ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

LUẬN VĂN THẠC SĨ

TÊN ĐỀ TÀI

NGHIÊN CỨU MỘT SỐ VẤN ĐỀ VỀ BIG DATA VÀ ỨNG DỤNG TRONG PHÂN TÍCH KINH DOANH

Giáo viên hướng dẫn : **GS.TS Vũ Đức Thi**
Học viên thực hiện : **Phạm Việt Anh**
Lớp : **CK16H**

Thái Nguyên, tháng 1 năm 2019

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

LUẬN VĂN THẠC SĨ

Tên đề tài ***Nghiên cứu một số vấn đề về Big Data và ứng dụng trong phân tích kinh doanh***

Giáo viên hướng dẫn : **GS.TS Vũ Đức Thi**
Học viên thực hiện : **Phạm Việt Anh**
Lớp : **CK16H**

Thái Nguyên, tháng 1 năm 2019

CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ BIG DATA VÀ QUY TRÌNH PHÂN TÍCH DỮ LIỆU LỚN

1.1 Giới thiệu tổng quan về Big Data

Trong 22 năm qua, dữ liệu đã tăng lên với một quy mô lớn trong các lĩnh vực khác nhau. Theo một báo cáo từ Tập đoàn Dữ liệu Quốc tế (IDC), trong năm 2011 dung lượng dữ liệu được tạo ra và sao chép trên toàn thế giới là 1.8ZB, tăng gần chín lần trong năm năm [1]. Con số này sẽ không dừng lại ở đó mà sẽ tăng gấp đôi ít nhất hai năm một lần trong tương lai gần.

Dưới sự phát triển mạnh mẽ của CNTT và sự gia tăng một cách bùng nổ của dữ liệu toàn cầu, thuật ngữ Big Data đã trở nên quen thuộc và thường được dùng để mô tả các hệ thống dữ liệu lớn. So với các tập dữ liệu truyền thống trước đây, dữ liệu lớn thường bao gồm các khối dữ liệu phi cấu trúc cần thêm phân tích trong thời gian thực. Ngoài ra, dữ liệu lớn cũng mang lại những cơ hội mới để khám phá giá trị mới, giúp chúng ta có được một sự hiểu biết một cách sâu sắc về các giá trị tiềm ẩn, cũng như những thách thức mới. Ví dụ là làm thế nào để tổ chức và quản lý các tập dữ liệu như vậy một cách hiệu quả.

Trong những năm trở lại đây, nhiều ngành công nghiệp đang trở nên quan tâm đến tiềm năng to lớn của Big Data, nhiều cơ quan chính phủ đã công bố kế hoạch lớn trong việc phát triển nghiên cứu và ứng dụng Big Data [2]. Không chỉ vậy, các vấn đề liên quan tới Big Data cũng luôn được nhắc đến trên các phương tiện truyền thông công cộng, chẳng hạn như Economist [3][4], New York Times [5] và Nation Public Radio [6][7]. Hai tạp chí khoa học đầu ngành là Nature và Science cũng đã mở mục riêng để thảo luận về những thách thức và các tác động của Big Data [8][9]. Tới đây, có thể nói rằng kỷ nguyên của Big Data đã đến [10].

Ngày nay, Big Data có liên quan đến dịch vụ của các công ty về Internet đều phát triển nhanh chóng. Ví dụ, Google xử lý dữ liệu khoảng hàng trăm Petabyte (PB), Facebook đã tạo khoảng hơn 10 PB dữ liệu log mỗi tháng, Taobao một công ty con của Alibaba tạo ra hàng chục Terabyte (TB) dữ liệu về giao dịch trực tuyến mỗi ngày.

1.1.1 Những định nghĩa và đặc trưng của Big Data

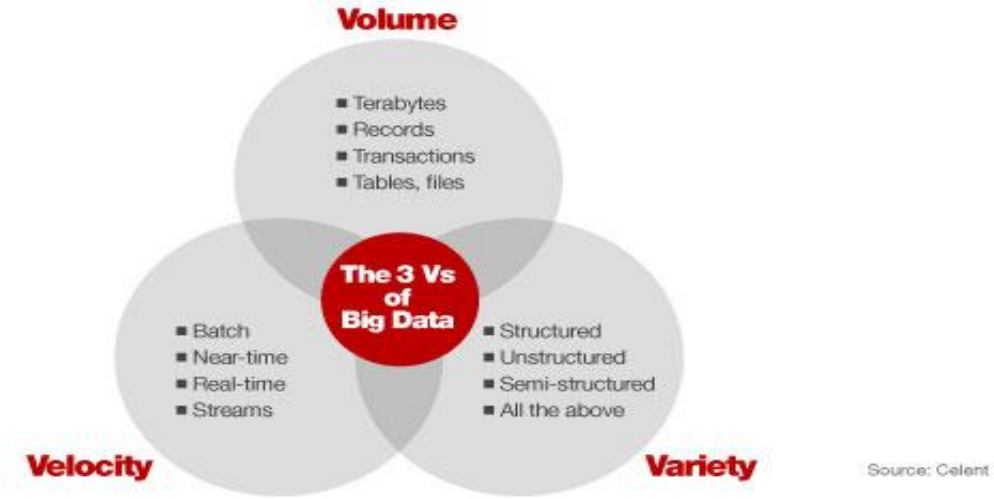
Big Data là một khái niệm trừu tượng và có rất nhiều định nghĩa về Big Data. Ngay như tên gọi là dữ liệu lớn hay dữ liệu khổng lồ thì nó còn có một số đặc trưng

khác trong đó xác định sự khác biệt giữa nó và "dữ liệu lớn" hay "dữ liệu rất lớn".

Hiện nay, mặc dù tầm quan trọng của Big Data đã được thừa nhận rộng rãi, nhưng vẫn có nhiều những ý kiến về định nghĩa của nó. Một cách tổng quát có thể định nghĩa rằng Big Data có nghĩa là các bộ dữ liệu không thể được nhận diện, thu hồi, quản lý và xử lý bằng CNTT truyền thống và các công cụ phần mềm/ phần cứng trong một khoảng thời gian có thể chấp nhận được. Phát sinh từ nhiều sự quan tâm, các doanh nghiệp khoa học và công nghệ, các nhà nghiên cứu, các nhà phân tích dữ liệu và các kỹ thuật viên có những định nghĩa khác nhau về Big Data. Sau đây là một số định nghĩa về Big Data mang tới một sự hiểu biết tốt hơn về những ý nghĩa xã hội, kinh tế và công nghệ rộng lớn của Big Data.

Năm 2010, Apache Hadoop định nghĩa dữ liệu lớn như "bộ dữ liệu mà không thể thu thập, quản lý và xử lý bởi các máy tính nói chung trong một phạm vi chấp nhận được". Cũng trên cơ sở đó, vào tháng 5 năm 2011, McKinsey & Company, một công ty tư vấn toàn cầu công bố Big Data như một địa hạt mới cho sự đổi mới, cạnh tranh và hiệu suất. Big Data có nghĩa là những bộ dữ liệu mà không có thể được thu lại, lưu trữ và quản lý bởi phần mềm cơ sở dữ liệu cổ điển. Định nghĩa này gồm hai ý nghĩa: Thứ nhất, dung lượng của các tập dữ liệu mà phù hợp với tiêu chuẩn Big Data đang thay đổi và có thể tăng trưởng theo thời gian hoặc với những tiến bộ công nghệ. Thứ hai, dung lượng của các tập dữ liệu mà phù hợp với tiêu chuẩn của Big Data trong các ứng dụng khác nhau trong mỗi ứng dụng. Hiện nay, Big Data thường từ vài TB đến vài PB [10]. Từ định nghĩa của McKinsey & Company, có thể thấy rằng dung lượng của một tập dữ liệu không phải là tiêu chí duy nhất cho Big Data. Quy mô dữ liệu ngày càng phát triển và việc quản lý nó mà không thể xử lý bằng công nghệ cơ sử dữ liệu truyền thống là hai đăng trưng quan trọng tiếp theo.

Dữ liệu lớn đã được định nghĩa từ sớm những năm 2001. Doug Laney, một nhà phân tích của META (nay có tên là công ty nghiên cứu Gartner) định nghĩa những thách thức và cơ hội mang lại của sự tăng trưởng dữ liệu với một mô hình "3Vs", tức là sự gia tăng của dung lượng, tốc độ và tính đa dạng trong một báo cáo nghiên cứu [11]. Mặc dù, mô hình này ban đầu không được sử dụng để xác định Big Data, tuy nhiên Gatrtner cùng nhiều doanh nghiệp khác bao gồm cả IBM và một số cơ sở nghiên cứu của Microsoft vẫn còn sử dụng mô hình "3Vs" để mô tả về dữ liệu lớn trong vòng 10 năm tiếp theo.



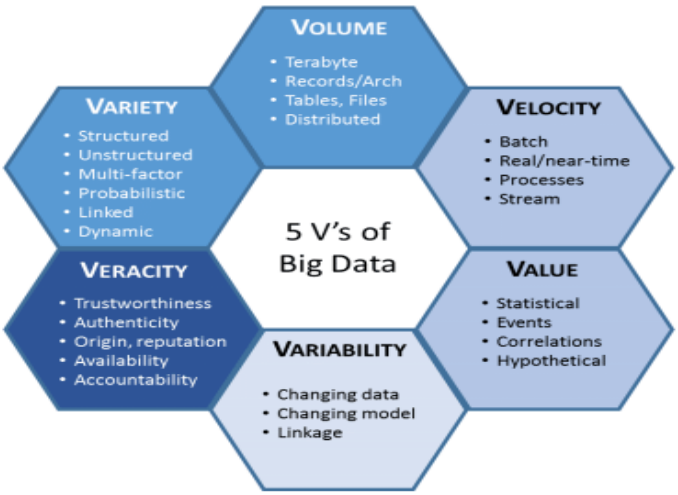
Hình 1.1: Mô hình 3Vs của Big Data

Mô hình "3Vs" được giải thích như sau:

- Dung lượng (Volume): Sự sản sinh và thu thập các dữ liệu lớn, quy mô dữ liệu trở nên ngày càng lớn.
- Tốc độ (Velocity): Tính kịp thời của dữ liệu lớn, cụ thể là việc thu thập và phân tích dữ liệu phải được tiến hành nhanh chóng và kịp thời để sử dụng một cách tối đa các giá trị thương mại của Big Data.
- Tính đa dạng (Variety): Các loại dữ liệu khác nhau bao gồm dữ liệu bán cấu trúc và phi cấu trúc như âm thanh, video, web, văn bản,…cũng như dữ liệu có cấu trúc truyền thống.

Đến năm 2011, định nghĩa về Big Data đã có sự thay đổi khi một báo cáo của IDC đã đưa ra một định nghĩa như sau: "Công nghệ Big Data mô tả một thế hệ mới của những công nghệ và kiến trúc, được thiết kế để lấy ra giá trị kinh tế từ dung lượng rất lớn của một loạt các dữ liệu bằng cách cho phép tốc độ cao trong việc thu thập, khám phá hoặc phân tích" [1]. Với định nghĩa này, dữ liệu lớn mang trong mình bốn đặc trưng và được hiểu như một mô hình "4Vs".

Năm 2014, Gartner lại đưa ra một khái niệm mới về Big Data qua mô hình "5Vs" với năm tính chất quan trọng của Big Data.



Hình 1.2: Mô hình 5vs của Big Data

Mô hình "5Vs" được giải thích như sau:

- Khối lượng (Volume): Sự sản sinh và thu thập các dữ liệu lớn, quy mô dữ liệu trở nên ngày càng lớn.

- Tốc độ (Velocity): Tính kịp thời của dữ liệu lớn, cụ thể là việc thu thập và phân tích dữ liệu phải được tiến hành nhanh chóng và kịp thời để sử dụng một cách tối đa các giá trị thương mại của Big Data.

- Tính đa dạng (Variety): Các loại dữ liệu khác nhau bao gồm dữ liệu bán cấu trúc và phi cấu trúc như âm thanh, video, web, văn bản,…cũng như dữ liệu có cấu trúc truyền thống.

- Tính chính xác (Veracity): Tính hỗn độn hoặc 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. Khối lượng dữ liệu lớn sẽ đi kèm với tính xác thực của dữ liệu.

- Giá trị (Value): Đây được coi là đặc điểm quan trọng nhất của dữ liệu lớn. Việc tiếp cận dữ liệu lớn sẽ không có ý nghĩa nếu không được chuyển thành những thứ có giá trị. Giá trị của dữ liệu là đặc điểm quan trọng nhất trong mô hình "5Vs" của Big Data.

Ngoài ra, Viện tiêu chuẩn và kỹ thuật quốc gia của Hoa Kỳ (NIST) định nghĩa "Dữ liệu lớn có nghĩa là các dữ liệu mà dung lượng dữ liệu, tốc độ thu thập hoặc biểu diễn dữ liệu hạn chế khả năng của việc sử dụng các phương pháp quan hệ truyền thống để tiến hành phân tích hiệu quả hoặc các dữ liệu mà có thể được xử lý một cách hiệu quả với các công nghệ". Định nghĩa này tập trung vào các khía cạnh công nghệ của

Big Data. Nó chỉ ra rằng phương pháp hay công nghệ hiệu quả cần phải được phát triển và được sử dụng để phân tích và xử lý dữ liệu lớn.

1.1.2 Sự phát triển của Big Data

Cuối những năm 1970, khái niệm "máy cơ sở dữ liệu" nổi lên, đó là một công nghệ đặc biệt sử dụng cho việc lưu trữ và phân tích dữ liệu. Với sự gia tăng của dung lượng dữ liệu, khả năng lưu trữ và xử lý của một hệ thống máy tính lớn duy nhất trở nên không đủ. Trong những năm 1980, hệ thống "không chia sẻ"- một hệ thống cơ sở dữ liệu song song được đề xuất để đáp ứng nhu cầu của dung lượng dữ liệu ngày càng tăng [12]. Kiến trúc hệ thống không chia sẻ được dựa trên việc sử dụng các cụm và mỗi máy có riêng bộ xử lý, lưu trữ và đĩa cứng. Hệ thống Teradata là hệ thống cơ sở dữ liệu song song thương mại thành công đầu tiên. Ngày 2 tháng 6 năm 1986, một sự kiện bước ngoặt xảy ra khi Teradata giao hệ thống cơ sở dữ liệu song song đầu tiên với dung lượng lưu trữ 1TB cho Kmart để giúp các công ty bán lẻ quy mô lớn tại Bắc Mỹ mở rộng kho dữ liệu [13]. Trong những năm 1990, những ưu điểm của cơ sở dữ liệu song song đã được công nhận rộng rãi trong lĩnh vực cơ sở dữ liệu.

Tuy nhiên, Big Data vẫn còn nhiều thách thức phát sinh. Với sự phát triển của dịch vụ Internet, các nội dung chỉ mục và truy vấn đã được phát triển nhanh chóng. Do đó, công cụ tìm kiếm của các công ty đều phải đối mặt với những thách thức của việc xử lý dữ liệu lớn. Google tạo ra mô hình lập trình GFS [14] và MapReduce [15] để đối phó với những thách thức mang lại về việc quản lý và phân tích dữ liệu ở quy mô Internet. Ngoài ra, nội dung được sinh ra bởi người sử dụng, cảm biến và các nguồn dữ liệu phổ biến khác cũng tăng, do đó yêu cầu một sự thay đổi cơ bản về kiến trúc tính toán và cơ chế xử lý dữ liệu quy mô lớn.

Vào tháng 1 năm 2007, Jim Gray là một nhà tiên phong về phần mềm cơ sở dữ liệu đã gọi sự biến đổi là "mô hình thứ tư" [16]. Ông nghĩ rằng cách duy nhất đối phó với mô hình như vậy là phát triển một thế hệ mới các công cụ máy tính để quản lý, trực quan hóa và phân tích dữ liệu khổng lồ. Trong tháng 6 năm 2011, một sự kiện bước ngoặt xảy ra khi EMC/IDC công bố một báo cáo nghiên cứu có tựa đề *Trích xuất giá trị từ sự hỗn độn*, đây là lần đầu tiên đưa ra khái niệm và tiềm năng của Big Data. Báo cáo nghiên cứu này gây ra mối quan tâm lớn trong cả công nghiệp và học thuật về Big Data.

Trong vài năm qua, gần như những công ty lớn bao gồm EMC, Oracle, IBM, Microsoft, Google, Amazon, Facebook,… đã bắt đầu cá dự án Big Data của họ. Từ năm 2005, IBM đã đầu tư 16 tỷ USD vào 30 sự tiếp nhận liên quan đến dữ liệu lớn. Về học thuật, Big Data cũng chiếm địa vị nổi bật. Trong năm 2008, Nature công bố một vấn đề đặc biệt về Big Data. Năm 2011, Science cũng đưa ra một vấn đề đặc biệt về công nghệ chủ chốt "xử lý dữ liệu" trong Big Data. Năm 2012, Tạp chí Hiệp hội Nghiên cứu châu Âu Tin học và Toán học (ERCIM) đăng một vấn đề đặc biêt về dữ liệu lớn. Vào đầu năm 2012, một báo cáo mang tên *Big Data, Big Impact* trình bày tại diễn đàn Davos ở Thụy Sĩ, đã thông báo rằng Big Data đã trở thành một loại tài sản kinh tế mới, giống như tiền tệ hoặc vàng.

Nhiều chính phủ quốc gia như Mỹ cũng đã rất quan tâm tới dữ liệu lớn. Trong tháng 3 năm 2012, chính quyền Obama đã công bố một khoản đầu tư 200 triệu USD để khởi động "Kế hoạch nghiên cứu và phát triển Big Data". Tháng 7 năm 2012 dự án "Đẩy mạnh công nghệ thông tin Nhật Bản" được ban hành bởi Bộ Nội vụ và Truyền thông Nhật Bản chỉ ra rằng sự phát triển Big Data nên có một chiến lược quốc gia và các công nghệ ứng dụng nên là trọng tâm. Cũng trong thời gian đó, Liên Hiệp Quốc đã đưa ra báo cáo *Big Data cho phát triển*, trong đó tóm tắt cách mà các chính phủ sử dụng Big Data để phục vụ và bảo vệ người dân một cách tốt hơn.

Công ty nghiên cứu thị trường IDC cho thấy doanh thu đến từ thị trường Big Data sẽ tăng lên 16,9 tỷ USD vào năm 2015 và sẽ tiếp tục tăng trưởng kép với tốc độ 27% và đạt đến 32,4 tỷ USD vào năm 2017.

1.1.3 Những thách thức mà Big Data mang lại

Với sự gia tăng một cách mạnh mẽ của dữ liệu trong kỷ nguyên Big Data đã mang tới những thách thức rất lớn về việc thu thập, lưu trữ, quản lý và phân tích dữ liệu. Hệ thống quản lý và phân tích dữ liệu truyền thống được dựa trên hệ thống quản lý cơ sở dữ liệu quan hệ (RDBMS). Tuy nhiên, RDBMS chỉ áp dụng cho các dữ liệu có cấu trúc, khác với những dữ liệu bán cấu trúc hoặc không có cấu trúc. Ngoài ra, RDBMS đang ngày càng sử dụng nhiều phần cứng đắt tiền. Các RDBMS truyền thống không thể xử lý dung lượng rất lớn và không đồng nhất của Big Data. Cộng đồng nghiên cứu đã đề xuất một số giải pháp theo các quan điểm khác nhau. Đối với các giải pháp lưu trữ vĩnh viễn và quản lý các tập dữ liệu quy mô lớn không có trật tự, hệ thống tập tin được phân phối và cơ sở dữ liệu NoSQL là những lựa chọn tốt. Những

frameworks lập trình như vậy đã đạt được thành công lớn trong các bài toán xử lý cụm, đặc biệt đối với lập thứ hạng trang web (webpage ranking). Nhiều ứng dụng dữ liệu lớn có thể được phát triển dựa trên những công nghệ hoặc nền tảng cách mạng này.

Các thách thức chính mà Big Data mang lại:

- *Biểu diễn dữ liệu:* Nhiều bộ dữ liệu có mức độ không đồng nhất trong kiểu, cấu trúc, ngữ nghĩa, tổ chức, độ chi tiết và khả năng tiếp cận. Biểu diễn dữ liệu nhằm mục đích làm cho dữ liệu có ý nghĩa hơn trong việc phân tích của máy tính và sự giải thích của người dùng. Tuy nhiên, việc biểu diễn dữ liệu không đúng cách sẽ làm giảm giá trị ban đầu của dữ liệu và thậm chí có thể gây cản trở cho việc phân tích dữ liệu. Biểu diễn dữ liệu hiệu quả sẽ phản ánh cấu trúc, lớp và kiểu dữ liệu cũng như các công nghệ tích hợp, để cho phép hoạt động hiệu quả trên các tập dữ liệu khác nhau.

- *Giảm sự dư thừa và nén dữ liệu:* Giảm sự dư thừa và nén dữ liệu là cách hiệu quả để giảm chi phí gián tiếp của toàn bộ hệ thống trên tiền đề rằng các giá trị tiềm năng của dữ liệu không bị ảnh hưởng. Ví dụ, hầu hết các dữ liệu được tạo ra bởi các mạng cảm biến là rất cần thiết, trong đó có thể được logic và nén ở các đơn đặt hàng của các cường độ.

- *Quản lý vòng đời của dữ liệu:* Vòng đời của dữ liệu là chuỗi các giai đoạn mà một đơn vị dữ liệu từ thế hệ ban đầu được thu thập, lưu trữ đến khi bị xóa bỏ và kết thúc vòng đời hữu ích của nó. So với tiến bộ của hệ thống lưu trữ tương ứng, cảm biến và máy tính đang tạo ra dữ liệu với quy mô và tốc độ chưa từng có. Điều này đã tạo ra rất nhiều thách thức, một trong số đó là hệ thống lưu trữ hiện đại không thể hỗ trợ dữ liệu lớn như vậy. Vì vậy, một nguyên tắc quan trọng liên quan đến các giá trị phân tích cần được phát triển để quyết định dữ liệu nào sẽ được lưu trữ và dữ liệu nào sẽ được loại bỏ.

- *Cơ chế phân tích:* Hệ thống phân tích Big Data sẽ xử lý khối lượng dữ liệu không đồng nhất trong một thời gian giới hạn. Tuy nhiên, RDBMS truyền thống được thiết kế với sự thiếu khả năng thay đổi và khả năng mở rộng, do đó không thể đáp ứng các yêu cầu về hiệu suất. Cơ sở dữ liệu không quan hệ đã chỉ ra những lợi thế riêng của mình trong việc xử lý dữ liệu phi cấu trúc và bắt đầu trở thành đề tài chủ đạo trong phân tích Big Data. Mặc dù vậy, vẫn còn một số vấn đề về cơ sở dữ liệu không quan hệ trong hoạt động và những ứng dụng cụ thể của chúng. Điều này dẫn tới việc cần tìm

một giải pháp thỏa hiệp giữa RDBMS và cơ sở dữ liệu không quan hệ. Ví dụ, một số doanh nghiệp đã sử dụng một kiến trúc cơ sở dữ liệu hỗn hợp mà tích hợp những ưu điểm của cả hai loại cơ sở dữ liệu như Facebook và Taobao.

- *Bảo mật dữ liệu:* Hầu như các nhà cung cấp dịch vụ hoặc chủ sở hữu dịch vụ Big Data có thể không duy trì và phân tích một cách hiệu quả các tập dữ liệu lớn như vậy vì khả năng hạn chế của họ. Họ phải dựa vào các chuyên gia hoặc các công cụ để phân tích dữ liệu như vậy, làm tăng rủi ro bảo mật.

- *Quản lý năng lượng:* Năng lượng tiêu thụ của hệ thống máy tính lớn đã thu hút nhiều sự quan tâm từ cả quan điểm kinh tế và môi trường. Với sự gia tăng của dung lượng dữ liệu và nhu cầu phân tích, xử lý, lưu trữ và truyền tải thì Big Data chắc chắn sẽ tiêu thụ ngày càng nhiều năng lượng điện. Vì vậy, cơ chế kiểm soát và quản lý điện năng tiêu thụ cấp hệ thống sẽ được thành lập với Big Data trong khi khả năng mở rộng và khả năng tiếp cận được đảm bảo.

- *Khả năng mở rộng và thay đổi:* Hệ thống phân tích Big Data phải hỗ trợ tập dữ liệu hiện tại và tương lai. Thuật toán phân tích phải có khả năng xử lý các tập dữ liệu ngày càng mở rộng và phức tạp hơn.

- *Sự hợp tác*: Phân tích các dữ liệu lớn là một nghiên cứu liên ngành, trong đó yêu cầu các chuyên gia trong các lĩnh vực khác nhau hợp tác để thu thập các dữ liệu. Một kiến trúc mạng lưới Big Data toàn diện phải được thiết lập để giúp các nhà khoa học và kỹ sư trong các lĩnh vực khác nhau truy cập các loại dữ liệu khác nhau và sử dụng đầy đủ chuyên môn của họ, phối hợp để hoàn thành các mục tiêu phân tích.

1.1.4 Những công nghệ trong Big Data

Có rất nhiều công nghệ gắn liền với Big Data, ở phần này sẽ giới thiệu một số công nghệ cơ bản liên quan chặt chẽ tới Big Data bao gồm điện toán đám mây, IoT, trung tâm dữ liệu và Hadoop.

- *Điện toán đám mây:*

Theo IBM thì điện toán đám mây là việc cung cấp tài nguyên máy tính cho người dùng tùy theo mục đích sử dụng thông qua Internet. Nguồn tài nguyên đó có thể là bất cứ thứ gì liên quan đến điện toán và máy tính, ví dụ như phần mềm, phần cứng, hạ tầng mạng cho tới các máy chủ và mạng lưới máy chủ cỡ lớn.

Điện toán đám mây có liên quan chặt chẽ với Big Data. Big Data là đối tượng của hoạt động tính toán chuyên sâu và nhấn mạnh khả năng lưu trữ của mỗi hệ thống