Slide 4

Có 4 vấn đề chính trong xử lý dữ liệu và phân tích dữ liệu. Four Vs bao gồm:

Volume (khối lượng): với sự phát triển của công nghệ, khối lượng dữ liệu ngày càng tăng nhanh chóng, đặt ra thách thức cho việc lưu trữ, quản lý và xử lý dữ liệu.

Variety (đa dạng) : là độ phong phú và đa dạng của nguồn dữ liệu, các loại dữ liệu khác nhau như dữ liệu có cấu trúc, dữ liệu văn bản, dữ liệu hình ảnh, dữ liệu âm thanh, dữ liệu video, ... có tới 90% dữ liệu được tạo ra ngày nay, chưa được cấu trúc

Velocity (tốc độ) : là tốc độ mà dữ liệu được tạo ra, thu thập và xử lý. Khối lượng dữ liệu tạo ra ngày càng lớn với tốc độ ngày càng nhanh yêu cầu phải xử lý nhanh chóng để đảm bảo tính hữu dụng của dữ liệu. Với những khối dữ liệu khổng lồ không cho phép hệ thống lưu trữ trước khi xử lý, yêu cầu khả năng xử lý dữ liệu theo thời gian thực.

Veracity (độ chính xác) : là độ tin cậy và độ chính xác của dữ liệu. Dữ liệu đến từ nhiều nguồn khác nhau và mỗi nguồn có thể không hoàn toàn đáng tin cậy. Việc làm sạch các dữ liệu cần được thực hiện để đảm bảo tính đáng tin cậy của dữ liệu. Một khía cạnh khác là độ chính xác của dữ liệu. Dữ liệu cần được đảm bảo rằng nó không bị thay đổi bởi bất kì tác nhân nào. Điểm cơ bản là dữ liệu cần phải đáng tin cậy.

Slide 5

Các đặc tính của big data yêu cầu các hệ thống, phương pháp và cách tiếp cận mới, khác với các hệ quản trị cơ sở dữ liệu DBMS truyền thống. Chương này và chương tiếp theo nhấn mạnh các phương pháp cơ sở hạ tầng hệ thống đang được phát triển để giải quyết ba điểm chính của big data: khối lượng, tốc độ và đa dạng. Quản lý big data phụ thuộc vào một bộ kết cấu phần mềm khác với lưu trữ phân tán, các khung xử lý dữ liệu và các công cụ để xây dựng phân tích big data phức tạp hơn. Phần còn lại của chương tập trung vào các thành phần của kiến trúc này, giải quyết các vấn đề về khả năng mở rộng, xử lý dữ liệu luồng, hệ thống đồ thị, các vấn đề đa dạng và các data lake.

Slide 6

Hệ quản trị cơ sở dữ liệu hiện tại sử dụng lưu trữ khối nhưng quản lý dữ liệu liệu lớn sử dụng một lớp lưu trữ phân tán, các dữ liệu được phân tán trên các nút của một cụm không chia sẻ.

Có 2 giải pháp để lưu trữ dữ liệu trên lớp lưu trữ phân tán, đó là lưu trữ đối tượng (object storage) và lưu trữ tập tin(file storage). Lưu trữ đối tượng quản lý dữ liệu như các đối tượng, bao gồm dữ liệu và siêu dữ liệu (metadata) của chúng, mỗi đối tượng được định danh bởi một chỉ số độc nhất. Nó giúp di chuyển đối tượng dữ liệu giữa các vị trí phân tán. Một đối tượng có thể được biểu diễn dưới dạng một bộ ba oid, dữ liệu và metadata. Sử dụng không gian đối tượng phẳng giúp quản lý lượng lớn cá đối tượng dữ liệu không cấu trúc. Các đối tượng có thể được truy cập dễ dàng bằng API dựa trên REST với các lệnh get/put dữ liệu dễ sử dụng trên các giao thức Internet. Object storage đặc biệt hữu ích để lưu một lượng lớn các đối tượng nhỏ như hình ảnh, tệp đính kèm theo thư... File storage quản lý dữ liệu trong các tệp không cấu trúc(các chuỗi byte) trên đó dữ liệu có thể được tổ chức thành các bản ghi có độ dài cố định hoặc có độ dài thay đổi. Hệ thống tập tin tổ chức các tệp trong một cấu trúc thư mục và duy trì cho mỗi tệp của nó các metadata. Với một khối dữ liệu lớn, một file cần phải được phân vùng và phân phối.

Slide 7

GFS là một hệ thống file phân tán được phát triển bởi gg để sử dụng trong các ứng dụng và hệ thống của gg chẳng hạn như Bigtable. Nó nhắm đến việc cung cấp hiệu suất, khả năng mở rộng, độ tin cậy và khả năng khả dụng. Tuy nhiên các hệ thống mục tiêu, ví dụ như các cụm không chia sẻ, có xác suất bị lỗi tại một thời điểm nhất định là cao. GFS giải quyết vấn đề này thông qua sao chép và chuyển giao.

GFS được tối ưu hóa cho các ứng dụng dữ liệu có tính quy mô lớn. Các ứng dụng có đặc điểm như sau : các tệp của chúng rất lớn, thường là vài gigabyte; các tải công việc chủ yếu bao gồm hoạt động đọc và ghi thêm, trong khi các cập nhật ngẫu nhiên là hiếm; vì các tải công việc chủ yếu là đọc và ghi thêm nên tốc độ xử lý cao quan trọng hơn là độ trễ thấp.

GFS tổ chức các tệp tin dưới dạng cây thư mục và xác định chúng bằng đường dẫn. Nó cung cấp một giao diện hệ thống tệp tin với các thao tác tệp tin truyền thống như tạo, mở, đọc, ghi, đóng và xóa tệp tin và 2 thao tác bổ sung là snapshot – cho phép tạo một bản sao của tệp tin haowcj của một cây thư mục và record – cho phép nối dữ liệu vào một tệp tin bởi các máy khách đồng thời một cách hiệu quả.

Slide 76

Hồ dữ liệu lưu trữ dữ liệu dưới định dạng gốc của nó. Mỗi phần tử dữ liệu có thể được lưu trữ trực tiếp, với một định danh duy nhất và siêu dữ liệu liên quan. Vì vậy với mỗi câu hỏi nghiệp vụ, người dùng có thể nhanh chóng tìmd thấy tập dữ liệu liên quan để phân tích.

Hồ dữ liệu được liên kết với hệ sinh thái phần mềm hadoop.

Slide 77

Các ưu điểm của hồ dữ liệu

Schema on read (đọc theo schema) : viết dữ liệu dưới dạng ban đầu của nó, đọc nó theo một biểu đồ được chỉ định. Giúp dữ liệu được truy vấn trong định dạng tự nhiên của nó. Điều này cung cấp tính linh hoạt khi dữ liệu mới có thểm được thêm bất cứ lúc nào

Xử lý dữ liệu nhiều công việc : bộ chồng phần mềm quản lý dữ liệu lớn cung cấp hỗ trợ cho nhiều phương thức truy cập đến từ cùng một dữ liệu. Ví dú như phân tích lô với một framework như MapReduce, OLAP tương tác hoặc phân tích kinh doanh với một framework