

TỔNG QUAN

Dữ liệu lớn

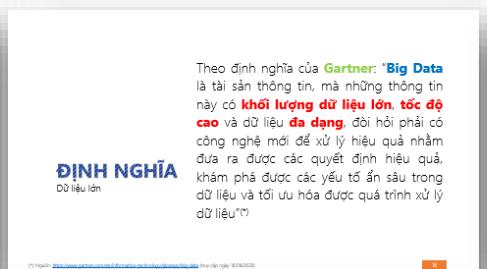
ThS. Nguyễn Hồ Duy Trí
trinhhd@uit.edu.vn



TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN

2020

NỘI DUNG



GIỚI THIỆU

Dữ liệu lớn

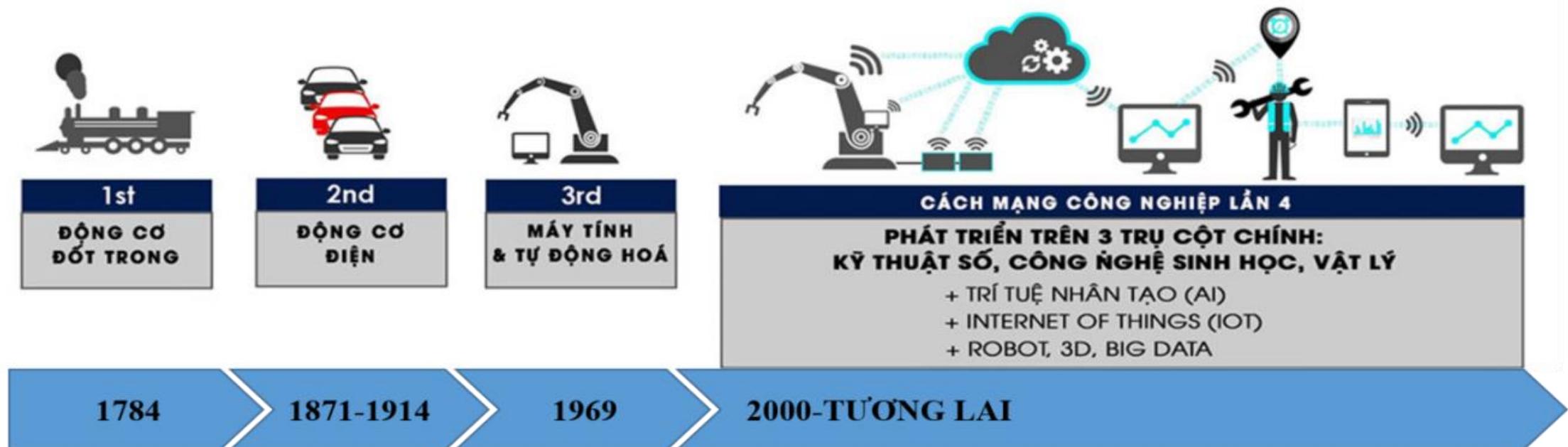


GIỚI THIỆU

- Bối cảnh
- Quá trình phát triển
- Dữ liệu hiện tại

Bối cảnh

Cách mạng công nghiệp 4.0



Cải cách số hóa (digital disruption) đẩy mỗi lĩnh vực đến những tương lai không tưởng và khác nhau.

- In 3D: in ra 1 cây đàn, 1 cái kính hàng hiệu, mô người... Từ sản xuất hàng loạt chuyển thành sản xuất cho từng cá nhân.
- Ngân hàng: điện thoại thông minh đang thách thức giao tiếp giữa khách hàng và ngân hàng. Vd: Timo, Momo...
- Bán lẻ: đơn ngành sang đa ngành (Tiki), thúc đẩy TMĐT phát triển mạnh mẽ.
- Doanh nghiệp có khả năng giao tiếp với từng cá nhân sẽ phát triển mạnh mẽ.

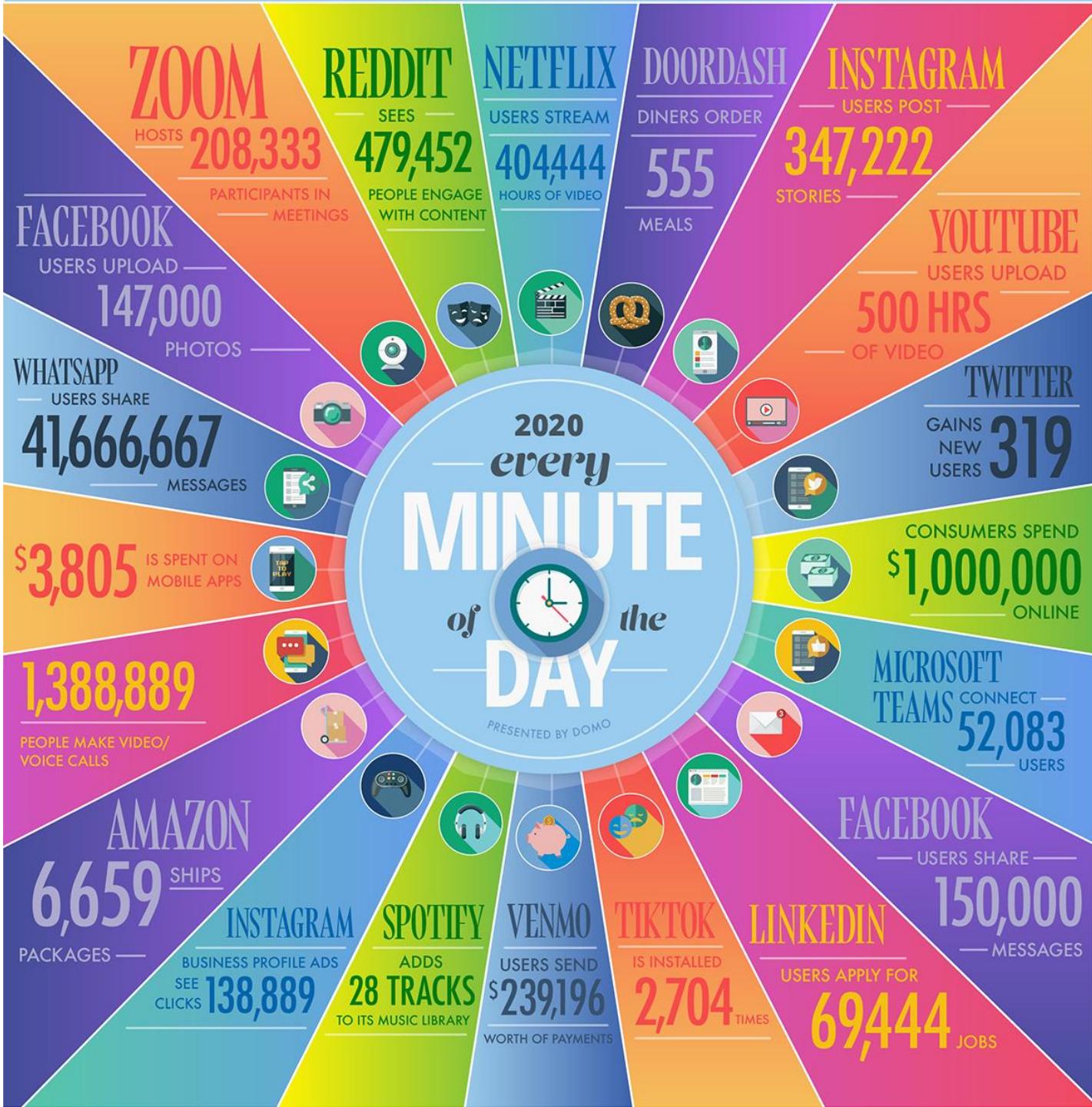
- **Dấu vết kỹ thuật số:** hầu như mỗi lần chúng ta sử dụng công nghệ trong thế giới số ngày nay, cho dù để trao đổi thông tin, mua bán, học tập, giải trí hay giao tiếp, chúng ta đều để lại một vết thông tin số nào đó. Nó sẽ phát triển thành một tấm gương phản chiếu cách chúng ta sử dụng thời gian, điều chúng ta quan tâm nhất, sở thích của chúng ta và thậm chí cái chúng ta cần.



- Google my activity: <https://myactivity.google.com/>
- Google maps timeline: <https://www.google.com/maps/timeline>

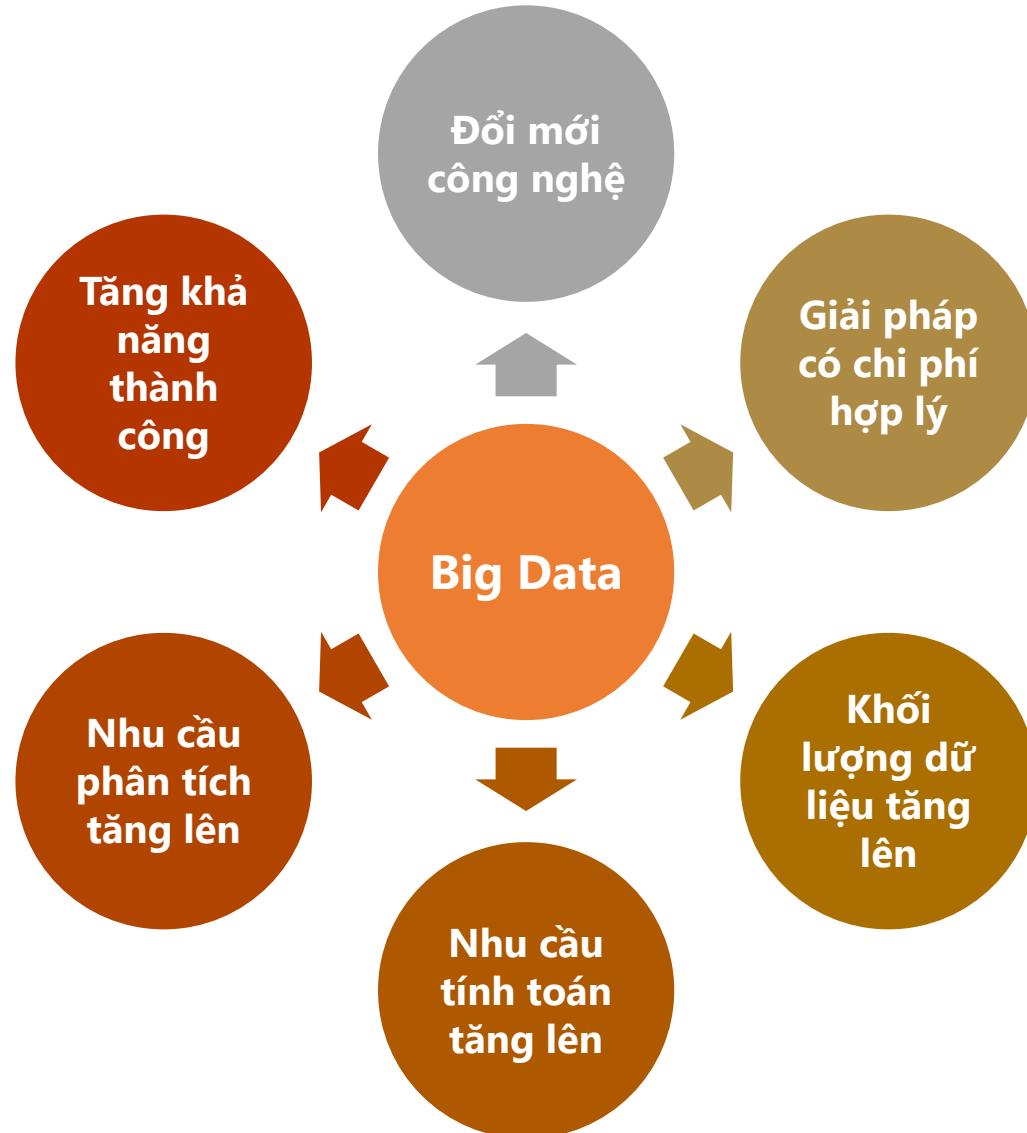


- Facebook activity log: <https://www.facebook.com/user/allactivity>



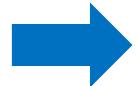
Nguồn:
<https://www.visualcapitalist.com/every-minute-internet-2020/>

Tại sao người ta cần đến Dữ liệu lớn?

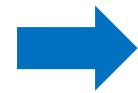


Những công nghệ phía sau Dữ liệu lớn

- Tính toán song song
- Lưu trữ khối lượng dữ liệu khổng lồ
- Phân tán dữ liệu
- Kết nối mạng tốc độ cao
- Máy tính hiệu suất cao
- Quản lý tác vụ và luồng xử lý
- Phân tích và khai thác dữ liệu
- Thu thập dữ liệu
- Học máy
- Trực quan hóa dữ liệu



Những công nghệ này đều
đã có từ lâu, vậy tại sao
đến nay Dữ liệu lớn mới
phát triển mạnh mẽ?

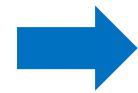


- Ngày càng nhiều dữ liệu được thu thập và lưu trữ
- Đã có những công cụ mã nguồn mở hỗ trợ
- Phần cứng hỗ trợ / Điện toán đám mây

Tại sao nhiều dữ liệu lại làm
nên sự thay đổi?

Câu chuyện bánh táo

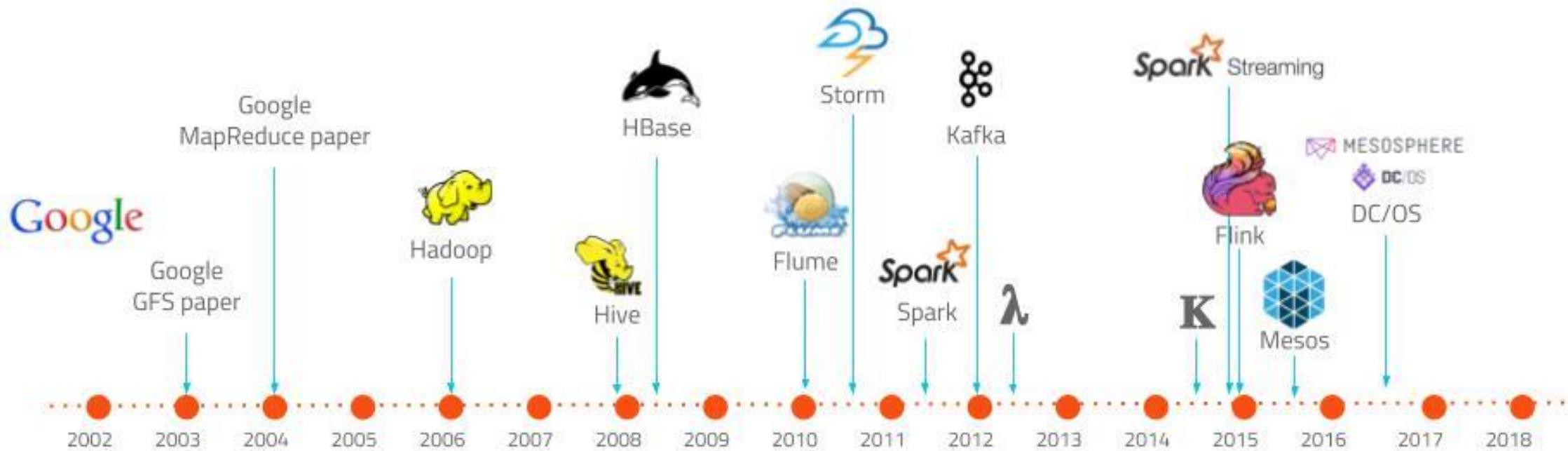




Với nhiều dữ liệu, chúng ta sẽ thấy được nhiều điều hơn:

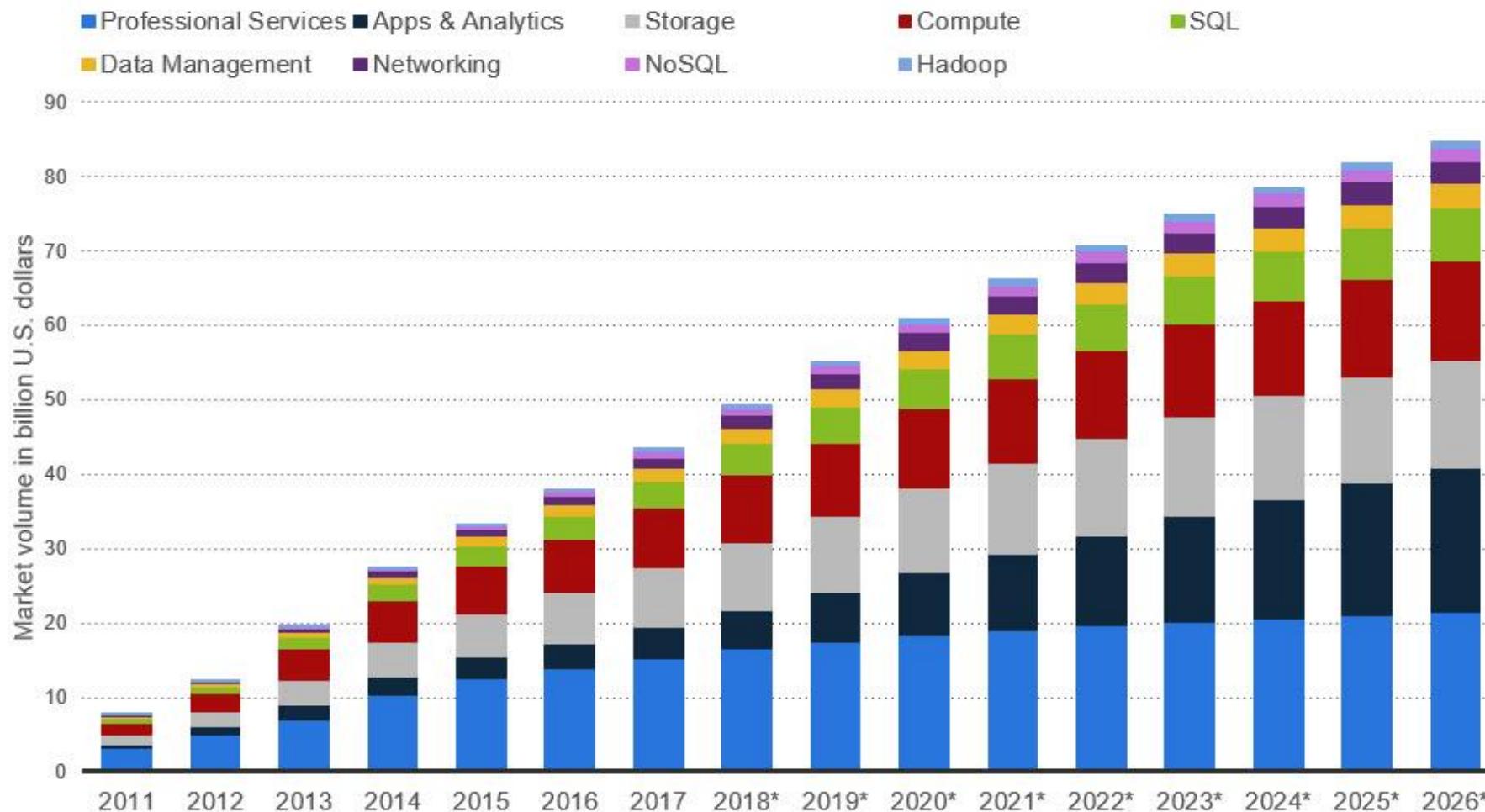
- Nhiều chi tiết mới hơn
- Nhiều điều tốt/có lợi hơn
- Những thông tin khác biệt

Quá trình phát triển



Big Data Market Worldwide Segment Revenue Forecast 2011-2026

Big Data Market Forecast Worldwide from 2011 to 2026, by segment (in billion U.S. dollars)



ĐỊNH NGHĨA

Dữ liệu lớn

Theo định nghĩa của **Gartner**: “**Big Data** là tài sản thông tin, mà những thông tin này có **khối lượng dữ liệu lớn, tốc độ cao** và dữ liệu **đa dạng**, đòi hỏi phải có công nghệ mới để xử lý hiệu quả nhằm đưa ra được các quyết định hiệu quả, khám phá được các yếu tố ẩn sâu trong dữ liệu và tối ưu hóa được quá trình xử lý dữ liệu”(*)

Đơn giản hơn thì 😎

Thuật ngữ “Big Data” là một **tập hợp dữ liệu rất lớn** mà các kỹ thuật điện toán thông thường **không** thể xử lý được.

Thuật ngữ “Big Data” không chỉ đề cập tới dữ liệu mà còn chỉ **cơ cấu tổ chức dữ liệu**, các **công cụ** và **công nghệ** liên quan.

N = Tất cả

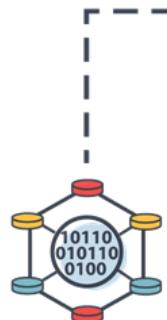
- Chúng ta không cần lấy mẫu nữa, chúng ta có toàn bộ tổng thể.
- Tránh được vấn đề sai số, độ lệch trong chọn mẫu thống kê truyền thống.

ĐẶC ĐIỂM

Dữ liệu lớn



BIG DATA



Volume



Value



Veracity



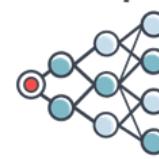
Visualization



Variety



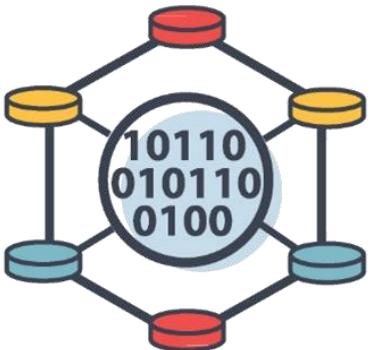
Velocity



Virality

3V's

Ba đặc điểm cơ bản nhất của BigData



Volume

Data Quantity



Variety

Data Types

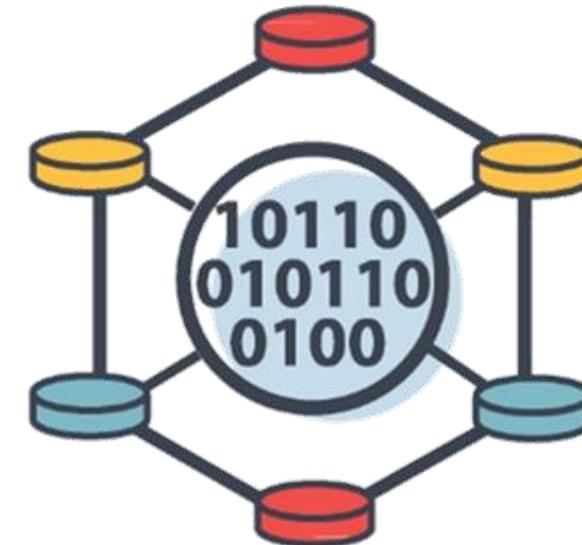


Velocity

Data Speed

Khối lượng dữ liệu **lớn**

Bao nhiêu thì được gọi là **lớn?**



Volume

Tốc độ mà dữ liệu **được sinh ra**

hay

Tốc độ mà dữ liệu cần **được xử lý** và **phân tích**



Velocity

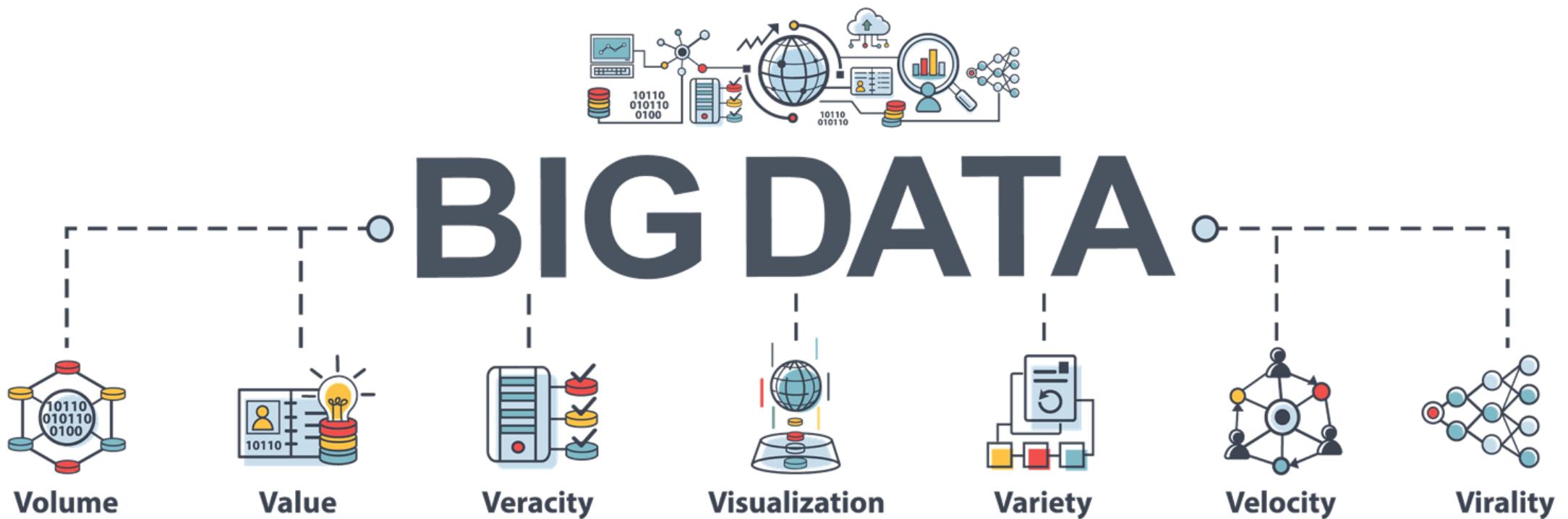
Tính **đa dạng** của dữ liệu



Variety

nV's

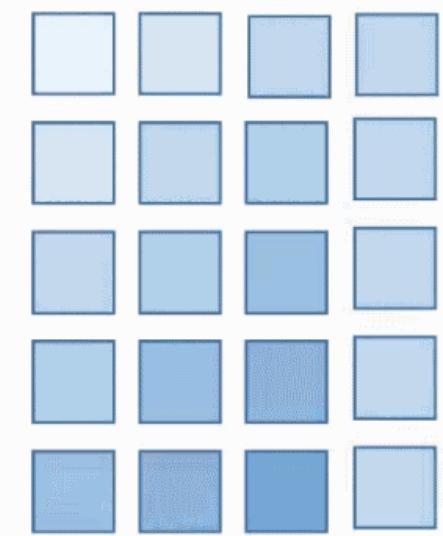
Các đặc điểm khác của BigData



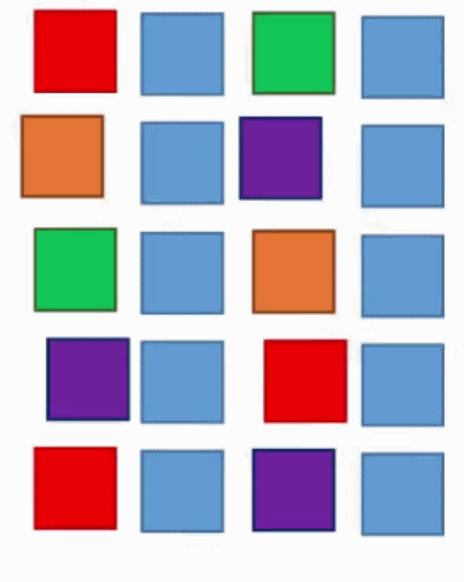
PHÂN LOẠI

Dữ liệu

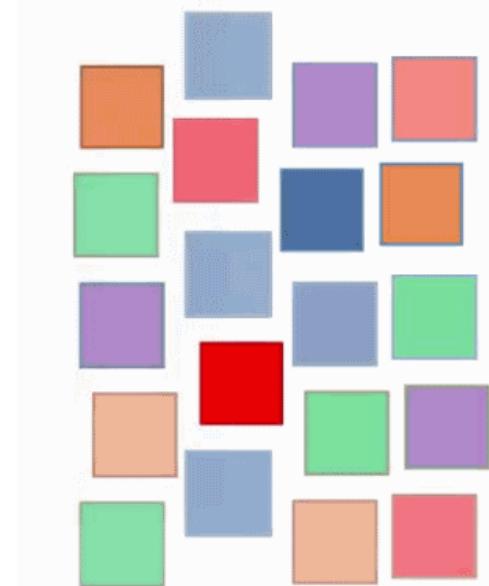




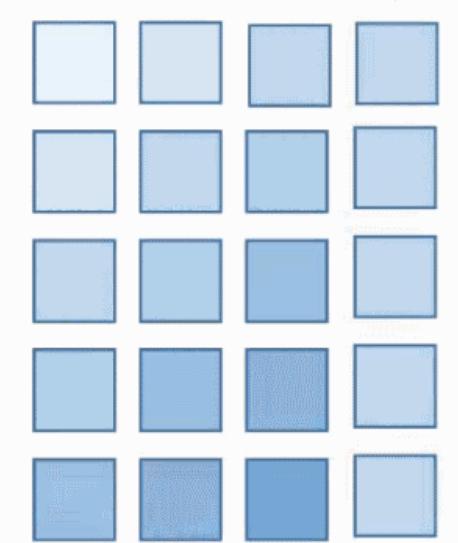
Có cấu trúc



Bán cấu trúc

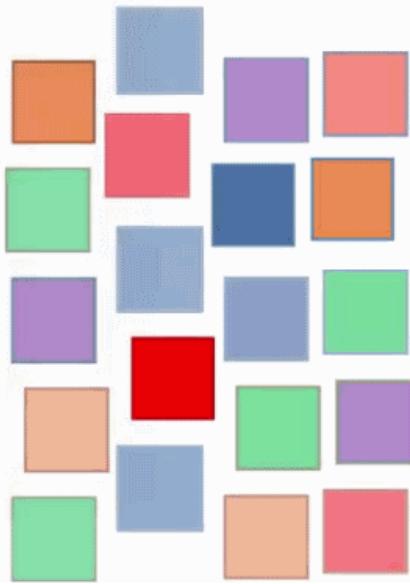


Phi cấu trúc



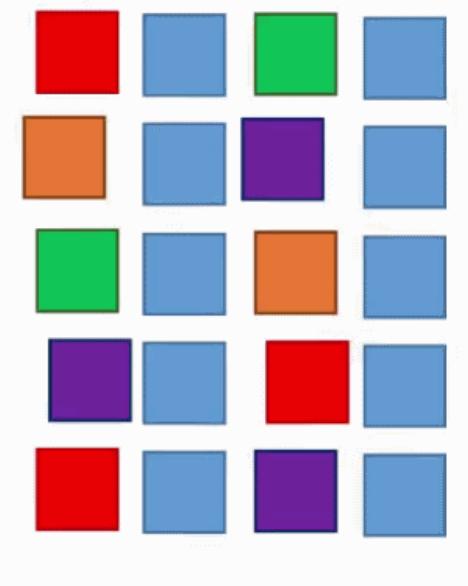
Có cấu trúc

- Dữ liệu có cấu trúc là dữ liệu có thể được lưu trữ và xử lý ở dạng cố định (lược đồ cố định).
- Ví dụ: cơ sở dữ liệu quan hệ, kho dữ liệu...



Phi cấu trúc

- Dữ liệu phi cấu trúc là những thông tin không được định nghĩa trước về mô hình dữ liệu hay cách thức tổ chức nội dung của dữ liệu.
- Phần lớn là những dữ liệu văn bản, tạo ra theo cách diễn đạt tự nhiên của con người. Tính bất thường và mơ hồ khiến dữ liệu phi cấu trúc khó xử lý.
- Ví dụ: văn bản, tài liệu, hình ảnh, âm thanh, video, logs, dữ liệu từ cảm biến...



Bán cấu trúc

- Dữ liệu bán cấu trúc không có mô hình dữ liệu chính thức, tuy nhiên nó có một số thuộc tính tổ chức như thẻ và các dấu hiệu khác để phân tách các phần tử ngữ nghĩa giúp dễ phân tích hơn.
- Ví dụ: tập tin XML, tài liệu JSON, email...

QUY TRÌNH XỬ LÝ

Dữ liệu lớn





Thu thập

Nguồn dữ liệu

- Chủ động, ví dụ: web crawler
- Bị động, ví dụ: video giám sát, luồng nhấp chuột

Truyền tải

- Chuyển đến nơi lưu trữ dữ liệu qua các kết nối tốc độ cao

Tiền xử lý

- Tích hợp, làm sạch, loại bỏ dư thừa...



Lưu trữ

Cơ sở hạ tầng lưu trữ

- Công nghệ lưu trữ, ví dụ: HDD, SSD
- Kiến trúc mạng, ví dụ: DAS, NAS, SAN

Quản lý dữ liệu

- Hệ thống tập tin (HDFS), key-value (Memcached), cơ sở dữ liệu hướng cột (Cassandra), cơ sở dữ liệu hướng tài liệu (MongoDB)...

Các mô hình lập trình

- MapReduce, xử lý luồng dữ liệu, xử lý đồ thị



Phân tích

Mục tiêu

- Mô tả, dự đoán

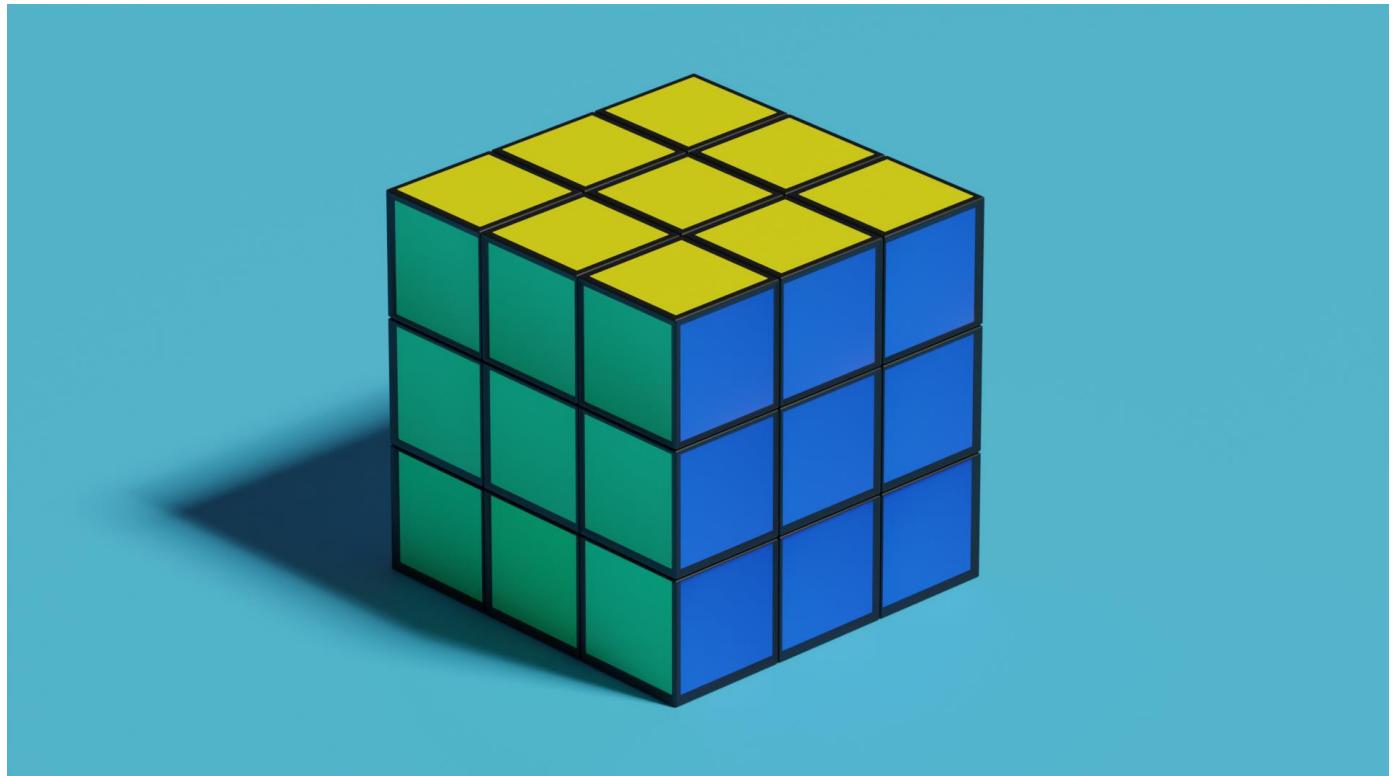
Phương pháp

- Phân tích thống kê, khai thác dữ liệu, khai thác văn bản, khai thác dữ liệu mạng và biểu đồ
- Gom cụm, phân lớp và hồi quy, phân tích liên kết

Các lĩnh vực đa dạng yêu cầu các kỹ thuật tùy chỉnh khác nhau

Những bài toán dữ liệu lớn

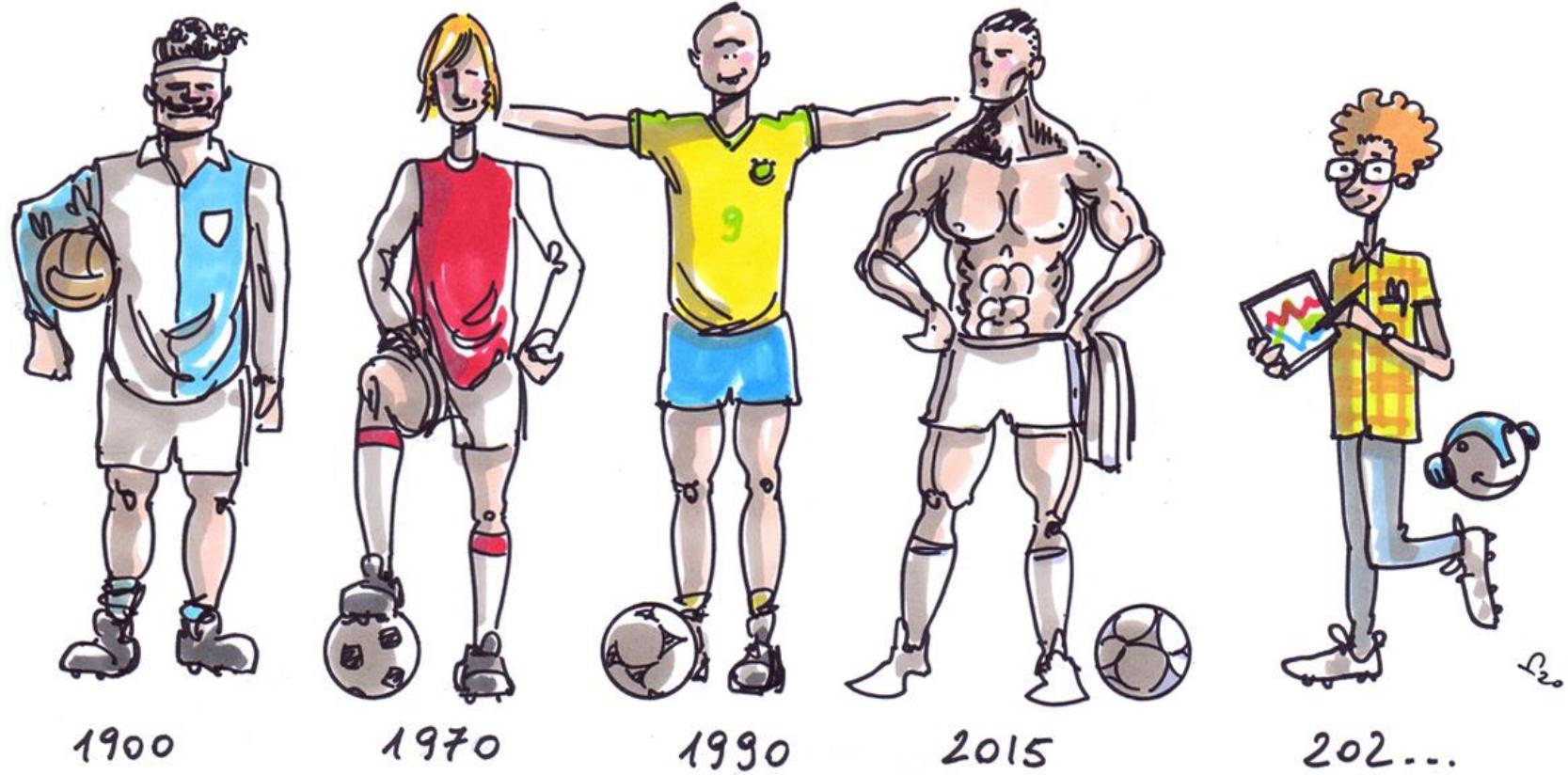
- Thu thập
- Quản lý
- Lưu trữ
- Tìm kiếm
- Chia sẻ
- Chuyển đổi
- Phân tích
- Trực quan hóa

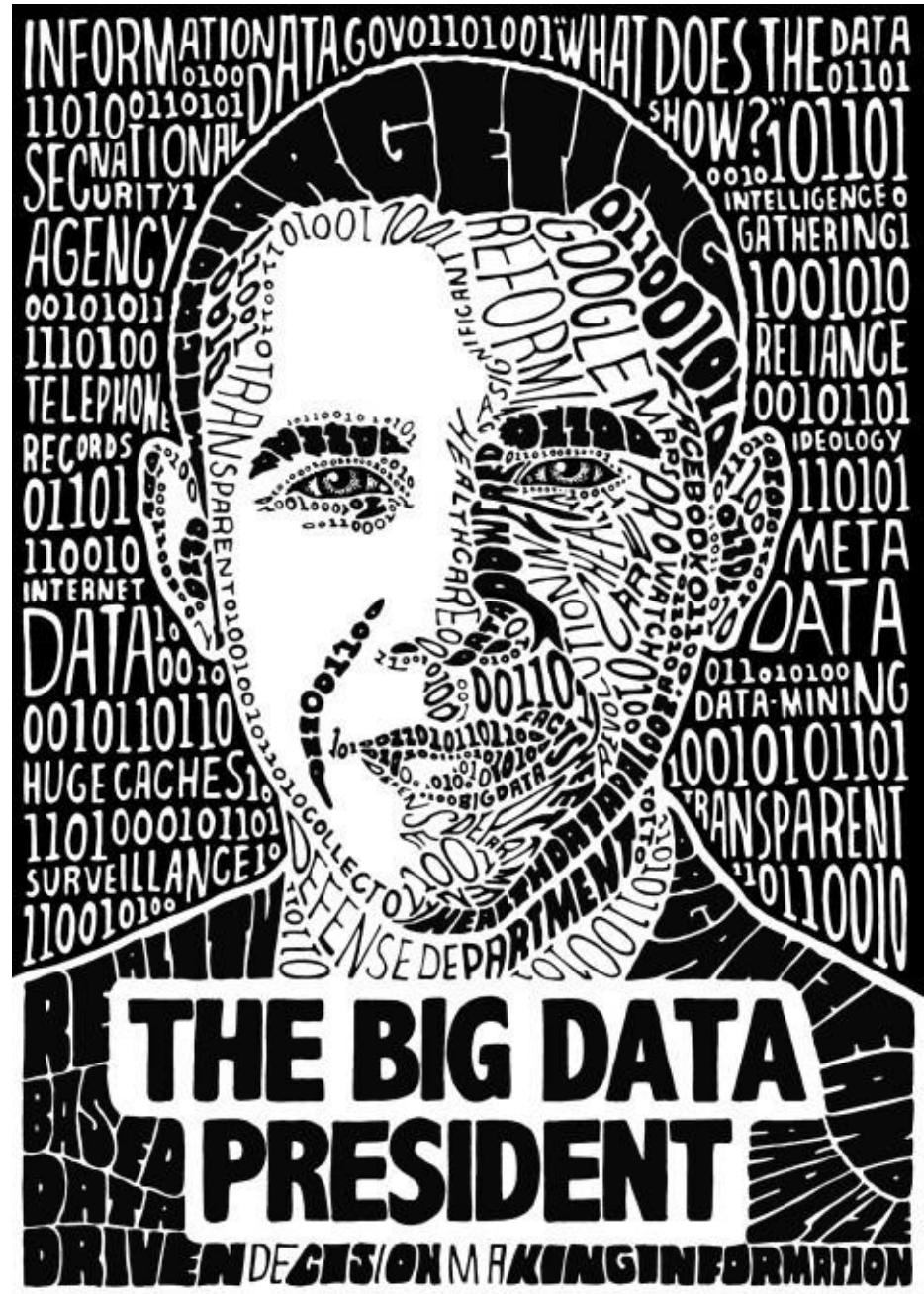


ỨNG DỤNG VÀ XU HƯỚNG

Dữ liệu lớn





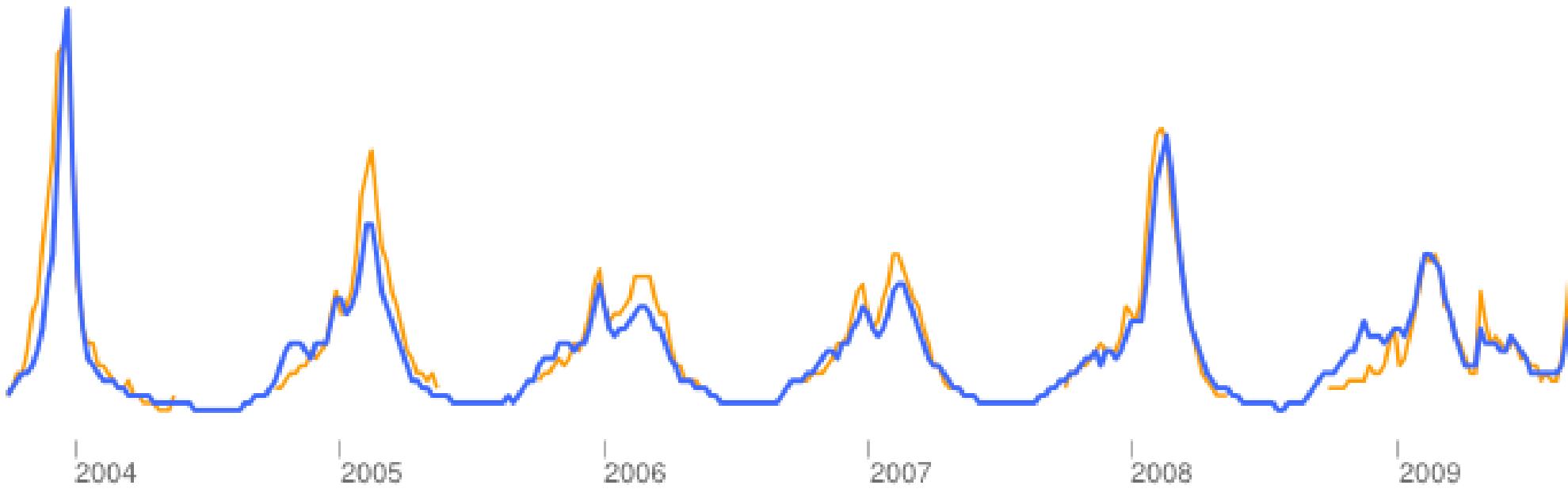




Bharatiya Janata Party भारतीय जनता पार्टी

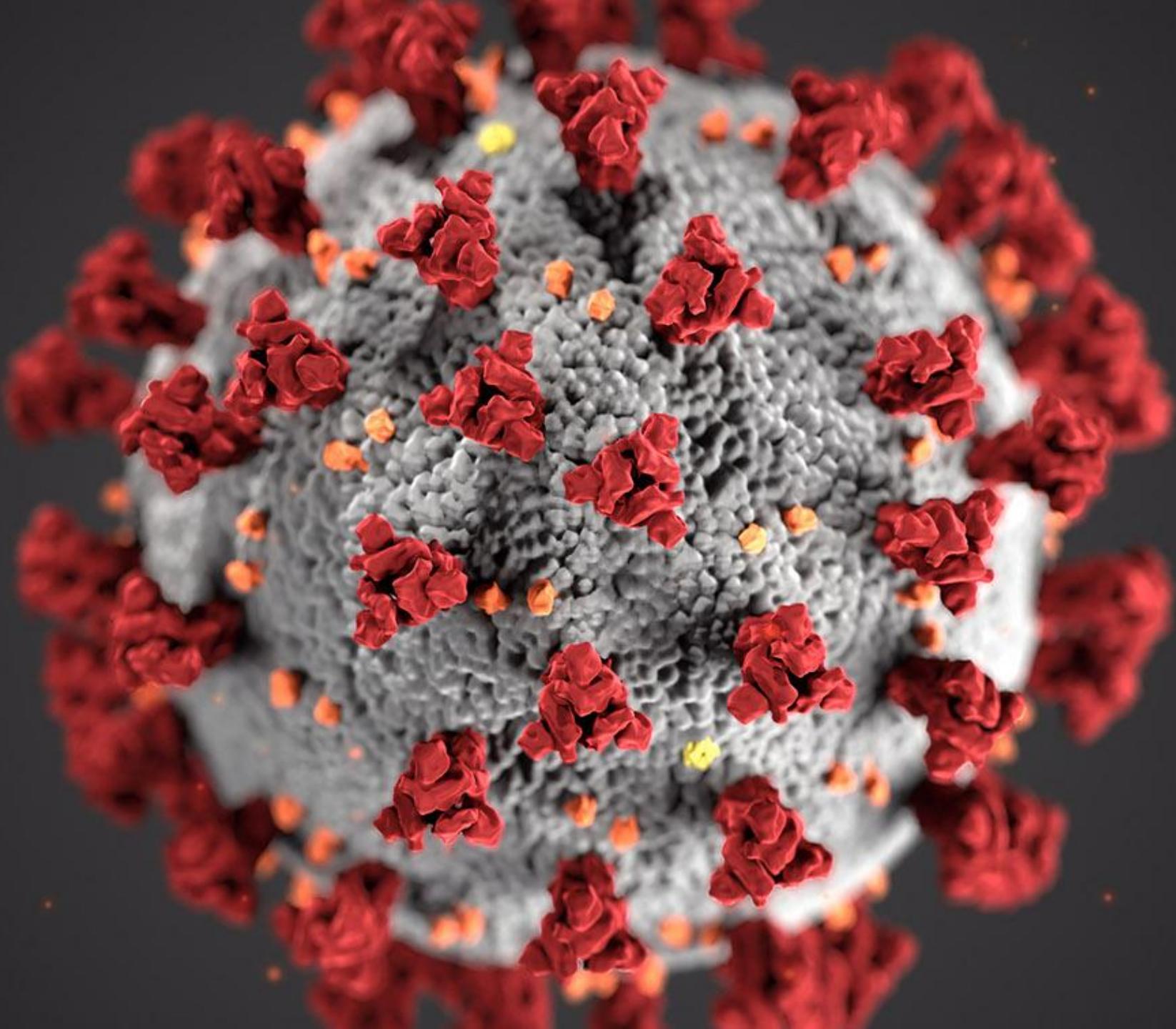
Đảng Nhân dân Ấn Độ (BJP) đã đánh bại Đảng Quốc đại trong cuộc Tổng tuyển cử quy mô lớn năm 2014, chấm dứt 10 năm cầm quyền của Đảng này, chiếm được đa số ghế tại Hạ viện.

Đây là chiến thắng vang dội nhất của Đảng BJP trong vòng 30 năm qua.



Google Flu Trends

<https://www.youtube.com/watch?v=6111nS66Dpk>





THÁCH THỨC & CƠ HỘI

Dữ liệu lớn

WHY ➔ **WHAT**

Thay đổi nhận thức và tổ chức xã hội



- Vấn đề Đạo đức học về dữ liệu: Ai là người sở hữu dữ liệu? Dữ liệu cá nhân sẽ được sử dụng như thế nào? Cần phải có những quy tắc quốc tế, tuy nhiên nếu quá khắt khe có thể gây trở ngại cho sự sáng tạo và việc sử dụng dữ liệu phục vụ cho các chính sách công.

Smart Cities



- Lợi ích và chiến lược phát triển quốc gia, hỗ trợ giải quyết những bài toán mang tầm vĩ mô như: môi trường, giao thông, dịch bệnh...
- Xây dựng xã hội thông minh



- Cạnh tranh, đổi mới, phát triển doanh nghiệp.
- Khởi nghiệp

- Khám phá mới cho khoa học



- Đòi hỏi con người phải tự trang bị những kiến thức, kỹ năng để khai thác được lợi thế từ dữ liệu lớn.
- Thay đổi, cập nhật liên tục.

Quá trình phơi bày “hồ sơ Panama”

Sau khi báo chí công bố 11,5 triệu tài liệu mang tên “hồ sơ Panama”, một loạt quốc gia trên thế giới đã tuyên bố tiến hành điều tra các thông tin liên quan tới vụ việc gây chấn động thế giới này.
Ngày 5/4, Thủ tướng Iceland đã tuyên bố từ chức sau vụ “Hồ sơ Panama”.



TÀI LIỆU THAM KHẢO

1. Tom White. 2015. **Hadoop: The Definitive Guide (4th ed.)**. O'Reilly Media, Inc.
2. Donald Miner and Adam Shook. 2012. **MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems (1st ed.)**. O'Reilly Media, Inc.
3. Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia. 2015. **Learning Spark: Lightning-Fast Big Data Analytics (1st ed.)**. O'Reilly Media, Inc.
4. Sandy Ryza, Uri Laserson, Sean Owen and Josh Wills. 2017. **Advanced Analytics with Spark: Patterns for Learning from Data at Scale (2nd ed.)**. O'Reilly Media, Inc.

Q & A