**TRƯỜNG ĐH CÔNG NGHỆ THÔNG TIN**

**KHOA CÔNG NGHỆ PHẦN MỀM**

-----🙞🙜🕮🙞🙜-----

****

**BÁO CÁO**

**Đề tài:**

**Xây dựng hệ thống recommendation**

**dựa trên Content base filtering**

**trên nền tảng Big Data với**

**Hadoop Mapreduce**

Giảng viên hướng dẫn:

Thạc sĩ: Phạm Thi Vương

Sinh viên thực hiện:

12520053 Nguyễn Viết Danh

Mục lục

[Chương 1 TÌM HIỂU VỀ BIG DATA 3](#_Toc438095496)

[1.1 Big data 3](#_Toc438095497)

[1.1.1 Giới thiệu 3](#_Toc438095498)

[1.1.2 Định nghĩa “big data” 4](#_Toc438095499)

[1.1.3 Một số đặc tính của big data 5](#_Toc438095500)

[1.2 Ứng dụng 6](#_Toc438095501)

[1.2.1 Chính phủ 6](#_Toc438095502)

[1.2.2 Phát triển quốc tế 7](#_Toc438095503)

[1.2.3 Sản xuất 7](#_Toc438095504)

[1.2.4 Chăm sóc sức khỏe 8](#_Toc438095505)

[1.2.5 Internet of Things (IoT) 8](#_Toc438095506)

[1.2.6 Công nghệ 8](#_Toc438095507)

[1.2.7 Bán lẻ 9](#_Toc438095508)

[1.2.8 Ngân hàng bán lẻ 9](#_Toc438095509)

[1.2.9 Bất động sản 9](#_Toc438095510)

[1.2.10 Khoa học và nghiên cứu 9](#_Toc438095511)

[1.2.11 Thể thao 10](#_Toc438095512)

[1.3 Các công cụ 10](#_Toc438095513)

[1.3.1 Hadoop 10](#_Toc438095514)

[1.3.2 MapReduce 11](#_Toc438095515)

[1.3.3 GridGain 11](#_Toc438095516)

[1.3.4 HPCC 11](#_Toc438095517)

[1.3.5 Storm 11](#_Toc438095518)

[1.3.6 Cassandra 11](#_Toc438095519)

[1.3.7 Hbase 12](#_Toc438095520)

[1.3.8 MongoDB 12](#_Toc438095521)

[1.3.9 Neo4j 12](#_Toc438095522)

[1.3.10 CouchDB 12](#_Toc438095523)

[Chương 2 TÀI LIỆU THAM KHẢO 13](#_Toc438095524)

# TÌM HIỂU VỀ BIG DATA

## Big data

### Giới thiệu

Big data là một thuật ngữ rộng cho bộ dữ liệu lớn hay phức tạp mà các ứng dụng xử lý dữ liệu truyền thống là không đủ để thực hiện. Những thách thức bao gồm phân tích, nắm bắt, quản lý dữ liệu, tìm kiếm, chia sẻ, lưu trữ, truyền, trực quan hóa, và bảo mật thông tin. Thuật ngữ này thường dùng chỉ đơn giản để sử dụng các phân tích hoặc các phương pháp tiên tiến khác để thu được chính xác giá trị từ dữ liệu và hiếm khi tập dữ liệu đó có kích thước cụ thể. Độ chính xác trong big data có thể hỗ trợ cho việc ra quyết định tự tin. Và quyết định tốt hơn có thể có nghĩa là hiệu quả hoạt động cao hơn, giảm chi phí và rủi ro.

Phân tích các tập dữ liệu có thể tìm thấy mối tương quan mới, để tìm ra xu hướng kinh doanh, ngăn ngừa bệnh tật, chống tội phạm,... Các nhà khoa học, doanh nhân, học viên truyền thông, quảng cáo cũng như chính phủ thường xuyên gặp khó khăn với các tập dữ liệu lớn trong các lĩnh vực bao gồm tìm kiếm trên Internet, tài chính và kinh doanh qua mạng. Các nhà khoa học gặp phải những hạn chế trong công tác e-Science, bao gồm khí tượng, gen, sự kết nối, mô phỏng vật lý phức tạp, nghiên cứu sinh học và môi trường.

Bộ dữ liệu có kích thước tăng dần trong một phần nào đó vì chúng đang được thu thập bởi các thiết bị giá rẻ ngày qua ngày và rất nhiều thông tin từ điện thoại di động, cảm biến từ xa, nhật ký phần mềm, máy ảnh, micro, thiết bị thu sóng radio (RFID) và các mạng cảm biến không dây. Hiện nay, khả năng lưu trữ thông tin bình quân đầu người gần gấp đôi 40 tháng kể từ những năm 1980. Như năm 2012, mỗi ngày có 2,5 exabyte (2,5 × 1018) dữ liệu được tạo ra. Những thách thức đối với các doanh nghiệp lớn là ai sẽ sở hữu các phát dữ liệu lớn để có thể giúp tổ chức bám trụ.

Làm việc với big data nhất thiết phải là không phổ biến. Hầu hết các phân tích là trên kích thước dữ liệu của máy tính cá nhân, máy tính để bàn hay máy xách tay. Đó là việc xử lý các tập dữ liệu có sẵn.

Các hệ thống quản lý cơ sở dữ liệu quan hệ, việc thống kê số liệu trên máy tính để bàn và các gói trực quan hóa thường gặp khó khăn trong việc xử lý big data. Công việc thay vào đó đòi hỏi nhiều phần mềm chạy song song trên hàng chục, hàng trăm, thậm chí hàng nghìn máy chủ. Những gì được coi là "big data" khác nhau tùy thuộc vào khả năng của người dùng và các công cụ họ sử dụng và khả năng mở rộng big data theo mục tiêu nhất định. Vì vậy, những gì được coi là "big" sẽ trở nên bình thường sau đó. Đối với một số tổ chức, họ phải đối mặt với hàng trăm gigabyte (GB) dữ liệu lần đầu tiên có thể làm họ cần phải xem xét lại phương án quản lý dữ liệu của mình. Đối với những tổ chức khác, nó có thể mất hàng chục hoặc hàng trăm terabyte (TB) trước khi kích thước dữ liệu trở thành một vấn đề quan trọng cần xem xét.

### Định nghĩa “big data”

Big data thường bao gồm các bộ dữ liệu với kích thước vượt quá khả năng của các công cụ phần mềm thường được sử dụng để nắm bắt, tổ chức, quản lý và xử lý dữ liệu trong một thời gian cho phép. Kích thước của big data là một mục tiêu liên tục di chuyển, như năm 2012 dao động từ vài chục terabyte đến hàng petabyte dữ liệu. Big data là một tập hợp các kỹ thuật và công nghệ đòi hỏi phải có các hình thức tích hợp mới để khám phá các giá trị ẩn đằng sau những tập dữ liệu lớn đa dạng, phức tạp và quy mô đồ sộ.

Trong một báo cáo nghiên cứu năm 2001 và các bài giảng liên quan, nhà phân tích Doug Laney, thành viên của META Group (nay là Gartner) đã xác định những thách thức của tăng trưởng dữ liệu và cơ hội là ba chiều, tức là tăng khối lượng (volume) (số lượng dữ liệu), vận tốc (velocity) (tốc độ dữ liệu vào và ra ) và sự đa dạng (variety) (nhiều loại dữ liệu và nguồn). Gartner và bây giờ nhiều ngành công nghiệp, tiếp tục sử dụng mô hình "3Vs" để mô tả dữ liệu lớn. Trong năm 2012, Gartner đã cập nhật lại định nghĩa của nó như sau: "Big data là số lượng lớn, tốc độ cao và đa dạng tài sản thông tin đòi hỏi những hình thức mới của xử lý để cho phép nâng cao việc ra quyết định, tạo cái nhìn sâu sắc và quá trình tối ưu hóa." Ngoài ra, sự đa dạng (veracity) được thêm vào bởi một vài tổ chức để mô tả khái niệm big data.

Định nghĩa 3Vs của Gartner vẫn được sử dụng rộng rãi và trong một thỏa thuận về định nghĩa thì "Big Data đại diện cho các tài sản thông tin đặc trưng bởi một lượng lớn dữ liệu (high volume), tốc độ xử lý (velocity) và đa dạng dữ liệu (variety) để đòi hỏi công nghệ và phương pháp phân tích đặc trưng để khai thác giá trị (value) của nó". Khái niệm 3Vs đã được mở rộng với đặc điểm bổ sung khác của big data:

* Volume (khối lượng): dữ liệu với khối lượng lớn và không phải là mẫu thử. Ở đây chú trọng vào việc quan sát và theo dõi những gì xảy ra.
* Velocity (tốc độ): big data thường được xử lý theo thời gian thực.
* Variety (sự đa dạng): big data rút ra từ văn bản, hình ảnh, âm thanh, video kết hợp với việc tổng hợp các dữ liệu liên quan.
* Mechine Learning (máy học): big data thường là việc phát hiện mẫu.
* Digital footprint (ghi dấu kỹ thuật số): big data thường là sản phẩm miễn phí của việc tương tác số.

Các trưởng thành phát triển của khái niệm thúc đẩy một sự khác biệt âm thanh hơn giữa dữ liệu lớn và Business Intelligence, liên quan đến dữ liệu và sử dụng:

* Kinh doanh thông minh sử dụng thống kê mô tả với dữ liệu với mật độ thông tin cao để đo lường sự vật, phát hiện các xu hướng…
* Big data sử dụng số liệu thống kê quy nạp và khái niệm từ nhận dạng hệ thống phi tuyến để suy ra luật (hồi quy, các mối quan hệ phi tuyến và các hiệu ứng nhân quả) từ các tập dữ liệu lớn với mật độ thông tin thấp để biển diễn mối quan hệ, sự phụ thuộc và thực hiện những dự đoán về kết quả và hành vi.

Trong một bài báo hướng dẫn được xuất bản trong tập chí IEEE Access (IEEE Access Journal), các tác giả đã phân loại các định nghĩa hiện tại của big data thành ba loại, cụ thể là: định nghĩa theo thuộc tính (Attribute Definition), định nghĩa so sánh (Comparative Definition) và định nghĩa theo kiến trúc (Architectural Definition). Các tác giả cũng trình bày một bản đồ công nghệ về big data để minh họa sự phát triển các công nghệ chủ chốt cho big data.

### Một số đặc tính của big data

Big data có thể được mô tả bởi các thuộc tính sau:

* **Khối lượng** (Volume): số lượng dữ liệu được tạo ra là quan trọng trong bối cảnh này. Kích thước của dữ liệu xác định giá trị và tiềm năng của các dữ liệu được xem xét. Cái tên 'big data ", bản thân nó đã chứa một thuật ngữ liên quan đến kích thước, và do đó nó là một đặc trưng của big data.
* **Sự đa dạng** (Variety): các loại nội dung và một thực tế cần thiết mà các nhà phân tích dữ liệu phải biết. Điều này giúp những người có liên quan và phân tích dữ liệu để sử dụng hiệu quả lợi thế của họ và sự đa dạng cũng rất quan trọng.
* **Tốc độ xử lý** (Velocity): trong bối cảnh này, tốc độ mà tại đó các dữ liệu được tạo ra và xử lý để đáp ứng nhu cầu và những thách thức trên con đường tăng trưởng và phát triển.
* **Sự biến đổi** (Variability): sự không đồng nhất các dữ liệu có thể hiển thị tại những thời điểm mà có thể cản trở quá trình xử lý và quản lý dữ liệu hiệu quả.
* **Độ chính xác** (Veracity): chất lượng của dữ liệu thu thập được, trong đó có thể khác nhau rất nhiều. Việc phân tích chính xác phụ thuộc vào tính xác thực của dữ liệu nguồn.
* **Độ phức tạp** (Complexity): quản lý dữ liệu có thể rất phức tạp, đặc biệt là khi một khối lượng lớn dữ liệu đến từ nhiều nguồn. Dữ liệu phải được liên kết, kết nối và tương quan để người dùng có thể nắm bắt các thông tin dữ liệu cần truyền đạt.

## Ứng dụng

Nhu cầu khai thác big data của các chuyên gia quản lý thông tin đang tăng lên rõ rệt. Các công ty như Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, HP và Dell đã dành hơn 15 tỷ USD vào các công ty phần mềm chuyên về quản lý dữ liệu và phân tích. Trong năm 2010, ngành công nghiệp này trị giá hơn 100 tỷ USD và đã được phát triển với gần 10% một năm - nhanh gấp 2 lần sự phát triển của các doanh nghiệp phần mềm.

Các nền kinh tế phát triển ngày càng sử dụng công nghệ dữ liệu chuyên sâu. Có 4,6 tỷ thuê bao điện thoại di động trên toàn thế giới, và khoảng 1 - 2 tỷ người truy cập internet. Từ năm 1990 đến năm 2005, hơn 1 tỷ người trên thế giới bước vào tầng lớp trung lưu, có nghĩa là nhiều người trở nên hiểu biết hơn, dẫn đến sự phát triển thông tin. Sự hiệu quả về trao đổi thông tin thông qua mạng viễn thông trên thế giới là 281 petabyte vào năm 1986, 471 petabyte trong năm 1993, 2.2 exabyte vào năm 2000, 65 exabyte trong năm 2007 và dự báo lượng truy cập internet lên đến 667 exabyte mỗi năm vào năm 2014. Theo một ước tính, một phần ba các thông tin được lưu trữ trên toàn cầu là ở dạng văn bản chữ và số, vẫn còn dữ liệu hình ảnh, đó là định dạng hữu ích nhất cho các ứng dụng big data. Điều này cũng cho thấy tiềm năng của dữ liệu chưa được sử dụng (trong các hình thức nội dung video và âm thanh).

Trong khi nhiều nhà sản xuất cung cấp giải pháp thương mại (off-the-shelf) cho Big Data, các chuyên gia khuyên nên phát triển các giải pháp tuỳ chỉnh riêng để giải quyết vấn đề của công ty nếu công ty có khả năng kỹ thuật đầy đủ.

### Chính phủ

Việc sử dụng và áp dụng big data trong các công việc của chính phủ là có lợi và hiệu quả về mặt chi phí, năng suất và sự đổi mới. Phân tích dữ liệu thường đòi hỏi nhiều bộ phận của chính quyền (trung ương và địa phương) để hợp tác làm việc và tạo ra các quy trình mới và sáng tạo để đạt được các kết quả mong muốn. Một số chính phủ đã ứng dụng big data như:

* Mỹ:
* Trong năm 2012, chính quyền Tổng thống Obama công bố sáng kiến nghiên cứu và phát triển Big Data, để khám phá làm thế nào big data có thể được sử dụng để giải quyết các vấn đề quan trọng mà chính phủ phải đối mặt. Sáng kiến này bao gồm 84 chương trình big data khác nhau trải rộng trên 6 phòng ban.
* Phân tích big data đóng một vai trò lớn trong chiến dịch tái tranh cử thành công năm 2012 của Barack Obama.
* Chính phủ Mỹ sở hữu 6 trong 10 siêu máy tính mạnh nhất thế giới.
* Trung tâm dữ liệu Utah là một trung tâm dữ liệu hiện đang được xây dựng bởi Cơ quan an ninh quốc gia Hoa Kỳ. Khi hoàn thành, trung tâm sẽ có thể xử lý một lượng lớn thông tin được thu thập bởi cơ quan an ninh quốc gia (National Security Agency) qua Internet. Số tiền chính xác của không gian lưu trữ là không rõ, nhưng các nguồn tin gần đây khẳng định nó sẽ được liệt vào khoảng một vài exabytes.
* Ấn Độ:
* Phân tích big data giúp ứng cử viên của đảng NDA giành chiến thắng trong cuộc Tổng tuyển cử Ấn Độ năm 2014.
* Chính phủ Ấn Độ sử dụng nhiều kỹ thuật để xác định làm thế nào các cử tri Ấn Độ đáp lại hành động của chính phủ, cũng như ý tưởng cải thiện chính sách.
* Anh: sử dụng big data cho dịch vụ công cộng.
* Dữ liệu về các loại thuốc kê toa: bằng cách kết nối nguồn gốc, địa điểm và thời gian của mỗi toa, một đơn vị nghiên cứu đã có thể để minh hoạ cho sự chậm trễ đáng kể giữa việc phát hành bất kỳ loại thuốc nào và sự tương thích với hướng dẫn của Viện Y tế Quốc gia và Chăm sóc sức khỏe Anh. Điều này giúp giảm thời gian cập nhật thuốc mới cho các bệnh nhân.

### Phát triển quốc tế

Nghiên cứu về việc sử dụng hiệu quả công nghệ thông tin và truyền thông cho sự phát triển cho thấy rằng công nghệ big data có thể có những đóng góp quan trọng nhưng cũng thách thức duy nhất hiện nay để phát triển quốc tế. Tiến bộ trong phân tích big data để cải thiện việc ra quyết định trong các lĩnh vực phát triển quan trọng như chăm sóc sức khỏe, việc làm, năng suất kinh tế, tội phạm, bảo mật, và thảm họa thiên nhiên và quản lý tài nguyên. Tuy nhiên, những thách thức lâu dài cho phát triển khu vực như cơ sở hạ tầng công nghệ ở mức trung bình và sự khan hiếm nguồn lực kinh tế và con người làm trầm trọng thêm mối quan tâm hiện tại với big data như sự riêng tư và vấn đề hội nhập.

### Sản xuất

Sự cải tiến trong việc lập kế hoạch cung ứng và cung cấp các sản phẩm chất lượng là lợi ích lớn nhất của big data cho sản xuất. Big data cung cấp một cơ sở hạ tầng cho sự minh bạch trong ngành công nghiệp sản xuất, đó là khả năng để làm sáng tỏ sự không chắc chắn như hiệu suất thành phần không phù hợp và tính sẵn sàng. Việc tiên đoán sản xuất như là một phương pháp áp dụng làm giảm thời gian chết và đòi hỏi số lượng lớn các dữ liệu và các công cụ dự báo tiên tiến cho một quá trình hệ thống các dữ liệu thành thông tin hữu ích. Một khung khái niệm về tiên đoán sản xuất bắt đầu với việc mua lại các loại dữ liệu khác nhau đến các dữ liệu có sẵn như âm thanh, sự rung động, áp suất, dòng điện, điện áp và dữ liệu điều khiển. Số lượng lớn các dữ liệu cảm giác trong các dữ liệu lịch sử xây dựng thành big data trong sản xuất. Hoạt động sản sinh big data là đầu vào cho các công cụ dự báo và phòng ngừa chiến lược như hệ thống quản lý dự báo và sức khỏe (Prognostics and Health Management).

### Chăm sóc sức khỏe

Phân tích big data đã giúp cải thiện việc chăm sóc sức khỏe bằng cách cung cấp thuốc cá nhân và quy tắc truyền lệnh, sự can thiệp nguy cơ lâm sàng và phân tích dự đoán, giảm lãng phí và chăm sóc đa dạng, báo cáo bên ngoài và nội bộ tự động cho các dữ liệu bệnh nhân, điều kiện y tế tiêu chuẩn và đăng ký bệnh nhân và giải pháp phân mảnh điểm.

### Internet of Things (IoT)

Để hiểu làm thế nào các phương tiện truyền thông sử dụng Big Data, điều này cần thiết đầu tiên để cung cấp một số bối cảnh vào những cơ chế được sử dụng cho quá trình truyền thông. Nó đã được đề xuất bởi Nick Couldry và Joseph Turow rằng các học viên tong đa phương tiện và quảng cáo tiếp cận big data như nhiều điểm hành động thông tin của hàng triệu cá nhân. Từ việc sử dụng các công cụ truyền thống như báo, tạp chí, hay chương trình truyền hình thì chuyển sang khai thác dữ liệu người tiêu dùng với các công nghệ, tiếp cận đối tượng mục tiêu vào những thời điểm tối ưu trong các địa điểm tối ưu. Mục đích cuối cùng là để phục vụ, hoặc truyền đạt, một tin nhắn hay những nội dung phù hợp với suy nghĩ người tiêu dùng. Ví dụ, môi trường xuất bản đang tăng cường chỉnh tin nhắn (quảng cáo) và nội dung (bài viết) để thu hút người tiêu dùng qua các hoạt động khai thác dữ liệu khác nhau.

Big Data và IoT làm việc cùng với sự liên kết. Từ góc độ phương tiện truyền thông, dữ liệu là các dẫn xuất chính của thiết bị liên kết nối và cho phép nhắm đến mục tiêu chính xác. Internet of Things, với sự giúp đỡ của big data, đã biến đổi ngành công nghiệp truyền thông đa phương tiện, các công ty và thậm chí cả các chính phủ, mở ra một kỷ nguyên mới của tăng trưởng kinh tế và khả năng cạnh tranh.

### Công nghệ

* **eBay.com** sử dụng hai kho dữ liệu ở mức 7.5 petabytes và 40PB cũng như một cụm 40PB Hadoop cho tìm kiếm, khuyến nghị người tiêu dùng và bán hàng.
* **Amazon.com** xử lý hàng triệu các hoạt động back-end mỗi ngày, cũng như truy vấn từ hơn nửa triệu người bán hàng của bên thứ ba. Các công nghệ cốt lõi giúp cho Amazon chạy ổn định là dựa trên Linux và đến năm 2005 họ đã có ba cơ sở dữ liệu Linux lớn nhất thế giới, với dung lượng 7.8 TB, 18.5 TB, và 24.7TB.
* **Facebook** xử lý 50 tỷ bức ảnh từ người dùng cơ sở của nó.
* Tính đến tháng 8 năm 2012, **Google** đã xử lý khoảng 100 tỷ lượt tìm kiếm mỗi tháng.

### Bán lẻ

Walmart (một cửa hàng bán lẻ nổi tiếng trên thế giới) xử lý hơn 1 triệu khách hàng giao dịch mỗi giờ, việc lưu thông tin vào cơ sở dữ liệu ước tính có hơn 2.5 petabytes (2560 terabytes) của dữ liệu tương đương với 167 lần các thông tin trong tất cả các sách trong Thư viện Quốc hội Hoa Kỳ.

### Ngân hàng bán lẻ

* Hệ thống phát hiện thẻ FICO bảo vệ tài khoản trên toàn thế giới.
* Khối lượng dữ liệu kinh doanh trên toàn thế giới, trên tất cả các công ty, ước tính tăng gấp đôi mỗi 1.2 năm.

### Bất động sản

Công ty bất động sảnWindermere sử dụng tín hiệu GPS vô danh từ gần 100 triệu lái xe để giúp người mua nhà mới xác định thời gian lái xe của họ đến nơi làm việc trong suốt các thời gian khác nhau trong ngày.

### Khoa học và nghiên cứu

* Khi Digital Sky Survey Sloan (SDSS) bắt đầu thu thập dữ liệu thiên văn vào năm 2000, nó tích lũy được nhiều hơn trong vài tuần đầu tiên của nó so với tất cả các dữ liệu thu thập được trong lịch sử thiên văn học trước đó. Tiếp tục với tốc độ khoảng 200 GB mỗi đêm, SDSS đã thu thập được hơn 140 terabytes thông tin. Đến lượt Large Synoptic Survey Telescope kế tục SDSS, sẽ hoạt động trong năm 2016, các nhà thiết kế mong đợi nó có thể thu thập được dữ liệu của mỗi năm ngày.
* Giải mã bộ gen của con người ban đầu mất 10 năm để xử lý, bây giờ nó có thể được xử lý trong vòng chưa đầy một ngày. Chi phí xác định trình tự DNA đã rẻ hơn 100 lần so với mức giảm chi phí dự đoán của định luật Moore.
* Trung tâm mô phỏng khí hậu NASA (NCCS) lưu 32 petabyte dữ liệu về quan sát và mô phỏng khí hậu trên các siêu máy tính cụm Discover.
* DNAStack của Google, biên soạn và tổ chức các mẫu DNA của dữ liệu di truyền từ khắp nơi trên thế giới để xác định bệnh và khuyết tật y tế khác. Những tính toán nhanh và chính xác loại bỏ những lỗi phát sinh khi tiến hành bởi con người cụ thể là các chuyên gia khoa học và sinh học làm việc với DNA. DNAStack, một phần của hệ gen Google (Google Genomics) cho phép các nhà khoa học sử dụng mẫu lớn các nguồn lực từ các máy chủ tìm kiếm của Google để mở rộng quy mô thí nghiệm mà thường sẽ phải mất nhiều năm mới có kết quả.

### Thể thao

Big data có thể được sử dụng để cải thiện đào tạo và tìm hiểu đối thủ. Bên cạnh đó, nó có thể dự đoán người chiến thắng trong một trận đấu bằng cách sử dụng phân tích big data. Hiệu suất trong tương lai của các cầu thủ có thể được tiên đoán một cách chính xác. Như vậy, giá trị và tiền lương của vận động viên được xác định bởi xem xét các dữ liệu thu thập trong suốt mùa giải.

MoneyBall cho thấy làm thế nào big data có thể được sử dụng để trinh sát các cầu thủ và cũng xác định được các cầu thủ bị đánh giá thấp.

Trong cuộc đua công thức một (Formula One races), chiếc xe đua với hàng trăm bộ cảm biến tạo ra hàng terabyte dữ liệu. Những cảm biến thu thập dữ liệu từ các điểm áp lực lốp xe đến hiệu suất đốt cháy nhiên liệu. Sau đó, dữ liệu này được chuyển giao cho tổng hành dinh Vương quốc Anh thông qua cáp sợi quang có thể truyền dữ liệu ở tốc độ của ánh sáng. Dựa trên các dữ liệu đó, các kỹ sư và các nhà phân tích dữ liệu quyết định các điều chỉnh để giành chiến thắng trong cuộc đua. Bên cạnh đó, sử dụng big data, các đội đua cố gắng dự đoán thời gian họ sẽ kết thúc cuộc đua, dựa trên mô phỏng bằng cách sử dụng dữ liệu thu thập được trong mùa giải.

## Các công cụ

### Hadoop

Hadoop - thư viện phần mềm được phát triển bởi Apache, là một framework cho phép phân phối các tập dữ liệu lớn trên các cụm máy tính sử dụng mô hình lập trình đơn giản. Nó được thiết kế để mở rộng quy mô từ các máy chủ đơn lẻ cho đến hàng ngàn máy chủ kết hợp với nhau, mỗi sản phẩm có tính toán và lưu trữ địa phương. Thay vì dựa vào phần cứng để cung cấp khả năng sẵn sàng ở mức cao, các thư viện được thiết kế để có thể tự phát hiện và xử lý lỗi ở lớp ứng dụng, do đó cung cấp một dịch vụ có tính sẵn sàng cao trên một cụm các máy tính hàng đầu.

Các dự án Apache™ Hadoop® phát triển phần mềm mã nguồn mở đáng tin cậy, có khả năng mở rộng và tính toán phân tán.

### MapReduce

Được phát triển bởi Google, các trang web MapReduce mô tả nó như là "một mô hình lập trình và framework phần mềm dùng để viết các ứng dụng xử lý nhanh một lượng lớn dữ liệu chạy song song trên các cụm nút tính toán lớn." Nó được sử dụng bởi Hadoop, cũng như rất nhiều các ứng dụng xử lý dữ liệu khác. Hệ điều hành: OS Independent.

### GridGain

GridGain cung cấp một sự thay thế cho Hadoop's MapReduce, tương thích với các hệ thống phân tán tập tin bằng Hadoop (Hadoop Distributed File System). Nó cung cấp tiến trình xử lý bộ nhớ trong để phân tích nhanh chóng các dữ liệu thời gian thực. Bạn có thể tải về các phiên bản mã nguồn mở của GridGrain từ GitHub hoặc mua một phiên bản hỗ trợ thương mại. Hệ điều hành: Windows, Linux, OS X.

### HPCC

Phát triển bởi LexisNexis Risk Solutions, HPCC là viết tắt của "high performance computing cluster" (cụm điện toán hiệu suất cao). Nó được cho là cung cấp hiệu suất vượt trội so với Hadoop. Cả hai phiên bản cộng đồng miễn phí và phiên bản doanh nghiệp đã có sẵn. Hệ điều hành: Linux.

### Storm

Storm cung cấp khả năng phân phối tính toán thời gian thực và thường được mô tả như là "Hadoop của thời gian thực" (hiện đang thuộc sở hữu của Twitter). Đó là khả năng mở rộng, mạnh mẽ, chịu lỗi và làm việc với gần như tất cả các ngôn ngữ lập trình. Hệ điều hành: Linux.

### Cassandra

Cassandra được phát triển bởi Facebook, cơ sở dữ liệu NoSQL này hiện đang được quản lý bởi Apache Foundation. Nó được sử dụng bởi nhiều tổ chức với các tập dữ liệu lớn đang hoạt động, bao gồm cả Netflix, Twitter, Urban Airship, Constant Contact, Reddit, Cisco và Digg. Hỗ trợ và dịch vụ thương mại có sẵn thông qua nhà cung cấp bên thứ ba. Hệ điều hành: OS Independent.

### Hbase

Là một dự án khác của Apache, HBase là kho dữ liệu không quan hệ cho Hadoop. Các tính năng bao gồm khả năng mở rộng mô-đun hóa và mở rộng tuyến tính, khả năng đọc và ghi thống nhất, hỗ trợ chuyển đổi dự phòng tự động và nhiều tính năng khác nữa. Hệ điều hành: OS độc lập.

### MongoDB

MongoDB được thiết kế để hỗ trợ cơ sở dữ liệu cực lớn. Đó là cơ sở dữ liệu NoSQL với tài liệu lưu trữ có định hướng, hỗ trợ chỉ mục đầy đủ, khả năng nhân rộng và tính sẵn sàng cao và nhiều tính năng khác nữa. Hệ điều hành: Windows, Linux, OS X, Solaris.

### Neo4j

Là cơ sở dữ liệu đồ thị hàng đầu thế giới, Neo4j tự hào có những cải tiến hiệu suất lên đến 1000 lần hoặc nhiều hơn so với cơ sở dữ liệu quan hệ. Các tổ chức quan tâm có thể mua phiên bản tiên tiến hoặc doanh nghiệp từ Neo Technology. Hệ điều hành: Windows, Linux.

### CouchDB

Được thiết kế cho Web, CouchDB lưu trữ dữ liệu trong các tài liệu JSON mà người dùng có thể truy cập thông qua môi trường web hoặc truy vấn bằng cách sử dụng JavaScript. Nó cung cấp khả năng mở rộng phân tán với khả năng lưu trữ chịu lỗi. Hệ điều hành: Windows, Linux, OS X, Android.

# TÀI LIỆU THAM KHẢO

Wikipedia

<https://en.wikipedia.org/wiki/Big_data>

50 Top Open Source Tools for Big Data

<http://www.datamation.com/data-center/50-top-open-source-tools-for-big-data-1.html>