ĐẠI HỌC KINH DOANH VÀ CÔNG NGHỆ HÀ NỘI

KHOA CNTT

Phân tích dữ liệu lớn

CHƯƠNG 5: Phân tích dữ liệu lớn

- **Hiểu về phân tích dữ liệu lớn**
- **Các kiểu phân tích Dữ liệu lớn**
  - Descriptive analysis (Phân tích miêu tả)
  - Diagnostic Analysis (Phân tích chẩn đoán)
  - Predictive Analysis (Phân tích dự đoán)
  - Prescriptive Analysis (Phân tích đề xuất)
- **Công cụ phân tích MapReduce**
- **Ứng dụng hiệu quả với một số lĩnh vực**
  - Ngành Bán lẻ
  - Nghành Ngân hàng
  - Nghành Chế tạo
  - Nghành Chăm sóc sức khỏe
  - Ngành Năng lượng
  - …

Phân tích dữ liệu lớn

Phân tích dữ liệu lớn là một quy trình hoàn chỉnh kiểm tra các tập hợp dữ liệu lớn thông qua các công cụ và quy trình khác nhau để khám phá các mẫu chưa biết, mối tương quan ẩn, xu hướng có ý nghĩa và các thông tin chi tiết khác để đưa ra quyết định dựa trên dữ liệu nhằm theo đuổi mục tiêu tốt hơn.

- Ra quyết định nhanh hơn, tốt hơn
- Giảm chi phí và tăng hiệu quả hoạt động
- Cải tiến theo định hướng dữ liệu cho thị trường

Lợi ích của phân tích dữ liệu lớn

Ra quyết định nhanh hơn, tốt hơn

Các doanh nghiệp có thể truy cập một lượng lớn dữ liệu và phân tích nhiều nguồn dữ liệu khác nhau để có được những hiểu biết mới và thực hiện hành động. Bắt đầu với quy mô nhỏ và quy mô để xử lý dữ liệu từ các bản ghi lịch sử và trong thời gian thực.

Quyết định về vị trí của hàng của Starbucks

Sau khi hàng trăm cửa hàng bị buộc phải đóng cửa vào năm 2008, giám đốc Starbucks Howard Schultz đã cam kết công ty sẽ sử dụng các phân tích định lượng hơn trong việc xác định các địa điểm mở cửa hàng trong tương lai.

Starbucks hiện là đối tác với một công ty phân tích vị trí địa điểm để xác định các địa điểm mở cửa hàng lý tưởng cho Starbucks. Điều được phân tích dựa trên tuổi tác, giới tính và hành vi tham gia giao thông của người dùng. Ngoài ra, Starbucks cũng cân nhắc các dữ liệu thông tin được gửi từ các chi nhánh bộ phận trước khi ra quyết định. Với việc sử dụng tất cả các dữ liệu này, Starbucks luôn định lượng khả năng thành công của một vị trí cửa hàng trước khi chính thức bắt tay vào đầu tư mở rộng vị trí đó.

Lợi ích của phân tích dữ liệu lớn

Giảm chi phí và tăng hiệu quả hoạt động

Các công cụ lưu trữ và xử lý dữ liệu linh hoạt có thể giúp các tổ chức tiết kiệm chi phí trong việc lưu trữ và phân tích một lượng lớn dữ liệu. Khám phá các mẫu và thông tin chi tiết giúp bạn xác định hoạt động kinh doanh hiệu quả hơn.

Gia tăng doanh thu ở Amazon

Amazon sử dụng dữ liệu về sản phẩm khách hàng vừa mua và các dữ liệu tìm kiếm của khách hàng để đưa ra các sản phẩm khuyến nghị mới cho khách. Thay vì giới thiệu sản phẩm một cách ngẫu nhiên, Amazon sử dụng các thuật toán phân tích để có thể cho ra kết quả kiến nghị sát nhất với nhu cầu của khách hàng. Vào năm 2017, Mckinsey đã chỉ ra 35% doanh thu của Amazon đến từ việc kiến nghị các sản phẩm mới cho khách hàng trên nền tảng này

Lợi ích của phân tích dữ liệu lớn

Cải tiến theo định hướng dữ liệu cho thị trường

Phân tích dữ liệu từ cảm biến, thiết bị, video, nhật ký, ứng dụng giao dịch, web và mạng xã hội giúp tổ chức hoạt động theo hướng dữ liệu. Đo lường nhu cầu của khách hàng và rủi ro tiềm ẩn và tạo ra các sản phẩm và dịch vụ mới.

Coca-Cola ứng dụng Big Data để thu hút và duy trì lượng khách hàng.

Trong năm 2015, hãng đồ uống này đã cố gắng tăng cường chiến lược dữ liệu của mình bằng cách xây dựng chương trình khách hàng thân thiết trên nền tảng kỹ thuật số. Big Data duy trì số lượng khách hàng ổn định tại Coca-Cola. Hãng đã theo đuổi 2 chiến lược: tiếp cận dựa trên dữ liệu khách hàng để phát triển sản phẩm mới và xác định kế hoạch phân phối, tiếp thị dựa trên các dữ liệu nghiên cứu thị trường.

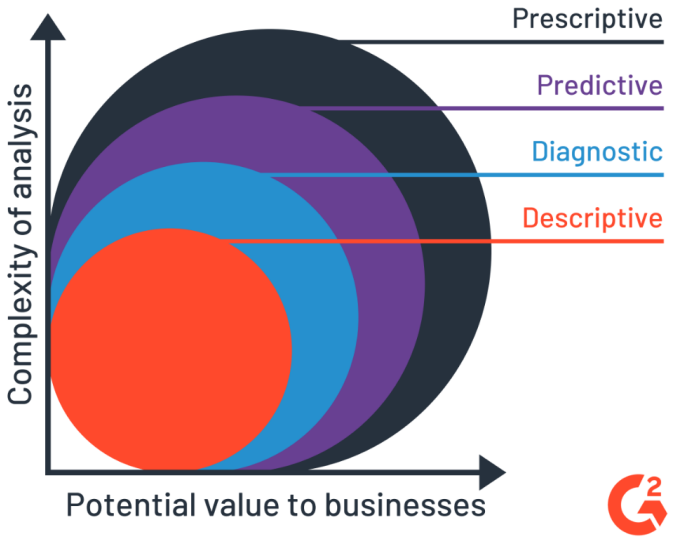
Dòng sản phẩm Coca-Cola vanilla là ví dụ điển hình của chiến lược này - được sáng chế tại Mỹ và nhanh chóng giành được sự quan tâm của công chúng.

Các kiểu phân tích Dữ liệu lớn

Big Data được hiểu là những dữ liệu khổng lồ, là nguồn tài sản thông tin có dung lượng lớn và đa dạng, có vận tốc cao. Tầm quan trọng của dữ liệu lớn không nằm ở lượng dữ liệu thô mà chúng ta có, nó nằm ở việc chúng ta làm gì với những dữ liệu đó.

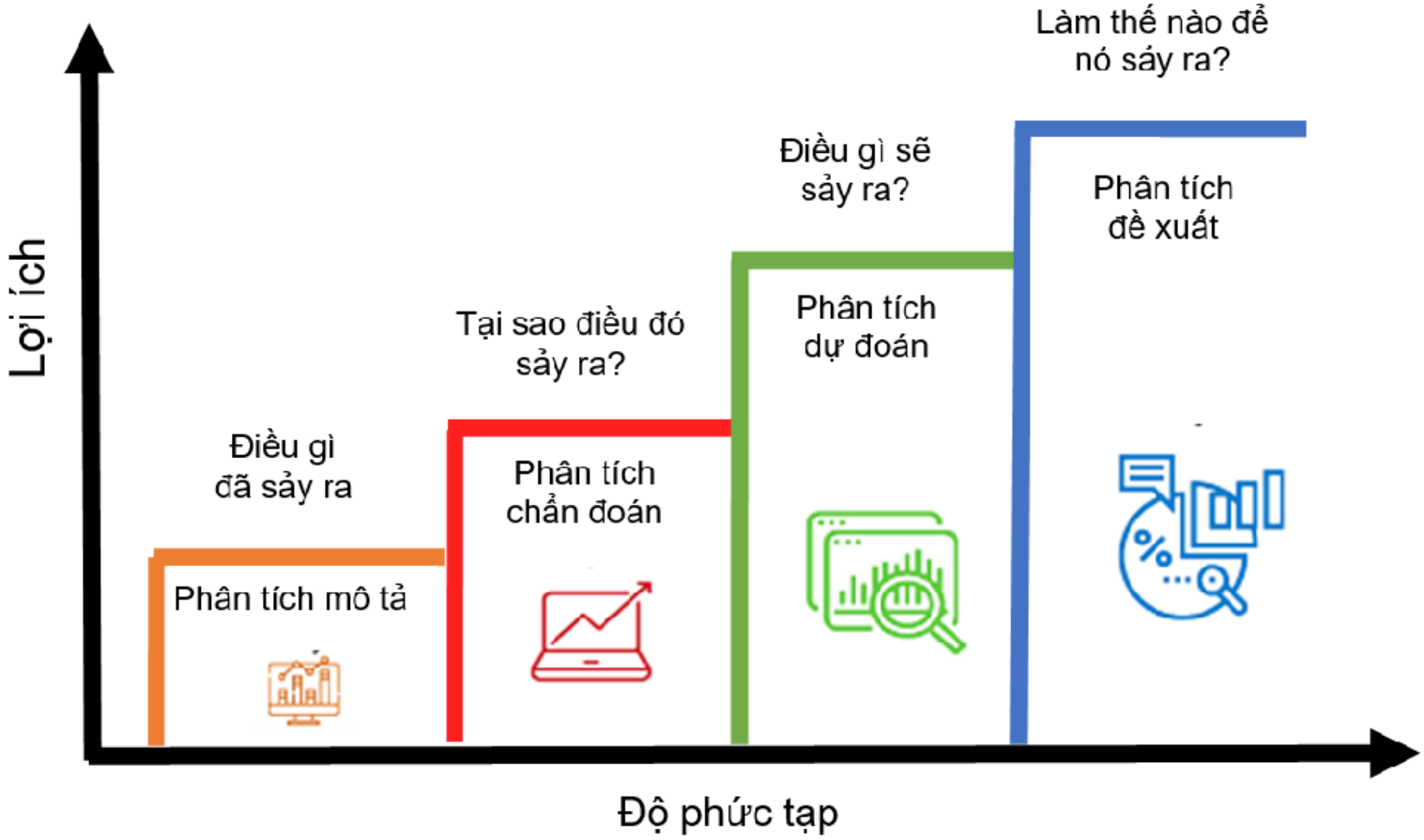
Các loại phân tích Dữ liệu lớn:

- Descriptive analysis (Phân tích mô tả)
- Diagnostic Analysis (Phân tích chẩn đoán)
- Predictive Analysis (Phân tích dự đoán)
- Prescriptive Analysis (Phân tích đề xuất)

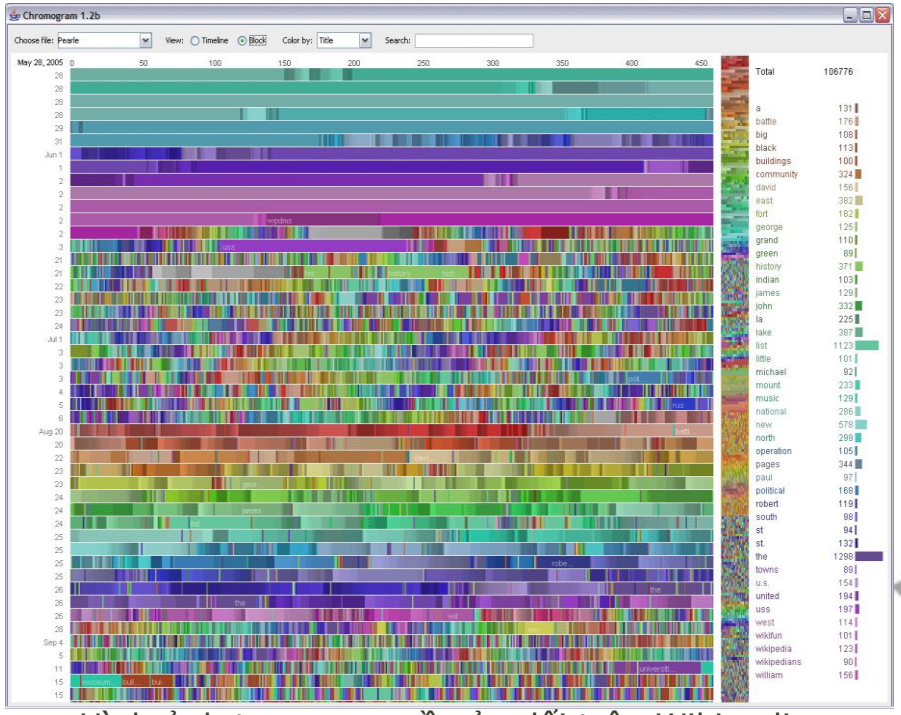


→ Hầu hết các doanh nghiệp, tổ chức sẽ sử dụng nguồn dữ liệu lớn phân tích để tìm ra câu trả lời cho các câu hỏi: giảm chi phí, giảm thời gian, phát triển mới và dịch vụ tối ưu, ra quyết định thông minh...

Phân tích Dữ liệu lớn



Các kiểu phân tích Dữ liệu lớn



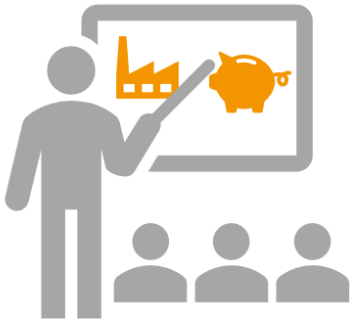
Hình ảnh trực quan về sửa đổi trên Wikipedia hằng ngày được tạo ra bởi IBM.
Với kích cỡ vài terabyte, các văn bản và hình ảnh trên Wikipedia là một ví dụ của dữ liệu lớn.

Mục tiêu của phân tích dữ liệu là biến dữ liệu thành thông tin chi tiết hữu ích. Có bốn loại phân tích dữ liệu chính: mô tả, chẩn đoán, dự đoán, đề xuất.
Chúng khác nhau về mức độ phức tạp và giá trị gia tăng của chúng.

Descriptive analysis - Phân tích miêu tả

- **Phân tích mô tả**

Phân tích mô tả tóm tắt dữ liệu quá khứ thành một biểu mẫu mà mọi người có thể dễ dàng đọc được. Phương pháp này phù hợp với việc xây dựng các báo cáo, chẳng hạn như doanh thu, lợi nhuận, doanh số bán hàng của công ty v.v. Ngoài ra, phân tích mô tả cũng được áp dụng để lập bảng số liệu truyền thông xã hội.



1. Xác định yêu cầu

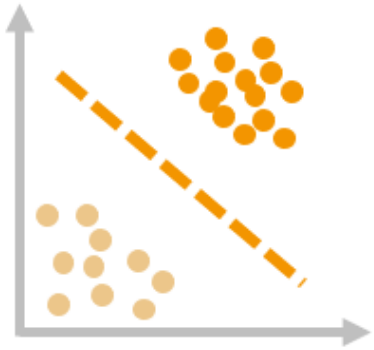
2. Thu thập dữ liệu

3. Mô tả dữ liệu

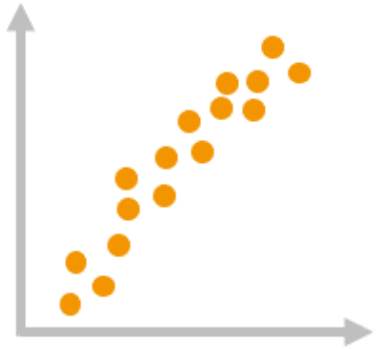
Diagnostic Analysis - Phân tích chẩn đoán

- **Phân tích chẩn đoán**

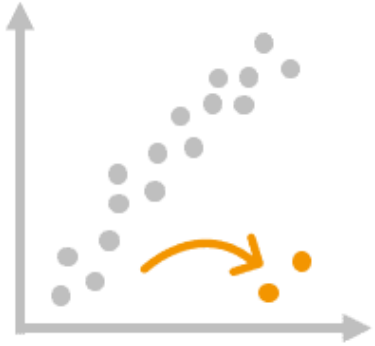
Phân tích chẩn đoán được thực hiện để xác định nguyên nhân gây ra sự cố, sử dụng một số kỹ thuật như phân tích chi tiết, khai phá dữ liệu (data mining) và khôi phục dữ liệu (data recovery). Phân tích chẩn đoán giúp cung cấp cái nhìn sâu sắc về một vấn đề cụ thể.



Phân tích mẫu
Xác định nhóm



Phân tích tương quan
Xác định xu hướng

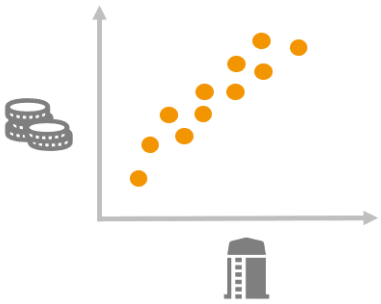


Phân tích bất thường
Xác định ngoại lệ

Diagnostic Analysis - Phân tích chẩn đoán

- **Phân tích chẩn đoán**

Phân tích chẩn đoán trả lời các câu hỏi về lý do tại sao mọi thứ lại xảy ra. Nó lấy những phát hiện từ phân tích mô tả và điều tra nguyên nhân gốc rễ. Do đó, sự khác biệt chính so với giai đoạn trước là chúng tôi quan tâm đến cách dữ liệu được kết nối với nhau.



1. Lấy mẫu

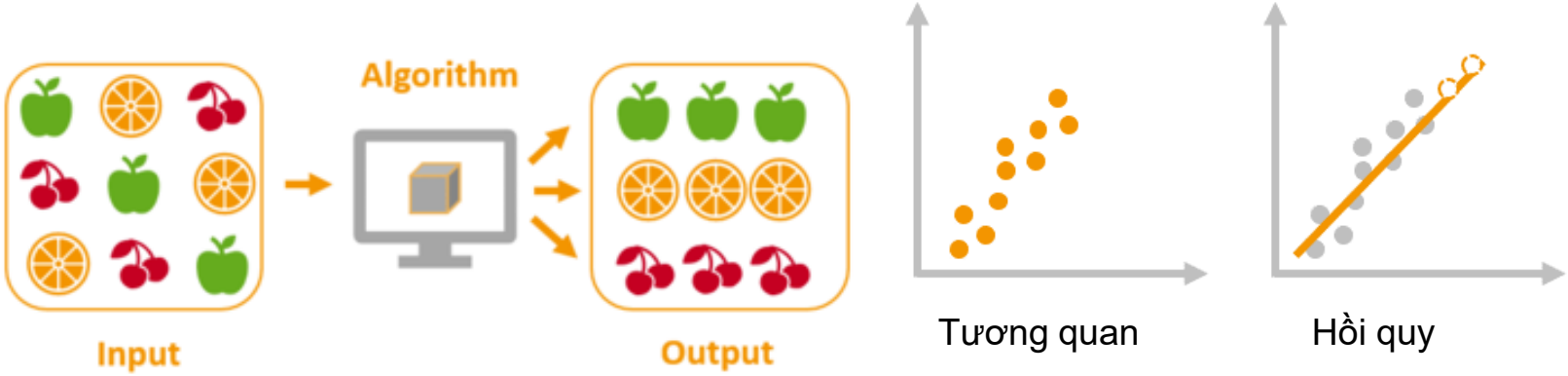
2. Tương quan phân bổ

3. Sự bất thường

Predictive Analysis - Phân tích dự đoán

- **Phân tích dự đoán**

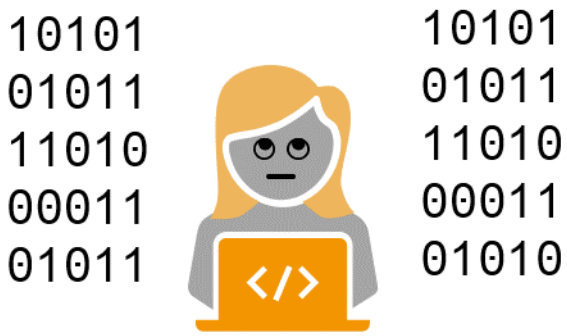
Phân tích dự đoán sử dụng **Data mining, AI và Machine learning** để phân tích dữ liệu trong quá khứ và hiện tại để đưa ra dự đoán về tương lai. Phương pháp này được ứng dụng nhằm dự đoán xu hướng của khách hàng, xu hướng thị trường, v.v.



Predictive Analysis - Phân tích dự đoán

- **Phân tích dự đoán**

Phân tích dự đoán giúp trả lời các câu hỏi về những gì có thể sẽ xảy ra trong tương lai. Vì lý do này, một mô hình dự đoán được xây dựng và liên tục tinh chỉnh. Cơ sở cho mô hình là dữ liệu lịch sử để xác định xu hướng và dự báo kết quả tiềm năng trong tương lai.



1. Xây dựng mô hình Merchine Learning – AI

2. Đưa ra dự đoán tối ưu

Prescriptive Analysis - Phân tích đề xuất

- **Phân tích đề xuất**

  Loại phân tích này đề xuất giải pháp cho một vấn đề cụ thể. Phân tích đề xuất được sử dụng kết hợp với cả phân tích mô tả và dự đoán. Phương pháp này phần lớn dựa vào AI và học máy.

- Biến các dự đoán thành dữ liệu hỗ trợ quyết định bằng cách chủ động đưa ra hành động tốt nhất tiếp theo cho hoạt động.
- Chính điều này cho phép các doanh nghiệp đưa ra lựa chọn tối ưu hơn khi đối mặt với sự không chắc chắn.

Prescriptive Analysis - Phân tích đề xuất

- **Phân tích đề xuất**

Loại phân tích này đề xuất giải pháp cho một vấn đề cụ thể. Phân tích đề xuất được sử dụng kết hợp với cả phân tích mô tả và dự đoán. Phương pháp này phần lớn dựa vào AI và học máy.

1. Ứng dụng kết quả hỗ trợ quyết định

2. Đạt kết quả tối ưu quy trình

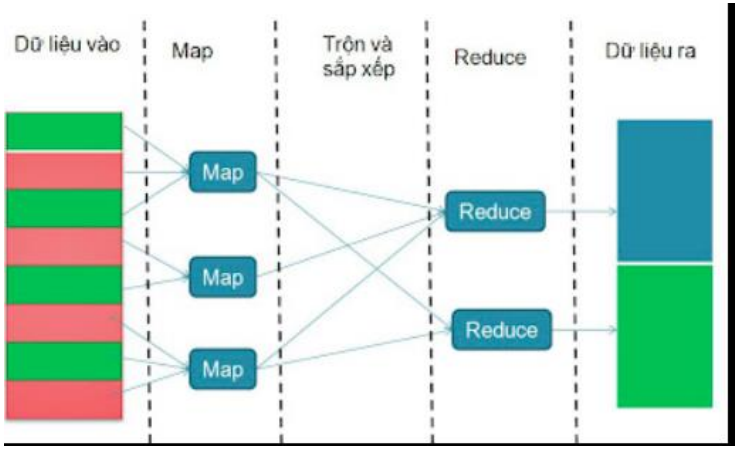
Công cụ phân tích MapReduce

- MapReduce là mô hình được thiết kế độc quyền bởi Google, nó có khả năng lập trình xử lý các tập dữ liệu lớn song song và phân tán thuật toán trên 1 cụm máy tính.
- MapReduce gồm một single master (máy chủ) JobTracker và các slave (máy trạm) TaskTracker trên mỗi cluster-node. Master có nhiệm vụ quản lý tài nguyên, theo dõi quá trình tiêu thụ tài nguyên và lập lịch quản lý các tác vụ trên các máy trạm, theo dõi chúng và thực thi lại các tác vụ bị lỗi. Những máy slave TaskTracker thực thi các tác vụ được master chỉ định và cung cấp thông tin trạng thái tác vụ (task-status) để master theo dõi.
- Áp dụng mô hình MapReduce chạy trên lượng lớn các machine cỡ hàng ngàn machine và data lên đến mức Terabytes.

Công cụ phân tích MapReduce

Hoạt động của MapReduce có thể được tóm tắt như sau:

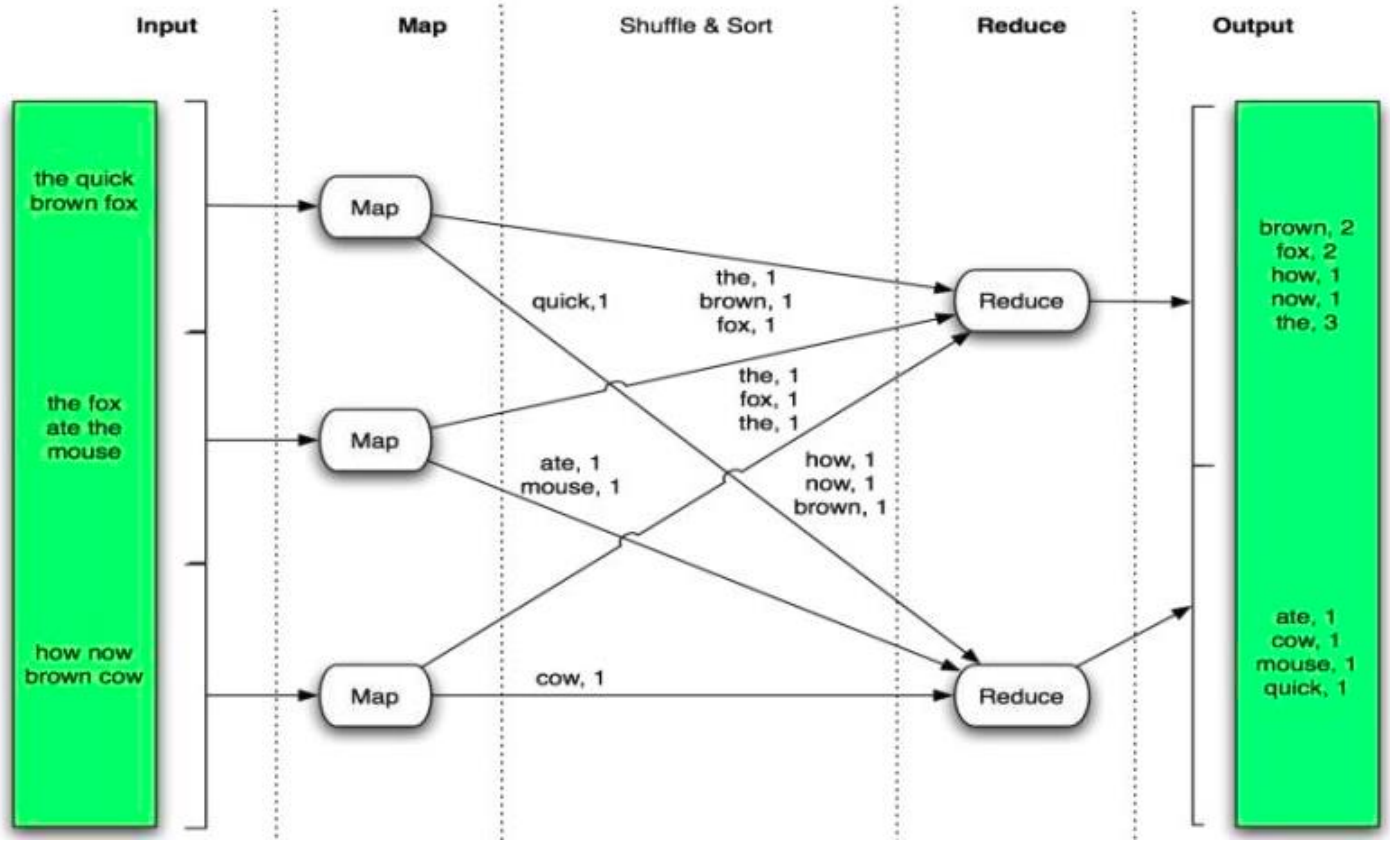
- Đọc dữ liệu đầu vào
- Xử lý dữ liệu đầu vào (thực hiện hàm map)
- Sắp xếp và trộn các kết quả thu được từ các máy tính phân tán thích hợp nhất.
- Tổng hợp các kết quả trung gian thu được (thực hiện hàm reduce)
- Đưa ra kết quả cuối cùng.



Các công việc sử dụng Mapreduce:

- Thống kê số từ khóa xuất hiện trong các documents.
- Thống kê số documents có chứa từ khóa.
- Thống kê số câu match với pattern trong các documents.
- Thống kê số URLs xuất hiện trong các web pages.
- Thống kê số lượt truy cập các URLs.
- Thống kê số từ khóa trên các hostnames.
- Distributed Sort.

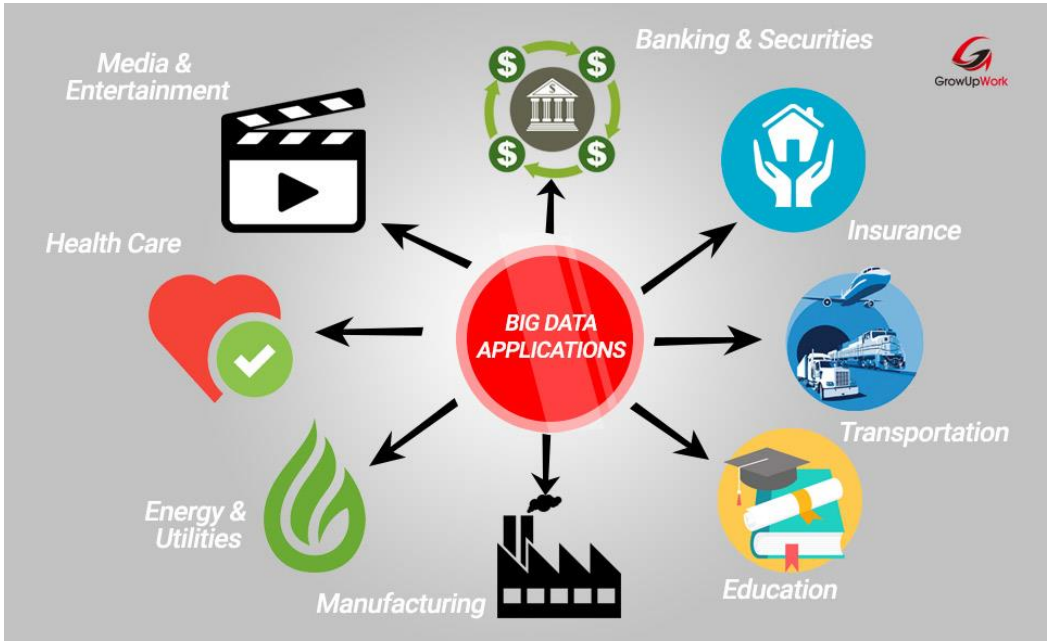
MapReduce với bài toán wordcount



Ứng dụng hiệu quả với một số lĩnh vực

Các doanh nghiệp sử dụng Big data nắm giữ lợi thế cạnh tranh tiềm năng hơn so với các doanh nghiệp khác về tốc độ đưa ra các ý tưởng quyết định kinh doanh hiệu quả, miễn là họ biết cách tận dụng những dữ liệu mình đã có.

- Bán lẻ
- Ngân hàng
- Sức khỏe
- Chế tạo
- Năng lượng



Ứng dụng Lĩnh vực Bán lẻ

Ứng dụng các công nghệ mới như AR, Big Data, AI, IoT,... doanh nghiệp có thể cung cấp dịch vụ chăm sóc khách hàng với những trải nghiệm và giá trị khác biệt, tối ưu hoá chiến lược quản lý và chuỗi cung ứng, từ đó gia tăng doanh số hoạt động.

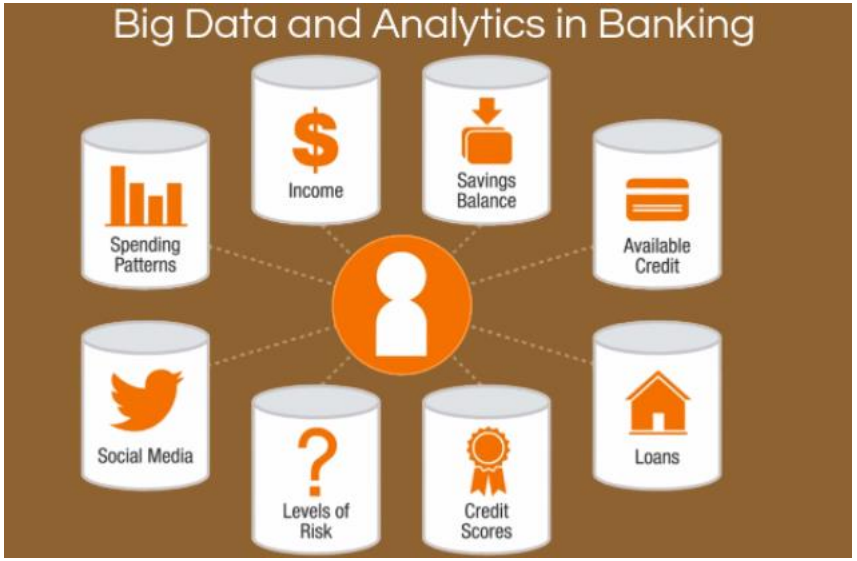
- Dự đoán hành vi mua sắm của khách hàng
- Hỗ trợ quản lý chuỗi cung ứng
- Phân tích hành trình của khách hàng
- Xây dựng mô hình chi tiêu cho từng khách hàng

→ Đối với các doanh nghiệp bán lẻ ngày càng phát triển và cạnh tranh khốc liệt, sử dụng dữ liệu lớn giúp doanh nghiệp nhanh chóng thay đổi được mặt hàng, cách thức giao vận và nắm bắt tâm lí khách hàng, thích ứng nhanh chóng và kịp thời với các hoàn cảnh xã hội thay đổi liên tục như hiện nay.

Ứng dụng ngành Ngân hàng

Trong bối cảnh ngân hàng kỹ thuật số bùng nổ, các ngân hàng cần xây dựng cơ sở hạ tầng dữ liệu lớn như: Định danh số, nhận biết khách hàng điện tử, an ninh mạng, công nghệ đám mây,...

- Big Data đã tham gia vào rất nhiều công đoạn của ngân hàng, từ thu tiền mặt, giao dịch điện tử đến quản lý tài chính. Cách thức ứng dụng Big Data.
- Khả năng phân tích dữ liệu khổng lồ và nhanh chóng của Big data cho phép các ngân hàng nắm bắt các gian lận ngay khi xảy ra, đảm bảo không có giao dịch trái phép nào được thực hiện nhờ vào các thuật toán phân tích dữ liệu và học máy → đảm bảo lợi ích của khách hàng và chính ngân hàng.



Ứng dụng Chăm sóc sức khỏe

Khi nói đến chăm sóc sức khỏe, mọi thứ cần phải được thực hiện nhanh chóng, chính xác – và, trong một số trường hợp, phải đủ minh bạch để đáp ứng các quy định nghiêm ngặt của ngành.

- Khi dữ liệu lớn được quản lý hiệu quả, các nhà cung cấp dịch vụ chăm sóc sức khỏe có thể rút ngắn thời gian khám bệnh, và tăng tính chính xác của những chẩn đoán thông qua việc khai thác những thông tin cần thiết qua nền tảng dữ liệu lớn.

- Các hoạt động khám - chữa bệnh có thể được cải thiện vượt bậc khi ứng dụng Big Data trong lĩnh vực này có thể kể tới như: Hồ sơ bệnh án, phác đồ điều trị, thông tin kê đơn,…

Ứng dụng Lĩnh vực Chế tạo

Có rất nhiều công nghệ có liên quan như Internet of Things, Robotics, AI, máy học và những công nghệ khác, nhưng xương sống của mỗi công nghệ này đều dựa trên Big Data Analytics.

- Big Data cung cấp cái nhìn sâu sắc giúp các nhà sản xuất có thể thay đổi quy trình sản xuất với mục tiêu tăng chất lượng và sản lượng trong khi giảm thiểu chất thải.
- Sử dụng Phân tích dữ liệu lớn, các nhà chế tạo có thể năm bắt được nhu cầu thị trường, cải thiện năng suất, nâng cao chất lượng, tối ưu hóa chuỗi cung ứng và quy trình hậu cần cũng như xây dựng nguyên mẫu trước khi ra mắt sản phẩm để hiểu và tiếp cận thị trường một cách linh hoạt hơn.
- Trong tất cả các bước này, Phân tích dữ liệu lớn đóng vai trò quan trọng cung cấp các số liệu, thông kê trên thực tế nhu cầu thị trường cho các nhà sản xuất.

Ứng dụng Lĩnh vực Năng lượng

Lĩnh vực năng lượng trong thời đại mới ứng dụng phổ biến các hệ thống Big Data Analytics cho việc triển khai, khai thác, định giá...

- Thị trường nhiên liệu hóa thạch rất dễ biến động đối với những ảnh hưởng về chính trị, tài chính thế giới nên việc sử dụng Phân tích dữ liệu là giải pháp tối ưu để tìm hiểu giá của một thùng dầu sẽ là bao nhiêu, sản lượng nên là bao nhiêu và liệu một giếng dầu có sinh lời hay không.
- Phân tích dữ liệu lớn từ IoT và cảm biến máy móc cũng được triển khai để tìm ra lỗi thiết bị, triển khai bảo trì dự đoán và sử dụng tối ưu các nguồn lực để giảm chi phí trong các hoạt động sản xuất phức tạp: mỏ khai thác, giếng khoan dầu...
- ...


Cảm ơn các bạn đã
lắng nghe!