



TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU LỚN

Giảng viên: Nguyễn Tu Trung, Trần Mạnh Tuấn
BM HTTT, Khoa CNTT, Trường ĐH Thủy Lợi

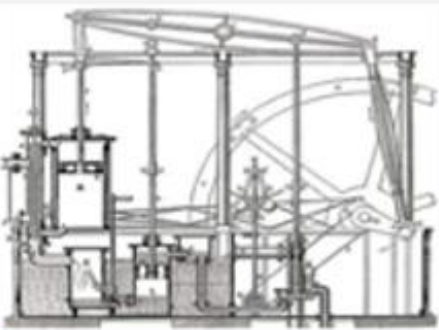
Hà Nội, 2019

Nội dung

- ❖ Cách mạng công nghiệp lần thứ 4
- ❖ Công nghệ số
- ❖ Dữ liệu lớn là gì
- ❖ Dữ liệu lớn đến từ đâu?
- ❖ Đặc trưng cơ bản của dữ liệu lớn
- ❖ Ứng dụng của dữ liệu lớn
- ❖ Tiếp cận dữ liệu lớn
- ❖ Công nghệ chính trong xử lý dữ liệu lớn

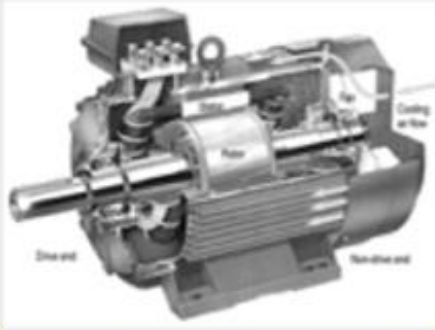
Cách mạng công nghiệp lần thứ 4

- ❖ Đặc trưng của một cuộc cách mạng công nghiệp:
 - ❖ Có đột phá của khoa học và công nghệ
 - ❖ Tạo ra sự thay đổi về bản chất của sản xuất
- ❖ Các cuộc cách mạng công nghiệp



Cách mạng công nghiệp lần thứ nhất về **sản xuất cơ khí** với máy móc dựa vào **động cơ hơi nước**.

Cuối thế kỷ 18 đầu thế kỷ 19



Cách mạng công nghiệp lần thứ hai về **sản xuất hàng loạt** với máy móc dựa vào **năng lượng điện**.

Cuối thế kỷ 19 đầu thế kỷ 20



Cách mạng công nghiệp lần thứ ba về **sản xuất tự động** với **máy tính, điện tử và cách mạng số hóa**.

Từ thập kỷ 70 của thế kỷ 20



Cách mạng công nghiệp lần thứ tư về **sản xuất thông minh** nhờ các **đột phá của công nghệ số**.

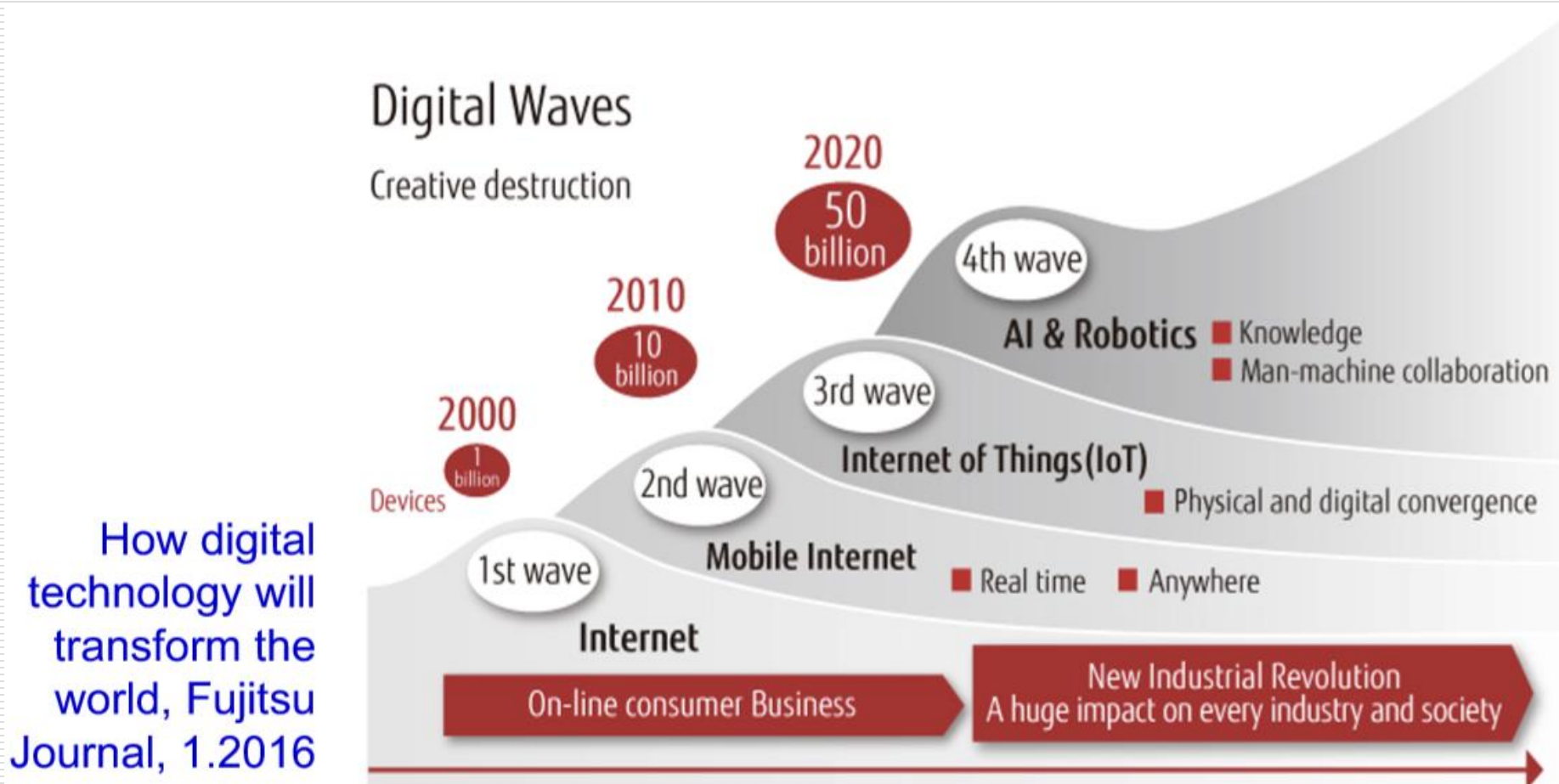
Bắt đầu từ bây giờ

Cách mạng công nghiệp lần thứ 4

- ❖ Cách mạng công nghiệp lần 4:
 - ❖ Sản xuất thông minh dựa trên tiến bộ của công nghệ thông tin, công nghệ sinh học, công nghệ nano...
 - ❖ Với nền tảng là các đột phá của công nghệ số trên Hệ kết nối không gian số-thực thể (cyber-physical systems)
- ❖ Cách mạng số hoá:
 - ❖ ‘Phiên bản số’ các thực thể: Biểu diễn các thực thể bằng ‘0’ và ‘1’ trên máy tính (digitalization)
 - ❖ Thí dụ: bệnh án điện tử...
- ❖ Hệ kết nối không gian số-thực thể (cyber-physical system): hệ kết nối các thực thể và ‘phiên bản số’ của chúng
- ❖ => Thay đổi phương thức sản xuất:
 - ❖ Hành động trong thế giới các thực thể
 - ❖ Tính toán, điều khiển trên không gian số

Công nghệ số

- ❖ Số hoá (thí dụ máy ảnh, in ấn, truyền hình...)
- ❖ Xử lý dữ liệu được số hoá



[illegible]

Dữ liệu lớn là gì

❖ Theo wikipedia:

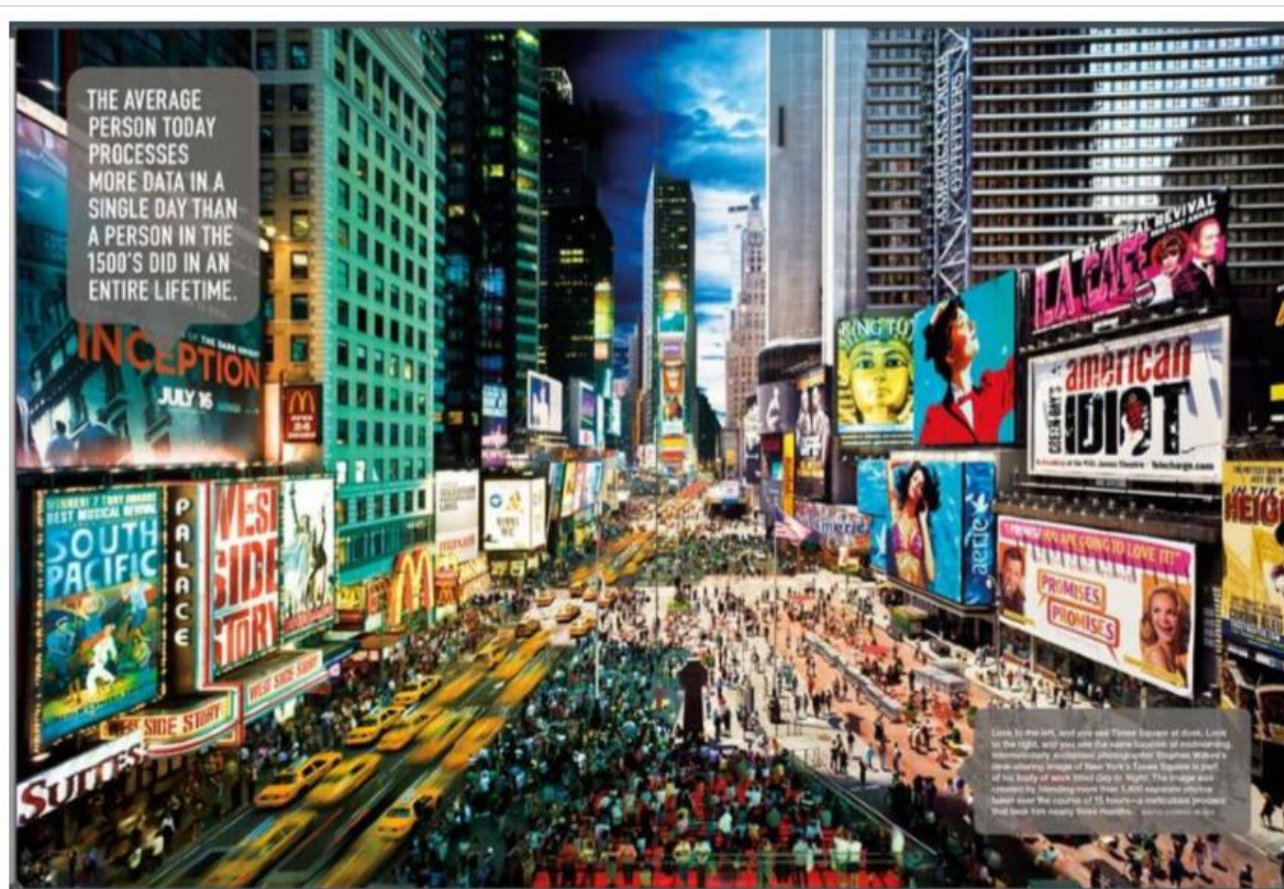
- ❖ Dữ liệu lớn (Big data) là một thuật ngữ chỉ bộ dữ liệu lớn hoặc phức tạp mà các phương pháp truyền thống không đủ các ứng dụng để xử lý dữ liệu này

❖ Theo Gartner:








- ❖ Dữ liệu lớn là những nguồn thông tin có đặc điểm chung khối lượng lớn, tốc độ nhanh và dữ liệu 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

Dữ liệu lớn đến từ đâu?

❖ Đến từ rất nhiều nguồn khác nhau



Every 60 seconds

-  98,000+ tweets
-  695,000 status updates
-  11 million instant messages
-  698,445 Google searches
-  168 million+ emails sent
-  1,820TB of data created
-  217 new mobile web users

Dữ liệu lớn đến từ đâu?

- ❖ “Chỉ trong ngày đầu tiên một em bé sinh ra đời, số lượng dữ liệu thu thập được tương đương với 70 lần thông tin trong Thư viện Quốc hội Mỹ (The Library of Congress)”



DURING THE FIRST DAY OF A BABY'S LIFE, THE AMOUNT OF DATA GENERATED BY HUMANITY IS EQUIVALENT TO 70 TIMES THE INFORMATION CONTAINED IN THE LIBRARY OF CONGRESS. ▼

Dữ liệu lớn đến từ đâu?



- Nhấp chuột
- Mua hàng
- Transactions
- Networks log
- ...
- Everything online
- ~ 8 hour / day

Dữ liệu từ mạng xã hội

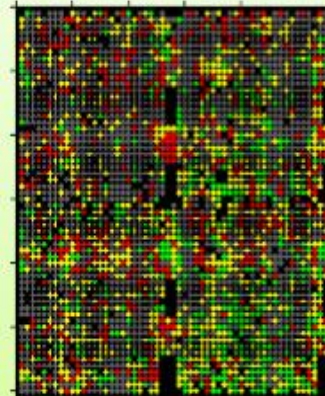


BIG DATA

Kết nối vạn vật và thiết bị thông minh



Dữ liệu từ nghiên cứu khoa học



Dữ liệu từ sinh học
(gene expression)
Nghiên cứu vũ trụ
Nông nghiệp

Dữ liệu lớn đến từ đâu?

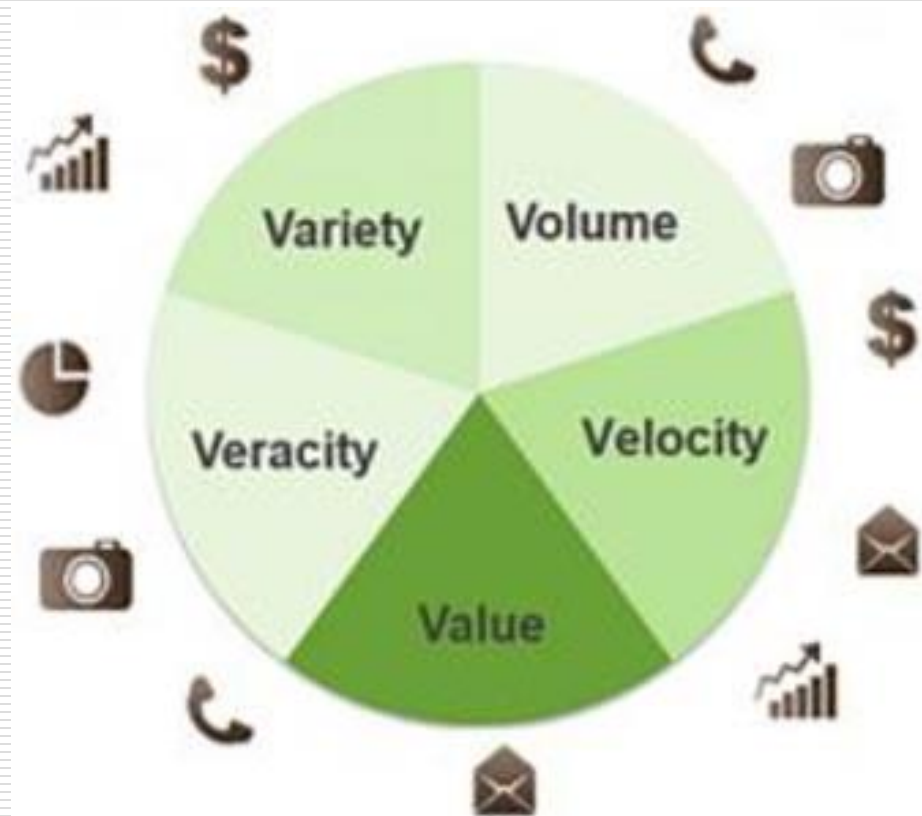
- ❖ Dữ liệu lớn được hình thành chủ yếu từ 6 nguồn:
- ❖ (1) Dữ liệu hành chính (phát sinh từ chương trình của một tổ chức, có thể là chính phủ hay phi chính phủ)
 - ❖ Ví dụ: hồ sơ y tế điện tử ở bệnh viện, hồ sơ bảo hiểm, hồ sơ ngân hàng...;
- ❖ (2) Dữ liệu từ hoạt động thương mại (phát sinh từ các giao dịch giữa hai thực thể)
 - ❖ Ví dụ: các giao dịch thẻ tín dụng, giao dịch trên mạng, bao gồm cả các giao dịch từ các thiết bị di động;
- ❖ (3) Dữ liệu từ các thiết bị cảm biến như thiết bị chụp hình ảnh vệ tinh, cảm biến đường, cảm biến khí hậu;

Dữ liệu lớn đến từ đâu?

- ❖ Dữ liệu lớn được hình thành chủ yếu từ 6 nguồn:
 - ❖ ...
 - ❖ (4) Dữ liệu từ các thiết bị theo dõi
 - ❖ Ví dụ theo dõi dữ liệu từ điện thoại di động, GPS;
 - ❖ (5) Dữ liệu từ các hành vi
 - ❖ Ví dụ như tìm kiếm trực tuyến (tìm kiếm sản phẩm, dịch vụ hay thông tin khác), đọc các trang mạng trực tuyến...;
 - ❖ (6) Dữ liệu từ các thông tin về ý kiến, quan điểm của các cá nhân, tổ chức, trên các phương tiện thông tin xã hội

Đặc trưng cơ bản của dữ liệu lớn

- ❖ Dữ liệu lớn có 5 đặc trưng cơ bản như sau (mô hình 5Vs về dữ liệu lớn):
- ❖ (1) Khối lượng dữ liệu (Volume)
- ❖ (2) Tốc độ (Velocity)
- ❖ (3) Đa dạng (Variety)
- ❖ (4) Độ tin cậy/chính xác (Veracity)
- ❖ (5) Giá trị (Value)



(1) Khối lượng dữ liệu (Volume)

- ❖ Là đặc điểm tiêu biểu nhất của dữ liệu lớn, khối lượng dữ liệu rất lớn
- ❖ Kích cỡ của Big Data đang từng ngày tăng lên, và tính đến năm 2012 thì nó có thể nằm trong khoảng vài chục terabyte cho đến nhiều petabyte (1 petabyte = 1024 terabyte) chỉ cho một tập hợp dữ liệu
- ❖ Dữ liệu truyền thống chúng ta có thể lưu trữ trên các thiết bị đĩa mềm, đĩa cứng
- ❖ Dữ liệu lớn sẽ sử dụng công nghệ “đám mây” mới có khả năng lưu trữ được dữ liệu lớn

(2) Tốc độ (Velocity)

- ❖ Tốc độ có thể hiểu theo 2 khía cạnh:
 - ❖ (a) Khối lượng dữ liệu gia tăng rất nhanh (mỗi giây có tới 72.9 triệu các yêu cầu truy cập tìm kiếm trên web bán hàng của Amazon)
 - ❖ (b) Xử lý dữ liệu nhanh ở mức thời gian thực (real-time), có nghĩa dữ liệu được xử lý ngay tức thời ngay sau khi chúng phát sinh (tính đến bằng mili giây)
- ❖ Các ứng dụng phổ biến trên lĩnh vực Internet, Tài chính, Ngân hàng, Hàng không, Quân sự, Y tế – Sức khỏe như hiện nay phần lớn dữ liệu lớn được xử lý real-time
- ❖ Công nghệ xử lý dữ liệu lớn ngày một tiên tiến cho phép chúng ta xử lý tức thì trước khi chúng được lưu trữ vào cơ sở dữ liệu

(3) Đa dạng (Variety)

- ❖ Đối với dữ liệu truyền thống chúng ta hay nói đến dữ liệu có cấu trúc
- ❖ Ngày nay hơn 80% dữ liệu được sinh ra là phi cấu trúc (tài liệu, blog, hình ảnh, vi deo, bài hát, dữ liệu từ thiết bị cảm biến vật lý, thiết bị chăm sóc sức khỏe...)
- ❖ Big Data cho phép liên kết và phân tích nhiều dạng dữ liệu khác nhau
- ❖ Ví dụ: với các comments/post của một nhóm người dùng nào đó trên Facebook với thông tin video được chia sẻ từ Youtube và Twitter

(4) Độ tin cậy/chính xác

- ❖ Một trong những tính chất phức tạp nhất của BigData là độ tin cậy/chính xác của dữ liệu
- ❖ Với xu hướng phương tiện truyền thông xã hội (Social Media) và mạng xã hội (Social Network) ngày nay và sự gia tăng mạnh mẽ tính tương tác và chia sẻ của người dùng Mobile làm cho bức tranh xác định về độ tin cậy và chính xác của dữ liệu ngày một khó khăn hơn
- ❖ Bài toán phân tích và loại bỏ dữ liệu thiếu chính xác và nhiễu đang là tính chất quan trọng của BigData

(5) 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 thì việc đầu tiên chúng ta cần phải làm đó là xác định được giá trị của thông tin mang lại như thế nào, khi đó chúng ta mới có quyết định nên triển khai dữ liệu lớn hay không
- ❖ Nếu chúng ta có dữ liệu lớn mà chỉ nhận được 1% lợi ích từ nó, thì không nên đầu tư dữ liệu lớn
- ❖ Kết quả dự báo chính xác thể hiện rõ nét nhất về giá trị của dữ liệu lớn mang lại
- ❖ Ví dụ: Từ khối dữ liệu phát sinh trong quá trình khám, chữa bệnh sẽ giúp dự báo về sức khỏe được chính xác hơn, sẽ giảm được chi phí điều trị và các chi phí liên quan đến y tế

Ứng dụng của dữ liệu lớn

- ❖ Dữ liệu lớn đã được ứng dụng trong nhiều lĩnh vực:
 - ❖ Hoạt động chính trị
 - ❖ Giao thông
 - ❖ Y tế
 - ❖ Thể thao
 - ❖ Tài chính
 - ❖ Thương mại
 - ❖ Thống kê...

Hoạt động chính trị

- ❖ Tổng thống Mỹ Obama đã sử dụng dữ liệu lớn để phục vụ cho cuộc tranh cử Tổng thống của mình
- ❖ Ông xây dựng đội chuyên thu thập thông tin và phân tích dữ liệu thu được



- ❖ Đội ngũ nhân viên này thu thập tất cả thông tin về người dân ở các khu vực, sau đó phân tích và chỉ ra một số thông tin quan trọng về người dân Mỹ như: Thích đọc sách gì, thích mua loại thuốc gì, thích sử dụng phương tiện gì...
- ❖ Thậm chí còn biết được cả thông tin về người đó đã bỏ phiếu tín nhiệm ai ở lần bầu cử trước

Hoạt động chính trị

- ❖ Trên cơ sở những thông tin này, Obama đưa ra kế hoạch vận động phù hợp, giúp ông tái đắc cử Tổng thống lần 2 của nước Mỹ
- ❖ Ngoài ra một số ứng dụng khác trong lĩnh vực chính trị mà dữ liệu lớn được áp dụng như:
 - ❖ Hệ thống chính phủ điện tử
 - ❖ Phân tích quy định và việc tuân thủ quy định
 - ❖ Phân tích, giám sát, theo dõi và phát hiện gian lận, mối đe dọa, an ninh mạng

Giao thông

- ❖ Sử dụng số liệu trong quá khứ để ước lượng các dòng giao thông trong thành phố vào các giờ cao điểm để:
 - ❖ Có những kế hoạch phân luồng giao thông chi tiết, hợp lý giúp giảm thiểu kẹt xe
 - ❖ Đưa ra thông tin cho người tham gia giao thông được biết nếu muốn đi từ nơi này đến nơi khác thì nên đi vào giờ nào để tránh kẹt xe, hoặc đi đường nào là ngắn nhất v.v...
 - ❖ Giúp phân tích định vị người dùng thiết bị di động, ghi nhận chi tiết cuộc gọi trong thời gian thực; và giảm thiểu tình trạng ùn tắc giao thông

Y tế

- ❖ Trong y học các bác sĩ dựa vào số liệu trong các bệnh án để đưa ra dự đoán về nguy cơ mắc bệnh và xu hướng lây lan của bệnh
- ❖ Ví dụ, ứng dụng Google Flu Trend
 - ❖ Dựa trên từ khóa tìm kiếm ở một khu vực nào đó, bộ máy phân tích của google sẽ phân tích và đối chiếu kết quả tìm kiếm đó và đưa ra dự báo về xu hướng dịch cúm tại khu vực đó
 - ❖ => Qua đó cho biết tình hình cúm tại khu vực đó sẽ diễn ra như thế nào để đưa ra các giải pháp phòng tránh
- ❖ Những kết quả mà Google Flu Trend đưa ra, hoàn toàn phù hợp với báo cáo của Tổ chức y tế thế giới WHO về tình hình bệnh cúm tại các khu vực đó

Thể thao

- ❖ Phân tích mô hình hệ thống cấu trúc sơ đồ chiến thuật của đội tuyển Đức đã đưa ra những điểm bất hợp lý trong cấu trúc của đội tuyển Đức
- ❖ => Giúp cho đội tuyển Đức khắc phục được điểm yếu và đã dành được World cup 2014



Tài chính

- ❖ Từ những dữ liệu chính xác, kịp thời thu thập được thông qua các giao dịch của khách hàng => tiến hành phân tích, xếp hạng và quản lý các rủi ro trong đầu tư tài chính, tín dụng

Thương mại

- ❖ Trong thương mại dữ liệu lớn giúp cho chúng ta thực hiện được một số công việc sau:
 - ❖ Phân khúc thị trường và khách hàng
 - ❖ Phân tích hành vi khách hàng tại cửa hàng
 - ❖ Tiếp thị trên nền tảng định vị
 - ❖ Phân tích tiếp thị chéo kênh, tiếp thị đa kênh
 - ❖ Quản lý các chiến dịch tiếp thị và khách hàng thân thiết
 - ❖ So sánh giá
 - ❖ Phân tích và quản lý chuỗi cung ứng
 - ❖ Phân tích hành vi, thói quen người tiêu dùng

Thống kê

- ❖ Nhận thấy những lợi ích to lớn và thách thức của Bigdata đối với thống kê nhà nước, Ủy ban Thống kê Liên hợp quốc cũng như các tổ chức thống kê khu vực và Cơ quan thống kê quốc gia của nhiều nước đã triển khai hàng loạt các hoạt động về Bigdata như:
 - ❖ Hàn Quốc sử dụng ảnh vệ tinh để thống kê nông nghiệp và một số lĩnh vực khác
 - ❖ Australia sử dụng ảnh vệ tinh để thống kê diện tích đất nông nghiệp và năng suất
 - ❖ Italia sử dụng dữ liệu điện thoại di động để thống kê di cư
 - ❖ Bhutan dùng thiết bị di động để tính toán chỉ số giá tiêu dùng
 - ❖ Estonia dùng điện thoại di động định vị vệ tinh để thống kê du lịch
 - ❖ EuroStat sử dụng dữ liệu về sử dụng điện thoại di động để thống kê du lịch

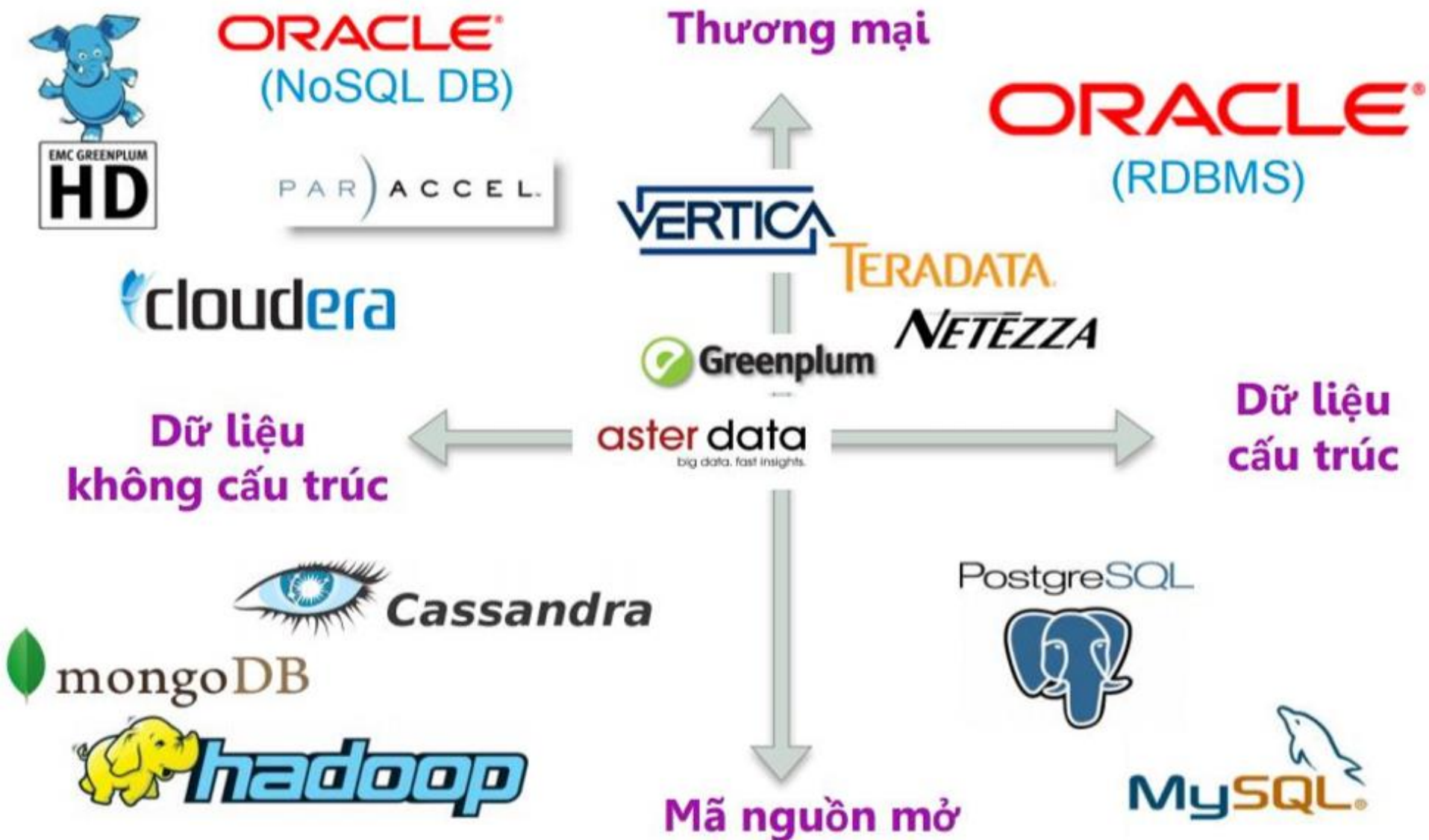
Tiếp cận dữ liệu lớn

- ❖ Nhiệm vụ khoa học công nghệ dữ liệu lớn
- ❖ Quản lý dữ liệu lớn
- ❖ Yêu cầu khi xử lý dữ liệu lớn

Nhiệm vụ khoa học công nghệ dữ liệu lớn

- ❖ Quản trị dữ liệu (DATA MANAGEMENT):
 - ❖ Lưu trữ, bảo trì và truy nhập các nguồn dữ liệu lớn
- ❖ Mô hình hóa và phân tích dữ liệu (DATA MODELING and ANALYTICS):
 - ❖ Tìm cách hiểu được dữ liệu và tìm ra các thông tin hoặc tri thức quý báu từ dữ liệu
- ❖ Trao đổi, hiển thị dữ liệu và kết quả phân tích dữ liệu (VISUALIZATION DECISIONS and VALUES) để tạo ra sản phẩm hay giá trị

Quản lý dữ liệu lớn



Yêu cầu khi xử lý dữ liệu lớn

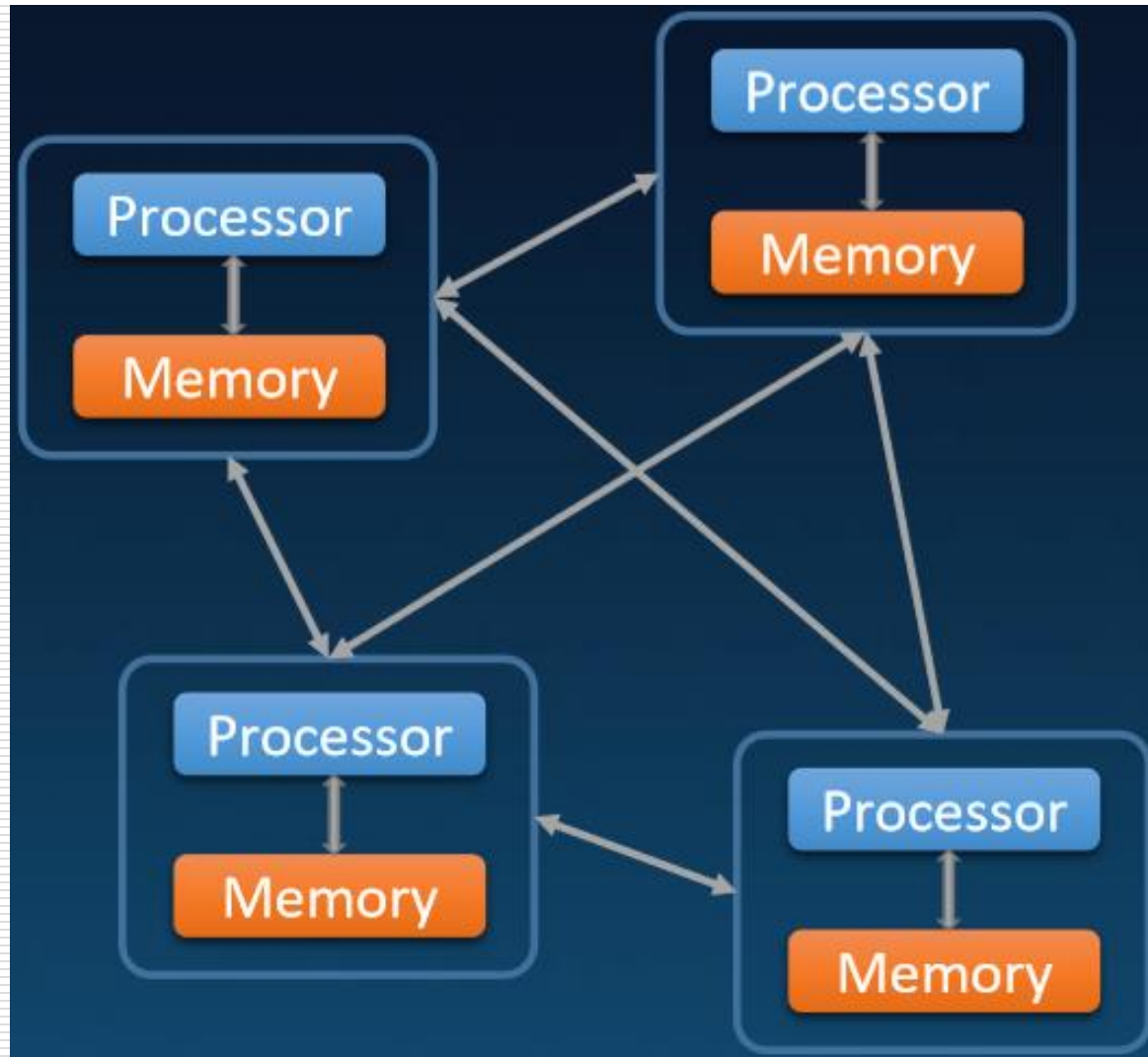
- ❖ Khả năng mở rộng
 - ❖ Hệ thống có khả năng đối phó với sự tăng trưởng của dữ liệu, tính toán và độ phức tạp
- ❖ Hiệu suất vào ra dữ liệu
 - ❖ Tốc độ truyền dữ liệu giữa hệ thống và thiết bị ngoại vi
- ❖ Khả năng chấp nhận lỗi
 - ❖ Khả năng tiếp tục hoạt động đúng trong trường hợp thất bại của một hay nhiều thành phần
- ❖ Xử lý thời gian thực
 - ❖ Khả năng xử lý dữ liệu và đưa ra kết quả chính xác trong những ràng buộc thời gian nhất định
- ❖ Hỗ trợ kích thước dữ liệu
 - ❖ Kích thước của tập dữ liệu mà hệ thống có thể xử lý hiệu quả
- ❖ Hỗ trợ tác vụ lặp: Hệ thống hỗ trợ hiệu quả tác vụ lặp

Công nghệ chính trong xử lý dữ liệu lớn

- ❖ Tính toán phân tán
- ❖ Tính toán song song
- ❖ Song song hóa bằng CPU đa nhân
- ❖ Song song hóa bằng GPU
- ❖ Xử lý phân tán với hệ thống cluster
- ❖ Xử lý phân tán trên cloud

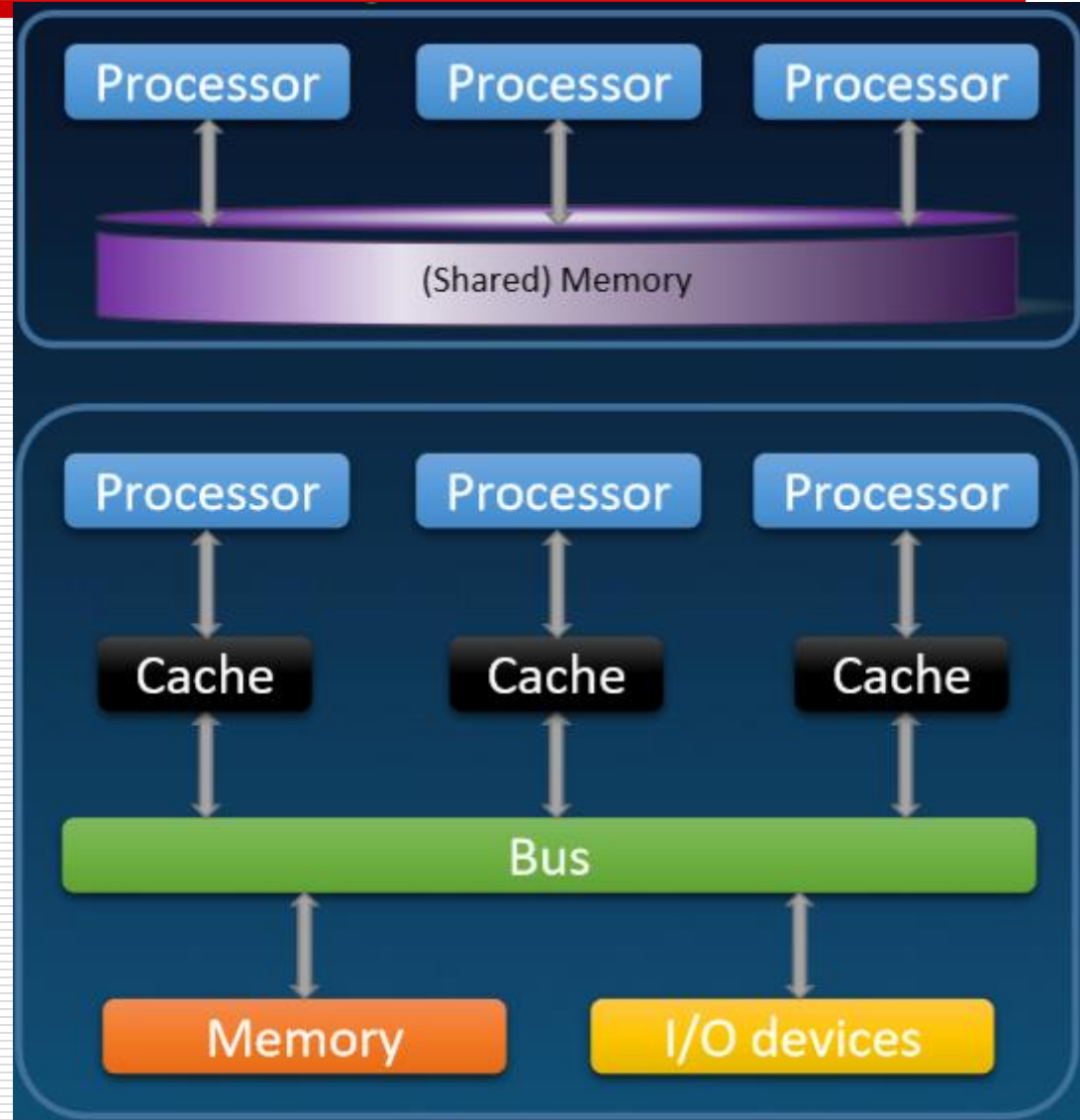
Tính toán phân tán

- ❖ Tính toán phân tán: bài toán được chia nhỏ thành cụm và phân tán vào nhiều máy khác nhau; mỗi máy có một bộ nhớ riêng



Tính toán song song

- ❖ Tính toán song song: bài toán có cấu trúc tính toán song song, được chia nhỏ vào nhiều bộ xử lý để tính song song có cùng bộ nhớ chung (chia sẻ)
- ❖ => Khác biệt chính so với tính toán phân tán: Bộ nhớ chia sẻ



Song song hóa bằng CPU đa nhân

- ❖ Máy tính với nhiều nhân xử lý
- ❖ Cơ chế song song đạt được thông qua đa luồng
- ❖ Số lượng nhân xử lý bị hạn chế: tiêu chuẩn thường có từ 4 đến 16 lõi xử lý xung nhịp từ 1 đến 4 GHz, CPU chuyên dụng có thể có đến 32 lõi xử lý

Song song hóa bằng GPU

- ❖ GPU (Graphics Processing Unit- là bộ xử lý đồ họa) là một loại bộ vi xử lý chuyên dụng
- ❖ Được tối ưu hóa để hiển thị đồ họa và thực hiện các tác vụ tính toán rất cụ thể
- ❖ Hiển thị video hoặc thực hiện các thao tác toán học đơn giản lặp đi lặp lại là “sở trường” của GPU
- ❖ Có hàng nghìn lõi xử lý chạy đồng thời (lên tới trên 2K lõi)
=> tốc độ cao hơn CPU rất nhiều
- ❖ Hạn chế: Ít phần mềm và thuật toán sẵn sàng với GPU

Xử lý phân tán với hệ thống cluster

- ❖ Hệ thống tính toán cụm: Tập các máy trạm hoặc PC kết nối chặt chẽ với nhau bởi mạng LAN tốc độ cao, chạy cùng một hệ điều hành
- ❖ Ưu điểm:
 - ❖ Kinh tế: rẻ hơn rất nhiều so với siêu máy tính truyền thống có cùng hiệu năng
 - ❖ Khả năng mở rộng: Dễ dàng nâng cấp, bảo trì
 - ❖ Tính tin cậy: Tiếp tục hoạt động thậm chí bị hỏng một phần (một vài máy tính hỏng)
- ❖ Hạn chế
 - ❖ Khi quản lý và tổ chức số lượng lớn máy tính
 - ❖ Hiệu suất vào/ra dữ liệu thấp
 - ❖ Không phù hợp cho xử lý thời gian thực

Xử lý phân tán trên cloud

- ❖ Được cung cấp bởi các công ty lớn
 - ❖ Google Cloud Platform
 - ❖ Amazon Web Services
 - ❖ Microsoft Azure
- ❖ Ưu điểm
 - ❖ Chi phí đầu tư và bảo trì thấp (dựa trên dịch vụ của nhà cung cấp)
 - ❖ Truy cập được mọi lúc, mọi nơi
 - ❖ Khả năng mở rộng cao
- ❖ Hạn chế
 - ❖ Vấn đề bảo mật dữ liệu không chắc được đảm bảo
 - ❖ Cần kết nối internet
 - ❖ Vấn đề di chuyển hệ thống (nếu cần)
 - ❖ => Phụ thuộc nhà cung cấp