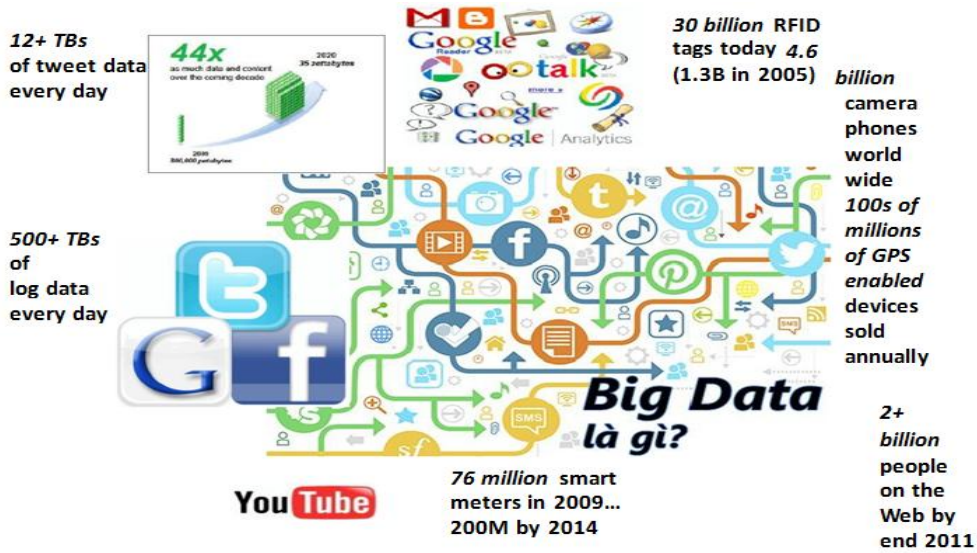
TÌM HIỂU VỀ BIG DATA

Nguyễn Gia Luyện
Giám đốc Trung tâm Tin học thống kê KVI, TCTK

Dữ liệu lớn (Bigdata) là tất cả các loại dữ liệu có dung lượng lớn; có giá trị lớn, nhưng khó khai thác và có thể rất nhạy cảm với thời gian. Bigdata đã vượt xa dữ liệu cấu trúc tiêu biểu (typical), nó có thể được truy vấn với hệ thống quản lý dữ liệu quan hệ - thường với những tệp phi cấu trúc (unstructured files), video kỹ thuật số, hình ảnh, dữ liệu cảm biến, tệp lưu nhật ký, thực sự bất cứ dữ liệu nào không có trong hồ sơ với các phạm vi tìm kiếm khác.

Bigdata được hình thành từ nhiều nguồn dữ liệu khác nhau. Quan sát ở hình dưới ta thấy đối với nguồn hình thành nên dữ liệu lớn ở khắp mọi nơi, đặc biệt có sẵn ở mạng xã hội như: Facebook, Twitter…



Theo tập đoàn SAS có một vài số liệu về Bigdata như sau:

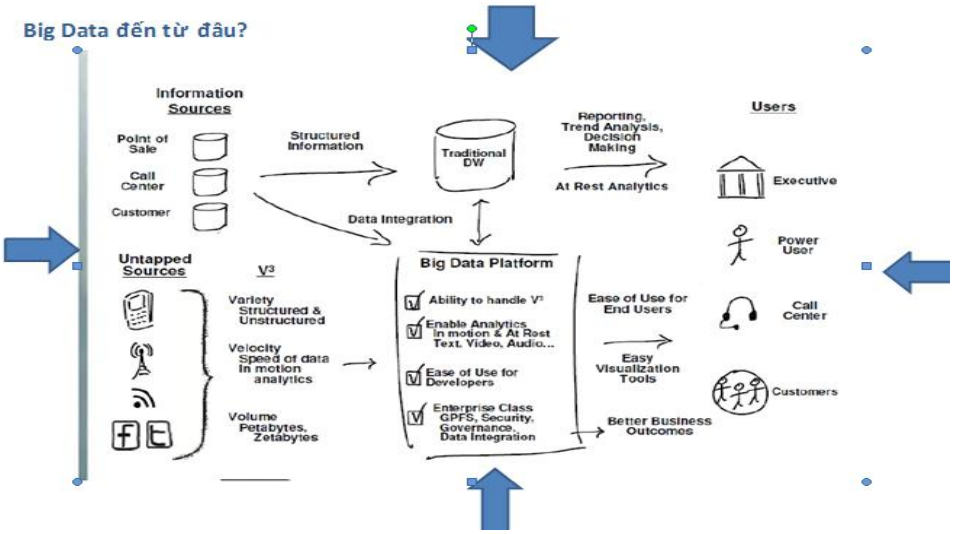
+ Trong vòng 4 giờ của ngày "Black Friday" năm 2012, cửa hàng Walmart đã phải xử lí hơn 10 triệu giao dịch tiền mặt, tức là khoản 5.000 giao diện mỗi giây.

+ Dịch vụ chuyển phát UPS nhận khoảng 39,5 triệu yêu cầu từ khách hàng của mình mỗi ngày

+ Dịch vụ thẻ VISA xử lí hơn 172.800.000 giao dịch thẻ chỉ trong vòng một ngày mà thôi

+ Trên Twitter có 500 triệu dòng tweet mới mỗi ngày, Facebook thì có 1,15 tỉ thành viên tạo ra một mớ khổng lồ dữ liệu văn bản, tập tin, video…

Theo Intel vào tháng 9/2013, hiện nay thế giới đang tạo ra 1 petabyte dữ liệu trong mỗi 11 giây - tương đương với một đoạn video HD dài 13 năm. Ví dụ, eBay thì sử dụng hai trung tâm d ữ liệu dung lượng lên đến 40 petabyte để chứa những truy vấn, tìm kiếm, đề xuất cho khách hàng cũng như thông tin về hàng hóa của mình . Amazon.com thì phải xử lí hàng triệu hoạt động mỗi ngày, Facebook cũng phải quản lí 50 tỉ bức ảnh từ người dùng t ải lên, YouTube hay Google thì phải lưu lại hết các lượt truy vấn và video của người dùng cùng nhiều loại thông tin khác có liên quan.



Mặc dù mới nghiên cứu và được đưa vào ứng dụng, sử dụng trên thực tế chưa lâu, nhưng không ai có thể phủ nhận được sự phát triển mạnh mẽ của Bigdata: từ dữ liệu truyền thống (data warehousing) đến dữ liệu phi cấu trúc (flat file). Dữ liệu truyền thống (dữ liệu dạng có cấu trúc) với sự tăng trưởng khổng lồ đang tạo ra thách thức lớn cho các tổ ch ức, các tổ ch ức phải đưa ra các giải pháp kho dữ liệu, nơi dữ liệu được lưu trữ và xử lý . Vì vậy, xu hướng kinh doanh thông minh (business intelligence) đã trở thành nhu cầu hàng ngày . Mô hình cơ sở d ữ liệu quan hệ và các khái nhiệm kho dữ liệu tất cả đều được xây dựng dựa trên mô hình cơ sở d ữ liệu quan hệ truyền thống nhưng hiện nay cũng gặp phải thách thức khi có sự xuất hiện của dữ liệu không có cấu trúc. Trong tương lai nhu cầu ngư ời dùng đòi h ỏi có nhiều thông tin, đa dạng hóa khai thác thông tin hơn. Còn đối với dạng dữ liệu phi cấu trúc, dữ liệu không có khả năng truy xuất dữ liệu hiệu quả và việc tích h ợp dữ liệu không thể thực hiện khi không có bất kỳ mô hình hay c ấu trúc xung quanh. Dữ liệu lưu trong flat file sẽ g ặp vấn đề về khai thác, mặc dù các ứng dụng đư ợc phát triển tại thời điểm đó là phù hợp với sự phát triển của kỹ thuật và dữ liệu.

Hiện nay, với sự phát triển ngày càng rộng của Bigdata sẽ tạo ra nhiều cơ hội, tuy nhiên đi đôi với việc đó, sẽ có nhiều thách thức chúng ta gặp phải…

Trước đây

- Mỗi tổ chức đều có chuyên môn để quản lý dữ liệu mình có
- Có thông tin thông minh trong video, photos, SMS, các tin nhắn mạng xã hội. Tất cả những điều này cần phải mang đến 1 nền tảng duy nhất và xây dựng 1 hệ thống thống nhất

Ngày nay

- Thế giới đã thay đổi thêm dữ liệu không có cấu trúc
- Người dùng yêu cầu những tính năng và công nghệ phải được xây dựng để hỗ trợ sửa dụng được những dữ liệu này

Ngày nay, có một lượng dữ liệu lớn (Volume) tốc độ cao (Velocity) khác nhau (Variety) là các thuộc tính của dữ liệu. Hệ thống cơ sở dữ liệu truyền thống có giới hạn khi giải quyết các thách thức mới. Đây là nhu cầu của khoa học Bigdata . Chúng ta đổi mới cách xử lý và quản lý dữ liệu, cần có những sáng tạo trong thu thập và trình bày dữ liệu cho người dùng tin. Vì vậy, nhu cầu ứng dụng thông minh thời gian thực từ các luồng dữ liệu tốc độ cao đang trở nên cần thiết.

Những lợi ích từ việc sử dụng Bigdata

Nhà nghiên cứu Danah Boyd đã đưa ra quan ngại của mình rằng việc sử dụng Bigdata trong việc chọn mẫu thống kê có thể gây ra sự chủ quan, và dù ít hay nhiều thì nó cũng có thể ảnh hưởn đến kết quả cuối cùng. Việc khai thác dữ liệu từ một số nguồn là Bigdata, trong khi những nguồn khác không phải là "dữ liệu lớn" thì đặt ra những thách thức khi phân tích dữ liệu.

Tóm lại, Bigdata là thách thức đặt ra cho các tổ chức, doanh nghiệp trong thời đại số hiện nay. Một khi làm chủ được dữ liệu lớn thì họ sẽ có cơ hội thành công lớn hơn trong bối cảnh cạnh tranh ngày nay, người dùng sẽ được hưởng lợi hơn từ việc trích xuất thông tin một cách chính xác hơn , hữu ích hơn với chi phí thấp hơn. Vẫn còn đó những chỉ trích xoay quanh Bigdata , tuy nhiên lĩnh vực này vẫn còn rất mới và chúng ta hãy chờ xem trong tương lai Bigdata sẽ phát triển như thế nào.