1. Tổng quan về BigData

Ngày nay, trước sự bùng nổ của internet, dữ liệu đã xuất hiện những dạng phức tạp hơn như phi cấu trúc hay bán cấu trúc. Các dạng dữ liệu đó được gọi chung cái tên là Big Data (dịch là “dữ liệu lớn”). Dữ liệu này có thể sinh ra từ bất cứ đâu như: hồ sơ, giao dịch, hình ảnh, nhắn tin,… của con người; nói cách khác, đó chính là những dữ liệu được sinh ra trong quá trình chia sẻ thông tin trực tiến liên tục của người sử dụng. Để cung cấp cái nhìn tổng quan, bài báo cáo xin giới thiệu tóm tắt những nét chính về dữ liệu lớn.

1. Khái niệm về dữ liệu lớn

* Theo wikipedia: Dữ liệu lớn là một thuật ngữ chỉ tập hợp dữ liệu với kích thước vượt xa khả năng của các công cụ phần mềm thông thường để thu thập, hiển thị, quản lý và xử lý dữ liệu trong một thời gian có thể chấp nhận được.
* Theo Gartner: Dữ liệu lớn là những nguồn thông tin có đặc điểm chung với khối lượng lớn, tốc độ nhanh và dữ liệu được định dạng dưới nhiều hình thức khác nhau, do đó muốn khai thác được đòi hỏi phải có hình thức xử lý mới để đưa ra quyết định, khám phá và tối ưu hóa quy trình.

1. Đặc trưng của dữ liệu lớn

Dữ liệu lớn có 5 đặc trưng cơ bản như sau:

1. Khối lượng dữ liệu (Volume)

Đây là đặc điểm tiêu biểu nhất của dữ liệu lớn. Kích cỡ của Big Data đang từng ngày tăng lên. Tính đến năm 2012, một tập dữ liệu lớn có khối lượng rơi vào khoảng vài chục terabyte cho tới vài petabyte (1 petabyte = 1024 terabyte). Để lưu trữ chúng, ta cần sử dụng công nghệ “đám mây” mới đủ khả năng đáp ứng để lưu trữ nó.

1. Tốc độ (Velocity)

Tốc độ ở đây được nói đến hai khía cạnh: (a) Khối lượng dữ liệu lớn tăng rất nhanh; (b) Xử lý dữ liệu nhanh ở mức thời gian thực. Theo tài liệu của Intel vào năm 2013, thế giới tạo ra 1 petabyte mỗi 11 giây. Nó tương đương một đoạn video HD dài 13 năm. Cùng với sự phát triền của nó, các công nghệ xử lý dữ liệu hiện nay đã cho phép xử lý dữ liệu tức thì trước khi chúng được lưu vào cơ sở dữ liệu

1. Đa dạng (Variety)

Đối với dữ liệu truyền thống,chúng ta hay nói đến dữ liệu có cấu trúc. Nhưng ngày nay, 80% dữ liệu được sinh ra là phi cấu trúc. Big Data cho phép liên kết và phân tích nhiều dạng dữ liệu khác nhau.

1. Độ tin cậy/chính xác (Veracity)

Một trong những đặc trưng phức tạp nhất của dữ liệu lớn là độ tin cậy/chính xác của dữ liệu. Với xu hướng phương tiện truyền thông xã hội và mạng xã hội ngày nay và sự gia tăng mạnh mẽ tính tương tác và chia sẻ của người dùng làm cho bức tranh xác định độ tin cậy và chính xác của dữ liệu ngày một khó khăn. Bài toán phân tích và loại bỏ dữ liệu thiếu chính xác và dữ liệu nhiễu đang là tính chất quan trọng của Big Data.

1. Giá trị (Value)

Giá trị là đặc điểm quan trọng nhất của dữ liệu lớn. Khi bắt đầu triển khai xây dựng dữ liệu lớn, ta cần xác định được giá trị mà nó mang lại. Nếu chúng ta có dữ liệu lớn nhưng chỉ khai thác được 1% giá trị của nó, thì ta không nên đầu tư phát triển dữ liệu lớn.

1. Sự khác biệt giữa dữ liệu lớn và dữ liệu truyền thống

Dữ liệu lớn khác với dữ liệu truyền thống ở 4 điểm cơ bản sau: Dữ liệu đa dạng hơn, lưu trữ nhiều dữ liệu hơn, truy vấn nhanh hơn, độ chính xác cao hơn.

1. Dữ liệu đa dạng hơn: Khi khai thác dữ liệu truyền thống (dữ liệu có cấu trúc), chúng ta cần trả lời các câu hỏi: lấy ra kiểu gì? Định dạng như thế nào? Đối với dữ liệu lớn, ta không cần quan tâm đến các câu hỏi trên mà chỉ quan tâm là giá trị mà dữ liệu mang lại có đáp ứng được cho công việc hiện tại và tương lai hay không.
2. Lưu trũ dữ liệu lớn hơn: Lưu trữ dữ liệu truyền thống vô cùng phức tạp và luôn phải đặt ra câu hỏi: lưu như thế nào? Dung lượng bao nhiêu là đủ? Kèm theo đó là chi phí đầu tư tương ứng. Với công nghệ lưu trữ đám mây hiện nay, việc lưu trữ dữ liệu lớn phần nào được giải quyết. Cùng với đó là việc phân phối lưu trữ dữ liệu phân tán và có thể kết hợp các dữ liệu phân tán lại với nhau một cách chính xác và xử lý nhanh trong thời gian thực.
3. Truy vấn dữ liệu nhanh hơn: Dữ liệu lớn được cập nhật liên tục trong khi đó kho dữ liệu truyền thống thì không và trong tình trạng không theo dõi thường xuyên dẫn đến tình trạng lỗi cấu trúc truy vấn khiến không tìm được thông tin đáp ứng theo yêu cầu.
4. Độ chính xác cao hơn: Dữ liệu lớn khi đưa vào sử dụng thường được kiểm định lại dữ liệu với những điều kiện chặt chẽ, số lượng thông tin được kiểm tra thông thường rất lớn, và đảm bảo về nguồn lấy dữ liệu không có sự tác động của con người vào thay đổi số liệu thu thập.

Trước những sự khác biệt đó, chúng ta cần những phương pháp phù hợp để xử lý dữ liệu lớn một cách nhanh chóng và chính xác.

1. Các phương pháp xử lý BigData
2. Phương pháp phân lớp
   1. Bài toán phân lớp
   2. Thuật toán Cây quyết định
   3. Các thuật toán khác
3. Phương pháp phân lớp