Giới thiệu về Big Data

Nguyễn Đình Hưng, PhD

January 2021

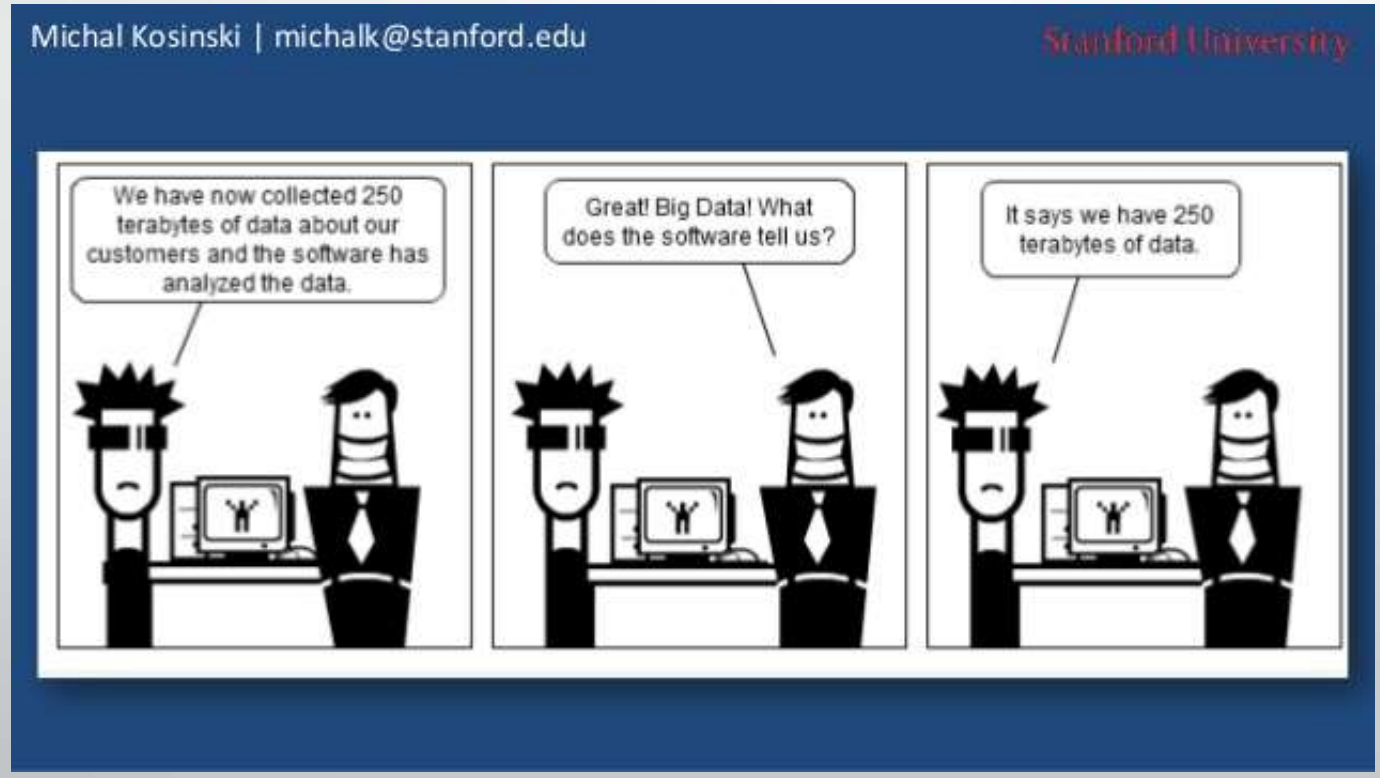
Nội dung

- Khái niệm Big Data
- Đặc trưng của Big Data
- Nguồn hình thành Big Data
- Lợi ích & thách thức khi xử lý Big Data

"

90% of the world's data was generated in the last few years.

"

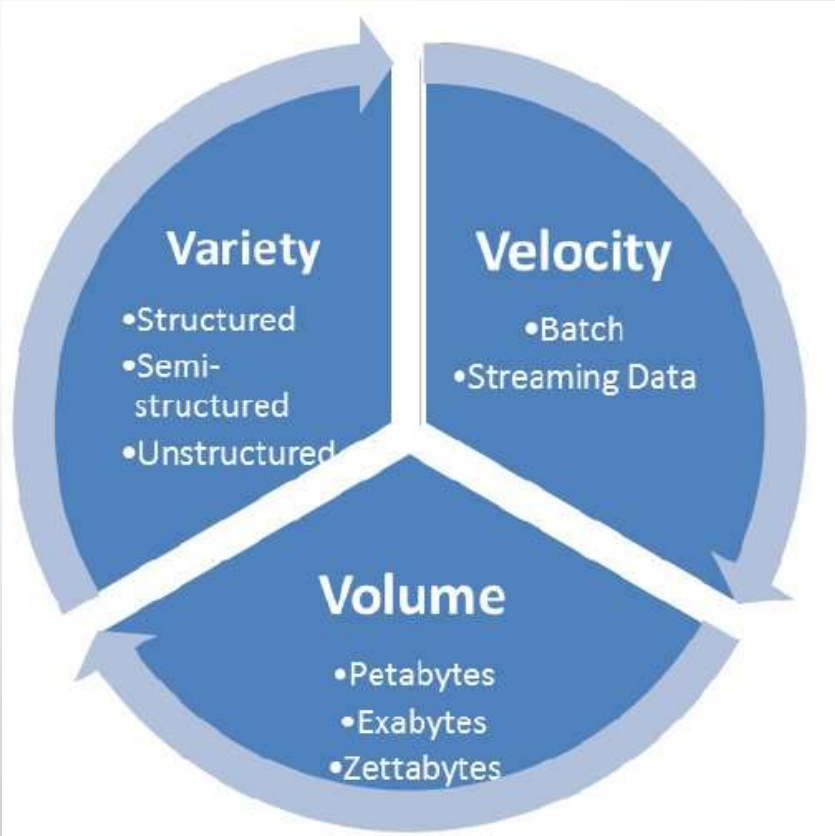


Dữ liệu lớn (Big Data) là gì?

- Vài năm gần đây, sự ra đời của nhiều công nghệ, thiết bị và phương tiện truyền thông mới đã tạo ra lượng dữ liệu khổng lồ và không ngừng tăng lên nhanh chóng:
  - Đến 2003, lượng dữ liệu sinh ra toàn cầu vào khoảng 5 tỷ gigabyte.
  - Dung lượng tương đương được tạo ra trong 2 ngày của 2011, và sau mỗi 10 phút trong 2013.
  - Đầu 2020, tổng dữ liệu toàn thế giới ước tính 44 zettabytes. (World Economic Forum)
- Dữ liệu lớn (Big Data) là các tập dữ liệu (datasets) có dung lượng rất lớn và/hoặc cấu trúc phức tạp. Việc xử lý dữ liệu lớn vượt quá khả năng của các phương pháp truyền thống.
- Khai thác dữ liệu lớn có thể rút ra thông tin hữu ích hỗ trợ ra quyết định, nhằm tăng hiệu quả hoạt động, giảm chi phí, giảm rủi ro cho cơ quan, doanh nghiệp.

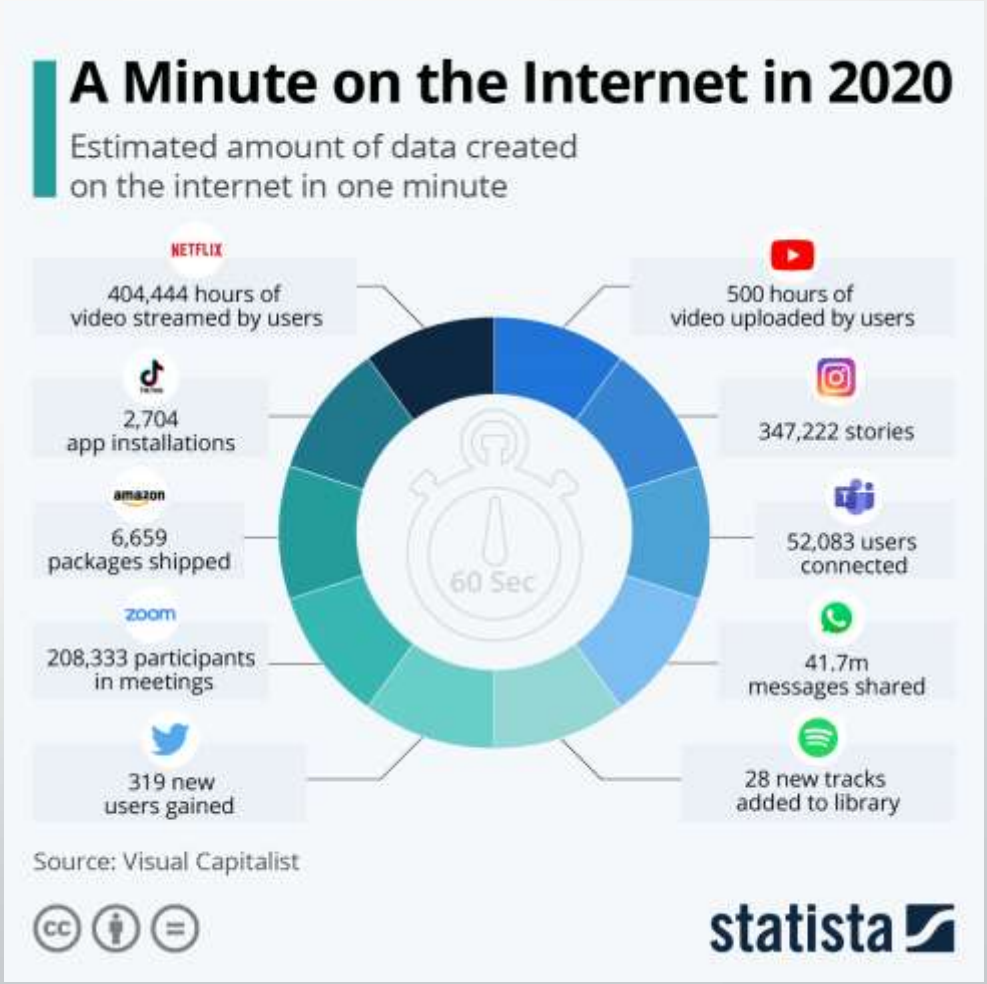
Đặc trưng của Big Data

- Volume
  - Dung lượng rất lớn
- Variety
  - Relational Data (Tables/Transaction)
  - Text Data (Web)
  - Semi-structured Data (XML)
  - Graph Data (Social Network)
  - Streaming Data
- Velocity
  - Dòng dữ liệu không ngừng chuyển động



Big Data đến từ đâu?

”640K ought to be enough for anybody” – Bill Gates, 1981



Big Data đến từ đâu? (cont.)

Processes 20 PB a day (2008)
Crawls 20B web pages a day (2012)
Search index is 100+ PB (5/2014)
Bigtable serves 2+ EB, 600M QPS (5/2014)

400B pages, 10+ PB (2/2014)

Hadoop: 365 PB, 330K nodes (6/2014)

300 PB data in Hive + 600 TB/day (4/2014)

Hadoop: 10K nodes, 150K cores, 150 PB (4/2014)

S3: 2T objects, 1.1M request/second (4/2013)

Big Data đến từ đâu? (cont.)

- Dữ liệu từ hộp đen (black box)
  - Ghi lại diễn biến hành trình của máy bay, tàu hỏa, …
- Giao dịch điện tử
  - Email
  - Mua bán, vận chuyển
- Mạng xã hội (social media)
  - Dữ liệu được hàng tỷ người dùng mạng xã hội tạo ra (user-generated data)
- Thị trường chứng khoán
  - Dữ liệu ghi nhận hoạt động mua bán của thị trường chứng khoán
- Các máy chủ tìm kiếm (search engines)
  - Các máy chủ tìm kiếm (như Google) lưu trữ lượng dữ liệu khổng lồ về các trang web toàn cầu
- Dữ liệu nghiên cứu khoa học
  - Máy gia tốc hạt LHC (Large Hadron Collider) tạo ra 25 petabytes dữ liệu mỗi năm (1 PB = 1024 TB)
- …

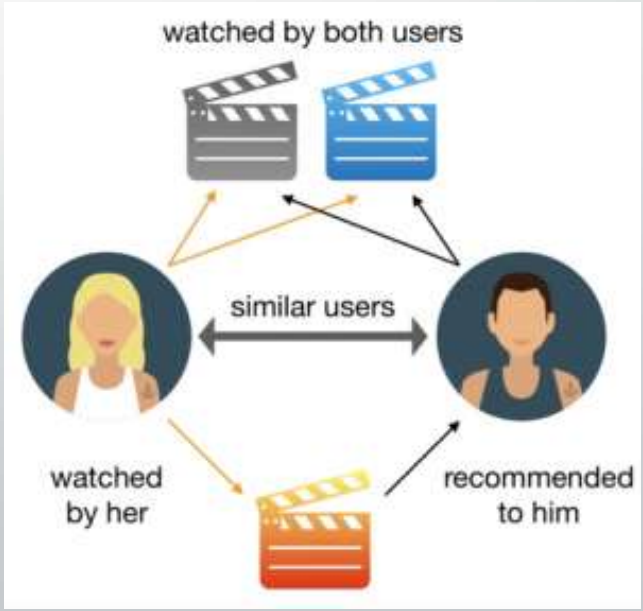
Ứng dụng Big Data

Ứng dụng Big Data (cont.)

- Bán lẻ (retail), sản xuất (production)
  - Phân tích dữ liệu khách hàng giúp hiểu rõ hơn sở thích, hành vi khách hàng, tư đó đưa ra sản phẩm, dịch vụ đáp ứng tốt hơn nhu cầu của họ
- Vận tải (logistics)
  - Trong lĩnh vực vận tải, chi phí xử lý sản phẩm bị trả lại cao gấp 1.5 lần chi phí giao hàng.
  - Phân tích dữ liệu vận tải giúp dự đoán những sản phẩm có khả năng bị trả lại cao, từ đó đưa ra giải pháp khắc phục nhằm giảm thiểu số hang hóa bị trả lại.
- Y tế (healthcare)
  - Phân tích dữ liệu y tế giúp đưa ra các giải pháp hiệu quả hơn trong điều trị, chăm sóc và nghiên cứu thuốc chữa bệnh

Ứng dụng Big Data (cont.)

- Giáo dục (education)
  - Một số trường đại học hàng đầu sử dụng Big Data làm công cụ để cải tiến chương trình đào tạo.
  - Một số trường phân tích dữ liệu sinh viên để tìm đặc điểm chung của các sinh viên bỏ học, từ đó đưa ra giải pháp giảm thiểu số sinh viên bỏ học.
- Thương mại điện tử (e-commerce)
  - Recommendation system
- Tài chính (finance)
  - Fraud detection
  - Risk analysis
  - Credit scoring



Những thách thức của Big Data

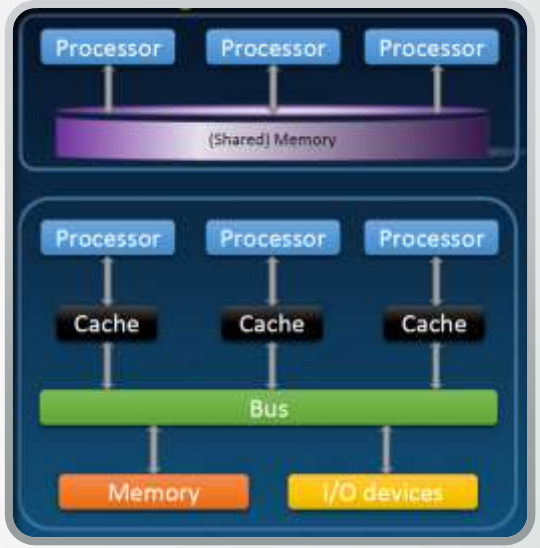
- Dữ liệu gia tăng nhanh chóng
  - Dữ liệu được tạo ra tăng theo lũy thừa
  - Phần lớn dữ liệu ở dạng phi cấu trúc: văn bản, hình ảnh, âm thanh
- Khó khăn
  - Thu thập dữ liệu (capturing data)
  - Quản lý, lưu trữ (storage)
  - Tìm kiếm (search)
  - Truyền dữ liệu (transfer)
  - Phân tích (analytics)

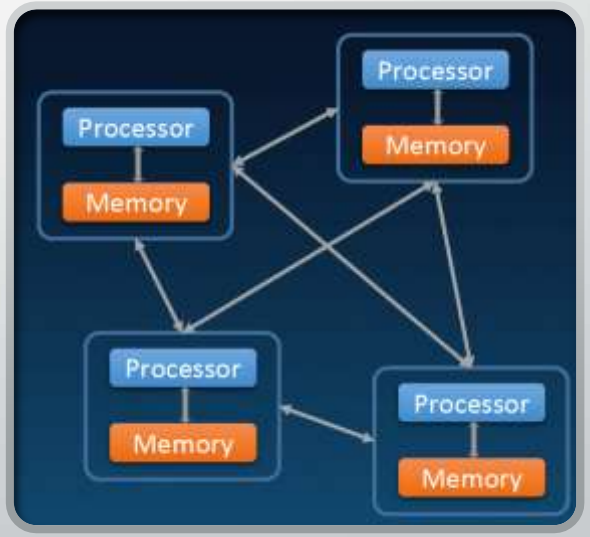
⇨ Phương pháp và công cụ phân tích dữ liệu là chìa khóa quan trọng.

- Vấn đề bảo mật thông tin
- Vấn đề đạo đức
  - Vi phạm đời sống riêng tư, thu thập dữ liệu trái phép

Các giải pháp then chốt xử lý Big Data

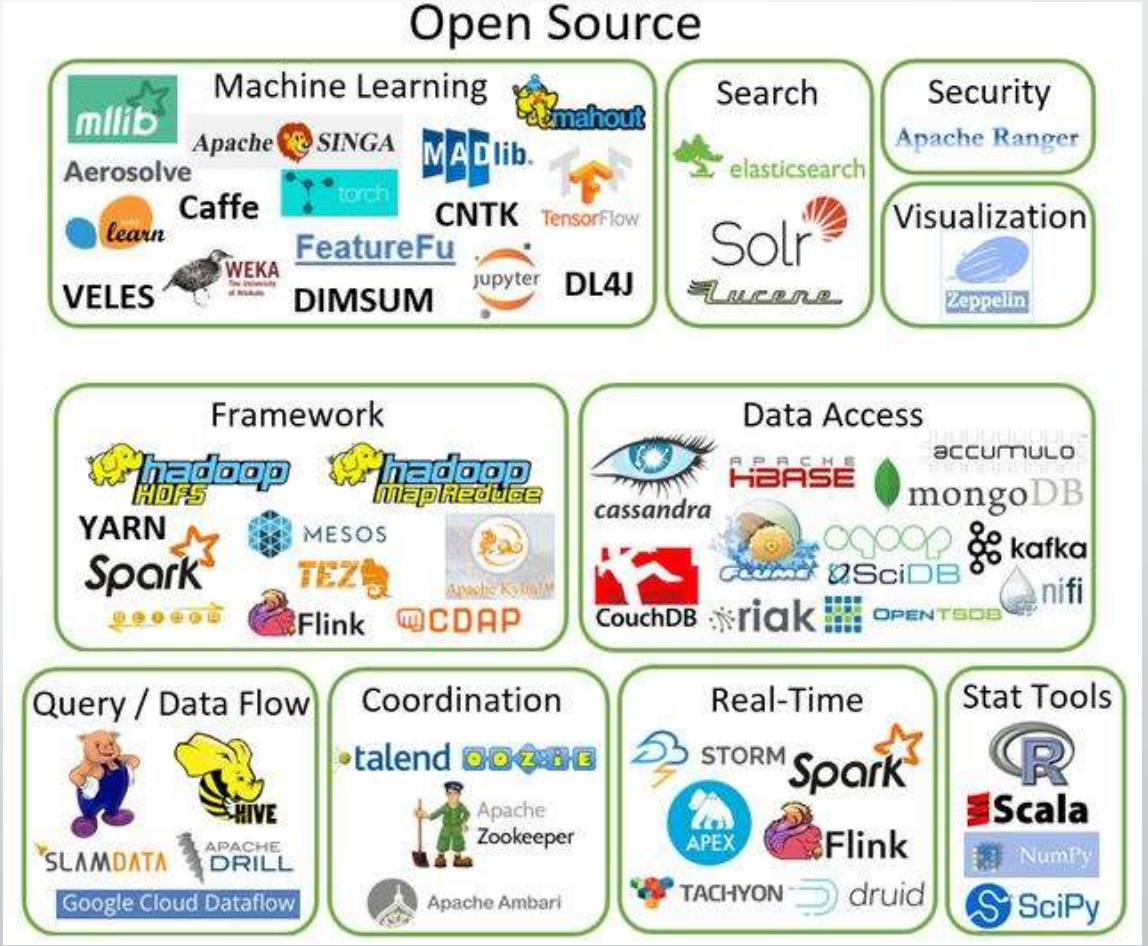
- Scaling out – Distributed computing
  - Khả năng bổ sung trạm xử lý khi cần (mở rộng theo chiều ngang – vertical scalability)
  - Mô hình phân tán: Bài toán được chia nhỏ thành các cụm và phân tán vào nhiều máy khác nhau
  - Mỗi máy có bộ nhớ riêng
- Scaling up - Parallel computing
  - Nâng cấp hiệu năng xử lý cho hệ thống hiện hữu (bộ xử lý, bộ nhớ)
  - Bài toán được chia nhỏ vào nhiều bộ xử lý để tính toán song song
  - Sử dụng bộ nhớ chung





Một số công cụ xử lý Big Data

- Frameworks
  - Hadoop
  - MapReduce
  - Spark
- Programming languages
  - Python
  - Java
  - Scala
  - R



Một số thuật ngữ

- Scalability
  - Khả năng đáp ứng tốt với sự gia tăng của dữ liệu và độ phức tạp tính toán
- Fault tolerance
  - Khả năng hoạt động liên tục, ổn định khi xảy ra lỗi ở một vài thành phần trong hệ thống
- Data I/O performance
  - Tốc độ đọc/ghi dữ liệu từ/lên một thiết bị ngoại vi
- Real-time processing
  - Khả năng xử lý dữ liệu và cho kết quả trong một khoảng thời gian ngắn
- Data size supported
  - Kích thước dữ liệu mà hệ thống hoạt động hiệu quả
- Iterative tasks support
  - Khả năng xử lý hiệu quả các tác vụ lặp lại