TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

**VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

-----🙞🙜🕮🙞🙜----- ****

**Xây dựng hồ sơ công nghệ cho hê sinh thái BigData: thu thập, lưu trữ, tính toán, chia sẻ, quản trị và các mô hình kinh doanh dữ liệu**

|  |  |  |
| --- | --- | --- |
| Giáo viên hướng dẫn: ***PGS. TS. Phạm Văn Hải*** | | |
| Thực hiện:  MSHV: | **Nguyễn Chí Hiếu**  **20151331** |  |

**Hà Nội 10/2019**

# DANH MỤC HÌNH ẢNH

[Hình 1: Chuỗi giá trị của Big data 11](file:///C:\Users\chihi\Downloads\BT%20ca%20nhan%20Tich%20hop%20HTTT%20HieuNV%2020082019.docx#_Toc21037446)

[Hình 2: Sự phát triển của khoa học dữ liệu 16](file:///C:\Users\chihi\Downloads\BT%20ca%20nhan%20Tich%20hop%20HTTT%20HieuNV%2020082019.docx#_Toc21037447)

# THUẬT NGỮ

|  |  |
| --- | --- |
| **Ký hiệu viết tắt** | **Ý nghĩa** |
| IoT | Internet of things |
| Big data | Dữ liệu lớn |

# MỤC LỤC

[DANH MỤC HÌNH ẢNH 2](#_Toc23345491)

[THUẬT NGỮ 3](#_Toc23345492)

[MỤC LỤC 4](#_Toc23345493)

[MỞ ĐẦU 5](#_Toc23345494)

[CHƯƠNG 1 Tổng quan về hệ sinh thái Big Data 7](#_Toc23345495)

[1.1 Big data: định nghĩa, đặc trưng, và nguồn dữ liệu 7](#_Toc23345496)

[1.2 Các bước trong quá trình khai phá Big data 11](#_Toc23345497)

[1.3 Các ông cụ thu thập, tiền xử lý và phân tích dữ liệu lớn 15](#_Toc23345498)

[1.4 Các công cụ xử lý dữ liệu lớn: 18](#_Toc23345499)

[CHƯƠNG 2 Mô hình kinh doanh dữ liệu lớn 20](#_Toc23345500)

[2.1 *Mô hình kinh doanh dữ liệu* 21](#_Toc23345501)

[2.2 *Các lớp trong mô hình kinh doanh dữ liệu lớn:* 25](#_Toc23345502)

[KẾT LUẬN 26](#_Toc23345503)

[TÀI LIỆU THAM KHẢO 30](#_Toc23345504)

# MỞ ĐẦU

Hiện nay, với sự số hóa của hầu hết các quy trình nghiệp vụ, sự xuất hiện các nền tảng mạng xã hội, sự triển khai đa dạng các loại cảm biến khác nhau, áp dụng các thiết bị kỹ thuật số cầm tay, thiết bị mang trên người và sự bùng nổ trong việc sử dụng Internet, một lượng dữ liệu khổng lồ đang được tạo ra một cách liên tục. Không ai có thể phủ nhận rằng Internet đã thay đổi cách thức hoạt động của các doanh nghiệp, chính phủ, thay đổi phương pháp giáo dục và lối sống của mọi người trên khắp thế giới. Trong giai đoạn biến đổi này, dữ liệu đóng vai trò vô cùng quan trọng và chứa đựng nhiều thông tin hơn bao giờ hết. Tuy nhiên tốc độ sản sinh dữ liệu là rất cao và các loại dữ liệu được tạo ra đã vượt qua khả năng của các kỹ thuật lưu trữ hiện có.

Trong hai thập kỷ qua, có một sự tăng trưởng to lớn về dữ liệu. Xu hướng này có thể được quan sát trong hầu hết các lĩnh vực. Theo báo cáo của công ty nghiên cứu International Data Corporation (IDC), từ năm 2012 đến 2020, lượng thông tin trong vũ trụ kỹ thuật số sẽ tăng thêm 35 nghìn tỷ gigabyte (1 gigabyte tương đương với 40 tủ hồ sơ chứa văn bản, hoặc hai đĩa nhạc CD). Số lượng đó ngang bằng với số lượng ngôi sao trong vũ trụ! (Forsyth, 2012).

Vào giữa những năm 2000, sự xuất hiện của phương tiện truyền thông xã hội, điện toán đám mây và sức mạnh xử lý (thông qua bộ xử lý đa lõi và GPU) đã thúc đẩy sự phát triển dữ liệu lớn (Manovich, 2011; Agneeswaran, 2012). Tính đến tháng 12 năm 2015, Facebook có trung bình 1,04 tỷ người dùng hoạt động hàng ngày, 934 triệu người dùng hoạt động hàng ngày trên thiết bị di động, có sẵn 70 ngôn ngữ, 125 tỷ kết nối bạn bè, 205 tỷ ảnh được tải lên mỗi ngày, 30 tỷ nội dung, 2,7 tỷ lượt thích và bình luận đang được đăng và 130 số bạn bè trung bình trên mỗi người dùng Facebook (Facebook, 2015). Điều này đã mở ra những hướng mới để nghiên cứu về các xu thế văn hóa và xã hội.

Việc khai thác và ứng dụng dữ lớn có thể giúp các tổ chức đưa ra quyết định sáng suốt và cung cấp lợi thế cạnh tranh. Trước đó, các tổ chức đã sử dụng các hệ thống xử lý giao dịch vốn đã sử dụng hệ thống quản lý cơ sở dữ liệu quan hệ (RDBMS) và các kỹ thuật phân tích dữ liệu đơn giản như ngôn ngữ truy vấn có cấu trúc (SQL) cho các hoạt động hàng ngày của họ giúp họ đưa ra quyết định và lập kế hoạch. Tuy nhiên, do sự gia tăng kích thước của dữ liệu, đặc biệt là dạng dữ liệu không có cấu trúc (ví dụ: đánh giá của khách hàng về các trang hoặc tweet trên Facebook của họ), gần như không thể xử lý các dữ liệu này bằng các kỹ thuật lưu trữ hiện có và các truy vấn đơn giản.

“Những ai có khả năng sử dụng Big Data sẽ có một lợi thế lớn thông qua việc không chỉ biết được ý kiến của công chúng mà còn có thể phát hiện ra những động thái trong chính tổ chức của họ, để cải tiến quy trình và có thông tin tốt hơn trong quá trình đưa ra các quyết định”, nhận định từ tờ báo uy tín của Pháp ParisTech Review. Chính vì vậy, chúng ta cần trang bị thông tin để làm sao tái cấu trúc hệ thống, thiết lập cơ sở hạ tầng thông tin, thu thập, xử lý, để dữ liệu lớn phát huy giá trị to lớn của nó”, Chính vì thế, dựa trên vai trò quan trọng của Big Data cũng như sự áp dụng còn chưa phổ biến và hạn chế của nguồn lực này trong các doanh nghiệp trên thế giới nói chung và VN nói riêng mà cần phải có thật nhiều hơn nữa những nghiên cứu có thể đưa ra những cái nhìn tổng quan về Big Data cũng như những minh chứng rõ nét về tầm quan trọng của nó đối với các doanh nghiệp.

# CHƯƠNG 1 Tổng quan về hệ sinh thái Big Data

## 1.1 Big data: định nghĩa, đặc trưng, và nguồn dữ liệu

**Định nghĩa**

Gần đây, cụm từ "dữ liệu lớn" (Big data) đã trở nên thông dụng. Nó đang được sử dụng bởi hầu hết mọi người bao gồm các học giả và chuyên gia trong ngành. Có nhiều định nghĩa khác nhau được nêu ra trong các tài liệu. Nhưng khái niệm dữ liệu lớn bắt nguồn từ năm 2001, trong đó những thách thức về tăng dữ liệu đã được giải quyết với mô hình 3V của Laney (2001). 3V, còn được gọi là quy mô của dữ liệu lớn, đại diện cho kích thước, sự đa dạng và tốc độ tạo ra của dữ liệu ngày càng tăng (Assunção et al., 2015). Mô hình ban đầu không được sử dụng cho dữ liệu lớn nhưng sau đó đã được các doanh nghiệp khác nhau bao gồm cả Microsoft và IBM sử dụng để định nghĩa khái niệm này (Meijer, 2011).

Vào năm 2010, Apache Hadoop đã định nghĩa dữ liệu lớn là các bộ dữ liệu, không thể lưu giữ, quản lý và xử lý bởi các máy tính thông thường trong phạm vi chấp nhận được (tr.173, Chen và cộng sự, 2014). Theo đó, vào năm 2011, Viện toàn cầu McKinsey đã định nghĩa dữ liệu lớn là "bộ dữ liệu có kích thước vượt quá khả năng của các công cụ phần mềm cơ sở dữ liệu thông thường để nắm bắt, lưu trữ, quản lý và phân tích" (tr.1 Manyika et al., 2011). International Data Corporation (IDC) định nghĩa các công nghệ dữ liệu lớn là một thế hệ công nghệ và kiến ​​trúc mới, được thiết kế để khai thác giá trị kinh tế từ khối lượng rất lớn của nhiều loại dữ liệu, bằng cách cho phép nắm bắt, khám phá và phân tích với tốc độ cao (trang 6, Gantz và Reinsel, 2011).

Các nhà học thuật định nghĩa dữ liệu lớn là kích thước khổng lồ của dữ liệu phi cấu trúc được tạo ra bởi nhóm ứng dụng không đồng nhất hiệu năng cao trải dài từ mạng xã hội đến các ứng dụng máy tính khoa học. Các bộ dữ liệu có phạm vi từ vài trăm gigabyte đến zetabyte vượt quá khả năng của các công cụ quản lý dữ liệu hiện có để nắm bắt, lưu trữ, quản lý và phân tích (Cuzzocrea et al., 2011; Qin et al., 2012; Agneeswaran, 2012).

Mặc dù dữ liệu lớn đã được xác định trong các hình thức khác nhau nhưng không có định nghĩa cụ thể. Vài người đã định nghĩa những gì nó làm trong khi rất ít người tập trung vào những gì nó là. Định nghĩa của dữ liệu lớn trên cơ sở 3V là tương đối. Điều gì được định nghĩa, vì dữ liệu lớn có thể không giống nhau vào ngày mai? Chẳng hạn, trong tương lai, với những tiến bộ trong công nghệ lưu trữ, dữ liệu được coi là dữ liệu lớn ngày nay có thể bị bắt giữ. Ngoài việc xác định dữ liệu lớn, cần phải hiểu cách sử dụng tốt nhất dữ liệu này để có được thông tin có giá trị cho việc ra quyết định.

**Các đặc trưng của Big Data**

Ban đầu, dữ liệu lớn được đặc trưng bởi các quy mô sau, thường được gọi là mô hình 3V:

1. **Kích thước (volume):** là độ lớn của dữ liệu đang được tạo và thu thập. Nó đang tăng với tốc độ nhanh hơn từ terabyte đến petabyte (1024 terabyte) (Zikopoulos và cộng sự, 2012; Singh và Singh, 2012). Với sự gia tăng về dung lượng lưu trữ, những gì không thể nắm bắt và lưu trữ bây giờ sẽ có thể trong tương lai. Việc phân loại dữ liệu lớn trên cơ sở khối lượng có liên quan đến loại dữ liệu được tạo và thời gian. Ngoài ra, loại dữ liệu (variety) cũng định nghĩa cho dữ liệu lớn. Ví dụ, hai loại dữ liệu, văn bản và video có cùng khối lượng có thể yêu cầu các công nghệ quản lý dữ liệu khác nhau (Gandomi và Haider, 2015).
2. **Tốc độ (velocity):** là tốc độ tạo dữ liệu. Phân tích dữ liệu truyền thống dựa trên các cập nhật định kỳ - hàng ngày, hàng tuần hoặc hàng tháng. Với tốc độ tạo dữ liệu ngày càng tăng, dữ liệu lớn cần được xử lý và phân tích theo thời gian thực hoặc gần thời gian thực để đưa ra quyết định sáng suốt. Vai trò của thời gian rất quan trọng ở đây (Singh và Singh, 2012; Gandomi và Haider, 2015). Một vài lĩnh vực bao gồm bán lẻ, viễn thông và tài chính tạo ra dữ liệu tốc độ cao. Dữ liệu được tạo thông qua các ứng dụng di động, ví dụ, nhân khẩu học, vị trí địa lý và lịch sử giao dịch, có thể được sử dụng trong thời gian thực để cung cấp dịch vụ được cá nhân hóa cho khách hàng. Điều này sẽ giúp giữ chân khách hàng cũng như tăng chất lượng dịch vụ.
3. **Chủng loại (variety):** đề cập đến các loại dữ liệu khác nhau đang được tạo và thu thập. Chúng vượt ra ngoài dữ liệu có cấu trúc và thuộc các loại dữ liệu bán cấu trúc và không cấu trúc (Zikopoulos và cộng sự, 2012; Singh và Singh, 2012; Gandomi và Haider, 2015). Dữ liệu có thể được tổ chức bằng mô hình dữ liệu được xác định trước được gọi là dữ liệu có cấu trúc. Dữ liệu dạng bảng trong cơ sở dữ liệu quan hệ và Excel là ví dụ về dữ liệu có cấu trúc và chúng chỉ chiếm 5% của tất cả dữ liệu hiện có (Cukier, 2010). Dữ liệu phi cấu trúc không thể được tổ chức bằng các mô hình được xác định trước này và các ví dụ bao gồm video, văn bản và âm thanh. Dữ liệu bán cấu trúc nằm giữa các loại dữ liệu có cấu trúc và không cấu trúc. Ngôn ngữ đánh dấu mở rộng (XML) thuộc danh mục này.

Sau đó, một vài tính chất khác của dữ liệu lớn đã được thêm vào, theo như liệt kê dưới đây:

1. **Độ chính xác (veracity):** Được đặt ra bởi IBM, tính chính xác đề cập đến sự không đáng tin cậy liên quan đến các nguồn dữ liệu (Gandomi và Haider, 2015). Chẳng hạn, phân tích tình cảm sử dụng dữ liệu truyền thông xã hội (Twitter, Facebook, v.v.) có thể không chắc chắn. Cần phân biệt dữ liệu đáng tin cậy với dữ liệu không chắc chắn và không chính xác và quản lý sự không chắc chắn liên quan đến dữ liệu.
2. **Sự thay đổi (variability):** Sự thay đổi và độ phức tạp đã được thêm vào dưới dạng kích thước bổ sung của SAS. Thông thường, sự không nhất quán trong tốc độ dữ liệu lớn dẫn đến sự thay đổi tốc độ dòng dữ liệu, được gọi là tính biến đổi (Gandomi và Haider, 2015). Dữ liệu được tạo từ nhiều nguồn khác nhau và có sự phức tạp ngày càng tăng trong việc quản lý dữ liệu từ dữ liệu giao dịch đến dữ liệu lớn. Dữ liệu được tạo từ các vị trí địa lý khác nhau có ngữ nghĩa khác nhau (Zikopoulos và cộng sự, 2012; Forsyth, 2012).
3. **Mật độ giá trị thấp (low-value density):** Dữ liệu ở dạng ban đầu của nó là không thể sử dụng. Dữ liệu được phân tích để khám phá giá trị rất cao (Sun và Heller, 2012). Ví dụ: nhật ký từ trang web không thể được sử dụng ở dạng ban đầu để có được giá trị kinh doanh. Nó phải được phân tích để dự đoán hành vi của khách hàng.

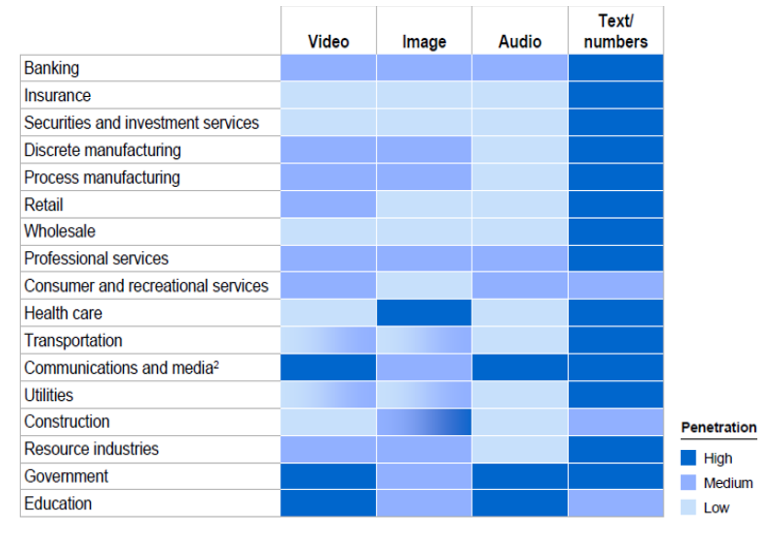
**Các nguồn dữ liệu lớn**

Ở những phần trước ta đã đề cập sơ bộ về dữ liệu lớn và các quy mô, tính chất của chúng. Phần này sẽ nêu tóm tắt về các nguồn dữ liệu lớn khác nhau. Số hóa nội dung theo các ngành là nguồn dữ liệu mới (Villars et al., 2011). Những tiến bộ trong công nghệ cũng dẫn đến tốc độ tạo dữ liệu rất lớn. Ví dụ, một trong những khảo sát lớn nhất trong thiên văn học, khảo sát bầu trời kỹ thuật số Sloan (SDSS) đã ghi lại tổng cộng 25TB dữ liệu trong các khảo sát đầu tiên (2000-2005) và thứ hai (2005-2008). Với những tiến bộ trong độ phân giải của kính thiên văn, lượng dữ liệu được thu thập vào cuối cuộc khảo sát thứ ba của họ (2008-14) là 100 TB. Sử dụng các thiết bị thông minh của người dùng cúng là một nguồn dữ liệu lớn. Đồng hồ thông minh trong ngành năng lượng ghi lại phép đo sử dụng điện cứ sau 15 phút so với số đọc hàng tháng trước đó.

Giao dịch kinh doanh, từ số liệu giá cả sản phẩm, thanh toán, dữ liệu chế tạo và phân bố... như số sản phẩm Amazon.com bán trong Quý 3 năm 2011 có giá trị 10 tỷ USD, như dãy các nhà hàng Domino bán pizza trên toàn nước Mỹ đạt 1 triệu khách mỗi ngày... Các phương tiện truyền thông xã hội, như mỗi ngày trên toàn thế giới có 230 triệu mẩu tin trao đổi trên các twitters, có 2,7 tỷ ý kiến trao đổi trên các facebooks, và số video mỗi ngày đưa lên Youtube cần đến 86.400 giờ để xem hết.

Ngoài phương tiện truyền thông xã hội, Internet of Things (IoT), giờ đây, trở thành nguồn dữ liệu mới. Dữ liệu có thể được thu thập từ nông nghiệp, công nghiệp, chăm sóc y tế, vv của các thành phố thông minh được phát triển dựa trên IoT. Các máy móc thu nhận dữ liệu, các thiết bị công nghiệp, các cảm biến (sensors), các dụng cụ giám sát... như máy gia tốc hạt lớn của CERN (tổ chức nghiên cứu nguyên tử châu Âu) tạo ra 40 terabytes dữ liệu mỗi giây…

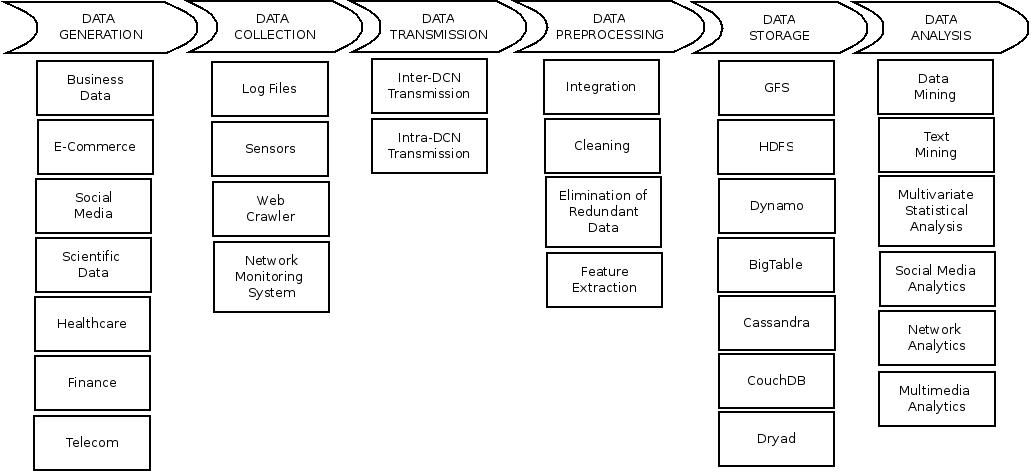
Bảng 1 tóm tắt các loại dữ liệu được tạo ra trong các lĩnh vực khác nhau.



Bảng 1: Các nguồn dữ liệu lớn

## 1.2 Các bước trong quá trình khai phá Big data

Các bước khai phá dữ liệu lớn, khái niệm được giới thiệu bởi Porter (1980), đề cập đến một tập hợp các hoạt động được thực hiện bởi một công ty để tăng thêm giá trị ở mỗi bước cung cấp sản phẩm / dịch vụ cho khách hàng của mình. Theo cách tương tự, chuỗi giá trị dữ liệu đề cập đến khung liên quan đến một tập hợp các hoạt động để tạo giá trị từ dữ liệu có sẵn. Nó có thể được chia thành bảy giai đoạn: tạo dữ liệu, thu thập dữ liệu, truyền dữ liệu, tiền xử lý dữ liệu, lưu trữ dữ liệu, phân tích dữ liệu và ra quyết định (Hình 2).



Hình 1: Các bước trong quá trình khai phá Big data

1. **Tạo dữ liệu:** Bước đầu tiên và quan trọng nhất của chuỗi giá trị dữ liệu lớn là tạo dữ liệu. Như đã thảo luận trong phần trước, dữ liệu được tạo từ nhiều nguồn khác nhau bao gồm dữ liệu từ Bản ghi chi tiết cuộc gọi (CDR), blog, Tweets và Trang Facebook.
2. **Thu thập dữ liệu:** Trong giai đoạn này, dữ liệu được lấy từ tất cả các nguồn dữ liệu có thể (Miller và Mork, 2013; Chen và cộng sự, 2014). Chẳng hạn, để dự đoán xu hướng khách hàng trong Viễn thông, dữ liệu có thể được lấy từ CDR và ý kiến / khiếu nại của khách hàng trên các Trang web Mạng xã hội như Twitter (dưới dạng tweet) và Facebook (ý kiến chia sẻ trên trang Facebook của công ty ). Các phương pháp được sử dụng phổ biến nhất là tệp nhật ký, cảm biến, trình thu thập dữ liệu web và phần mềm giám sát mạng (Chen et al., 2014).
3. **Truyền dữ liệu**: Sau khi dữ liệu được thu thập, nó được chuyển đến cơ sở hạ tầng lưu trữ và xử lý dữ liệu để xử lý và phân tích thêm. Nó có thể được thực hiện theo hai giai đoạn: Truyền mạng liên động (DCN) và truyền Intra-DCN. Truyền liên DCN liên quan đến việc chuyển dữ liệu từ nguồn dữ liệu sang trung tâm dữ liệu trong khi truyền sau giúp truyền trong trung tâm dữ liệu. Ngoài việc lưu trữ dữ liệu, trung tâm dữ liệu giúp thu thập, tổ chức và quản lý dữ liệu.
4. **Tiền xử lý dữ liệu:** Dữ liệu được thu thập từ nhiều nguồn dữ liệu khác nhau có thể là dư thừa, bị nhiễu và không nhất quán, do đó, trong giai đoạn này; dữ liệu được xử lý trước để cải thiện chất lượng dữ liệu cần thiết để phân tích. Điều này cũng giúp cải thiện tính chính xác của phân tích và giảm chi phí lưu trữ. Dữ liệu có thể được xử lý trước với sự trợ giúp của các bước sau: Điều này cũng giúp cải thiện tính chính xác của phân tích và giảm chi phí lưu trữ
   1. Tích hợp: Dữ liệu từ nhiều nguồn khác nhau được kết hợp để cung cấp một cái nhìn thống nhất về dữ liệu có sẵn. Liên kết dữ liệu và lưu trữ dữ liệu là hai phương pháp truyền thống thường được sử dụng. Kho dữ liệu thực hiện quy trình trích xuất (Extract), chuyển đổi (Transform) và tải (Load), gọi tắt là quy trình ETL. Trong quá trình trích xuất, dữ liệu được chọn, thu thập, xử lý và phân tích. Quá trình chuyển đổi dữ liệu được trích xuất sang định dạng chuẩn được gọi là quá trình chuyển đổi. Trong quá trình tải, dữ liệu được trích xuất và chuyển đổi được nhập vào cơ sở hạ tầng lưu trữ. Để tích hợp dữ liệu động, dữ liệu có thể được tổng hợp từ nhiều nguồn dữ liệu khác nhau bằng cơ sở dữ liệu ảo. Nó không chứa bất kỳ dữ liệu nào nhưng có thể thu được các chi tiết liên quan đến thông tin liên quan đến dữ liệu gốc hoặc siêu dữ liệu (Miller và Mork, 2013; Chen và cộng sự, 2014).
   2. Làm sạch: Dữ liệu được kiểm tra tính chính xác, đầy đủ và nhất quán. Trong quá trình này, dữ liệu có thể bị xóa và sửa đổi để cải thiện chất lượng dữ liệu. Quy trình chung bao gồm năm quy trình sau: loại lỗi được xác định và xác định, lỗi được xác định từ dữ liệu, lỗi được sửa, loại lỗi và ví dụ tương ứng được ghi lại và quy trình nhập dữ liệu có thể được sửa đổi để tránh lỗi trong tương lai (Maletic và Marcus , 2000).
   3. Loại bỏ dữ liệu dư thừa: Nhiều bộ dữ liệu có dữ liệu dư thừa hoặc lặp lại dữ liệu và được gọi là dự phòng dữ liệu. Điều này làm tăng chi phí lưu trữ, dẫn đến sự không thống nhất dữ liệu và ảnh hưởng đến chất lượng dữ liệu. Để khắc phục điều này, các phương pháp giảm dữ liệu khác nhau như lọc và nén dữ liệu, được sử dụng. Hạn chế của các kỹ thuật giảm dữ liệu này là chúng làm tăng chi phí tính toán. Do đó, phân tích lợi ích chi phí nên được thực hiện trước khi sử dụng các kỹ thuật giảm dữ liệu.
5. **Lưu trữ dữ liệu:** Các hệ thống lưu trữ dữ liệu lớn sẽ cung cấp không gian lưu trữ đáng tin cậy và quyền truy cập mạnh mẽ vào dữ liệu. Các hệ thống lưu trữ phân tán cho dữ liệu lớn nên xem xét các yếu tố như tính nhất quán (C), tính khả dụng (A) và dung sai phân vùng (P). Theo lý thuyết CAP được đề xuất bởi Brewer (2000), các hệ thống lưu trữ phân tán có thể đáp ứng đồng thời hai yêu cầu, đó là tính nhất quán và tính sẵn có hoặc tính khả dụng của phân vùng hoặc dung sai phân vùng nhưng không phải tất cả các yêu cầu đồng thời (Gilbert và Lynch, 2002 ). Nghiên cứu đáng kể vẫn đang diễn ra trong lĩnh vực cơ chế lưu trữ dữ liệu lớn. Sự tiến bộ về mặt này là Google File System (GFS), Dynamo, BigTable, Cassandra, CouchDB và Dryad.
6. **Phân tích dư liệu:** Sau khi dữ liệu được thu thập, chuyển đổi và lưu trữ, quy trình tiếp theo là khai thác dữ liệu hoặc phân tích dữ liệu, được liệt kê bằng các bước sau:
   1. Xác định số liệu: Dựa trên dữ liệu được thu thập và chuyển đổi, một bộ số liệu được xác định cho một vấn đề cụ thể. Ví dụ, để xác định một khách hàng tiềm năng sắp ra mắt, có thể xem xét một số lần anh ấy / cô ấy liên lạc (có thể thông qua một cuộc gọi thoại, tweet hoặc khiếu nại trên trang Facebook). (Miller và Mork, 2013).
   2. Chọn kiến trúc dựa trên kiểu phân tích: Dựa trên tính kịp thời của phân tích sẽ được thực hiện, kiến trúc phù hợp được chọn. Phân tích thời gian thực được sử dụng trong miền mà dữ liệu liên tục thay đổi và cần phải phân tích nhanh để thực hiện các hành động. Các tính toán dựa trên bộ nhớ và các hệ thống xử lý song song là các kiến trúc hiện có. Phát hiện gian lận trong lĩnh vực bán lẻ và gian lận viễn thông là những ví dụ về phân tích thời gian thực. Các ứng dụng không yêu cầu thời gian phản hồi cao được thực hiện bằng phân tích ngoại tuyến. Dữ liệu có thể được trích xuất, lưu trữ và phân tích tương đối muộn hơn. Kiến trúc thường được sử dụng là nền tảng Hadoop (Chen et al., 2014).
   3. Lựa chọn các thuật toán và công cụ phù hợp: Một trong những bước quan trọng nhất của phân tích dữ liệu là lựa chọn các kỹ thuật phù hợp để phân tích dữ liệu. Vài kỹ thuật phân tích dữ liệu truyền thống như phân tích cụm, phân tích hồi quy và thuật toán khai thác dữ liệu, vẫn giữ tốt cho phân tích dữ liệu lớn. Phân tích cụm là một kỹ thuật không giám sát, nhóm các đối tượng dựa trên một số tính năng. Kỹ thuật khai thác dữ liệu giúp trích xuất thông tin chưa biết, ẩn và hữu ích từ một tập dữ liệu khổng lồ. 10 thuật toán khai thác dữ liệu mạnh nhất đã được đưa vào danh sách rút gọn và thảo luận trong Wu et al. (2007). Các công cụ khác nhau có sẵn để phân tích dữ liệu bao gồm phần mềm nguồn mở và phần mềm thương mại. Một vài ví dụ về phần mềm nguồn mở là R để khai thác dữ liệu và trực quan hóa, Weka / Pentaho cho học máy và RapidMiner cho học máy và phân tích dự đoán.
   4. Trực quan hóa dữ liệu: Nhu cầu kiểm tra chi tiết ở nhiều tỷ lệ và chi tiết phút đã dẫn đến trực quan hóa dữ liệu. Giao diện trực quan cùng với các phân tích thống kê và bối cảnh liên quan giúp xác định các mẫu trong dữ liệu lớn theo thời gian (Fisher và cộng sự, 2012). Phân tích trực quan (VA) được định nghĩa là khoa học về lý luận phân tích được tạo điều kiện bởi các giao diện tương tác trực quan, (Thomas và Cook, 2005). Rất ít công cụ trực quan hóa là Tableau, QlikView, Spotfire, JMP, Jaspersoft, Visual Analytics, Ly tâm, Visual Mining và Board. Một so sánh các công cụ trực quan dựa trên chức năng xử lý dữ liệu, phương pháp phân tích và kỹ thuật trực quan của chúng đã được thảo luận trong Zhang et al. (2012).
7. **Ra quyết định:** Dựa trên phân tích và kết quả trực quan, những người ra quyết định có thể quyết định xem và làm thế nào để thưởng cho một hành vi tích cực và thay đổi hành vi tiêu cực. Các chi tiết của một vấn đề cụ thể có thể được phân tích để hiểu nguyên nhân của các vấn đề đưa ra các quyết định và kế hoạch sáng suốt cho các hành động cần thiết (Miller và Mork, 2013).

Trên đây đã trình bày về cách có thể trích xuất giá trị từ dữ liệu lớn. <ột ngành bất kể ngành nào nên xem xét ba tiêu chí trước khi triển khai phân tích dữ liệu lớn: có thể thu được thông tin hữu ích ngoài các thông tin thu được từ các hệ thống hiện tại, sẽ có bất kỳ cải thiện nào về độ chính xác về thông tin thu được bằng cách sử dụng phân tích dữ liệu lớn và cuối cùng, sẽ triển khai phân tích dữ liệu lớn giúp cải thiện tính kịp thời của phản ứng (Villars et al., 2011).

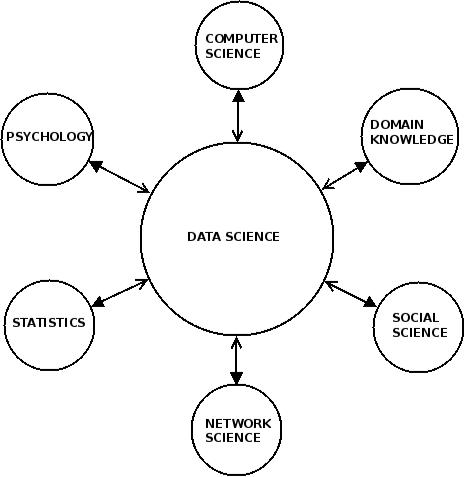
## 1.3 Các ông cụ thu thập, tiền xử lý và phân tích dữ liệu lớn

Sự tiến bộ trong kiến ​​trúc điện toán là cần thiết để xử lý cả các yêu cầu lưu trữ dữ liệu và xử lý máy chủ nặng cần thiết để phân tích khối lượng lớn và nhiều loại dữ liệu kinh tế (Villars et al., 2011). Phần này cung cấp một cái nhìn tổng quan về các công nghệ được áp dụng để phân tích dữ liệu lớn.

Với khả năng tính toán cao với chi phí tương đối rẻ cho phép các nhà nghiên cứu khám phá các cơ hội tiềm ẩn của dữ liệu lớn với lĩnh vực khoa học dữ liệu. Càng ngày, dữ liệu không chỉ được ghi lại để lưu giữ hồ sơ, mà để khám phá chúng bằng các hệ thống thông minh để có được những hiểu biết mới, có thể chưa được hình dung tại thời điểm thu thập dữ liệu. Bằng cách khởi xướng những câu hỏi thú vị và tinh chỉnh chúng mà không cần sự can thiệp của chuyên gia, nó sẽ có khả năng tự khám phá thông tin mới (Dhar, 2013). Ví dụ, nếu một thông tin có thể được dẫn xuất rằng một nhóm người cụ thể dễ bị ung thư và cũng cung cấp các thông tin khác như chế độ ăn uống, thói quen hàng ngày và bản chất của thuốc gây ra hiệu ứng này. Sẽ thật tuyệt vời khi phát triển các loại mô hình này, có thể trông giống như một câu chuyện viễn tưởng vào thời điểm này.

Những tiến bộ trong các công nghệ này giúp các nhà quản lý trong phân tích xây dựng kịch bản. Phân tích dữ liệu lớn có ứng dụng rất lớn trong các lĩnh vực khác nhau bao gồm thiên văn học, y tế và viễn thông. Mặc dù có lợi thế, phân tích dữ liệu lớn có những hạn chế và thách thức riêng. Vấn đề bảo mật và quyền riêng tư là mối quan tâm chính cho các nhà nghiên cứu. Những tiến bộ cũng như những hạn chế dẫn đến một lĩnh vực liên ngành mới gọi là khoa học dữ liệu, sử dụng các môn kiến ​​thức bao gồm tâm lý học, thống kê, kinh tế, khoa học xã hội, khoa học mạng và khoa học máy tính (Hình 2).

Các công cụ đang được sử dụng để thu thập dữ liệu bao gồm nhiều thiết bị kỹ thuật số khác nhau (ví dụ: thiết bị di động, máy ảnh, thiết bị đeo được và đồng hồ thông minh) và các ứng dụng tạo dữ liệu khổng lồ dưới dạng nhật ký, văn bản, giọng nói, hình ảnh và video. Để xử lý những dữ liệu này, một số nhà nghiên cứu đã đưa ra các kỹ thuật mới giúp thể hiện tốt hơn các dữ liệu phi cấu trúc, điều này có ý nghĩa trong bối cảnh dữ liệu lớn để có được những hiểu biết hữu ích mà có thể không được hình dung trước đó.



Hình 2: Sự phát triển của khoa học dữ liệu

**Ngôn ngữ truy vấn phi cấu trúc (NoSQL)**

Hệ thống quản lý cơ sở dữ liệu quan hệ (RDBMS) là phương pháp truyền thống để quản lý dữ liệu có cấu trúc. RDBMS sử dụng cơ sở dữ liệu và lược đồ quan hệ để lưu trữ và truy xuất dữ liệu. Kho dữ liệu được sử dụng để lưu trữ và truy xuất các bộ dữ liệu lớn. Ngôn ngữ truy vấn có cấu trúc (SQL) là ngôn ngữ truy vấn cơ sở dữ liệu được sử dụng phổ biến nhất. Dữ liệu được lưu trữ trong kho dữ liệu bằng cách sử dụng phương pháp thứ nguyên và phương pháp chuẩn hóa (Bakshi, 2012). Theo cách tiếp cận, dữ liệu được chia thành bảng thực tế và bảng thứ nguyên hỗ trợ bảng thực tế. Theo cách tiếp cận chuẩn hóa, dữ liệu được chia thành các thực thể tạo ra một số bảng trong cơ sở dữ liệu quan hệ.

Do ràng buộc về tính nguyên tố, tính nhất quán, tính độc lập và tính bền vững (ACID), việc mở rộng một khối lượng lớn dữ liệu là không thể. RDBMS không có khả năng xử lý dữ liệu bán cấu trúc và không cấu trúc (Qin et al., 2012; Mukherjee et al., 2012; Zikopoulos et al., 2012). Những hạn chế này của RDBMS đã dẫn đến khái niệm về NoQuery.

NoQuery lưu trữ và quản lý dữ liệu phi cấu trúc. Các cơ sở dữ liệu này còn được gọi là cơ sở dữ liệu không có lược đồ của Wikipedia vì chúng cho phép nâng cấp nhanh chóng cấu trúc dữ liệu mà không cần viết lại bảng. NoQuery hỗ trợ lưu trữ tài liệu, lưu trữ giá trị khóa, BigTable và cơ sở dữ liệu đồ thị. Nó sử dụng mô hình nhất quán lỏng hơn so với các cơ sở dữ liệu truyền thống. Các chức năng quản lý dữ liệu và lưu trữ dữ liệu được tách biệt trong cơ sở dữ liệu NoSQL (Bakshi, 2012). Nó cho phép khả năng mở rộng của dữ liệu. Một vài ví dụ về cơ sở dữ liệu NoSQL là HBase, MangoDB và Dynamo.

**Hadoop**

Năm 2005, một dự án Apache Hadoop mã nguồn mở đã được hình thành và triển khai trên cơ sở Google File System và mô hình lập trình Map Reduce (Prekopcsàk et al., 2011; Bakshi, 2012; Minelli et al., 2013).

**Hadoop Distributed File System (HDFS)**

HDFS là hệ thống lưu trữ phân tán có khả năng chịu lỗi, có tính khả mở, có khả năng cấu hình cao cho cụm Hadoop. Dữ liệu trong cụm Hadoop được chia thành nhiều phần bởi HDFS và được phân phối trên các máy chủ khác nhau trong cụm Hadoop. Một đoạn nhỏ của toàn bộ tập dữ liệu được lưu trữ trên máy chủ.

**Hadoop MapReduce**

MapReduce là một framwork để xử lý phân tán một lượng lớn dữ liệu theo cách tin cậy, chịu lỗi. Hai giai đoạn riêng biệt của MapReduce là:

1. **Map Phase:** Trong Map phase, khối lượng công việc được chia thành các khối lượng công việc phụ nhỏ hơn. Các tác vụ được gán cho Mapper, xử lý từng khối dữ liệu đơn vị để tạo ra một danh sách các cặp (khóa, giá trị) được sắp xếp. Danh sách này, là đầu ra của mapper, được chuyển sang giai đoạn tiếp theo. Quá trình này được gọi là xáo trộn.
2. **Reduce:** Trong Reduce phase, đầu vào được phân tích và hợp nhất để tạo ra đầu ra cuối cùng được ghi vào HDFS trong cụm.

Bảng 2 tóm tắt các khả năng dữ liệu lớn và các công nghệ chính có sẵn (Sun và Heller, 2012).

*Table 2: Big Data Capabilities and their Primary Technologies*

**Các hạn chế của Hadoop**

Mặc dù có những lợi thế của Hadoop so với RDBMS, nó vẫn chịu những hạn chế sau (ParAccel, 2012):

1. **Nhiều bản sao dữ liệu:** HDFS không hiệu quả dẫn đến việc tạo ra nhiều bản sao của dữ liệu (tối thiểu 3 bản).
2. **Hỗ trợ SQL có giới hạn:** Hadoop cung cấp hỗ trợ SQL có giới hạn và họ thiếu các chức năng cơ bản như truy vấn phụ, phân tích “group by” v.v.
3. **Thực thi không hiệu quả:** Thiếu trình tối ưu hóa truy vấn dẫn đến kế hoạch dựa trên chi phí không hiệu quả để thực hiện do đó dẫn đến cụm lớn hơn so với cơ sở dữ liệu tương tự.
4. **Hạn chế về Framework:** Logic chuyển đổi phức tạp không thể được sử dụng bằng MapReduce framwork.
5. **Đòi hỏi kỹ năng:** Cần có kiến thức về thuật toán và kỹ năng để phát triển MapReduce phân tán để thực hiện đúng.

Một trong những thách thức lớn nhất là có một cơ sở hạ tầng điện toán có thể phân tích dữ liệu khối lượng lớn và đa dạng (có cấu trúc và không cấu trúc) từ nhiều nguồn và cho phép phân tích thời gian thực nội dung không thể đoán trước mà không có lược đồ hoặc cấu trúc rõ ràng (Villars et al., 2011).

## 1.4 Các công cụ xử lý dữ liệu lớn:

Có nhiều công cụ giúp đạt được các mục tiêu này và giúp các nhà khoa học dữ liệu xử lý dữ liệu để phân tích chúng. Nhiều ngôn ngữ, khung và công nghệ lưu trữ dữ liệu mới đã xuất hiện hỗ trợ xử lý dữ liệu lớn.

**R:** là một ngôn ngữ tính toán thống kê nguồn mở cung cấp nhiều kỹ thuật thống kê và đồ họa để rút ra những hiểu biết sâu sắc từ dữ liệu. Nó có một cơ sở lưu trữ và xử lý dữ liệu hiệu quả và hỗ trợ các hoạt động véc tơ với một bộ các toán tử để xử lý nhanh hơn. Nó có tất cả các tính năng của ngôn ngữ lập trình tiêu chuẩn và hỗ trợ các đối số có điều kiện, các vòng lặp và các hàm do người dùng định nghĩa. R được hỗ trợ bởi một số lượng lớn các gói thông qua Mạng lưu trữ toàn diện R (CRAN). Nó có sẵn trên các nền tảng Windows, Linux và Mac. Nó có một tài liệu mạnh mẽ cho mỗi gói. Nó có một sự hỗ trợ mạnh mẽ cho việc trộn dữ liệu, khai thác dữ liệu và thuật toán học máy cùng với sự hỗ trợ tốt cho việc đọc và viết trong môi trường phân tán, giúp nó phù hợp để xử lý dữ liệu lớn. Tuy nhiên, quản lý bộ nhớ, tốc độ và hiệu quả có lẽ là thách thức lớn nhất mà R. R Studio phải đối mặt là Môi trường phát triển tích hợp được phát triển để lập trình bằng ngôn ngữ R. Nó được phân phối cho các máy tính để bàn độc lập cũng như hỗ trợ kiến ​​trúc máy khách-máy chủ, có thể được truy cập từ bất kỳ trình duyệt nào.

**Python:** là một ngôn ngữ lập trình phổ biến khác, là nguồn mở và được hỗ trợ bởi các nền tảng Windows, Linux và Mac. Nó lưu trữ hàng ngàn gói từ các mô-đun đóng góp của bên thứ ba hoặc cộng đồng. NumPy, Scikit và Pandas hỗ trợ một số gói phổ biến để học máy và khai thác dữ liệu để xử lý trước dữ liệu, tính toán và mô hình hóa. NumPy là gói cơ sở cho máy tính khoa học. Nó bổ sung hỗ trợ cho các mảng và ma trận lớn, đa chiều với Python. Scikit hỗ trợ phân loại, hồi quy, phân cụm, giảm kích thước, lựa chọn tính năng, và các thuật toán lựa chọn mô hình và tiền xử lý.

Pandas hỗ trợ trong việc trộn dữ liệu và chuẩn bị cho phân tích dữ liệu và mô hình hóa. Nó hỗ trợ mạnh mẽ cho phân tích biểu đồ với thư viện NetworkX và nltk để phân tích văn bản và xử lý ngôn ngữ tự nhiên. Python rất thân thiện và tuyệt vời để phân tích nhanh về một vấn đề. Nó cũng tích hợp tốt với spark thông qua thư viện pyspark.

**Scala**: là một ngôn ngữ hướng đối tượng và là từ viết tắt của “Scalable Language” - ngôn ngữ có thể mở rộng. Đối tượng và mọi hoạt động trong Scala là một cuộc gọi phương thức, giống như bất kỳ ngôn ngữ hướng đối tượng nào. Nó đòi hỏi môi trường máy ảo java. Spark, một khung tính toán cụm trong bộ nhớ được viết bằng Scala. Scala đang trở thành công cụ lập trình phổ biến để xử lý các vấn đề dữ liệu lớn.

**Apache Spark:** là một công nghệ điện toán cụm trong bộ nhớ được thiết kế để tính toán nhanh, được triển khai trong Scala. Nó sử dụng Hadoop cho mục đích lưu trữ vì nó có khả năng quản lý cụm riêng. Nó cung cấp các API tích hợp cho Java, Scala và Python. Gần đây, nó cũng đã bắt đầu hỗ trợ R. Nó đi kèm với 80 toán tử cấp cao để truy vấn tương tác. Tính toán trong bộ nhớ được hỗ trợ với khung Dữ liệu phân tán đàn hồi (RDD), phân phối khung dữ liệu thành các phần nhỏ hơn trên các máy khác nhau để tính toán nhanh hơn. Nó cũng hỗ trợ Map và Giảm để xử lý dữ liệu. Nó hỗ trợ SQL, truyền dữ liệu, thuật toán xử lý đồ thị và thuật toán học máy. Mặc dù Spark có thể được truy cập bằng Python, Java và R, nhưng nó có hỗ trợ mạnh mẽ cho Scala và ổn định hơn tại thời điểm này. Nó hỗ trợ học sâu với sparkling water trong H2O.

**Apache Hive:** là một nền tảng nguồn mở cung cấp các phương tiện để truy vấn và quản lý các bộ dữ liệu lớn nằm trong bộ lưu trữ phân tán (Ví dụ: HDFS). Nó tương tự như SQL và nó được gọi là HiveQL. Nó sử dụng MapReduce để xử lý các truy vấn và cũng hỗ trợ các nhà phát triển cắm mã trình ánh xạ và trình giảm tùy chỉnh của họ khi HiveQL thiếu thể hiện logic mong muốn.

**Apache Pig:** à một nền tảng cho phép các nhà phân tích phân tích các tập dữ liệu lớn. Đây là ngôn ngữ lập trình cấp cao, được gọi là Pig Latin để tạo các chương trình MapReduce yêu cầu Hadoop để lưu trữ dữ liệu. Mã Pig Latin được mở rộng với sự trợ giúp của các Hàm do người dùng xác định có thể được viết bằng Java, Python và một số ngôn ngữ khác. Có thể tuân thủ song song đáng kể, điều này cho phép họ xử lý các tập dữ liệu rất lớn.

**Amazon Elastic Compute Cloud (EC2)**: à một dịch vụ web cung cấp khả năng tính toán trên đám mây. Nó cho phép kiểm soát hoàn toàn các tài nguyên máy tính và cho phép các nhà phát triển chạy tính toán của họ trong môi trường điện toán mong muốn. Đây là một trong những nền tảng điện toán đám mây thành công nhất. Nó hoạt động theo nguyên tắc của mô hình pay-as-you-go.

Một vài framework khác hỗ trợ xử lý dữ liệu lớn là MongoDB, BlinkDB, Tachyon, Cassandra, CouchDB, Clojure, Tableau, Splunk và một số framework khác.

# CHƯƠNG 2 Mô hình kinh doanh dữ liệu lớn

Một ví dụ điển hình về hiệu quả của việc khai thác Big Data trong lĩnh vực kinh doanh có thể kể đến Amazon. Ông lớn trong lĩnh vực thương mại điện tử này có thể dự đoán được tại từng khu vực địa lý khác nhau, sản phẩm nào sẽ có nhu cầu cao trong thời gian tới và chuẩn bị hàng ngay từ khi khách hàng còn chưa có ý định mua. Nhờ vào cơ sở dữ liệu khổng lồ về thói quen mua sắm, thời tiết, văn hóa tiêu dùng và thậm chí cả tình hình chính trị, Amazon luôn đi trước một bước so với đối thủ. Mới đây, doanh nghiệp còn triển khai chương trình chào hàng tận nhà và tỷ lệ khách mua hàng đạt con số ấn tượng là 65% từ đó cho thấy việc phân tích dữ liệu của Amazon đã thật sự mang lại hiệu quả.

Với hoạt động bán lẻ trực tuyến, Amazon.com xử lí hàng triệu hoạt động mỗi ngày cũng như những yêu cầu từ khoảng nửa triệu đối tác bán hàng. Thông qua những hoạt động đó, doanh nghiệp này đã tiếp cận được một lượng rất lớn những dữ liệu về khách hàng, như: tên, địa chỉ, thông tin thanh toán và lịch sử tìm kiếm đều được lưu trữ trong ngân hàng dữ liệu.

Google và Facebook cũng đã sử dụng Big Data để xác định mục tiêu quảng cáo mà chúng ta có thể thấy xuất hiện gần đây trên màn hình và các ứng dụng của họ. Ngoài ra, một số tập đoàn khác, chẳng hạn như các công ty bảo hiểm, bất động sản, ở các nước mà luật pháp cho phép cũng đã thu thập các dữ liệu cá nhân để xác định phương hướng kinh doanh và đối tượng khách hàng.

## 2.1 *Các mô hình kinh doanh dữ liệu*

Thật dễ dàng để tìm thấy một ví dụ về một công ty có thể được coi là con đẻ của dữ liệu lớn. Thông thường, những câu chuyện thành công như vậy sẽ liên quan đến việc sử dụng sáng tạo dữ liệu lớn để cung cấp các sản phẩm mới hoặc để đạt được hiệu quả lớn theo một cách nào đó. Trong phần này, chúng tôi lùi lại từ các thực tiễn cụ thể để có cái nhìn trừu tượng hơn về các cách mà dữ liệu có thể đóng vai trò là thành phần trung tâm của mô hình kinh doanh và các cơ hội mà những điều này ngụ ý. Cần nhấn mạnh rằng các mô hình kinh doanh này không loại trừ lẫn nhau và nhiều tổ chức trong mẫu của chúng tôi tham gia vào nền kinh tế dữ liệu thông qua nhiều hơn một trong các kênh này.

#### 2.1.1. Đưa ra quyết định kinh doanh

Trước khi thảo luận về cách kiếm tiền trực tiếp từ dữ liệu độc quyền, ta thấy rằng trong nhiều trường hợp, dữ liệu không cần phải được mua bán trực tiếp. Thật vậy, các công ty đã sử dụng dữ liệu trong nội bộ để thông báo các quyết định chiến lược và hoàn thiện quy trình kinh doanh từ rất lâu trước khi kinh doanh kích hoạt dữ liệu trở thành mốt (Beniger, 1986). Trong những trường hợp này, dữ liệu được sử dụng làm đầu vào cho quy trình quản lý. Hiệu quả của dữ liệu lớn đã được chứng minh mô hình này. Trong các công ty/tổ chức lớn, dữ liệu của bên thứ nhất được sử dụng để thông báo các quyết định kinh doanh giữa các năm với chất lượng rất tốt. Đối với các nhà cung cấp từ doanh nghiệp đến doanh nghiệp như Rolls Royce, mô hình kinh doanh chính là bán thiết bị cho các khách hàng như Boeing hoặc Virgin Atlantic. Tuy nhiên, bằng cách chạy trong nền, dữ liệu được thu thập thông qua các cảm biến từ xa được cài đặt trên thiết bị của họ cảnh báo công ty về các vấn đề bảo trì cho phép họ cung cấp dịch vụ tốt hơn và nghiên cứu và phát triển thông tin. Tương tự như vậy, các nhà bán lẻ như Tesco và Starbucks đã đi tiên phong trong việc sử dụng thẻ thưởng để thu thập dữ liệu về khách hàng của họ và phù hợp với hành vi mua hàng trực tuyến và ngoại tuyến của họ. Những dữ liệu này cũng thông báo các quyết định về sản phẩm, giá cả, chương trình khuyến mãi, giữ cổ phiếu và chiến lược kinh doanh tổng thể. Trong khi mô hình kinh doanh chính cho các công ty này là bán lẻ, dữ liệu của bên thứ nhất cung cấp nhiều quyết định quan trọng.

#### 2.1.2. Môi giới dữ liệu

Một cách rõ ràng để kiếm tiền từ dữ liệu của bên thứ nhất là coi nó như một sản phẩm thương mại và bán cho các bên khác. Do đó, dữ liệu của bên thứ nhất được coi là đầu ra theo đúng nghĩa của nó. Một ví dụ tương đối thuần túy của loại mô hình kinh doanh này có thể thấy trong Nielsen, một công ty nghiên cứu thị trường, chuyên cung cấp dữ liệu và phân tích về hành vi của khán giả. Nielsen thu thập dữ liệu của bên thứ nhất bằng bảng đối tượng dựa trên các lĩnh vực nghiên cứu riêng hoặc nghiên cứu theo hợp đồng. Mô hình kinh doanh cho Nielsen là cung cấp dữ liệu liên quan đến nghiên cứu đối tượng ở các định dạng khác nhau dựa trên thông số kỹ thuật của khách hàng.

Trong các trường hợp khác, các công ty đang khám phá ra rằng dữ liệu họ tạo ra thông qua hoạt động hàng ngày của doanh nghiệp có thể có giá trị thị trường theo đúng nghĩa của nó. Các công ty truyền thông xã hội như Twitter, để kiểm tra, bán quyền truy cập vào dữ liệu họ lưu trữ cho các bên thứ ba sử dụng nó cho nhiều mục đích khác nhau như phân tích thị trường và phân tích tình cảm. Tương tự như vậy, các tổ chức tin tức và nền tảng truyền thông trực tuyến thu thập dữ liệu từ khách truy cập vào trang web của họ. Dữ liệu của bên thứ nhất này chủ yếu phản ánh các hành vi web (tìm kiếm, lượt xem, nhấp chuột, tải xuống và bài đăng) và thông tin về vị trí và thiết bị. Mặc dù dữ liệu này thông báo cho việc ra quyết định nội bộ, chủ sở hữu trang web cũng đóng vai trò là nhà môi giới dữ liệu và bán dữ liệu này cho bên thứ ba. Họ cũng có tùy chọn cộng tác với các doanh nghiệp khác như nhà quảng cáo để chạy các chiến dịch dựa trên dữ liệu này.

Theo Davis, Giám đốc CNTT của Analytics tại Willis Group, lý do tại sao các nhà môi giới dữ liệu rất quan trọng đối với nền kinh tế dữ liệu đang phát triển là: một dải dữ liệu mở rộng mà chúng tôi muốn sử dụng và phân tích, và phạm vi nguồn đang ngày càng tăng. Các phạm vi này từ nguồn cấp dữ liệu trực tiếp thông thường như các dịch vụ web được lưu trữ bởi các công ty và cơ quan thông qua các số liệu thống kê, nhân khẩu học và dữ liệu rủi ro từ số lượng bên thứ ba ngày càng tăng. Khi việc kết hợp dữ liệu từ các nguồn khác nhau trong quá trình phân tích trở nên ngày càng phổ biến, nhiều khả năng các tổ chức sẽ chuyển sang các bên thứ ba để cung cấp dữ liệu đó. Rõ ràng là không có hiệu quả hoặc thiết thực cho mọi tổ chức để thu thập dữ liệu của riêng mình so với mọi tổ chức để tự khoan dầu hoặc tự tạo ra điện.

Môi giới dữ liệu không phải là mới. Danh sách khách hàng từ lâu đã được xem là tài sản độc quyền có giá trị và có thể bán được, trong khi các công ty khác nhau đã cung cấp dữ liệu giá cổ phiếu theo thời gian thực từ sàn của Sàn giao dịch chứng khoán New York kể từ cuối thế kỷ XIX. Tuy nhiên, hai xu hướng mới quan trọng đang nổi lên. Thứ nhất, khi các kỹ thuật phân tích và năng lực tính toán mở rộng sang văn bản mã hóa và các dạng dữ liệu phi cấu trúc khác, phạm vi cấu thành dữ liệu (và do đó, tạo thành một nhà môi giới dữ liệu) đã tăng lên. Do văn bản hoàn chỉnh của kho lưu trữ báo giờ có thể được coi là dữ liệu, nên một công ty nắm giữ kho lưu trữ đó có thể trở thành nhà môi giới dữ liệu chỉ bằng cách truy cập thông qua API. Thứ hai, đã có sự tăng trưởng lớn về số lượng các hoạt động kinh doanh, giao dịch và tương tác được trung gian kỹ thuật số. Điều này có nghĩa là dữ liệu trước đây đã bị loại bỏ hoặc không bao giờ bị bắt ở nơi đầu tiên hiện được lưu trữ kỹ thuật số từ Sinh ra. Do đó, các công ty không quan tâm đến dữ liệu sẽ thấy mình sở hữu dữ liệu có thể có giá trị đối với người khác.

#### 2.1.3. Dịch vụ phân tích dữ liệu

Nhiều chuyên gia đã nhấn mạnh rằng giá trị của dữ liệu không nằm ở giá trị nội tại của nó, mà nằm ở các hành động do phân tích. Người dân muốn có câu trả lời; họ không muốn có nhiều dữ liệu, như Vivienne Ming của Gild đã nói. Tuy nhiên, nhiều tổ chức không phải là công ty dữ liệu hiện không có chuyên môn hoặc năng lực nội bộ để thực hiện phân tích đó. Kết quả là, một mô hình kinh doanh phổ biến cho các công ty trong lĩnh vực dữ liệu lớn là việc cung cấp các phân tích dưới dạng dịch vụ. Các hình thức tiềm năng để phân tích như một dịch vụ rất đa dạng. Các ví dụ trong mẫu của chúng tôi từ các tổ chức lớn như Experian, dựa trên các bộ dữ liệu khổng lồ để cung cấp điểm tín dụng tiêu dùng, cho các công ty khởi nghiệp như Gild, tuyển dụng thay mặt các công ty công nghệ sử dụng dữ liệu web để lập hồ sơ nhân viên tiềm năng. Trong mọi trường hợp, tính năng xác định của các doanh nghiệp dịch vụ phân tích là họ lấy dữ liệu đầu vào (dữ liệu độc quyền của riêng mình, dữ liệu do khách hàng cung cấp, một số nguồn dữ liệu của bên thứ ba hoặc bất kỳ kết hợp nào của chúng) và tạo ra như một đầu ra một bản tóm tắt dữ liệu, phân tích, cái nhìn sâu sắc, lời khuyên hoặc một số sản phẩm khác có nguồn gốc từ dữ liệu đó.

Phân tích dữ liệu lớn cũng đang trở nên có sẵn cho sử dụng cá nhân với sự gia tăng của các công ty phân tích đối mặt với người tiêu dùng. Mappiness, một ứng dụng di động cho phép người dùng báo cáo mức độ hạnh phúc của họ và nhận phản hồi, thu thập dữ liệu cá nhân, phân tích và báo cáo dưới dạng có thể sử dụng cho người dùng, trong khi BrightScope cung cấp xếp hạng kế hoạch nghỉ hưu và phân tích đầu tư dựa trên sự kết hợp mở và phân tích đầu tư dựa trên sự kết hợp của mở và dữ liệu cá nhân. Một phần quan trọng trong mô hình kinh doanh của các công ty như Amazon và Netflix cũng là việc cung cấp các khuyến nghị dựa trên dữ liệu cho người tiêu dùng, vừa nâng cao trải nghiệm của khách hàng và cải thiện khả năng giữ chân khách hàng.

#### 2.1.4. Tư vấn và gợi ý

Nhận thức đầy đủ lợi ích của dữ liệu lớn đòi hỏi phải có chuyên môn về công nghệ, phân tích dữ liệu, chiến lược kinh doanh và tổ chức, đạo đức và một loạt các lĩnh vực khác. Hàng tá câu hỏi phải được giải quyết trong việc xây dựng và thực hiện chiến lược dữ liệu mạch lạc. Các ví dụ bao gồm kiến trúc tốt nhất cho cơ sở hạ tầng lưu trữ dữ liệu vật lý, nhân viên dữ liệu nên được đặt trong hệ thống phân cấp quản lý như thế nào, nên đưa ra các giao thức bảo mật nào để bảo vệ tính toàn vẹn của dữ liệu và quan điểm đạo đức phù hợp trong việc xử lý cá nhân dữ liệu? Tuy nhiên, giống như một số công ty không có vị trí tốt để thực hiện phân tích dữ liệu của riêng họ, những công ty khác thiếu chuyên môn nội bộ để giải quyết tất cả các khía cạnh của vấn đề chiến lược này. Điều này đã tạo ra một ngành công nghiệp của các công ty, như IBM, cung cấp tư vấn và chuyên môn về chính xác những vấn đề này.

#### 2.1.5. Cung cấp cơ sở hạ tầng

Phương tiện lưu trữ, máy chủ và máy trạm, máy quét mã vạch, phần mềm phân tích thống kê và trực quan hóa, phần mềm cơ sở dữ liệu, cảm biến từ xa, công nghệ mã hóa và thiết bị mạng và nhiều ví dụ khác về phần cứng và phần mềm tạo thành công cụ thương mại cho một doanh nghiệp tích hợp dữ liệu. Do đó, các nhà sản xuất các công cụ này là một phần quan trọng của nền kinh tế dữ liệu lớn. Ví dụ về các công ty cung cấp các công cụ dữ liệu lớn bao gồm IPONweB, nhà cung cấp cơ sở hạ tầng và công nghệ cho ngành quảng cáo trực tuyến và ESRI, cung cấp nền tảng phân tích phần mềm không gian địa lý.

## 2.2 *Các lớp trong mô hình kinh doanh dữ liệu lớn:*

Các mô hình kinh doanh dữ liệu lớn được mô tả ở trên có thể được nhóm thành ba loại để mang lại một phân loại mô hình kinh doanh dữ liệu mới. Danh mục đầu tiên là những gì có thể được gọi là người dùng dữ liệu. Đây là những tổ chức sử dụng dữ liệu để thông báo quyết định kinh doanh hoặc làm đầu vào cho các sản phẩm và dịch vụ khác như báo cáo tín dụng hoặc chiến dịch quảng cáo được nhắm mục tiêu. Đây là những tổ chức tham gia trả lời câu hỏi: làm thế nào dữ liệu có thể được sử dụng để tạo ra giá trị trong doanh nghiệp của họ?

Lớp thứ hai của mô hình kinh doanh bao gồm các nhà cung cấp dữ liệu. Đây là các tổ chức tạo ra dữ liệu có giá trị nội tại và do đó có thể bán được, hoặc nếu không thì đóng vai trò môi giới bằng cách cung cấp quyền truy cập vào tổng hợp dữ liệu của bên thứ nhất và bên thứ ba. Các công ty như vậy không cần chuyên cung cấp dữ liệu. Thật vậy, nhiều tổ chức đang thấy rằng họ nắm giữ dữ liệu có giá trị có thể xem xét khi một số bên thứ ba đưa nó vào sử dụng khác với dữ liệu ban đầu được thu thập. Vì, giống như hầu hết các hàng hóa thông tin, chi phí cố định của sản xuất dữ liệu thường cao so với chi phí phân phối thay đổi, có khả năng đạt được hiệu quả lớn từ việc sử dụng lại dữ liệu này.

Lớp thứ ba của mô hình kinh doanh bao gồm một loạt các hoạt động hỗ trợ các bên thứ ba thiếu cơ sở hạ tầng hoặc chuyên môn. Những người hỗ trợ dữ liệu này thực hiện một loạt các dịch vụ tư vấn không rõ ràng về cách tận dụng dữ liệu lớn, cung cấp cơ sở hạ tầng vật lý và chứng minh các dịch vụ phân tích thuê ngoài. Các tổ chức này đang đóng một vai trò đặc biệt quan trọng trong thời gian chuyển đổi hiện nay khi một số lượng lớn các công ty đang tổ chức lại để làm cho dữ liệu tập trung hơn vào hoạt động kinh doanh của họ, nhưng vẫn thiếu chuyên môn hoặc năng lực để thực hiện hoàn toàn trong nội bộ.

Người dùng dữ liệu phụ thuộc vào cơ sở hạ tầng và dữ liệu được cung cấp bởi người hỗ trợ dữ liệu và nhà cung cấp dữ liệu, tương ứng. Tương tự như vậy, người hỗ trợ dữ liệu đóng một vai trò thực tế quan trọng trong việc cho phép thu thập và tổng hợp dữ liệu của các nhà cung cấp dữ liệu. Cuối cùng, cả nhà cung cấp dữ liệu và người hỗ trợ dữ liệu đều phụ thuộc vào cộng đồng người dùng dữ liệu tích cực để tạo thị trường cho sản phẩm của họ. Do các loại hình kinh doanh dữ liệu lớn khác nhau phụ thuộc rất nhiều vào nhau để thành công, nên bất kỳ chính sách nào nhằm tăng cường nền kinh tế dữ liệu đều cần có lập trường tương đối toàn diện, nhằm mục đích thúc đẩy các hoạt động trên toàn bộ mô hình kinh doanh dữ liệu lớn. Các loại hình được tóm tắt trong Bảng 2.

Bảng 2

|  |  |  |
| --- | --- | --- |
| Type | Example functions | Dependencies |
| Data users | Using data to inform strategic decisions; building data into products | Depend on suppliers for raw data, and on facilitators for infrastructure and skills |
| Data suppliers | Gathering primary data; aggregating and packaging data for sale | Depend on facilitators for infrastructure and skills, and on users both as customers and as sources of data |
| Data facilitators | Supplying infrastructure; consultancy; outsourced analysis | Depend on users and on suppliers as customers |

# KẾT LUẬN

Big Data và các công nghệ phân tích có khả năng làm thay đổi hoàn toàn bộ mặt của các ngành kinh tế và các nghề nghiệp. Chính vì thế mà những mô hình kinh doanh mới dựa trên Big Data đang được hình thành một cách rộng rãi để giúp các tổ chức, doanh nghiệp tận dụng được nguồn dữ liệu này. Đây cũng là xu thế tất yếu trong tương lai không xa với doanh nghiệp VN đặc biệt là trong bối cảnh hội nhập sâu vào rộng vào nền kinh tế quốc tế như hiện nay. Dù vẫn còn tồn tại những thách thức và hạn chế nhưng Big Data vẫn là một xu thế không thể chối bỏ của thế giới. Tuy nhiên, các doanh nghiệp cần phải hết sức thận trọng trong việc sử dụng và ứng dụng nó. Một khi làm chủ được Big Data thì họ sẽ có cơ hội thành công lớn hơn trong bối cảnh cạnh tranh ngày nay, thế giới sẽ được hưởng lợi nhiều hơn từ việc trích xuất thông tin một cách chính xác, hữu ích với những chi phí thấp hơn. Vẫn còn đó những chỉ trích xoay quanh Big Data, tuy nhiên lĩnh vực này vẫn còn đang rất mới và chúng ta hãy chờ xem trong tương lai Big Data sẽ tiến hóa như thế nào.

# CÂY CÔNG NGHỆ

|  |  |  |  |
| --- | --- | --- | --- |
| **Hệ sinh thái Big Data** | Data Generation | Business Data |  |
| E-Commerce |  |
| Social Media |  |
| Scientific Data |  |
| Healthcare |  |
| Fiannce |  |
| Telecom |  |
| Data Collection | Log Files |  |
| Sensors |  |
| Web Crawler |  |
| Network Monitoring System |  |
| Data Transmission | Inter-DCN Transmission |  |
| Intra-DCN Transmission |  |
| Data Preprocessing | Integration |  |
| Cleaning | * Open Refine   (Larsson, 2013) [33] |
| * Data Cleaner   (Larsson, 2013) [33] |
| Elimination of Redundant Data |  |
| Feature Extraction |  |
| Data Storage | GFS | Hệ thống file của Google |
| HDFS | Là khung để lưu trữ lớn thiết lập dữ liệu bằng cách phân phối nó trên cụm máy tính.  (Bhosale and Gadekar, 2014) [7]  Ứng dụng: Yahoo, Facebook, Twitter,... |
| Cloudera | Xây dựng trung tâm dữ liệu cho doanh nghiệp doanh nghiệp và giúp đỡ tổ chức tiếp cận tốt hơn với dữ liệu.  (Pol, 2014) [55]  Ứng dụng: HP, Oracle,... |
| MongoDB | MongoDB là một kho lưu trữ dữ liệu và công cụ quản lý cho cấu trúc và dữ liệu bán cấu trúc thay đổi thường xuyên.  (Zhao, 2013) [80]  Ứng dụng: Uber, Codeacademy,... |
| MDM | MDM kết hợp dữ liệu thời gian thực của sản phẩm và khách hàng trên các doanh nghiệp và tích hợp chúng.  (Myers, 2016) [47]  Ứng dụng: |
| Cassandra | NoSQL được phát triển bởi Facebook.  Ứng dụng: eBay, GitHub, Instagram, Netflix, Reddit,... |
| Data Processing | R | R là ngôn ngữ lập trình và môi trường phần mềm cho tính toán thống kê và đồ họa.  (Goyal and Singh, 2015) [24] |
| Python | Python là một ngôn ngữ mã nguồn mở để xử lý và phân tích các dữ liệu.  (Babu and Babu, 2016) [3] |
| Apache Samza | Apache Samza là một khung xử lý luồng được gắn chặt với hệ thống nhắn tin Apache Kafka.  Ứng dụng: LinkedIn, Netflix, TripAdvisor |
| Apache Spark | Apache Spark là một khung xử lý hàng loạt thế hệ tiếp theo với khả năng xử lý luồng.  Ứng dụng: Pinterest, Conviva |
| Apache Storm | Apache Storm là một khung xử lý luồng tập trung vào độ trễ cực thấp.  Ứng dụng: Twitter, Wego, Navisite |
| Apache Flink | Ứng dụng: Alibaba, Amazon, Ebay |
| Amazon Elastic Compute Cloud (EC2) | Ứng dụng: Amazon |
| Data Analysis | Data Mining | * Rapid Miner   (Rangra, and Bansal, 2014) [57] |
| * Oracle Data Mining |
| * Teradata |
| * Kaggle |
| Text Mining |  |
| Multivariate Statistical Analytics |  |
| Network Analytics |  |
| Multimedia Analytics |  |
| Data Visualization | Tableau | Là công cụ kinh doanh thông minh để phân tích dữ liệu một cách trực quan.  (Pradhananga et. al, 2015) [56] |
| Platfora | Platfora chuyển đổi dữ liệu lớn thô trong Hadoop vào dữ liệu tương tác cơ chế xử lý  (Wang et. al, 2015a) [65] |
| Datawrapper | Datawrapper là một nguồn mở công cụ giúp mọi người sáng tạo đơn giản, chính xác và có thể nhúng biểu đồ trong vài phút.  (Babu and Babu, 2016) [3] |

# HỒ SƠ CÔNG NGHỆ

|  |  |  |
| --- | --- | --- |
| Tên công nghệ | Công nghệ cho hệ sinh thái Big data | |
| Lớp công nghệ | 3 | |
| **Phạm vi ứng dụng:**  Hệ sinh thái Big data là công nghệ mới sử dụng cho nguồn dữ liệu lớn, ngày càng tăng về kích cỡ, nhằm mục tiêu lưu trữ một cách hiệu quả, an toàn, có khả năng chịu lỗi, phục vụ cho quá trình xử lí dữ liệu tiếp theo được nhanh chóng, hiệu quả, tiết kiệm tài nguyên. | | |
| **Mô tả công nghệ:**  Dữ liệu phi cấu trúc lớn từ các nguồn khác nhau tạo ra chính vấn đề phức tạp. Để giải quyết vấn đề này, dữ liệu lớn cần một số công nghệ đặc biệt [28]. Giới thiệu dữ liệu lớn công nghệ khác nhau và các kỹ thuật khác nhau để phân tích dữ liệu, thao tác và trực quan hóa dữ liệu. Như thế đó nhiều cách để xử lý dữ liệu lớn, Hadoop [7] rất nổi tiếng khung lập trình nguồn mở. Nó được sử dụng bởi các nhà nghiên cứu hoặc bởi các tổ chức dữ liệu lớn. Có khác nhau kỹ thuật được sử dụng để phân tích dữ liệu. Phần mềm Hadoop [33] là được sử dụng cho dữ liệu phi cấu trúc và bán cấu trúc.  Môi trường Hadoop Apache hiện tại bao gồm Kernel, HDFS, Map Reduce và các thành phần khác như Apache Hive, Zookeeper và HBase. Dữ liệu được phân tích bằng cách sử dụng NoQuery, do đódữ liệu thay đổi liên tục thay vì cố định như trong cơ sở dữ liệu truyền thống môi trường. | | |
| **Các thách thức của công nghệ:**  Sự đa dạng thông tin (Heterogeneity): Trên thực tế thì sự tồn tại của một dữ liệu đồng nhất và có cấu trúc tốt với cùng kích thước và cấu trúc là rất cần thiết. Chính vì vậy, việc làm sạch và xử lý các lỗi từ các nguồn dữ  liệu không đầy đủ hay các dữ liệu gây nhiễu là tối quan trọng và không thể thiếu.  Quy mô (Scale): Thực tế cho thấy rằng khối lượng dữ liệu đang gia tăng khá lớn và gần như đã vượt qua khả năng sức mạnh xử lý và lưu trữ của máy tính.  Yếu tố thời gian (Timeliness): Các kết quả phân tích thì cần phải được đưa ra một cách tức thì nghĩa là trong thời gian thực nên cần phải có sự phát triển các quy trình xử lý dựa vào sự gia tăng của các cấu trúc dữ liệu nhằm đạt được mục đích cuối cùng đó là đáp ứng yêu cầu truy cập thông tin một cách tức thì của người sử dụng.  Tính cá nhân (Privacy): Các yếu tố luật pháp liên quan đến các hạn chế về quyền cá nhân hiện đang trở thành một mối đe dọa vì nó có thể hạn chế các ích lợi tiềm năng. Chính vì lý do đó mà các chính sách an ninh mạng  hay cụ thể là an ninh thông tin chia sẻ các dữ liệu liên quan đến người dùng cá nhân cần phải được xem xét một cách thận trọng và nghiêm túc.  Sự kết hợp của con người  (Human collaboration): Các đánh giá của chuyên gia là hết sức cần thiết để có thể hiểu các nguồn dữ liệu và để khai thác một cách tối ưu các tiềm năng của nó đòi hỏi sự phối hợp nhịp nhàng và nhuần nhuyễn giữa máy móc và con người | | |
| **Hệ thống lớn sử dụng Big Data** |  |  |
| Amazon Web services (ERM) | Cung cấp khả năng tương tác với các AWS lưu trữ dữ liệu khác như Amazon S3, Amazon DynamoDB |
| MapR Technologies | Nền tảng NFS được viết trên C++, với khả năng đọc ghi hiện đại.Tích hợp cả phương pháp lưu trữ tệp tin phân tán và cơ sở dữ liệu dạng bảng trển cùng một lớp hợp nhất. Cho phép các công ty hỗ trơ cả các ứng dụng tính toán(Hbase) và phân tích(Hive, Impala) |
| Cloudera(CDH) | Chứa nhân của Hadoop cung cấp khả năng mở rộng xử lý dữ liệu phân tán trên tập dữ liệu lớn( MapReduce, HDFS). Đồng thời bổ sung thêm các thành phần đảm bảo tính khả dụng và bảo mật cũng như việc tích hợp với phần cứng và các sản phẩm phần mềm. |
| Microsoft Azure | Cung cấp cá dịch vụ phân tích xây dựng trên nền tảng YARN. |
| IBM | Cung cấp kiên trúc hiệu quả cho công việc với dữ liệu lớn dựa trên MapReduce. Bao gồm cá thành phần qua trong đối với ứng dụng sử dụng dữ liệu lớn như: servers, hệ thống lưu trữ, runtime libraries, hệ điều hành, các ứng dụng tương thích với Hadoop. Các Cluster được tích hợp với Cluster Manager, InfoSphere. |
| **Benchmark** |  |  |
|  | |
|  |  |  |
| **Triển khai hệ thống** | Cấu hình tối thiểu | * Intel Core 2 Duo/Quad/Hex/Octa hoặc cao hơn, máy tính PC hoặc latop bộ xử lý 64 bit( xung tối thiểu 2.5GHz) * HDD: 1-4TB * 64-512GB RAM * 10 Gigabit Ethernet hoặc Nonded Gigabit Ethernet. |
| Phiên bản Hadoop ổn định | * Apache Hadoop 3.2.1 * Link: <https://hadoop.apache.org/releases.html> |
| * Hệ sinh thái Hadoop | * [**HDFS**](https://www.edureka.co/blog/hadoop-ecosystem#hdfs) * [**YARN**](https://www.edureka.co/blog/hadoop-ecosystem#yarn) * [**MapReduce**](https://www.edureka.co/blog/hadoop-ecosystem#mapreduce)-> Data processing using programming * [**Spark**](https://www.edureka.co/blog/hadoop-ecosystem#apache_spark)-> In-memory Data Processing * [**PIG**](https://www.edureka.co/blog/hadoop-ecosystem#apache_pig)**,**[**HIVE**](https://www.edureka.co/blog/hadoop-ecosystem#apache_hive)-> Data Processing Services using Query (SQL-like) * [**HBase**](https://www.edureka.co/blog/hadoop-ecosystem#apache_hbase)-> NoSQL Database * [**Mahout**](https://www.edureka.co/blog/hadoop-ecosystem#apache_mahout)**, Spark MLlib** -> Machine Learning * [**Apache Drill**](https://www.edureka.co/blog/hadoop-ecosystem#apache_drill) -> SQL on Hadoop * [**Zookeeper**](https://www.edureka.co/blog/hadoop-ecosystem#apache_zookeeper)-> Quản lý Cluster * [**Oozie**](https://www.edureka.co/blog/hadoop-ecosystem#apache_oozie)-> Lập lịch job * [**Flume**](https://www.edureka.co/blog/hadoop-ecosystem#apache_flume)**,**[**Sqoop**](https://www.edureka.co/blog/hadoop-ecosystem#apache_sqoop) -> Data Ingesting Services * [**Solr & Lucene**](https://www.edureka.co/blog/hadoop-ecosystem#apache_solr_and_lucene) -> Searching & Indexing * [**Ambari**](https://www.edureka.co/blog/hadoop-ecosystem#apache_ambari)-> Provision, Monitor and Maintain cluster |
|  |  |  |

# TÀI LIỆU THAM KHẢO

662-679. (2000). Towards robust distributed systems (abstract). *Proceedings of the Nineteenth Annual ACM Symposium on Principles of Distributed Computing (PODC '00)* (s. 7). New York: ACM.

Bhadani, A. a. (2010). Performance evaluation of web servers using central load balancing policy over virtual machines on cloud. *Proceedings of the Third Annual ACM Conference.* Bangalore: ACM.

Blount, M. E. (2010). Real-time analysis for intensive care: Development and deployment of the Artemis analytic system. *IEEE Engineering in Medicine and Biology Magazine*, 110-118.

Boyd, D. C. (2011). ix provocations for big data. In: A Decade in Internet Time. *Symposium on the Dynamics of the Internet and Society.*

Boyd, D. C. (2012). Critical questions for big data. Information. *Communication & Society*, 662-679.

Brown, B. C. (2015, Jan 20). *Are you ready for the era of big data?McKinsey Quarterly*. Hentet fra McKinsey Quarterly: http://www.mckinsey.com/insights/strategy/are you ready for the era of big data,

Chen, M. M. (2014). Big data: A survey. *Mobile Networks and Applications 19 (2)*, 171-209.

Coumaros, J. d. (2015, March 20). *Big data alchemy: How can banks maximize the value of their customer data?* Hentet fra Capgemini Consulting: https://www.capgemini.com/resources/big-data-customer-analytics-in-banks

Cukier, K. (2015, June 15). *Data, data everywhere: A special report on managing information*. Hentet fra The Economist: http://www.economist.com/node/15557443

Cuzzocrea, A. S.-Y. (2011). Analytics over large-scale multidimensional data: the big data revolution! *Proceedings of the ACM 14th International workshop on Data Warehousing and OLAP (DOLAP '11)* (ss. 101-104). New York: ACM.

Facebook. (2016, February 20). Hentet fra http://www.statisticbrain.com/facebook-statistics/

Fisher, D. D. (2012). Interactions with big data analytics. *Interactions 19 (3)*, 50-59.

Forsyth, C. (2012, January). *For big data analytics there‟s no such thing as too big*. Hentet fra Cisco White paper: http://www.cisco.com/en/US/solutions/ns340/ns517/ns224/big data wp.pdf

Foster, I. Z. (2008). Cloud computing and grid computing 360-degree compared. *Grid Computing Environments Workshop, 2008 (GCE '08)*, (ss. 1-10).

Gandomi, A. H. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management 35 (2)*, 137-144.

Gantz, J. R. (2011). *Extracting value from chaos.* IDC.

Garlasu, D. S. (2013). A big data implementation based on grid computing. *11th Roedunet International Conference (RoEduNet)*, (ss. 1-4). Sinaia.

Gilbert, S. L. (2002). Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. *SIGACT News 33 (2)*, (ss. 51-59).

Jin, X. W. (2015). Significance and challenges of big data research. *Big Data Research 2 (2)*, 59-64.

Jothimani, D. B. (2015). Towards Understanding the Cynicism of Social Networking Sites: An Operations Management Perspective. *Procedia - Social and Behavioral Sciences*, 117–132.

Kaisler, S. A. (2013). Big data: Issues and challenges moving forward. *46th Hawaii International Conference on System Sciences (HICSS)*, (ss. 995-1004). Hawaii.

Khan, N. Y. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges. *The Scientific World Journal*.

Laney, D. (2001). *3D data management: Controlling data volume, velocity and variety.* Tech. Rep. 949, META Group.

Li, H. L. (2014). Challenges and trends of big data analytics. *Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*, (ss. 566-567). Guangzhou China.

Maletic, J. M. (2000). Data cleansing: Beyond integrity analysis. *Proceedings of the Conference on Information Quality (IQ2000).*

V., D. (2013). Data science and prediction. *Communications of ACM*, 64–73.