu sử, sở thích, công việc, bạn bè • Sử dụng các kỹ thuật khai phá dữ liệu	Obama và đội ngũ của mình có những hoạt động vận động thích hợp với cử tri, góp phần đáng kể vào chiến thắng cuối cùng.	[17]
An ninh và an toàn công cộng	Sở cảnh sát Los Angeles sử dụng ứng dụng đám mây của PredPol Inc. trong dự báo tội phạm	Xác định thời gian, vị trí có thể xảy ra các hành vi phạm tội như trộm cắp, bạo hành	Dựa trên dữ liệu lịch sử để dự đoán thời gian, địa điểm của hoạt động phạm tội	Trong 6 tháng, số vụ trộm cắp giảm 33%, hành vi bạo lực giảm 21%	[18]

Bảng 2 Một số ví dụ tiêu biểu về ứng dụng Big Data

PHẦN 4 NHẤN MẠNH ỨNG DỤNG CỦA BIG DATA TRONG LĨNH VỰC TÀI CHÍNH, THƯƠNG MẠI, BẢO HIỂM VÀ QUẢN LÝ CÔNG

4.1 Tài chính ngân hàng, bảo hiểm

Nhiều cuộc khảo sát được thực hiện để xác định vai trò của Big Data trong hoạt động của tổ chức. Khảo sát của Gartner FEI năm 2013 nhấn mạnh tầm quan trọng của BI&A[1] trong công việc của các giám đốc tài chính [19]. Nhờ khung nhìn tổng quan, rõ ràng vào dữ liệu của tổ chức, các giám đốc tài chính có thể có những quyết định tốt hơn, làm tăng hiệu quả hoạt động của tổ chức, tăng tính liên kết giữa tài chính và hoạt động kinh doanh chung, cũng như tăng cường tính linh hoạt của tổ chức. Một ví dụ từ ngân hàng China Merchants Bank (CMB), Trung Quốc cho thấy hiệu quả của việc ứng dụng Big Data. Để thu hút khách hàng, ngân hàng sử dụng dịch vụ tích điểm và đổi điểm[2]. Ngân hàng cũng sử dụng mô hình cảnh báo khả năng người dùng ngưng sử dụng dịch vụ để xây dựng các gói dịch vụ tín dụng lãi suất cao nhằm giữ chân khách hàng. Đồng thời, thông qua việc phân tích dữ liệu các giao dịch, các khách hàng tiềm năng, là các doanh nghiệp nhỏ, cũng được xác định một cách hiệu quả [13].

[1] BI&A (Business Intelligence and Analytics): chỉ những kỹ thuật, công nghệ, phương pháp, ứng dụng, hệ thống phân tích các dữ liệu nghiệp vụ nhằm giúp nhà quản lý hiểu rõ hơn về tình hình hoạt động kinh doanh của tổ chức, cũng như tình hình thị trường, từ đó ra các quyết định kịp thời nhằm đạt được mục tiêu của tổ chức (Xem Hsinchun, C., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS Quarterly).*

[2] Tích điểm: Multi-times score accumulation

Đổi điểm: score exchange in shops

Có nhiều nguyên nhân dẫn đến quyết định đầu tư vào các dự án Big Data. Bài báo này tổng hợp ứng dụng tiêu biểu của Big Data trong ngành tài chính, ngân hàng và bảo hiểm và chia thành 3 nhóm chính: quản lý rủi ro, tư vấn và dự báo. Trên thực tế, nhiều ứng dụng về Big Data được nghiên cứu và phát triển nhằm cải tiến hiệu quả hoạt động của các tổ chức tín dụng và bảo hiểm [20].

4.1.1 Quản lý rủi ro

Hoạt động quản lý rủi ro được cải thiện đáng kể nhờ những tác động của Big Data. Trước đây, hoạt động phân tích các tình huống rủi ro chủ yếu phụ thuộc vào việc phân tích khách hàng, các danh mục đầu tư, độ tin cậy tín dụng. Hiện nay, với những nguồn dữ liệu từ các phương tiện truyền thông xã hội cho phép tạo ra những hiểu biết mới về các danh mục rủi ro của khách hàng. Các dữ liệu thu được từ nhiều nguồn không liên kết làm tăng khả năng phát hiện các hoạt động gian lận sớm hơn so với các phương pháp hiện hành [5].

Hiểu về rủi ro và làm thế nào để quản lý rủi ro tốt hơn là mối quan tâm chính của các công ty bảo hiểm. Phân tích rủi ro bao gồm việc đanh giá khả năng rủi ro xảy ra và chi phí phải bỏ ra trong từng trường hợp rủi ro. Những dữ liệu như mưa đá, cháy rừng, bão lụt, tội phạm và các yếu tố khác cần được khai thác và tận dụng để đánh giá rủi ro. Các dữ liệu từ các thiết bị viễn thông, thiết bị cảm biến được cài đặt trong các phương tiện giao thông có thể thu thập những dữ liệu như địa điểm, tốc độ, quãng đường đi, tình trạng vận hành của phương tiện trong thời gian thực, giúp cải thiện khả năng đánh giá rủi ro, từ đó, doanh nghiệp có thể tạo ra nhiều chiến lược giá khác nhau.

4.1.2 Tư vấn

Big Data và các ứng dụng liên quan cho phép các tổ chức tài chính thu thập và tổ chức các dữ liệu như sở thích của khách hàng, lịch sử giao dịch, phương thức giao dịch, vị trí địa lý, thông tin gia đình, v.v... Từ đó, hệ tư vấn sẽ dựa vào mục tiêu kinh doanh của ngân hàng, nhu cầu của Khách hàng để từ đó đưa ra các kiến nghị về bán chéo[1], bán thêm[2] hoặc cung cấp các dịch vụ, tốt hơn cho khách hàng. Thông qua việc phân tích dữ liệu khách hàng ở cấp độ tinh vi hơn, các tổ chức còn có thể tạo ra những cơ hội mới từ việc tạo ra những sản phẩm mục tiêu mới.

4.1.3 Dự báo

Các kỹ thuật thống kê trên dữ liệu lịch sử cho phép dự đoán các hành động tiếp theo của khách hàng. Nền tảng phân tích dữ liệu lớn thông qua việc sử dụng các kỹ thuật xử lý phân tán (Map-Reduce) cho phép tổ chức tài chính, ngân hàng có thể lưu trữ, xử lý khối lượng dữ liệu rất lớn. Nhờ vậy, các mô hình dự báo có thể chạy trên toàn bộ các tập dữ liệu, giúp rút ngắn thời gian trích xuất, khám phá những thông tin quý giá còn tiềm ẩn. [1]

---

[1] Bán chéo (Cross-selling): là một thuật ngữ để chỉ cách thức giới thiệu những sản phẩm hoặc dịch vụ có liên quan đến sản phẩm khách hàng đang hoặc đã mua. Ví dụ, nếu khách hàng đã mua điện thoại, thì thuyết phục khách hàng mua thêm vỏ điện thoại.

[2] Bán thêm (Up-selling): là một thuật ngữ để chỉ cách thức giới thiệu những sản phẩm hoặc dịch vụ có giá cao hơn, hay nâng cấp sản phẩm, dịch vụ với những tính năng bổ sung

4.2 Thương mại

Các phân tích trên lượng dữ liệu lớn còn góp phần cải tiến và tối ưu hóa quá trình ra quyết định, giảm thiểu rủi ro, tạo ra những giá trị gia tăng cho doanh nghiệp. Bằng việc khai thác nền tảng phân tích dữ liệu lớn, các doanh nghiệp có thể khám phá các giá trị tiềm ẩn to lớn, thông qua các khung nhìn tổng hợp về hành vi mua hàng của khách hàng. Chẳng hạn, các công ty kinh doanh qua mạng chẳng những có thể theo dõi để biết được không chỉ những thông tin như khách hàng mua gì, mà còn biết được họ xem những mặt hàng nào, họ xem những gì, làm gì mỗi lần họ truy cập vào trang web, hay mức độ khách hàng bị tác động bởi những chính sách khuyến mãi hay bình luận từ những khách hàng khác; từ đó phát hiện ra được những điểm chung của những nhóm khách hàng.

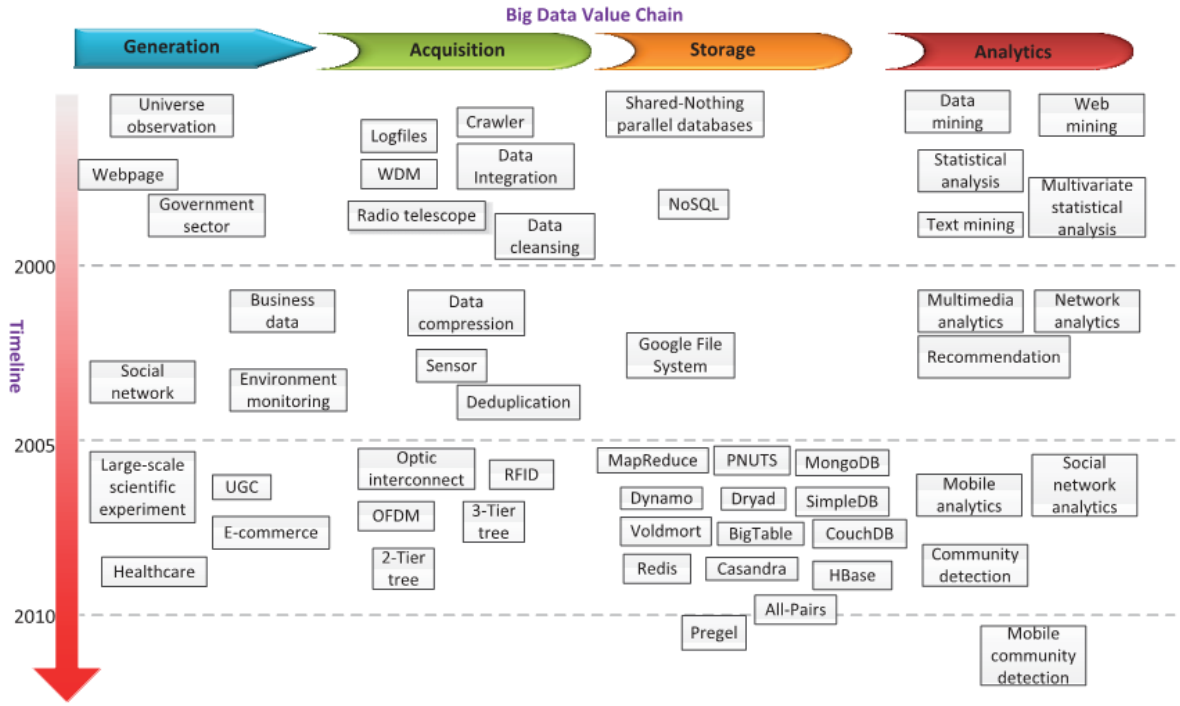
Ngoài ra, sự phát triển của Internet, web 2.0, các thiết bị di động cho phép tổ chức sử dụng nhiều phương thức khác nhau để tương tác với khách hàng bên cạnh các phương tiện truyền thống. Việc phân tích các giao dịch của khách hàng qua các kênh khác nhau này cho phép tổ chức hiểu hành vi khách hàng, phân cụm nhóm khách hàng, từ đó có thể cung cấp các sản phẩm và dịch vụ phù hợp với yêu cầu khách hàng.

Big Data còn mang lại lợi ích cho các doanh nghiệp trong việc lên kế hoạch bán hàng. Bằng việc so sánh các yếu tố khác nhau từ nguồn dữ liệu khổng lồ, doanh nghiệp có thể tối ưu hóa việc định giá cho các sản phẩm. [13]. Việc sử dụng Big Data trong quản lý chuỗi cung ứng cho phép các doanh nghiệp tối ưu hóa dự trữ kho, vận chuyển, phối hợp với nhà cung cấp nhằm giảm thiểu khoảng cách giữa cung và cầu, kiểm soát ngân sách, và cải thiện dịch vụ. [5, 13]

PHẦN 5 KHÍA CẠNH KỸ THUẬT

5.1 Luồng dữ liệu trong Big Data

Hệ thống Big data thường lớn và phức tạp, nó cung cấp các chức năng để xử lý Big Data từ lúc hình thành đến lúc kết thúc. Thường luồng dữ liệu trong Big Data được phân làm 4 giai đoạn: Nguồn tạo ra dữ liệu, Thu thập dữ liệu, Lưu trữ dữ liệu và Phân tích dữ liệu[13]. Hình 1 bên dưới mô tả các công nghệ liên quan đến 4 giai đoạn của luồng dữ liệu:



Hình 1: Bản đồ công nghệ của Big data theo luồng dữ liệu [12]

5.1.1 Nguồn tạo ra dữ liệu

Do sự phát triển vượt bậc của các công nghệ hiện đại nên nguồn tạo ra dữ liệu ngày càng phát triển mạnh mẽ. Thật vậy, IBM ước tính rằng 90% dữ liệu trong thế giới ngày nay đã được tạo ra trong 2 năm qua [21]. Nguyên nhân của sự bùng nổ dữ liệu này cũng có nhiều tranh cãi. Theo [12] sự bùng nổ dữ liệu có liên hệ mật thiết đến sự phát triển của công nghệ thông tin, được chia làm 3 giai đoạn như sau:

Giai đoạn 1: bắt đầu từ những năm 1990. Khi công nghệ số và những hệ thống cơ sở dữ liệu được áp dụng rộng rãi, nhiều tổ chức đã sử dụng chúng để lưu trữ những dữ liệu lớn của họ như các giao dịch trong lãnh vực ngân hàng hay các trung tâm tài chính, các tài liệu của chính phủ… Đây là những dữ liệu có cấu trúc và được phân tích thông qua hệ thống CSDL quan hệ.

Giai đoạn 2: giai đoạn 2 bắt đầu bằng sự bùng nổ của Internet. Vào những năm cuối của thập niên 90, hệ thống Web 1.0, đặc trưng bởi các công cụ tìm kiếm và thương mại điện tử, tạo ra 1 lượng lớn dữ liệu bán cấu trúc và/hoặc không cấu trúc, bao gồm các trang web và lịch sử giao dịch. Kể từ những năm 2000, rất nhiều các ứng dụng Web 2.0 đã tạo ra một lượng dữ liệu phong phú các dữ liệu do người dùng đóng góp từ các diễn đàn, nhóm, blog, các trang web, mạng xã hội.

Giai đoạn 3: được kích thích bởi các thiết bị di động như điện thoại thông minh, máy tính bảng, cảm biến và các thiết bị hỗ trợ Internet dựa trên cảm biến.

Với cách phân loại này, chúng ta thấy rằng các mô hình tạo dữ liệu phát triển 1 cách nhanh chóng, từ lưu trữ thụ động trong giai đoạn 1 đến tạo dữ liệu tích cực trong giai đoạn 2 và tạo dữ liệu tự động trong giai đoạn 3. Ba loại dữ liệu này chính là nguồn dữ liệu chính của Big Data, trong đó các dữ liệu tự động sẽ đóng góp nhiều nhất trong tương lai gần.

Bảng 3 bên dưới mô tả sự phát triển nhanh chóng của các dữ liệu và giao dịch của các dịch vụ phổ biến trên mạng hiện nay. Đây là dữ liệu rất quan trọng đối với các doanh nghiệp, thông qua việc khai thác và phân tích các loại dữ liệu này, những thông tin hữu ích như thói quen và sở thích của người sử dụng có thể được xác định, và nó thậm chí có thể dự đoán hành vi và trạng thái cảm xúc của người sử dụng.

Dịch vụ	Mô tả
YouTube	(i) Mỗi phút người dùng tải lên 100 giờ video (ii) Mỗi tháng, hơn 1 tỷ người sử dụng truy cập vào YouTube
Facebook	(i) Mỗi phút, 34.722 Likes (ii) 100 terabytes (TB) dữ liệu được tải lên mỗi ngày (iii) Hiện nay, có 1,4 tỷ người sử dụng
Twitter	(i) Có hơn 645 triệu người sử dụng (ii) Mỗi ngày có 175 triệu tweet
Google	(i) Hơn 2 triệu truy vấn tìm kiếm mỗi phút (ii) Mỗi ngày, 25 petabyte (PB) được xử lý
Apple	Khoảng 47.000 ứng dụng được tải xuống mỗi phút
Flickr	3.125 người dùng tải lên các bức ảnh mới mỗi phút
WordPress	Mỗi phút có gần 350 blogs mới

Bảng 3: Sự phát triển của các dịch vụ phổ biến trên mạng hiện nay[22]

Big Data được tạo ra trong nhiều lãnh vực khác nhau. Bảng 4 liệt kê các nguồn Big Data từ các lãnh vực khác nhau cùng với các thuộc tính quan trọng của các loại dữ liệu này. Ta dễ dàng nhận thấy phần lớn các nguồn dữ liệu là không có cấu trúc, với độ lớn là PB và đòi hỏi phải phân tích nhanh chóng, chính xác với 1 lượng lớn người dùng

Nguồn	Lãnh vực	Độ lớn	Loại	Thời gian đáp ứng	Số lượng người dùng	Độ chính xác	Tham khảo
Walmart	Bán lẻ	2.5 PB/giờ	Có cấu trúc	Rất nhanh	1 triệu/giờ	Rất cao	[23]
Amazon	TMĐT	Nhiều PB/ngày	Bán cấu trúc	Rất nhanh	0.5 triệu/ ngày	Rất cao	[12]
Tìm kiếm Google	Tìm kiếm	25 PB/ngày	Bán cấu trúc	Nhanh	2 triệu/phút	Cao	[12]
Facebook	Mạng xã hội	100 TB/ngày	Có cấu trúc, không cấu trúc	Nhanh	1.4 tỉ	Cao	[24]

AT&T	Mạng di động	323 TB	Có cấu trúc	Nhanh	Rất lớn	Cao	[24]
SDSS	Khoa học	20 TB/ngày	Không cấu trúc	Chậm	Nhỏ	Rất cao	[12]

Bảng 4: Một số nguồn Big Data điển hình

5.1.2 Thu thập dữ liệu

Thu thập dữ liệu trong Big data thường gồm 3 bước: thu nhận dữ liệu, truyền dữ liệu và tiền xử lý dữ liệu. Đôi khi việc thu nhận dữ liệu sẽ chứa những dữ liệu dư thừa hoặc không cần thiết làm tăng dung lượng lưu trữ và cũng ảnh hưởng đến tốc độ xử lý của giai đoạn phân tích tiếp theo. Do đó, hoạt động tiền xử lý là không thể thiếu để đảm bảo cho việc lưu trữ và khai thác dữ liệu hiệu quả.

i. Thu nhận dữ liệu

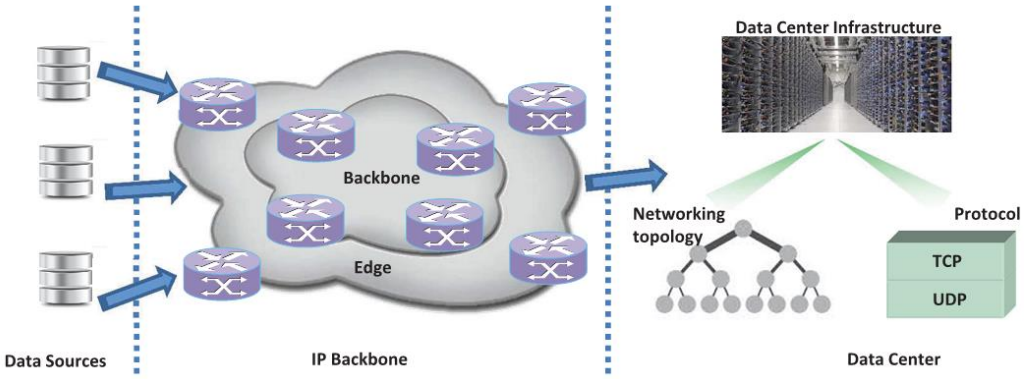
Thu nhập dữ liệu là quá trình lấy dữ liệu thô từ các đối tượng trong thế giới thực. Quá trình này cần được thiết kế tốt vì phương phát thu nhận dữ liệu không chỉ phụ thuộc vào các đặc tính vật lý của các nguồn dữ liệu, mà còn phụ thuộc vào mục tiêu của phân tích dữ liệu. Bảng 5 mô tả các phương pháp thu nhận dữ liệu phổ biến hiện nay: Kết quả là có nhiều loại phân tích dữ liệu như bảng bên dưới:

Phương pháp	**Loại**	**Độ lớn**	**Độ phức tạp**	**Ứng dụng**	**Tham khảo**
Log file	Cấu trúc hoặc bán cấu trúc	Nhỏ	Dễ	Nhật ký web, tài chính	[12]
Cảm biến	Cấu trúc hoặc bán cấu trúc	Trung bình	Cao	Video giám sát, nghiên cứu môi trường	[12]
Web crawler	Hỗn hợp	Lớn	Trung bình	Tìm kiếm	[12]
Libpcap-based hoặc Zero-copy	Cấu trúc	Trung bình	Trung bình	Giám sát mạng	[13]
Mobile	Hỗn hợp	Nhỏ	Trung bình	Mobile gián điệp	[13]

Bảng 5: Các phương pháp thu nhận dữ liệu

ii. Truyền dữ liệu

Sau khi thu nhận dữ liệu, chúng ta phải chuyển nó vào trung tâm dữ liệu để chuẩn bị cho các bước tiếp theo. Cơ chế truyền dữ liệu có thể được chia làm 2 giai đoạn: truyền qua IP backbone và truyền về trung tâm dữ liệu:



Hình 2: Cơ chế truyền dữ liệu [12]

iii. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một giai đoạn quan trọng trong quá trình thu thập dữ liệu để tăng chất lượng của dữ liệu trong những hệ thống Big Data. Theo [13] có 3 phương pháp chính để thực hiện tiền xử lý dữ liệu, được trình bày ở Bảng 6

Kỹ thuật	Mô tả	Phương pháp	Ứng dụng
Tích hợp	Kết hợp dữ liệu từ nhiều nguồn khác nhau và cung cấp cho người dùng một khung nhìn thống nhất	• Kho dữ liệu: cũng được gọi là ETL (chiết, chuyển đổi và nạp) [25] • Dữ liệu liên kết (data federation): một CSDL ảo được tạo ra để truy vấn và tổng hợp dữ liệu	• Công cụ xử lý lưu lượng • Công cụ tìm kiếm [26]
Làm sạch	Xác định dữ liệu không chính xác, không đầy đủ, hoặc không hợp lý, sau đó chỉnh sửa hay xóa dữ liệu để tăng chất lượng của dữ liệu	Thường gồm 5 bước: xác định loại lỗi, tìm kiếm lỗi, sửa lỗi, tài liệu hóa loại lỗi, thay đổi cách nhập dữ liệu để giảm lỗi trong tương lai [27]	• Thương mại điện tử [28] • RFID • Sinh học
Loại bỏ dư thừa	Loại bỏ những dữ liệu dư thừa và bị lặp lại	Phát hiện dư thừa, lọc dữ liệu, nén dữ liệu	• Tìm kiếm đa phương tiện • Phân tích ADN

Bảng 6: Các phương pháp chính để thực hiện tiền xử lý dữ liệu [13]

5.1.3 Lưu trữ dữ liệu

Cơ chế lưu trữ trong Big Data cũng khác so với cách lưu trữ bình thường. Vì số lượng dữ liệu quá lớn, phải cần 1 cơ chế lưu trữ sao cho hiệu quả về mặt kinh tế, cũng như hiệu quả về mặt kỹ thuật, phục vụ cho việc phân tích và thống kê sau này. Hiện nay có 3 cách lưu trữ thường sử dụng: (i) Hệ thống tập tin; (ii) Cơ sở dữ liệu và (iii) mô hình lập trình

i. Hệ thống tập tin

Cách lưu trữ kiểu tập tin đã phát triển từ lâu và tương đối trưởng thành sau 1 thời gian dài áp dụng. Có thể kể đến 1 số kỹ thuật nổi tiếng và sử dụng phổ biến hiện nay như Bảng 7:

Kỹ thuật	Hãng	Đặc điểm
GFS	Google	Có thể hoạt động trên những máy chủ rẻ tiền để cung cấp khả năng chịu lỗi và hiệu suất cao cho một số lượng lớn máy khách. Nó phù hợp cho ứng dụng với kích thước file và hoạt động đọc nhiều hơn ghi
Cosmos	Microsoft	Hỗ trợ tìm kiếm và quảng cáo
Haystack	Facebook	Hỗ trợ lưu trữ số lượng lớn những file ảnh nhỏ

Bảng 7: Một số kỹ thuật lưu trữ dữ liệu

ii. Cơ sở dữ liệu

Kỹ thuật CSDL đã được phát triển hơn 30 năm. Rất nhiều hệ thống CSDL được phát triển để những loại dữ liệu khác nhau và dễ dàng mở rộng. CSDL quan hệ truyền thống không thể đáp ứng được tính chất 4V của Big Data. Do đó, CSDL NoSQL đang trở thành công nghệ cốt lõi cho Big Data. Bởi vì nó có nhiều ưu điểm vượt bậc và rất thích hợp với Big Data như: lược đồ linh hoạt, API đơn giản, thống nhất, hỗ trợ 1 số lượng lớn dữ liệu. Hiện tại có 3 loại NoSQL chính: Key-value, hướng cột và hướng tài liệu. Bảng 8 mô tả sơ lược về 3 loại CSDL này cùng với các thư viện nổi tiếng hỗ trợ:

Kỹ thuật	Đặc điểm	Thư viện	Kiểu lưu trữ	Ứng dụng
Key-value	Được thiết lập bởi 1 mô hình dữ liệu đơn giản, dữ liệu được lưu trữ theo kiểu khóa và giá trị	• Dynamo (Amazon) • Voldemort (LinkedIn)	• Plug-in • RAM	• Y tế [2] • Không gian [19] • Mạng xã hội, thương mại điện tử [7]
Hướng cột	Lưu trữ và xử lý dữ liệu theo cột. Cả hàng và cột được phân đoạn và lưu trữ trong nhiều máy chủ để tăng khả năng mở rộng	• BigTable (Google) • Cassandra (Facebook) • Hbase (Apache)	• GFS • Ổ đĩa • HDFS	• Google Earth, Analystics [29] • Nhiều lãnh vực1 • Chính phủ điện tử [30], nhiều lãnh vực

[1] Rất nhiều công ty từ các lãnh vực khác đã áp dụng và sử dụng Casandra cũng như Hbase, có thể tham khảo thêm tại: http://planetcassandra.org/companies/ và http://wiki.apache.org/hadoop/Hbase/PoweredBy

Hướng tài liệu	Hỗ trợ lưu trữ kiểu dữ liệu phức tạp hơn kiểu Key-value. Không có ràng buộc nào về mẫu của tài liệu lưu trữ	• SingleDB (Amazon) • MongoDB (10gen) • CouchDB (Couchbase)	• S3 • Ổ đĩa • Ổ đĩa	• Logging, game online • Sinh học[31] • Rất nhiều lãnh vực khác[1]

Bảng 8: Các loại Cơ sở dữ liệu NoSQL

Ngoài những CSDL trên, nhiều dự án được phát triển để hỗ trợ những kiểu dữ liệu khác nhau như: biểu đồ (Neo4j, DEX) và PNUTS. Bởi vì CSDL quan hệ và CSDL NoSQL có những ưu và khuyết điểm riêng nên nếu kết hợp được 2 loại CSDL này thì sẽ sinh ra một loại CSDL mới vừa mạnh mẽ trong truy vấn giống CSDL quan hệ vừa linh hoạt và dễ dàng mở rộng như CSDL NoSql. Hiện tại, Google đang đi tiên phong phát triển 1 loại các loại CSDL mới theo hướng này như: Megastore, Spanner và F1.

Không có 1 loại CSDL nào là phù hợp cho mọi tình huống và mọi loại dữ liệu. Tùy theo từng bài toán cụ thể mà chúng ta nên chọn CSDL cho phù hợp, vì mỗi loại CSDL cũng có những ưu và khuyết riêng. Nhiều khi chúng ta phải đánh đổi giữa hiệu suất đọc và hiệu suất ghi, đồng bộ và không đồng bộ, độ trễ và độ bền, phân vùng dữ liệu [11]

iii. Những mô hình lập trình

Mặc dù NoSQL có rất nhiều điểm mạnh và phù hợp với Big Data nhưng nó vẫn còn những mặc hạn chế về truy vấn và phân tích dữ liệu. Mô hình lập trình rất phù hợp với các ứng dụng logic và phân tích dữ liệu. Tuy nhiên, những mô hình song song truyền thống (Message Passing Interface (MPI) and Open Multi-Processing (OpenMP)) vẫn còn những mặt hạn chế để giải quyết các bài toán song song trên quy mô Big Data, tức là hàng trăm thậm chí hàng ngàn máy chủ trên diện rộng. Nhiều mô hình lập trình song song mới cho Big Data đã được đề xuất. Bảng bên dưới so sánh tính năng của những mô hình lập trình hiện nay cho Big Data[12]:

	MapReduce	**Dryad**	**Pregel**	**GraphLab**	**Storm**	**S4**
Mô tả	Xử lý song song ở quy mô lớn	Xử lý song song ở quy mô lớn	Xử lý dạng đồ thị ở quy mô lớn	Khai phá dữ liệu và máy học ở quy mô lớn	Xử lý phân phối thời gian thực	Xử lý phân phối thời gian thực
Mô hình lập trình	Map và Reduce	Đồ thị phi chu trình trực tiếp	Đồ thị trực tiếp	Đồ thị trực tiếp	Đồ thị phi chu trình trực tiếp	Đồ thị phi chu trình trực tiếp

[1] Rất nhiều lãnh vực khác, có thể tham khảo thêm tại: https://www.mongodb.com/industries

Xử lý dữ liệu	Hệ thống tập tin phân phối	Nhiều kiểu lưu trữ khác nhau	Hệ thống tập tin phân phối	Bộ nhớ hay đĩa	Bộ nhớ	Bộ nhớ
Kiến trúc	Chủ-Khách	Chủ-Khách	Chủ-Khách	Chủ-Khách	Chủ-Khách	Phân cấp và đối xứng
Chịu lỗi	Cấp node	Cấp node	Checkpoint[1]	Checkpoint	Một phần	Một phần

Bảng 9: Tổng hợp những mô hình lập trình

Theo bảng so sánh ở trên, ta nhận thấy rằng: i) Mặc dù xử lý thời gian thực đang tập trung nghiên cứu hiện nay, nhưng phần lớn các mô hình vẫn tập trung vào xử lý hàng loạt; ii) Hầu hết các mô hình đều sử dụng đồ thị bởi vì đồ thị có thể thể hiện các tác vụ phức tạp hơn; iii) Xử lý thời gian thực sử dụng bộ nhớ như là phương tiện lưu trữ dữ liệu để đạt được tốc độ truy cập và xử lý cao hơn, trong khi mô hình hàng loạt sử dụng hệ thống tập tin hay đĩa để lưu trữ dữ liệu lớn hơn và hỗ trợ nhiều client; iv) Kiến trúc của các mô hình thường là Chủ-Khách; v) Các chiến lược khả năng chịu lỗi là khác nhau

5.1.4 Phân tích dữ liệu

Phân tích dữ liệu là giai đoạn quan trọng nhất trong luồng dữ liệu của Big Data với mục đích rút trích những dữ liệu có ích, cung cấp các đề xuất và các quyết định. Tùy từng lãnh vực khác nhau mà việc phân tích dữ liệu sẽ mang lại những giá trị tiềm năng khác nhau [5]. Tuy nhiên, phân tích dữ liệu là 1 một lãnh vực rất rộng lớn, thường xuyên thay đổi và vô cùng phức tạp. Nên bài viết này chỉ tập trung vào các phương pháp, kiến trúc và các công cụ để phân tích Big Data.

i. Những phương pháp phân tích chung

Mặc dù các kiểu dữ liệu, mục đích và ứng dụng khác nhau, nhưng một số phương pháp phân tích chung vẫn hữu ích cho các loại khác nhau. Dưới đây là 3 loại phân tích phổ biến hay dùng hiện nay:

- *Dữ liệu trực quan (Data Visualization)*: là phương pháp nhằm truyền đạt thông tin rõ ràng và hiệu quả thông qua các phương tiện đồ họa. Hiển thị trực quan cho Big Data là một lãnh vực nghiên cứu đang được quan tâm hiện nay [3, 5]
- *Phân tích thống kê*: là dựa trên lý thuyết thống kê, mà là một nhánh của toán học ứng dụng. Phân tích thống kê có thể phục vụ cho 2 mục đích của Big Data: mô tả và suy luận.
- *Khai phá dữ liệu*: là quá trình tính toán để phát hiện các mô hình trong Big Data. Tại hội nghị quốc tế IEEE 2006 vê khai phá dữ liệu, 10 thuật

[1] Checkpoint: ám chỉ kỹ thuật lưu 1 snapshot trạng thái của ứng dụng, có thể phục hồi lại từ snapshot này trong trường hợp thất bại

toán thường được xử dụng nhất là: C4.5, K-means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART [18].

Hiện tại có khá nhiều công cụ hỗ trợ các phương pháp trên, cả những phần mềm chuyên nghiệp, nghiệp dư, thương mại và mở nguồn mở. 5 công cụ hàng đầu được sử dụng rộng rãi hiện nay phải kể đến[8]: Rapid-I RapidMiner/RapidAnalytics (39.2%), R (37.4%), Excel (28%), Weka / Pentaho (14.3%), Python (13.3%)

ii. Những phương pháp phân tích mới trong Big Data

Do sự phát triển mạnh mẽ của Internet và các thiết bị công nghệ cao, đặc biệt trong các lãnh vực kinh doanh, mạng và khoa học, đã đẩy mạnh việc nghiên cứu các phương pháp phân tích Big Data mới nhằm phục vụ cho việc khai thác những giá trị tiềm ẩn trong các lãnh vực trên. Bảng 10 mô tả về những phương pháp phân tích dữ liệu phổ biến hiện nay trong Big Data [12]:

Lãnh vực phân tích	**Nguồn dữ liệu**	**Tính chất**	**Giải pháp**
Phân tích cấu trúc dữ liệu	• Giao dịch của khách hàng • Dữ liệu thí nghiệm khoa học	• Có cấu trúc • Khối lượng ít và có tính thời gian thực	• Khai phá dữ liệu • Phân tích thống kê
Phân tích văn bản	• Log • Email • Tài liệu công ty • Quy tắc và quy định của chính phủ • Nội dung trang Web • Thông tin phản hồi và ý kiến	• Không cấu trúc • Văn bản phong phú • Theo ngữ cảnh • Có ngữ nghĩa • Phụ thuộc ngôn ngữ	• Trình bày tài liệu • NLP (xử lý ngôn ngữ tự nhiên) • Chiết lọc thông tin • Mô hình chủ đề • Tóm tắt • Phân loại • Phân cụm • Hỏi và trả lời • Khai thác quan điểm
Phân tích Web	• Nhiều loại trang Web	• Tích hợp văn bản và liên kết • Có các biểu tượng • Siêu dữ liệu	• Khai thác nội dung • Khai thác cấu trúc • Khai thác sử dụng
Phân tích đa phương tiện	• Phim trường • Người dùng • Thiết bị giám sát	• Hình ảnh, âm thanh video • Lớn • Có dữ liệu dư thừa • Có ngữ nghĩa theo thời gian	• Tóm tắt • Chú thích • Lập chỉ mục và tìm kiếm • Khuyến nghị • Phát hiện sự kiện

Phân tích mạng	• Mạng xã hội	• Nội dung phong phú • Các mối quan hệ xã hội • Nhiễu và có tính dư thừa • Tiến hóa nhanh	• Phát hiện cộng đồng • Tiến hóa mạng • Phân tích ảnh hưởng • Tìm kiếm từ khóa • Phân loại • Phân cụm • Học tập chuyển giao
Phân tích di động	• Ứng dụng di động • Cảm biến • RFID	• Dựa trên vị trí • Cá nhân • Thông tin bị phân đoạn	• Giám sát • Khai thác vị trí

Bảng 10: Những phương pháp phân tích dữ liệu trong Big Data

5.2 Nền tảng Hadoop

Hadoop đã được tạo ra vào năm 2005 bởi Doug Cutting và Mike Cafarella để giải quyết các vấn đề của Big Data. Đến năm 2011, Hadoop được sử dụng rộng rãi trong các trong ty lớn. Hơn 50% các công ty trong nhóm Fortune 50 đã sử dụng Hadoop. Ventana Research đã công bố những kết quả điều tra khá ấn tượng về việc sử dụng Hadoop trong các doanh nghiệp [9]: khoảng 63% các tổ chức sử dụng Hadoop để quản lý Big Data không cấu trúc; 94% người sử dụng Hadoop để phân tích Big Data, điều mà trước đây không thể thực hiện; 88% phân tích dữ liệu chi tiết hơn; trong khi 82% các doanh nghiệp có thể lưu trữ nhiều dữ liệu hơn.

Hadoop là một framework mã nguồn mở hỗ trợ lưu trữ và xử lý Big Data với các cấu trúc khác nhau (kể cả không cấu trúc) trên những máy chủ bình thường. Hadoop có nhiều lợi thế so với các framework khác:

- *Khả năng mở rộng*: cho phép thay đổi số lượng phần cứng mà không cần thay đổi định dạng dữ liệu hay khởi động lại hệ thống
- *Hiệu quả chi phí*: hỗ trợ lưu trữ và xử lý song song trên những máy chủ bình thường
- *Linh hoạt*: hỗ trợ bất kỳ loại dữ liệu từ bất kỳ nguồn nào
- *Chịu lỗi*: thiếu dữ liệu và phân tích thất bại là hiện tượng thường gặp trong phân tích tích Big Data. Hadoop có thể phục hồi và phát hiện nguyên nhân thất bại do tắc nghẽn mạng

Hadoop gồm nhiều module kết hợp với nhau hỗ trợ tất cả các giai đoạn trong luồng Big Data từ giai đoạn thu thập đến phân tích và quản lý dữ liệu.

Giai đoạn	Module	Mô tả
Thu thập dữ liệu	Flume	Thu thập, tập hợp và chuyển 1 lượng lớn dữ liệu từ các nguồn khác nhau về trung tâm lưu trữ
	Sqoop	Cho phép dễ dàng nhập và xuất dữ liệu giữa Hadoop và các kho dữ liệu có cấu trúc

Lưu trữ dữ liệu	HDFS	Hệ thống file phân phối có thể chạy trên những máy chủ bình thường, dựa trên thiết kế của GFS. Gồm 1 NameNode để quản lý file metadata và nhiều DataNode để lưu trữ dữ liệu thực tế. Một file được chia làm nhiều khối và các khối sẽ lưu trong các DataNode
	Hbase	CSDL hướng cột dựa trên Bigtable của Google
Tính toán	MapReduce	Là cốt lõi tính toán để phân tích Big Data. MapReduce framework sẽ gồm 1 master và 1 slave trên mỗi node. Master có trách nhiệm lập kế hoạch cho những slave, theo dõi và thực hiện lại các nhiệm vụ thất bại. Các slave thực hiện các nhiệm vụ theo chỉ dẫn của của master. Gồm 2 chức năng chính: map và reduce
Phân tích dữ liệu	Pig Latin	Ngôn ngữ cho xử lý dữ liệu
	Hive	Tổng hợp dữ liệu và truy vấn adhoc
	Mahout	Thư viện khai phá dữ liệu và máy học, gồm 4 nhóm: lọc tập hợp, gom cụm, phân loại, khai phá mô hình theo hướng song song
Quản lý	Zokeeper	Là 1 trung tâm dịch vụ cho việc bảo trì cấu hình, đặt tên, đồng bộ phân phối và cung cấp các dịch vụ theo nhóm
	Chukwa	Chịu trách nhiệm theo dõi tình trạng hệ thống và có thể hiển thị, giám sát và phân tích các dữ liệu thu thập được

Bảng 11: Những Module chính trong Hadoop

Với những lợi thế trên, Hadoop đã được sử dụng ở nhiều dự án và mỗi công ty sử dụng cho những nhu cầu riêng của mình. Như Yahoo đã chạy Hadoop trên 42.000 máy chủ tại 4 trung tâm dữ liệu vào tháng 7, 2012 để hỗ trợ chức năng lọc thư rác và tìm kiếm. Hay Facebook dùng Hadoop để lưu trữ và xử lý 100 PB cả dữ liệu có cấu trúc và phi cấu trúc. Bảng 12 mô tả việc việc sử dụng Hadoop trong các công ty hàng đầu và mục đích của họ:

Chức năng	**Được sử dụng bởi**
Tìm kiếm	Yahoo, Amazon, Zvents, Facebook, Yahoo, ContexWeb.Joost, Last.fm
Xử lý Log	
Phân tích ảnh và Video	NewYorkTimes,Eyelike
Kho dữ liệu	Facebook, AOL
Khuyến nghị	Facebook

Bảng 12: Sử dụng Hadoop

Hadoop được thiết kế cho các ứng dụng loại batch. Trong nhiều ứng dụng thời gian thực, Storm là ứng viên thích hợp cho cơ chế xử lý luồng dữ liệu liên tục. Storm có thể được sử dụng để phân tích thời gian thực, tính toán liên tục… Gần đây Twitter đang phát triển 1 dự án mã nguồn mở của họ là Summingbird [32] để tích hợp Hadoop và Storm.

PHẦN 6 NHỮNG THÁCH THỨC CẦN GIẢI QUYẾT

6.1 Về kỹ thuật, công nghệ

Việc phân tích Big Data đang đối mặt với nhiều thách thức, nhưng các nghiên cứu vẫn đang trong giai đoạn đầu. Những nỗ lực nghiên cứu tiếp theo là cần thiết để nâng cao hiệu quả trong việc hiển thị, lưu trữ và phân tích dữ liệu:

Truyền dữ liệu: Như đã thảo luận, truyền dữ liệu lớn thường phải gánh chịu chi phí cao, đây là nút cổ chai của việc tính toán Big Data. Tuy nhiên, truyền dữ liệu là không thể tránh khỏi trong các ứng dụng Big Data. Nâng cao hiệu quả truyền dữ liệu lớn là một yếu tố quan trọng để nâng cao tính toán Big Data.

Tốc độ xử lý trong các yêu cầu thời gian thực: khi dữ liệu số lượng dữ liệu tăng nhanh chóng, gây ra 1 thách thức rất lớn đối với các ứng dụng thời gian thực. Nên việc tìm các phương pháp hiệu quả trong suốt luồng dữ liệu là cần thiết để đáp ứng yêu cầu về thời gian thực.

Nền tảng Big Data: Mặc dù Hadoop đã trở thành một trụ cột trong nền tảng phân tích Big Data, nó vẫn còn trong giai đoạn phát triển, so với CSDL quan hệ (hơn 30 năm phát triển. Đầu tiên, Hadoop phải tích hợp với thời gian thực cho việc thu thập và truyền Big Data, và cung cấp xử lý nhanh hơn dựa trên các mô hình xử lý hàng loạt. Thứ hai, Hadoop nên cung cấp một giao diện lập trình ngắn gọn, và ẩn những tiến trình xử lý phức tạp bên dưới. Thứ ba, trong những hệ thống Hadoop lớn, số lượng máy chủ lên hàng ngàn thâm chí hàng trăm ngàn, có nghĩa là năng lượng tiêu thụ đáng kể. Nên Hadoop nên có cơ chế sử dụng năng lượng hiệu quả. Có khá nhiều nghiên cứu để cải thiện cũng như khắc phục những điểm yếu của Hadoop được thảo luận tại [33, 34]

Bảo mật dữ liệu và quyền riêng tư: là vấn đề rất quan trọng. Một số ví dụ trong thực tế cho thấy không chỉ thông tin cá nhân người tiêu dùng, thông tin mật của các tổ chức mà ngay cả các bí mật an ninh quốc gia cũng có thể bị xâm phạm. Do vậy, giải quyết các vấn đề an ninh dữ liệu bằng các công cụ kỹ thuật và các chính sách trở nên vô cùng cấp bách. Các nền tảng Big Data nên cân bằng tốt giữa việc truy cập dữ liệu và xử lý dữ liệu [5].

6.2 Về tổ chức:

Thiếu hụt nguồn lực có kiến thức sâu về thống kê, công nghệ thông tin, cũng như nguồn nhân lực có kỹ năng phân tích và quản lý cho các dự án Big Data. Bởi vì phần lớn nguồn dữ liệu có giá trị nằm ngoài phạm vi của tổ chức, các nhà quản lý phải đối mặt với khó khăn trong việc đặt đúng câu hỏi và sử dụng các kết quả phân tích dữ liệu một cách hiệu quả. Do đó, nếu đầu tư lớn vào Big Data, số lượng dữ liệu thu thập được nhiều, nhưng không được tận dụng để đưa ra những giá trị thông tin tiềm ẩn thì sẽ dẫn đến sự lãng phí tài nguyên [8]. Theo tổ chức McKinsey

Global, đến năm 2018, Hoa Kỳ có thể phải đối mặt sự thiếu hụt từ 140,000 đến 190,000 nhân lực có kỹ năng phân tích. [5]

PHẦN 7 KẾT LUẬN

Big Data đóng vai trò quan trọng trong việc mang lại những giá trị to lớn, không chỉ cho các tổ chức doanh nghiệp mà còn cho nền kinh tế quốc gia và cho các công dân trong nền kinh tế đó. Thông tin được thu thập ngày càng minh bạch, chi tiết, chính xác, giúp các nhà lãnh đạo có thể ra những quyết định đúng và hợp lý hơn, giảm thiểu các rủi ro có thể xảy ra, giúp cho các cá nhân được trải nghiệm các dịch vụ mà các tổ chức và chính phủ mang lại một cách tốt hơn. Tuy nhiên, đây vẫn là lĩnh vực còn rất mới, đặt ra nhiều vấn đề và thách thức mà các tổ chức và các nhà nghiên cứu cần giải quyết.

Tài liệu tham khảo

[1] Mohanty, S., et al., *Big Data Imperatives*. 2013: Apress.

[2] Chui, M., M. Löffler, and R. Roberts. *The Internet of Things*. 2010 [cited 2014 7/11]; Available from: http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_internet_of_things.

[3] Gantz, J. and D. Reinsel, *Extracting value from chaos.* IDC iview, 2011: p. 1-12.

[4] O'Leary, D.E., *Exploiting Big Data from Mobile Device Sensor-Based Apps: Challenges and Benefits.* MIS Quarterly Executive, 2013. 12(4): p. 179-187.

[5] Manyika, J., et al., *Big data: The next frontier for innovation, competition, and productivity.* 2011.

[6] Clayton, R., *CFOs Take Notice Big Data May Be Your New Best Friend.* Financial Executive, 2013. 29(10): p. 22-25.

[7] Rivera, J. and R.v.d. Meulen. *Gartner Survey Reveals That 73 Percent of Organizations Have Invested or Plan to Invest in Big Data in the Next Two Years*. 2014 [cited 7 11]; Available from: http://www.gartner.com/newsroom/id/2848718.

[8] Luan, D. *Big Data là gì và người ta khai thác, ứng dụng nó vào cuộc sống như thế nào?* 2013 [cited 2014 11/10]; Available from: https://www.tinhte.vn/threads/big-data-la-gi-va-nguoi-ta-khai-thac-ung-dung-no-vao-cuoc-song-nhu-the-nao.2210939/.

[9] Tuan, D.Q. *Facebook xử lí hơn 500 TB dữ liệu mỗi ngày*. 2012 [cited 2014 10/10]; Available from: http://www.thongtincongnghe.com/article/37841.

[10] Hsinchun, C., R.H.L. Chiang, and V.C. Storey, *BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT.* MIS Quarterly, 2012. 36(4): p. 1165-1188.

[11] Quân, Đ. *Khai thác Big Data trong lĩnh vực thông tin - viễn thông*. 2013 [cited 2014 10/10]; Available from: http://www.pcworld.com.vn/articles/kinh-doanh/giai-phap/2013/09/1234258/khai-thac-big-data-trong-linh-vuc-thong-tin-vien-thong/.

[12] Han, H., et al., *Toward Scalable Systems for Big Data Analytics: A Technology Tutorial.* Access, IEEE, 2014. 2: p. 652-687.

[13] Chen, M., S. Mao, and Y. Liu, *Big Data: A Survey.* Mobile Networks and Applications, 2014. 19(2): p. 171-209.

[14] Mangalindan, J., *Amazon's recommendation secret.* Cable News Network. A Time Warner Company. Online referred to, 2012. 4: p. 2013.

[15] Cook, S., et al., *Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic.* PloS one, 2011. 6(8): p. e23610.

[16] HP. *Ứng dụng của Big Data trong lĩnh vực y tế*. 2014 [cited 2014 7/11]; Available from: http://vht.com.vn/ung-dung-cua-big-data-trong-linh-vuc-y-te/.

[17] Lampitt, A., *'The real story of how Big Data analytics helped Obama win'.* Info World, 2013. 14.
[18] Goll, D. *Santa Cruz firm PredPol helps predict, prevent property crimes*. 2012 [cited 2014 10/11]; Available from: http://www.bizjournals.com/sanjose/news/2012/06/11/new-santa-cruz-company-to-plead-case.html?page=all.
[19] Meulen, R.v.d. *Gartner Says Business Intelligence/Analytics Is Top Area for CFO Technology Investment Through 2014*. 2013 [cited 2014 10/11]; Available from: http://www.gartner.com/newsroom/id/2488616.
[20] Lewis, H. *Big data - Time for a lean approach in financial services*. 2012 [cited 2014 10/10]; Available from: http://www2.deloitte.com/content/dam/Deloitte/global/Documents/Financial-Services/dttl-fsi-uk-mi-da-big-data.pdf.
[21] IBM. *What is big data?* [cited 2014 7/10]; Available from: http://www-01.ibm.com/software/data/Big Data/what-is-big-data.html.
[22] Khan, N., et al., *Big data: survey, technologies, opportunities, and challenges.* ScientificWorldJournal, 2014. 2014: p. 712826.
[23] Cukier, K., *Data, data everywhere: A special report on managing information*. 2010: Economist Newspaper.
[24] Wikibon, *A Comprehensive List of Big Data Statistics [Online]*. 2013.
[25] Lenzerini, M. *Data integration: A theoretical perspective*. in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2002. ACM.
[26] Cafarella, M.J., A. Halevy, and N. Khoussainova, *Data integration for the relational web.* Proceedings of the VLDB Endowment, 2009. 2(1): p. 1090-1101.
[27] Maletic, J.I. and A. Marcus. *Data Cleansing: Beyond Integrity Analysis*. in *IQ*. 2000. Citeseer.
[28] Kohavi, R., et al., *Lessons and challenges from mining retail e-commerce data.* Machine Learning, 2004. 57(1-2): p. 83-113.
[29] Chang, F., et al., *Bigtable: A distributed storage system for structured data.* ACM Transactions on Computer Systems (TOCS), 2008. 26(2): p. 4.
[30] Xie, X.L., Z.X. Sun, and Z. Xiong, *The research of the key tech-application based on HBase for e-government cloud of minority areas.* Applied Mechanics and Materials, 2014. 530: p. 827-831.
[31] Manyam, G., et al., *Relax with CouchDB—Into the non-relational DBMS era of bioinformatics.* Genomics, 2012. 100(1): p. 1-7.
[32] Boykin, O., et al., *Summingbird: A Framework for Integrating Batch and Online MapReduce Computations.* Proceedings of the VLDB Endowment, 2014. 7(13).
[33] Lee, K.-H., et al., *Parallel data processing with MapReduce: a survey.* AcM sIGMoD Record, 2012. 40(4): p. 11-20.
[34] Rao, B.T. and L. Reddy, *Survey on improved scheduling in hadoop mapreduce in cloud environments.* arXiv preprint arXiv:1207.0780, 2012.

Thông tin tác giả

Lê Thị Quỳnh Nga,
Khoa Hệ Thống Thông Tin Kinh Doanh – ĐH Kinh Tế HCM,
Email: nga.lethiquynh@ueh.edu.vn

Nguyễn Mạnh Tuấn,
Khoa Hệ Thống Thông Tin Kinh Doanh – ĐH Kinh Tế HCM,
Email: tuannm@ueh.edu.vn