CHƯƠNG 2

PHÂN TÍCH DỮ LIỆU CÓ CẤU TRÚC

NỘI DUNG BÀI HỌC

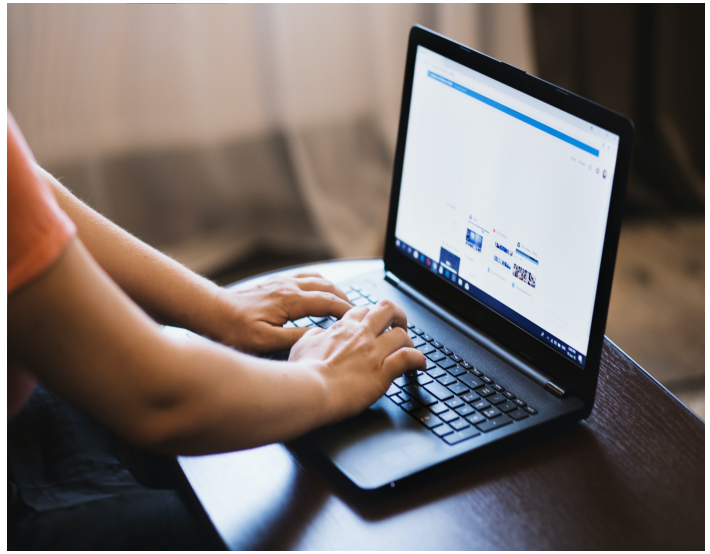
- CÁC HÌNH THỨC PHÂN TÍCH DỮ LIỆU
- PHÂN TÍCH CƠ BẢN
- PHÂN TÍCH NÂNG CAO
- GIẢI THUẬT PHÂN TÍCH DỮ LIỆU
- CƠ SỞ HẠ TẦNG ĐỂ PHÂN TÍCH DỮ LIỆU LỚN
- ỨNG DỤNG PHÂN TÍCH DỮ LIỆU LỚN

CÁC HÌNH THỨC PHÂN TÍCH DỮ LIỆU

HÌNH THỨC PHÂN TÍCH	CÔNG DỤNG
Phân tích cơ bản	Phân chia dữ liệu, báo cáo, trình bày, theo dõi và giám sát, nhận dạng bất thường
Phân tích nâng cao	Phân tích tinh vi bao gồm các mô hình dự đoán và các kỹ thuật kết hợp khác
Phân tích hoạt động	Phân tích quy trình kinh doanh
Phân tích doanh thu	Phân tích nhằm thúc đẩy kinh doanh

PHÂN TÍCH CƠ BẢN

- **PHÂN CHIA DỮ LIỆU:**
  - Chia dữ liệu thành các phần nhỏ để dễ thao tác, khám phá.
  - Tìm sự liên quan giữa các yếu tố (thuộc tính) trong dữ liệu.
  - Thực hiện thủ công thông qua các phương pháp thống kê.
  - Thường được biểu diễn thông qua các biểu đồ.
  - Ví dụ:
    - Phân tích mối quan hệ giữa nhiệt độ và độ pH trong nước.
    - Phân tích thông tin doanh số bán hàng của vùng vào 1 thời điểm trong năm



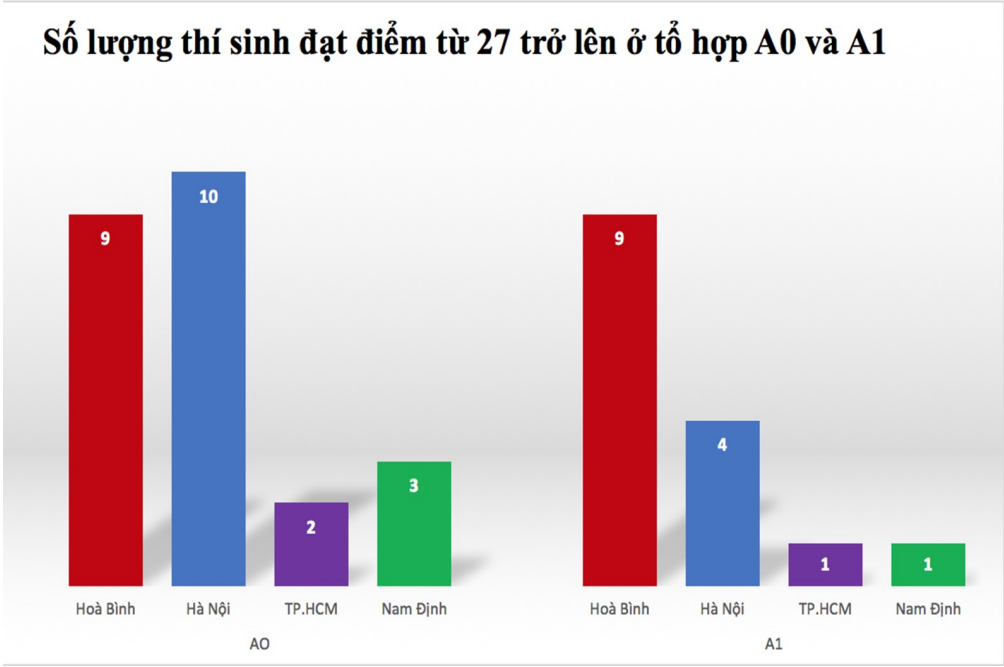
PHÂN TÍCH CƠ BẢN

- **THEO DÕI VÀ GIÁM SÁT CƠ BẢN:**
  - Mục tiêu là theo dõi và giám sát khối lượng lớn dữ liệu được thu thập theo thời gian thực.
  - Đưa ra các giải pháp thời vụ khi phân tích dữ liệu.
  - Ví dụ:
    - Theo dõi tình trạng internet của một công ty viễn thông.
    - Theo dõi lượng truy cập cùng lúc của các từ khoá, để xây dựng một chiến dịch quảng cáo hợp lý.

Google Ads

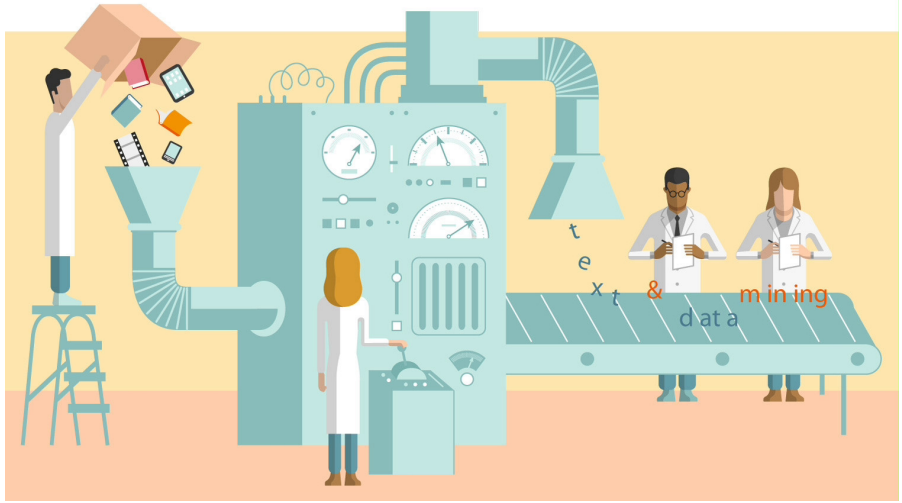
PHÂN TÍCH CƠ BẢN

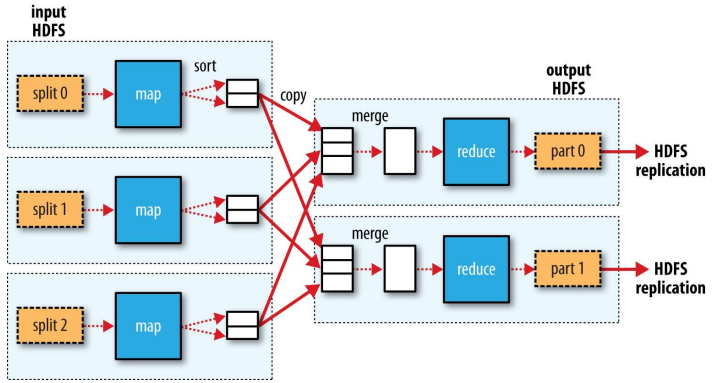
- **NHẬN DẠNG BẤT THƯỜNG:**
  - Mục tiêu là xác định các bất thường, chẳng hạn như một sự kiện mà các quan sát thực tế khác với những gì chúng ta mong đợi.
  - Ví dụ:
    - Khi năng suất giảm thì các biểu đồ dữ liệu sẽ giảm từ đó có thể kịp thời đưa ra giải pháp.
    - Phát hiện gian lận thi cử ở Sơn La, Hoà Bình.



PHÂN TÍCH NÂNG CAO

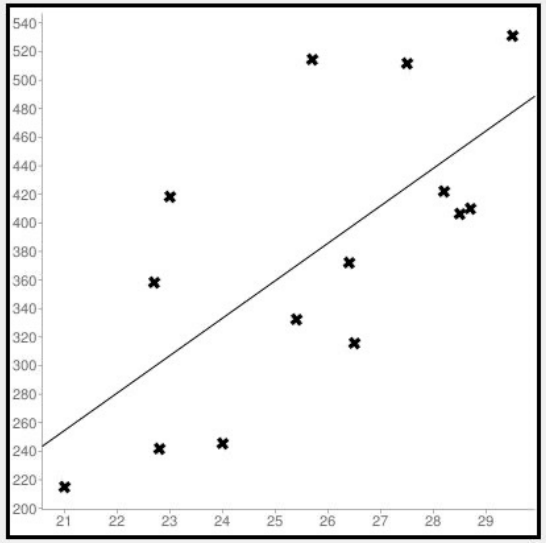
- Phân tích nâng cao gồm các giải thuật phân tích tinh vi các dữ liệu phức tạp có cấu trúc hoặc phi cấu trúc.
- Khác với phân tích cơ bản được các nhà thống kê và các nhà toán học áp dụng để sử dụng, phân tích nâng cao sử dụng các kỹ thuật và mô hình phức tạp hơn.
- Trong nhưng năm gần đây các kỹ thuật phân tích dữ liệu được áp dụng và nghiên cứu ngày càng chuyên sâu hơn, tiêu biểu như: Khai phá dữ liệu, các mô hình dự báo và các giải thuật phân tích thống kê.





PHÂN TÍCH NÂNG CAO

- **KHAI PHÁ DỮ LIỆU:**
  - khám phá và phân tích một lượng lớn dữ liệu để tìm các mẫu có ích trong dữ liệu.
  - Khai phá dữ liệu sử dụng các kỹ thuật thống kê, máy học và quản trị CSDL.
  - Mục tiêu chính của khai phá dữ liệu là phân loại và dự đoán
    - ***Phân loại***: là cố gắng sắp xếp dữ liệu thành các nhóm.
    - ***Dự đoán:*** là tìm kiếm giá trị của 1 biến liên tục dựa vào các biến độc lập.
  - Các giải thuật tiêu biểu:
    - Luật kết hợp.
    - Cây phân loại.
    - Hồi quy logistic.
    - Mạng nơron.
    - Kỹ thuật gom cụm và k-láng giềng gần nhất.



PHÂN TÍCH NÂNG CAO

- **MÔ HÌNH DỰ BÁO:**
  - Là một trong những mô hình phổ biến nhất trong các trường hợp sử dụng phân tích dữ liệu nâng cao.
  - Mô hình dự báo sử dụng kỹ thuật thống kê hoặc khai phá dữ liệu bao gồm các kỹ thuật và giải thuật có thể áp dụng trên dữ liệu có cấu trúc và phi cấu trúc để xác định kết quả trong tương lai.
  - Được sử dụng trong các hệ thống như: ngân hàng, bảo hiểm, quảng cáo, viễn thông ...
- **CÁC GIẢI THUẬT PHÂN TÍCH THỐNG KÊ KHÁC:**
  - Bao gồm gợi ý, tối ưu hoá, phân tích cụm hoặc phân tích vi mô hoặc phân tích mối quan hệ.
  - Phân tích nâng cao có thể không yêu cầu dữ liệu lớn, nhưng với dữ liệu lớn có thể sẽ cho ra kết quả hợp lý hơn

PHÂN TÍCH HOẠT ĐỘNG VÀ HIỆU QUẢ KINH DOANH

- **Phân tích dữ liệu là một phần của hoạt động kinh doanh.**
  - Các nhà thống kê tại một cty bảo hiểm có thể xây dựng 1 mô hình dữ báo khả năng khiếu nại.
  - Kết quả giao dịch của các khách hàng tại các ngân hàng có thể đưa ra dự báo về khả năng thanh toán các khoản nợ của các khách hàng đó.
- **Phân tích hiệu quả kinh doanh được sử dụng để tối ưu hoá hoạt động của một doanh nghiệp nhằm tạo ra quyết định tốt hơn và giúp nâng cao doanh thu.**
  - Sử dụng dữ liệu là các món hàng trong các đơn hàng mà siêu thị có thể đưa ra phương pháp xếp các kệ hàng sao cho khách hàng dễ tìm nhất.
  - Dữ liệu kinh doanh của các chuỗi cửa hàng có thể gợi ý cho các nhà quản trị các sản phẩm cần đưa xuống chi nhánh nào để đạt doanh số tối đa.

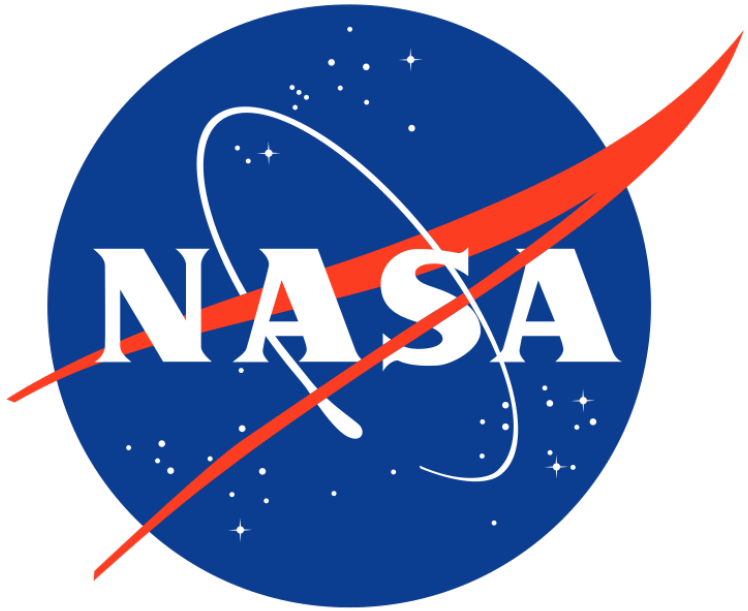
GIẢI THUẬT PHÂN TÍCH DỮ LIỆU

- Các giải thuật phân tích dữ liệu cần được thiết kế theo nhu cầu của công ty
- Ngày nay các giải thuật phân tích cần phải được thiết kế sao cho có thể tương thích được với các loại dữ liệu lớn (hàng tỷ dòng dữ liệu)
- Cần phân tích dư thừa để có thể giữ lại các dữ liệu có giá trị.

CƠ SỞ HẠ TẦNG ĐỂ PHÂN TÍCH BIG DATA

- **Tích hợp công nghệ:** cơ sở hạ tầng cần tích hợp với công nghệ dữ liệu lớn với công nghệ truyền thống để có thể xử lý tất cả các loại dữ liệu lớn
- **Lưu trữ lượng lớn dữ liệu khác nhau:** hệ thống Hadoop là hệ thống mà doanh nghiệp cần để quản lý và xử lý lượng lớn dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc
- **Dữ liệu quy trình đang chuyển động:** cần có khả năng tính toán luồng dữ liệu (data stream) để xử lý dữ liệu đang chuyển động liên tục từ nhiều nguồn khác nhau trong thời gian thực để hỗ trợ ra quyết định.
- **Kho dữ liệu:** cần có giải pháp tối ưu hoá để lưu trữ và quản lý lượng dữ liệu có chiều hướng ngày càng tăng về số lượng.

CÁC ỨNG DỤNG PHÂN TÍCH DỮ LIỆU LỚN



NỘI DUNG BÀI HỌC

- CÁC HÌNH THỨC PHÂN TÍCH DỮ LIỆU
- PHÂN TÍCH CƠ BẢN
- PHÂN TÍCH NÂNG CAO
- GIẢI THUẬT PHÂN TÍCH DỮ LIỆU
- CƠ SỞ HẠ TẦNG ĐỂ PHÂN TÍCH DỮ LIỆU LỚN
- ỨNG DỤNG PHÂN TÍCH DỮ LIỆU LớN