DỮ LIỆU LỚN
(BIG DATA)

GVGD: VÕ THỊ HỒNG TUYẾT

Số tiết

- Số tiết lý thuyết: 45 tiết
- Số tiết thực hành: 30 tiết
- Số giờ tự học: 90 giờ

Nội dung môn học

- **Chương 1**. Cơ bản về Dữ liệu lớn
- **Chương 2**. Hadoop cơ bản
- **Chương 3**. Lập trình MapReduce
- **Chương 4**. Các công cụ phát triển ứng dụng dữ liệu lớn
- **Chương 5**. Phân tích Dữ liệu Lớn

Hình thức đánh giá

Đánh giá	Trọng số	Hình thức đánh giá	Nội dung	Phương pháp đánh giá
Kiểm tra thường xuyên	20%	Từng buổi học	Điểm danh 100% các buổi học	Điểm danh các buổi học
		Bài tập + báo cáo tiến độ	- Hoàn thành các hoạt động được giao trên Moodle và trên lớp. - Hoàn thành báo cáo tiến độ đề tài.	Đánh giá bài tập trên lớp và bài tập online
Đánh giá phần thực hành	30%	Sinh viên hoàn thành bài thực hành trong buổi thực hành	Hoàn thành các hoạt động được giao trên Moodle và trên lớp	Đánh giá bài tập trên lớp và bài tập online
Đánh giá cuối kỳ	50%	Đồ án môn học	Đồ án môn học	Đánh giá đồ án

CHÍNH SÁCH VỚI HỌC PHẦN

- + Mỗi mốc thời gian đi trễ 15' so với giờ vào lớp sẽ nhận 1 dấu trừ (tương ứng 0.25) vào tổng điểm quá trình.
- + Nộp bài tập đầy đủ, đúng hạn theo quy định. Mọi trường hợp trễ hạn phải có minh chứng lý do và phải được phê duyệt. Nếu không sẽ nhận điểm 0 cho bài tập đó.
- + Các trường hợp xin vắng có phép theo quy định của Nhà Trường (có minh chứng cụ thể và được chấp thuận).
- + Vắng quá 20% số buổi học sẽ mất điểm chuyên cần.
- + Vắng buổi báo cáo tiến độ đề tài (không có lý do và minh chứng được chấp thuận) sẽ nhận điểm 0 tại cột báo cáo tiến độ đề tài.
- + Thời hạn giải đáp thắc mắc điểm và chốt điểm quá trình: trực tiếp vào buổi học cuối.

Tài liệu tham khảo

TT	Tên tác giả	Tên sách	Nhà xuất bản
Giáo trình chính			
1	O'Reilly Media, Inc	Big Data Now	O'Reilly Media, Inc (2017).
Sách, giáo trình tham khảo			
2	Jeffrey Needham	Disruptive Possibilities: How Big Data Changes Everything	2013
3	Hadoop	http://hadoop.apache.org (cập nhật 05/05/2023)	

Bài 1

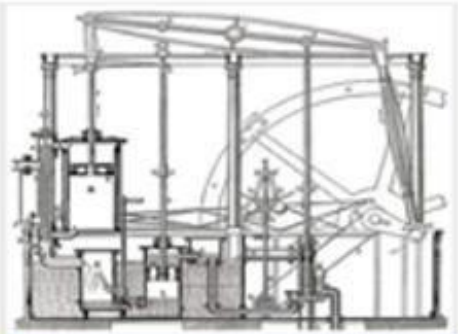
TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU LỚN

Nội dung

1. Cách mạng công nghiệp lần thứ 4
2. Công nghệ số
3. Dữ liệu lớn là gì
4. Dữ liệu lớn đến từ đâu?
5. Đặc trưng cơ bản của dữ liệu lớn
6. Ứng dụng của dữ liệu lớn
7. Tiếp cận dữ liệu lớn
8. Công nghệ chính trong xử lý dữ liệu lớn
9. Một số nền tảng công nghệ quan trọng

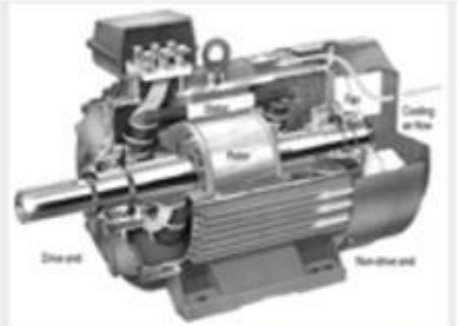
Cách mạng công nghiệp lần thứ 4

- 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 hoá*.

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…
  - Các đột phá của công nghệ số trên Hệ kết nối không gian số-thực thể (cyber-physical systems)
- Cách mạng số hoá:
- Hệ kết nối không gian số-thực thể (cyber-physical system): hệ kết nối các thực thể và 'phiên bản số' của chúng

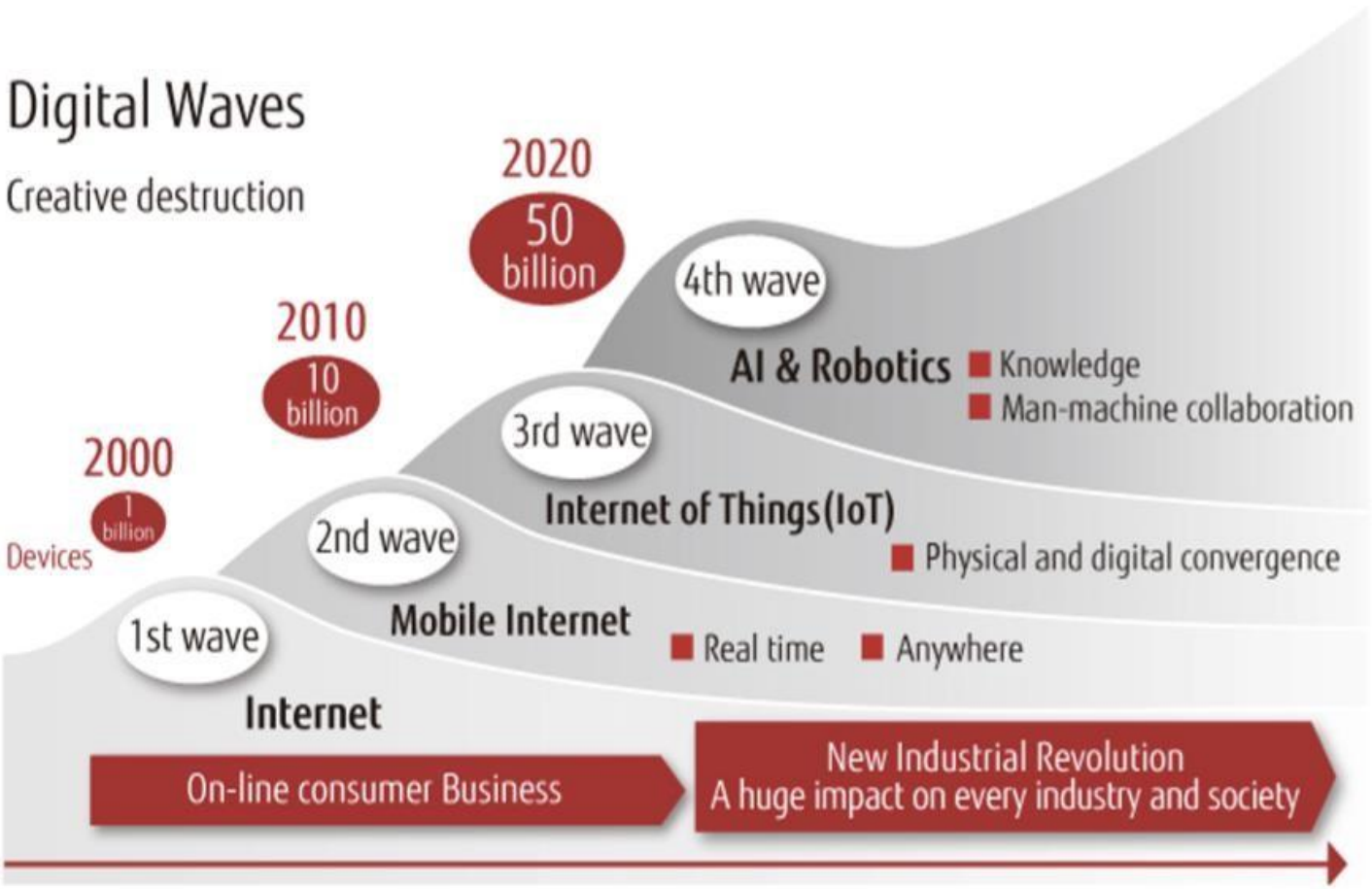
=> Thay đổi phương thức sản xuất:

  - Hành động trong thế giới các thực thể
  - Tính toán, điều khiển trên không gian số

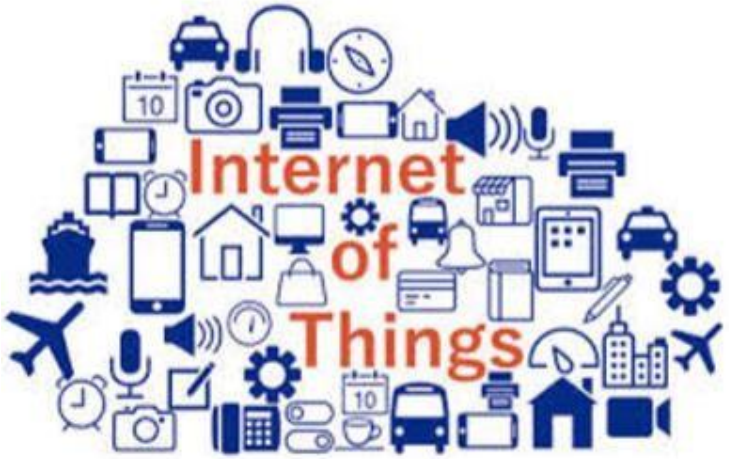
Công nghệ số

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

How digital technology will transform the world, Fujitsu Journal, 1.2016



Công nghệ số - Đột phá gần đây



Dữ liệu lớn ?

- **Theo wikipedia:**
  - 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:**
  - 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.

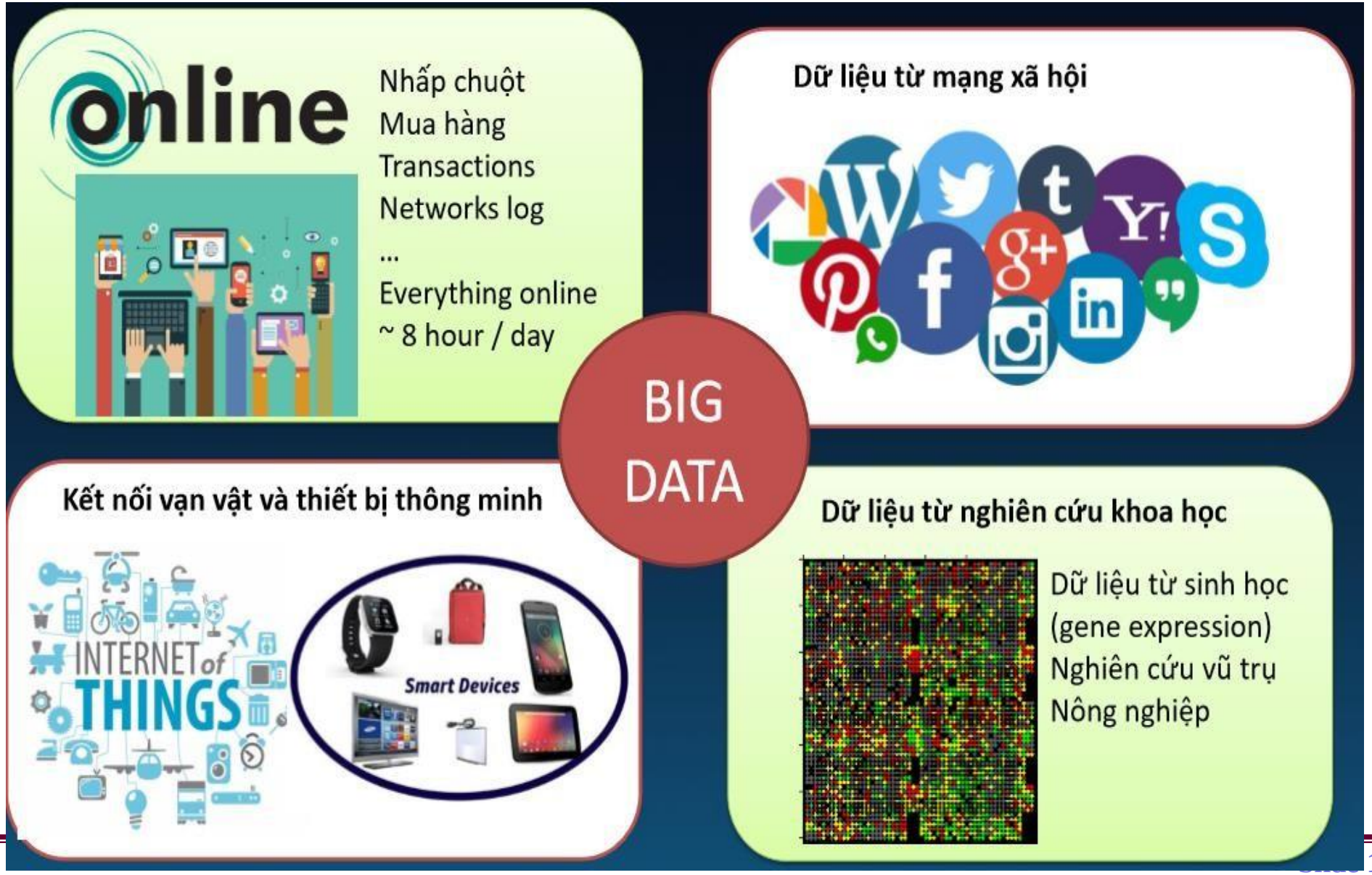
Dữ liệu lớn đến từ đâu?

- 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?



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)
2. Dữ liệu từ hoạt động thương mại (phát sinh từ các giao dịch giữa hai thực thể)
3. Dữ liệu từ các thiết bị cảm biến như thiết bị chụp hìnhảnh vệ tinh, cảm biến đường, cảm biến khí hậu;
4. Dữ liệu từ các thiết bị theo dõi
5. Dữ liệu từ các hành vi
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 (mô hình 5Vs về dữ liệu lớn):

1. Độ lớn dữ liệu (Volume)
2. Tốc độ (Velocity)
3. Đa dạng (Variety)
4. Độ tin cậy/chính xác (Veracity)
5. Giá trị (Value)

Độ lớn của dữ liệu (Volume)

- Độ lớn của dữ liệu rất lớn
- Kích cỡ của Big Data đang từng ngày tăng lên
- 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

Tốc độ (Velocity)

- Tốc độ có thể hiểu theo 2 khía cạnh:
  - Khối lượng dữ liệu gia tăng rất nhanh
  - 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
- Các ứng dụng phổ biến trên lĩnh vực Internet, Tài chính, Ngân hàng, Hàng không, Quân sự, Y tế – Sức khỏe 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

Đa dạng (Variety)

- Đối với dữ liệu truyền thống => 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

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 => 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.

Giá trị (Value)

- 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

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...

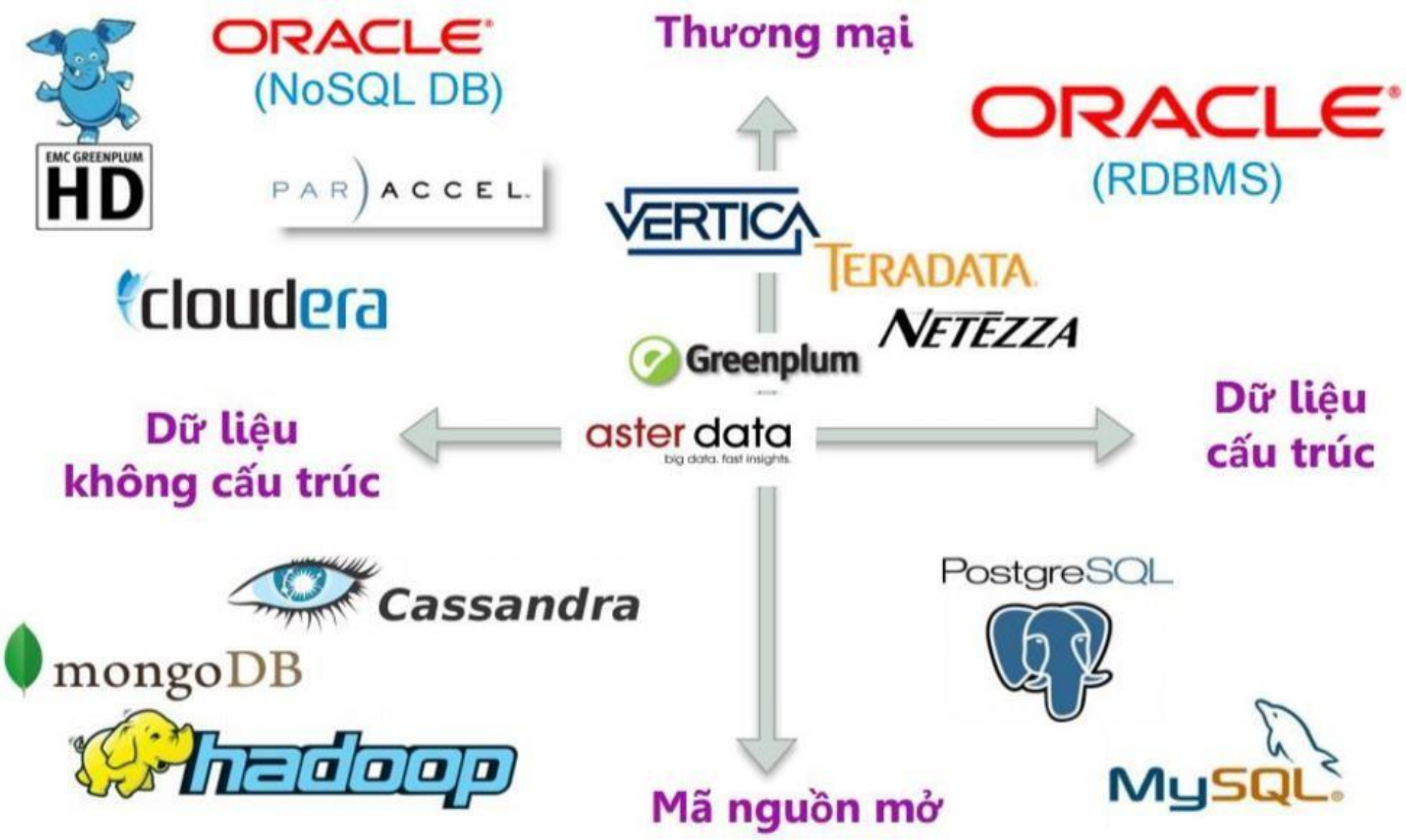
Tiếp cận dữ liệu lớn

- 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
- Trao đổi, hiển thị dữ liệu và kết quả phân tích dữ liệu để tạo ra sản phẩm hay giá trị

Quản lý dữ liệu lớn



Yêu cầu khi xử lý dữ liệu lớn

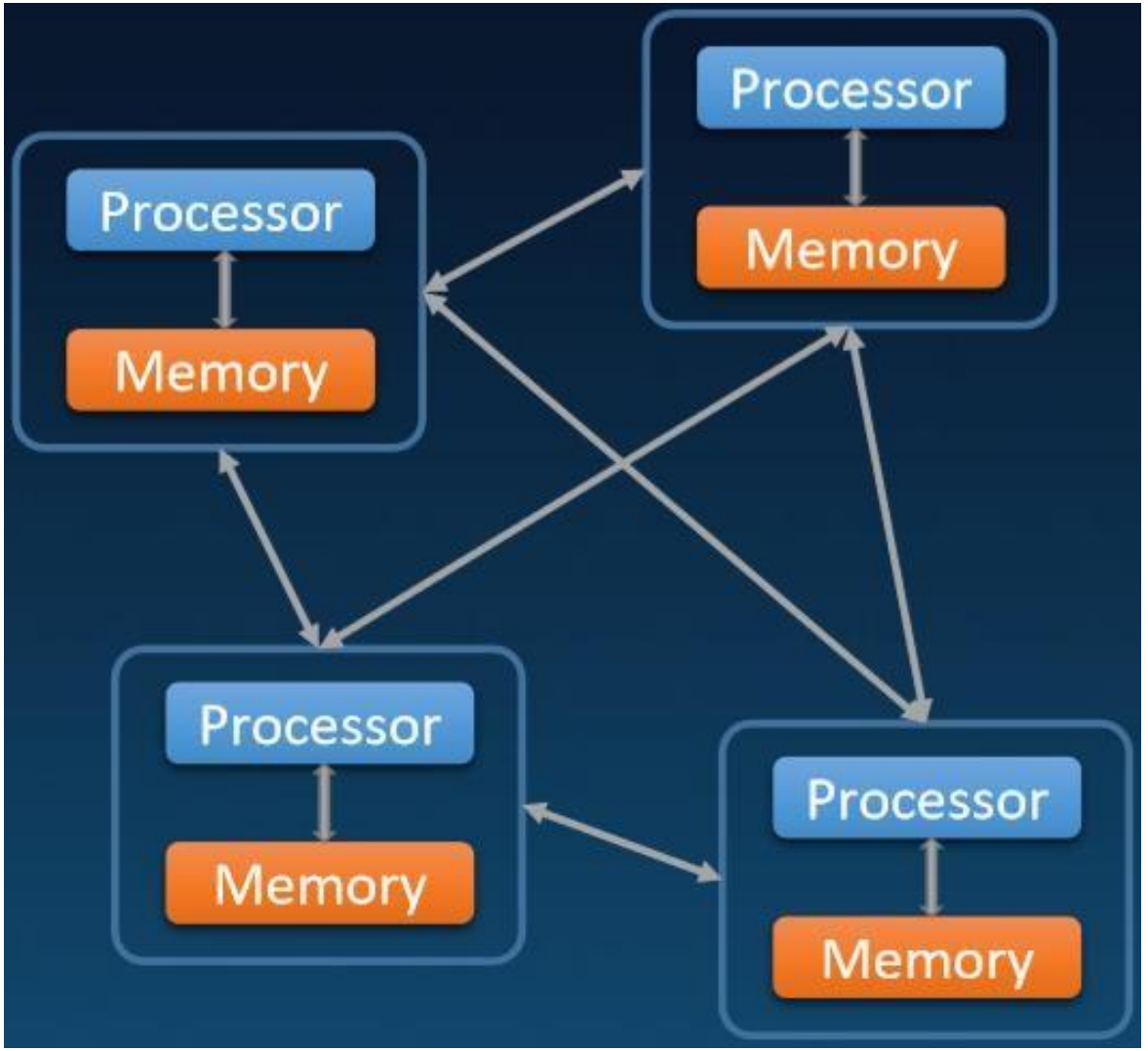
- Khả năng mở rộng
  - Hệ thống có khả năng đối phó với sự tăng trưởng của dữ liệu, tính toán và độ phức tạp
- Hiệu suất vào ra dữ liệu
  - Tốc độ truyền dữ liệu giữa hệ thống và thiết bị ngoại vi
- Khả năng chấp nhận lỗi
  - Khả năng tiếp tục hoạt động đúng trong trường hợp thất bại của một hay nhiều 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

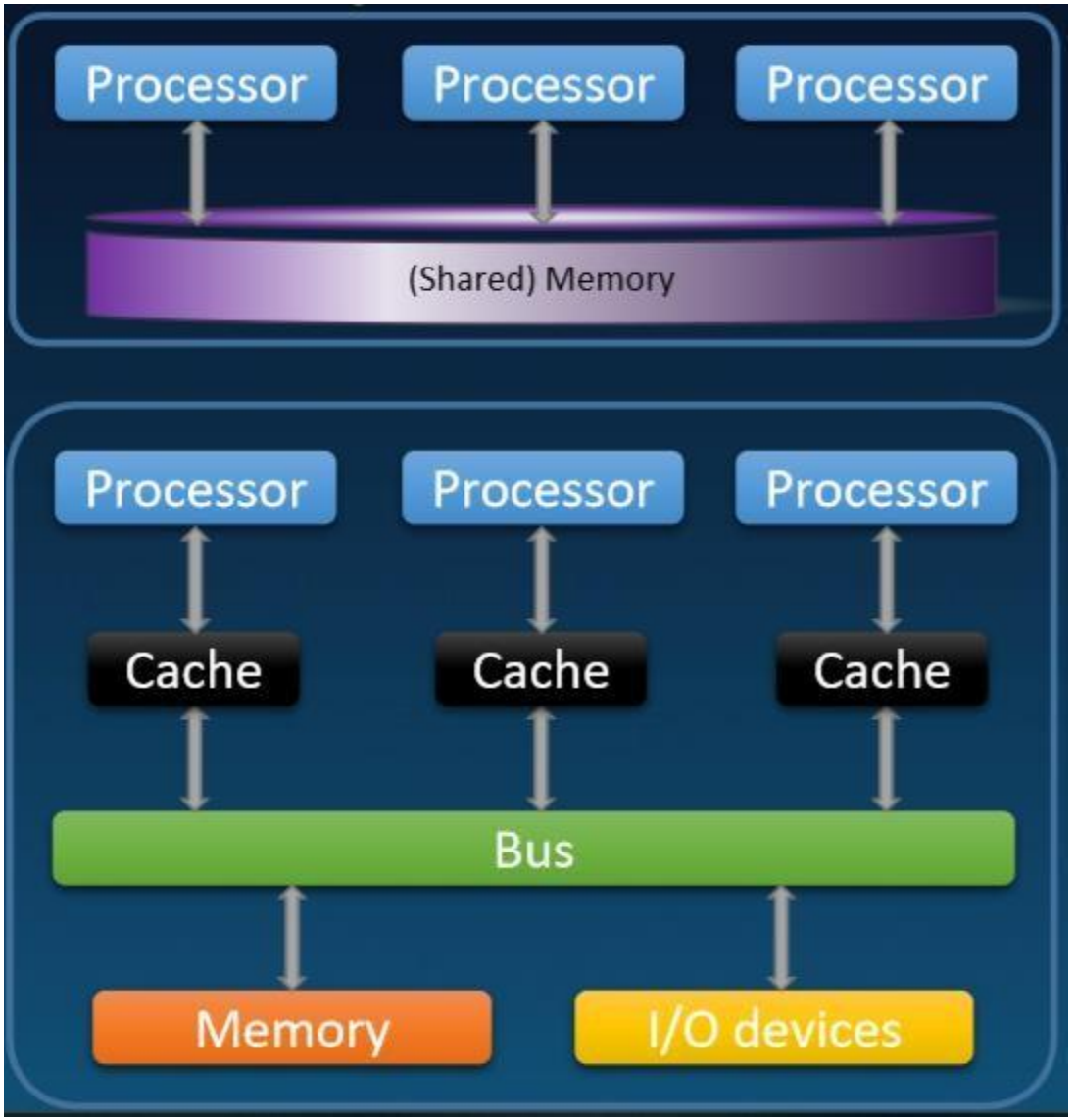
- Tính toán phân tán: bài toán được chia nhỏ thành cụm và phân tán vào nhiều máy khác nhau; mỗi máy có một bộ nhớ riêng



Tính toán song song

- 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
  - Khí 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

Một số công nghệ quan trọng

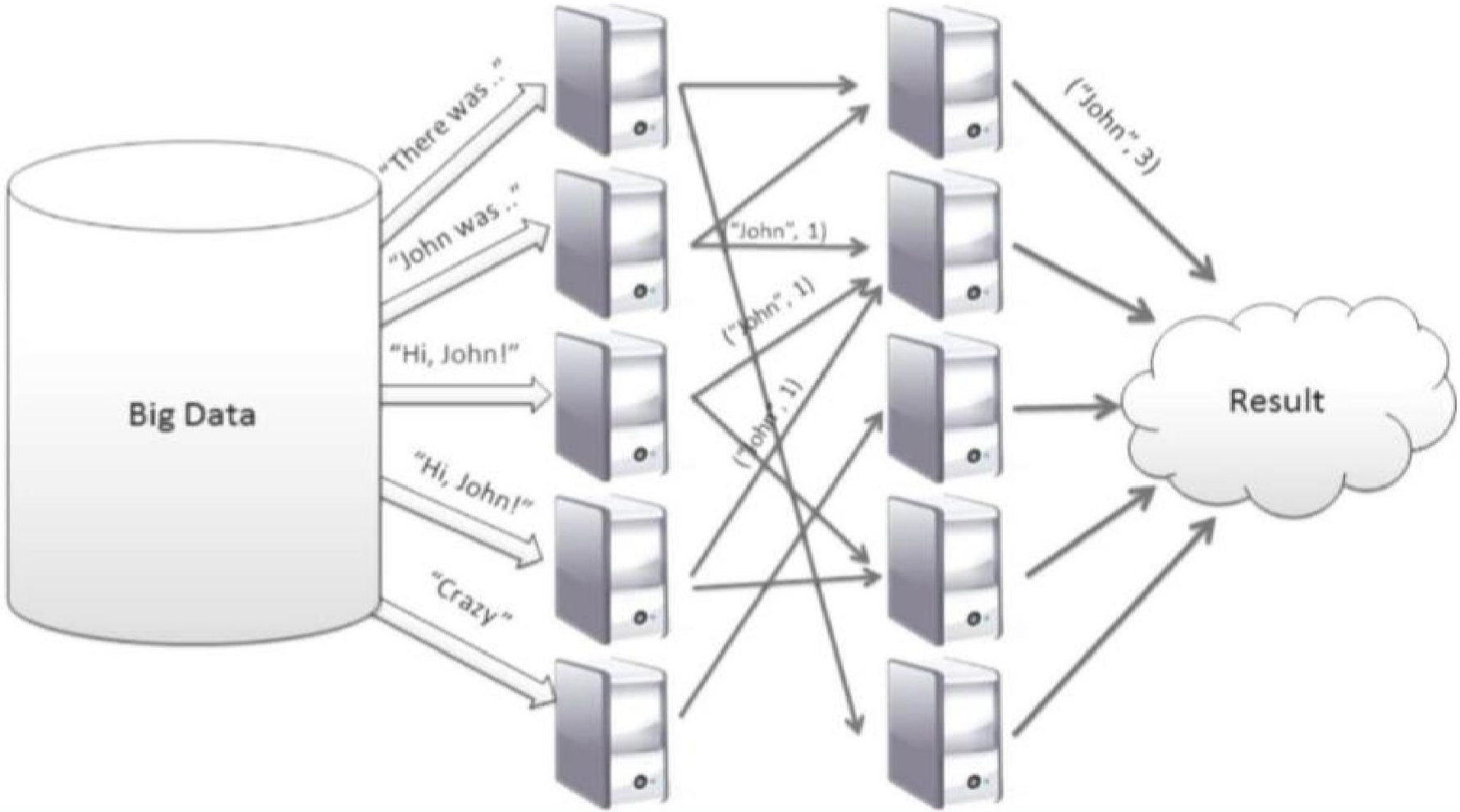
- MapReduce
- Hadoop
- Spark
- TensorFlow

Công nghệ MapReduce

- Nhu cầu xử lý DLL tăng nhanh
- Phát minh năm 2004 bởi Google
- Hỗ trợ xử lý song song và phân tán
- Được sử dụng trong Hadoop
- Chia nhiệm vụ thành 2 phần chính: map and reduce
- **Map:** Đọc dữ liệu từ HDFS, xử lý dữ liệu và cho kết quả trung gian => thực hiện song song
- **Reduce:** Tổng hợp kết quả trung gian để sinh kết quả cuối cùng => thực hiện song song
- **Hạn chế:**
  - Thiếu hiệu quả trong việc chạy các thuật toán lặp
  - Mapper đọc dữ liệu tương tự nhau nhiều lần từ đĩa

Công nghệ MapReduce

- Song song hóa dữ liệu, Xử lý phân tán, Tổng hợp kết quả



Hadoop

- Là framwork opensource để lưu trữ và xử lý tập dữ liệu lớn sử dụng cụm phần cứng
- Khả năng chấp nhận lỗi cao: khắc phục lỗi khi một vài node phần cứng bị hỏng
- Có thể lên tới 100 hay 1000 nodes
- Thành phần chính:
  - **Common:** Cung cấp tiện ích hỗ trợ các module, thành phần khác
  - **YARN:** framework để lập lịch công việc và quản lý tài nguyên cụm
  - **HDFS:** Hệ thống file phân tán
  - **MapReduce:** Mô hình tính toán cho xử lý song song các tập dữ liệu lớn

Spark

- Tập dữ liệu phân tán linh hoạt (Resilient Distributed Datasets-RDD)
  - Tập các bản ghi chỉ đọc, đã phân vùng, được phân tán trên các cụm, lưu trữ trong bộ nhớ hoặc đĩa
  - Xử lý dữ liệu dựa trên mô hình đồ thị chuyển đổi trong đó các node là các RDD, các cạnh là các chuyển đổi
- **Ưu điểm:**
  - Khả năng chấp nhận lỗi cao: do các RDD
  - Khả năng lưu trữ: Lưu trữ một số RDD trong RAM, vì thế nhanh hơn Hadoop MapReduce đối với vấn đề lặp
  - Hỗ trợ MapReduce

TensorFlow

- Hỗ trợ xử lý song song
- Là framework nguồn mở cho học sâu, được phát triển bởi team GoogleBrain
- Cung cấp tính năng cơ bản để định nghĩa các hàm trên các Tensor và tính toán
- Ngôn ngữ lập trình thân thiện: Python hoặc C++
- Hỗ trợ đa nền tảng: Linux CPU, Linux GPU, Mac OS CPU, Windows CPU, Android
- Nhiều GPU trên một máy tính hoặc phân phối trên nhiều máy tính

HẾT