Mục lục

DANH SÁCH HÌNH VẼ

LỜI MỞ ĐẦU

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. Trong kỷ nguyên của IoT (Internet of Things), 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. Theo một báo cáo của IDC, năm 2011, lượng dữ liệu được tạo ra trên thế giới là 1.8ZB , tăng gần 9 lần chỉ trong 5 năm. 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, được thu thập từ nhiều nguồn khác nhau.

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.

Mục tiêu của nghiên cứu này nhằm đưa cái nhìn toàn cảnh về Big Data đồng thời nhấn mạnh vào 2 vấn đề là xu hướng phát triển của công nghệ Big Data và ảnh hưởng của nó đến phát triển kinh tế xã hội và quản lý nhà nước.

Bên cạnh các nghiên cứu cơ bản, đề tài cũng tập trung vào nghiên cứu các công nghệ nền tảng để xây dựng các ứng dụng xử lý dữ liệu lớn (tập trung vào Apache Hadoop). Thêm vào đó, nhóm đề tài cũng thực hiện ứng dụng thử nghiệm nền tảng này trong việc xử lý dữ liệu ảnh văn bản. Việc xây dựng ứng dụng thực tế này vừa giúp nhóm đề tài nắm bắt được kỹ thuật, công nghệ nền tảng, vừa ứng dụng vào nhu cầu thực tế của Viện CNPM & NDS và gắn liền với nhiệm vụ về Kho dữ liệu của Viện.

Cuối cùng, đề tài đưa ra một số đề xuất về các ứng dụng dữ liệu lớn nên được triển khai và phân tích một số nền tảng công nghệ xử lý dữ liệu lớn để có những đánh giá và lựa chọn phù hợp.

CHƯƠNG 1: TỔNG QUAN VỀ DỮ LIỆU LỚN

1.1. Mở đầu

Một nửa thế kỷ sau khi máy tính bước vào xã hội chính thống, dữ liệu bắt đầu được tích lũy nhiều tới mức mà một điều gì đó mới mẻ và đặc biệt sắp xảy ra. Không những thế giới tràn ngập thông tin nhiều hơn bao giờ hết, mà thông tin còn tăng nhanh hơn. Sự thay đổi về quy mô đã dẫn đến một sự thay đổi về trạng thái. Thay đổi về lượng đã dẫn tới thay đổi về chất. Các khoa học như thiên văn, gen, mới được trải nghiệm sự bùng nổ trong những năm 2000, đã đưa ra thuật ngữ "dữ liệu lớn", khái niệm mà nay đã di trú vào tất cả các lĩnh vực của đời sống con người.

Không có một định nghĩa chính xác cho dữ liệu lớn. Ban đầu ý tưởng là dung lượng thông tin đã tăng quá lớn tới mức số lượng cần khảo sát không còn vừa vào bộ nhớ các máy tính dùng để xử lý, do vậy các kỹ sư cần cải tạo các công cụ họ dung để có thể phân tích được tất cả thông tin. Đó là xuất xứ của các công nghệ xử lý mới như MapReduce của Google và nguồn mở tương đương của nó, Hadoop, khởi đầu từ Yahoo. Những công nghệ này cho phép ta quản lý những khối lượng dữ liệu lớn hơn nhiều so với trước đây, và quan trọng là không cần đưa dữ liệu vào các hàng ngăn nắp hoặc các bảng cơ sở dữ liệu cổ điển. Các công nghệ nghiền dữ liệu khác, bỏ qua các cấu trúc phân cấp và đồng nhất cứng nhắc cổ điển, cũng ở trong tầm ngắm. Đồng thời, do các công ty Internet có thể thu thập được vô số dữ liệu quý giá và có động cơ kinh tế lớn để khai thác chúng, nên các công ty này trở thành người sử dụng hàng đầu của các công nghệ xử lý hiện đại nhất, vượt qua các công ty truyền thống, đôi khi có tới hàng chục năm kinh nghiệm nhiều hơn.

Dữ liệu lớn đề cập tới những thứ người ta có thể làm với một quy mô lớn mà không thể làm với một quy mô nhỏ hơn, để trích xuất những hiểu biết mới hoặc tạo ra những dạng giá trị mới, theo những cách thức có thể làm thay đổi các thị trường, các tổ chức, mối quan hệ giữa các công dân và các chính phủ, và hơn thế nữa.

Nhưng đó chỉ là bước khởi đầu. Thời đại của dữ liệu lớn thách thức cách chúng ta sống và tương tác với thế giới. Nổi bật nhất, xã hội sẽ phải cắt giảm một số nỗi ám ảnh của nó về quan hệ nhân quả để đổi lấy mối tương quan đơn giản, không biết *tại sao* mà chỉ biết *cái gì*. Điều đó làm đổ vỡ hàng thế kỷ các tập quán đã được thiết lập và thách thức hiểu biết cơ bản nhất của chúng ta về việc làm thế nào để đưa ra được quyết định và hiểu được thực tế.

Dữ liệu lớn đánh dấu bước khởi đầu của một biến đổi lớn. Đúng như kính thiên văn tạo điều kiện cho chúng ta hiểu biết được vũ trụ và kính hiển vi cho phép chúng ta hiểu biết được vi trùng, các kỹ thuật mới để thu thập và phân tích những tập hợp lớn dữ liệu sẽ giúp chúng ta tìm ra ý nghĩa của thế giới theo những cách thức mà chúng ta mới chỉ vừa bắt đầu ưa thích.

Cuộc cách mạng thật sự không phải ở những chiếc máy tính toán dữ liệu mà ở chính dữ liệu và cách ta sử dụng chúng. Để đánh giá mức độ một cuộc cách mạng thông tin đã tiến triển tới đâu, ta hãy xem xét các xu hướng xuyên xuốt các lĩnh vực của xã hội. Lấy ví dụ thiên văn học. Khi Sloan Digital Sky Survey (SDSS – Trạm quan sát bầu trời bằng kỹ thuật số Sloan) bắt đầu hoạt động vào năm 2000, kính thiên văn của nó tại New Mexico trong mấy tuần đầu tiên đã thu thập nhiều dữ liệu hơn những gì được thu thập trong toàn bộ lịch sử của ngành thiên văn. Đến năm 2010, lưu trữ của trạm đã bạt ngàn với con số khổng lồ 140 tera (10 mũ 12) byte thông tin. Nhưng kẻ kế nhiệm, kính thiên văn của Large Synoptic Survey (LSST) ở Chile, dự kiến vận hành vào năm 2016, cứ mỗi năm ngày sẽ thu thập được lượng dữ liệu tương đương như thế.

Những số lượng vô cùng to lớn như vậy cũng có thể được tìm thấy ngay xung quanh chúng ta. Khi các nhà khoa học lần đầu giải mã gen người vào năm 2003, họ đã mất một thập kỷ làm việc miệt mài để xác định trình tự cho ba tỷ cặp cơ sở. Bây giờ, sau một thập kỷ, một thiết bị đơn lẻ cũng có thể xác định trình tự cho số lượng DNA như vậy chỉ trong một ngày.

Trong ngành tài chính, khoảng 7 tỷ cổ phiếu được mua bán mỗi ngày trên các thị trường chứng khoán Mỹ, trong số đó khoảng hai phần ba được giao dịch bằng các thuật toán máy tính dựa trên các mô hình toán học xử lý hàng núi dữ liệu để dự đoán lợi nhuận trong khi cố gắng giảm thiểu rủi ro.

Các công ty Internet đặc biệt bị tràn ngập. Google xử lý hơn 24 peta (10 mũ 15) byte dữ liệu mỗi ngày, một khối lượng gấp hàng ngàn lần tất cả các ấn phẩm trong Thư viện Quốc hội Mỹ. Facebook, một công ty không hề tồn tại một thập kỷ trước, nhận hơn 10 triệu ảnh mới được tải lên mỗi giờ. Các thành viên Facebook nhấp nút "like" hoặc gửi lời bình luận gần ba tỷ lần mỗi ngày, tạo một dấu vết số để công ty có thể "đào xới" nhằm biết được các sở thích của người sử dụng. Trong khi đó, 800 triệu người sử dụng dịch vụ Youtube của Google tải lên hơn một giờ video mỗi giây. Thành viên của mạng Twitter tăng khoảng 200 phần trăm mỗi năm và đến năm 2012 đã có hơn 400 triệu *tweet* mỗi ngày.

Từ khoa học tới y tế, từ ngân hàng tới Internet, các lĩnh vực có thể khác nhau,

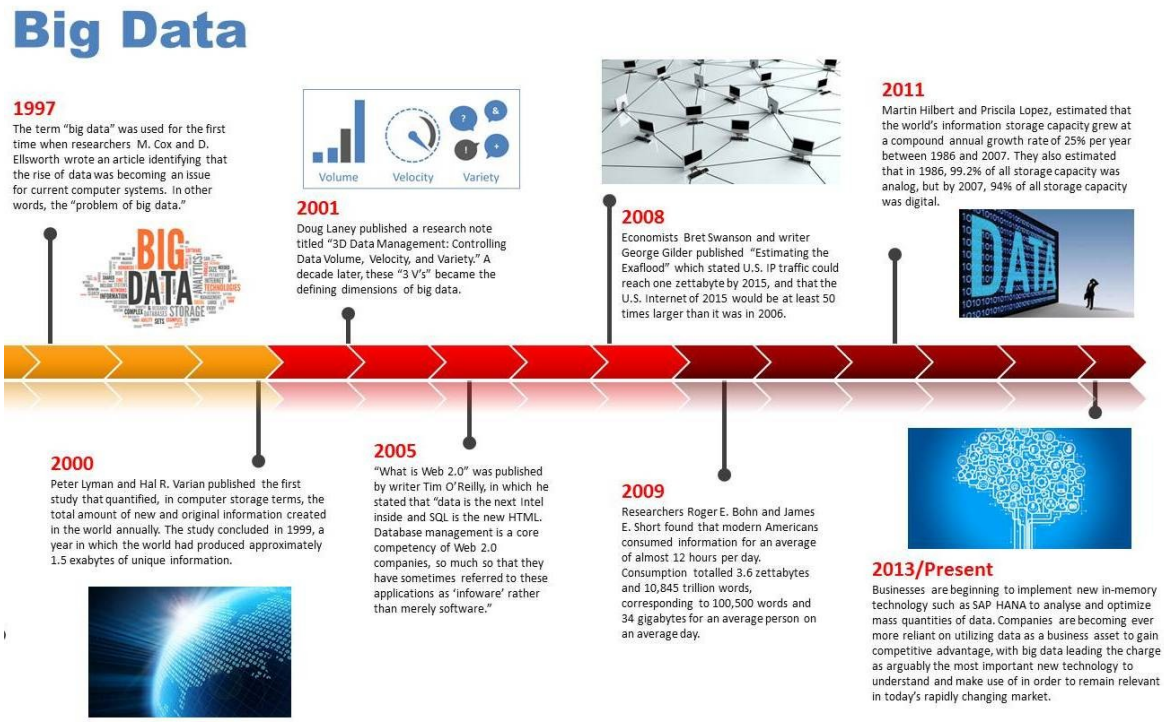
nhưng cùng nhau chúng đều có một câu chuyện tương tự: số lượng dữ liệu trong thế giới đang tăng rất nhanh, vượt sức không chỉ những chiếc máy tính mà cả trí tưởng tượng của chúng ta.

Nhiều người đã thử đưa ra một con số thực tế về lượng thong tin xung quanh chúng ta và tính toán xem nó tăng như thế nào. Họ đã có những mức độ thành công khác nhau bởi họ đo lường những thứ khác nhau.. Một trong những nghiên cứu toàn diện hơn được Martin Hilbert của Trường Truyền thông và Báo chí Annenberg thuộc Đại học Nam California thực hiện. Ông đã nỗ lực đưa ra một con số cho mọi thứ đã từng được sản xuất, lưu trữ và truyền tải. Chúng không chỉ bao gồm sách, tranh, email, ảnh, nhạc, và phim (cả dạng analog và digital), mà còn có trò chơi điện tử, cuộc gọi điện thoại, thậm chí các hệ thống điều hướng xe và thư gửi qua bưu điện. Ông cũng bao gồm các phương tiện truyền thông phát sóng như truyền hình và radio, dựa trên tiếp cận khán giả. Theo ước lượng của Hilbert, hơn 300 exa (10 mũ 18) byte dữ liệu lưu trữ đã tồn tại vào năm 2007. Để dễ hình dung ý nghĩa của nó, thử nghĩ thế này. Một bộ phim dài ở dạng kỹ thuật số có thể được nén vào một tập tin 1 giga byte. Một exa byte là 1 tỷgiga byte. Tóm lại là vô cùng nhiều. Điều thú vị là năm 2007 chỉ khoảng 7 phần trăm dữ liệu ở dạng analog (giấy, sách, ảnh in,vân vân). Phần còn lại là ở dạng digital – kỹ thuật số. Nhưng mới gần đây, bức tranh đã rất khác. Mặc dù những ý tưởng của cuộc"cách mạng thông tin" và "thời đại kỹ thuật số" đã xuất hiện từ những năm 1960, chúng mới chỉ trở thành hiện thực ở vài khía cạnh. Tới tận năm 2000, mới chỉ có một phần tư thông tin lưu trữ của thế giới được số hóa. Ba phần tư còn lại vẫn ở trên giấy,phim, đĩa nhựa, băng từ, và những thứ tương tự. Lượng thông tin kỹ thuật số lúc đó chưa nhiều. Nhưng vì dữ liệu kỹ thuật số phát triển rất nhanh – cứ hơn ba năm lại tăng gấp đôi, theo Hilbert – nên tình hình đã nhanh chóng tự đảo ngược. Thông tin analog, ngược lại, không hề tăng. Do vậy vào năm 2013 lượng thông tin lưu trữ trong thế giới ước lượng khoảng 1.200 exa byte, trong đó chưa đến 2 phần trăm là phi kỹ thuật số.

Chẳng có cách nào phù hợp để hình dung kích thước như vậy của dữ liệu là có ý nghĩa gì. Nếu tất cả được in thành sách, chúng có thể phủ kín bề mặt của nước Mỹ với chiều dày 52 lớp. Nếu được ghi vào CD-ROM và xếp chồng lên nhau, chúng có thể tạo thành 5 cột vươn cao tới mặt trăng. Vào thế kỷ thứ ba trước Công nguyên, khi Ptolemy II của Ai Cập cố gắng lưu trữ một bản của mỗi tác phẩm từng được viết ra, Thư viện lớn của Alexandria đã tượng trưng cho toàn bộ tri thức của thế giới. Trận lũ lớn kỹ thuật số hiện đang quét qua trái đất tương đươngvới việc cung cấp cho mỗi người sống trên trái đất hôm nay 320 lần nhiều hơn thông tin như ước lượng đã được lưu trữ ở Thư viện Alexandria.

1.2. Lược sử về sự hình thành Dữ liệu lớn

Tốc độ bùng nổ thông tin (thuật ngữ được sử dụng lần đầu tiên năm 1941, theo The Oxford English Dictionary) buộc con người phải có những đánh giá về kích thước dữ liệu cũng như những đổi mới cơ bản trong ý tưởng xây dựng các ứng dụng có liên quan đến dữ liệu. Sự hình thành thuật ngữ Dữ liệu lớn được ghi nhận lần đầu tiên trong báo cáo của Michael Cox và David Ellsworth vào tháng 10 năm 1997 trình bày trong bài viết "Application-controlled demand paging for out-of-core visualization" tại Hội nghị IEEE lần thứ 8.



Hình 1.1: Lược sử về sự hình thành Dữ liệu lớn – Nguồn Internet

Tháng 8 năm 1999 Steve Bryson, David Kenwright, Michael Cox, David Ellsworth, và Robert Haimes xuất bản "Visually exploring gigabyte data sets in real time" trên tờ Communications of the ACM. Đây là bài viết CACM đầu tiên sử dụng thuật ngữ "Big Data" (tên của một trong những phần của bài viết là "Big Data for Scientific Visualization"). Bài báo mở đầu bằng nhận định: "Những chiếc máy tính mạnh là lợi thế cho việc khảo sát nhiều lĩnh vực, cũng có thể là bất lợi; tính toán nhanh chóng tạo ra một lượng lớn dữ liệu. Nếu trước kia bộ dữ liệu megabyte đã từng được coi là lớn, thì bây giờ chúng ta có thể tìm thấy những bộ dữ liệu của cá nhân vào khoảng 300GB. Tuy nhiên hiểu biết các dữ liệu thu được từ tính toán cao cấp là một nỗ lực đáng kể. Nhiều nhà khoa

học cho biết khó khăn xuất hiện khi xem xét tất cả các con số. Còn theo Richard W. Hamming, nhà toán học và cũng là người tiên phong trong lĩnh vực khoa học máy tính, lại chỉ ra rằng mục đích của máy tính là thấu hiểu sự vật, chứ không phải chỉ dừng lại ở các con số".

Tháng 10 năm 1999, Bryson, Kenwright và Haimes cùng với David Bank, Robert van Liere, và Sam Uselton trình bày báo cáo "Automation or interaction: what's best for big data?" tại hội nghị IEEE năm 1999.

Tháng 11 năm 2000, Francis X. Diebold trình bày với Đại hội Thế giới lần thứ VIII của Hiệp hội kinh tế lượng một tài liệu có tiêu đề "Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting". Trong đó ông khẳng định rằng: "Gần đây, nhiều ngành khoa học, như vật lý, sinh học, khoa học xã hội, vốn đang buộc phải đương đầu với khó khăn – đã thu được lợi từ hiện tượng Big Data và đã gặt hái được nhiều thành công. Big Data chỉ sự bùng nổ về số lượng (và đôi khi, chất lượng), khả năng liên kết cũng như độ sẵn sàng của dữ liệu, chủ yếu là kết quả của những tiến bộ gần đây và chưa từng có trong việc ghi lại dữ liệu và công nghệ lưu trữ".

Tháng 2 năm 2001, Doug Laney, một nhà phân tích của Tập đoàn Meta, công bố một nghiên cứu có tiêu đề "3D Data Managment: controlling Data Volume, Velocity, and Variety". Một thập kỷ sau, "3Vs" đã trở thành thuật ngữ được chấp nhận rộng rãi trong xác định dữ liệu lớn ba chiều, mặc dù thuật ngữ này không xuất hiện trong nghiên cứu của Laney.

Tháng 9 năm 2008, A special issue of Nature on Big Data nghiên cứu ý nghĩa của các bộ dữ liệu lớn đối với khoa học hiện đại.

Tháng 12 năm 2008, Randal E. Bryant, Randy H. Katz, và Edward D. Lazowska đưa ra bài viết "Big-Data Computing: Creating Revolutionary breakthroughs in Commerce, Science and Society", trong đó mô tả : "Cũng như công cụ tìm kiếm đã làm thay đổi cách chúng ta tiếp cận thông tin, các hình thức khác của sử dụng dữ liệu lớn có thể sẽ làm thay đổi cách hoạt động của các công ty, các nhà nghiên cứu khoa học, các học viên y tế, quốc phòng và tình báo của đất nước ta… Sử dụng dữ liệu lớn có lẽ là đổi mới lớn nhất trong công nghệ máy tính suốt một thập kỷ qua. Chúng tôi chỉ mới bắt đầu nhìn thấy tiềm năng của nó trong việc thu thập, sắp xếp và xử lý dữ liệu của tất cả các tầng lớp xã hội. Một khoản đầu tư dù khiêm tốn của chính phủ liên bang sẽ thúc đẩy phát triển và mở rộng nó. "

Tháng 2 năm 2010, Kenneth Cukier đăng trên tờ The Economist a Special Report

bài viết có tựa đề "Data, data everywhere". Cukier viết: "…thế giới chứa một số lượng thông tin số lớn đến mức không tưởng, và càng ngày càng được nhân rộng với tốc độ nhanh hơn bao giờ hết… Hiệu quả đã được thể hiện ở khắp mọi nơi, từ kinh doanh đến khoa học, từ chính phủ cho nghệ thuật. Các nhà khoa học và kỹ sư máy tính đã đặt ra một thuật ngữ mới cho hiện tượng này: Big Data".

Tháng 5 năm 2011, James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, và Angela Hưng Byers của Viện toàn cầu McKinsey công bố nghiên cứu "Big data: The next frontier for inovation, competition, and productivity". Trong nghiên cứu, họ tính toán rằng đến năm 2009, gần như tất cả các lĩnh vực trong nền kinh tế Mỹ đã đạt mức lưu trữ trung bình là 200 terabyte (gấp hai lần kích thước dữ liệu của nhà bán lẻ Mỹ Wal-Mart năm 1999) đối với công ty có hơn 1.000 nhân viên trong đó các chứng khoán và đầu tư khu vực dịch dẫn đầu về lượng dữ liệu lưu trữ. Tổng cộng, nghiên cứu ước tính rằng khối lượng lưu trữ là khoảng 7,4 exabyte đối với các doanh nghiệp và 6,8 exabyte đối với người tiêu dùng trong năm 2010.

Tháng 5 năm 2012, Danah Boyd và Kate Crawford đưa ra luận điểm của họ trong bài "Critical Question for Big Data" trên tờ Information, Communications and Society. Họ định nghĩa Big Data như là "một hiện tượng văn hóa, công nghệ và học thuật dựa trên sự tương tác của: (1) Công nghệ tối đa hóa sức mạnh tính toán và độ chính xác thuật toán để thu thập, phân tích, liên kết, và so sánh các tập dữ liệu lớn. (2) Phân tích: tạo ra trên dữ liệu lớn để xác định mô hình để làm cho tuyên bố kinh tế, xã hội, kỹ thuật và pháp lý. (3) Thần thoại: Niềm tin phổ biến rằng dữ liệu lớn cung cấp một hình thức cao hơn của trí thông minh và kiến thức có thể tạo ra mà những hiểu biết mà trước đây không thể, với hào quang của sự thật, khách quan, chính xác."

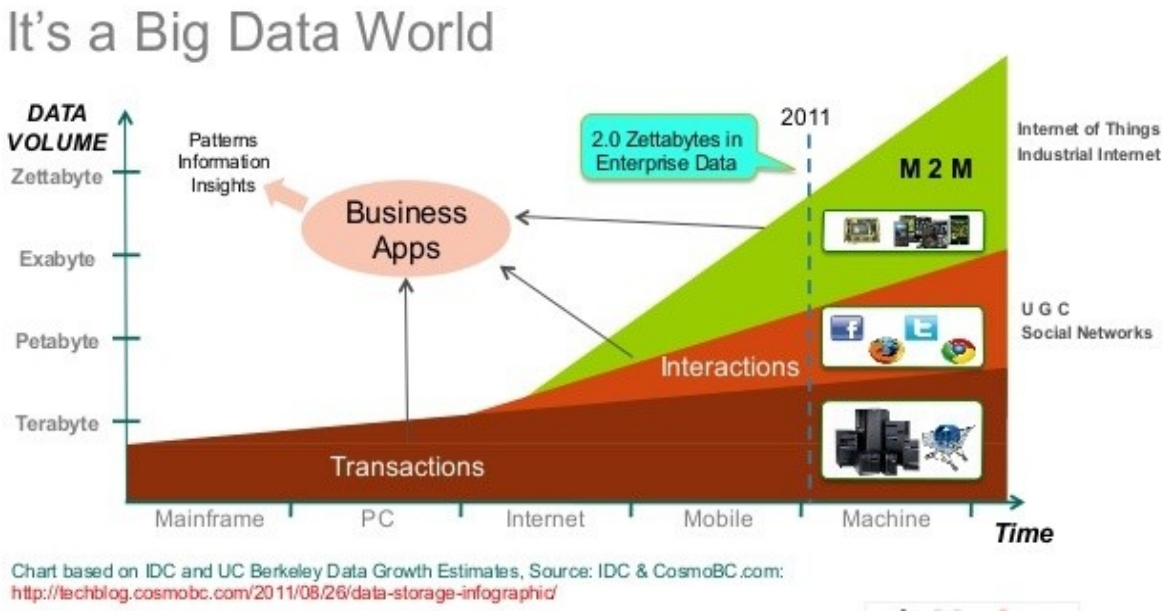
1.3. Định nghĩa về Dữ liệu lớn

Có nhiều định nghĩa về Dữ liệu lớn như của Forrester:

"Big Data is the frontier of a firm's ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers." -- Forrester

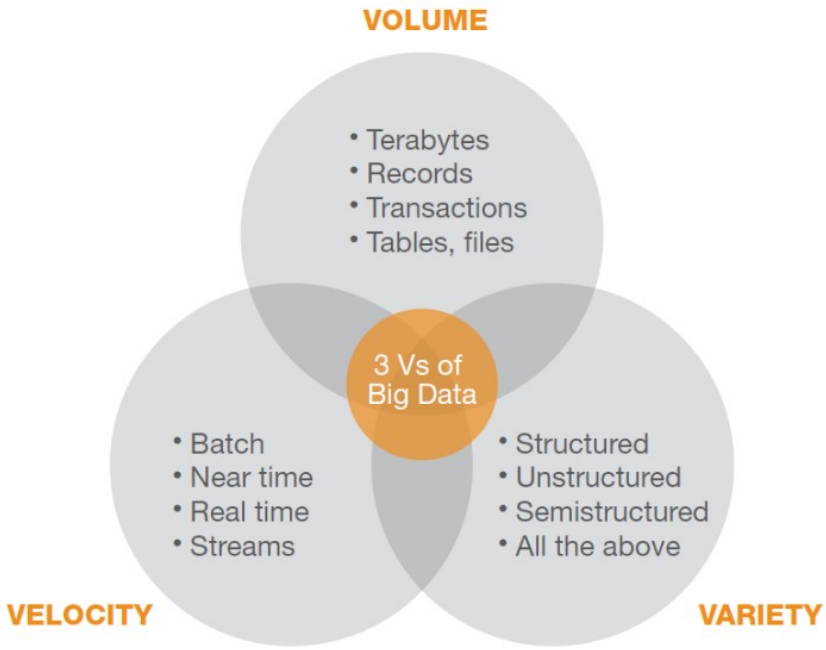
Nhưng định nghĩa để có thể đặc tả đúng nhất mà được nhiều nguồn trích dẫn nhất là của Gartner:

"Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" -- (Gartner 2012)



Hình 1.2 : Đồ thị về lượng dữ liệu được tạo ra trên thế giới năm 2011- Báo cáo IDC

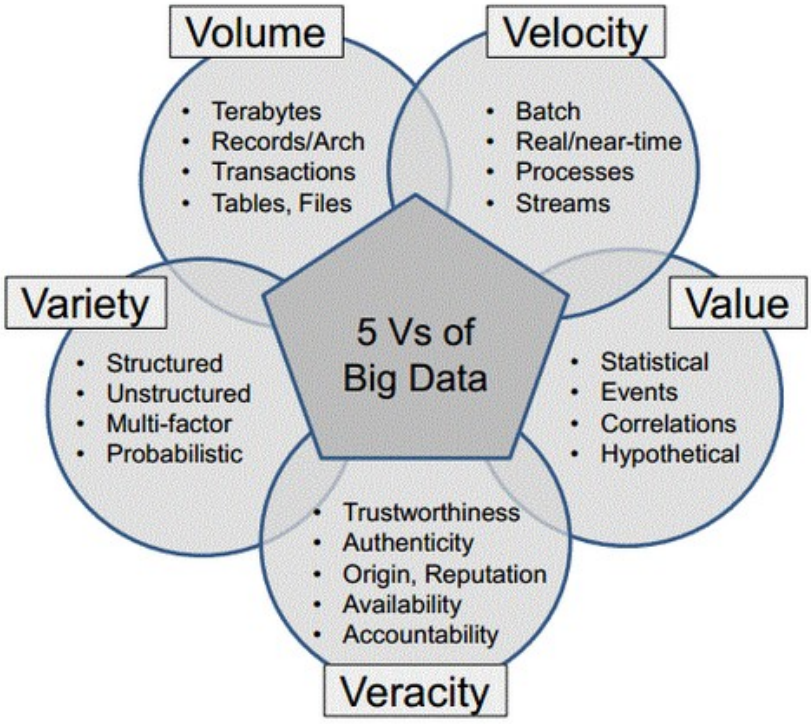
Trên thế giới có nhiều định nghĩa về Big Data. Vào năm 2001, nhà phân tích Doug Laney của hãng META Group (bây giờ chính là công ty nghiên cứu Gartner) đã nói rằng những thách thức và cơ hội nằm trong việc tăng trưởng dữ liệu có thể được mô tả bằng ba chiều "3V": tăng về số lượng lưu trữ (volume), tăng về tốc độ xử lý (velocity) và tăng về chủng loại (variety). Giờ đây, Gartner cùng với nhiều công ty và tổ chức khác trong lĩnh vực công nghệ thông tin tiếp tục sử dụng mô hình "3V" này để định nghĩa nên Big Data. Đến năm 2012, Gartner bổ sung thêm rằng Big Data ngoài ba tính chất trên thì còn phải "cần đến các dạng xử lí mới để giúp đỡ việc đưa ra quyết định, khám phá sâu vào sự vật/sự việc và tối ưu hóa các quy trình làm việc".



Hình 1.3: Mô hình "3Vs" của Big Data – Nguồn Internet

Dữ liệu lớn (Big Data) là khối lượng dữ liệu rất lớn được tạo ra từ mọi thứ xung quanh chúng ta, từ các thiết bị kỹ thuật số như di động, video, hình ảnh, tin nhắn tới các thiết bị cảm biến, các máy móc được kết nối (ví dụ như ôtô, máy bay hoặc các thiết bị giám sát từ xa) tới các trang web và mạng xã hội. Dữ liệu lớn có đặc điểm là được sinh ra với khối lượng (volume), tốc độ (velocity), độ đa dạng (variety) và tính xác thực (veracity) rất lớn. Ước tính 95% dữ liệu trên thế giới là được sinh ra trong vòng 2 năm trở lại đây. [Tan Jee Toon, Tổng Giám đốc IBM Việt Nam]

Sau đây là khái niệm mới về Big Data 2014 của Gartner về mô hình "5Vs" - năm tính chất quan trọng nói lên Big Data:



Hình 1.4: Mô hình "5Vs" của Big Data – Nguồn Internet

- **Volume (Khối lượng):** nói đến một lượng dữ liệu lớn được tạo ra mỗi giây. Hãy hình dung đó là tất cả các emails, các thông điệp twitter, các bức ảnh, các đoạn video, dữ liệu từ các cảm biến v.v… mà chúng ta tạo và chia sẽ mỗi giây. Chúng ta không phải nói về dữ liệu hàng terabyte mà là những dữ liệu hàng Zettebyte hay Brontobytes. Riêng trên Facebook, chúng ta gửi 10 tỉ thông điệp một ngày, click nút "like" 4.5 tỉ lần và tải lên 350 triệu bức ảnh mới hàng ngày. Nếu so sánh với tất cả dư liệu của thế giới từ trước nay đến năm 2008 thì lượng dữ liệu này chỉ bằng lượng dữ liệu được tạo ra trong mỗi phút hiện nay. Việc tăng trưởng này khiến cho dữ liệu trở nên quá lớn để có thể lưu trữ và phân tích theo công nghệ CSDL truyền thống. Với công nghệ dữ liệu lớn, chúng ta đã có thể lưu trữ và sử dụng những tập dữ liệu này với sự giúp đỡ của các hệ thống phân tán, nơi mà dữ liệu chỉ được lưu trữ một phần tại các địa điểm khác nhau và được tập hợp bởi phần mềm

- **Velocity (tốc độ):** nói đến tốc độ mà dữ liệu mới được tạo ra và tốc độ mà dữ liệu chuyển động. Hãy tưởng tượng đó là các thông điệp của mạng xã hội lan truyền theo đơn vị giây. Hay đó là tốc độ mà các giao dịch thẻ tín dụng gian lận được kiểm tra. Công nghệ dữ liệu lớn cho phép chúng ta có thể phân tích dữ liệu ngay khi chúng đang được tạo ra mà không cần lưu trữ chúng trong các CSDL.

- **Variety (đa dạng)** :nói đến các kiểu khác nhau của dữ liệu hiện giờ chúng ta đang sử dụng. Trong quá khứ, chúng ta tập trung chủ yếu vào các dữ liệu có cấu trúc được lưu trữ trong các bảng hoặc các CSDL quan hệ. Thực tế, có tới 80% dữ liệu trên thế giới ngày nay là phi cấu trúc (vd: hình ảnh, đoạn video, các thông điệp của mạng xã hội) và vì thể không thể đặt chùng vào các bảng. Với công nghệ Big Data, chúng ta có thể lưu trữ các loại dữ liệu khác nhau (cấu trúc và phi cấu trúc) bao gồm các thông điệp, trao đổi của mạng xã hội, các hình ảnh, dữ liệu cảm biến, video, tiếng nói cùng với các dữ liệu có cấu trúc truyền thống.

- **Veracity (Chính xác):** nói đến tính hỗn độn hoặc tính tin cậy của dữ liệu. Với rất nhiều dạng thức khác nhau của dữ liệu lớn, chất lượng và tính chính xác của dữ liệu rất khó kiểm soát. Tuy nhiên, công nghệ dữ liệu lớn và phân tích dữ liệu ngày nay cho phép chúng ta làm việc với những loại dữ liệu này. Khối lượng lớn thường đi kèm với việc thiết chính xác và chất lượng của dữ liệu.

- **Value (giá trị):** Đặc điểm cuối cùng và cũng được coi là quan trọng nhất của dữ liệu lớn là “giá trị”. Việc tiếp cận được dữ liệu lớn sẽ chẳng có ý nghĩa gì nếu chúng ta không chuyển được chúng thành những thứ có giá trị. Chính vì vậy, có thể nói “giá trị” là chữ V quan trọng nhất của Big Data.

Thách thức trong việc xử lý những khối lượng lớn dữ liệu thực chất đã tồn tại từ khá lâu. Trong gần hết lích sử, chúng ta đã làm việc với một ít dữ liệu vè các công cụ để thu thập, tổ chức, lưu trữ và phần tích nó rất nghèo nàn. Chúng ta sàng lọc thông tin, giữ lại mức tối thiểu vừa đủ để có thể khảo sát được dễ dàng hơn. Lấy mẫu ngẫu nhiên làm giảm những vấn đề dữ liệu lớn xuống thành những vấn đề dữ liệu dễ quản lý hơn. Lấy mẫu ngẫu nhiên đã là một thành công lớn và là xương sống của đo lường hiện đại có quy mô lớn. Nhưng nó chỉ là một đường tắt, một lựa chọn tốt thứ 2 để thu thập và phân tích tập dữ liệu đầy đủ. Nó đi kèm với điểm yếu cố hữu. Độ chính xác của nó phụ thuộc vào việc đảm bảo tính ngẫu nhiên. Những thành kiến có hệ thống trong cách thức dữ liệu được thu thập có thể dẫn đến các kết quả ngoại suy rất sai. Việc lấy mẫu đi kèm với một hạn chế đã được thừa nhận từ lâu đó là nó làm mất đi chi tiết. Tuy nhiên, ngày nay, trong nhiều lĩnh vực đang diễn ra một sự thay đổi từ thu nhập một số dữ liệu sang thu thập càng nhiều càng tốt và nếu có thể thì lấy tất cả mọi thứ.

Sử dụng tất cả có nghĩa là chúng ta có thể đi sâu vào dữ liệu; mẫu không thể làm được điều đó. Vì vậy, dữ liệu toàn diện hơn sẽ thay thế con đường tắt lấy mẫu ngẫu nhiên. Làm như vậy đòi hỏi phải có sức mạnh xử lý và lưu trữ phong phú cũng như các công cụ tiên tiến để phân tích tất cả. Nó cũng đòi hỏi những cách thức để dễ dàng và chi phí thấp để thu thập dữ liệu. Trong có khứ mỗi yếu tố này đều là thách thức về công nghệ và giá cả. Tuy nhiên hiên nay chi phí và độ phức tạp của tất cả các mảnh ghép này đã giảm đáng kể. Nhứng gì trước đây là phạm vi của chỉ các công ty lớn nhất thì bây giờ lại khả thi cho hầu như tất cả.

Sử dụng tất cả dữ liệu cho phép phát hiện các kết nối và chi tiết mà bình thường sẽ bị che giấu trong sự bao la của thông tin. Ví dụ, việc phát hiện các gian lận thẻ tín dụng hoạt động bằng cách tìm kiếm những bất thường, và cách tốt nhất để tìm ra chúng là xử lý tất cả các dữ liệu thay vì một phần. Các giá trị ngoại lai là những thông tin thú vị nhất, và chỉ có thể nhận ra chúng khi so sánh với hàng loạt giao dịch bình thường, nó là một vấn để về dữ liệu lớn. Và bởi vì các giao dịch thẻ tín dụng xảy ra tức thời nên việc phân tích thường phải được thực hiện theo thời gian thực.

Sử dụng tất cả dữ liệu không nhất thiết phải là một công việc rất lớn, dữ liệu lớn không cần thiết phải lớn một cách tuyệt đối, mặc dù thường thì nó là như vậy.

Vì dữ liệu lớn dựa trên tất cả thông tin, hoặc nhiều thông tin nhất có thể, nên nó cho phép chúng ta nhìn vào các chi tiết hoặc thử nghiệm các phân tích mới mà không ngại rủi ro bị mất chất lượng. Chúng ta có thể kiểm tra các giải thuyết mới ở nhiều cấp độ chi tiết.

Với sự phát triển của công nghệ, ngày càng có nhiều cơ hội trong đó việc sử dụng tất cả các dữ liệu có sẵn là khả thi. Tuy nhiên nó đi kèm với hạn chế, tăng khối lượng sẽ mở cánh cửa cho sự thiếu chính xác. Điều chắc chắn là những số liệu sai sót và bị hỏng đã luông luôn len lỏi vào các bộ dữ liệu. Chúng ta đã luông luông xem chúng như những rắc rồi và cố gắng loại bỏ chúng. Những gì chúng ta chưa bao giờ muốn làm là xem chúng như những điều không thể tránh khỏi và học cách sống chung với chúng. Đây là một trong những thay đổi cơ bản khi chuyển từ dữ liệu nhỏ sang dữ liệu lớn. Các sai sót về dữ liệu gây ra sự hỗn độn, hỗ độn có thể đơn giản là khả năng sai sót tăng lên khi thêm điểm dữ liệu. Khi số lương tăng lên gấp hàng nghìn lần thì khả năng một số trong đó có thể sai cũng tăng lên. Nhưng cũng có thể làm tăng sự hỗn độn bằng cách kết hợp nhiều loại thông tin khách nhau và từ nguồn khác nhau, không luôn tương thích với nhau một các hoàn hảo.

Ví dụ khi đo nhiệt độ trong một khu vườn, nếu chỉ có một cảm biến nhiệt độ cho toàn bộ khu vườn, ta phải chắc chắn rằng nó chính xác và hoạt động tốt tại mọi thời điểm. Ngược lại, nếu có hàng trăm cảm biến cho mỗi cây trong khu vườn, chúng ta có thể sử dụng các cảm biến rẻ hơn, ít phức tạp hơn (miễn là chúng không phát sinh một sai số có hệ thống). Rất có thể tại một thời điểm, một vài cảm biến sẽ báo dữ liệu không chính xác, tạo ra một bộ dữ liệu ít chính xác hoặc hỗn độn hơn so với bộ dữ liệu từ một cảm biến chính xác. Bất kỳ phép đọc cụ thể nào đó cũng đều có thể không chính xác, nhưng tổng hợp của nhiều phép độc sẽ cung cấp một bức tranh toàn diện hơn. Bởi các bộ dữ liệu này bao gồm nhiều điểm dữ liệu hơn, nó cung cấp giá trị lớn hơn nhiều và có thể bù đắp cho sự hỗn độn của nó.

Tất nhiên dữ liệu không được phép sai hoàn toàn, nhưng chúng ta sẵn sàng hy sinh một chút trong sự chính xác để đổi lại hiểu biết về xu hướng chúng. Dữ liệu lớn biến đổi các con số thành một cái gì đó mang tinh xác suất nhiều hơn là tính chính xác.

Sự phát triển của công nghệ đã làm máy tính nhanh hơn, lưu trữ được nhiều hơn, đồng thời hiệu suất của các thuật toán điều khiển cũng tăng với mức tăng còn nhanh hơn có mức tăng của năng lực xử lý của máy tính. Tuy nhiên, nhiều lợi ích cho xã hội từ dữ liệu lớn lại xảy ra không phải vì các chip nhanh hơn hay vì các thuật toán tốt hơn mà vì có nhiều dữ liệu hơn.

Ví dụ, thuật toán chơi cờ chỉ thay đổi chút ít trong vài thập kỷ qua, bởi các quy tắc của cờ vua đã được biết đầy đủ và bị giới hạn một cách chặt chẽ. Lý do các chương trình cơ vua ngày nay chơi tốt hơn trước đây rất nhiều là một phần bởi chúng được cung cấp dữ liệu nhiều hơn. Thực tế các thế cờ đã được phân tích một cách hoàn toàn đầy đủ và tất

cả các bước đi có thể đã được thể hiện trong một bảng lớn, khi không nén dữ liệu này chiếm hơn một tera byte dữ liệu. Điều này cho phép các máy tính có thể chơi cờ một cách hoàn hảo và con người không bao giờ có thể chơi thắng được máy tính.

Một ví dụ khác về việc "có nhiều dữ liệu hơn sẽ hiệu quả hơn việc có các thuật toán tốt hơn" là trong lĩnh vực xử lý ngôn ngữ tự nhiên. Khoảng năm 2000, Microsoft cố gắng cải thiện bộ kiểm tra ngữ pháp trong chương trình Microsoft word. Họ không chắc liệu sẽ hữu ích hơn nếu cố gắng cải thiện các thuật toán sắn có hay tìm kiếm một kỹ thuật mới. Trước khi đi theo bất kỳ hướng nào, họ quyết định xem xét những gì sẽ xảy ra khi họ cung cấp thêm rất nhiều dữ liệu cho các phương pháp hiện có. Hầu hết các thuật toán học tập của máy dựa trên những tập sao lục văn bản đạt tới một triệu từ hoặc ít hơn. Họ đã lấy bốn thuật toán thông thường và cung cấp dữ liệu nhiều hơn ở ba mức khác nhau: 10 triệu từ, 100 triệu từ và 1 tỷ từ. Kết quả là khi có nhiều dữ liệu đi vào, hiệu suất của tất cả bốn thuật toán đều được cải thiện đáng kể. Trong thực tế, một thuật toán đơn giản hoạt động kém hiệu quả nhất với nửa triệu từ lại hoạt động tốt hơn những thuật toán khác khi có một tỷ từ. Ngược lại, thuật toán làm việc tốt nhất với ít dữ liệu lại hoạt động kém nhất với lượng dữ liệu lớn hơn, mặc dù chúng đều cải thiện đáng kể.

Năm 2006, Google đã nhảy vào lĩnh vực dịch thuật, thay vì dịch các trang văn bản thành hai ngôn ngữ, Google tự giúp mình với một bộ dữ liệu lớn hơn nhưng cũng hỗn độn hơn nhiều: toàn bộ mạng internet và hơn thế nữa. Hệ thống của google đã thu lượng bất kể bản dịch nào có thể tìm thấy, để huấn luyện máy tính. Chúng bao gồm các trang web của các công ty viết ở nhiều ngôn ngữ khác nhau, các bản dịch đồng nhất của các văn bản chính thức và các báo cáo của các tổ chức liên chính phủ như liên hợp quốc, liên minh châu âu. Thậm chí các bản dịch sách từ dự án sách của Google cũng được thu nhận. Bất chấp sự hỗn độn của đầu vào, dịch vụ của Google hoạt động tốt nhất. Các bản dịch của nó là chính xác hơn so với của các hệ thống khác và nó phong phú hơn rất nhiều. Vào giữa năm 2012, bộ dữ liệu của nó bao gồm hơn 60 ngôn ngữ. Nó thâm chí có thể chấp nhận nhập văn bản vào bằng giống nói trong 14 ngôn ngữ để dịch. Và vì nó xử lý ngôn ngữ đơn giản như là dữ liệu hỗn độn để đánh giá xác suất, nó thậm chí có thể dịch giữa các ngôn ngữ. Trong trường hợp này, nó sẽ sử dụng tiếng Anh như một cầu nối. Nó linh hoạt hơn rất nhiều so với những cách tiếp cận khác vì nó có thể thêm và bớt cá từ qua kinh nghiệm chúng được hay không được sử dụng. Lý do hệ thống dịch thuật của Google hoạt động tốt không phải vì nó có một thuật toán thông minh hơn. Nó hoạt động tốt bởi vì nó có nhiều dữ liệu hơn và không chỉ dữ liệu chất lượng cao. Việc sử dụng bộ dữ liệu lớn hơn cho phép nhưng bước tiến lớn trong xử lý ngôn ngữ tự nhiên mà các hệ thống nhận dạng tiếng nói và dịch máy dựa vào. Mô hình đơn giản và rất nhiều dữ liệu thắng thế những mô hình phức tạp hơn nhưng dựa vào ít dữ liệu hơn.

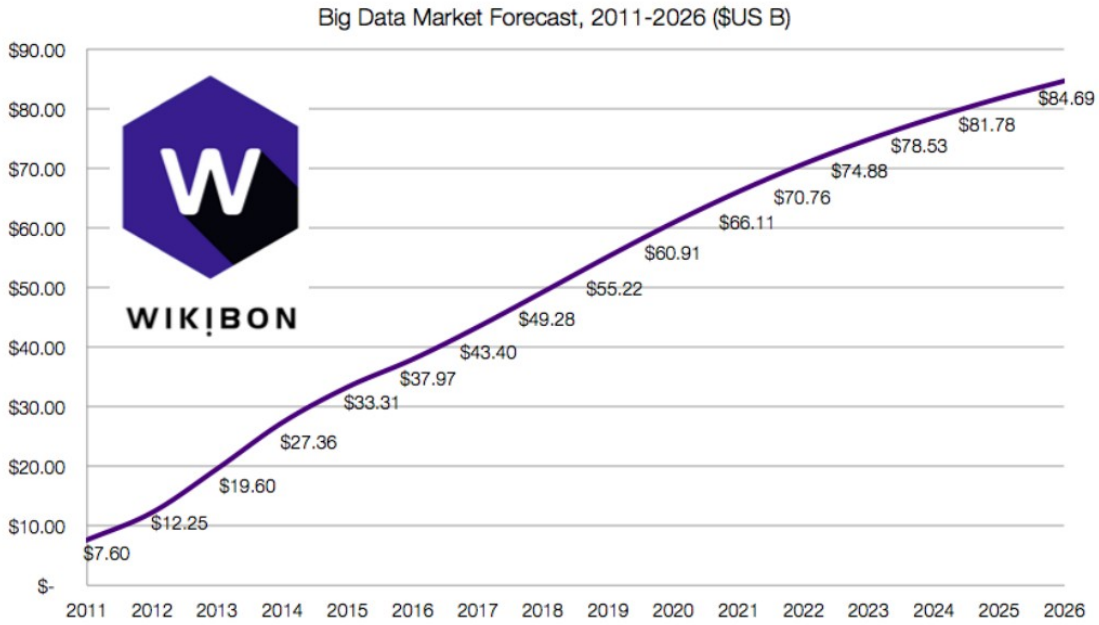
Trong nhiều lĩnh vực công nghệ và xã hội, dữ liệu lớn đã chứng tỏ xu thế nhiều hơn và hỗn độn chứ không phải ít hơn và chính xác. Hãy xem xét trường hợp của việc phân loại nội dung. Trong nhiều thế kỷ con người đã phát triển các nguyên tắc phân loại và chỉ số để lưu trữ và tìm kiếm tài liệu. Trong thế giới dữ liệu nhỏ thì chúng hoạt động tốt, tuy nhiên khi tăng quy mô lên nhiều cấp độ, những hệ thống này lại sụp đổ. Năm 2011, trang web chia sẻ hình ảnh Flickr có chưa hơn 6 tỷ hình ảnh từ hơn 75 triệu người dùng. Việc cố gắng gán nhãn cho từng bức ảnh theo những thể loại định trước đã tỏ ra vô ích. Thay vào đó, nguyên tăc phân loại sạch được thay thế bằng cơ chế hỗn độn hơn nhưng linh hoạt hơn và dễ thích nghi hơn. Khi tải ảnh lên Flickr, người dùng “gán thẻ” (tag) cho chúng. Có nghĩa là người dùng gán một số bất kỳ các nhãn văn bản và sử dụng chúng để tổ chức và tìm kiếm các tư liệu. Thẻ được tạo ra và gán một cách đặc biệt, không có phân loại sẵn để chúng ta phải tuân thủ. Thay vào đó, bất cứ ai cũng có thể thêm các thẻ mới bằng cách gõ chúng vào. Gắn thẻ đã nôi rleen như một tiêu chuẩn thực tế để phân loại nội dung trên internet, được sử dụng trên các trang mạng xã hội như Twitter, các blog... Nó làm cho người dùng dễ dàng di chuyển hơn trong sự bao la của nội dung các trang web, đặc biệt là cho nhứng thứ như hình ảnh, phim, và âm nhạc không

dựa trên văn bản nên việc tìm kiếm bằng từ không thể hoạt động. Tất nhiên, một số thẻ có thể bị viết sai chính tả, và những lỗi như vạy sẽ tạo ra sự không chính xác, không chỉ đối với chính dữ liệu mà còn đối với việc chúng được tổ chức ra sao. Nhưng bù lại cho sự hỗn độn trong cách tổ chức các bộ sưu tập ảnh, chúng ta cố một vũ trụ phong phú hơn nhiều của cá nhãn mác, và mở rộng ra là sự truy cập sâu hơn, rộng hơn tới các ảnh của chúng ta. Chúng cũng cho phép phối hợp các thẻ tìm kiếm để lọc các bức ảnh theo những cách không thể làm được trước đây.

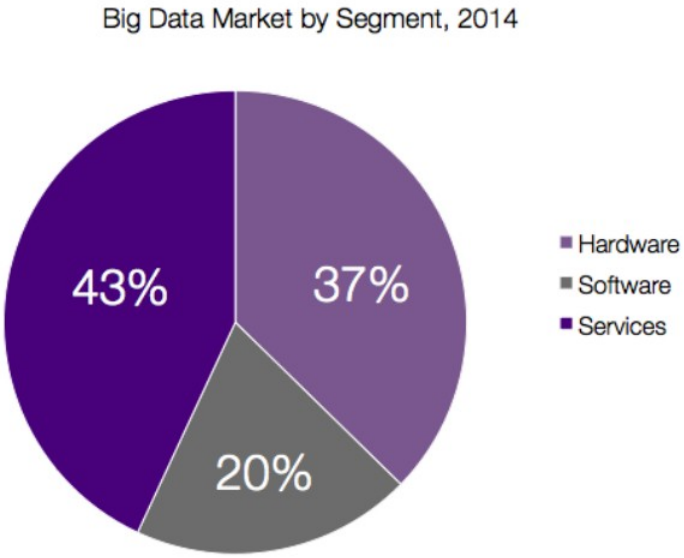
1.4. Xu hướng phát triển của công nghệ dữ liệu lớn.

Năm 2014, thị trường công nghệ về Big Data tiếp tục trên đà phát triển dựa trên các tiêu chí về doanh thu liên quan đến việc bán sản phẩm, dịch vụ và việc áp dụng các công nghệ Big Data của các doanh nghiệp lớn trên thị trường.

Theo dự báo thị trường Wikibon, đối với năm 2014, thị trường Big Data - được đo bằng doanh thu liên quan đến việc bán phần cứng, phần mềm và các dịch vụ chuyên nghiệp, đạt $27.36 tỷ cao hơn năm 2013 ($19.6 tỷ). Tuy vậy tốc độ tăng trưởng chung của thị trường của Big Data đã chậm lại trong năm qua năm từ 60% năm 2013 và 40% vào năm 2014. Wikibon cũng mở rộng dự báo thị trường Big Data đến năm 2026. Wikibon hy vọng thị trường Big Data đạt $84 tỷ vào năm 2026, với tỷ lệ tăng trưởng hàng năm khoảng 17% trong giai đoạn 15 năm bắt đầu từ 2011.

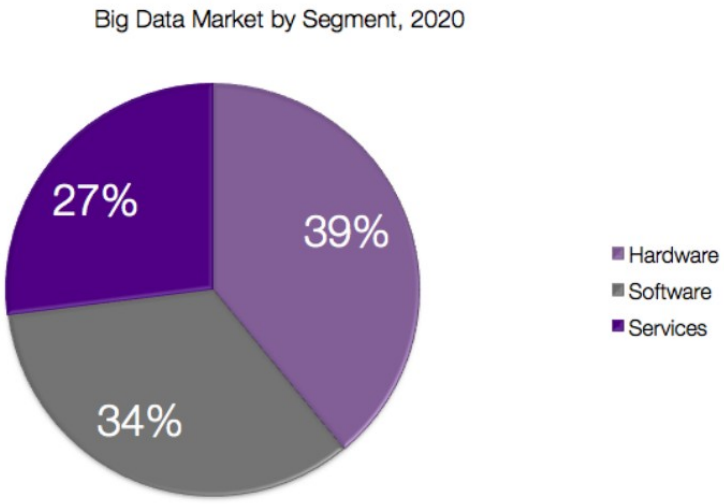


Hình 1.5: Dự báo thị trường Big Data đến năm 2026 – Nguồn Wikibon

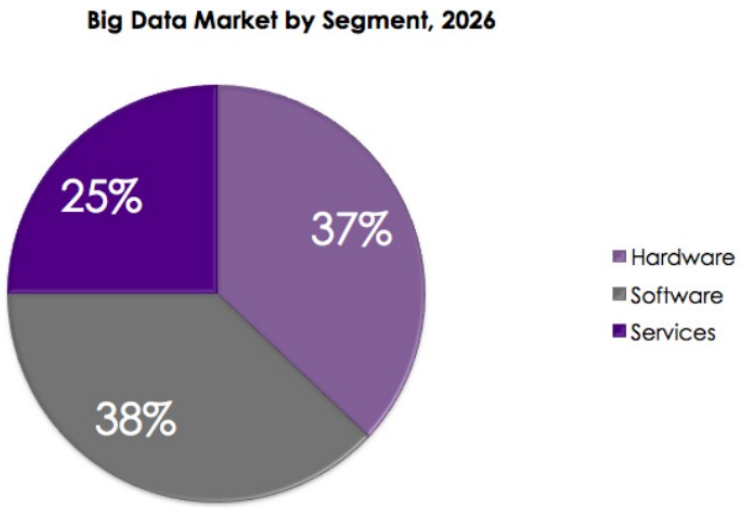


Hình 1.6: Phân khúc thị trường Big Data năm 2014 – Nguồn Wikibon

Wikibon tin rằng một sự thay đổi đáng kể trong doanh thu từ các dịch vụ chuyên nghiệp với các phần mềm trong những năm tới.



Hình 1.7: Dự báo phân khúc thị trường Big Data năm 2020 – Nguồn Wikibon



Hình 1.8: Dự báo phân khúc thị trường Big Data năm 2026 – Nguồn Wikibon

Doanh thu từ Big Data:

Lợi ích từ việc ứng dụng ứng dụng Big Data vào việc phân tích dữ liệu, thói quen, tâm lý, nhu cầu của khách hàng để làm cơ sở cho các hoạt động kinh doanh, marketing của các doanh nghiệp trên thế giới, đã đem lại cho họ một các khoản doanh thu lớn. Wikibon đã theo dõi và phân tích doanh thu từ Big Data của hơn 60 nhà cung cấp năm 2014.

Bảng doanh thu từ Big Data của 60 hãng công nghệ lớn năm 2014:

Nhà cung cấp	Doanh thu từ Big Data	% Big Data Hardware Revenue	% Big Data Software Revenue	% Big Data Services Revenue
IBM	$1,601	26%	35%	39%
HP	$932	43%	14%	43%
SAP	$923	0%	79%	21%
Teradata	$687	29%	40%	31%
Dell	$685	85%	0%	15%
Palantir	$544	0%	35%	65%
SAS Institute	$533	0%	67%	33%
Microsoft	$532	0%	70%	30%
Accenture	$498	0%	0%	100%
Oracle	$493	29%	40%	31%
Splunk	$451	0%	74%	26%
Amazon	$440	0%	0%	100%
PwC	$406	0%	0%	100%
Deloitte	$375	0%	0%	100%
Informatica	$353	0%	87%	13%
Cisco	$321	85%	0%	15%

Systems				
EMC	$315	71%	0%	29%
Intel	$268	81%	4%	15%
Google	$225	0%	0%	100%
Mu Sigma	$225	0%	0%	100%
CSC	$210	0%	0%	100%
Microstrategy	$192	0%	75%	25%
NetApp	$184	73%	0%	27%
Red Hat	$169	0%	74%	26%
Pivotal	$159	0%	77%	23%
Cap Gemini	$145	0%	0%	100%
Opera Solutions	$130	0%	0%	100%
TCS	$124	0%	0%	100%
VMware	$106	0%	79%	21%
MarkLogic	$102	0%	80%	20%
Qlik	$101	0%	89%	11%
Rackspace	$95	0%	0%	100%
Actian	$94	0%	89%	11%
Cloudera	$91	0%	53%	47%
Tableau Software	$90	0%	89%	11%
DDN	$86	84%	0%	16%
TIBCO	$64	0%	67%	33%
Guavus	$62	0%	68%	32%
Alteryx	$55	0%	87%	13%
1010data	$53	0%	89%	11%
Hortonworks	$43	0%	63%	37%
MapR	$42	0%	83%	17%
Syncsort	$35	0%	86%	14%
MongoDB	$35	0%	71%	29%
DataStax	$34	0%	76%	24%
Attivio	$32	0%	63%	38%
GoodData	$27	0%	78%	22%
Fractal Analytics	$25	0%	0%	100%
Datameer	$25	0%	80%	20%
Sumo Logic	$25	0%	0%	100%
Talend	$25	0%	68%	32%
Attunity	$24	0%	83%	17%
Pentaho	$24	0%	79%	21%
Couchbase	$18	0%	78%	22%
SiSense	$15	0%	67%	33%
Basho	$14	0%	79%	21%
Aerospike	$13	0%	85%	15%
Neo Technology	$13	0%	85%	15%
Revolution Analytics	$12	0%	67%	33%

Think Big Analytics	$12	0%	0%	100%
Digital Reasoning	$12	0%	67%	33%
Paxata	$11	0%	82%	18%
Tresata	$12	0%	83%	17%
Trifacta	$10	0%	90%	10%
ODM	$5,814	100%	0%	0%
Other	$7,891	22%	6%	72%
Total	**$27,361**	**37%**	**20%**	**43%**

BigData là nhu cầu đang tăng trưởng lớn đến nỗi từ năm 2010, Software AG, Oracle, IBM, Microsoft, SAP, EMC, HP và Dell đã chi hơn 15 tỷ USD cho các công ty chuyên về quản lý và phân tích dữ liệu.

Interactions Marketing, một công ty tiếp thị theo hình thức tận dụng ngay chính khách hàng của mình, đã tiến hành kiểm soát dữ liệu lớn bằng cách sử dụng dữ liệu giao dịch điểm bán hàng và dữ liệu thông tin thời tiết khu vực từ nhiều nguồn khác nhau để có được những hiểu biết nhanh nhất về hành vi mua sắm.

Mọi khía cạnh trong đời sống của chúng ta đều sẽ bị ảnh hưởng bởi dữ liệu lớn. Các ứng dụng dữ liệu lớn được sử dụng phổ biến nhất cũng như tạo ra được những lợi ích cao nhất trong 10 lĩnh vực.

1.4.1. Sự hiểu biết và khách hàng mục tiêu

Đây là một trong những lĩnh vực lớn nhất và được công bố công khai nhất cách dữ liệu lớn được sử dụng ngày nay. Ở đây, dữ liệu lớn được sử dụng để hiểu rõ hơn về khách hàng và hành vi cũng như sở thích của họ.

Các công ty đều mong muốn mở rộng tập hợp dữ liệu truyền thống với các dữ liệu truyền thông xã hội, trình duyệt web cũng như phân tích văn bản và dữ liệu cảm biến để có được một bức tranh hoàn chỉnh hơn về khách hàng của họ. Trong nhiều trường hợp, mục tiêu lớn hơn là để tạo ra mô hình dự báo.

Bạn có thể ghi nhớ về ví dụ của nhà bán lẻ Target (Mỹ), những người có thể dự đoán rất chính xác khi nào một khách hàng của họ sẵn sàng mua. Sử dụng dữ liệu lớn, các công ty viễn thông có thể dự đoán tốt hơn về việc khách hàng rời mạng. Hay WalMart có thể dự đoán sản phẩm gì sẽ được bán ra, và các công ty bảo hiểm xe hơi hiểu khách hàng của họ lái xe như thế nào.

Interactions Marketing, một công ty tiếp thị theo hình thức tận dụng ngay chính khách hàng của mình, đã tiến hành kiểm soát dữ liệu lớn bằng cách sử dụng dữ liệu giao dịch điểm bán hàng và dữ liệu thông tin thời tiết khu vực từ nhiều nguồn khác nhau để có được những hiểu biết nhanh nhất về hành vi mua sắm. Bài thử nghiệm này sử dụng Google BigQuery, một dịch vụ web để phân tích sự tương tác của các bộ dữ liệu cực lớn, và công cụ phân tích hình ảnh Tableau Software để nhanh chóng kiểm tra số lượng lớn thông tin. Sự kết hợp của các công cụ cho phép Interactions cắt giảm thời gian phân tích từ khoảng một tuần xuống còn một vài giờ hay thậm chí chỉ còn vài phút, Giovanni DeMeo, Phó Chủ tịch phân tích và tiếp thị toàn cầu của Interactions, cho biết. Chương trình phân tích các hành động của người mua hàng qua đó giúp các nhà bán lẻ và các nhà sản xuất lên kế hoạch chương trình khuyến mãi tại cửa hàng trước khi những sự kiện này xảy ra. Kết quả mà phân tích dữ liệu tìm thấy trong dự án này là: Một ngày trước khi sự kiện thời tiết tương tự như thống kê xảy ra, doanh số bán hàng của 28 loại sản phẩm đã tăng từ 20% lên 261% so với cùng thời điểm năm ngoái.

Các nhà bán lẻ có thể tối ưu hóa giá cả và lượng hàng hóa của họ dựa trên các dự đoán được tạo ra từ dữ liệu phương tiện truyền thông xã hội, xu hướng tìm kiếm web và dự báo thời tiết. Một quy trình kinh doanh với rất nhiều phân tích dữ liệu lớn là chuỗi cung ứng hoặc cung cấp lộ trình tối ưu hóa. Ở đây, cảm biến nhận dạng tần số vô tuyến định vị và địa lý được sử dụng để theo dõi hàng hóa, phương tiện giao hàng và các tuyến đường tối ưu bằng cách tích hợp dữ liệu giao thông trực tiếp.

Ngay cả chiến dịch bầu cử của Mỹ cũng có thể được tối ưu hóa bằng việc sử dụng phân tích dữ liệu lớn. Các chuyên gia cho rằng, ông Obama giành chiến thắng trong chiến dịch bầu cử năm 2012 là do khả năng vượt trội của đội ngũ sử dụng khả năng phân tích dữ liệu lớn.

Lĩnh vực nhân sự cũng đang được cải thiện bằng cách sử dụng phân tích dữ liệu lớn. Điều này bao gồm việc tối ưu hóa của việc ‘săn’ tài năng, cũng như đánh giá nền văn hóa công ty và sự tham gia của nhân viên trong việc sử dụng công cụ dữ liệu lớn.

1.4.2. Định lượng cá nhân và tối ưu hóa hiệu suất

Dữ liệu lớn không chỉ dành cho các công ty và chính phủ mà còn cho từng cá nhân. Ngày nay chúng ta có thể được hưởng lợi từ dữ liệu được tạo ra từ các thiết bị đeo như đồng hồ thông minh hoặc vòng đeo tay thông minh.

Lấy sợi dây Up của Jawbone làm ví dụ: Sợi dây thu thập dữ liệu về việc tiêu thụ

calo của chúng ta, mức độ hoạt động, và mô hình giấc ngủ. Ngoài việc mang lại cho cá nhân những hiểu biết phong phú, giá trị hơn cả là trong việc phân tích các dữ liệu thu thập được[2].

Trong trường hợp Jawbone, công ty hiện thu thập giá trị của dữ liệu giấc ngủ mỗi đêm trong vòng 60 năm. Phân tích khối lượng dữ liệu lớn này sẽ mang lại cái nhìn hoàn toàn mới để phản hồi cho người dùng cá nhân. Các lĩnh vực khác, nơi mà chúng ta được hưởng lợi từ phân tích dữ liệu lớn chính là việc tìm kiếm tình yêu trực tuyến. Các trang web hẹn hò trực tuyến lớn nhất đang áp dụng công cụ dữ liệu lớn và các thuật toán để tìm thấy người phù hợp nhất cho chúng ta.

Các thiết bị đeo tay sẽ thu thập dữ liệu thông tin của người sử dụng, mục đích ban đầu là có được các số liệu thông báo với người dùng là họ đã có những hoạt động gì (đi bộ, leo cầu thang, đi nhanh,...), giúp người dùng có thể kiểm soát được năng lượng tiêu thụ trong ngày, kiểm soát được thời gian nghỉ ngơi (ngủ, tĩnh dưỡng - không vận động). Nhưng mục tiêu cuối cùng đối với các nhà cung cấp thiết bị đeo tay thông minh là có thể thu thập được dữ liệu của nhiều người nhất, tất nhiên là những dữ liệu có tính cá nhân, không vi phạm nguyên tắc bảo mật và vi phạm quyền cá nhân. Với những dữ liệu đó, các công ty có thể thực hiện phân tích với lượng dữ liệu lớn. Họ có thể phân tích những gì từ các hoạt động, từ các bài tập thể dục của người dùng? Đó có thể là những cách tập thể dục phổ biến, các xu hướng tập thể dục, các bài hát được sử dụng khi thực hiện các bài tập. Họ xác định được ngày nào trong tuần sẽ có ít người tham gia các lớp luyện tập nhất, các đối tượng tham gia thích hợp vào các khoảng thời gian nào để có thể tư vấn với người dùng về chương trình tập luyện, hay đưa ra các chương trình khuyến mãi.

1.4.3. Cải thiện chăm sóc sức khỏe và y tế công

Khả năng tính toán, phân tích dữ liệu lớn cho phép chúng ta giải mã toàn bộ chuỗi DNA trong vài phút và tìm ra những phương pháp chữa trị mới, nhằm hiểu rõ hơn cũng như dự đoán mô hình bệnh. Hãy nghĩ về điều gì sẽ xảy ra khi tất cả các dữ liệu cá nhân, từ đồng hồ thông minh và các thiết bị đeo, có thể được sử dụng để áp dụng cho hàng triệu người và các căn bệnh khác nhau của họ. Các thử nghiệm lâm sàng trong tương lai sẽ không bị giới hạn bởi kích thước mẫu nhỏ mà sẽ có khả năng bao quát tất cả mọi người!

Kỹ thuật dữ liệu lớn đã được sử dụng để giám sát trẻ sơ sinh trong chuyên khoa chăm sóc trẻ sinh non và khoa bệnh nhi. Bằng cách ghi lại và phân tích từng nhịp tim và mô hình thở của mỗi bé, các nhà khoa học đã có thể phát triển những thuật toán có thể dự đoán nhiễm trùng trong vòng 24 giờ trước khi các triệu chứng vật lý xuất hiện. Bằng cách

đó, nhóm nghiên cứu có thể can thiệp sớm và giữ lại mạng sống cho những đứa trẻ mà thời gian sống chỉ tính bằng giờ.

Hơn nữa, phân tích dữ liệu lớn cho phép chúng ta theo dõi, dự đoán sự phát triển của dịch bệnh và sự bùng phát dịch bệnh. Tích hợp dữ liệu từ hồ sơ y tế với phân tích phương tiện truyền thông xã hội cho phép chúng ta giám sát dịch cúm trong thời gian thực, chỉ đơn giản bằng cách lắng nghe những gì mọi người đang đề cấp đến, ví dụ như: "Cảm giác như người thừa hôm nay – trên giường với bệnh cảm lạnh".

Hệ chuẩn đoán y học bao gồm những hệ thống có sự hỗ trợ của hệ chuyên gia dựa trên luật (gọi là DSSes: Rule-based Expert Decision Support Systems), nhưng với dữ liệu lớn, bằng chứng tồn tại những hệ thống này có thể ra khỏi nghiên cứu và trở thành những người phụ tá y tế chính.

1.4.5. Cải thiện hiệu suất thể thao

Hầu hết các môn thể thao hiện đại đều áp dụng phân tích dữ liệu lớn. Chúng ta có công cụ SlamTracker của IBM dành cho các giải đấu quần vợt. Chúng ta sử dụng phân tích video để theo dõi hiệu suất của mỗi cầu thủ trong bóng đá hoặc bóng chày, và công nghệ cảm biến trong các thiết bị thể thao như bóng rổ hay các câu lạc bộ golf cho phép chúng ta có được thông tin phản hồi (thông qua điện thoại thông minh và các máy chủ điện toán đám mây) về hiệu suất thi đấu của mình và làm thế nào để cải thiện nó.

Nhiều đội thể thao có tiếng còn theo dõi các vận động viên bên ngoài của môi trường thể thao, như sử dụng công nghệ thông minh để theo dõi chế độ dinh dưỡng và giấc ngủ, cũng như các cuộc hội thoại truyền thông xã hội để nhận biết tâm tư, tình cảm.

Gần đây nhất là mùa Worldcup năm 2014 diễn ra tại Brasil, đội tuyển Đức có một chiến thuật hợp lý, vượt trội cho từng trận đấu với từng đối thủ cũng như cho cả vòng loại? Bí mật này nằm ở công nghệ phân tích big data mà đội tuyển Đức áp dụng từ những năm 2012. Công nghệ này giúp phân tích từng cầu thủ đối phương, đồng thời đưa ra giải pháp tối ưu cho từng cầu thủ trong đội tuyển Đức.

1.4.6. Nâng cao khoa học và nghiên cứu

Khoa học và nghiên cứu hiện đang biến đổi rất nhanh bởi các khả năng mới mà dữ liệu lớn mang lại. Lấy ví dụ, CERN, phòng thí nghiệm vật lý hạt nhân Thụy Sĩ với chiếc máy gia tốc hạt lớn nhất và mạnh nhất thế giới, Large Hadron Collider. Với những thí nghiệm để mở khóa những bí mật của vũ trụ, cách hình thành và vận hành ra sao, đã tạo

ra một lượng lớn dữ liệu.

Các trung tâm dữ liệu của CERN có 65.000 bộ vi xử lý để phân tích 30 petabyte dữ liệu. Tuy nhiên, nó sử dụng các quyền hạn tính toán của hàng nghìn máy tính phân phối tại 150 trung tâm dữ liệu trên toàn thế giới để phân tích. Quyền hạn tính toán như vậy có thể được thừa hưởng và làm biến đổi rất nhiều lĩnh vực khác của khoa học và nghiên cứu.

1.4.7. Tối ưu hóa hiệu suất máy móc và thiết bị

Phân tích dữ liệu lớn giúp máy móc và thiết bị trở nên thông minh và độc lập hơn. Ví dụ, các công cụ dữ liệu lớn được sử dụng để vận hành xe hơi tự lái của Google. Toyota Prius được trang bị máy ảnh, GPS cũng như các máy tính mạnh mẽ và bộ cảm biến để lái xe an toàn trên đường mà không có sự can thiệp của con người. Công cụ dữ liệu lớn cũng được sử dụng để tối ưu hóa lưới điện năng lượng sử dụng dữ liệu từ công-tơ thông minh. Chúng ta thậm chí có thể sử dụng công cụ dữ liệu lớn để tối ưu hóa hiệu suất của máy tính và các kho dữ liệu.

1.4.8. Cải thiện an ninh và thực thi pháp luật

Dữ liệu lớn được áp dụng rất nhiều trong việc cải thiện an ninh và cho phép thực thi pháp luật. Cơ quan An ninh Quốc gia Mỹ (NSA) sử dụng phân tích dữ liệu lớn để chống âm mưu khủng bố (và có thể gián điệp trên tất cả chúng ta). Các đơn vị khác sử dụng kỹ thuật dữ liệu lớn để phát hiện và ngăn chặn các cuộc tấn công không gian mạng. Lực lượng cảnh sát sử dụng các công cụ dữ liệu lớn để bắt tội phạm và thậm chí dự đoán hoạt động tội phạm, và những công ty thẻ tín dụng sử dụng dữ liệu lớn dùng nó để phát hiện các giao dịch gian lận.

1.4.9. Cải thiện và tối ưu hóa các thành phố, quốc gia

Dữ liệu lớn được sử dụng để cải thiện nhiều khía cạnh của các thành phố và quốc gia. Ví dụ như nó cho phép các thành phố tối ưu hóa luồng giao thông dựa trên thông tin giao thông trong thời gian thực cũng như dữ liệu trên các phương tiện truyền thông xã hội và dữ liệu thời tiết. Một số thành phố đang thực hiện thí điểm phân tích dữ liệu lớn với mục đích biến mình thành thành phố thông minh, nơi mà cơ sở hạ tầng giao thông và các quy trình tiện ích đều được kết nối với nhau. Nơi một chiếc xe buýt sẽ chờ một đoàn tàu đến trễ và nơi tín hiệu giao thông dự đoán khối lượng giao thông và hoạt động đểgiảm thiểu ùn tắc.

1.4.10. Kinh doanh tài chính

Thể loại cuối cùng về ứng dụng dữ liệu lớn đến từ các giao dịch tài chính. Tần số giao dịch cao (HFT) là một lĩnh vực nơi dữ liệu lớn được sử dụng rất nhiều ngày nay. Ở đây, thuật toán dữ liệu lớn được sử dụng để đưa ra các quyết định giao dịch. Ngày nay, phần lớn các giao dịch cổ phiếu diễn ra thông qua các thuật toán dữ liệu dựa ngày càng nhiều vào tín hiệu tài khoản từ các mạng truyền thông xã hội và các trang web tin tức để đưa ra quyết định mua và bán trong từng giây.

Phân tích tâm lý thị trường chứng khoán sử dụng Google Trends đã chỉ ra được tương quan tốt cho những tăng giảm chỉ mục theo thời gian, mà có lẽ không đáng ngạc nhiên nhưng những thú vị về tính trọng đại như một ứng dụng dữ liệu lớn. Bài viết "Quantifying Trading Behavior in Financial Markets Using Google Trends (Dự đoán xu hướng thương mại trong thị trường tài chính sử dụng Google Trends)" cung cấp bằng chứng rằng việc sử dụng phân tích tâm lý để kéo dài hay rút ngắn quyết định mua và bán cổ phiếu nắm giữ có thể tốt hơn việc mua và nắm giữ những chiến lược đơn giản và quỹ đầu tư index. Nghiên cứu này có thể được phân tích chi tiết hơn nhưng những kết quả của nó cũng khá thuyết phục. Một nghiên cứu thú vị dự đoán những khả năng xảy ra của một hệ thống cho một lĩnh vực hình thức kinh doanh thực tế.

Uber, có thể nói là một trong những ứng dụng đầu tiên của công nghệ dữ liệu lớn (big data) vào kinh tế. Nhờ ứng dụng thuật toán thông minh xử lý dữ liệu lớn, Uber biết thời điểm "cầu" lệch xa "cung" như ngày nghỉ lễ, thời tiết xấu hay thành phố có biến để điều chỉnh hệ số tăng giá sốc (surge). Bằng hình thức này, nhu cầu hành khách đi xe giảm xuống, trong khi đó lại kích thích thêm nhiều tài xế Uber tham gia vận chuyển hành khách, bài toán cung - cầu được cân đối.

Sự xuất hiện của Uber còn đánh dấu một cột mốc liên quan đến nền kinh tế chia sẻ (sharing economy), nó có lợi cho nền kinh tế, xét trên góc độ phân bổ và sử dụng nguồn lực. Có thể nói đây là một tác động rất lớn của cuộc cách mạng công nghệ thông tin viễn thông đến kinh tế. "Bàn tay vô hình hay" là thông tin về giá (price information) của Adam Smith có thể sẽ dần dần được dòng chảy thông tin/dữ liệu lớn (information flow/big data) thay thế. Ngày đó chắc chắn còn rất xa, nhưng khởi điểm của nó đã bắt đầu bằng những dịch vụ/doanh nghiệp như Uber.

CHƯƠNG 2: CÔNG NGHỆ DỮ LIỆU LỚN TẠI VIỆT NAM

2.1. Hiện trạng và xu hướng phát triển công nghệ dữ liệu lớn tại Việt Nam

Việt Nam tuy là một nước đang phát triển nhưng lại có tốc độ tăng trưởng trong lĩnh vực viễn thông và công nghệ thông tin rất nhanh. Với hơn 44 triệu người dùng internet, trong đó có đến 26 triệu người tham gia các mạng xã hội có thể nói Việt Nam là một thị trường rất tiềm năng cho các công ty, tổ chức triển khai và khai thác các lợi ích từ công nghệ dữ liệu lớn. Công nghệ dữ liệu lớn đã bắt đầu được nghiên cứu và đưa vào ứng dụng từ trước năm 2000, tuy nhiên đến 2011 cùng với sự bùng nổ của thông tin thì công nghệ dữ liệu lớn cũng thực sự phát triển mạnh mẽ và được phổ biến trên toàn thế giới.

Các công ty lớn tại Việt Nam cũng sớm tiếp thu được xu thế này và đã bắt đầu nghiên cứu và ứng dụng công nghệ Big Data vào các dự án và sản phẩm của mình. Năm 2012, FPT đã triển khai thành công cho cục Quản lý Giám sát Bảo Hiểm và Hiệp Hội Bảo Hiểm Việt Nam dự án "Xây dựng CSDL về bảo hiểm bắt buộc trách nhiệm dân sự của chủ xe cơ giới". Đây là dự án đầu tiên tại Việt Nam sử dụng BI-GIS. Dự án yêu cầu tích hợp dữ liệu tự động từ nhiều nguồn với sự đa dạng của các hệ thống thông tin tại các doanh nghiệp bảo hiểm. Hệ thống thông tin quản trị (Business Intelligence - BI) mạnh cho phép các cơ quan, doanh nghiệp tập hợp và làm sạch dữ liệu từ nhiều nguồn khác nhau, quản lý khối lượng dữ liệu lớn (big data), các chiều phân tích số liệu để phục vụ quản trị, phân tích và hoạch định chính sách. BI-GIS tích hợp nhiều công nghệ mới nhất phục vụ cho công tác quản lý và khai thác thông tin: Data Integration, Data Warehouse, Analytics Dashboard, Map Data, Cloud Services, v.v. Với xu thể hiện đại hóa quản lý theo định hướng "xã hội hóa thông tin" và "dịch vụ công điện tử" của Chính phủ.

VCCorp một trong những công ty truyền thông, thương mại điện tử hàng đầu tại Việt Nam cũng đã sớm đưa công nghệ Big Data vào ứng dụng và đã đạt được những thành công đáng kể. Với hơn 20 sản phẩm trong lĩnh vực Truyền thông, Thương mại điện tử và Mạng xã hội, VCCorp hiện đang hợp tác với hơn 20 báo điện tử và hơn 200 trang web của Việt Nam tạo thành sức mạnh bó đũa với độ phủ khoảng 31 triệu độc giả - tương đương 90% người dùng Internet Việt Nam. VCCorp đã sớm nắm bắt xu hướng phát triển ứng dụng công nghệ điện toán đám mây và dữ liệu lớn, triển khai công nghệ phân tích dữ liệu, hành vi để phát hiện người dùng đang quan tâm vấn đề, mặt hàng gì để có quảng cáo phù hợp, đạt hiệu quả cao nhất. Quy mô xử lý dữ liệu của VCCorp hiện đã ngang với Yahoo, ước tính có tới 30 tỷ lượt hiển thị quảng cáo/tháng, 1.000 tỷ bản ghi, 20

Terabytes/ngày. Độ chính xác trong phán đoán hành vi người dùng ngang với Google (82%). VCCorp đang cạnh tranh trực diện với Google, Facebook tại Việt Nam trên quy mô, sản phẩm, công nghệ quảng cáo và các kênh bán hàng nội địa, hiện quy mô gấp 1,5 lần Google, và 2 lần Facebook tại Việt Nam.

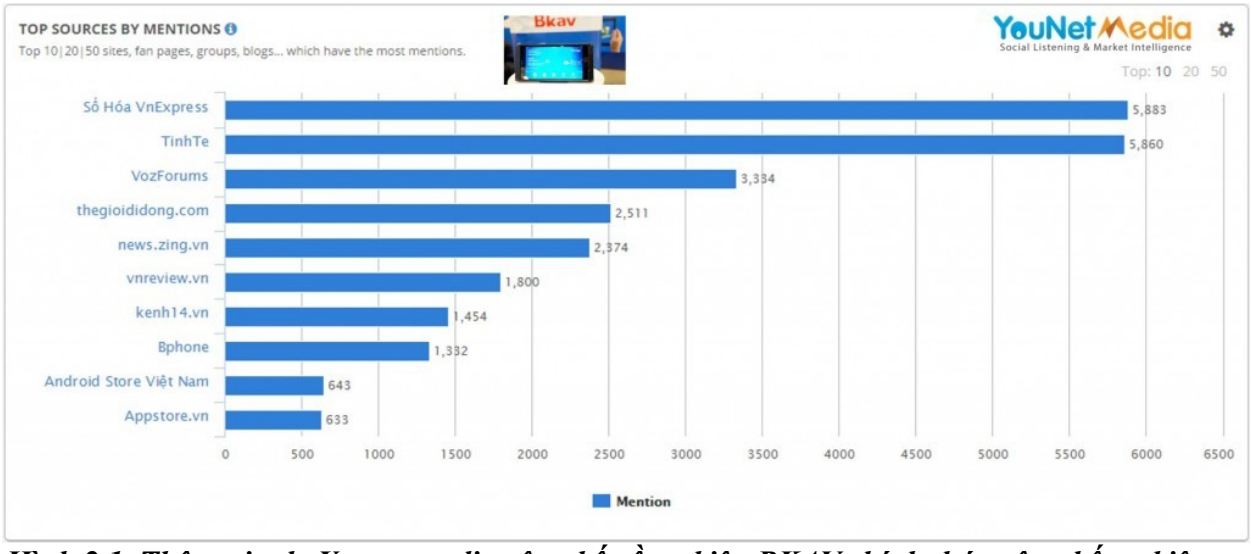
Một trong những những công ty nội dung số hàng đầu ở Việt Nam – VNG - sở hữu ứng dụng OTT (Over The Top) nổi tiếng Zalo chat có 37 triệu người dùng – với tỉ lệ người dùng thường xuyên (active users) là 60% và được dự đoán là sẽ đạt con số 40 triệu người dùng vào tháng 11 năm 2015. Với một lượng khách hàng khổng lồ như vậy đã đặt Zalo ở vị trí tốp đầu trong cuộc đua công nghệ tại Việt Nam với các tên tuổi OTT thế giới, bỏ xa những ứng dụng nội địa khác, và đây cũng là mỏ vàng đáng mơ ước của bất kỳ một công ty công nghệ nào ở Việt Nam. Với việc tận dụng tập dữ liệu rất lớn, cùng với công nghệ dữ liệu lớn, công ty có thể vừa cải thiện dịch vụ khách hàng vừa tạo thêm lợi nhuận mới cho mình. Ví dụ, thông qua việc phân tích nội dung cuộc trò chuyện của người dùng, công ty có thể dự đoán được mối quan tâm của khách hàng và từ đó có thể gửi đến người dùng vài thông tin bổ sung cũng như thông tin quảng cáo phù hợp.

Ngoài ra các công ty, tập đoàn viễn thông cũng là những người tiên phong trong lĩnh vực ứng dụng công nghệ dữ liệu lớn. Với việc nắm giữ trong tay một lượng lớn dữ liệu về khách hàng, các công ty viễn thông đã sớm nghĩ đến việc ứng dụng công nghệ dữ liệu lớn để xử lý khối lượng dữ liệu. Ví dụ, mạng di động Viettel Telecom với khoảng 55 triệu thuê bao di động phát sinh một lượng dữ liệu khổng lồ và liên tục. Băng thông kết nối Internet của các thuê bao data trong mạng bằng với băng thông cho phép trình diễn 50,000 bộ phim chất lượng cao HD. Mỗi một giờ, có khoảng 10 triệu cuộc gọi được gửi đi, mỗi cuộc gọi chứa khoảng 5MB thông tin. Trong khối lượng dữ liệu khổng lồ phát sinh liên tục trong mạng di động, việc chiết xuất ra các loại thông tin đem lại giá trị, không chỉ với nhà mạng mà còn giá trị với các doanh nghiệp, tổ chức kinh doanh bên ngoài. Nhà mạng muốn biết số lượng thuê bao có khả năng rời mạng vào tháng sau, cách thức để giữ họ lại với mạng của mình. Doanh nghiệp kinh doanh ô tô muốn tìm kiếm lớp khách hàng giàu có, thích sử dụng ô tô đắt tiền để gửi các thông tin khuyến mại, ưu đãi nhằm tiếp cận lớp khách hàng này. Hay một siêu thị muốn gửi các thông tin giảm giá khuyến mại đến lớp thuê bao đang ở xung quanh khu vực siêu thị trong một buổi sáng chủ nhật.

Các công ty hàng đầu trong lĩnh vực internet hay nội dung số luôn là các công ty đi đầu trong việc ứng dụng và triển khai các công nghệ mới. Đới với dữ liệu lớn, điều kiện đầu tiên để có thể bắt đầu xem xét, nghiên cứu và triển khai là phải có dữ liệu. Các công ty nêu trên đều là những công ty đang nắm trong tay lượng dữ liệu rất lớn và có nhu

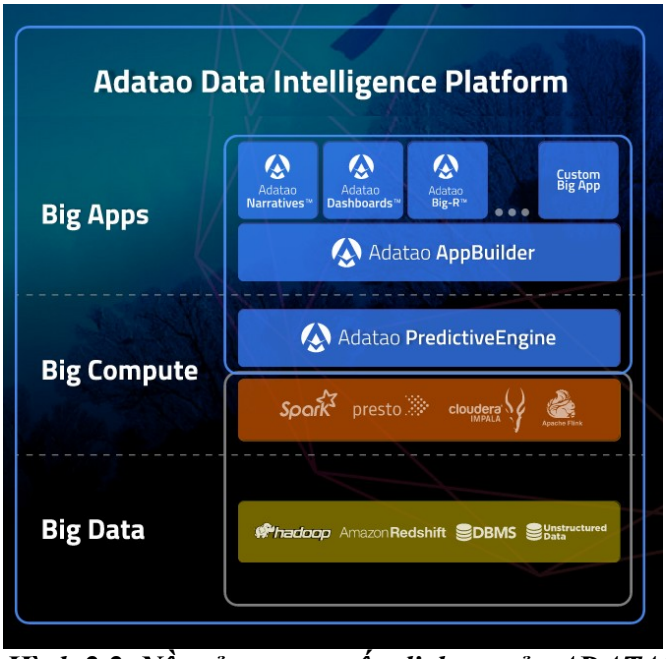
cầu phải xử lý chúng để nâng cao chất lượng dịch vụ của chính mình và sau đó là tìm kiếm thêm lợi nhuận từ việc phân tích dữ liệu. Tuy nhiên, trong thời kỳ bùng nổ thông tin trên internet như hiện nay, đặc biệt là các trang thông tin điện tử và mạng xã hội thì cơ hội để có được một lượng thông tin khổng lồ để phục vụ cho các bài toán dữ liệu lớn là không khó. Năm 2013 – 2014 có thể thấy sự phát triển nhanh chóng của các công ty khởi nghiệp trong lĩnh vực phân tích dữ liệu lớn. Có rất nhiều các công ty đã đưa ra các sản phẩm, dịch vụ dựa trên việc khai thác lượng dữ liệu lớn từ các mạng xã hội, các thông tin công cộng trên internet.

Đầu tiên, có thể kể đến công ty ANTS mới được thành lập từ năm 2014 nhưng đã hợp tác với hơn 100 khách hàng lớn như Sendo.vn, FPTshop.com.vn, 24H, Zing, Tiền Phong, Thanh Niên, Tuổi Trẻ,… và doanh thu của ANTS đã lên tới vài chục tỷ đồng/năm. ANTS là sàn giao dịch mua bán quảng cáo trực tuyến (ANTS Ad Exchange) đấu giá theo thời gian thực (Real-time Bidding) đầu tiên của Việt Nam, hoạt động dựa trên nền tảng công nghệ dữ liệu lớn hay nói một các đơn giản là có thể giúp khách hàng quảng cáo đúng người, đúng chỗ, đúng thời điểm bằng công nghệ mới. Một trong những khách hàng đầu tiên và điển hình của ANTS là trang bán hàng trực tuyến Lazada. ANTS thuyết phục được Lazada là giúp họ nâng số lượng đơn hàng đến từ các click quảng cáo lên gấp hàng chục lần so với việc đặt banner quảng cáo trên các website. Bài toán tiếp theo mà Lazada đặt ra cho ANTS là làm cách nào đó để trung bình mỗi ngày lượt click xem quảng cáo của Lazada tăng gấp 10 lần. Kết quả sau 1 tháng, Lazada đã có được con số như kỳ vọng và liên tục tăng lên vào các khoảng thời gian sau đó. Từ sự thành công với Lazada, đến thời điểm này, ANTS đã có hơn 100 khách hàng lớn, 3 tỷ lượt hiển thị quảng cáo/tháng, 10.000 vị trí quảng cáo/tháng được quản lý, 3.000 quảng cáo banner/tháng được xử lý đấu giá theo thời gian thực và khớp với người dùng. Dựa trên việc phân tích dữ liệu người dùng, ANTS đã giúp khách hàng tìm ra chính xác các nhóm người quan tâm. Đơn cử như với trường hợp của một khách hàng trong lĩnh vực thẩm mỹ, ANTS đã xây dựng một giải pháp tổng thể cho quảng cáo đa kênh (ANTS Multichannel Marketing Platform), giúp họ phân tích được dữ liệu để tìm ra chính xác các nhóm khách hàng mục tiêu và cá nhân hóa thông điệp quảng cáo theo đúng kênh, thời điểm truy cập mạng của những nhóm khách hàng này. Giải pháp này đã giúp khách hàng giảm 70% chi phí từ việc quảng cáo banner.



Hình 2.1. Thông tin do Younet media công bố về sự kiện BKAV chính thức công bố sự kiện ra mắt Bphone ngày 26/05/2015.

Một công ty khác đã khởi nghiệp rất thành công từ việc tận dụng nguồn dữ liệu lớn từ internet đó là YouNet Media. Được thành lập từ năm 2013 bởi công ty YouNetCo, đến nay YouNet Media đã trở thành một trong những công ty hàng đầu Việt Nam về theo dõi – quản trị và phân tích thương hiệu, thị trường và người dùng trên môi trường Internet. Khởi đầu của YouNet Media là xây dựng một nền tảng đám mây (Cloud Platform) để thu thập thông tin trên mạng xã hội (Facebook, YouTube…), diễn đàn, thương mại điện tử, tin tức trực tuyến… nhằm giúp doanh nghiệp theo dõi, quản trị thương hiệu, phòng chống khủng hoảng, định hướng chiến lược tiếp thị và đo lường thị trường. Đến nay hệ thống Social Listening & Market Intelligence của YouNet Media có khả năng thu thập và phân tích thông tin theo thời gian thực (real-time), tự động (automatic sentiment) và bao phủ trên 90% các nguồn tin tức và thảo luận từ các mạng xã hội (800 nghìn fanpages & nhóm và 20 triệu người dùng Facebook Việt Nam, Youtube, …), diễn đàn, cộng đồng, tin tức trực tuyến…và cho ra các thống kê cụ thể về thương hiệu, sản phẩm, sự kiện… chỉ trong vòng 1 tiếng đồng hồ. Dựa trên các kết quả phân tích này, doanh nghiệp sẽ biết được người dùng cảm thấy thế nào về sản phẩm của mình cũng như những đánh giá của khách hàng về công ty, sản phẩm cùng phân khúc… để từ đó kịp thời điều chỉnh hoặc phát triển các chiến lược marketing, bán hàng hay cải tiến sản phẩm, dịch vụ…Không chỉ lắng nghe, thu thập dữ liệu nội dung, ứng dụng của YouNet Media còn có thể giúp doanh nghiệp chăm sóc khách hàng trên tất cả các kênh trực tuyến, một giải pháp hữu hiệu hơn rất nhiều so với một tổng đài chăm sóc khác hàng truyền thống. Ngoài tính ưu việt về thời gian, chỉ trong vòng 1 tiếng, so với công cụ nghiên cứu thị trường truyền thống phải mất vài tuần, thậm chí là vài tháng, YouNet Media còn có thể giúp doanh nghiệp kịp thời xử lý khủng hoảng truyền thông trên mạng xã hội.



Hình 2.2. Nền tảng cung cấp dịch vụ của ADATAO

Ngoài các doanh nghiệp được thành lập trong nước thì các công ty do các Việt kiều thành lập ở nước ngoài rồi quay về Việt Nam xây dựng đội ngũ phát triển cũng góp phần làm phong phú thêm thị trường phân tích dữ liệu lớn và thúc đẩy xu hướng phát triển công nghệ dữ liệu lớn tại Việt Nam, nổi bật có thể kể tới công ty ADATAO. Đây là công ty được sáng lập bởi một người gốc Việt Nam (Christopher Nguyen). Mặc dù mới chính thức ra mắt tháng 12-2013 nhưng Adatao đã hoàn thiện 2 sản phẩm của mình cho hai nhóm đối tượng khác nhau. Nhóm thứ nhất bao gồm những chuyên gia nghiên cứu dữ liệu và kỹ sư phần mềm, Adatao cung cấp hệ thống pAnalytics dựa trên nền tảng Spark Apache nhằm giúp tương tác, chỉnh sửa và xây dựng các ứng dụng dữ liệu. Thông qua đó, các nhà phân tích dữ liệu doanh nghiệp có thể truy cập vào cơ sở dữ liệu trong một môi trường thân thiện và dễ dàng hơn với những ngôn ngữ phổ biến như R, Python, SQL và Java. Nhóm khách hàng thứ hai gồm những người dùng phổ thông sẽ sử dụng hgiải pháp pInsights được thiết kế giúp doanh nghiệp có thể truy cập dữ liệu một cách khá dễ dàng, dữ liệu sẽ được trích xuất ra ra file có định dạng text và đồ thị biểu diễn trực quan, giúp khách hàng nhanh chóng có được điều mong muốn giống như đang tra cứu Google. Ngay trong lần đầu tiên "demo" sản phẩm, Christopher Nguyễn và Adatao đã gây được chú ý với quỹ đầu tư Andreessen Horowitz và chính thức được đầu tư serie A với số tiền là 13 triệu USD.

Lĩnh vực ngân hàng luôn là lĩnh vực tiên phong trong việc ứng dụng công nghệ thông tin trong việc nâng cao hiệu quả quản lý và chất lượng dịch vụ, chính vì vậy, ngân hàng cũng là một nhân tố giúp thúc đẩy ứng dụng công nghệ dữ liệu lớn tại Việt Nam.

Tháng 3 năm 2015, IBM đã công bố công nghệ dữ liệu lớn và phân tích của hãng này đã được Ngân hàng Việt Nam Thịnh Vượng (VPBank) ứng dụng để đồng bộ hóa các dữ liệu khách hàng, theo mô hình một tổ chức kinh doanh định hướng dữ liệu tạo sự khác biệt trong các dịch vụ tài chính. Sự kiện này đã đưa VPBank thành một trong những ngân hàng đầu tiên tại VN triển khai công nghệ tiên tiến trong lĩnh vực Dữ liệu lớn để tạo ra sự khác biệt trong cách tiếp cận khách hàng và nâng cao hiệu quả kinh doanh nói chung. VPBank hiện đang cung cấp dịch vụ thông qua tất cả những ứng dụng phổ biến trên thị trường, từ hệ thống ngân hàng lõi (core banking), Internet banking, mobile banking đến hệ thống các loại thẻ quốc tế, thẻ nội địa, thẻ trả trước… Mỗi ngày có tới hàng triệu giao dịch được xử lý tại nhiều chi nhánh, điểm giao dịch và phòng ban khác nhau trên toàn hệ thống. Kho dữ liệu khổng lồ này đã nhanh chóng được VPBank nhìn nhận là nguồn tài sản quý giá và cần có chiến lược quản lý và chuyển đổi nguồn dữ liệu này thành những thông tin hữu dụng. Trọng tâm của chiến lược này là trang bị cho các chuyên gia tài chính, các giám đốc quan hệ khách hàng và các cán bộ tín dụng những thông tin chất lượng về sản phẩm và dịch vụ khách hàng, ví dụ như thói quen sử dụng từng loại sản phẩm, dịch vụ của từng đối tượng khách hàng, chi phí trực tiếp tính theo sản phẩm hay theo kênh cung cấp dịch vụ, các đối tượng khách hàng tiềm năng của từng loại sản phẩm ngân hàng, doanh thu dự kiến theo từng chiến dịch marketing trong tương lai, v.v.

Ứng dụng công nghệ dữ liệu lớn trong phục vụ quản lý nhà nước đang là một xu hướng chung trên thế giới. Nhiều nước trên thế giới như Mỹ, Nhật, Hàn quốc, EU đang rất thành công trong việc triển khai công nghệ dữ liệu lớn phục vụ cho quản lý nhà nước, đem lại nhiều lợi ích to lớn về kinh tế và xã hội, giúp cải thiện và thay đổi hoàn toàn cách thức cung cấp dịch vụ công cho người dân cũng như quản lý và giải quyết các vấn đề của chính phủ. Tại Việt Nam, vấn đề này cũng đã được đề cập đến trong các cuộc thảo luận, hội thảo khoa học về chính phủ điện tử.

Tiêu biểu trong việc chủ động tham gia vào ứng dụng công nghệ dữ liệu lớn và phục vụ quản lý nhà nước là công ty cổ phần giải pháp phần mềm Hanel. Hanel quan tâm đến việc tập hợp lại các nguồn dữ liệu mà các bộ, ngành, địa phương, doanh nghiệp sẵn có, để từ đó hình thành nguồn dữ liệu lớn để đưa ra các giải pháp giao thông tối ưu. Theo Hanel, hiện Việt Nam đã có một khối lượng dữ liệu thông tin khổng lồ, với hàng tỷ dữ liệu thông tin mỗi ngày, từ các nguồn dữ liệu đơn lẻ. Các dữ liệu đó có thể lấy từ các nguồn như hệ thống cân tải trọng mà Hanel đang xây dựng, hệ thống camera giám sát, xử lý hình ảnh của lực lượng công an, hệ thống giám sát hành trình ôtô của Tổng cục Đường bộ VN, thống kê mật độ giao thông dựa trên dữ liệu thuê bao di động của Viettel, thông tin quản lý hàng hóa, hành khách của (của Hải quan và đơn vị vận tải); chưa kể những

thông tin được chia sẻ trên mạng xã hội, trên facebook, video giao thông…Theo Hanel, khi các nguồn dữ liệu được tập hợp thành dữ liệu lớn và ứng dụng công nghệ thông tin để xử lý dữ liệu sẽ giúp người tham gia giao thông và cơ quan quản lý đưa ra quyết sách tối ưu để giao thông trở nên thông minh hơn. Công nghệ dữ liệu lướn có ý nghĩa đặc biệt quan trọng, mang lại những lợi ích cụ thể cho giao thông thông minh như: giúp dự báo được các khả năng trong tương lai; cụ thể hóa các xu hướng, hiện trạng để ứng dụng công nghệ thông tin giải quyết vấn đề và phục vụ phát triển xã hội; tối ưu hóa các dữ liệu gốc, có thể ứng dụng cho Vận tải đa phương thức. Khi dữ liệu lớn được hình thành, sẽ nói cho chúng ta chính xác thông tin, mang đến kết luận chứ không phải suy luận nữa. Ví dụ như dữ liệu lớn sẽ cho ta những chỉ dẫn tức thời về giao thông, hay cho thông tin chính xác để biết rằng con đường làm đó cần mở rộng ra 6 làn hay 8 làn, đoạn đường đó cần lắp các camera giám sát ở khoảng cách bao nhiêu. Đề xuất việc tích hợp và kết nối các nguồn dữ liệu đơn lẻ, Hanel đồng thời khẳng định có thể triển khai và thực hiện được dự án về giải pháp về BigData, với sự hợp tác từ các doanh nghiệp, Bộ GTVT, cũng như cần thiết có thêm chính sách hỗ trợ. Đây là thời kỳ cần thiết tập hợp các nguồn dữ liệu đơn lẻ sẵn có thành BigData để tối ưu hóa giải pháp và ứng dụng công nghệ thông tin, điện tử, viễn thông vào giao thông thông minh. Để thực hiện có hiệu quả và mang tính thực tiễn cao, cần có Một khung chính sách và chế tài triển khai gồm: Sự hợp tác để thu thập được đầy đủ dữ liệu từ các nguồn khác nhau (Dữ liệu hóa); Tích hợp và kết nối các nguồn dữ liệu đang bị rời rạc, chia cắt; Chọn lọc và sử dụng dữ liệu có giá trị; Kiểm soát dữ liệu và tính an toàn thông tin.

Bên cạnh các đề xuất, các hội nghị, hội thảo khoa học về ứng dụng dữ liệu lớn trong quản lý nhà nước, một số đơn vị cũng đã chủ động ứng dụng công nghệ dữ liệu lớn nhằm nâng cao khả năng điều hành quản lý lĩnh vực của mình. Tổng cục du lịch Việt Nam đã sử dụng dịch vụ của công ty InfoRe - một công ty chuyên về phân tích dữ liệu – để phân tích sắc thái thông tin du lịch tự động dựa trên mọi thông tin xuất bản trên báo điện tử, diễn đàn và mạng xã hội Facebook về vấn đề du lịch.

Ngoài các công ty đang ứng dụng công nghệ dữ liệu lớn trong việc cung cấp các sản phẩm, dịch vụ của mình hay các công ty cung cấp dịch vụ trực tiếp trên công nghệ dữ liệu lớn thì còn có các công ty tuy không trực tiếp triển khai hạ tầng, công cụ xử lý dữ liệu lớn nhưng cũng đóng góp vào bức tranh chung của hiện trạng ứng dụng công nghệ dữ liệu lớn tại Việt Nam. Các công ty này sở hữu trong tay một lượng dữ liệu lớn từ việc cung cấp dịch vụ của mình tới khách hàng, tuy nhiên chưa thực sự có nhu cầu khai thác lượng dữ liệu này mà thường hợp tác với bên thứ 3 để khai thác. Điển hình có thể kể đến đó là dịch vụ GrabTaxi đang rất phổ biến tại các nước Đông Nam Á, trong đó có Việt

Nam. GrabTaxi là một công ty công nghệ, khởi nghiệp trong ngành vận tải Đông Nam Á; cung cấp các giải pháp "vận tải thông minh" cho các thị trường tăng trưởng nhanh trong khu vực thông qua việc tạo ra các sàn giao dịch điện tử cho dịch vụ vận tải. Công ty GrabTaxi cung cấp nhiều dịch vụ kết nối vận tải khác nhau tại 6 quốc gia, bao gồm Malaysia, Indonesia, Philippines, Singapore, Thái Lan, Việt Nam. Dịch vụ GrabTaxi có mặt tại hơn 20 thành phố trên khắp Đông Nam Á, trong đó có Thành phố Hồ Chí Minh và Hà Nội. GrabTaxi đã hợp tác với Ngân hàng Thế giới (World Bank) để chống tắc nghẽn giao thông và cải tiến an toàn đường bộ cho hơn 620 triệu cư dân trong khu vực Đông Nam Á. Sự hợp tác giữa World Bank và GrabTaxi cung cấp miễn phí nền tảng nguồn dữ liệu mở OpenTraffic. Sự hợp tác giữa World Bank và GrabTaxi được chính quyền địa phương thử nghiệm tại các thành phố GrabTaxi đang hoạt động như: Cebu, Manila, Davao City, Jakarta, TP.HCM và Hà Nội. Thông qua việc hợp tác này, GrabTaxi sẽ giúp các cơ quan giao thông vận tải địa phương giám sát tình trạng giao thông theo thời gian thực và thu thập dữ liệu lịch sử di chuyển. Theo đó, chính phủ sẽ có thể ra những quyết định đúng đắn và có cơ sở hơn về những vấn đề tưởng chừng ngoài tầm tay, bao gồm kế hoạch đèn giao thông, điều khoản vận tải công cộng, nhu cầu cơ sở hạ tầng đường phố, quản lí giao thông khi xảy ra tình huống khẩn cấp và quản lí nhu cầu đi lại.

Kết luận

Với hiện trạng ứng dụng công nghệ dữ liệu lớn tại Việt Nam hiện nay, có thể nói công nghệ dữ liệu lớn tại Việt Nam mới chỉ bắt đầu phát triển. Với sự tham gia sớm của các công ty lớn hoạt động trong lĩnh vực cung cấp nội dung, thương mại điện tử, viễn thông đã giúp Việt Nam sớm tiếp cận được với công nghệ dữ liệu lớn. Tuy nhiên, các công ty này mới chỉ dừng lại ở việc ứng dụng công nghệ dữ liệu lớn để duy trì lợi thế cạnh tranh của mình so với đối thủ và nâng cao lợi nhuận. Các dữ liệu này chưa thực sự được khai thác hết, hơn nữa nếu các dữ liệu mà các công ty sở hữu được chia sẻ để cùng nhau khai thác bằng công nghệ dữ liệu lớn có thể tạo ra những giá trị mới giúp công ty có thể thay đổi hoàn toàn về chất.

Tuy nhiên với sự tham gia vào thị trường dữ liệu lớn, các công ty này đã giúp thúc đẩy sự phát triển mạnh mẽ của các công ty khởi nghiệp tham gia vào lĩnh vực phân tích dữ liệu lớn. Các công ty khởi nghiệp này đã rất thành công trong việc ứng dụng công nghệ dữ liệu lớn để giải quyết các nhu cầu của khách hàng, đem lại lợi ích cho cả công ty và khách hàng. Với các lợi ích đem lại rõ rệt như vậy, chắc chắn rằng các công ty này sẽ tạo thêm động lực, nguồn cảm hứng để công nghệ dữ liệu lớn có thể nở rộ tại Việt Nam trong thời gian tới. Tuy nhiên, xu hướng hiện này chủ yếu là tập trung vào việc phân tích các thông tin công cộng trên internet như mạng xã hội, các trang thông tin điện tử, diễn

đàn để nhằm nâng cao hiệu quả marketting trực tuyến. Hi vọng trong thời gian tới sẽ có nhiều sản phẩm, dịch vụ phân tích dữ liệu lớn trong các lĩnh vực khác như y tế, giáo dục, giao thông thông minh và đặc biệt là quản lý nhà nước.

Trong quản lý nhà nước, công nghệ dữ liệu lớn gần như chưa được triển khai và ứng dụng. Việc đưa công nghệ dữ liệu lớn vào phục vụ quản lý nhà nước mới dừng lại ở mức đề xuất và ứng dụng thử nghiệm, đơn lẻ. Đây là một trong lĩnh vực mà công nghệ dữ liệu lớn cần được ứng dụng sớm. Nếu được triển khai và ứng dụng đúng thì công nghệ dữ liệu lớn có thể trở thành đòn bẩy để giúp cải cánh một cách toàn diện hệ thống quản lý nhà nước, góp phần thúc đẩy phát triển kinh tế xã hội, đưa Việt Nam thoát khỏi bẫy thu nhập trung bình, vượt lên so với các nước khác trong khu vực.

Việt Nam đang ngày càng gia tăng tốc độ phát triển và hội nhập với các xu hướng công nghệ thế giới. Với hơn 30 triệu người dùng Internet và hơn 15 triệu ngươi dùng Mobile Internet làm cho Việt Nam đang đứng trước một cơ hộ vô cùng lớn về khai thác dữ liệu lớn. Sẽ có những doanh nghiệp Việt Nam khai thác thành công dữ liệu lớn với doanh số hàng trăm triệu USD trong vòng 5 năm tới. Đặc biệt, giai đoạn 2014-2016, xu hướng Mobile và lượng người dùng Internet 3G sẽ tiếp tục tăng mạnh. Các dịch vụ kết nối OTT (Over-the-top) và truyền thông xã hội đóng góp hơn 80% phương thức giao tiếp online, video online và nội dung số mobile. Điều này góp phần đẩy mạnh xu hướng truyền thông số đa phương tiện, đa màn hình (PC, smartphone , tablet, smart TV) sẽ bùng nổ với độ phủ hơn 50% dân số Việt Nam. Việt Nam là một kho "vàng" dữ liệu vô cùng lớn cho việc ứng dụng Big Data.

2.2. Ảnh hưởng của công nghệ dữ liệu lớn đến phát triển kinh tế xã hội

Công nghệ xử lý dữ liệu lớn (BigData) không đem lại các ảnh hưởng trực tiếp tới việc sản xuất, kinh doanh của các tổ chức, doanh nghiệp, hay nói một các khác việc ứng dụng công nghệ BigData không tạo ra lợi nhuận trực tiếp cho các tổ chức doanh nghiệp này. Tuy nhiên, công nghệ Big-data lại ảnh hưởng tới các tổ chức doanh nghiệp ở mức chiến lược và điều hành, giúp tạo ra lợi thế cạnh tranh so với các đối thủ.

2.2.1. Big Data ành hưởng đến định hướng mục tiêu thị trường

Dữ liệu lớn có thể thay đổi cách thức các công ty xác định thị yếu khách hàng của họ, các công ty có thể đẩy mạnh các chiến lược tiếp thị cũ bằng cách sử dụng các công cụ dữ liệu lớn mới. Chiến lược thâm nhập thị trường có thể tận dụng dữ liệu lớn để tạo ra các thông tin quảng bá giúp giữ khách hàng hiện có và nâng cao doanh số. Tương tự như

vậy đối với khách hàng mới, giúp cải thiện được mức độ tin tưởng.

Việc thúc đẩy sự hấp dẫn của một công ty, và tăng cường sự hiểu biết về thị trường nhằm bán ra các sản phẩm khác nhau cho cùng một đối tượng khách hàng. Các công ty không chỉ bắt đầu phân tích một lượng lớn các giao dịch có liên quan đến các phương tiện truyền thông xã hội để hiểu sở thích khách hàng của họ, mà họ còn tạo ra các dịch vụ mới cho khách hàng.

Rõ ràng, khi tham gia vào một thị trường mới cần phải nắm bắt được sức mạnh của dữ liệu lớn, và nó là một thách thức thật sự. Các công ty không còn cần phải tốn nhiều công sức cho việc tiếp cận một thị trường rộng lớn. Thay vào đó họ có thể sử dụng phân tích dữ liệu để xác định thị trường ngách mới hoặc thậm chí chia nhỏ thị trường hiện có thành các thị trường nhỏ hơn để tăng sức cạnh tranh. Kết hợp với những tiến bộ trong tiếp thị truyền thông tự động, chúng ta đang hướng tới thời đại của quảng cáo đại chúng. Như vậy, mục tiêu cuối cùng của các nhà tiếp thị là quảng cáo đại chúng. Tổng hợp và phân tích dữ liệu lớn hứa hẹn cung cấp cho các doanh nghiệp có cái nhìn thực tế về thị yếu của khách hàng. Dữ liệu của mạng truyền thông xã hội được thu thập một cách bí mật, bởi vì hầu hết chúng ta đều đưa ra các bình luận và nhận xét trong mạng xã hội, trả lời các câu hỏi trong các cuộc điều tra. Điều này giúp giảm bớt các chi phí nghiên cứu thị trường, các sai lầm trong bán hàng, tiếp thị, chiến lược kinh doanh của công ty. Bằng việc phân tích tâm lý về những bài viết về các lính vực của đời sống của khách hàng trên mạng xã hội có thể đưa ra được các sản phẩm và dịch vụ mới. Phân tích hành vi của khách hàng để đưa ra dự đoán, cho phép các nhà tiếp thi phát hiện lệch lạc trong mô hình kinh doanh. Khả năng truy cập các thông tin cá nhân trên mạng xã hội tweets , Facebook và LinkedIn, làm giảm nhẹ tính hoài nghi về nguồn gốc của thông tin.

Do dó, dữ liệu lớn đã đưa ra được công cụ để quản lý các mối quan hệ thị trường, là một công cụ tuyệt vời giúp có được các thông tin chính xác, từ đó có các chiến lược tiếp thị tại các thời điểm thích hợp giúp khách hàng đưa ra quyết định mua hàng. Mashable.com là dịch vụ với hơn 20 API (Application Programming Interface) có thể giúp mọi người thu thập thông tin mà họ mong muốn từ nhiều nguồn khác nhau như Facebook và Twitter hoặc thậm chí các văn bản trong các bài báo và blog.

Chỉ với một công cụ, họ có thể biết được thông tin phản hồi về các chương trình khuyến mãi và các hoạt động quảng bá khác một cách nhanh nhất (Provost and Fawcett 2013). Dựa trên các dấu vết về địa chỉ IP của máy tính về các hoạt động trên mạng, các sản phẩm khách hàng đã mua , đánh giá và quan tâm, vị trí địa lí của khách hàng, các thông tin cá nhân, các nhà tiếp thị sẽ đưa ra các sản phẩm phủ hợp cho khách hàng, hay

một nhóm khách hàng.

Tiền năng của xu hướng Internet of Things (IoT) là rất lớn, nó giúp cho các công ty tạo ra lợi thế cạnh tranh và mô hình kinh doanh thật sự khác biệt.

2.2.2. BigData tạo ra sự đổi mới trong định hướng thiết kế

Sự kết hợp của các nguồn dữ liệu lớn với các công nghệ mới nổi khác có thể truyền cảm hứng cho các xu hướng thiết kế. Những sáng tạo mang tính đột phá ban đầu sẽ khiến khách hàng không hứng thú nhưng sau đó họ sẽ thích hơn.

Ví dụ như Apple không thay đổi cách chúng ta thực hiện cuộc gọi từ điện thoại di động, nhưng nó thay đổi cách nhìn của chúng ta về điện thoại. Nó có thể là tất cả những gì bạn muốn, từ thiết kế phối màu cho căn phòng của những đứa trẻ, tới giết thời gian với trò chơi "angry birds", tới kiểm tra những bản tin mới, tới xem một bộ phim, tới đo kích thước căn phòng của bạn. Một chiếc Iphone không còn là chiếc điện thoại thông thường nữa, nó là một công cụ đa tính năng và Apple không phải là một công ty điện thoại, đó là công ty đã làm thay đổi cuộc sống của chúng ta, và hầu hết mọi người đều thích sản phẩm này. Các sản phẩm của Apple không còn là những sản phẩm của quy trình công nghiệp nữa, nó là biểu tượng và tạo ra nét đặc trưng khác biệt. Mua một sản phẩm để khẳng định bản thân. Do đó đổi mới sản phẩm không còn chỉ là về bản chất sản phẩm, mà nó là chiến lược chia sẻ ý tưởng sản phẩm với khách hàng để trở thành một phần của cộng đồng.

Cũng với cách đổi mới sản phẩm không chỉ là về bản chất sản phẩm, mà còn chia sẻ ý nghĩa, mô hình thiết kế cũng như chia sẻ ý nghĩa về những gì một tổ chức đại diện cho. Ví dụ, Asos.com là một nhà bán lẻ thời trang nhưng không chỉ là về quần áo. Công ty đã đầu tư vào một trang web không chỉ là về thời trang mà nó còn là nơi giao dịch thời trang, cho phép bất cứ ai, bất cứ nới nào trên thế giới có thể bán các sản phẩm thời trang đến tất cả mọi người trên toàn thế giới, và công ty sẽ thu một khoản hoa hồng khoảng 10% cho mỗi sản phẩm, đó là một mô hình kinh doanh khá độc đáo.

Điều gì sẽ tạo ra xu hướng thiết kế từ Big Data? Đến nay, mọi người đều hiểu "Big data = Social Data" , tuy nhiên các sáng tạo đổi mới đều xuất phát từ Internet of Things (IoT). Các hệ thống tiên tiến được trang bị các cảm biến và hệ thống tự động hỗ trợ giải quyết, chứ không chỉ là phải chỉ là tự động. Các hệ thống tiên tiến đã thay đổi mô hình của chúng ta, làm thay đổi các giá trị cốt lõi cái gì đúng, cái gì sai. Chúng hứa hẹn sẽ trở thành thế giới các cỗ máy cùng nhau hoạt động "always-on, always-aware,

always-connected, always-controllable". Điều này sẽ ảnh hưởng đến hầu hết các lĩnh vực cơ sở hạ tầng. Với công nghệ này sẽ biến những thứ bình thường thành các dịch vụ mới. Ví dụ như việc đi lại sẽ có những thay đổi đáng kể trong vài thập kỷ tới. Chiếc xe tương lai của bạn có thể trở thành người lái xe cẩn thận. Bạn có thể yêu cầu nó đến đón và đưa bạn tới nơi làm việc, chiếc xe sẽ giao tiếp với các xe khác trên đường đi để có thể lái xe một cách an toàn, và tất nhiên nó sẽ tự điều chỉnh lượng tiêu thụ năng lượng, hay sử dụng các nguồn năng lượng xanh như năng lượng mặt trời, hydro. Và có lẽ, bạn thậm chí sẽ không cần phải sở hữu nó.

Từ năm 2008 nhiều thiết bị được kết nối với internet hơn, mở ra một cơ hội kinh doanh rất lớn. Theo như bộ tài chính của chính phủ Anh, thị trường toàn cầu về các giải pháp thành phố thông minh sẽ đem lại hơn 400 tỷ USD mỗi năm vào năm 2020.Điều này nghe có vẻ rất lớn, nhưng nó chỉ là một phần nhỏ của chi tiêu cơ sở hạ tầng toàn cầu (Townsend 2013). Mặc dù, sự ra tăng các hành động của các civic hacker, công nghệ mã nguồn mở, và dữ liệu chính phủ vẫn đang làm việc với nhau tạo ra những công nghệ thông minh giúp cho các thành phố an toàn, dân chủ và thân thiện hơn (Townsend 2013). Và trong khi điều này là một thách thức cho các tổ chức thu lợi nhuận, nhưng nó là vô cùng quan trọng cho xã hội, là mục tiêu cho những tổ chức phi lợi nhuận muốn làm thay đổi xã hội chứ không phải vì tiền.

Có lẽ ý tưởng đổi mới không còn nằm bên trong tổ chức, chúng ta đã bước vào kỷ nguyên của sự đổi mới peer-to-peer, nơi những ý tưởng và giải pháp được bắt nguồn và xây dựng bởi số đông.

2.2.3. Big-data kích thích sự sáng tạo tập thể

Dữ liệu lớn không những làm thay đổi cách chúng ta tiếp cận thị trường với một sản phẩm hoặc dịch vụ, mà còn thay đổi cách chúng ta thiết kế và sáng tạo ra các sản phẩm, dịch vụ.

"Sáng tạo mở" được dựa trên nguyên tắc là các ý tưởng sáng tạo không chỉ bị hạn chế bên trong một tổ chức. Quan điểm này cho thấy một số nguyên tắc rất khác nhau về cách các tổ chức thành công nên đối xử ra sao. Ví dụ như, nó xóa bỏ khái niệm "non-invented here", các ý tưởng hữu ích có thể ở bất cứ nơi nào, chúng có thể đến từ các trường đại học, các nhà cung cấp, khách hàng, các công ty khác, công chúng. Sở hữu trí tuệ (IP) là một tài sản kinh doanh có thể được mua và bán để thu lợi nhuận. Nó là một vấn đề của người tạo ra và người bên ngoài vì lợi ích chung. Dữ liệu lớn có thể trở thành một khái niệm ở một cấp độ mới. Được xem như yêu cầu về sản phẩm, truyền thông xã

hội có thể tìm ra các khiếu nại của khách hàng và danh sách các sản phẩm mong muốn. Nhưng đó không phải điều duy nhất, nó giúp chúng ta có cái nhìn sâu sắc hơn về thị trường để có những đáp ứng nhanh chóng. Sự sáng tạo mở được hỗ trợ bởi các sáng tạo trung gian, như nền tảng Innocen tive, nó phù hợp với các công ty đang tìm kiếm các giải pháp. Các công ty lớn có thể tận dụng lợi thế của những phát triển bên ngoài thành của mình. Một phần, AstraZeneca đã thành lập một quỹ sáng tạo khoảng $100,000 để tìm ra một giải pháp triệt để cho căn bệnh Targeted Delivery of Oligonucleotides, với mong muốn sẽ giúp điều trị hiệu quả các khối u. Các chuyên gia thuê ngoài thì luôn có sẵn cho các công ty nhỏ. Đây là một ý tưởng tốt, các kinh nghiệm gia công phần mềm là bình đẳng giữa các công ty lớn và bé, và dữ liệu lớn giúp cho các tổ chức nhỏ có chố đứng. Trong thời đại dữ liệu lớn, không chỉ dữ liệu và ý kiến là mở cho tất cả, mà còn là các ý tưởng, hay thậm chí là ý tưởng kinh doanh cũng hoàn toàn mở. Các trung tâm sáng tạo mọc lên khắp toàn cầu cung cấp sự hỗ trợ cho những ai có ý tưởng gây dựng một doanh nghiệp, tư vấn và hỗ trợ tài chính. Các kênh huy động tài chính Crowdfunding tạo ra các nguồn tài trợ đến từ cộng đồng, bằng cách huy động vốn cho các hoạt động dựa trên sự ủng hộ của công chúng. Ví dụ như Kickstarter.com là cộng đồng của những người cùng làm việc với nhau nó dựa trên nền tảng kênh huy động tài chính Crowdfunding cho phép mọi người quyên góp, đặt hàng trước, hoặc nhận một cổ phần trong công ty (Kickstarter 2014).

Crowdfunding hoạt động trong nhiều lĩnh vực từ sáng tác truyện tranh, tranh ảnh tới thực phẩm và các ý tưởng kinh doanh công nghệ. Ví dụ như, Lix, một ý tưởng về bút in 3D cần £30,000 thì đã được quyên góp tới £485,249 từ 5388 người ủng hộ trong 26 ngày , hầu hết trong số họ đều đặt mua bút trước.

Phương tiện truyền thông và dữ liệu lớn hỗ trợ lẫn nhau. Xác định các ý tưởng, thử nghiệm, các sản phẩm và các kịch bản được đem ra thảo luận, phát triển và cập nhật liên tục trong cộng đồng và được kiểm tra bằng cách sử dụng dữ liệu lịch sử và hiện tại để dự đoán phản ứng của thị trường. Sử dụng phân tích dự báo , ví dụ như sáng tạo có thể biết được các kịch bản tốt nhất và so sánh với các lựa chọn khác nhau.

2.2.4. Big Data ảnh hưởng đến việc định hướng mô hình kinh doanh

Sự xuất hiện của điện thoại di động là một nền tảng phương tiện truyền thông có khả năng sinh lợi cho các nhà tiếp thị. Các thiết bị di động cho phép các giao dịch buôn bán tức thời như một dạng đại lý trung gian mới, như các kênh phân phối hàng ngày tận dụng những lợi thế của tiến bộ công nghệ. Groupon, dịch vụ dựa trên địa điểm, là tên tuổi tiềm năng nhất trong lĩnh vực quảng cáo, khuyến mãi, nó là nơi mà người dân tại mỗi

vùng địa lý có thể tìm thấy giao dịch tại địa phương về bất cứ điều gì, bất cứ lúc nào, bất cứ nơi nào. Họ theo dõi vị trí của hàng triệu thuê bao của họ trên toàn cầu để kết hợp chúng với những giao dịch tại địa phương trong khu vực của họ dựa trên lợi ích của họ. Với sự phát triển của các kênh phân phối, các kênh phân phối tổng hợp như Yipit là một kênh phân phối phục vụ tất cả các nhu cầu của khách hàng .

Tuy nhiên, chúng ta vẫn chưa thực sự có nhiều sáng tạo khi sử dụng dữ liệu lớn. Cho đến nay, mong muốn của mọi người đều bị điều khiển bởi các quan nhiệm cũ kỹ, vẫn tìm cách sử dụng dữ liệu lớn để làm những điều vấn đề cũ cho hiệu quả và linh hoạt hơn. Với dữ liệu lớn đang làm thay đổi luật lệ ưu đãi đối với những người có kỹ năng công nghệ và phân tích, không thể tránh khỏi việc các công ty công nghệ sẽ đa dạng hóa các loại hình, tham gia vào các lĩnh vực truyền thống cho dù không phải là lĩnh vực của họ hay mua lại các công ty nhỏ hơn và có tham vọng trên thị trường. Lưu trữ dữ liệu và khả năng phân tích tạo ra lợi thế cạnh tranh cho những công ty muốn tham giavào các lĩnh vực khác.

Ví dụ như trong lĩnh vực bán lẻ, quảng cáo trực tuyền và cross-selling là những động lực chính cho việc sử dụng dữ liệu lớn. Nó sẽ không có gì đáng nói, tuy nhiên cuộc chơi bị điều khiển bởi các công ty công nghệ lớn đang làm thay đổi mô hình kinh doanh trong lịnh vực bán lẻ. IBM vào tháng 4 năm 2014 đã công bố việc mua lại Fluid để phát triển một ứng dụng di động ảo mua sắm cá nhân, dựa trên công nghệ trực quan Watson giúp cho mọi người có thể tương tác bằng chính ngôn ngữ của họ. Điều đó đã giúp chúng ta hiểu được tiền năng của việc kết hợp trí thông minh nhân tạo AI với phân tích dữ liệu lớn.

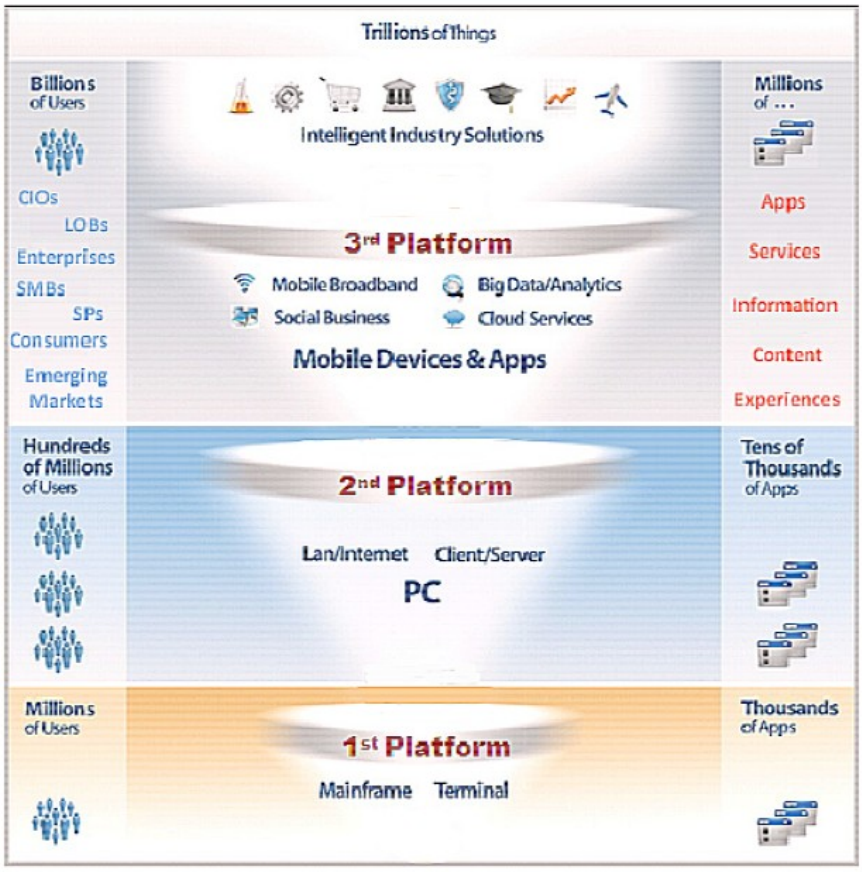
Ngoài ra, Google cũng đa tham gia vào lĩnh vực du lịch, một phần tài trợ cho các đại lý du lịch của họ thông qua các chi phí quảng cáo trực tuyến. Gần 70% lượng thẻ du lịch được thực hiện trực tuyến. Khoảng 70-90% các quảng cáo của các cơ quan du lịch trực tuyến được chi cho Google. Google biết cách các đối thủ cạnh tranh thực hiện, sở hữu các kênh gần với khách hàng, sử dụng sự phân tích, giữ khách hàng, có khả năng lưu trữ và với số lượng người dùng trung thành. Với khả năng của các trung tâm dữ liệu, Google có thể tham gia vào bất kỳ lĩnh vực nào mà họ muốn.

Có lẽ sẽ có nhiều sáng tạo ấn tượng hơn trong lĩnh vực thành phố thông minh liên quan đến “Internet of Things” (IoT). Dữ liệu được tạo ra từ các cảm biến được gắn vào các vật dụng quen thuộc như thùng rác, bánh xe đạp, đường ống nước, đèn giao thông, kết hợp với trí tuệ nhân tạo AI có thể tạo thành một mạng lưới cơ sở hạ tầng thành phố tự quản. IBM, Microsoft và Cisco đều đang chạy đua giành các dự án thành phố thông minh

(Ratti 2014). Đó là vai trò của dữ liệu lớn trong việc tạo ra những đổi mới đột phá mà sẽ thay đổi cách chúng ta sống.

Liệu các công ty lớn có dành được lợi thế cạnh tranh trong lĩnh vực này? Có hay không. Một mô hình cạnh tranh đang nổi lên phù hợp với quan niệm mở của mô hình thị trường người dùng doanh nghiệp prosumer ví dụ như là một khách hàng cũng là người sản xuất ra sản phẩm và dịch vụ. Đây là quan niệm về dịch vụ Service Mashups–compositions , các phân hệ dịch vụ được ghép nối với nhau bởi chính người tiêu dùng. Do đó, vai trò của công ty là đưa ra các phân hệ dịch vụ mà bằng cách nào đó có thể dễ dàng kết hợp với các phân hệ khác để tạo thành một dịch vụ. Các doanh nghiệp đang cùng phát triển web và điện toán đám mây dựa trên môi trường tương tác, nơi dịch vụ IOT có thể dễ dàng kết hợp với nhau.

2.3. Ảnh hưởng của công nghệ dữ liệu lớn đối với chính phủ



Những tiến bộ trong công nghệ và sự gia tăng khối lượng thông tin đang thay đổi cách thức vận hành các hoạt động nghiệp vụ trong nhiều lĩnh vực, trong đó có Quản lý Nhà nước. Lượng dữ liệu phát sinh và được số hóa lưu trữ trong hoạt động Quản lý Nhà nước đang gia tăng do sự tăng trưởng nhanh chóng của điện thoại di động, các ứng dụng, các thiết bị cảm biến thông minh, giải pháp điện toán đám mây, các cổng thông tin đối

thoại trực tiếp với người dân. Thông tin số hóa ngày càng mở rộng và trở nên phức tạp hơn đồng nghĩa với sự phức tạp cũng gia tăng trong các công tác quản lý, xử lý, lưu trữ, bảo mật và phân phối thông tin. Các tổ chức cần một công cụ mới giúp họ nắm bắt, tìm kiếm, phát hiện và phân tích những dữ liệu phi cấu trúc của mình. Chính phủ Mỹ cũng nhận ra rằng, thông tin là một tài sản chiến lược, cần phải bảo vệ, nâng tầm và phân tích cả hai loại dữ liệu thông tin có cấu trúc và phi cấu trúc để có thể đáp ứng tốt hơn những yêu cầu của các nhiệm vụ đặt ra. Các tổ chức nỗ lực phát triển theo định hướng dữ liệu để thực hiện thắng lợi nhiệm vụ đồng thời cũng là đặt nền tảng cho sự tương quan phụ thuộc giữa các đối tượng sự kiện, con người, quy trình và thông tin. Hình dưới đây minh họa cho khái niệm nền kinh tế thông minh, với các giải pháp công nghiệp thông minh được triển khai ở lớp thứ 3, nơi công nghệ định hướng sự tăng trưởng bùng nổ của thông tin số hóa như thu thập và phổ biến thông tin từ hàng nghìn tỷ thiết bị như điện thoại thông minh và cảm biến nhúng.

Các giải pháp có tính ứng dụng cao của Chính phủđược tạo ra từ một sự kết hợp của hầu hết các công nghệ mang tính đột phá:

- Thiết bị di động và các ứng dụng
- Dịch vụ đám mây
- Mạng xã hội
- Công nghệ lưu trữ và phân tích Big Data

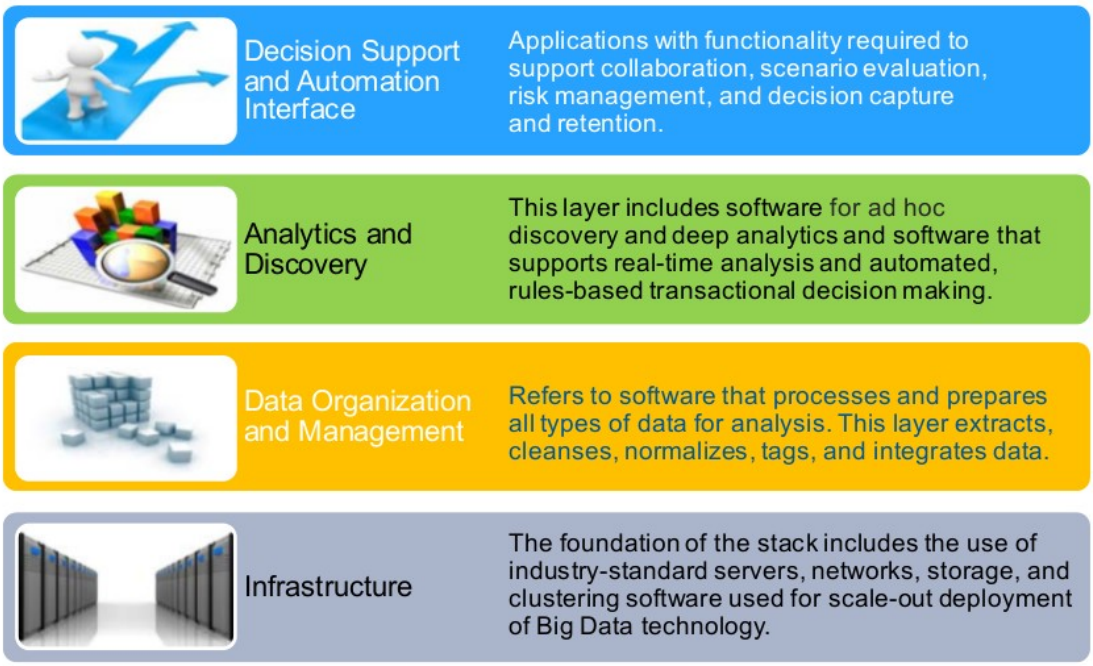
Việc sử dụng máy tính xách tay, điện thoại thông minh, máy tính bảng trong hoạt động Quản lý Nhà nước sẽ tiếp tục gia tăng. Điện toán di động cho phép làm việc từ xa hiệu quả và đảm bảo tính liên tục của các hoạt động và hiệu suất làm việc trong khắc phục thiên tai. Các công nghệ truyền thông mạng xã hội tạo điều kiện cho công dân thực hiện quyền làm chủ trong công tác Quản lý Nhà nước. Khi các kỹ năng truyền thông mạng xã hội được cải thiện tốt hơn, giá trị của sự hợp tác này càng được thấy rõ, đặc biệt là tạo nên tính mở, minh bạch cho công tác Quản lý Nhà nước, tạo điều kiện thuận lợi cho việc cung cấp các nghiệp vụ Quản lý Nhà nước đến người dân.

Về lâu dài, các công nghệ này sẽ là công cụ quan trọng để đối phó với sự phức tạp của việc gia tăng các thông tin số. Big Data là một trong những giải pháp thông minh, cho phép Chính phủ đưa ra những quyết định tốt hơn dựa trên việc phân tích một khối lượng lớn dữ liệu - có liên quan và không liên quan, có cấu trúc và phi cấu trúc.

Nghiên cứu cho thấy rằng công nghệ Big Data là hạt nhân của nền kinh tế thông minh và các giải pháp công nghệ khác. Công nghệ Big Data là một thế hệ công nghệ và kiến trúc mới, được thiết kế để trích xuất giá trị từ một khối lượng rất lớn và đa dạng về loại dữ liệu bằng cách cho phép nắm bắt, phát hiện và phân tích dữ liệu ở tốc độ cao. Các công cụ nắm bắt, tìm kiếm, phát hiện và phân tích dữ liệu giúp các tổ chức có được thông tin hữu ích từ các dữ liệu phi cấu trúc - chiếm đến hơn 90% số lượng dữ liệu.

Big Data được định nghĩa bao gồm phần cứng, dịch vụ và các phần mềm tích hợp nhằm tổ chức, quản lý, phân tích và trình diễn dữ liệu được đặc trưng bởi dung lượng(Volume), tốc độ xử lý (Velocity), tính đa dạng (Variety), giá trị (Value) khai thác và thừa hưởng từ dữ liệu phục vụ cho các nhiệm vụ tiếp theo.

Kiến trúc hạ tầng nền tảng của công nghệ Big Data bao gồm các máy chủ chuẩn công nghiệp, mạng và phần mềm clustering. Kho lưu trữ đang trở thành một vấn đề chiến lược khi mà sự bùng nổ của dữ liệu có cấu trúc và phi cấu trúc dẫn đến sự lo ngại về việc dự phòng, phục hồi và lưu trữ dữ liệu phát sinh trong các hoạt động Quản lý Nhà nước. Ngân sách có hạn cùng với sự gia tăng khối lượng thông tin trong hoạt động Quản lý Nhà nước và mục tiêu hướng tới số hóa toàn bộ thông tin khiến các tổ chức Chính phủ khó khăn trong việc thay đổi giữa phương thức lưu trữ truyền thống và chuyển sang sử dụng công nghệ lưu trữ Big Data. Rất nhiều tổ chức đã xem xét lại vai trò của phương thức lưu trữ truyền thống và triển khai một giải pháp tiết kiệm chi phí hiệu quả với phương thức kết hợp sử dụng băng từ để bảo quản lưu trữ, sử dụng ổ đĩa cho các thông tin thường xuyên được yêu cầu, và sử dụng lưu trữ đám mây cho các dữ liệu Big Data.



2.3.1. Những thay đổi của chính phủ trong thời đại dữ liệu lớn:

Dữ liệu lớn có thể làm thay đổi mô hình cung cấp dịch vụ công trong tương lai. Trong quá trình đó, vai trò của chính phủ trong xã hội cũng sẽ thay đổi cho phù hợp. Các hỗ trợ công nghệ đối với kết nối trực tiếp và việc bắt đầu sử dụng dữ liệu lớn sẽ tạo ra các lợi thế cạnh tranh để thu hút nguồn lực và tài năng để phát triển, duy trì nền kinh tế thông minh, hội nhập toàn cầu.

Việc ứng dụng dữ liệu lớn có thể tác động đến lĩnh vực công thông qua việc cung cấp các dịch vụ công và cơ hội mới cho các tổ chức cung cấp dịch vụ công đồng thời cấu trúc đó có thể chuyển đổi vai trò của chính phủ trong các hoạt động xã hội. Việc sử dụng công nghệ thông tin để cải thiện dịch vụ công cộng bắt đầu bằng chính phủ điện tử. Chuyển đổi phương thức cung cấp dịch vụ công sang công nghệ thông tin mà chính phủ đang sử dụng là một nhiệm vụ phức tạp và tốn kém, thường gắn liền với việc tự động hóa dịch vụ công cộng và tích hợp các hệ thống xử lý. Bên cạnh các dự án chính phủ điện tử nhắm nâng cao hiệu quả hoạt động, các hoạt động khác cũng cần được xem xét để thúc đẩy hoạt động công khai minh bạch, sự tham gia của công dân, và sự phối hợp liên ngành. Điều này có thể đạt được bằng cách chia sẻ cơ sở hạ tầng của khu vực công, chia sẻ thông tin với các cơ quan khác, đánh giá lại năng lực cốt lõi để nâng cao chất lượng dịch vụ cung cấp và để thu hút các tổ chức bên ngoài, chẳng hạn như các trường đại học và doanh nghiệp.

Sử dụng công nghệ dữ liệu lớn làm nền tảng cho sự tiện bộ của chính phủ sẽ giúp những thay đổi này thực sự hiệu quả đồng thời cũng làm thay đổi căn bản mối quan hệ giữa chính phủ và công dân. Từ năm 2012, cả EU và Mỹ đang tìm cách thay đổi luật pháp và chính sách, loại bỏ những trở ngại trong việc sử dụng dữ liệu lớn đồng thời khẳng định tính hiệu quả của dữ liệu lớn trong quản lý và hứa hẹn chi phí sử dụng sẽ thấp hơn đối với khu vực công. Dữ liệu lớn cũng hứa hẹn cung cấp đồng bộ các dịch vụ công cộng. Thiết bị thông minh, chẳng hạn như đèn giao thông thông minh có thể thông báo cho đơn vị quản lý bảo trì trước thời hạn và các vấn đề bất ổn, vì vậy công việc sửa chữa có thể được sắp xếp hợp lý mà không làm gián đoạn dịch vụ.

Cách thức mới trong việc cung cấp dịch vụ công:

Theo cách truyền thống, mối quan hệ giữa người dân và dịch vụ công khá đơn giản. Thường người dân nộp thuế và đổi lại họ được phục vụ trong các lĩnh vực khác nhau, sức khỏe, giáo dục, bảo trì đường bộ và các loại tương tự.

Tuy nhiên gần đây, mối quan hệ này đang dần thay đổi và công dân không chỉ đơn

giản là người hưởng dịch vụ và đã trở thành một đối tác trong việc cung cấp dịch vụ. Ý tưởng chính ở đây là việc cung cấp dịch vụ công cộng là trách nhiệm của tất cả mọi người, từ cá nhân cho đến chính phủ và ngay cả các tổ chức phi chính phủ.

Ở Anh, năm 2010, "hợp tác với công dân và cộng đồng" là một nội dung quan trọng chương trình "Big Society" được nêu trong trong tuyên ngôn của Đảng Bảo Thủ. Năm 2009, sáng kiến US Open goverment của Mỹ cũng nhấn mạnh vai trò của công dân trong việc tham gia vào dịch vụ công. Cả hai đều được thành lập trên ý tưởng tạo điều kiện cho mọi người để tự quan tâm tới bản thân và người khác. Các phương tiện truyền thông và điện thoại thông minh có thể tạo thuận lợi cho sự tương tác giữa người dân và chính phủ. Họ cũng có thể khuếch đại truyền thông và kêu gọi sự tham gia của công chúng. Ví dụ trong lĩnh vực chống tội phạm, các công dân cần liên kết với cảnh sát trong việc theo dõi và báo cáo hoạt động đáng ngờ và việc này đã được nhiều chính phủ thực hiện một cách khá hiệu quả.

Ở Mỹ, các nhà phát triển ứng dụng đã xây dựng trang web Citysourced.com tạo điều kiện cho công dân và cư dân có thể báo cáo và cung cấp thông tin cho chính quyền địa phương về tình hình dân sự, như các ổ gà trên đường, các bức vẽ bậy, vỉa hè hông hay đèn đường hỏng. Mọi người có thể báo cáo một cách công khai hoặc ẩn danh, họ có thể tải lên hình ảnh, và gắn chúng vào một bản đồ đường phố. Báo cáo được gửi đến chính quyền địa phương và được theo dõi quá trình xử lý trực tuyến. Đây là một ví dụ điển hình về các công dân tham gia vào quá trình cải thiện dịch vụ công. Khoa học công nghệ đã tạo điều kiện để nâng cao vai trò của chính quyền địa phương và hoạt động này là một dịch vụ miễn phí cho cộng đồng và cho cả chính phủ bằng cách giảm thiểu chi phí kiểm tra, theo dõi xã hội, cộng đồng.

Sự cải tiến của công nghệ cũng sẽ đem đến một thay đổi cơ bản khác trong xã hội đó là công dân có thể tham gia vào quá trình ra quyết định một cách trực tiếp. Việc này có được là nhớ tận dụng các lợi thế của việc thu thập các thông tin ra quyết định trực tuyến. Chính phủ có thể thu thập ý kiến của người dân về một vấn đề thông qua các mạng xã hội, diễn đàn...sau đó phân tích để biết được xu hướng của người dân, đây chính là một vấn đề của dữ liệu lớn.

Siêu đô thị và vấn đề quản lý:

Từ năm 2011, lần đầu tiên trong lịch sử số người sống ở các thành phố nhiều hơn ở nông thôn. Siêu đô thị - thành phố lớn hơn 10 triệu người đã thành một vấn đề cấp bách. Theo Liên Hiệp Quốc, số lượng các siêu đô thị tăng từ 5 (năm 1975) đến 26 (2006), với 24 trong số đó nằm ở các nước đang phát triển. Các thành phố lớn không còn là vấn

đề của địa phương hay quốc gia. Nó đã trở thành vấn đề toàn cầu, ảnh hưởng đến sự thịnh vượng trong tương lai và sự ổn định của toàn bộ thế giới.

Việc duy trì các siêu đô thị chủ yếu là làm thế nào để nâng cao mức sống khi phải đối mặt với tỷ lệ tăng trưởng dân số cao. Thành phố thông minh (một phần của giải pháp IoT) được coi là một trong cách hiệu quả để quản lý hạ tầng đối với các thành phố này. Lượng dữ liệu sản sinh ra đối với một thành phố thông minh sẽ rất lơn và cần đến công nghệ Big Data để giải quyết các nhu cầu trong quản lý và điều hành.

Nguồn lực từ đám đông (crowdsourcing)

Crowdsourcing đang trở thành một thuật ngữ ngày càng phổ biến và mở ra con đường mới cho việc tạo ra giá trị công cộng miễn phí, cam kết dân sự, và minh bạch. Nó được thể hiện dưới nhiều hình thức. ví dụ, "Báo cáo từ đám đông" (Crowdreporting) là một hình thức phổ biến của crowdsourcing trong lĩnh vực công cộng và phù hợp với các quan niệm mới là công dân là một đối tác. Ví dụ, tại Mỹ "SeeClickFix.com" là một ví dụ điển hình. Nó là một dịch vụ trực tuyến được thiết kế để giúp các công dân báo cáo những vấn đề không khẩn cấp trong họ khu vực thông qua một giao diện web, Facebook hay các ứng dụng điện thoại thông minh. Quá trình xử lý vấn đề được theo dõi trực tuyến. Sau khi vấn đề được báo cáo, nó được theo dõi trực tuyến tương tự như cách các công ty chuyển phát nhanh theo dõi các gói hàng. Các thông tin tương tự như vậy được công bố thông qua Twitter và Facebook để thông báo cho công chúng. Các báo cáo các vấn đề như thế này vẫn được thực hiện trong quá khứ bằng cách sử dụng các phương tiện khác, như gọi điện thoại hoặc viết một lá thư. Tuy nhiên sự khác biệt ở đây là các vấn đề được báo cáo tức thì, trực tiếp và minh bạch, và không thể từ chối lẩn tránh được.

Công chúng không cần phải đi ra khỏi nhà của mình để báo cáo các vấn đề như vậy nữa. Có một ứng dụng dành cho việc đó hoặc chỉ đơn giản là đăng nhập vào Facebook. Không phải chờ đợi trên điện thoại để tiếp cận với một công chức, không phải mất thời gian cho việc ghi chép. Bây giờ tất cả mọi người với một điện thoại thông minh có thể đi xung quanh và báo cáo các vấn đề dân sự. Ngoài ra, thông tin phản hồi trực tiếp và khả năng theo dõi quá trình xử lý có thể sẽ làm mọi người hài lòng và cho họ cảm giác đã đóng góp tới lợi ích chung của xã hội. Trong quá khứ, việc báo cáo thông tin không được ghi nhận xảy ra phổ biến và việc theo dõi quá trình xử lý là không thể thực hiện được. Vì vậy, sự tương tác trực tiếp giữa công dân và cơ quan chính phủ đã giúp chính phủ đạt được 3 mục đích:

(i) công dân được tham gia vào quá trình quản lý nhà nước;

(ii) Giảm chi phí của các dịch vụ công.

(iii) Cải thiện tính minh bạch của các dịch vụ công cộng.

Nguồn thông tin mới: Internet of Things (IoTs)

Số lượng các thiết bị có kết nối M2M (máy tới máy) đã tăng trưởng mạnh mẽ nhờ sự giảm giá thành của các thiết bị và sự tiến bộ của công nghệ và viễn thông. Thành phố thông minh với các cảm biến để đo đạc và thu thập các dữ liệu từ môi trường, giao thông cho đến sức khỏe. IoT sẽ đẩy việc lưu trữ dữ liệu, kết nối, và xử lý chúng lên một giới hạn mới cao hơn và chắc chắn rằng nó cũng sẽ đem lại nhiều lợi ích to lớn cho xã hội.

Đây sẽ là một nguồn lực quý giá để chính phủ có thể thay đổi toàn diện cách thức quản lý xã hội. Một trong các công nghệ sẽ phát huy vai trò trong việc thúc đẩy quá trình đó chính là dữ liệu lớn.

Nhà nước định hướng xu thế phát triển thị trường

Dễ dàng và kịp thời thu thập và phân tích những thông tin có liên quan và không liên quan là rất quan trọng đối với Nhà nước trong việc đáp ứng và cải thiện những yêu cầu nhiệm vụ thay đổi liên tục giữa các cơ quan. Dữ liệu liên tục được tạo ra và được lưu trữ số hóa thông qua những hoạt động Quản lý Nhà nước, từ các cảm biến, tương tác với người dân, và các chương trình trao đổi thông tin khác. Các tổ chức Chính phủ đang tiến hành triển khai công nghệ Big Data để phân tích các tập lớn dữ liệu phục vụ mục đích nghiên cứu khoa học đồng thời cũng từ đó khai thác thông tin để ngăn chặn ảnh hưởng xấu đến từ các hoạt động khủng bố, các hành vi tham nhũng, lãng phí, gian lận.

Tháng 3 năm 2012, Chính phủ Mỹ đã khởi động dự án Nghiên cứu đề xuất và triển khai sáng kiến Big Data với nguồn kinh phí đầu tư lên đến 200 triệu USD nhằm mục đích cải thiện những công cụ/kỹ thuật cần thiết để theo dõi, truy cập, tổ chức, lưu trữ, mô hình hóa, phân tích dữ liệu và thu thập thông tin trích xuất được từ khối lượng lớn dữ liệu số. Ý tưởng này tập trung vào các mục tiêu quan trọng của Chính phủ bao gồm các hoạt động nghiên cứu khoa học, nghiên cứu môi trường và y sinh học, giáo dục, an ninh quốc gia và bao gồm nhiều lĩnh vực.

Bộ Quốc phòng Mỹ đã được đầu tư gần 60 triệu USD mỗi năm cho các dự án mới sẽ khai thác và tận dụng tối đa dữ liệu theo phương thức mới để tăng độ nhạy cảm, tăng khả năng nhận thức và hỗ trợ ra quyết định, tạo nên hệ thống thực sự tự động, có thể học hỏi kinh nghiệm, chủ động ra quyết định, và nhận thức được các giới hạn. Bộ Quốc phòng đồng thời cũng lên kế hoạch để nâng cao nhận thức về tình huống cho các chiến sĩ

và chuyên gia phân tích, gia tăng hỗ trợ cho các hoạt động quân sự.

Trung tâm nghiên cứu dự án công nghệ cao Bộ Quốc phòng Mỹ đang tiến hành chương trình XDATA với nguồn kinh phí 25 triệu USD mỗi năm, kéo dài trong vòng 4 năm để phát triển kỹ thuật điện toán và các công cụ phần mềm cho phân tích khối lượng lớn dữ liệu, bao gồm cả dữ liệu bán cấu trúc và dữ liệu phi cấu trúc (dạng văn bản và tin nhắn trao đổi).

Bộ Y tế và Bộ Khoa học công nghệ đang đầu tư vào các công trình nghiên cứu khoa học kỹ thuật Big Data. Những nghiên cứu này tập trung vào việc quản lý, phân tích, ảo hóa và trích xuất thông tin hữu ích từ các tập lớn dữ liệu; Bộ Y tế đặc biệt quan tâm đến những vấn đề liên quan đến sức khỏe và các dịch bệnh như phân tử, nguyên tử, điện sinh, hóa chất, hành vi, dịch tễ học và lâm sàng.

Bộ Năng lượng sẽ dành 25 triệu USD trong ngân sách để thành lập Ban Quản lý quy mô, phân tích và mô hình hóa dữ liệu. Với đầu tàu là phòng thí nghiệm quốc gia Lawrence Berkely kết hợp cùng với các chuyên gia đến từ 6 phòng thí nghiệm quốc gia khác và 7 trường đại học, tất cả sẽ cùng nhau nghiên cứu và phát triển các công cụ mới để giúp các nhà khoa học quản lý và mô hình hóa dữ liệu trên các siêu máy tính.

Cơ quan Khảo sát địa chất Mỹ (USCS) áp dụng công nghệ Big Data trong các nghiên cứu khoa học về Trái đất. Sáng kiến này gia tăng sự hiểu biết, nhằm ứng phó với sự biến đổi khí hậu, các trận động đất và sự thay đổi của các chỉ số sinh thái.

CHƯƠNG 3: NỀN TẢNG CÔNG NGHỆ PHÂN TÍCH DỮ LIỆU LỚN

3.1. Bộ công cụ phân tích dữ liệu lớn

Hiện nay có rất nhiều công cụ dùng để xử lý dữ liệu lớn đã và đang được nghiên cứu và phát triển bởi các viện nghiên cứu lớn trên Thế Giới. Các công cụ này giúp cho việc xử lý một lượng dữ liệu khổng lồ một cách nhanh chóng, giúp người dùng có thể dễ dàng tìm được thông tin cần thiết trong thời gian thực, nhất là khi kỷ nguyên của exabytes đang đến gần.

3.1.1. Apache Hadoop

Hadoop là một dự án phần mềm mã nguồn mở được phát triển bởi Apache, nhằm

thu các giá trị có ích từ khối lượng, tốc độ và tính đa dạng của dữ liệu (cấu trúc/phi cấu trúc). Apache Hadoop là một khuôn khổ cho phép để xử lý và phân phối các bộ dữ liệu lớn trên các cụm máy tính sử dụng mô hình lập trình đơn giản. Nó được thiết kế để mở rộng từ một máy chủ duy nhất đến hàng ngàn máy, mỗi máy cung cấp tính toán và lưu trữ địa phương.

3.1.2. Apache Spark

Apache Spark, như đúng cái tên mà Apache Software Foundation đặt tên cho nó. Spark là một engine xử lý rất nhanh một lượng lớn dữ liệu. Nó cung cấp nhiều API hỗ trợ cho các lập trình Java, Scala và Python. Ngoài ra, còn một tập hợp các thư viện hỗ trợ xử lý stream, machine và phân tích hình ảnh.

Apache Spark là một mã nguồn mở cụm hệ thống máy tính nhằm mục đích để phân tích dữ liệu nhanh – để chạy và viết nhanh. Để chạy các chương trình nhanh hơn, Spark cung cấp nguyên thủy cho tính toán cluster trong bộ nhớ: có thể tải dữ liệu vào bộ nhớ và truy vấn nó lặp đi lặp lại nhanh hơn nhiều so với các hệ thống dựa trên đĩa như Hadoop MapReduce.

Apache Spark bước đầu đã cho các kết quả khả quan khi có những đo đạc về hiệu suất, nó có thể chạy nhanh hơn so với Hadoop MapReduce lên đến 100 lần. Kết quả đo đạc gần đây của DataBricks chỉ ra rằng Apache Spark giúp tiết kiệm tài nguyên hơn mà lại hiệu quả hơn.

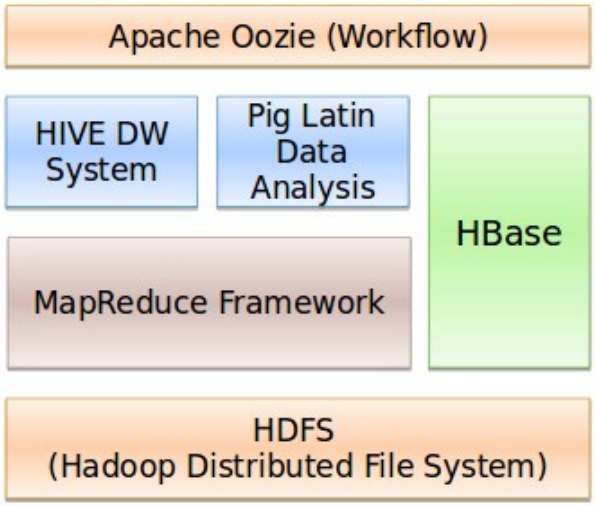
Tháng 12 năm 2014, DataBricks thực hiện các phép đo đạc ánh xạ dữ liệu 100TB dữ liệu (1000 tỷ bản ghi), họ sử dụng nền tảng Apache Spark trên 206 máy chủ EC2 thuê của Amazon đã cho những kết quả rất khả quan (chi tiết tham khảo bài viết “Spark the fastest open source engine for sorting a petabyte” trên https://databricks.com). Kỷ lục thế giới lúc đó là 72 phút được lập bởi Yahoo khi sử dụng nền tảng Hadoop MapRecude trên 2100 nodes (xem thêm tại “GraySort and MinuteSort at Yahoo on Hadoop 0.23” tại http://sortbenchmark.org/Yahoo2013Sort.pdf). Databricks đã lập được kỷ lục mới là 23 phút. Điều này có nghĩa là Spark chỉ sử dụng số lượng máy ít hơn gấp 10 lần nhưng lại cho kết quả nhanh hơn gấp 3 lần.

3.2. Kiến trúc Apache Hadoop

Apache Hadoop là một framework dùng để chạy những ứng dụng trên 1 cluster lớn được xây dựng trên những phần cứng thông thường. Hadoop hiện thực mô hình Map/Reduce,

đây là mô hình mà ứng dụng sẽ được chia nhỏ ra thành nhiều phân đoạn khác nhau, và các phần này sẽ được chạy song song trên nhiều node khác nhau.

3.2.1. Thành phần của Apache Hadoop



Hình 3.1: Hệ sinh thái của Apache Hadoop v1.x (nguồn skillspeed.com)

- HDFS: hệ thống tập tin phân tán HDFS (viết tắt từ Hadoop Distributed File System) giúp cho việc lưu trữ dữ liệu lớn được thuận lợi hơn

- MapReduce: Đây là mô hình lập trình cho Hadoop. Được chia làm giai đoạn là Map và Reduce. Thực tế là một quá trình shuffle-sort (một quá trình mà hệ thống thực hiện sắp xếp và chuyển các kết quả đầu ra của bộ ánh xạ tới các đầu vào của các bộ rút gọn) giữa hai giai đoạn Map và Reduce.

- Hadoop Streaming: Một tiện ích để tạo nên mã MapReduce bằng bất kỳ ngôn ngữ nào: C, Perl, Python, C++, Bash,...

- Hive và Hue: Hive chuyển đổi lệnh SQL thành một tác vụ MapReduce. Hue cung cấp một giao diện đồ họa dựa trên trình duyệt để làm công việc Hive yêu cầu.

- Pig Latin: Một môi trường lập trình mức cao hơn để viết mã MapReduce.

- Sqoop: Cung cấp việc truyền dữ liệu hai chiều giữa Hadoop và cơ sở dữ liệu quan hệ (nếu có sử dụng).

- Oozie: Quản lý luồng công việc Hadoop. Oozie không thay thế trình lập lịch biểu hay công cụ BPM của bạn, nhưng nó cung cấp cấu trúc phân nhánh if-then-else và điều khiển trong phạm vi tác vụ Hadoop của bạn.

- Hbase: Một kho lưu trữ key-value có thể mở rộng quy mô rất lớn. Nó hoạt động rất giống như một hash-map để lưu trữ lâu bền. Nó không phải là một cơ sở dữ liệu quan hệ.

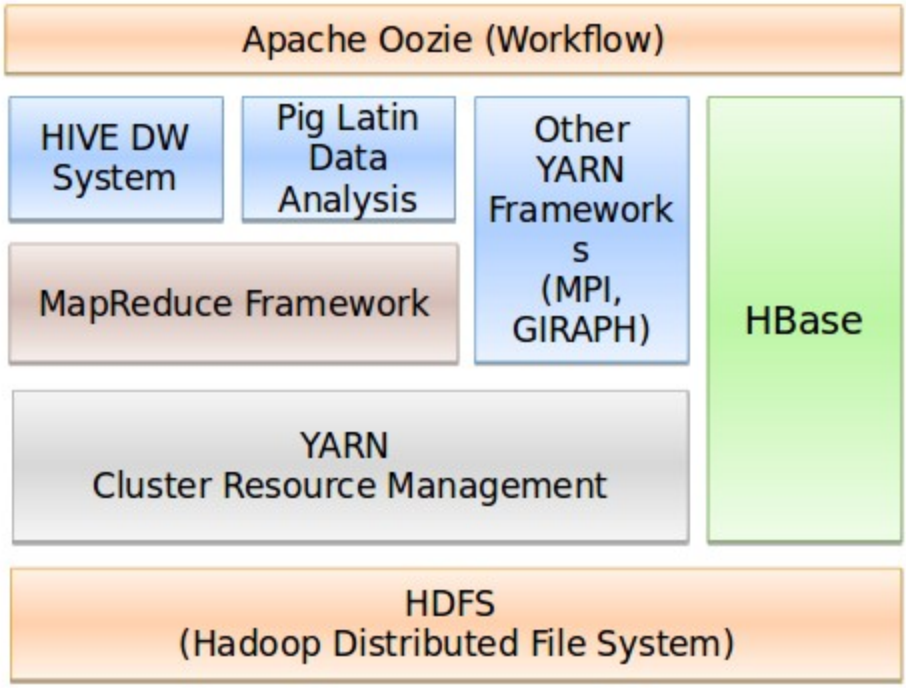
- FlumeNG: Trình nạp thời gian thực để tạo luồng dữ liệu vào Hadoop. Nó lưu trữ dữ liệu trong HDFS và HBase.

- Whirr: Cung cấp Đám mây cho Hadoop. Giúp cho việc khởi động một hệ thống chỉ trong vài phút với một tệp cấu hình rất ngắn.

- Mahout: Máy học dành cho Hadoop. Được sử dụng cho các phân tích dự báo và phân tích nâng cao khác.

- Fuse: Làm cho hệ thống HDFS trông như một hệ thống tệp thông thường, do đó có thể sử dụng lệnh ls, cd, rm và những lệnh khác với dữ liệu HDFS.

- Zookeeper: Được sử dụng để quản lý đồng bộ cho hệ thống.



Hình 3.2: Hệ sinh thái của Apache Hadoop v2.x (nguồn skillspeed.com)

3.2.2. Hệ thống tập tin phân tán Hadoop

Khi kích thước của tập dữ liệu vượt quá khả năng lưu trữ của một máy tính, tất yếu sẽ dẫn đến nhu cầu phân chia dữ liệu lên trên nhiều máy tính. Các hệ thống tập tin quản lý việc lưu trữ dữ liệu trên một mạng nhiều máy tính gọi là hệ thống tập tin phân tán. Do hoạt động trên môi trường liên mạng, nên các hệ hống tập tin phân tán phức tạp hơn rất

nhiều so với một hệ thống file cục bộ.

Hadoop mang đến cho chúng ta hệ thống tập tin phân tán HDFS (viết tắt từ Hadoop Distributed File System) với nỗ lực tạo ra một nền tảng lưu trữ dữ liệu đáp ứng cho một khối lượng dữ liệu lớn và chi phí rẻ.

HDFS ra đời trên nhu cầu lưu trữ dữ liệu của Nutch, một dự án Search Engine nguồn mở. HDFS kế thừa các mục tiêu chung của các hệ thống file phân tán trước đó như độ tin cậy, khả năng mở rộng và hiệu suất hoạt động. Tuy nhiên, HDFS ra đời trên nhu cầu lưu trữ dữ liệu của Nutch, một dự án Search Engine nguồn mở, và phát triển để đáp ứng các đòi hỏi về lưu trữ và xử lý của các hệ thống xử lý dữ liệu lớn với các đặc thù riêng. Do đó, các nhà phát triển HDFS đã xem xét lại các kiến trúc phân tán trước đây và nhận ra các sự khác biệt trong mục tiêu của HDFS so với các hệ thống file phân tán truyền thống.

- Thứ nhất: các lỗi về phần cứng sẽ thường xuyên xảy ra. Hệ thống HDFS sẽ chạy trên các cluster với hàng trăm hoặc thậm chí hàng nghìn node. Các node này được xây 18 dựng nên từ các phần cứng thông thường, giá rẻ, tỷ lệ lỗi cao. Chất lượng và số lượng của các thành phần phần cứng như vậy sẽ tất yếu dẫn đến tỷ lệ xảy ra lỗi trên cluster sẽ cao. Các vấn đề có thể điểm qua như lỗi của ứng dụng, lỗi của hệ điều hành, lỗi đĩa cứng, bộ nhớ, lỗi của các thiết bị kết nối, lỗi mạng, và lỗi về nguồn điện… Vì thế, khả năng phát hiện lỗi, chống chịu lỗi và tự động phục hồi phải được tích hợp vào trong hệ thống HDFS.

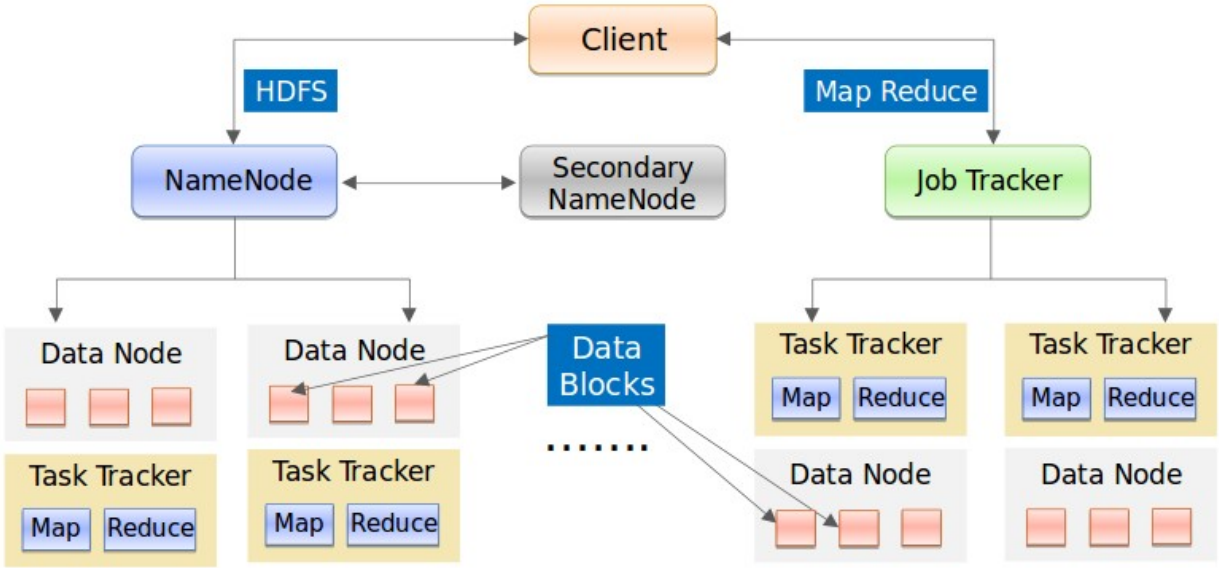
- Thứ hai: kích thước file sẽ lớn hơn so với các chuẩn truyền thống, các file có kích thước hàng GB sẽ trở nên phổ biến. Khi làm việc trên các tập dữ liệu với kích thước nhiều TB, ít khi nào người ta lại chọn việc quản lý hàng tỷ file có kích thước hàng KB, thậm chí nếu hệ thống có thể hỗ trợ. Điều chúng muốn nói ở đây là việc phân chia tập dữ liệu thành một số lượng ít file có kích thước lớn sẽ là tối ưu hơn. Hai tác dụng to lớn của điều này có thể thấy là giảm thời gian truy xuất dữ liệu và đơn giản hoá việc quản lý các tập tin.

- Thứ ba: hầu hết các file đều được thay đổi bằng cách thêm dữ liệu vào cuối file hơn là ghi đè lên dữ liệu hiện có. Việc ghi dữ liệu lên một vị trí ngẫu nhiên trong file không hề tồn tại. Một khi đã được tạo ra, các file sẽ trở thành file chỉ đọc (read-only), và thường được đọc một cách tuần tự. Có rất nhiều loại dữ liệu phù hợp với các đặc điểm trên. Đó có thể là các kho dữ liệu lớn để các chương trình xử lý quét qua và phân tích dữ liệu. Đó có thể là các dòng dữ liệu được tạo ra một

cách liên tục qua quá trình chạy các ứng dụng (ví dụ như các file log). Đó có thể là kết quả trung gian của một máy này và lại được dùng làm đầu vào xử lý trên một máy khác. Và do vậy, việc thêm dữ liệu vào file sẽ trở thành điểm chính để tối ưu hoá hiệu suất.

Các thành phần hệ thống Hadoop

Hệ thống Hadoop có một kiến trúc master/slave, trên một cluster chạy HDFS. Mỗi khi chạy HDFS có nghĩa là chạy một tập các trình nền - daemon, hoặc các chương trình thường trú, trên các máy chủ khác nhau trên hạ tầng. Những trình nền có vai trò cụ thể, một số chỉ tồn tại trên một máy chủ, một số có thể tồn tại trên nhiều máy chủ.



Hình 3.3: Các dịch vụ bên trong một hệ thống HDFS phiên bản 1.x

(nguồn: skillspeed.com)

Các trình nền bao gồm:

- **Namenode** đóng vai trò là master, chịu trách nhiệm duy trì thông tin về cấu trúc cây phân cấp các file, thư mục của hệ thống file và các metadata khác của hệ thống file. Cụ thể, các Metadata mà Namenode lưu trữ gồm có:
    - **File System Namespace:** là hình ảnh cây thư mục của hệ thống file tại một thời điểm nào đó. File System namespace thể hiện tất các các file, thư mục có trên hệ thống file và quan hệ giữa chúng.
    - **Thông tin để ánh xạ từ tên file ra thành danh sách các block:** với mỗi file, ta có một danh sách có thứ tự các block của file đó, mỗi Block đại diện bởi

Block ID.

- **Nơi lưu trữ các block:** các block được đại diện một Block ID. Với mỗi block ta có một danh sách các DataNode lưu trữ các bản sao của block đó.

Chức năng của NameNode là nhớ (memory) và I/O chuyên sâu. Như vậy, máy chủ lư trữ NameNode thường không lưu trữ bất cứ dữ liệu người dùng hoặc thực hiện bất cứ một tính toán nào cho một ứng dụng MapReduce để giảm khối lượng công việc trên máy.

NameNode theo dõi cách các tập tin của được phân chia thành các khối (block), những node nào lưu các khối đó, và "kiểm tra sức khỏe" tổng thể của hệ thống tệp phân tán.

- **DataNode**: Mỗi máy slave trong cluster của bạn sẽ lưu trữ (host) một trình nền DataNode để thực hiện các công việc nào đó của hệ thống file phân tán - đọc và ghi các khối HDFS tới các file thực tế trên hệ thống file cục bộ (local filesytem). Khi đọc hay ghi một file HDFS, file đó được chia nhỏ thành các khối và NameNode sẽ nói cho các client nơi các mỗi khối trình nền DataNode sẽ nằm trong đó. Client của bạn liên lạc trực tiếp với các trình nền DataNode để xử lý các file cục bộ tương ứng với các block.

  Hơn nữa, một DataNode có thể giao tiếp với các DataNode khác để nhân bản các khối dữ liệu của nó để dự phòng.

  Các DataNode thường xuyên báo cáo với các NameNode. Sau khi khởi tạo, mỗi DataNode thông báo với NameNode của các khối mà nó hiện đang lưu trữ. Sau khi Mapping hoàn thành, các DataNode tiếp tục thăm dò ý kiến NameNode để cung cấp thông tin về thay đổi cục bộ cũng như nhận được hướng dẫn để tạo, di chuyển hoặc xóa các blocks từ đĩa địa phương (local).

- **Secondary NameNode:** là một trình nền hỗ trợ giám sát trạng thái của các cụm HDFS. Giống như NameNode, mỗi cụm có một Secondary NameNode, và nó thường trú trên một máy của mình.

  Namenode có nhược điểm là khi có sự cố với nó, hệ thống sẽ mất kiểm soát, không thể truy cập được dữ liệu nên cần phải có Secondary NameNode.

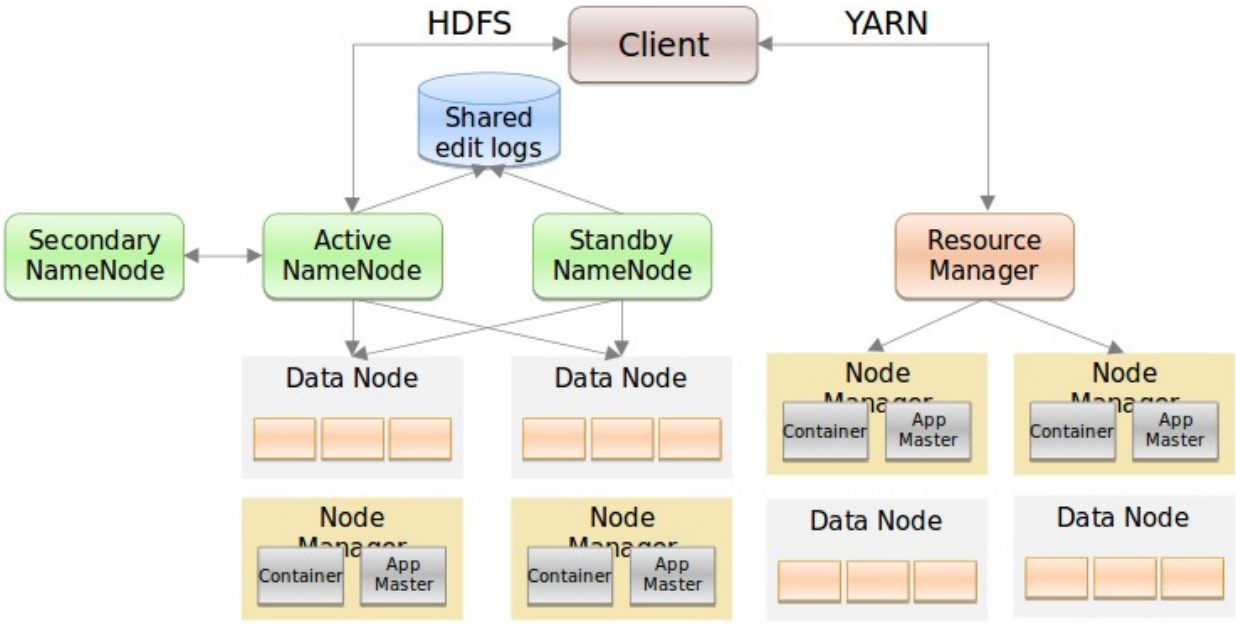
- **JobTracker:** làm nhiệm vụ kiểm soát tài nguyên tính toán của toàn cluster và chia việc cho các node tính toán. Trình nền JobTracker là một liên lạc giữa ứng dụng

với Hadoop. Một khi mã nguồn được gửi tới các cụm (cluster), JobTracker sẽ quyết định kế hoạch thực hiện bằng cách xác định những tập tin nào sẽ xử lý, các nút được giao các nhiệm vụ khác nhau, và theo dõi tất cả các nhiệm vụ khi dúng đang chạy. Nếu một nhiệm vụ (task) thất bại (fail), JobTracker sẽ tự động chạy lại nhiệm vụ đó, có thể trên một node khác, cho đến một giới hạn nào đó được định sẵn của việc thử lại này.

- **TaskTracker:** làm nhiệm vụ thông báo tình trạng tính toán của node trong cluster đồng thời chạy các Task được giao bởi JobTracker. Mỗi Task này có thể là một Mapper hay 1 Reducer.

  Như với các trình nền lưu trữ, các trình nền tính toán cũng phải tuân theo kiến trúc master/slave: JobTracker là giám sát tổng việc thực hiện chung của một công việc MapRecude và các task Tracker quản lý việc thực hiện các nhiệm vụ riêng trên mỗi node slave.

  Một trong những trách nhiệm của các TaskTracker là liên tục liên lạc với JobTracker. Nếu JobTracker không nhận được nhịp đập từ một TaskTracker trong vòng một lượng thời gian đã quy định, nó sẽ cho rằng TaskTracker đã bị treo (cashed) và sẽ gửi lại nhiệm vụ tương ứng cho các nút khác trong cluster.



Hình 3.4: Các dịch vụ bên trong một hệ thống HDFS phiên bản 2.x

(nguồn: skillspeed.com)

Mô hình này có một số nhược điểm:

- Vấn đề mở rộng (scalability)

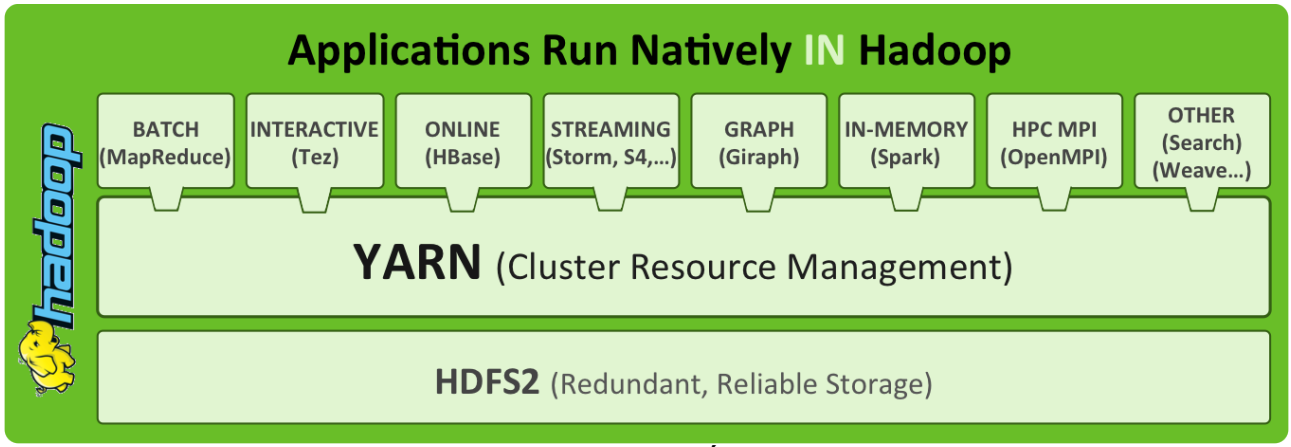
- Vấn đề tính sẵn sàng (availibility)
- Vấn đề về sử dụng tài nguyên

Do chỉ có một máy làm nhiệm vụ quản lý tài nguyên, đồng thời quản lý tình trạng thực hiện Job nên năng lực tính toán của cluster bị phụ thuộc vào năng lực của máy tính chạy JobTracker. (Kiến trúc này có cluster tối đa là khoảng 5000 node)

Tính sẵn sàng của hệ thống cũng không cao khi chỉ có một máy tính vừa lo chạy job vừa lo lập lịch. Khi JobTracker trở nên bận rộn, Job queue có thể bị hủy khiến cho các job đang thực thi lỗi và không thành công.

Tài nguyên tính toán trong mô hình này là slot. Số lượng Task chạy xong xong trên 1 máy là số lượng cores trên máy đó vì vậy lúc khởi động cluster, người quản trị phải tự tay cài đặt số slot dành cho mapper và reducer. Tùy vào cách sử dụng cluster mà cách cấu hình này nhiều khi không đạt hiệu quả cao. Việc điều chỉnh các slot này bằng tay đòi hỏi kinh nghiệm và nhiều khi không đáp ứng được nhu cầu ngay trước mắt của job. Ngoài ra các tài nguyên như DiskIO hay GPUs v.v không được sử dụng.

Vì vậy kiến trúc Apache Hadoop thế hệ thứ hai được ra đời với nhiều cải tiến để dễ dàng phục vụ các ứng dụng dữ liệu lớn hơn.



Hình 3.5: Các dịch vụ bên trong một hệ thống Apache Hadoop phiên bản 2.x

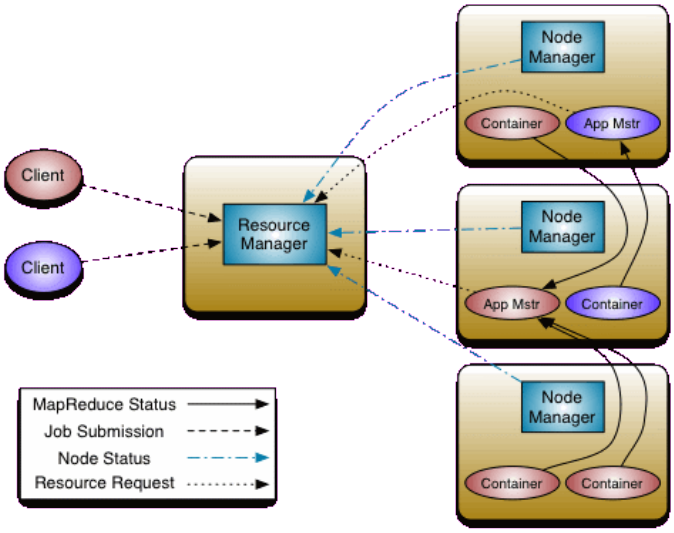
(nguồn: Hortonwork)

Mô hình hạ tầng dữ liệu lớn thế hệ thứ 2

YARN là viết tắt của cụm từ "Yet-Another-Resource-Negotiator" là một framework hỗ trợ phát triển ứng dụng phân tán. YARN cung cấp daemons và APIs cần thiết cho việc phát triển ứng dụng phân tán, đồng thời xử lý và lập lịch sử dụng tài nguyên tính toán

(CPU hay memory) cũng như giám sát quá trình thực thi các ứng dụng đó. YARN tổng quát hơn MapReduce thế hệ đầu tiên (gồm mô hình JobTracker/TaskTracker).

MapReduce2 (MR2) là mô hình MapReduce được viết lại để chạy như là một ứng dụng trên YARN. Mô hình chạy job vẫn tương tự như MapReduce 1, ngoại trừ rằng các job bây giờ chạy ApplicationMaster của riêng chúng. Bản thân ApplicationMaster không được chạy trên NameNode (như mô hình JobTracker cũ nữa) mà sẽ được gửi và chạy trên máy tính toán (DataNode).



Hình 3.6: Mô hình MapRecude thế hệ thứ 2

Kiến trúc mới chia 2 chức năng chính của JobTracker - quản lý tài nguyên và quản lý job thành 2 components riêng biệt:

- Resource Manager (RM): quản lý toàn bộ tài nguyên tính toán của cluster.
- Application Master (AM): đơn vị là trên 1 ứng dụng và quản lý vòng đời của Job.

Do vậy đối với YARN, MapReduce sẽ là 1 ứng dụng chạy trên YARN, sử dụng tài nguyên do RM cấp phát. Các node tính toán trong cluster bây giờ sẽ chạy NodeManager quản lý các tiến trình chạy trên máy đó. Resource Manager và Node Manager trở thành xương sống của tính toán phân tán trong YARN. Việc mỗi ứng dụng được tách ra riêng cho phép các process chạy lâu (long running process) cũng có thể được khởi động trên YARN.

ApplicationMaster trên 1 ứng dụng là một thư viện cho phép yêu càu tài nguyên từ Resource Manager và giao tiếp với Node Manager để chạy và thực thi các tasks. Trong YARN, MapReduce2 chỉ là một ứng dụng thay vì là thành phần không thể thiếu của hadoop như ở hadoop phiên bản 1. Application Master cho phép xây dựng các ứng dụng

khác MR chạy trên YARN.

Hiện tại có rất nhiều ứng dụng BigData được port chạy trên YARN, trong đó có một số ứng dụng nổi tiếng như Spark hay H2O.

Resource Manager dùng một module lập lịch ở đó các Job sẽ được cho vào hàng đợi. Module này có thể tháo lắp tự do. Hiện tại YARN có 2 schedulers là:

- CapacityScheduler
- FairScheduler

3.3. Kiến trúc Apache Spark

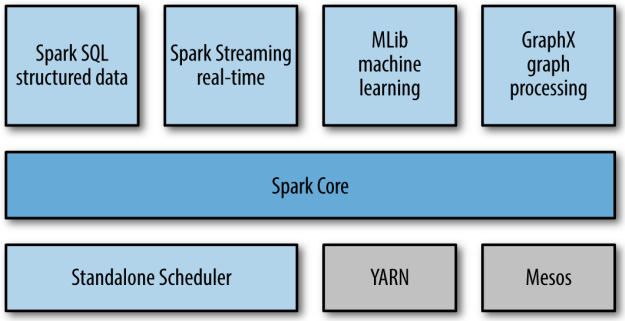
3.3.1. Các thành phần Apache Spark

Apache Spark là hệ thống nền tảng gồm nhiều thành phần được tích hợp chặt chẽ với nhau. Ở thành phần cốt lõi, Spark được xem như là một công cụ tính toán có trách nhiệm lập lịch, phân phối và giám sát các ứng dụng bao gồm nhiều nhiệm vụ tính toán trên nhiều máy thực thi hoặc một cụm các máy thực thi. Bởi vì thành phần cốt lõi của Spark được thiết kế với mục đính tính toán nhanh và sử dụng chung cho nhiều mục đích khác nhau (chẳng hạn phục vụ SQL, máy học - machine learning) nên các thành phần được thiết kế để có thể tương thích chặt chẽ, giúp cho việc sử dụng như là một thành phần thư viện phần mềm, giúp cho việc phát triển dự án được thuận lợi và nhanh chóng hơn.

Apache Spark tích hợp chặt chẽ nhiều thành phần mang tới nhiều lợi ích hơn. Các thư viện và các thành phần ở lớp cao hơn trong mô hình kiến trúc phát triển sản phẩm phân tích dữ liệu lớn được hưởng lợi từ những cải tiến ở lớp thấp hơn, sự thay đổi mang tính nâng cao của tầng thấp sẽ độc lập được với tầng cao. Ví dụ như nếu thành phần lõi của Spark được tối ưu hóa thì các thành phần ở tầng trên (như thư viện SQL, máy học ML) sẽ tự động được cải thiện tốc độ.

Kiến trúc tích hợp của Spark giúp khả năng xây dựng ứng dụng được đa dạng hơn, giúp kết hợp từ nhiều mô hình xử lý khác nhau. Ví dụ, một ứng dụng máy học để phân loại dữ liệu trong thời gian thực khi nhận dữ liệu trực tuyến. Đồng thời, các nhà phân tích có thể truy vấn dữ liệu kết quả cũng bằng thời gian thực thông qua SQL. Ngoài ra các kỹ sư dữ liệu, các nhà khoa học dữ liệu có thể truy cập vào cùng một dữ liệu bằng một ngôn ngữ lập trình nào đó (ví dụ Python) để phân tích

dữ liệu kiểu ad-hoc.



Hình 3.7: Kiến trúc thành phần lõi Apache Spark

Lõi Spark

Lõi Spark (spark core) chứa các chức năng cơ bản của Spark: bộ lập lịch, quản lý bộ nhớ, phục hồi lỗi, tương tác với hệ thống lưu trữ,... Spark Core cũng cung cấp các API tương tác với các tập dữ liệu phân tán có khả năng phục hồi (resilient distributed datasets – RDDs), giúp trừu tượng hóa các chương trình của Spark.

Spark SQL

Spark SQL là gói thư viện để làm việc với dữ liệu có cấu trúc, giúp việc truy vấn dữ liệu thông qua SQL cũng như các bộ công cụ khác tương đương như Apache Hive, hỗ trợ được nhiều nguồn dữ liệu khác nhau như bảng Hive, Parquet, JSON.

Spark SQL cung cấp cách thức kết hợp giữa các truy vấn SQL với các chương trình được hỗ trợ bởi RDD. Điều này giúp cho Spark SQL giống như các công cụ nguồn mở để khai thác kho dữ liệu.

Luồng Spark

Luồng Spark (Spark streaming) là một thành phần của Spark cho phép xử lý các dòng dữ liệu trực tuyến. Ví dụ: các dòng dữ liệu trực tuyến này là các bản ghi lịch sử (log) hoạt động được tạo ra bởi các máy chủ dịch vụ, hay trạng thái sử dụng dịch vụ Web của người dùng.

Luồng Spark cung cấp các API để thao tác với luồng dữ liệu tương ứng chặt chẽ với các API của thành phần RDD trong Lõi Spark, làm cho việc dễ dàng lập trình tạo các ứng dụng phân tích dữ liệu mới.

Mlib

Spark có chứa các thư viện máy học (ML) gồm nhiều thuật toán máy học để phân

loại, hồi quy, lọc cộng tác, hỗ trợ các chức năng như đánh giá mô hình, nhập dữ liệu

GraphX

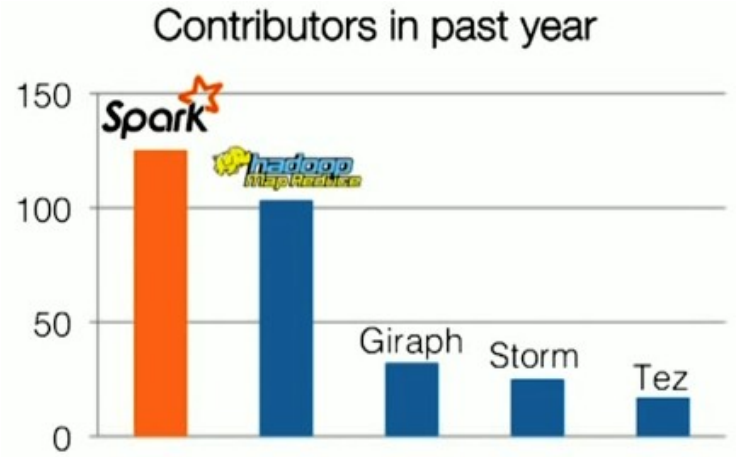
GraphX là một thư viện cho các thao tác liên quan đến đồ thị (ví dụ: đồ thị hoạt động của người dùng trên mạng xã hội), và thực hiện các tính toán đồ thị song song song. Giống như Spark Streaming và Spark SQL, GraphX mở rộng các API của RDD, cho phép các nhà phát triển tạo ra một biểu đồ có sự linh hoạt giữa các đối tượng giá trị và trục tọa độ.

3.3.2. Một số ứng dụng của Apache Spark

Apache Spark sau khi được công bố dã nhanh chóng trở thành một dự án nguồn mở được quan tâm nhất, công nghệ này đã sớm được các hãng công nghệ sử dụng chẳng hạn như Conviva, ClearStory, và Yahoo.

Mục tiêu của Spark là thay thế MapReduce: dễ dàng sử dụng hơn, chạy nhanh hơn. Spark cung cấp các thư viên machine learning (ML) và các thuật toán đồ thị, đồng thời cũng h ỗ trợ các ứng dụng streaming theo thời gian thực và SQL thông qua Spark Streaming và Shark. Các ứng dụng Spark có thể được viết bằng Java, Scala, hoặc Python, và có tốc độ xử lý nhanh hơn từ 10 tới 100 lần so với các ứng dụng MapReduce.

Matei Zaharia, tác giả của Spark và CTO của Databricks chia sẻ: “Trong những năm qua, Spark đã thực sự vượt qua về mức độ quan tâm đối với Hadoop MapReduce và các engine khác, điều đó được thể hiện qua số lượng người đóng góp cho dự án này”



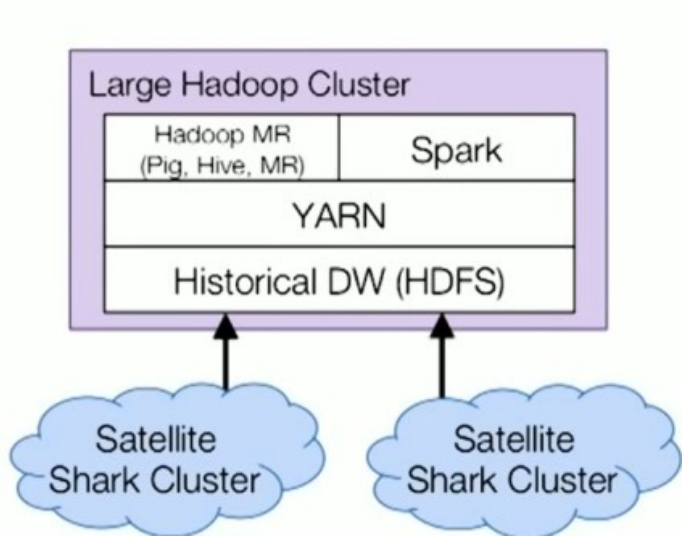
Zaharia cũng chia sẻ rằng: “Có rất nhiều nơi để thu thập dữ liệu nhưng không phải các nơi đó có khả năng khai thác được thông tin, phải là những công ty chuyên biệt và

lớn như Google mới có được những thuật toán để đạt được khả năng khai thác thông tin tốt nhất" và ông nói: "Spark được thiết kế để giải quyết vấn đề này, Spark mang đến các phương thức phân tích dữ liệu đầu cuối, mang tới một giá trị tương đương về hiệu suất và sự tinh tế (sophistication) so với các hệ thống đắt tiền khác, giúp bạn có thể làm được nhiều việc hơn với khối dữ liệu của bạn".

Triển khai Apache Spark tại Yahoo:

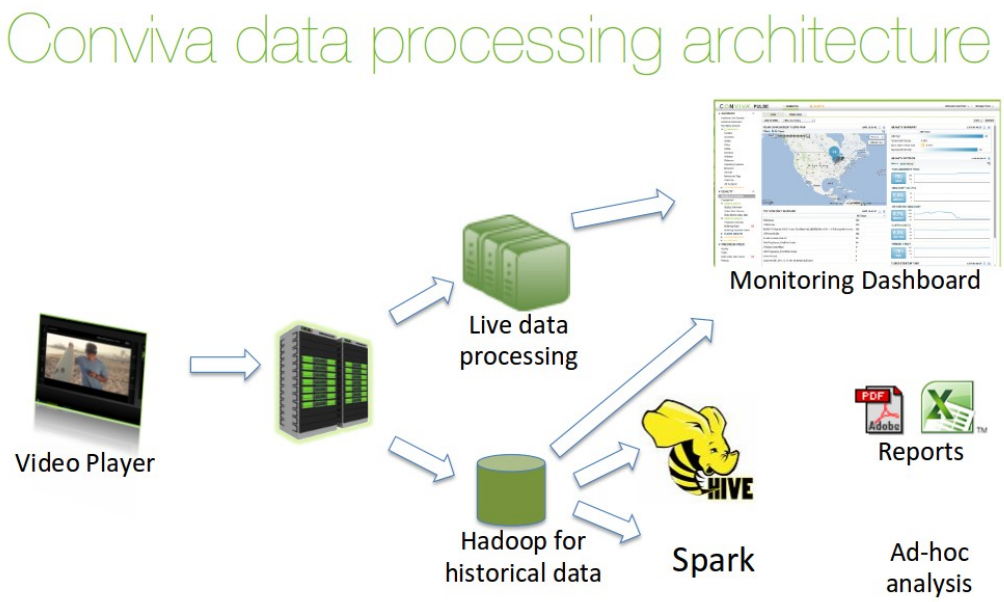
Yahoo hiện có 2 dự án sử dụng Spark trong quá trình hoạt động, một cho cá nhân hóa các trang tin tức dành cho khách hàng duyệt Web và một cho phân tích các quảng cáo. Đối với dự án cá nhân hóa các thông tin thời sự, Yahoo sử dụng giải thuật ML chạy trên Spark để tìm ra được những chủng loại tin tức mà người dùng cá nhân hay quan tâm tới, và cũng để phân loại các câu chuyện tin tức đang diễn ra được quan tâm nhiều nhất. Zaharia chia sẻ về mục tiêu dự án này: "Khi bạn làm cá nhân tin tức, bạn cần phải phản ứng nhanh chóng với những gì người dùng đang làm và những sự kiện xảy ra ở thế giới bên ngoài. Nếu bạn nhìn vào trang chủ của Yahoo, các mục tin tức nào sẽ hiển thị cho bạn? Bạn cần phải tìm hiểu những mục tin tức sẽ mang đến cho mỗi người để họ thích đọc những tin tức đó. Và bạn cần phải tìm hiểu về người dùng, khi họ nhấp chuột vào thì cũng phải tìm ra được họ đang quan tâm đến chủ đề nào".

Đối với dự án phân tích quảng cáo, Yahoo sử dụng Hive trên Spark (gọi là Shark). Hãng muốn sử dụng các công cụ BI hiện có để xem và truy vấn dữ liệu phân tích quảng cáo của họ thu thập được trong Hadoop. Yahoo chọn Hive là vì các API của Hive giúp các công cụ khác có thể tương tác trực tiếp được với máy chủ Shark.



Triển khai Apache Spark tại Conviva

Conviva là một trong những công ty streaming video lớn nhất trên Internet, cung cấp nguồn dữ liệu 4 tỷ video mỗi tháng (chỉ đứng thứ hai trên YouTube). Để có thể cung cấp được dịch vụ như thế, đòi hỏi phải có một hạ tầng công nghệ phía sau để đảm bảo được chất lượng dịch vụ. Conviva đã sử dụng Spark để cung cấp dịch vụ QoS nhằm đảm bảo chất lượng bộ đệm video, giúp cho video chạy mượt hơn. Zaharia nói rằng: "Conviva sử dụng Spark Streaming để tìm hiểu điều kiện mạng trong thời gian thực bằng cách tương tác trực tiếp với các trình xem video, để có thể tối ưu hóa tốc độ cho từng trường hợp."

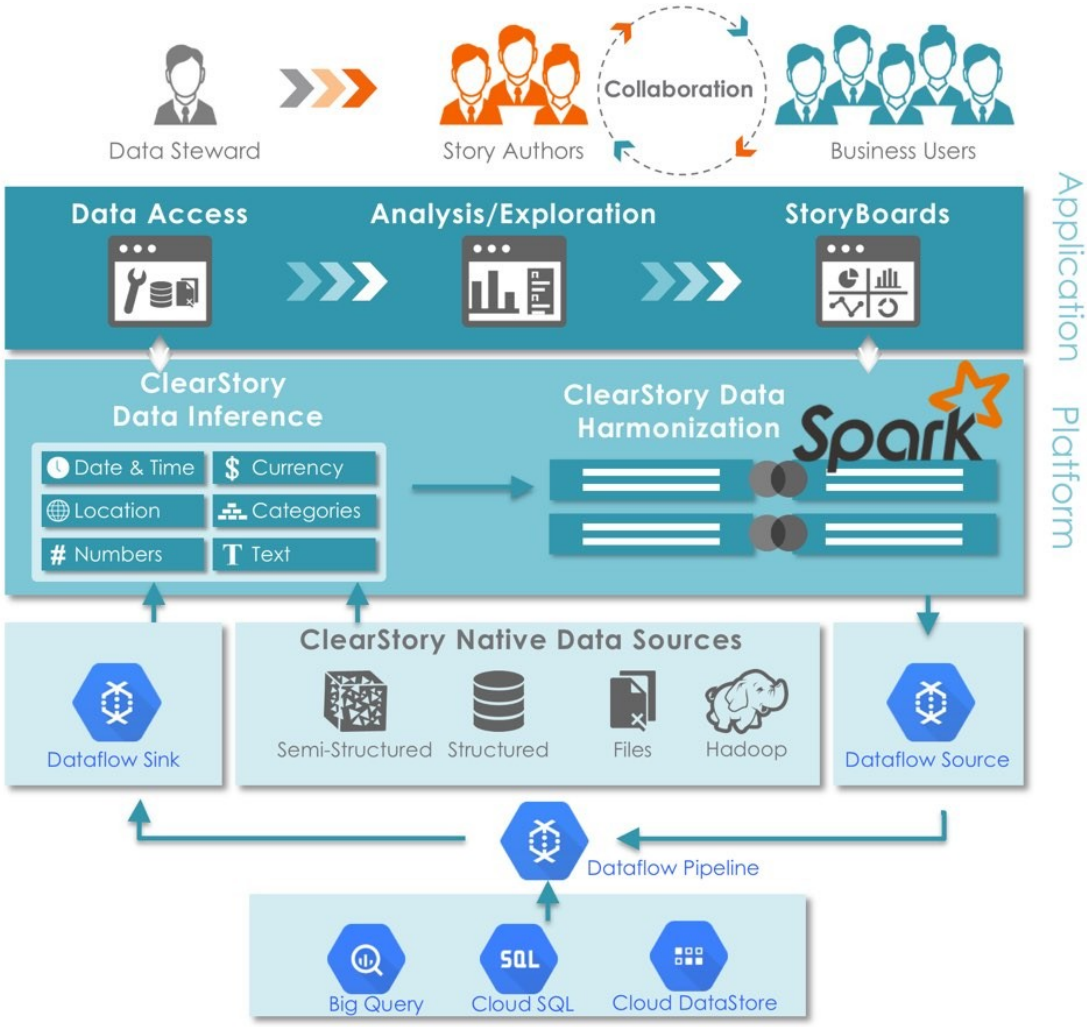


Mô hình kiến trúc xử lý dữ liệu của Conviva (Nguồn: Conviva - the pulse of online video)

Triển khai Apache Spark tại ClearStory

ClearStory cung cấp giải pháp giúp người dùng doanh nghiệp có thể hợp nhất nguồn dữ liệu nội bộ với các nguồn bên ngoài, chẳng hạn như pha trộn nguồn cấp dữ liệu từ các hoạt động phương tiện truyền thông xã hội và dữ liệu công cộng, mà không yêu cầu một mô hình hóa dữ liệu phức tạp.

ClearStory là một trong những khách hàng đầu tiên của Databricks, dựa vào công nghệ Spark đã hình thành nên nền tảng cốt lõi của dòng sản phẩm tương tác thời gian thực của ClearStory. Vaibhav Nivargi là người sáng lập của ClearStory chia sẻ: "Nếu không có Spark sẽ không có dòng sản phẩm của chúng tôi như bây giờ. Spark có khái niệm về sự cư trú phân phối bộ dữ liệu, khi đó các đơn vị dữ liệu trong bộ nhớ sẽ được phân phối trải dài trên nhiều máy của hệ thống cluster, điều này giúp cho chiến lược phân tích của ClearStory khả thi hơn."



Mô hình kiến trúc xử lý dữ liệu của ClearStory

CHƯƠNG 4: ỨNG DỤNG THỬ NGHIỆM CÔNG NGHỆ DỮ LIỆU LỚN TRONG XỬ LÝ ẢNH VĂN BẢN

4.1. Đặt vấn đề

Hoạt động của cơ quan nhà nước luôn gắn liền với văn bản, hồ sơ, giấy tờ. Đây là các đối tượng, là sản phẩm từ các quy trình hành chính cần phải được quản lý, lưu trữ và khai thác theo đúng quy định chung, cụ thể là Luật Lưu trữ do Quốc hội ban hành 11/11/2011, các Nghị định của Chính phủ, Thông tư do Bộ Nội vụ ban hành.

Để ứng dụng CNTT trong hoạt động của các cơ quan nhà nước góp phần nâng cao hiệu quả quản lý, điều hành tác nghiệp, giảm thời gian và chi phí giải quyết các thủ tục hành chính, tăng cường sử dụng văn bản điện tử theo Chỉ thị 15/CT-TTg của Thủ tướng Chính phủ, sau thời gian khảo sát, nghiên cứu hiện trạng thực tế và một số bất cập trong lưu trữ, quản lý, trao đổi hồ sơ - văn bản bằng các công cụ phần mềm, mạng máy tính trong nhiều cơ quan nhà nước tại Hà nội, TP. HCM... Viện CNPM & NDS Việt Nam đã đề xuất và được Lãnh đạo Bộ TT&TT chấp thuận cho thực hiện thí điểm Đề án "Xây dựng, vận hành và khai thác các Kho hồ sơ số hóa - văn bản điện tử phục vụ quản lý nhà nước" tại Văn phòng Bộ TT&TT, một số đơn vị trực thuộc Bộ, một số Sở TT&TT trong cả nước.*(Quyết định số 1182/QĐ-BTTTT ngày 18/08/2014 phê duyệt Đề án)*

Theo các quy định về hành chính, văn bản trong cơ quan nhà nước được phân chia thành hai loại chính là *văn bản phát hành* và *văn bản soạn thảo:*

- **Văn bản phát hành** là loại văn bản đã *hoàn chỉnh* về nội dung, được trình bày (in ấn trên giấy) theo thể thức quy định, được cá nhân và tổ chức có thẩm quyền *ký tên, đóng dấu*, có giá trị pháp lý để trao đổi, thực thi và cần phải được *lưu trữ, quản lý, khai thác* sử dụng theo các quy định chung về văn thư - lưu trữ, như khi chuyển giao phải ghi tên người nhận, người gửi, chế độ gửi văn bản; phải đảm bảo các yêu cầu về bảo mật và an toàn nội dung. Để quản lý lưu trữ hồ sơ - văn bản phát hành, văn thư cơ quan phải lập các bộ hồ sơ, cấp mã số lưu trữ văn bản trong hồ sơ, đăng ký thành các bộ hồ sơ và lưu vào trong kho lưu trữ văn thư của cơ quan...

- **Văn bản soạn thảo** là loại văn bản đang trong quá trình hoàn thiện, được sử dụng trong nội bộ cơ quan hoặc giữa các cơ quan có liên quan, chủ yếu được dùng để trao đổi thông tin trong nội bộ. Việc lưu trữ và trao đổi *văn bản soạn thảo* có thể được thực hiện bằng nhiều hình thức (không nhất thiết phải thông qua hệ thống văn thư lưu trữ

cơ quan và theo các quy định văn thư lưu trữ). Tuy nhiên trong mỗi cơ quan lãnh đạo cần ban hành quy định về lưu trữ và trao đổi phù hợp với hoạt động của cơ quan mình.

Mục tiêu và nội dung hoạt động chính của Đề án là nhằm xây dựng các Kho Dữ liệu phù hợp với yếu cầu sử dụng văn bản điện tử (văn bản phát hành, văn bản soạn thảo) trong các cơ quan nhà nước, cụ thể là:

+ **Kho dữ liệu hồ sơ số hóa** phục vụ chủ yếu nhu cầu lưu trữ và trao đổi (nội bộ, liên cơ quan) các *văn bản phát hành* (thực tế là các *bản sao y điện tử* có giá trị pháp lý tương đương văn bản gốc) bằng các phương tiện điện tử như phần mềm số hóa văn bản và tạo lập thông tin mô tả, các giải pháp phần mềm lưu trữ, xác thực, quản lý quy trình nhận và chuyển văn bản qua mạng máy tính đảm bảo tính *pháp lý*, tuân thủ các yêu cầu *nghiệp vụ văn thư*, *an toàn* và *bảo mật* nội dung văn bản trong quá trình trao đổi văn bản điện tử.

+ **Kho dữ liệu văn bản điện tử** phục vụ chủ yếu nhu cầu lưu trữ và trao đổi (nội bộ) các *văn bản soạn thảo* thông qua các phương tiện điện tử như ứng dụng phần mềm soạn thảo văn bản hay số hóa, các công cụ phần mềm quản lý, tìm kiếm xem văn bản trên máy tính trạm, trên thiết bị di động, thông qua trình duyệt web, gửi nhận qua email, sao chép, chuyển và nhận văn bản điện tử qua mạng máy tính. Việc lưu trữ và trao đổi các loại văn bản điện tử soạn thảo này *không cần đòi hỏi* tính pháp lý đầy đủ, các yêu cầu nghiệp vụ văn thư lưu trữ cũng như bảo mật nội dung thông tin.

Việc thống nhất khái niệm và phân loại Kho Dữ liệu (*hồ sơ số hóa và văn bản điện tử*) như thuyết minh ở trên khi xây dựng, vận hành và khai thác các Kho Dữ liệu sẽ góp phần nâng cao hiệu quả ứng dụng CNTT, giảm độ phức tạp, tăng cường tính khả thi và phạm vi ứng dụng của văn bản điện tử trong cơ quan nhà nước, nếu so sánh tỷ lệ trên 80% văn bản soạn thảo với khoảng 20% văn bản phát hành dùng để trao đổi giữa các cơ quan nhà nước.

Việc xây dựng các kho dữ liệu bằng cách số hóa các văn bản là cần thiết và bắt đầu được triển khai rộng rãi dưới nhiều hình thức. Việc xây dựng các kho dữ liệu này mới dừng lại ở việc số hóa và lưu trữ các văn bản thành và tổ chức thành có kho lưu trữ. Nếu chỉ dừng lại ở đó thì sẽ chưa thấy được các giá trị mà kho dữ liệu đem lại. Mục đích của kho dữ liệu là tập hợp dữ liệu để có thể cung cấp dữ liệu cho các ứng dụng khai thác khai thác và tạo ra các giá trị mới. Để có thể thực hiện được mục tiêu đó thì trước hết kho dữ liệu phải được dữ liệu hóa.

Sự khác biệt giữa số hóa và dữ liệu hóa trở nên rõ ràng khi chúng ta xem xét một

lĩnh vực mà cả hai hiện tượng đã xảy ra và so sánh kết quả của chúng. Năm 2004 Google đã công bố một kế hoạch táo bạo, họlấy tất cả các trang sách của tất cả các cuốn sách mà họ có được và cho phép tất cả mọi người trên toàn thế giới tìm kiếm và truy cập miễn phí qua Internet. Để đạt được điều này công ty đã hợp tác với một số thư viện lớn nhất và uy tín nhất trên thế giới và phát triển những máy quét có thể tự động lật các trang, để việc quét hàng triệu cuôn sách vừa có thể thực hiện được vừa khả thi về mặt tài chính.

Đầu tiên google số hóa văn bản, từng trang được quét và ghi vào một tập tin hình ảnh có độ phân giải cao, được lưu trữ trên máy chủ của google. Trang sách được chuyển thành một bản sao điện tử để có thể dễ dàng truy cập thông qua Web. Tuy nhiên, việc truy cập sẽ đòi hỏi người đọc phải biết cuốn sách nào có thông tin mình quan tâm, hoặc phải đọc nhiều để tìm ra thông tin cần thiết. Người ta không thể tìm kiếm văn bản theo từ khóa, hoặc phân tích nó bởi vì văn bản chưa được dữ liệu hóa. Tất cả những gì google có là hình ảnh mà chỉ con người mới có thể biến đổi thành thông tin hữu ích bằng cách đọc.

Google muốn nhiều hơn nữa, họ hiểu rằng thông tin chưa đựng những giá trị mà chỉ có thể được chuyển tải một khi nó được dữ liệu hóa. Và do vậy Google đã sử dụng phần mềm nhận dạng kỹ tự quang học để nhận dạng ra các chữ cái, từ, câu và đoạn văn. Kết quả là văn bản đã được dữ liệu hóa chứ không chỉ làm một ảnh quét của trang sách.

Bây giờ các thông tin trên trang sách mới có thể được sử dụng không chỉ cho người đọc, mà còn cho các máy tính để xử lý và cho các thuật toán để phân tích. Dữ liệu hóa là cho văn bản có thể lập chỉ mục và do đó có thể tìm kiếm được. Chúng ta có thể so sánh phong cách văn bản và xác định được tác giả khi có tranh chấp tác quyền. Dữ liệu hóa cũng giúp cho việc phát hiện đạo văn trong các công trình hàn lâm trở nên dễ dàng hơn.

Để có thể có được những lợi ích to lớn từ việc ứng dụng công nghệ dữ liệu lớn, việc đầu tiên là phải dữ liệu hóa được kho dữ liệu. Đề tài đã lựa chọn ứng dụng công nghệ xử lý dữ liệu lớn vào việc dữ liệu hóa kho dữ liệu tại viện CNPM & NDS Việt Nam. Trong khuôn khổ đề tài, nhóm nghiên cứu đã xây dựng ứng dụng thử nghiệm cho việc nhận dạng văn bản theo mẫu được triển khai trên nền tảng xử lý dữ liệu lớn Hadoop.

4.2. Nhận dạng văn bản theo mẫu

4.2.1. Các phương pháp nhận dạng ảnh

Có nhiều phương pháp nhận dạng mẫu khác nhau được áp dụng rộng rãi trong các hệ thống nhận dạng kí tự. Các phương pháp này có thể được tích hợp trong các hướng tiếp cận sau: Đối sánh mẫu, thống kê, cấu trúc, mạng nơ ron và SVM

- Máy vecto hỗ trợ (SVM) : ý tưởng chính của phương pháp này là tìm một siêu phẳng phân cách sao cho khoảng cách lề giữa hai lớp đạt cực đại. Khoảng cách này được xác định bởi các véc tơ tựa (SV – Support Vector), các SV này được lọc ra từ tập mẫu huấn luyện bằng cách giải một bài toán tối ưu lồi.

- Phương pháp tiếp cận cấu trúc: dựa vào việc mô tả đối tượng nhờ một số khái niệm biểu diễn đối tượng cơ sở trong ngôn ngữ tự nhiên. Để mô tả đối tượng người ta dùng một số dạng nguyên thuỷ như đoạn thẳng, cung,… Mỗi đối tượng được mô tả như một sự kết hợp của các dạng nguyên thuỷ.

- Phương pháp ngữ pháp (Grammatical Methods): Các phương pháp ngữ pháp khởi tạo một số luật sinh để hình thành các ký tự từ một tập các công thức ngữ pháp nguyên thủy. Các luật sinh này có thể kết nối bất kỳ kiểu đặc trưng thống kê và đặc trưng hình thái nào dưới một số cú pháp hoặc các luật ngữ nghĩa Giống như lý thuyết ngôn ngữ, các luật sinh cho phép mô tả các cấu trúc câu có thể chấp nhận được và trích chọn thông tin theo ngữ cảnh về chữ viết bằng cách sử dụng các kiểu ngữ pháp khác nhau.

- Phương pháp đồ thị (Graphical Methods): Các đơn vị chữ viết được mô tả bởi các cây hoặc các đồ thị. Các dạng nguyên thủy của ký tự (các nét) được lựa chọn bởi một hướng tiếp cận cấu trúc. Đối với mỗi lớp, một đồ thị hoặc cây được thành lập trong giai đoạn huấn luyện để mô tả các nét, các ký tự hoặc các từ. Giai đoạn nhận dạng gán một đồ thị chưa biết vào một trong các lớp bằng cách sử dụng một độ đo để so sánh các đặc điểm giống nhau giữa các đồ thị.

- Mô hình Markov ẩn (Hidden Markov Model): Mô hình Markov ẩn (HMM) là một trong những mô hình máy học quan trọng nhất trong xử lý ngôn ngữ tự nhiên và nhận dạng. Mô hình này là trường hợp mở rộng của máy hữu hạn trạng thái có hướng, có trọng số. HMM thường được sử dụng để xử lý những sự kiện không quan sát trực tiếp được (sự kiện ẩn). Do vậy, HMM được ứng dụng để giải quyết

những bài toán có độ nhiễu lớn, chẳng hạn, dự báo, nhận dạng tiếng nói,…

- Phương pháp đối sánh mẫu: Kỹ thuật nhận dạng chữ đơn giản nhất dựa trên cơ sở đối sánh các nguyên mẫu (prototype) với nhau để nhận dạng ký tự hoặc từ. Nói chung, toán tử đối sánh xác định mức độ giống nhau giữa hai véc tơ (nhóm các điểm, hình dạng, độ cong...) trong một không gian đặc trưng.

Với phương pháp đối sánh mẫu qua việc sử dụng một thư viện được xây dựng sẵn (tiếng Việt, tiếng Anh, ký tự toán học, …) có khả năng mở rộng chỉnh sửa cao, tạo được những "key word" (từ khóa) áp dụng cho việc tìm kiếm thông tin của ảnh văn bản.

4.2.2. Nhận dạng ảnh văn bản theo mẫu

Việc nhận dạng ảnh của một văn bản hiện nay thường được xử lý và nhận dạng với các ký tự quang học độc lập (OCR - Optical Character Recognition), sau đó dùng các phương pháp phục hồi để chuyển thành dạng văn bản có thể đọc. Tuy nhiên phương pháp này không hoạt động tốt trên chữ viết tay, gặp lỗi khi thực hiện ghép thành câu từ hoàn chỉnh và có nghĩa.

Một phương pháp khác là nhận biết các từ bỏ qua giai đoạn nhận dạng ký tự bằng cách sử dụng những bộ từ điển mẫu từ để so sánh sự tương đồng. Phương pháp này cũng giảm được độ nhiễu của hình ảnh và tăng tốc độ xử lý so với phương pháp nhận dạng ký tự riêng lẻ do số từ trung bình nhỏ hơn nhiều số ký tự trên một ảnh văn bản. Nhưng về mặt chính xác kém hơn so với phương pháp nhận dạng ký tự quang học độc lập.

Trong Hình 4.1 ví dụ một số mẫu trong thư viện như cùng một từ "program" có nhiều mẫu chữ khác nhau, để đảm bảo việc nhận dạng chính xác thì bộ thư viện đòi hỏi một sự đa dạng, phong phú. Trong khuôn khổ bài toán tìm kiếm ảnh văn bản, không đặt nặng vấn đề nhận dạng chính xác toàn văn của ảnh văn bản mà chỉ đòi hỏi rút trích được những từ có trong ảnh văn bản.

Hình 4.1: Một số mẫu nhận dạng trong các thư viện

Từ việc thống kê ngôn ngữ sự lặp lại của những từ thông dụng trên một trang (báo, tài liệu) xảy ra thường xuyên. Việc sử dụng mẫu của những thừ phổ biến này cũng cải thiện đáng kể tốc độ xử lý, dễ dàng trong việc đánh chỉ mục, phân nhóm để cải thiện hiệu năng tìm kiếm ảnh văn bản.

word	freq	word	freq	word	freq	word	freq
THE	0.071	IN	0.022	ON	0.008	AS	0.006
OF	0.032	FOR	0.010	HE	0.007	BY	0.006
AND	0.024	THAT	0.009	AT	0.007	IT	0.005
A	0.024	IS	0.008	WITH	0.006	HIS	0.005
TO	0.023	WAS	0.008	BE	0.006	SAID	0.004

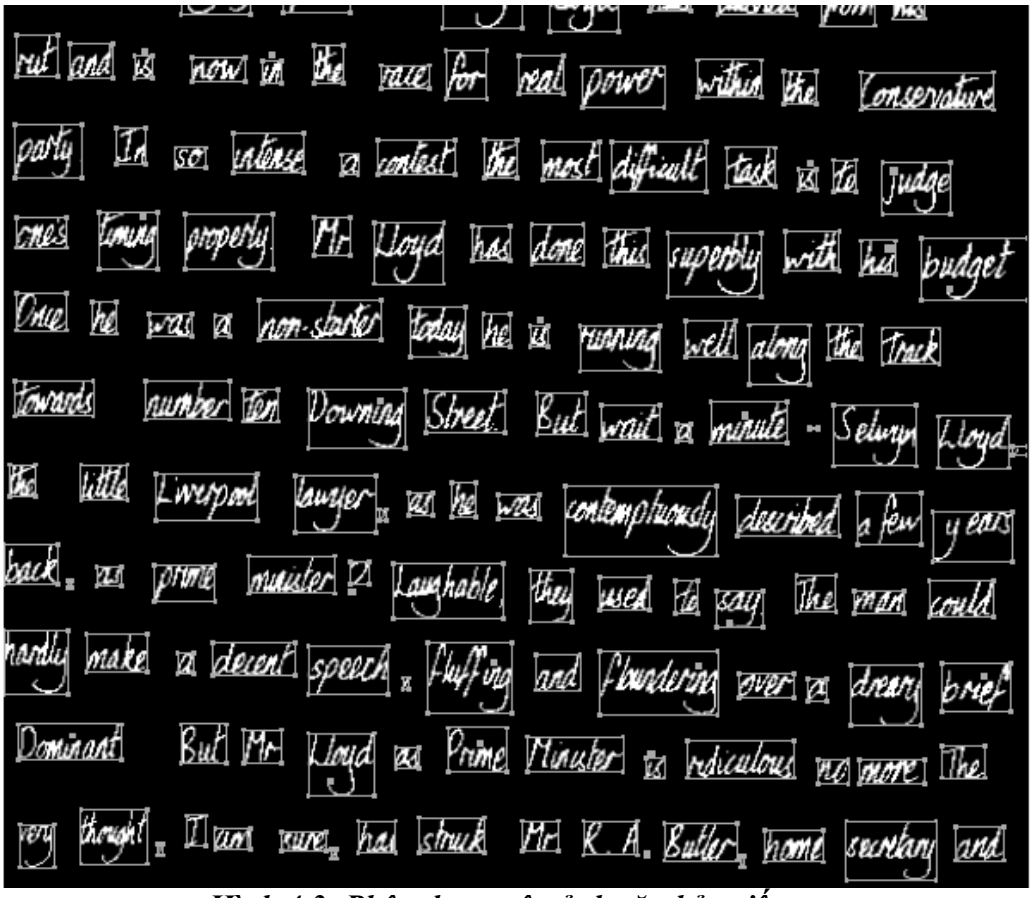
Hình 4.2: Thống kê 20 từ xuất hiện nhiều nhất trong 90 000 bài báo tiếng Anh

Theo thống kê như hình 4.2, trung bình trong một tài liệu tiếng Anh có sự xuất hiện tới 7% từ "THE" , 3% từ "OF".Và với hai mươi từ thông dụng ở ví dụ trên chiếm tới 29% của mẫu trong tài liệu tiếng Anh. Điều đó cũng có nghĩa một phần ba số từ trong ảnh văn bản có thể được nhận dạng chỉ với hai mươi từ thông dụng.

Để dễ dàng cho việc đánh chỉ mục và tăng cao hiệu năng của việc tìm kiếm trong ảnh văn bản, đề tài chỉ thử nghiệm việc sử dụng phương pháp nhận dạng theo từ cho việc nhận dạng và trích xuất ảnh văn bản.

Để nhận dạng từ trên ảnh văn bản cần sử dụng kỹ thuật phân đoạn, kỹ thuật này giả định trên ảnh văn bản chỉ có hai màu trắng và đen (ký tự là màu trắng và nền đen), khoảng cách giữa các ký tự liền kề nhỏ hơn so với khoảng cách giữa các từ liền kề. Từ đó

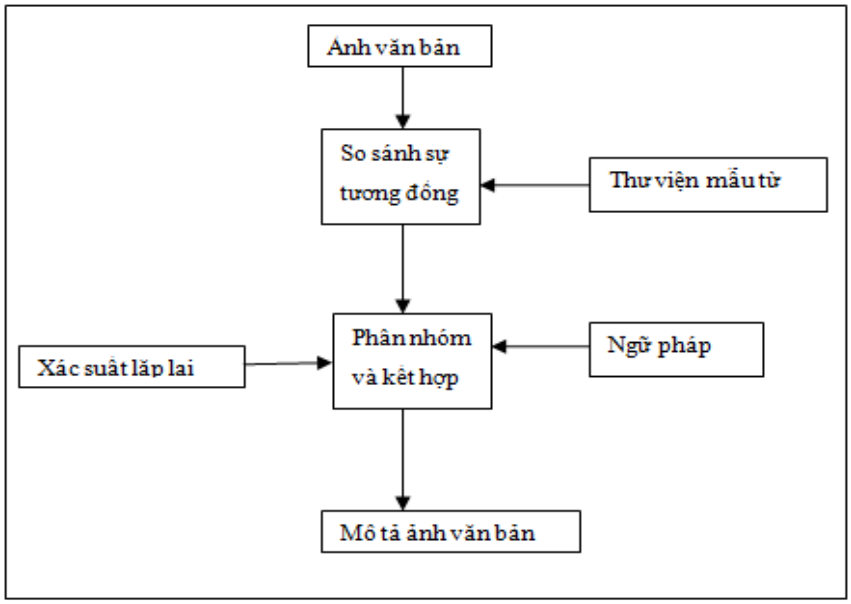
xây dựng một hình ảnh mới qua các quá trình xử lý độ nghiêng, làm giảm nhiễu, nếu khoảng cách giữa hai điểm ảnh màu trắng liền kề nhỏ hơn một một số k (khoảng cách trung bình giữa hai từ) thì tất các các điểm ảnh giữa hai điểm ảnh đó là màu trắng. Bằng phương pháp này khi thực hiện theo phương dọc và phương ngang có thể tạo được một vùng bao ngoài tối thiểu là một hình chữ nhật trắng quanh từ:



Hình 4.3: Phân đoạn trên ảnh văn bản viết tay

Trong hình 4.3 minh họa cho việc phân đoạn ảnh theo phương pháp đã trình bày ở trên. Một số lỗi có thể sảy ra như những dấu chấm, từ viết sai, thừa. Điều này cũng có thể bỏ qua khi đã xác định được kích thước tối thiểu ảnh của một từ. Như khi văn bản được đánh trên cùng một định dạng chữ hoặc văn bản viết tay được viết bởi cùng một người, những lỗi sai sẽ được giảm thiểu. Điều này cũng hoàn toàn phù hợp trên thực tế.

Sau quá khi phân đoạn ảnh, tiếp theo đem so sánh vùng đã được phân đoạn với những mẫu trong thư viện. Để việc nhận dạng chính xác hơn có thể kết hợp với cú pháp ngữ pháp mà văn bản sử dụng.



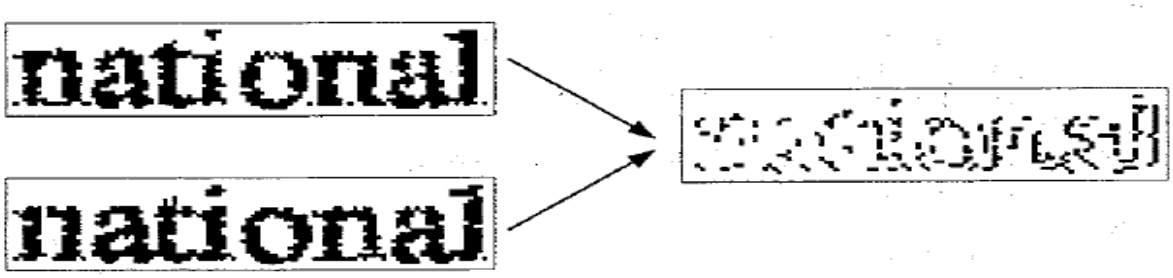
Hình 4.4: Mô tả quá trình nhận dạng ảnh văn bản bằng phương pháp mẫu từ

Một hình ảnh văn bản được chia nhỏ thành các từ được phân cách bởi khoảng trắng trên ảnh, dựa vào xác xuất lặp lại của các từ đăng trưng (dựa vào ngôn ngữ, chuyên ngành, loại dữ liệu, của các bộ từ điển mẫu khác nhau) để tìm kiếm việc tái xuất của từ trên hình ảnh, bươc tiếp nhận dạng từ riêng lẻ sau dựa vào hình ảnh của mẫu và xác nhận các câu, đoạn từ ngữ pháp, cú pháp chính tả để đưa ra mô tả ảnh văn bản phù hợp với dữ liệu ban đầu.

Các cơ sở để nhận dạng mẫu chữ:

- Đầu tiên xác định hình ảnh của từ được đưa vào so sánh bằng cách căn lề theo phương ngang và thẳng dứng dựa vào đường cơ sở. Đường cơ sở được tính bằng cách xác định bằng phương pháp phân đoạn đã trình bày trên.

- Sau đó hình ảnh sẽ được chuyển đổi thành một vecto đặc trưng bằng cách chia hình ảnh thành một ô lưới 4 x 8. Sau đó tính gradient, cấu trúc và tính lồi lõm của mỗi ô của lưới. Kết quả là một vector nhi phân với độ dài là 1024. Ở ví dụ trong Hình 4.5 thể hiện kết quả khi XOR hai ảnh “national”. Kết quả của phép tính này được so sánh với một ngưỡng trung bình để xác nhận hình ảnh. Để tăng độ chính xác có thể áp dụng các thuật toán xử lý đồ họa, xử lý về mặt ngữ pháp, kết cấu câu từ trong đoạn văn.

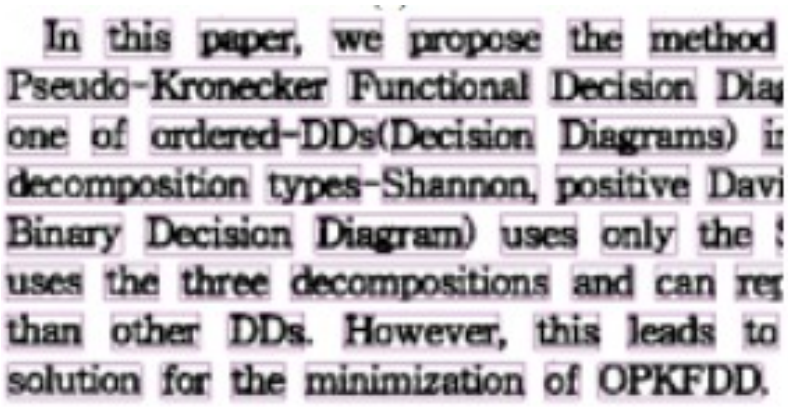


Hình 4.5: Kết quả khi thực hiện so sánh hai mẫu ảnh của một chữ

Theo phương pháp này, với bộ thư viện càng đầy đủ và đa dạng, thì khả năng nhận dạng càng tối ưu. Không chỉ các định dạng là ảnh của các mẫu chữ được đánh máy mà còn các văn bản được viết tay, các ngôn ngữ phức tạp như tiếng Trung Quốc, tiếng A Rập,..

4.2.3. Lập chỉ mục từ trong văn ảnh văn bản sử dụng mẫu từ tương đồng

Để áp dụng vào bài toán tìm kiếm phương pháp nhận dạng theo mẫu, việc lập chỉ mục không chỉ trên những thông tin văn bản thuần túy như tiêu đề tác giả, ngày tháng lập,.. mà còn trên những chữ sau quá trình phân đoạn và nhận dạng theo phương pháp nhận dạng theo mẫu.


In this paper, we propose the method
Pseudo-Kronecker Functional Decision Dia
one of ordered-DDs(Decision Diagrams) i
decomposition types-Shannon, positive Davi
Binary Decision Diagram) uses only the
uses the three decompositions and can re
than other DDs. However, this leads to
solution for the minimization of OPKFDD.

Hình 4.6: Ví dụ về phân đoạn từ trên ảnh

Quá trình tiền xử lý hình ảnh áp dụng phương pháp sử dụng mẫu:

- Các văn bản in được quét, chụp thành các file ảnh được lưu trong ổ cứng.
- Sau đó các file này được nhị phân hóa theo ngưỡng của hình ảnh (thành các hình ảnh tối giản chỉ có trắng và đen).

- Phân đoạn các hình ảnh thành các từ, đối sánh với mẫu trong bộ thư viện mẫu phù hợp.

- Ghi nhớ những mẫu từ thích hợp (những từ thường được sử dụng nhất) được lưu lại làm mẫu đặc trưng để gom nhóm tất cả những từ nào phù hợp với nó trong tất cả các tài liệu bằng cách dựa vào diện tích của vùng xuất hiện và tỉ lệ của các từ. Tiếp theo kết hợp với việc so sánh khoảng cách tối thiểu bằng phép XOR hình ảnh có thể dễ dàng tính tần số xuất hiện của một từ và phân lớp nó.

- Phần đánh chỉ mục: Đối với những từ phù hợp với lớp đặc trưng thường xuất hiện ta có thể bỏ qua, và đánh chỉ mục theo mẫu đó.

Sau bước tiền xử lý các tài liệu ảnh văn bản được mô tả dưới dạng một danh sách theo các mẫu chữ dạng chuẩn ASCII kèm theo tọa độ, số lần lặp lại các chữ trong hình ảnh, vị trí trong ảnh.

Token	Word	Area	E_{EDmin}	Xshift	Yshift
280	Standard	1530	0.000	0	0
239	comment	1722	0.203	-4	0
94	come to	1241	0.212	1	0
45	whether	1258	0.212	1	0
186	branch	1743	0.218	0	0
56	subscribes	1900	0.228	-4	0
283	substances	1479	0.231	1	0
167	Standard	1440	0.231	1	0

Hình 4.7: Văn bản được đánh chỉ mục theo vùng và tọa độ

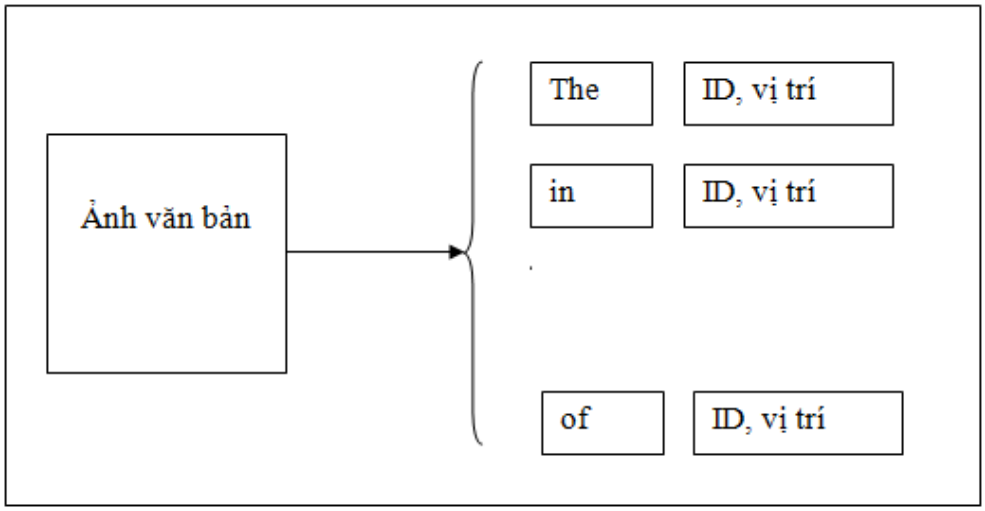
4.3. Ứng dụng công nghệ dữ liệu lớn để xử lý ảnh văn bản

4.3.1. Tìm kiếm ảnh văn bản áp dụng công nghệ dữ liệu lớn

Hadoop MapReduce là một mô hình lập trình hỗ trợ đa dạng các loại dữ liệu. Nhưng giải pháp MapReduce không phải là mô hình áp dụng cho mọi vấn đề, trên thực tế giải pháp này áp dụng tốt cho co các trường hợp lớn được xử lý phân tán song song. Để tìm kiếm và ảnh văn bản có nhiều phương pháp nhưng với khối lượng lớn dữ liệu và đặc

biệt không phải là dạng dữ liệu có cấu trúc, nên sử dụng công nghệ dữ liệu lớn (Hadoop) để tìm kiếm dữ liệu chỉ phụ thuộc vào các tập dữ liệu được phân tích của ảnh văn bản.

Độ tương đồng giữa nội dung được truy vấn và ảnh văn bản phụ thuộc vào tần số lặp lại của từ khóa trong nội câu truy vấn trong dữ liệu mô tả ảnh văn bản. Qua quá trình xử lý dữ liệu ảnh văn bản thô, mỗi ảnh văn bản được mô tả dưới dạng một tập các mẫu từ đã được trích xuất.



Hình 4.8: Dữ liệu ảnh văn bản được trích xuất

Để hoàn thành được yêu cầu của người tìm kiếm (nhập từ khóa tìm kiếm có liên qua đến ảnh văn bản) và nhận được môt danh sách kết quả (ảnh dữ liệu chứa từ khóa tìm kiếm) được xắp xếp với một tiêu chí nào đó

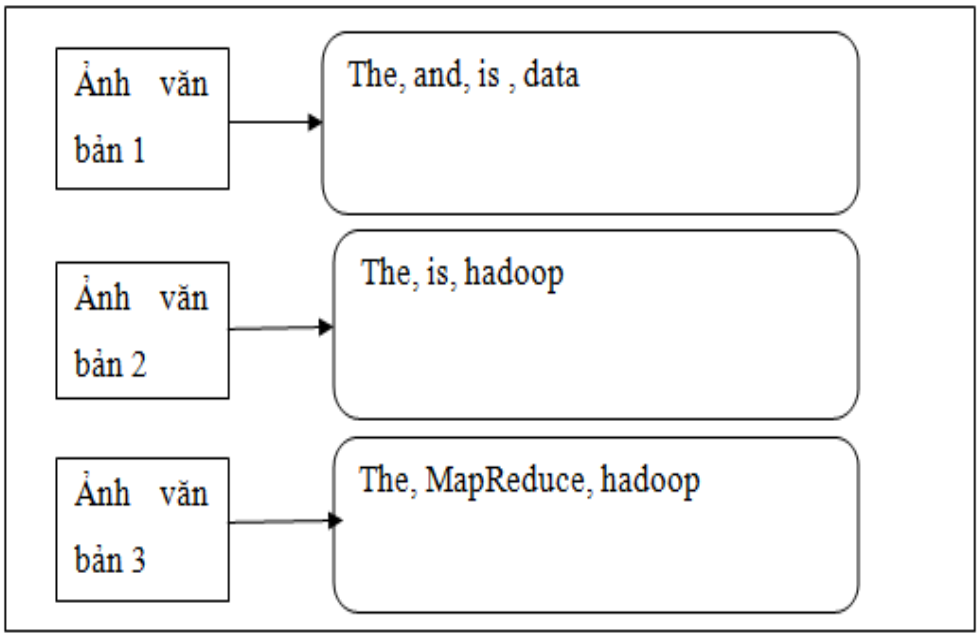
Đánh chỉ mục ngược (INVERTED INDEXING)

Các dữ liệu ảnh văn bản sau khi được trích xuất sẽ được chương trình tự động phân tách và tạo chỉ mục ngược (reverse index): chỉ mục với khoá là từ khoá và value là danh sách các tài liệu có mặt từ khoá). Kết quả của quá trình này là một khối chỉ mục ngược.

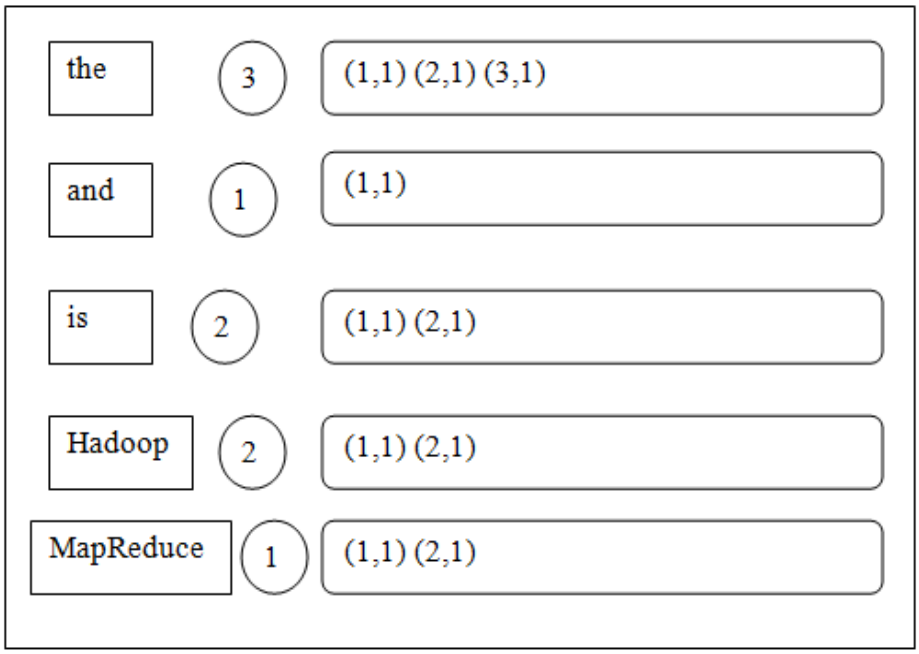
Tìm kiếm

Khi người dùng nhập một câu query (thông thường là một từ khoá), hệ thống sẽ thực hiện tìm kiếm trên khối chỉ mục để tìm ra những tài liệu phù hợp nhất (khớp nhất) với query.

Về cơ bản việc đánh chỉ mục là một danh sách thông tin được liên kết với dữ liệu gốc. Một ảnh văn bản sau khì được trích xuất thành các từ (word) và được đánh chỉ mục theo từ với thông tin kèm theo là vị trí, số lần xuất hiện, mã của tài liệu.



Hình 4.9: Các từ xuất hiện được trong các ảnh văn bản



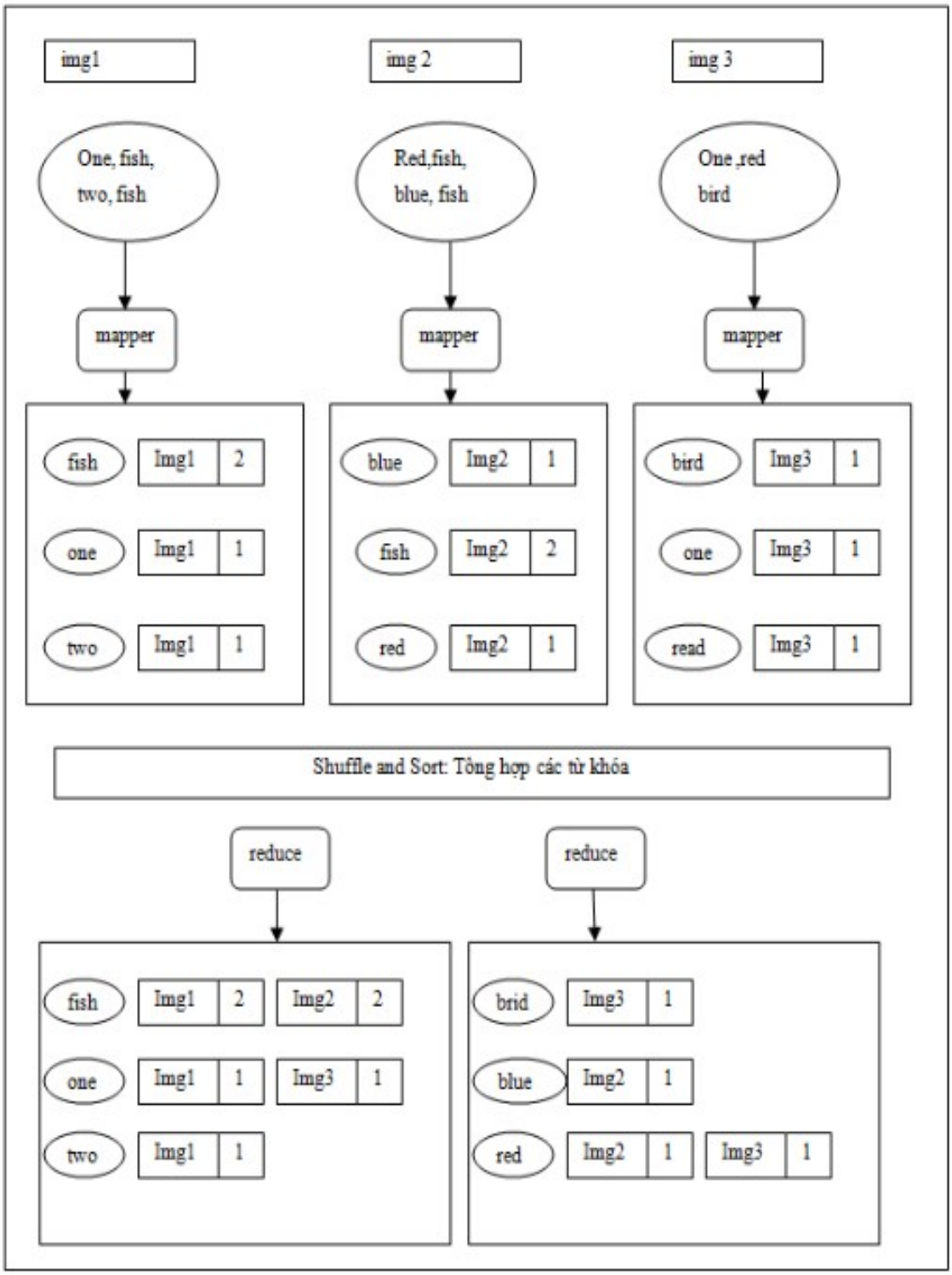
Hình 4.10: Minh họa chỉ số ngược

Ở trên minh họa cách đánh chỉ mục các từ khóa trong ảnh văn bản gồm số lần xuất hiện, mã (ID) của ảnh văn bản, tần số xuất hiện trong ảnh văn bản. Để việc thực hiện truy vấn có hiệu quả nhất thông tin được gán vào khóa chỉ cần mã của ảnh văn bản. Để có thể thực hiện truy vấn một cách tối ưu hơn có thể mở rộng thông tin cho khóa bằng cách thêm các trường như vị trí, mô tả từ khóa.

Để giải quyết vấn đề tìm kiếm ảnh văn bản, các tập ảnh phải được trích xuất và tạo được tập tài liệu chỉ mục hoàn chỉnh, bằng phương pháp lập này những thông tin về từ

khóa, tần số lặp lại trong tài liệu ảnh được lưu thành nhiều bản ghi, mỗi bản ghi lưu trữ thông tin về chỉ số lặp lại của một từ có trong ảnh.

Sau quá trình Tổng hợp và sắp xếp (Shuffle and Sort) các thông tin của các khóa được được lưu trữ thành nhiều bản ghi mỗi bản ghi lưu trữ từ khóa và thông tin cũng như trọng số xuất hiện trong mỗi ảnh văn bản.

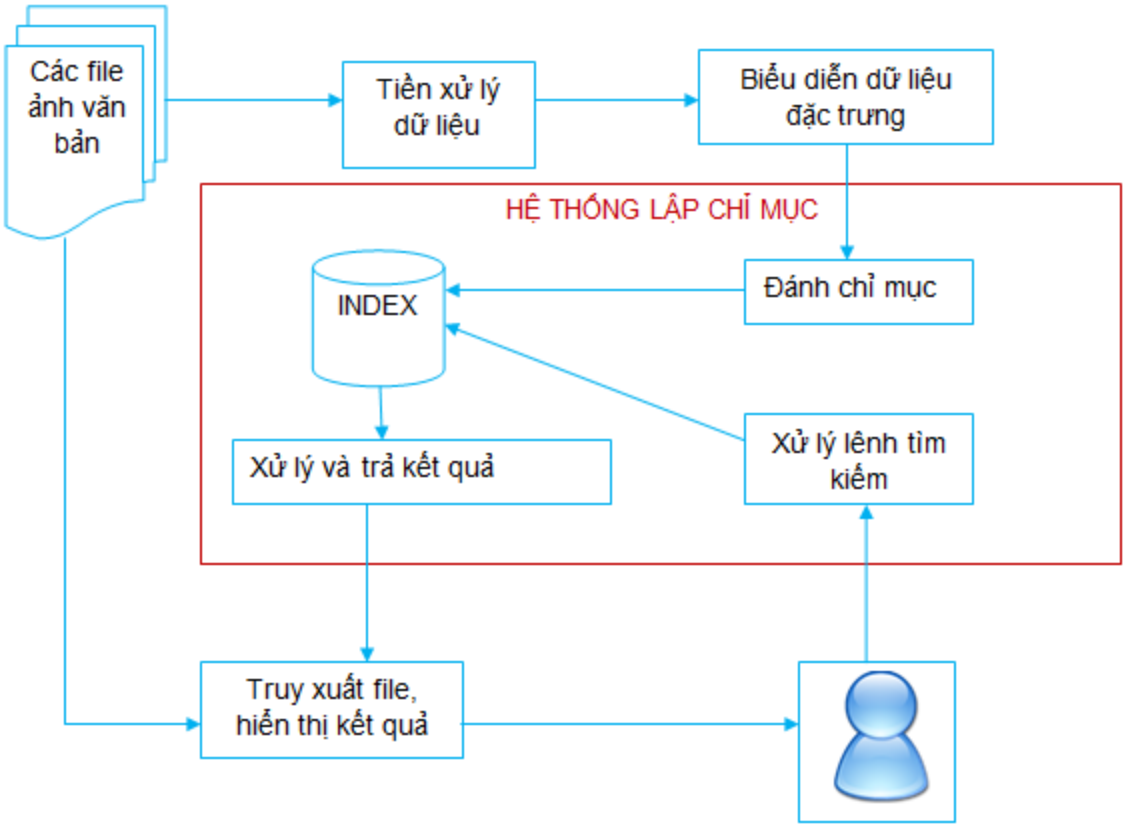


Hình 4.11: Hình minh họa thuật lập chỉ mục đơn giản với 3 maper và 2 reduce

Nếu xét trong một node của hệ thống và với một yêu cầu tìm kiếm thì số lượng phép toán xử lý không nhiều hoàn toàn không cần thiết áp dụng nền tảng Big Data (ở đây là mô hình Hadoop MapReduce). Nhưng nếu với một lượng dữ liệu lớn, và lượng người dùng nhiều thì việc áp dụng hệ thống phân tán, nhiều luồng cùng xử lý một lúc thì mô hình này hòa toàn thích hợp. Với một hệ thống lớn để đảm bảo tốc độ, cần tổ chức xử lý phân tán, song song công việc này, mỗi máy tính sẽ chịu trách nhiệm sắp xếp các tài liệu có độ tương đồng nằm trong một khoảng nhất định. Để làm được điều đó yếu cầu phải chuẩn hóa dữ liệu và tổ chức kho dữ liệu trong kho dữ liệu một cách khoa học.\

4.4. Xây dựng ứng dụng tìm kiếm ảnh văn bản

4.4.1. Các bước thực hiện xây dựng chương trình tìm kiếm

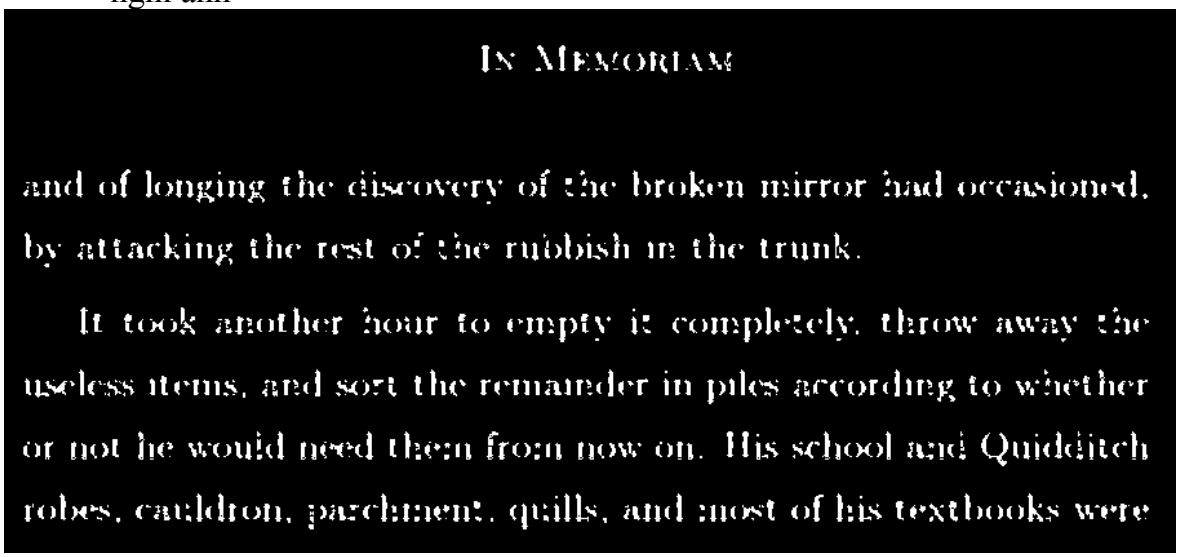


Hình 4.12: Các bước xử lý của chương trình tìm kiếm

Bước 1: Chuẩn bị dữ liệu ảnh văn bản, các tệp tin ảnh văn bản được lưu trữ trên ổ đĩa với định dạng ảnh (JPG, PNG,…)
Bước 2: Tiền xử lý:

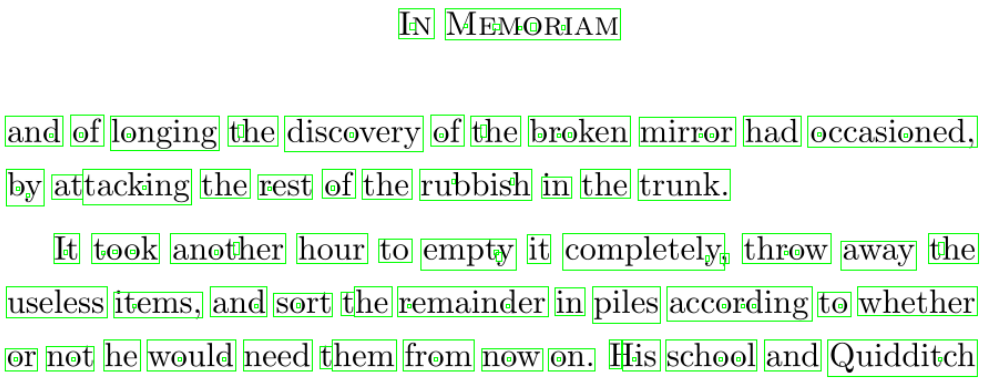
- Xử lý nhiễu
- Chuyển ảnh ban đầu thành dạng ảnh đa mức xám và áp dụng ngưỡng thích

nghi ảnh



Hình 4.13: Dạng ảnh xám

- Phân đoạn ảnh



Hình 4.14 Minh họa phân đoạn ảnh văn bản

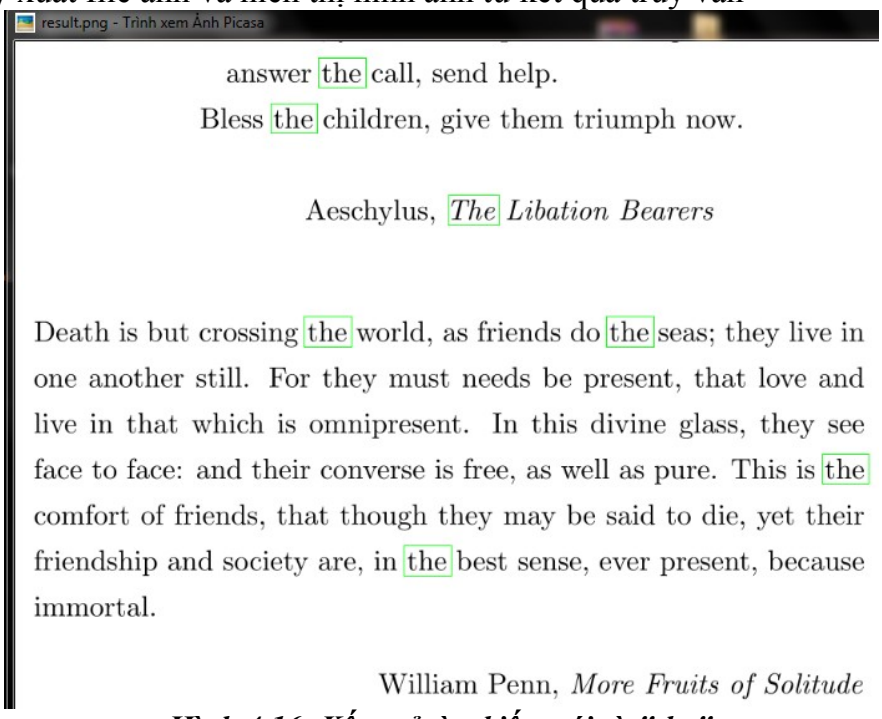
- Nhận dạng mẫu chữ
- Lưu đặc trưng của ảnh ra file xml

```
<doc>
  <field name="key">indeed</field>
  <field name="template">template/528.PNG</field>
  <field name="image">image/HarryPotterAnd TheDeathlyHallows/(Book 7) Harry Potter And The Deathly Hallows-034.png</field>
  <field name="count">1</field>
  <field name="xy1">1345;1377;376;475</field>
</doc>
<doc>
  <field name="key">brilliance</field>
  <field name="template">template/613.PNG</field>
  <field name="image">image/HarryPotterAnd TheDeathlyHallows/(Book 7) Harry Potter And The Deathly Hallows-034.png</field>
  <field name="count">1</field>
  <field name="xy1">545;577;385;513</field>
</doc>
<doc>
  <field name="key">chapters</field>
  <field name="template">template/637.PNG</field>
  <field name="image">image/HarryPotterAnd TheDeathlyHallows/(Book 7) Harry Potter And The Deathly Hallows-034.png</field>
  <field name="count">1</field>
  <field name="xy1">343;379;413;532</field>
</doc>
```

Hình 4.15: Biểu diễn dữ liệu tiền xử lý

Hình trên minh họa dữ liệu tiền xử lý của ảnh văn bản, đặc trưng của ảnh văn bản được lưu trữ dưới dạng file xml.Trong đó các trường: “key” chứa từ trích xuất được trong ảnh, “template” chứa đường dẫn của mẫu nhận dạng, “image” chứa đường dẫn của ảnh,

“count” chứa số tần xuất của từ trong ảnh văn bản, “xy1” chứa tọa độ của khu vực xuất hiện từ.
Bước 3: Dữ liệu tiền xử lý được đưa vào hệ thống đánh chỉ mục (Apache lucene)
Bước 4: Xây dựng công cụ giao tiếp để thực hiện truy vấn với hệ thống (Apache lucene)
Bước 5: Truy xuất file ảnh và hiển thị hình ảnh từ kết quả truy vấn



Hình 4.16: Kết quả tìm kiếm với từ "the"

4.4.2. Kết quả

Kết quả chạy với hệ thống tuần tự:

CPU core	4 (2.93GHz)	
RAM	4GB	
OS	Ubuntu 14.04	
Số file ảnh văn bản	339	
Thời gian tiền xử lý (phút)	350	
Dữ liệu trích xuất (mẫu)	65534	
Thời gian thực hiện tìm kiếm (ms)	889	

Hình 4.17: Kết quả thực hiện với hệ thống tuần tự

Kết quả khi chạy với Hadoop:

CPU core	4 (2.93GHz)	
RAM	4GB	
OS	Ubuntu 14.04	
Hadoop version	2.6.0	
Số file ảnh văn bản	339	
Thời gian tiền xử lý (phút)	197	
Dữ liệu trích xuất (mẫu)	65534	
Thời gian thực hiện tìm kiếm (ms)	889	

Hình 4.18: Kết quả thực hiện với hệ thống Hadoop

4.5. Đánh giá và khuyến cáo

Theo kết quả thử nghiệm cho thấy trích xuất 339 file ảnh theo phương pháp thông thường cần thời gian tiền xử lý 350 phút, khoảng 6 giờ. Khi áp dụng Hadoop để xử lý song song, vẫn cấu hình máy tính như vậy cho thời gian xử lý chỉ 197 phút (hơn 3 giờ) giảm gần ½ thời gian. Trong thực tế để có thể xử lý một khối lượng văn bản lớn (hàng triệu trang băn bản) cần có một hệ thống cluster với nhiều máy tính để xử lý. Điều này hoàn toàn khả thi khi ứng dụng nền tảng xử lý dữ liệu Hadoop:

- Do Hadoop là một nền tảng mã nguồn mở ban đầu được phát triển để phục vụ tìm kiếm hơn nên việc áp dụng cho mô hình tìm kiếm ảnh văn bản là hoàn toàn thích hợp. Có thể tải các bản phân phối và hướng dẫn cài đặt tương đối chi tiết.

- Hơn nữa Hadoop cung cấp một mô hình khai thác dữ liệu mới, hiện đại mà trước đây chưa từng có. Cho phép sử dụng khai thác các loại dữ liệu đa dạng nói chung , trong khuôn khổ là ảnh văn bản nói riêng theo cách mà trước đây chưa làm được. Có thể mở rộng, tích hợp phần trích xuất dữ liệu ảnh văn bản để nâng cao hiệu năng xử lý.

- Về chi phí, Hadoop là một nền tảng dữ liệu được thiết kế để chạy trên phần cứng chi phí thấp thay cho các phần cứng chuyên môn hóa đắt tiền

- Ngoài ra với mô hình lập trình Map Reduce trong hệ thống Hadoop, các hệ thống điện toán đám mây dựa trên mã nguồn mở có thể xử lý dữ liệu trên nhiều máy tính trong môi trường đã được phân phối.

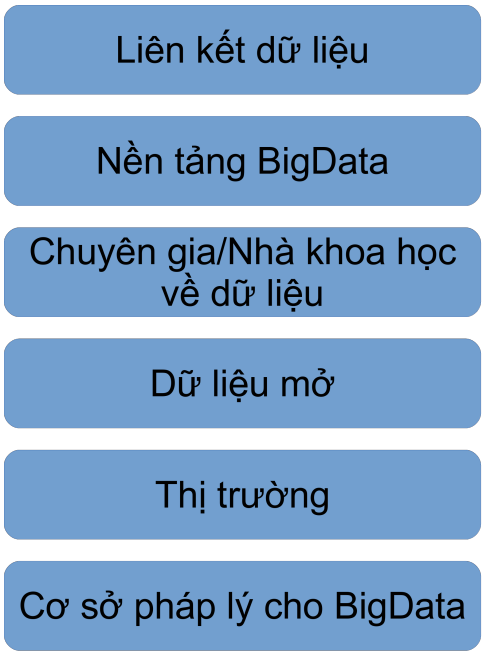
Cùng với việc xây dựng các Kho dữ liệu để lưu trữ và tập hợp các văn bản phục vụ cho xử lý, ứng dụng phân tích và nhận dạng ảnh văn bản theo mẫu dựa trên nền tảng Hadoop đã góp phần tăng thêm nhiều giá trị cho dữ liệu văn bản. Đây chưa phải là một ứng dụng dữ liệu lớn hoàn thiện nhưng đó là một bước tiếp cận gần hơn đến mục đích cuối cùng là xây dựng một ứng dụng dữ liệu lớn hoàn thiện. Dữ liệu nhận dạng văn bản cùng với các dữ liệu từ các nguồn khác (CSDL quan hệ, mạng xã hội...) sẽ là đầu vào cho các ứng dụng khác giúp tìm ra các giá trị mới nhờ việc khai phá dựa trên công nghệ dữ liệu lớn. Kết quả của ứng dụng sẽ được sử dụng để nhóm nghiên cứu tiếp tục phát triển hoàn thiện và đưa và ứng dụng thực tế trong Kho dữ liệu tại Viện CNPM & NDS cũng như các đơn vị đã triển khai trong Đề án Kho dữ liệu do Viện chủ trì.

CHƯƠNG 5: MỘT SỐ KIẾN NGHỊ VÀ ĐỀ XUẤT

5.1. Đề xuất xây dựng chiến lược phát triển công nghệ dữ liệu lớn

Công nghệ dữ liệu lớn mới chỉ bắt đầu phát triển tại Việt Nam, còn rất nhiều cơ hội cho các tổ chức, công ty tham gia vào thị trường này. Tuy nhiên, nếu cứ để phát triển một cách tự do mang nhiều tính tự phát thì lâu dài sẽ dẫn đến sự không thống nhất trong việc chia sẻ dữ liệu, hạ tầng và nền tảng kỹ thuật và nghiêm trọng hơn là sự vi phạm tính riêng tư và tạo ra sự "độc tài dữ liệu". Để có thể định hướng đúng cũng như thúc đẩy nhanh sự phát triển, bắt kịp xu hướng của thế giới về công nghệ dữ liệu lớn thì nhà nước cần có những chính sách phù hợp ngay từ bây giờ.

Để có được một chiến lược phát triển công nghệ dữ liệu lớn toàn diện, đề tài đề xuất cần phải xem xét, xây dựng các chính sách theo 6 vấn đề sau:



Liên kết dữ liệu:

Việc xây dựng các ứng dụng dữ liệu lớn phụ thuộc vào cách chúng ta thu thập và lưu trữ dữ liệu. Một khi chúng ta đã có dữ liệu thì việc xây dựng các ứng dụng dữ liệu lớn chỉ bị hạn chế bởi nhu cầu và trí tưởng tượng của chúng ta. Chính vì vậy, để công nghệ dữ liệu lớn có thể thực sự phát triển được thì trước hết phải thu thập và liên kết được các loại dữ liệu. Chúng ta phải tạo ra một hạ tầng để có thể liên kết các dữ liệu từ địa phương đến trung ương, giữa công và tư. Để thực hiện cần phải xây dựng các chính sách phù hợp, bao gồm:

- Xây dựng các tiêu chuẩn, quy chuẩn trong việc tạo lập, lưu trữ và trao đổi dữ liệu giữa các đơn vị từ trung ương đến địa phương.

- Xây dựng tiêu chuẩn, quy chuẩn cho hạ tầng kết nối, chia sẻ và lưu trữ dữ liệu quốc gia.

- Xây dựng các chính sách nhằm khuyến khích hay bắt buộc triển khai việc tạo lập, lưu trữ và trao đổi các loại dữ liệu.

Nền tảng kỹ thuật xử lý dữ liệu lớn:

Để triển khai được công nghệ dữ liệu lớn chúng ta cần có công cụ, đó chính là một nền tảng kỹ thuật để xử lý dữ liệu lớn. Đối với một quốc gia đang phát triển như Việt Nam thì việc xây dựng hạ tầng kỹ thuật xử lý dữ liệu lớn dựa trên mã nguồn mở là lựa chọn tối ưu, có như vậy mới không bị lệ thuộc công nghệ nước ngoài, tiết kiệm được chi phí và có thể làm chủ hoàn toàn về công nghệ. Một số công việc cần thực hiện:

- Xây dựng trung tâm xử lý dữ liệu lớn quốc gia để nghiên cứu, thiết kế và phát triển kiến trúc nền tảng xử lý dữ liệu lớn dựa trên các phần mềm mã nguồn mở đồng thời phục vụ nhu cầu xử lý dữ liệu của nhà nước và xã hội.

- Công bố và hỗ trợ nền tảng xử lý dữ liệu lớn dưới dạng nguồn mở để các doanh nghiệp có thể tiếp cận, giúp giảm bớt chi phí và rút ngắn thời gian phát triển các ứng dụng dữ liệu lớn, góp phần đẩy mạnh sự phát triển của các doanh nghiệp tham gia vào thị trường xử lý dữ liệu lớn.

Chuyên gia/Nhà khoa học về dữ liệu:

Trong thời đại dữ liệu lớn, một nghề chuyên môn mới đã xuất hiện đó là "nhà khoa học dữ liệu", kết hợp các kỹ năng của nhà thống kê, người lập trình phần mềm, nhà thiết kế thông tin đồ họa, và người kể chuyện. Học viện McKinsey Global - Hoa Kỳ đã đưa ra những dự đoán bi quan về sự khan hiếm các nhà khoa học dữ liệu trong cả hiện tại và tương lai. Có thể nói yếu tố con người chính là yếu tố quan trọng nhất giúp đẩy nhanh sự phát triển của công nghệ dữ liệu lớn tại Việt Nam, là yếu tố chúng ta có thể cạnh tranh được với thế giới. Chính vì vậy, ưu tiên tập trung vào phát triển nguồn nhân lực phải là chinhs sách hàng đầu:

- Khuyến khích các cơ sở đào tạo thiết kế các chương trình giảng dạy về dữ liệu lớn, và tập trung vào đào tạo nguồn nhân lực cho ngành khoa học dữ liệu lớn.

- Thành lập các trung tâm, viện nghiên cứu chuyên sâu về dữ liệu lớn.

Dữ liệu mở:

Ý tưởng về "dữ liệu mở" không phải là mới nhưng đến khi công nghệ dữ liệu lớn bùng nổ thì "dữ liệu mở" trở nên quan trọng hơn rất nhiều. Với "dữ liệu mở" mọi thành phần trong xã hội đều có thể tìm kiếm các giá trị từ nguồn dữ liệu này, tạo động lực cho sự sáng tạo, phát triển của bất kỳ cá nhân, tổ chức nào. Có thể lấy một ví dụ về dự án "dữ liệu mở" mà Việt Nam cũng đang tham gia và hưởng lợi đó là dự án Open Transport (Giao thông Mở) của World Bank Transport & ICT Global Practice cho phép chính phủ các nước truy cập miễn phí vào dữ liệu về mật độ giao thông trên đường giúp chính phủ các nước phân tích tình hình giao thông, từ đó đưa ra các chính sách quản lý hợp lý tại các quốc gia đó. Để xây dựng được hạ tầng "Dữ liệu mở" quốc gia, cần thực hiện:

- Xây dựng các chính sách cho phép việc chia sẻ dữ liệu giữa các đơn vị (công và tư)
- Xây dựng nền tảng dữ liệu mở và công cộng, cho phép các đơn vị (công và tư) chia sẻ và truy xuất dữ liệu công cộng này.

Thị trường:

- Tạo ra một ngành công nghiệp về dữ liệu lớn thông qua các khu công nghệ cao hoặc các trung tâm vườn ươm cho các công ty khởi nghiệp.

Cơ sở pháp lý:

- Tạo ra hành lang pháp lý để quản lý và khuyến khích việc phát triển công nghệ dữ liệu lớn, đồng thời đảm bảo sự công bằng, minh bạch trong chia sẻ dữ liệu, bảo vệ dữ liệu và sự riêng tư.

5.2. Đề xuất các ứng dụng dữ liệu lớn

Sự phát triển mạnh mẽ của công nghệ trong những thập kỷ gần đây đã giúp chúng ta đạt được nhiều thành tựu. Việc công nghệ dữ liệu lớn được sử dụng rộng rãi và có ảnh hưởng trong mọi lĩnh vực của đời sống dường là một xu hướng tất yếu. Việc chúng ta có thêm các công cụ mới, các con chip mạnh hơn, phần mềm tốt hơn đã phần nào đẩy nhanh xu hướng đó. Tuy nhiên đó không phải là nguyên nhân trực tiếp dẫn đến xu hướng phổ biến của ứng dụng công nghệ dữ liệu lớn. Lý do sâu xa hơn của những xu hướng này là chúng ta có nhiều dữ liệu hơn rất nhiều. Và lý do chúng ta có nhiều dữ liệu hơn là vì

chúng ta đã đưa nhiều khía cạnh hơn của thực tế vào một định dạng dữ liệu. Việc có nhiều dữ liệu giúp chúng ta khám phá được những giá trị tiền ẩn từ việc phân tích toán học đối với dữ liệu.

Dựa trên việc chúng ta có dữ liệu ra sao sẽ quyết định chúng ta ứng dụng công nghệ dữ liệu lớn như thế nào. Hiện tại, Việt Nam đang có thế mạnh trên một số lĩnh vực cho phép thu thập tương đối đầy đủ dữ liệu để có thể bắt đầu xây dựng các ứng dụng dữ liệu lớn.

5.2.1. Ứng dụng dữ liệu lớn dựa trên các hồ sơ và dữ liệu thống kê.

Ứng dụng công nghệ dữ liệu lớn vào các dữ liệu văn bản đã được nhắc đến trong Chương 5 của đề tài như là một ứng dụng thử nghiệm công nghệ Big Data, cho thấy tầm quan trọng của việc thu thập, lưu trữ và phân tích dữ liệu văn bản.

Tuy nhiên, suy rộng ra, các dữ liệu này không chỉ là các ảnh quét của các văn bản, bên cạnh đó còn có các email, tin nhắn tức thời, các file dữ liệu, file văn bản, tất cả các dữ liệu này dẫn tới tốc độ tăng trưởng của Big Data, việc quản lý và lưu trữ thông tin này - và tăng trưởng của nó - không phải là nhiệm vụ tầm thường.

Các hoạt động nghiệp vụ trong các cơ quan nhà nước phát sinh ra một lượng văn bản giấy khổng lồ, các văn bản này đã và đang tiếp tục được số hóa và lưu trữ tại các cơ quan phát hành ra văn bản. Bên cạnh đó, các email trao đổi công việc, các file văn bản soạn thảo cũng đang được lưu trữ phân tán tại các cơ quan. Ngoài ra, một số ngành đã ứng dụng CNTT trong việc cung cấp dịch vụ công đã tạo ra các CSDL chuyên ngành. Ta có thể liệt kế một số loại dữ liệu điển hình đang được tạo ra và lưu trữ lại như sau:

- **Tại các cơ quan quản lý nhà nước:** Các văn bản quy phạm pháp luật số hóa, email trao đổi, các file văn bản soạn thảo. Các CSDL nằm trong các phần mềm quản lý điều hành

- **Các cơ sở y tế, bệnh viện:** Hồ sơ bệnh án của các bệnh nhân được số hóa, các CSDL về bệnh nhân nằm trong các phần mềm quản lý điều hành, phần mềm quản lý bệnh nhân.

- **Các cơ sở giáo dục:** Hồ sơ học tập của các học sinh, sinh viên, các CSDL về điểm thi của sinh viên.

- **Thuế, Hải quan:** Các báo cáo thuế dạng điện tử, tờ khai hải quan điện tử, CSDL

veeg thuế và hải quan.

- **Các CSDL chuyên ngành khác:** CSDL về công dân, CSDL về tài nguyên, CSDL về môi trường, CSDL về đất đai...

Tất cả các dữ liệu trên một khi được tập hợp, chia sẻ và khai thác (bằng công nghệ dữ liệu lớn) sẽ giúp chúng ta khám phá ra nhiều giá trị bổ ích giúp cải thiện việc quản lý và điều hành xã hội. Ví dụ, việc phân tích các hồ sơ sức khỏe, CSDL về môi trường và hồ sơ quản lý xây dựng có thể cho ta mối tương quan giữa sức khỏe của người dân ở một khu vực nhất định với môi trường ở khu vực đó. Việc này giúp các cơ quan quản lý nhà nước có những hành động phù hợp để điều chỉnh việc quản lý và quy hoạch.

Lợi ích từ việc khai thác các dữ liệu trên bằng công nghệ dữ liệu lớn là rõ ràng và to lớn, tuy nhiên để có thể thực sự triển khai được các ứng dụng thì còn nhiều việc phải làm. Bản thân các dữ liệu đã được thu thập và lưu trữ tuy nhiên lại bị phân tán và rời rạc. Chưa có cơ chế chia sẻ các thông tin này giữa các đơn vị sở hữu dữ liệu, cũng chưa có đơn vị chuyên trách nào đứng ra để tập hợp và xây dựng các ứng dụng dữ liệu lớn phục vụ cho các ngành. Nên nhớ, dữ liệu lớn chỉ thực sự phát huy khi lượng dữ liệu phải đủ lớn.

5.2.2. Ứng dụng dữ liệu lớn dựa trên thông tin vị trí

Ngày nay, dữ liệu thông tin vị trí là một trong những loại dữ liệu được thu thập và lưu trữ tự động một cách phổ biến nhất. Các điện thoại thông minh hiện nay thu thập dữ liệu vị trí và gửi nó trở lại các hãng sản xuất như Apple, Google hay Microsoft một cách tự động mà người dùng không hề biết. Ngoài ra, với sự giảm giá thành mạnh mẽ, các module định vị GPS được dùng rất phổ biến để theo dõi vị trí của các xe máy, xe ô tô từ cá nhân cho đến nhà nước.

Ở Việt Nam, các dữ liệu về vị trí đã và đang được thu thập hàng ngày, tuy nhiên các dữ liệu này chưa được quan tâm và sử dụng đúng để tạo ra giá trị mới, đặc biệt trong lĩnh vực quản lý nhà nước. Ở khu vực tư nhân, đã có nhiều dịch vụ khai thác các dữ liệu này nhằm tạo ra một loại hình kinh doanh mới. Điển hình là các dịch vụ kinh doanh vận tải như GrabTaxi và Uber. Đây là hai dịch vụ rất thành công trong việc cung cấp dịch vụ dựa trên thông tin vị trí. Ngoài việc sử dụng để cung cấp dịch vụ, toàn bộ vị trí (lộ trình của người sử dụng) cũng được thu thập và lưu trữ. Tuy nhiên, các thông tin này chủ yếu được khai thác bởi các tổ chức nước ngoài.

Các công ty cung cấp dịch thông tin di động (viettel, Vinaphone, mobifone...) cũng

đang nắm trong tay một lượng dữ liệu lớn về thông tin vị trí của các khách hàng được xác định nhờ các trạm phát sóng của họ. Tuy nhiên, việc khai thác các thông tin này còn hạn chế và chưa được chia sẻ để tạo ra các ứng dụng dữ liệu lớn hiệu quả.

Nếu các dữ liệu về thông tin vị trí được khai thác hiệu quả thì có thể đem lại nhiều lợi ích to lớn cho việc quản lý của nhà nước và hỗ trợ các doanh nghiệp. Ví dụ việc tích lũy dữ liệu vị trí cho phép phát hiện ùn tắc giao thông mà không cần trông thấy những chiếc xe, nhờ số lượng và tốc độ của các máy điện thoại di chuyển trên một đường tiết lộ thông tin này. Các bản ghi thông tin vị trí địa lý mỗi ngày từ sự di chuyển của hàng triệu thuê bao điện thoại di động có thể tạo ra các báo cáo giao thông thời gian thực các thành phố trên khắp cả nước. Các dữ liệu vị trí cũng có thể cho biết các khu vực của một thành phố có cuộc sống về đêm nhộn nhịp nhất, hoặc để ước tính có bao nhiêu người đã có mặt tại một cuộc biểu tỉnh. Các nghiên cứu cũng chỉ ra rằng, việc phân tích các chuyển động và các mô hình cuộc gọi cũng cho phép xác định những người đã mắc bệnh cúm trước khi bản thân họ biết rằng họ bị bệnh.

Đối với các doanh nghiệp kinh doanh vận tải, dũ liệu định vị cho biết một cách chi tiết về thời gian, địa điểm, và khoảng cách xe chạy thực tế, nó cho phép công ty biết nơi chốn của xe trong trường hợp chậm trễ, để giám sát nhân viên, và theo dõi hành trình của họ để tối ưu hóa các tuyến đường. Việc phân tích các dữ liệu này cũng có thể giúp các công ty tối ưu các tuyến đường để giảm được các chi phí giúp ổn định giá thành vận tải. Đối với cơ quan quản lý nhà nước, các dữ liệu này khi kết hợp với các dữ liệu về tai nạn giao thông cũng giúp đưa ra các dự báo giúp cải thiện tình hình tai nạn giao thông.

5.2.3. Ứng dụng dữ liệu lớn dựa trên thông tin tương tác

Các diễn đàn, mạng xã hội không chỉ đơn giản cung cấp cho chúng ta một cách để tìm và giữ liên lạc với bạn bè và đồng nghiệp, chúng lấy các yếu tố vô hình trong cuộc sống hàng ngày của chúng ta và biến thành dữ liệu có thể được sử dụng để làm những điều mới mẻ. Facebook dữ liệu hóa các mối quan hệ. Twitter giúp dữ liệu hóa cảm xúc bằng cách tạo ra một cách dễ dàng cho người dùng ghi lại và chia sẻ những điều bận tâm của họ.

Việc thu thập và phân tích dữ liệu này có thể đem lại những hiệu quả không ngờ. Thông qua việc phân tích các mạng xã hội, diễn đàn, các trang thông tin điện tử có thể đánh giá được phản ứng của người dân với một chính sách mới đưa ra hay đối với một hiện tượng xã hội.

5.3. Đề xuất nền tảng công nghệ dữ liệu lớn

Với các ưu điểm của mình, Apache Hadoop là trung tâm của các phiên bản mới nhất của các giải pháp Big Data, vì vậy lựa chọn nền tảng Hadoop để xây dựng các ứng dụng dữ liệu lớn là một lựa chọn phù hợp.

Nền tảng Big Data phân loại theo hướng mà người dùng tiếp cận với Hadoop. Một số doanh nghiệp cung cấp dịch vụ Big Data tích hợp với một phiên bản của Hadoop, trong khi các doanh nghiệp khác cung cấp Hadoop kết nối với hệ thống phân tích dữ liệu đã có sẵn. Cách thức thứ hai này thường bao gồm việc xử lý song song dữ liệu đã tạo nên thương hiệu của doanh nghiệp trong lĩnh vực Big Data từ trước khi Hadoop xuất hiện: Vertica and Aster Data. Sức mạng của Hadoop trong trường hợp này chính là xử lý dữ liệu phi cấu trúc song song với khả năng phân tích cơ sở dữ liệu hiện có, bao gồm cả dữ liệu có cấu trúc và phi cấu trúc.

Thực tế việc triển khai Big Data không hẳn là chỉ gói gọn trong hai thể loại dữ liệu cấu trúc và phi cấu trúc. Ta sẽ luôn tìm thấy sự tồn tại của Hadoop như là một phần của hệ thống hoạt động với cơ sở dữ liệu quan hệ hoặc MPP.

Cũng giống như Linux, không có giải pháp Hadoop nào chỉ sử dụng mã nguồn của Apache Hadoop. Thay vào đó, nó được đóng gói thành các bản phân phối. Ở mức tối thiểu, những bản phân phối này đã trải qua một quá trình thử nghiệm, và thường bao gồm những thành phần bổ sung như các công cụ quản lý và giám sát. Bản phân phối thường được dùng nhất hiện nay là của Cloudera, Hortonworks and MapR. Không phải mọi bản phân phối đều sẽ thương mại hóa, dự án BigTop mục tiêu nhằm tạo ra một bản phân phối Hadoop dưới sự bảo trợ của Apache.

5.3.1. Những hệ thống tích hợp Hadoop

Các nhà cung cấp phần mềm Hadoop hàng đầu đã liên kết các sản phẩm Hadoop của họ với phần còn lại của cơ sở dữ liệu và các dịch vụ phân tích. Những nhà cung cấp này không yêu cầu khách hàng phải tìm kiếm Hadoop từ bên thứ ba, mà cung cấp Hadoop như một phần cốt lõi của giải pháp Big Data, tăng cường bởi các công cụ phân tích và workflow.

EMC Greenplum

- Database : Greenplum Database
- Deployment options :
    - Appliance: Modular Data Computing Appliance
    - Software: Enterprise Linux)
- Hadoop : Bundled distribution (Greenplum HD); Hive, Pig, Zookeeper, Hbase
- NoSQL component : Hbase

Được mua lại bởi EMC, và ngay lập tức trở thành trung tâm chiến lược của công ty, Greenplum là công ty tương đối mới, so với các công ty khác trong lĩnh vực này. Họ đã biến đó thành lợi thế trong việc tạo ra một nền tảng phân tích, với đội ngũ nghiên cứu khoa học linh hoạt.

Greenplum Unified Analytics Platform (UAP) bao gồm ba yếu tố: cơ sở dữ liệu Greenplum MPP, cho dữ liệu có cấu trúc; một bản phân phối Hadoop, Greenplum HD; một lớp làm việc nhóm hiệu suất dành cho đội ngũ khoa học dữ liệu.

Greenplum HD xây dựng dựa trên bản phân phối Hadoop tương thích với MapR, nhằm thay thế hệ thống tập tin, tăng tốc độ triển khai và tốc độ cung cấp dữ liệu, và những tính năng bền vững khác. Khả năng tương tác giữa HD và cơ sở dữ liệu Greenplum MPP cho phép một truy vấn có thể đồng thời truy cập cả hai cơ sở dữ liệu và dữ liệu Hadoop.

Chorus là một tính năng độc đáo, cho thấy sự quan tâm của Greenplum đến ý tưởng của khoa học dữ liệu và tầm quan trọng của đội ngũ nghiên cứu linh hoạt đến việc khai thác hiệu quả Big Data. Nó hỗ trợ nhiều vai trò trong tổ chức, từ các nhà phân tích, các nhà khoa học dữ liệu và các DBA nhằm điều hành kinh doanh các bên liên quan.

Nhằm đảm bảo vai trò của EMC trong thị trường Data Center, Greenplum UAP được cung cấp như một module có sẵn trong các cấu hình.

IBM

- Database : DB2
- Deployment options : Software (Enterprise Linux), Cloud
- Hadoop : Bundled distribution (InfoSphere BigInsights); Hive, Oozie, Pig, Zoo-

keeper, Avro, Flume, HBase, Lucene

- NoSQL component : Hbase

InfoSphere BigInsights là bản phân phối Hadoop của IBM, và là một phần của bộ sản phẩm được cung cấp dưới quản lý thông tin thương hiệu "InfoSphere". Mọi vấn đề Big Data tại IBM nhằm nhấn mạnh vào "Big", khiến IBM được biết đến với tên gọi "Big Blue".

BigInsights bổ sung cho Hadoop với một loạt tính năng, bao gồm các công cụ quản lý và quản trị. Nó cũng cung cấp các công cụ phân tích văn bản nhằm nhận dạng các thực thể: xác định người, địa chỉ, số điện thoại, … .

Ngôn ngữ truy vấn Jaql của IBM cung cấp sự kết nối giữa Hadoop và các sản phẩm IBM khác, như là cơ sở dữ liệu quan hệ hoặc kho dữ liệu Netezza.

InfoSphere BigInsights tương thích với cơ sở dữ liệu và các sản phẩm kho dữ liệu khác của IBMm bao gồm DB2, Netezza và kho dữ liệu của bản thân InfoSphere và các dòng phân tích. Để hỗ trợ thăm dò phân tích, BigInsights cung cấp với BigSheets, một công cụ trình diễn Big Data dạng bảng dữ liệu.

IBM đánh địa chỉ luồng dữ liệu Big Data thông qua InfoSphere. BigInsights tạm thời chưa được cung cấp dưới dạng thiết bị nhưng có thể được sử dụng trong đám mây qua Rightscale, Amazon, Rackspace, và đám mây IBM Smart Enterprise.

Microsoft

- Database : SQL Server
- Deployment options : Software (Windows Server), Cloud (Windows Azure Cloud)
- Hadoop : Bundled distribution (Big Data Solution); Hive, Pig

Microsoft đã sử dụng Hadoop như là nhân tố chính của dịch vụ Big Data, và đang theo đuổi một cách tiếp cận nhằm làm tăng khả năng đáp ứng của Big Data thông qua bộ công cụ phân tích, bao gồm những công cụ quen thuộc của Excel và Power Pivot.

Giải pháp Big Data của Microsoft mang Hadoop đến với nền tảng của Windows Server và nền tảng của đám mây Windows Azure. Microsoft đã đóng gói thành bản phân phối Hadoop của riêng họ, tích hợp với Window System Center và Active Directory. Họ dự định đóng góp những phản hồi với Apache Hadoop nhằm đảm bảo rằng một phiên bản

mã nguồn mở của Hadoop sẽ chạy được trên Windows.

Về phía server, Microsoft cung cấp sự tích hợp với cơ sở dữ liệu SQL Server và sản phẩm kho dữ liệu. Tuy nhiên không bắt buộc phải sử dụng các giải pháp kho dữ liệu của Microsoft. Kho dữ liệu Hadoop Hive là một phần của giải pháp Big Data, bao gồm kết nối từ Hive đến ODBC và Excel.

Microsoft tập trung đội ngũ lập trình tạo ra JavaScript API cho Hadoop. Sử dụng JavaScript, các lập trình viên có thể tạo ra các Hadoop jobs cho MapReduce, Pig hoặc Hive thậm chí từ môi trường dựa trên trình duyệt. Microsoft đồng thời cung cấp Visual Studio và .NET tích hợp với Hadoop.

Việc triển khai có thể dựa trên hệ thống máy chủ hoặc trong đám mây hoặc một sự kết hợp hybrid.

Oracle

- Deployment options : Appliance (Oracle Big Data Appliance)
- Hadoop : Bundled distribution (Cloudera's Distribution including Apache Hadoop); Hive, Oozie, Pig, Zookeeper, Avro, Flume, HBase, Sqoop, Mahout, Whirr
- NoSQL component : Oracle NoSQL Database

Tham gia vào thị trường Big Data vào cuối năm 2011, Oracle tiếp cận theo hướng dựa trên công cụ. Bộ công cụ Big Data của Oracle tích hợp Hadoop, công cụ phân tích R, bộ CSDL Oracle NoSQL, và một kết nối với CSDL của Oracle và dòng sản phẩm kho dữ liệu Exadata.

Cách tiếp cận của Oracle hướng tới thị trường doanh nghiệp cao cấp, và đặc biệt chú trọng đến việc triển khai nhanh chóng, hiệu suất cao. Đây là nhà cung cấp duy nhất tích hợp ngôn ngữ phân tích phổ biến R với Hadoop, và vận hành CSDL NoSQL theo một lối riêng trái ngược với Hadoop Hbase.

Thay vì phát triển bản phân phối Hadoop riêng, Oracle đã hợp tác với Cloudera để nhận được sự hỗ trợ về Hadoop, mang lại cho Oracle một giải pháp Hadoop hoàn thiện và đã được chứng nhận. Kết nối CSDL một lần nữa thúc đẩy sự tích hợp của dữ liệu có cấu trúc Oracle với các dữ liệu phi cấu trúc được lưu trữ trong Hadoop HDFS.

CSDL NoSQL của Oracle là một CSDL dạng key-value có thể mở rộng, xây dựng dựa

trên công nghệ Berkeley DB (Mike Olson, CEO của Cloudera đã từng là giám đốc điều hành của SleepyCat, công ty sáng tạo nên Berkeley DB). Oracle định vị CSDL NoSQL của mình như là một phương tiện trong việc thu thập dữ liệu Big Data trước khi phân tích.

Sản phẩm Oracle R Enterprise cung cấp tích hợp trực tiếp vào CSDL Oracle, cũng như Hadoop, cho phép các đoạn script R thực thi trên dữ liệu mà không cần phải lấy nó ra khỏi kho dữ liệu.

Phân tích CSDL với kết nối Hadoop

Xử lý song song khối lượng lớn (MPP) dữ liệu được dành cho xử lý dữ liệu Big Data có cấu trúc, còn với Hadoop là dữ liệu phi cấu trúc. Cùng với Greenplum, Aster Data và Vertica là hai sản phẩm Big Data đi đầu trước thời kỳ của Hadoop.

Những giải pháp MPP này chủ yếu xử lý dữ liệu nhằm phân tích khối lượng công việc và tích hợp dữ liệu, và cung cấp kết nối đến Hadoop và kho dữ liệu. Trong các thương vụ mua lại gần đây cho thấy các sản phẩm này trở thành bộ công cụ phân tích của các nhà cung cấp dịch vụ kho và lưu trữ dữ liệu: Teradata mua lại Aster Data, EMC mua lại Greenplum, và HP mua lại Vertica.

So sánh một số giải pháp:

	Aster Data	ParAccel	Vertica
Database	CSDL phân tích MPP	CSDL phân tích MPP	CSDL phân tích MPP
Deployment options	-Appliance: Aster MapReduce Appliance -Software: Enterprise Linux -Cloud: Amazon EC2, Terremark and Dell Clouds	-Software: Enterprise Linux -Cloud: Cloud Edtion	-Appliance: HP Vertica Appliance -Software: Enterprise Linux -Cloud: Cloud and Virtualized
Hadoop	Cho phép kết nối đến Hadoop	Cho phép tích hợp Hadoop	Cho phép kết nối đến Hadoop và Pig

5.3.2. Các công ty Hadoop

Trực tiếp sử dụng Hadoop là một con đường để tạo ra giải pháp Big Data, đặc biệt khi mà cơ sở hạ tầng của bạn chưa đáp ứng được với dòng sản phẩm của các nhà cung

cấp lớn. Thực tế, mọi CSDL hiện tại đều có tính năng kết nối với Hadoop, và có rất nhiều bản phân phối Hadoop để lựa chọn.

Căn cứ vào đặc tính định hướng người phát triển của thế giới Big Data, các bản phân phối của Hadoop thường xuyên được cung cấp dưới dạng một phiên bản chỉnh sửa công cộng. Những bản chỉnh sửa này thiếu chức năng quản trị, nhưng chứa toàn bộ những chức năng cần thiết cho việc đánh giá và phát triển.

Phiên bản phân phối đầu tiên của Hadoop đến từ Cloudera và IBM, tập trung vào khả năng sử dụng và quản lý. Phiên bản này bổ sung các cải tiến định hướng hiệu suất cho Hadoop, giống như MapR và Platform Computing. Trong khi duy trì khả năng tương thích API, những nhà cung cấp này đã thay thế những thành phần gây trì trệ hoặc những điểm yếu trong bản phân phối của Apache bằng những thành phần có hiệu suất tốt hơn và xử lý mạnh hơn.

Cloudera

Là nhà cung cấp các phiên bản Hadoop lâu năm nhất, Cloudera cung cấp cho một giải pháp Hadoop dành cho các công ty, cùng với các dịch vụ, đào tạo và hỗ trợ. Cùng với Yahoo, Cloudera đã có những đóng góp đáng kể đối với Hadoop, và thông qua nhiều cuộc hội nghị ngành công nghiệp để nâng tầm Hadoop đến vị trí hiện nay.

Hortonworks

Mặc dù mới đặt chân vào thị trường, Hortonworks vẫn có một lịch sử lâu dài với Hadoop. Tách ra từ Yahoo, nơi khởi nguồn của Hadoop, Hortonworks gắn bó chặt chẽ và thúc đẩy công nghệ cốt lõi Apache Hadoop. Hortonworks cũng là một đối tác của Microsoft nhằm hỗ trợ và thúc đẩy quá trình tích hợp Hadoop.

5.3.3. Đánh giá một số bản phân phối Hadoop:

	Cloudera	**EMC Greenplum**	**Hortonworks**	**IBM**
Tên sản phẩm	Cloudera's Distri-bution including Apache Hadoop	Greenplum HD	Hortonworks Data Platform	InfoSphere BigIn-sights
Bản miễn phí	CDH (Đã tích hợp và kiểm thử bản phân phối của Apache Hadoop)	Community Edition (chứng nhận 100% mã nguồn mở, hỗ trợ bản phân phối của Apache Hadoop stack)		Basic Edition (tích hợp một bản phân phối của Hadoop)
Bản thương mại	Cloudera Enterprise (Bổ sung thêm lớp phần mềm quản lý nằm trên CDH)	Enterprise Edition (Tích hợp phiên bản MapR's M5 tương thích với Hadoop, thay thế cho MapR's C++ dựa trên hệ thống tập tin; bổ sung thêm công cụ quản lý MapR)		Enterprise Edition (Bản phân phối Hadoop, bổ sung giao diện bảng tính BigSheets, lập lịch, phân tích văn bản, đánh chỉ mục, kết nối JDBC, hỗ trợ bảo mật)
Thành phần Hadoop	Hive, Oozie, Pig, Zookeeper, Avro, Flume, HBase, Sqoop, Mahout, Whirr	Hive, Pig, Zookeeper, HBase	Hive, Pig, Zoo-keeper, HBase, None, Ambari	Hive, Oozie, Pig, Zoo-keeper, Avro, Flume, HBase, Lucene
Bảo mật	- Cloudera Manager - Kerberos			Chức năng bảo mật: xác thực LDAP , ủy quyền dựa theo vai trò, đảo ngược proxy
Giao diện quản trị	- Cloudera Manager	- Giao diện quản trị	- Apache Ambari	- Giao diện quản trị

	- Quản lý tập trung và cảnh báo	- Bộ cung cụ quản trị MapR Heatmap cluster	- Giám sát, quản trị, quản lý vòng đời cho các Hadoop cluster	- Chức năng quản trị bao gồm quản trị Hadoop HDFS và MapReduce, quản trị cluster và server, xem nội dung tập tin HDFS
Quản lý Job	-Cloudera Manager - Phân tích, giám sát Job, tìm kiếm log	- Công cụ quản lý Job có tính sẵn sàng đáp ứng cao - JobTracker HA và hệ phân tán NameNode HA ngăn ngừa khả năng thất lạc Job, khởi chạy lại và dự phòng sự cố	- Apache Ambari - Giám sát, quản trị, quản lý vòng đời cho các Hadoop cluster	- Chức năng quản lý Job - Khởi tạo Job, thay thế, hủy, kiểm tra trạng thái, ghi log
Kết nối CSDL		Greenplum Database		DB2, Netezza, Info-Sphere Warehouse
Tính năng tương tác				
Truy cập HDFS	- Fuse-DFS - Mount HDFS tương tự hệ thống tập tin truyền thống	- NFS - Truy cập HDFS tương tự hệ thống tập tin mạng	- WebHDFS - Sử dụng REST API để truy cập HDFS	
Cài đặt	Cài đặt dựa trên trình cài đặt			- Cài đặt nhanh - Công cụ cài đặt hướng giao diện
APIs bổ sung				Jaql (Ngôn ngữ truy vấn chức năng, khai báo được thiết kế để xử lý bộ dữ liệu lớn)

	MapR	Microsoft	Platform Computing
Tên sản phẩm	MapR	Big Data Solution	Platform MapReduce
Bản miễn phí	MapR M3 Edition (bản phân phối miễn phí kết hợp với cải tiến hiệu năng MapR)		Platform MapReduce Developer Edition (Phiên bản không bao gồm chức năng quản lý tài nguyên)
Bản thương mại	MapR M5 Edition (Tăng cường bản M3 với tính đáp ứng cao và tính năng bảo vệ dữ liệu)	Big Data Solution (phiên bản Hadoop trên Windows, tích hợp CSDL của Microsoft và các sản phẩm phân tích)	Platform MapReduce (nâng cao khả năng thực thi của Hadoop MapReduce, cung cấp API tương thích với Apache Hadoop)
Thành phần Hadoop	Hive, Pig, Flume, HBase, Sqoop, Mahout, None, Oozie	Hive, Pig	
Bảo mật		tích hợp Active Directory	
Giao diện quản trị	- Giao diện quản trị - Công cụ quản trị MapR Heatmap cluster	- Tích hợp System Center	- Giao diện quản trị - Quán lý Platform MapReduce Workload
Quản lý Job	- Công cụ quản lý Job có tính sẵn sàng đáp ứng cao - JobTracker HA và hệ phân tán NameNode HA ngăn ngừa khả năng thất lạc Job, khởi chạy lại và dự phòng sự cố		
Kết nối CSDL		SQL Server, SQL Server Parallel Data Warehouse	
Tính năng tương tác		Hive ODBC Driver, Excel Hive Add-in	
Truy cập HDFS	- NFS - Truy cập HDFS tương tự hệ thống tập tin mạng		
Cài đặt			
APIs bổ sung	REST API	JavaScript API	Bao gồm R, C/C+

			+,C#, Java, Python
Quản lý phiên bản	Mirroring, snapshots		

KẾT LUẬN

Những lợi ích mà dữ liệu lớn đang đem lại cho chúng ta là không thể chối cãi, Những ảnh hưởng của nó là khá lớn và rõ ràng trên mọi mặt của đời sống xã hội. Vì vậy, hiểu và nắm bắt được đúng các khái niệm về dữ liệu lớn là quan trọng và cần thiết để có thể nhanh chóng đưa dữ liệu lớn vào ứng dụng trong các công việc hàng ngày, nâng cao hiệu quả công việc, năng suất lao động. Có thể nói dữ liệu lớn sẽ là chìa khóa để thúc đẩy nền kinh tế phát triển trong giai đoạn tiếp theo và là đòn bẩy để giúp chúng ta thoát khỏi bẫy thu nhập trung bình để vươn lên và đuổi kịp sự phát triển của thế giới.

Với mục tiêu bước đầu nắm bắt được chính xác và đầy đủ về dữ liệu lớn, đồng thời làm chủ được các công nghệ nền tảng trong việc xử lý dữ liệu lớn để có thể xây dựng được các ứng dụng thực tế. Nhóm nghiên cứu đã thực hiện các nghiên cứu cơ bản về dữ liệu lớn, các xu hướng và ảnh hưởng của nó lên đời sống kinh tế xã hội, qua đó giúp nhóm thực hiện đề tài hiểu và nắm rõ được các khái niệm và xây dựng được các bài toàn dữ liệu lớn phù hợp. Nghiên cứu này cũng có thể là tài liệu tham khảo tốt cho các cá nhân, đơn vị bắt đầu tìm hiểu về dữ liệu lớn.

Bên cạnh các nghiên cứu về lý thuyết, thì một trong những mục tiêu đạt được của đề tài là đã xây dựng được ứng dụng thử nghiệm dựa trên các nghiên cứu đã thực hiện trong đề tài. Ứng dụng này có tính thực tiễn cao, gắn liền với các hoạt động khác của Viện CNPM & NDS. Với các nghiên cứu này, chắc chắn rằng sau khi đề tài kết thúc, nhóm sẽ tiếp tục hoàn thiện ứng dụng xử lý và tìm kiếm ảnh văn bản trên công nghệ dữ liệu lớn để tích hợp vào kho dữ liệu giúp nâng cao được hiệu quả khai thác của kho dữ liệu cho các đơn vị đã triển khai. Hoàn thiện ứng dụng này cũng sẽ tạo tiền đề để nhóm có thể tiếp tục nghiên cứu và xây dựng các ứng dụng phân tích dữ liệu khác để tận dụng nguồn dữ liệu văn bản số hóa một cách tối đa và hiệu quả.

TÀI LIỆU THAM KHẢO

1. Per-Erik Danielsson (1980), "Euclidean distance", Computer Graphics and Image Processing, volume 14, 1980, 11, 227-248.
2. Jonathan J. Hull, Siamak Khoubyari, and Tin Kam Ho (1992), Word Image Matching in a Methodology for Degraded Text Recognition, University of New York.
3. Fatima EL Jamiy, Abderrahmane Daif, Mohamed Azouazi and Abdelaziz Marzak (2014), "The potential and challenges of Big data - Recommendation systems next level application", IJCSI International Journal of Computer Science Issues, Volume 11, 2014 .
4. Jimmy Lin and Chris Dyer (2010), Data-Intensive Text Processing with MapReduce, University of Maryland
5. R. Manmatha, Chengfen Han, E.M. Risenman, and W.B. Croft (1996), "Indexing Handwriting Using Word Matching", Proceeding of Digital Libraries'96, 1st ACM International Conference on Digital Libraries, 1996, pp. 151-159.
6. Allen Wittenauer, Deploying Grid Services Using Hadoop, ApacheCon EU 2008, April 2008.
7. Kenneth Cukier and Viktor Mayer-Schönberger, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Hachette UK, Mar 14, 2013.
8. Vincenzo Morabito, Big Data and Analytics Strategic and Organizational Impacts, Springer
9. Edward J. Yoon, An Introduction to Bulk Synchronization Parallel on Hadoop, HUG, Korea, December 2009.
10. Hadoop wiki (http://wiki.apache.org/hadoop/)
11. http://www.ibm.com/developerworks/vn/library/data/2013Q1/dm-1209hadoopbigdata/
12. *https://hbr.org/2015/06/inventory-management-in-the-age-of-big-data*
13. https://en.wikipedia.org/wiki/Apache_Spark

14. http://techkites.blogspot.com/2015/02/implementing-real-time-trending-engine.html

15. http://blog.brakmic.com/data-science-for-losers-part-3-scala-apache-spark/

16. https://www.linkedin.com/pulse/apache-spark-next-big-data-thing-navdeep-singh-gill

17. http://www.infoq.com/articles/apache-spark-introduction