**TRƯỜNG ĐH CÔNG NGHỆ THÔNG TIN**

**KHOA CÔNG NGHỆ PHẦN MỀM**

-----🙞🙜🕮🙞🙜-----

****

**BÁO CÁO**

**Đề tài:**

**Xây dựng hệ thống recommendation**

**dựa trên Content base filtering**

**trên nền tảng Big Data với**

**Hadoop Mapreduce**

Giảng viên hướng dẫn:

Thạc sĩ: Phạm Thi Vương

Sinh viên thực hiện:

12520053 Nguyễn Viết Danh

# TÌM HIỂU VỀ BIG DATA

## Big data

### Giới thiệu

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

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

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

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

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

### Định nghĩa “big data”

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

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

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

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

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

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

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

### Một số đặc tính của big data

## Ứng dụng

## Các công cụ

# tài liệu tham khảo

Wikipedia

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