CHƯƠNG 1

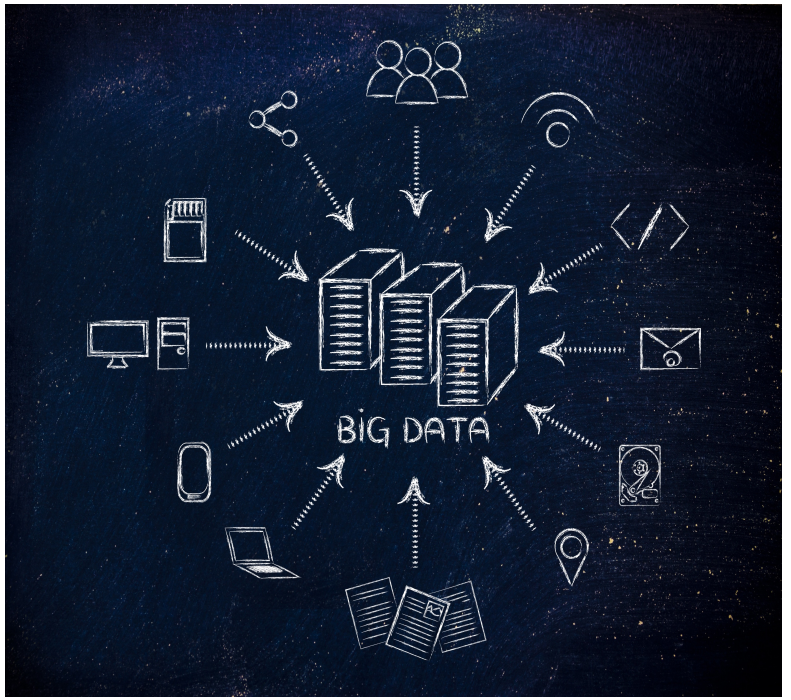
HIỂU VỀ DỮ LIỆU LỚN

NỘI DUNG BÀI HỌC

- TỔNG QUAN
- ĐẶC ĐIỂM VÀ VÀI TRÒ CỦA DỮ LIỆU LỚN
- CÁC GIAI ĐOẠN CỦA QUẢN LÝ DỮ LIỆU
- QUẢN LÝ DỮ LIỆU LỚN
- CÁC LOẠI DỮ LIỆU
- MỘT SỐ ỨNG DỤNG CỦA DỮ LIỆU LỚN

TỔNG QUAN VỀ DỮ LIỆU LỚN

- Hiện nay dữ liệu đang ngày càng bùng nổ về số lượng và thể loại (âm thanh, hình ảnh, văn bản, các dữ liệu có cấu trúc và phi cấu trúc …)
- Các kho dữ liệu tại các công ty đang ngày càng phong phú, con người bắt đầu tìm thấy giá trị của nguồn dữ liệu khổng lồ đó
- Các công ty công nghệ lớn hiện nay đang tận dụng tối đa lợi ích từ nguồn dữ liệu lớn này và kiếm về lợi nhuận khổng lồ
- Các dữ liệu y khoa được khai thác để xây dựng các hệ thống chuẩn đoán và phát hiện sớm ...

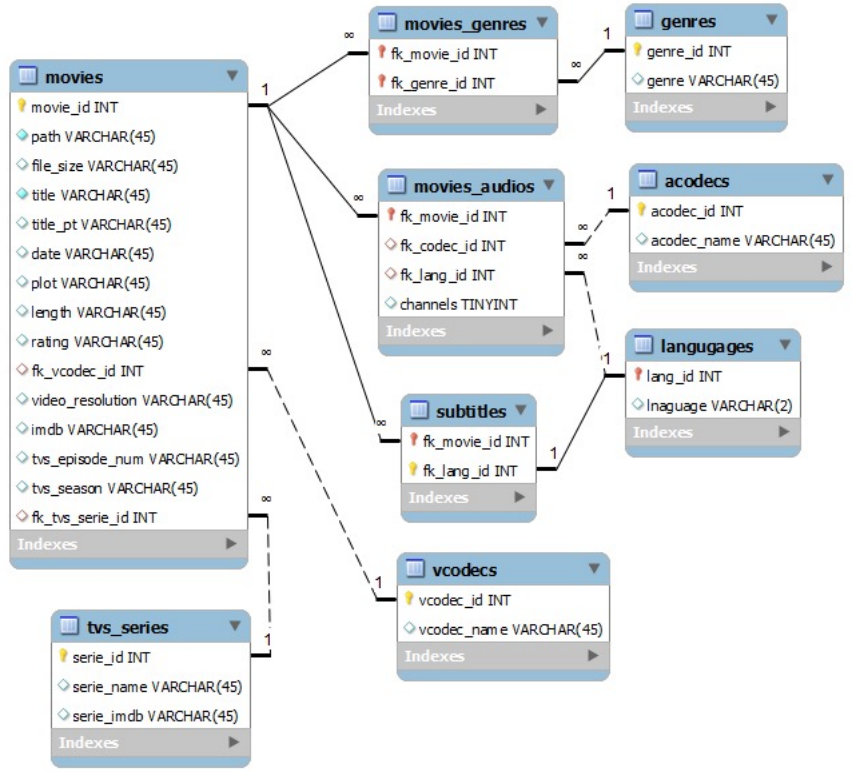


ĐẶC ĐIỂM VÀ VÀI TRÒ CỦA DỮ LIỆU LỚN

- **Dữ liệu lớn có 5 đặc điểm chung:**
  - ***Khối lượng dữ liệu (Volume)*** : thể hiện sự lớn của dữ liệu về mặt khối lượng, dữ liệu thường có kích thước lên đến vài Terabyte hoặc petabyte.
  - ***Tốc độ (Velocity):*** Được thể hiện theo 2 khí cạnh
    - Tốc độ gia tăng khối lượng dữ liệu
    - Tốc độ xử lý dữ liệu ***(theo thời gian thực)***
  - ***Đa dạng (Variety):*** có 2 dạng là có cấu trúc và phi cấu trúc, hiện nay gần 80% dữ liệu phát sinh là phi cấu trúc (tài liệu, hình ảnh, âm thanh, video, dữ liệu từ các thiết bị cảm biến ...)
  - ***Độ tin cậy (Veracity):*** Tính chất phức tạp nhất chính là xác định độ tin cậy của dữ liệu lớn, do nguồn dữ liệu đặc biệt quá lớn
  - ***Giá trị (Value):*** Đây là đặc điểm quan trọng nhất của dữ liệu lớn, trước khi xây dựng 1 ứng dụng dựa trên dữ liệu lớn ta cần đánh giá các giá trị mà dữ liệu lớn mang lại cho ứng dụng (cân nặng có thể được xây dựng các ứng dụng ước lượng mật độ xương hoặc khả năng mắc bên tim mạch, tiểu đường)

CÁC GIAI ĐOẠN CỦA QUẢN LÝ DỮ LIỆU

- Việc quản lý dữ liệu đã được phát triển từ khi máy tính bắt đầu hình thành, nhưng sự thay đổi của các dạng dữ liệu bắt con người phải nâng cấp các hệ thống quản lý dữ liệu lên để có thể đáp ứng được nhu cầu lưu trữ.
- Việc quản lý dữ liệu có thể được thể hiện qua 3 giai đoạn:
  - ***Giai đoạn 1:*** Quản lý dữ liệu trong các cơ sở dữ liệu
    - Bắt đầu từ những năm 1960, các dạng dữ liệu được lưu là dữ liệu phẳng sau được nâng cấp lên thành các dạng dữ liệu quan hệ và được quản trị bằng các hệ quản trị CSDL
    - Khi dữ liệu bắt đầu bùng nổ, mô hình dữ liệu quan hệ đang xuất hiện các điểm yếu như lưu trữ tốn kém, hiệu suất truy cập thấp, dữ liệu trùng lập cao ...

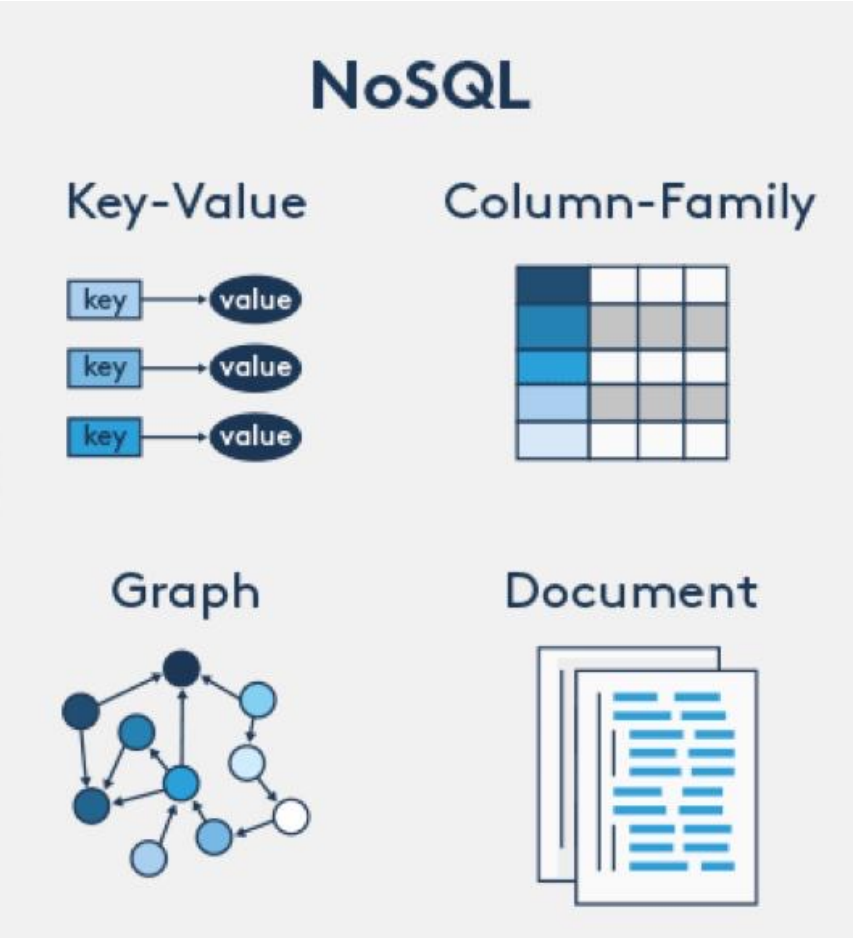


CÁC GIAI ĐOẠN CỦA QUẢN LÝ DỮ LIỆU

- ***Giai đoạn 2:*** Quản lý nội dung và web
  - Bắt đầu từ những năm 80, các doanh nghiệp có nhu cầu lưu trữ dữ liệu phi cấu trúc như: hình ảnh, tài liệu, âm thanh ... Ngày càng tăng cao.
  - Các nên tảng web cũng ngày càng phát triển, đòi hỏi phải có giải pháp lưu trữ và quản lý hợp lý
  - Những công nghệ như điện toán đám mây, ảo hoá, các hệ thống tích hợp công nghệ web ... Giúp giải quyết bài toán quản lý.

CÁC GIAI ĐOẠN CỦA QUẢN LÝ DỮ LIỆU

- ***Giai đoạn 3:*** Quản lý dữ liệu lớn
  - Việc ảo hoá dữ liệu và lưu trữ trên các đám mây giúp cho việc khai thác dữ liệu lớn ngày càng đơn giản hơn.
  - Việc độ tin cây, tốc độ internet và sức mạnh xử lý của máy tính ngày càng tăng cộng thêm giá thành của các thiết bị ngày càng rẻ giúp cho việc tiếp cận với việc quản lý dữ liệu lớn của doanh nghiệp ngày càng dễ dàng
  - Ngoài việc quản lý, các nhà quản trị còn có thể tìm thấy được các tri thức trong các tập dữ liệu lớn thông qua các thuật toán



QUẢN LÝ DỮ LIỆU LỚN

- Để quản lý dữ liệu lớn, các doanh nghiệp cần thực hiện các công việc sau:
  1. Bắt đầu với việc nắm bắt, tổ chức, tích hợp, phân tích và hoạt động của doanh nghiệp.
  2. Thiết lập nền tảng kiến trúc.
  3. Cơ sở hạ tầng hỗ trợ, dư thừa dữ liệu.
  4. Cơ sở hạ tầng an ninh.
  5. Nguồn dữ liệu tác nghiệp.
  6. Vấn đề hiệu suất.
  7. Tổ chức các dịch vụ và công cụ dữ liệu.
  8. Kho dữ liệu phân tích và dữ liệu thị trường.
  9. Phân tích dữ liệu lớn.
  10. Lập báo cáo và biểu diễn trực quan.

CÁC LOẠI DỮ LIỆU

- **Dữ liệu có cấu trúc:** là loại dữ liệu có khuôn mẫu nhất định như chiều dài, kiểu dữ liệu, miền giá trị. Các khuôn mẫu này sẽ được áp dụng để thu thập dữ liệu. Chiếm khoản 20% số lượng dữ liệu hiện nay. Có 2 nguồn để tạo ra loại dữ liệu này là:
  - ***Dữ liệu do máy tính tạo ra:***
    - Dữ liệu từ các cảm biến.
    - Web log data.
    - Dữ liệu tài chính.
  - ***Dữ liệu do con người tạo ra:***
    - Dữ liệu nhập.
    - Dữ liệu nhấp chuột.
    - Dữ liệu trò chơi.
    - Các dòng dữ liệu trên mạng xã hội, diễn đàn, điện thoại ...

CÁC LOẠI DỮ LIỆU

- **Dữ liệu phi cấu trúc:** là loại dữ liệu không có khuôn mẫu rõ ràng. Dữ liệu phi cấu trúc rất đa dạng và khó phân tích hơn dữ liệu có cấu trúc. 80% dữ liệu trên internet là dữ liệu phi cấu trúc.
- Các nguồn sinh ra dữ liệu phi cấu trúc:
  - Ảnh vệ tinh.
  - Dữ liệu khoa học.
  - Hình ảnh và video.
  - Radar hay dữ liệu sóng siêu âm.
  - Dữ liệu văn bản.
  - Dữ liệu về môi trường mạng xã hội.
  - Dữ liệu di động
  - Nội dung website.

MỘT SỐ ỨNG DỤNG CỦA DỮ LIỆU LỚN

- Chăm sóc sức khoẻ.
- Quản lý sản xuất.
- Quản lý giao thông
- Hệ thống gợi ý trong các lĩnh vực.
- Hệ hỗ trợ ra quyết định.
- Trí tuệ nhân tạo.

NỘI DUNG BÀI HỌC

- TỔNG QUAN
- ĐẶC ĐIỂM VÀ VÀI TRÒ CỦA DỮ LIỆU LỚN
- CÁC GIAI ĐOẠN CỦA QUẢN LÝ DỮ LIỆU
- QUẢN LÝ DỮ LIỆU LỚN
- CÁC LOẠI DỮ LIỆU
- MỘT SỐ ỨNG DỤNG CỦA DỮ LIỆU LỚN