



# Samsung Innovation Campus

| Khóa học Big Data

Together for Tomorrow!  
**Enabling People**

Education for Future Generations

Chương 1.

# Giới thiệu về Big Data

| **Khóa học Big Data**

# Mô tả chương

## Mục tiêu:

- ✓ Đầu tiên chúng ta sẽ khám phá hành trình của con người với dữ liệu và xử lý dữ liệu trong suốt lịch sử của chúng ta.
  - Chúng ta sẽ có được góc nhìn về cách con người tiếp tục phát triển những cách mới để lưu trữ và xử lý dữ liệu
  - Những sự đổi mới và thay đổi quan trọng đã cho phép con người phát triển từ thời đại nông nghiệp sang thời đại thông tin
- ✓ Chúng ta đang ở đỉnh điểm của một sự chuyển dịch mô hình lớn khác và thay đổi cuộc chơi, điều này một lần nữa cách mạng hóa cách chúng ta làm việc với dữ liệu. Tại mỗi điểm nối như vậy, con người có thể tiến tới bước tiến hóa tiếp theo thông qua việc phát triển các giải pháp sáng tạo cho phép chúng ta thu được lợi ích
  - Big Data và quá trình xử lý dữ liệu được thiết lập để mang lại lợi ích mang tính cách mạng cho những người có thể khai thác các giá trị khó tiếp cận bên trong

## Nội dung:

1. Tổng quan và nền tảng của Big Data
2. Xu hướng hiện tại trong Big Data

Bài 1.

# Tổng quan và nền tảng của Big Data

| Giới thiệu về Big Data

Bài 1.

# Tổng quan và nền tảng của Big Data

## | 1.1. Dữ liệu trong lịch sử loài người

| 1.2. Dữ liệu đang biến đổi thế giới

| 1.3. Tại sao lại cần Big Data

| 1.4. Thay đổi cách xử lý Big Data

# Dữ liệu là gì?

[Thảo luận]

# Con người bắt đầu sử dụng nó từ khi nào?

Sự thật hoặc thông tin, đặc biệt khi được kiểm tra và sử dụng để tìm hiểu sự việc hoặc để đưa ra quyết định

-Từ điển Oxford Learners-



# Hành trình với dữ liệu và thông tin

| 150,000 năm trước

- ▶ Con người đã lưu giữ "hồ sơ" và thu thập "dữ liệu" một cách hiệu quả từ thời tiền sử.

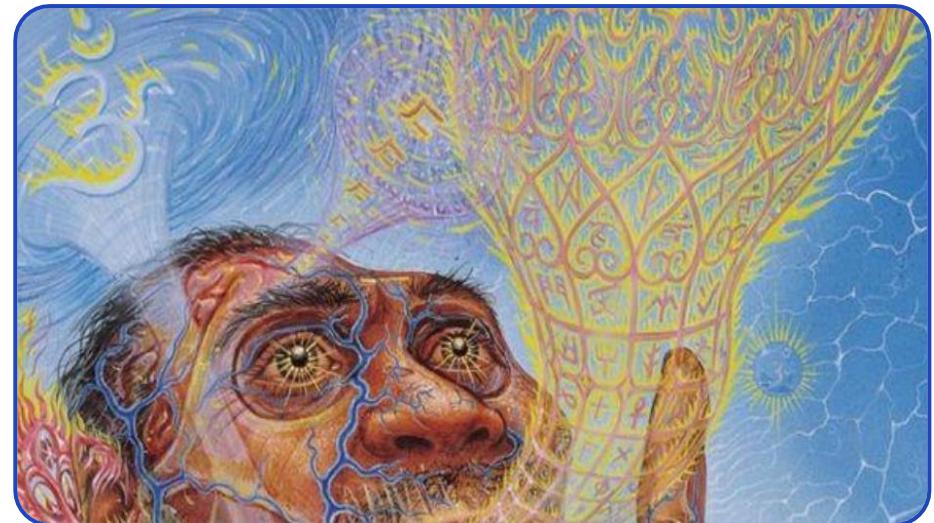
[Hình ảnh](#)

[Kể truyện](#)

[Thần thoại](#)

[Tập thể xã hội](#)

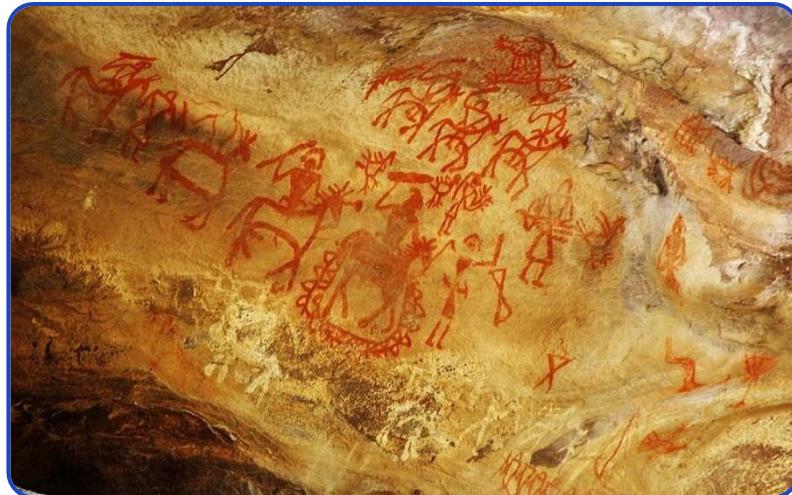
[Sự hợp tác](#)



# Hành trình với dữ liệu và thông tin

| 150.000 ~ 40.000 năm trước

- ▶ Tranh hang động thời tiền sử



| 40,000 ~ 12,000 năm trước

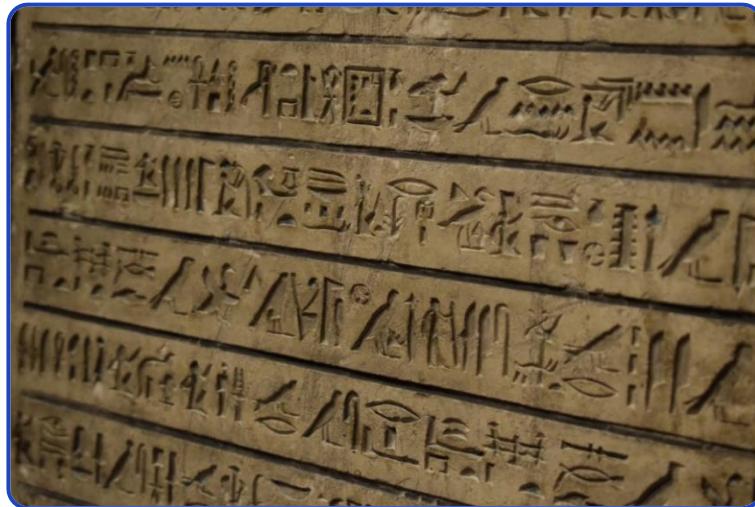
- ▶ Cách mạng nông nghiệp thông qua các bản vẽ thời tiền sử



### Các công cụ cổ xưa để lưu giữ hồ sơ và xử lý dữ liệu

| 12,000 ~ 5,000 năm trước

▶ Phiến đá



▶ Bàn tính cổ đại



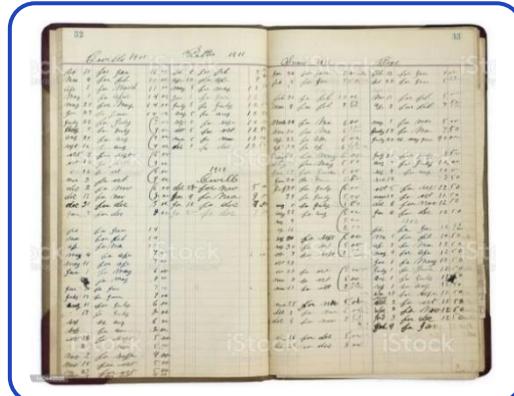
# Cuộc cách mạng công nghiệp

| 250 năm trước

### ► Sổ kế toán



### ► Hồ sơ kế toán

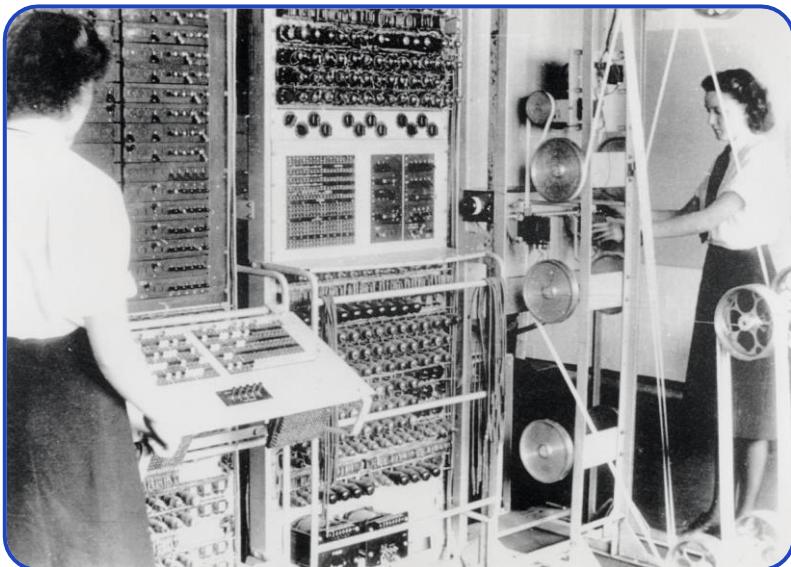


### ► Bản lưu trữ hồ sơ



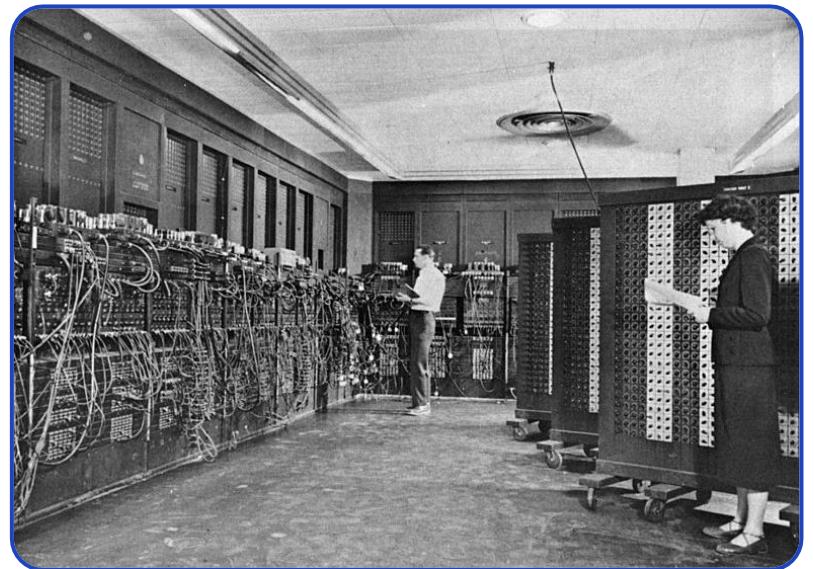
# Sự ra đời của Máy tính

| 250 ~ 80 năm trước



「Colossus」

Máy tính lập trình kỹ thuật số điện tử đầu tiên trên thế giới. Nó đã được sử dụng để phân tích mật mã.



「Eniac」

Máy tính kỹ thuật số thực sự đầu tiên. Nó được Quân đội Hoa Kỳ sử dụng để tính toán các bảng bắn pháo binh.

# Thời đại Thông tin

| Hiện tại: Chúng ta đang sống trong một Thế giới được kết nối!

- ▶ Dữ liệu và công nghệ thông tin đang thúc đẩy sự tiến bộ của con người trong mọi khía cạnh của cuộc sống.
- ▶ Lượng dữ liệu được tạo ra là không thể dò được.



# Tại sao Big Data xuất hiện

## | Trung tâm dữ liệu khổng lồ

- ▶ Cung cấp chi phí tính toán thấp hơn và lượng dữ liệu khổng lồ để phân tích

## | IoT

- ▶ Tạo ra lượng dữ liệu khổng lồ
- ▶ Điện thoại thông minh ở khắp mọi nơi

## | Điện toán đám mây

- ▶ Cung cấp khả năng mở rộng gần như vô hạn
- ▶ Cho phép xử lý Big Data theo thời gian thực



Bài 1.

# Tổng quan và nền tảng của Big Data

- | 1.1. Dữ liệu trong lịch sử loài người
- | **1.2. Dữ liệu đang biến đổi thế giới**
- | 1.3. Tại sao cần Big Data
- | 1.4. Thay đổi cách xử lý Big Data

# Dữ liệu trở thành nhiên liệu mới như thế nào?



# Các trường hợp sử dụng Big Data

Trường hợp nghiên cứu  
1



Công cụ tìm  
kiếm

Trường hợp nghiên cứu  
2



Thương mại  
điện tử

Trường hợp nghiên cứu  
3

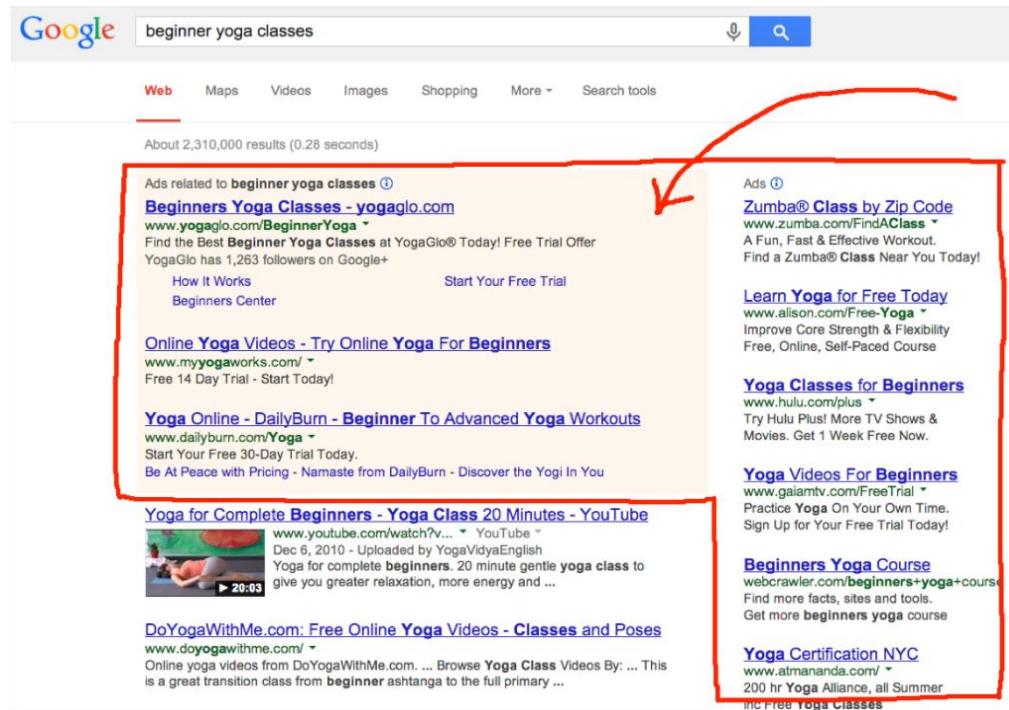


Nội dung kinh  
doanh

## [Trường hợp nghiên cứu 1] Công cụ tìm kiếm

### Google Ads

- Cho phép doanh nghiệp tiếp cận mọi người khi họ tìm kiếm từ khóa hoặc duyệt các trang web có chủ đề liên quan đến doanh nghiệp đó



## Google đã kiếm được bao nhiêu tiền?

Câu hỏi



Trong năm tài chính 2020, Google đã kiếm được bao nhiêu tiền?

Google đã kiếm được bao nhiêu tiền?

Câu trả lời



Trong năm tài chính 2020, Google đã kiếm được bao nhiêu tiền?

**181.69 Tỷ Đô la**

## [Trường hợp nghiên cứu 2] Thương mại điện tử

Amazon sử dụng dữ liệu để bán hàng theo nhu cầu của bạn

- ▶ Amazon hiện đang sử dụng tính năng lọc cộng tác giữa các mặt hàng, giúp chia tỷ lệ thành các tập dữ liệu lớn và đưa ra các đề xuất chất lượng cao theo thời gian thực
- ▶ 35% doanh số bán hàng từ các đề xuất sản phẩm

**35% doanh thu**



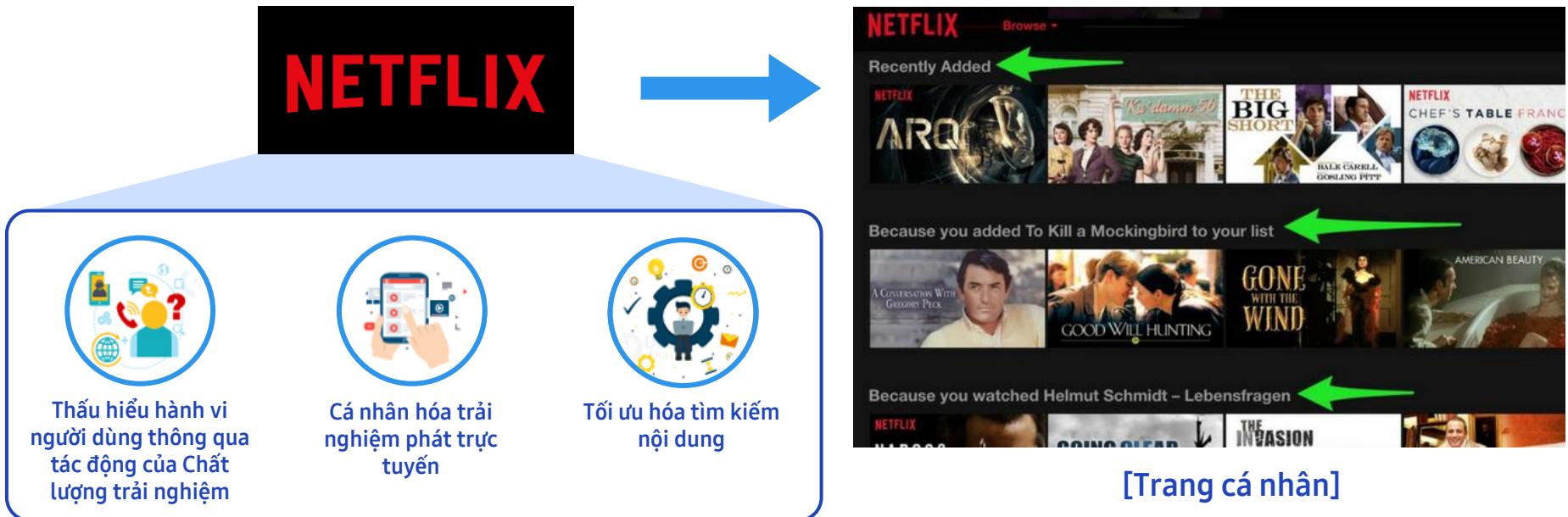
The screenshot shows the Amazon.com homepage with the search bar at the top. Below it, a large blue banner on the right side displays the text "Recommended for You". Underneath this, there are three book covers with "LOOK INSIDE!" buttons:

- Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop**
- Google Apps Administrator Guide: A Private-Label Web Workspace**
- Googlepedia: The Ultimate Google Resource (3rd Edition)**

### [Trường hợp nghiên cứu 3] Nội dung kinh doanh

#### Đề xuất video của Netflix

- ▶ 75% video phát trực tuyến từ các đề xuất



## Thay đổi người bảo vệ – Vốn hóa thị trường



Tài nguyên quý giá nhất thế giới không còn là dầu mỏ mà là dữ liệu

Bài 1.

# Tổng quan và nền tảng của Big Data

- | 1.1. Dữ liệu trong lịch sử loài người
- | 1.2. Dữ liệu đang biến đổi thế giới
- | **1.3. Tại sao cần Big Data**
- | 1.4. Thay đổi cách xử lý Big Data

# Big Data là gì?

[Thảo luận]

Thu thập dữ liệu từ SNS,  
Blog, Chat, Web?

Một cách để dự đoán  
tương lai?

Khai thác dữ  
liệu?

Công cụ giúp tăng doanh số  
và doanh thu?

Một cách để hiểu rõ hơn  
về khách hàng?



## Ý nghĩa của Big Data

Các tập dữ liệu cực lớn có thể được phân tích bằng máy tính để tiết lộ các mẫu, xu hướng và mối liên hệ, đặc biệt liên quan đến hành vi và tương tác của con người.

- Oxford Languages -

Big Data là tài sản thông tin có khối lượng lớn, tốc độ cao và/hoặc có tính đa dạng cao, đòi hỏi các hình thức xử lý thông tin đổi mới, hiệu quả về chi phí cho phép nâng cao hiểu biết sâu sắc, ra quyết định và tự động hóa quy trình.

- Gartner Glossary -

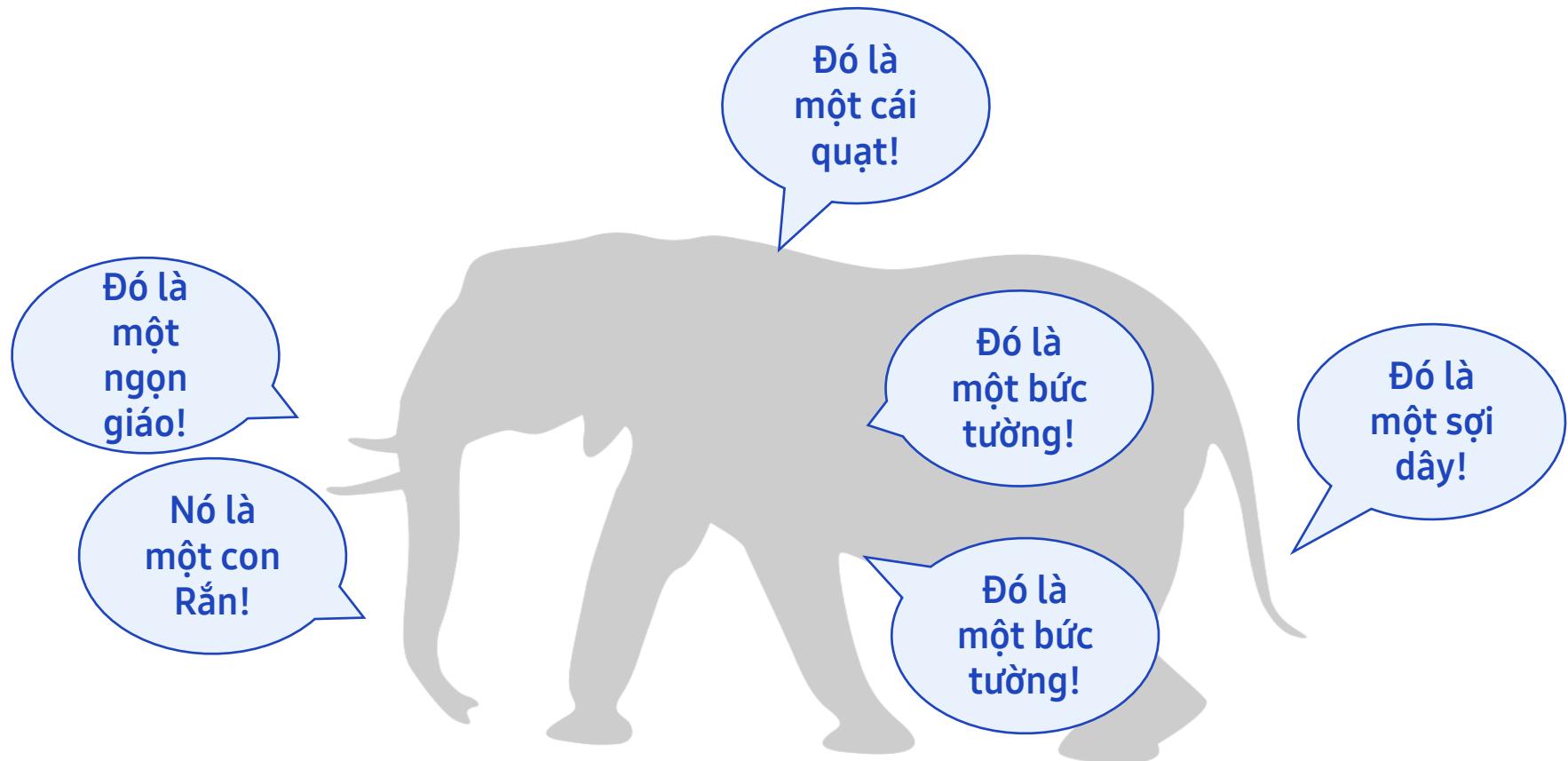
# Hiểu về Big Data

## I Big Data là gì? → Tại sao lại là Big Data?

- ▶ Cố gắng trả lời Big Data là gì có thể là một câu hỏi sai
- ▶ Cách tốt nhất để hiểu Big Data là gì để hỏi- "Tại sao Big Data ra đời?"
- ▶ Nó đang cố gắng giải quyết những điểm đau nào?

# Quan điểm về Big Data

[Thảo luận]



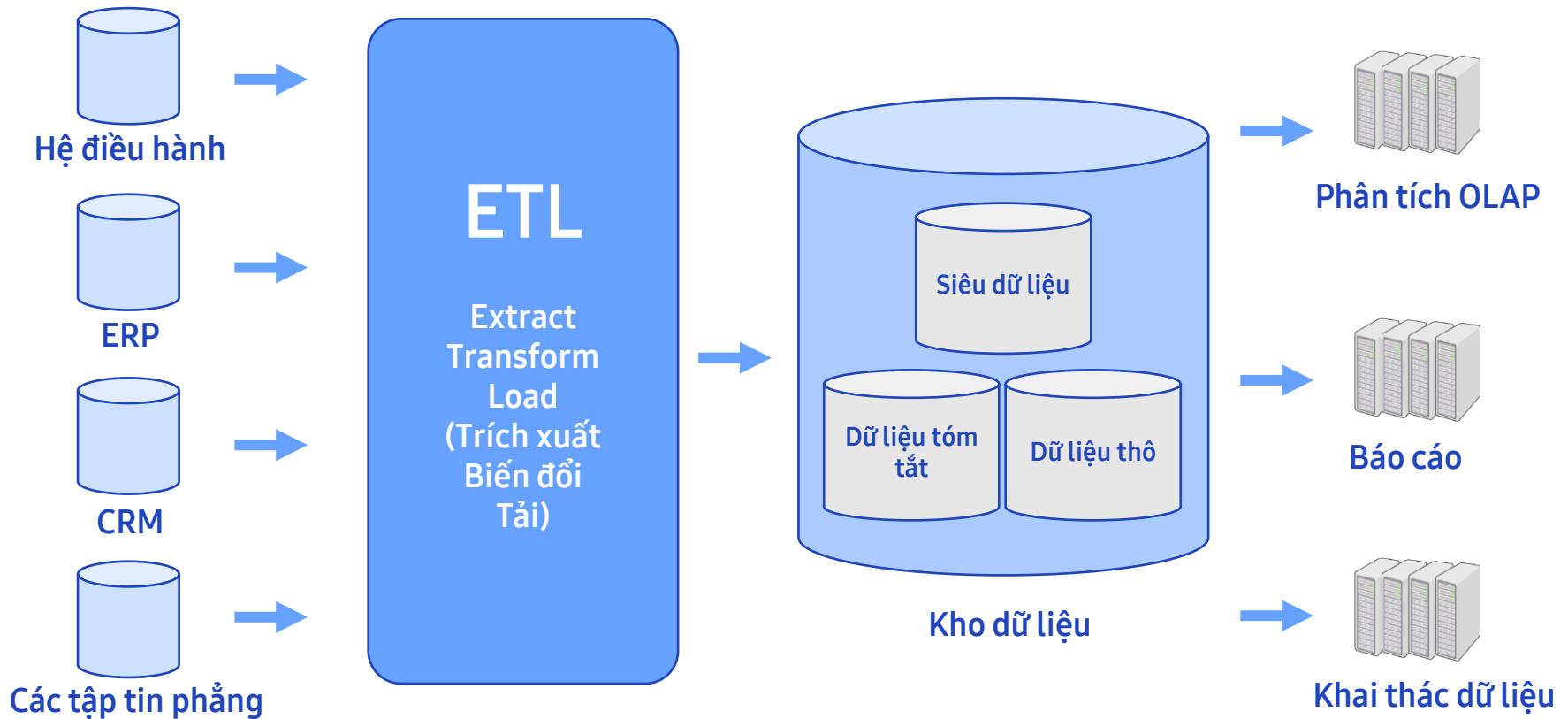
# Điểm đau là gì?

## | Điểm đau chính của trường dữ liệu

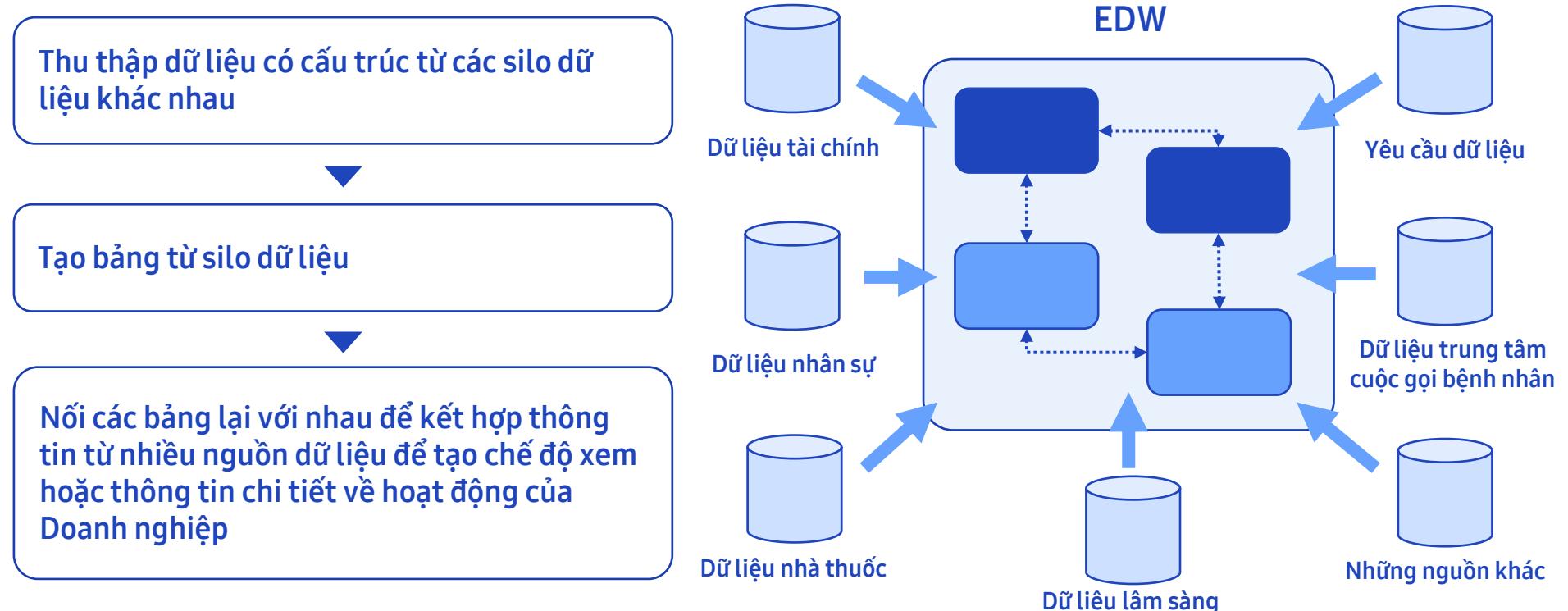
- ▶ Phương tiện lưu trữ
- ▶ Khả năng mở rộng
- ▶ Kết nối mạng
- ▶ Chuyển đổi dữ liệu dưới mức tối ưu
- ▶ Bảo mật dữ liệu



## Xử lý dữ liệu trước Big Data



## Silo dữ liệu và kho dữ liệu doanh nghiệp



# Hệ thống quản lý cơ sở dữ liệu quan hệ

### Dữ liệu có cấu trúc

- ▶ Nó tuân thủ một mô hình dữ liệu được xác định trước và do đó rất dễ phân tích.
- ▶ Nó phù hợp với định dạng bảng có mối quan hệ giữa các hàng và cột khác nhau.

### Ứng dụng RDBMS



## Sự phát triển của kỷ nguyên Dot Com

| World Wide Web trở nên phổ biến và các công ty Internet

Google



facebook

amazon



YouTube

Instagram

TikTok

yahoo!

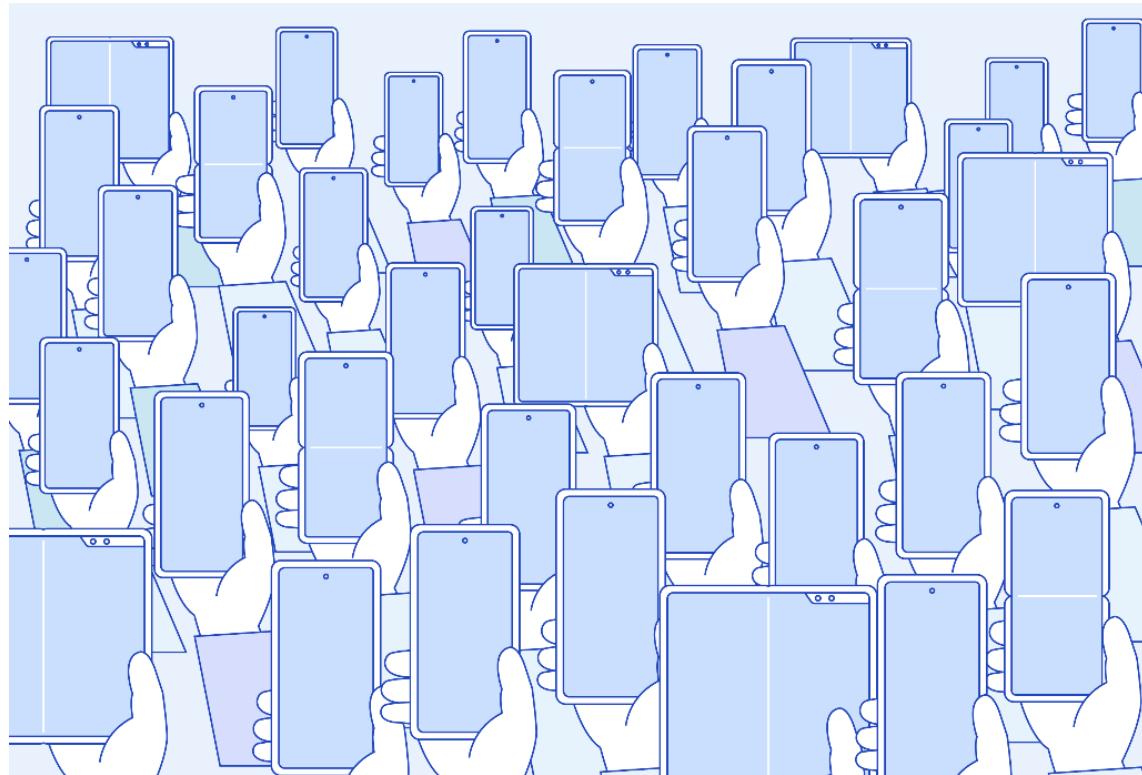
## Các trang web hàng đầu tạo ra các loại dữ liệu khác nhau

- ▶ Văn bản và Siêu văn bản: hộp cuộn, Siêu văn bản, Đồng bộ hóa với âm thanh, Đã sao chép, Bản nhạc có thể tìm kiếm, Chú thích cho phim
- ▶ Video: video tiêu chuẩn, ảnh toàn cảnh thực tế ảo và các đối tượng thực tế ảo
- ▶ Hoạt ảnh
- ▶ Đồ họa: Đồ họa bitmap 2D, đồ họa dựa trên vector 2D, hình ảnh 3D
- ▶ Âm thanh: Âm thanh số hóa, MIDI

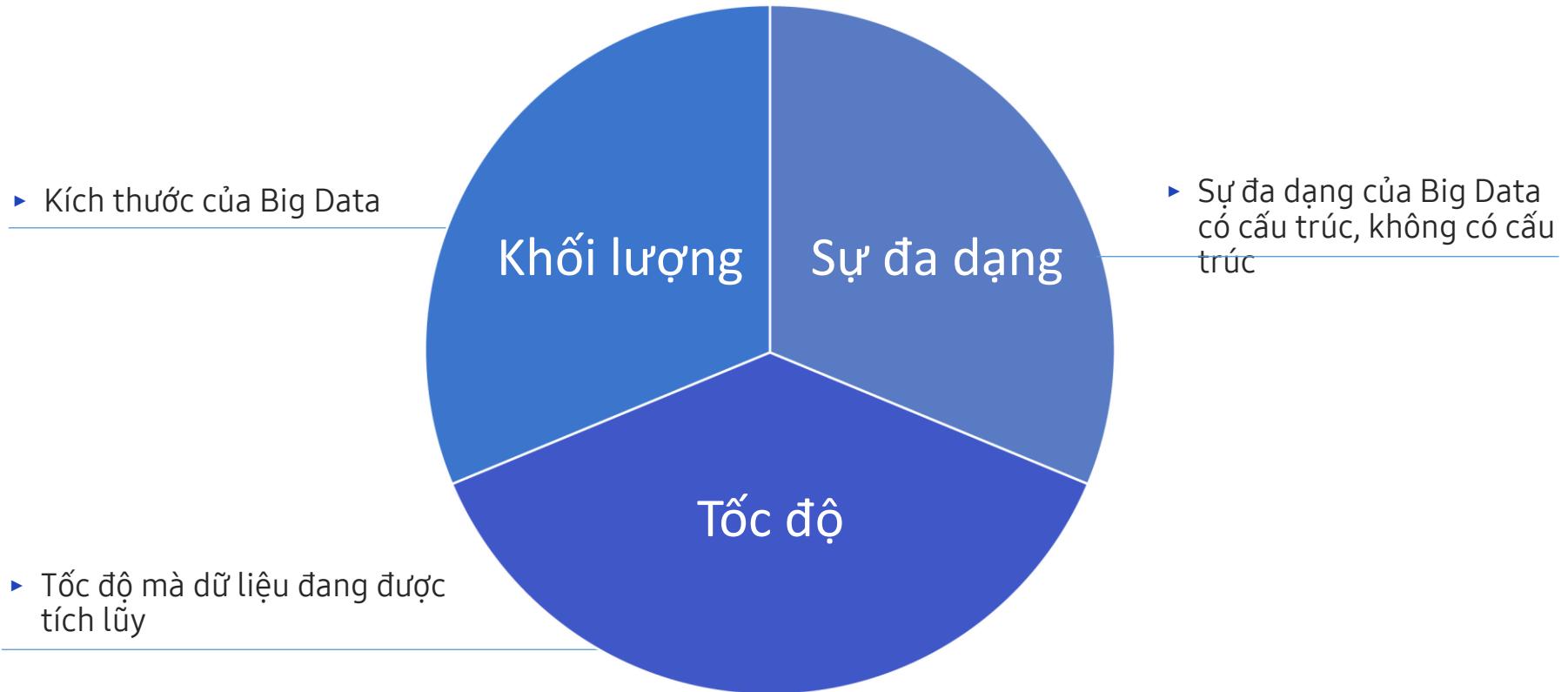


# Điện thoại thông minh ở mọi nơi

| Lượng dữ liệu di động cũng tăng vọt



## Đặc điểm của Big Data



# Tăng Định lượng Dữ liệu - Khối lượng

## I Bao nhiêu dữ liệu được tạo ra mỗi ngày

- ▶ Vào năm 2020, mọi người tạo ra 1,7 MB dữ liệu mỗi giây.
- ▶ Đến năm 2022, 70% GDP toàn cầu sẽ được số hóa.
- ▶ Vào năm 2021, 68% người dùng Instagram xem ảnh của các thương hiệu.
- ▶ Đến năm 2025, hơn 200 zettabyte dữ liệu sẽ được lưu trữ trên đám mây trên toàn cầu.
- ▶ Vào năm 2020, người dùng đã gửi khoảng 500.000 Tweet mỗi ngày.
- ▶ Đến cuối năm 2020, 44 zettabyte sẽ tạo nên toàn bộ vũ trụ kỹ thuật số.
- ▶ Mỗi ngày có 306,4 tỷ email được gửi và 500 triệu Tweet được tạo.

Dữ liệu được tạo ra hàng ngày, ước tính hiện tại là 1,145 nghìn tỷ MB mỗi ngày.

# Thay đổi hồ sơ dữ liệu – Sự đa dạng

## Thay đổi hồ sơ dữ liệu

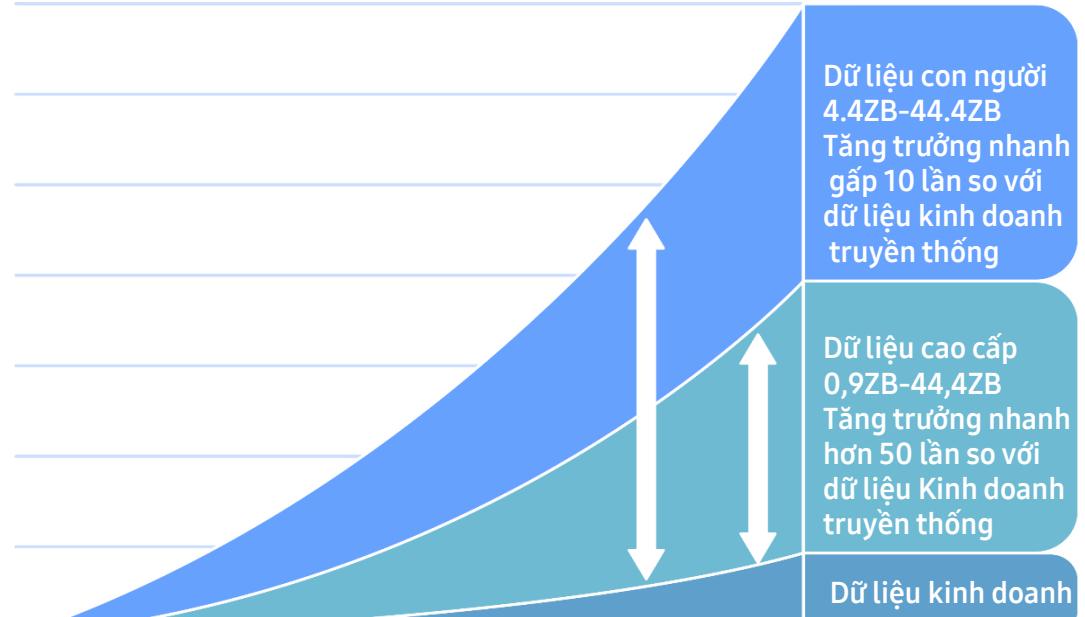
- ▶ Dữ liệu giao dịch
- ▶ Dữ liệu khoa học
- ▶ Dữ liệu cảm biến
- ▶ Dữ liệu truyền thông xã hội
- ▶ Dữ liệu doanh nghiệp
- ▶ Dữ liệu công cộng



# Tràn ngập dữ liệu – Tốc độ

## Tăng tốc độ tạo dữ liệu

- ▶ Dữ liệu do con người và máy tạo ra đang có tốc độ tăng trưởng nhanh hơn gấp 10 lần so với dữ liệu kinh doanh truyền thống
- ▶ Dữ liệu máy thậm chí còn tăng nhanh hơn với tốc độ tăng trưởng gấp 50 lần



Nguồn: Bên trong Big Data

Điểm đau là việc xử lý dữ liệu truyền thống thông qua RDBMS / EDW không còn khả thi nữa

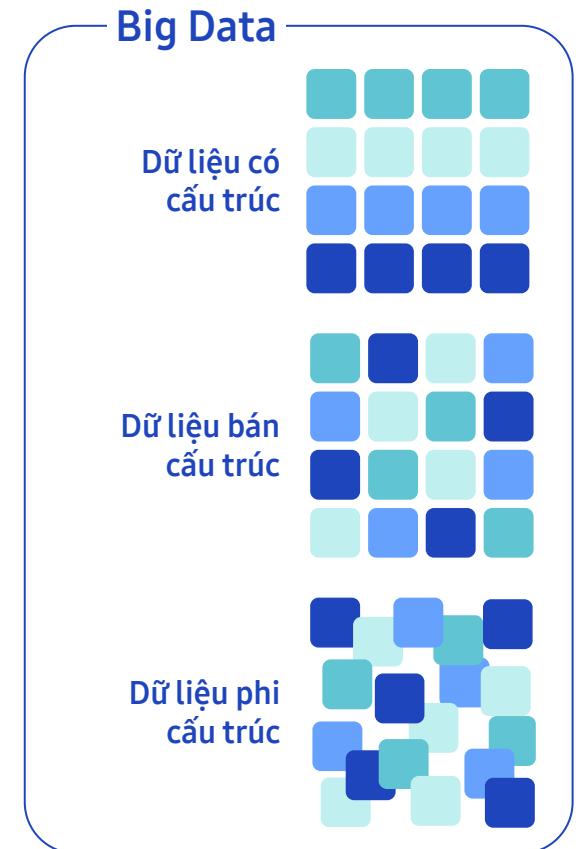
# Sự đa dạng làm xáo trộn cấu trúc thống nhất

## | Cấu trúc, không cấu trúc, bán cấu trúc chuẩn mực mới

- ▶ Các hệ thống RDBMS không thể xử lý dữ liệu bán cấu trúc và phi cấu trúc
- ▶ Dữ liệu có cấu trúc: Địa chỉ, Ngày, Số (Điện thoại, Mã Zip, v.v.), Văn bản, Hầu hết Dữ liệu CRM
- ▶ Dữ liệu bán cấu trúc: E-Mails, Ngôn ngữ đánh dấu, EDI, XML
- ▶ Dữ liệu phi cấu trúc: Âm thanh, Tệp văn bản, Dữ liệu mạng xã hội, Dữ liệu nhật ký, Hoạt động trên thiết bị di động

## | Dữ liệu phi cấu trúc chiếm 80% và hơn thế nữa dữ liệu doanh nghiệp

- ▶ Dữ liệu có cấu trúc theo truyền thống dễ dàng hơn đối với các ứng dụng Big Data, nhưng các giải pháp phân tích dữ liệu ngày nay đang có những bước tiến lớn trong lĩnh vực dữ liệu phi cấu trúc



# Khối lượng làm giảm hiệu suất

## Hiệu suất theo quy mô

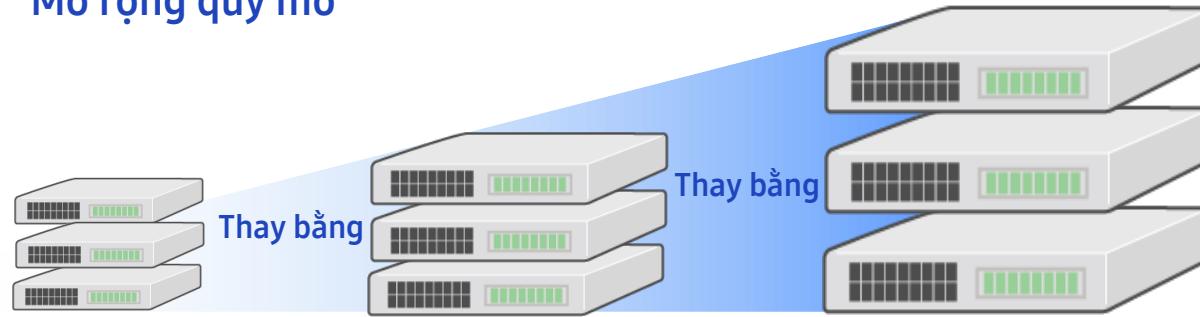
- ▶ Các hệ thống RDBMS chưa bao giờ được thiết kế để xử lý lượng dữ liệu thực sự lớn
- ▶ Chúng hoạt động tốt nhất khi kích thước của dữ liệu nằm trong phạm vi Gigabyte hoặc nhiều nhất là Terabyte
- ▶ Doanh nghiệp đứng trước hàng Petabyte dữ liệu cần xử lý
- ▶ Hiệu suất truy cập và xử lý trở nên tồi tệ hơn khi nói đến Dữ liệu lớn

## Tốc độ tạo nên nhu cầu về khả năng mở rộng

### Khả năng mở rộng

- ▶ Để theo kịp tốc độ ngày càng tăng của dữ liệu được tạo, Big Data yêu cầu các hệ thống lưu trữ cần mở rộng.
- ▶ Đây không phải là vấn đề trong quá trình xử lý EDW truyền thống: thông thường, dữ liệu cũ hơn và ít liên quan hơn sẽ bị xóa để nhường chỗ cho dữ liệu mới hơn thay vì chia tỷ lệ.
- ▶ RDBMS phụ thuộc nhiều vào phần cứng đắt tiền, vì vậy có thể thực hiện được nhưng không phù hợp để mở rộng quy mô.

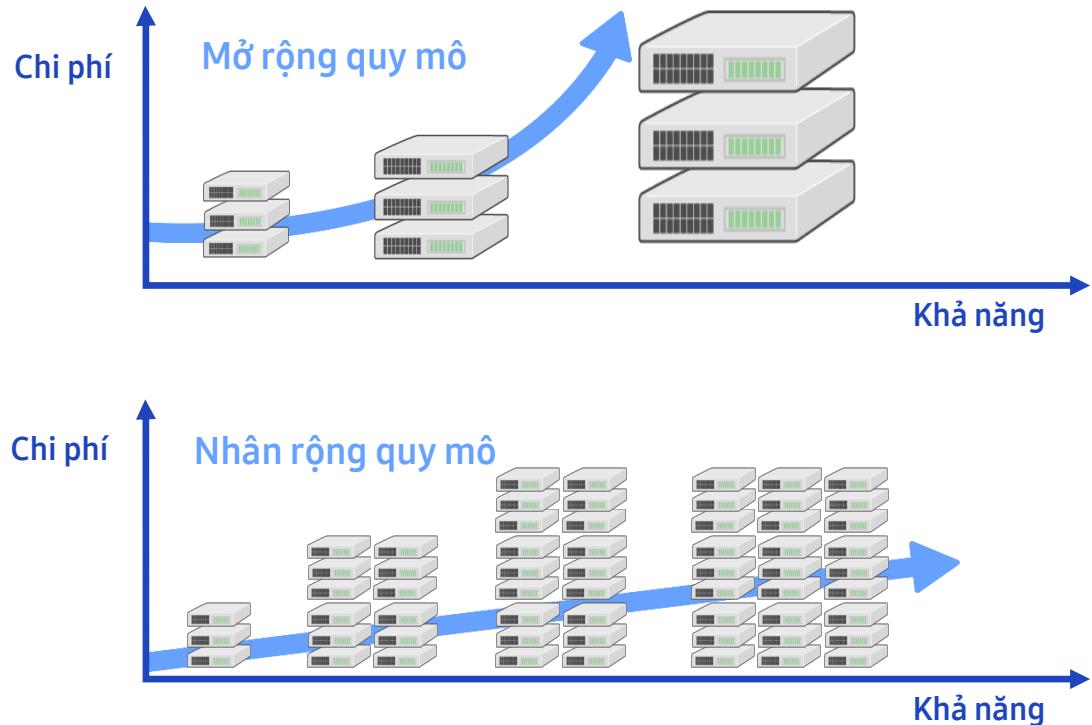
### Mở rộng quy mô



# Tốc độ tạo ra chi phí theo cấp số nhân

## Chi phí

- ▶ Các hệ thống RDBMS được coi là hệ thống Mở rộng quy mô: mỗi thế hệ cần được thay thế bằng các hệ thống phần cứng phức tạp và đắt tiền hơn bao giờ hết
- ▶ Hệ thống mở rộng quy mô có hiệu quả chi phí thấp hơn đáng kể khi có phạm vi dữ liệu được lưu trữ từ 100TB đến 1PB
- ▶ Ngược lại, hệ thống mở rộng quy mô duy trì hiệu quả chi phí không đổi bất kể kích thước của dữ liệu



Bài 1.

# Tổng quan và nền tảng của Big Data

- | 1.1. Dữ liệu trong lịch sử loài người
- | 1.2. Dữ liệu đang biến đổi thế giới
- | 1.3. Tại sao cần Big Data
- | 1.4. Thay đổi cách xử lý Big Data**

# Vai trò của một kỹ sư dữ liệu

## | Kỹ sư giống như một đầu bếp

- ▶ Các kỹ sư tận dụng các nguyên liệu sẵn có hiện tại để tạo ra những sản phẩm tốt nhất
- ▶ Một sản phẩm hoặc hệ thống thường không thể hỗ trợ các thành phần đắt tiền nhất hiện có
- ▶ Các thành phần mới liên tục có sẵn hoặc trở nên có sẵn với chi phí hợp lý
- ▶ Thỏa hiệp phải được thực hiện đối với các lựa chọn thành phần và sau đó các giải pháp phải được thiết kế để khắc phục mọi thiếu sót vốn có trong các lựa chọn đó



# Cách đối phó với Big Data

## Yêu cầu của nền tảng dữ liệu lớn

- ▶ Chi phí ban đầu phải hợp lý và có thể quản lý được và khi hệ thống phát triển
- ▶ Khả năng lưu trữ và xử lý lượng dữ liệu khổng lồ (petabyte++) trong thời gian hợp lý
- ▶ Vì quá trình dồn dữ liệu không chậm lại nên có thể phát triển khi các yêu cầu về lưu trữ và xử lý tăng lên trong tương lai
- ▶ Dữ liệu không thể bị mất. Hệ thống phải có khả năng chịu lỗi.

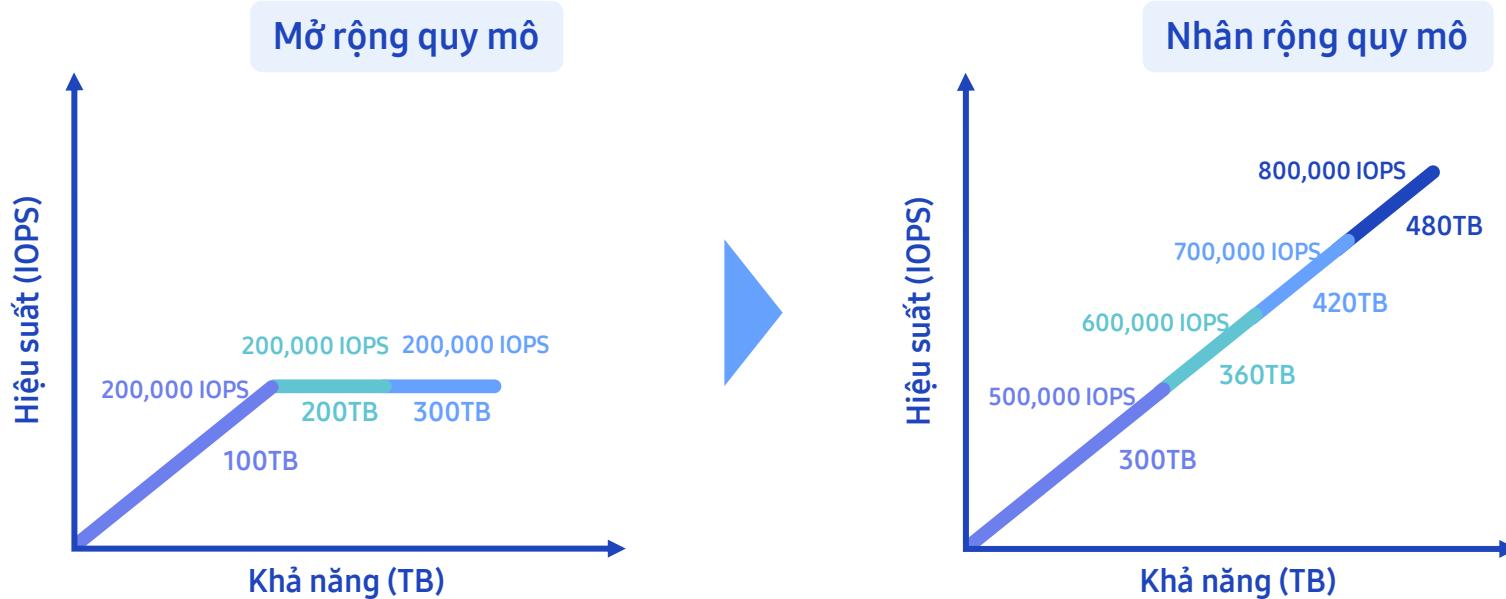
Giải pháp là kiến trúc phân tán song song

Hiệu quả chi phí	Hiệu suất
Khả năng mở rộng	Khả dụng

# Hiệu quả chi phí

## Sử dụng phần cứng tiết kiệm

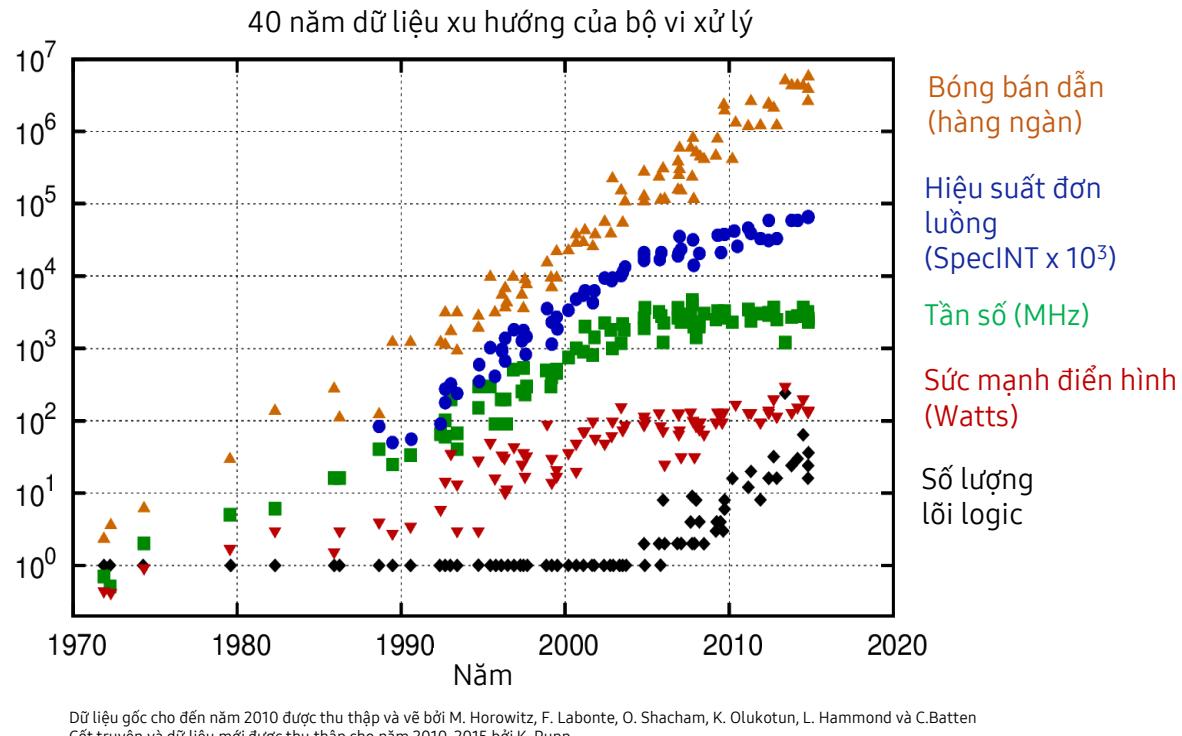
- Chuyển đổi từ Mở rộng quy mô (Scale-up) sang Nhân rộng quy mô (Scale-out) là cách duy nhất để hỗ trợ tăng trưởng dung lượng lưu trữ theo cách tiết kiệm chi phí



# Hiệu suất

## Bộ xử lý đa lõi và nhanh hơn

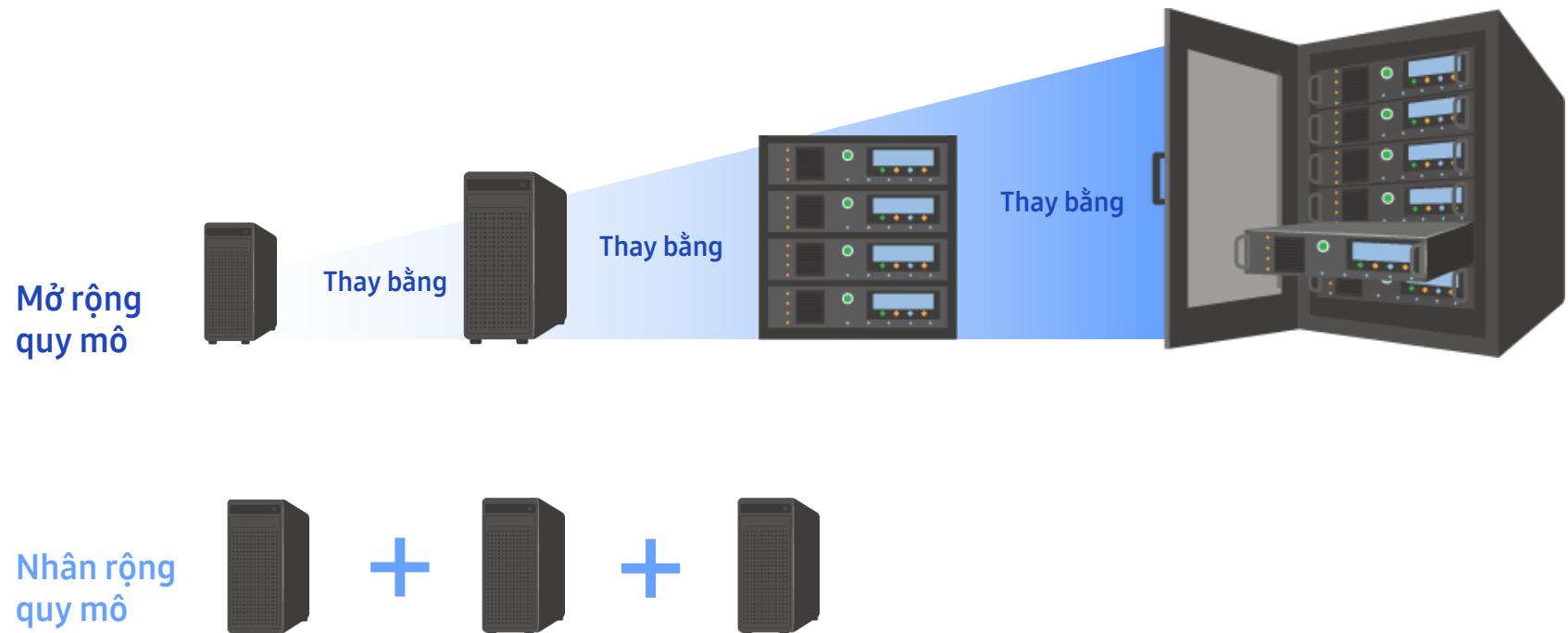
- ▶ Tần suất không còn tăng do giới hạn vật lý.
- ▶ Điều này liên quan đến thực tế là hiệu suất của một luồng ngừng tăng cùng một lúc.
- ▶ Thay vào đó, nhiều lõi logic được giới thiệu trong bộ xử lý.



# Khả năng mở rộng

[Thảo luận]

| Tại sao bạn nghĩ rằng quy mô lưu trữ chuyển sang Nhân rộng quy mô từ Mở rộng quy mô?

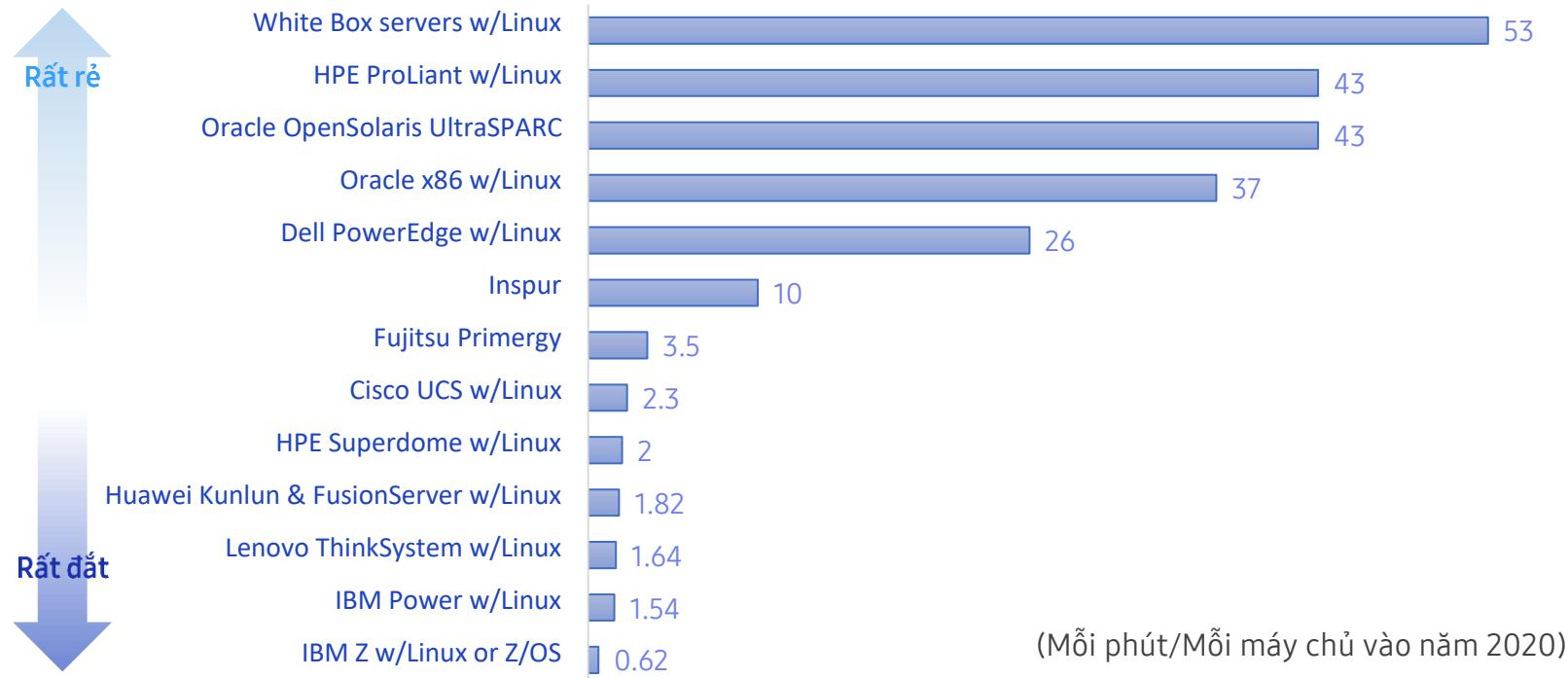


# Khả dụng (1/2)

[Thảo luận]

| Bạn sẽ sử dụng máy chủ nào trong số những máy chủ này để xây dựng kiến trúc Big Data?

「 Thời gian ngừng hoạt động ngoài kế hoạch do nền tảng phần cứng máy chủ 」



# Khả dụng (2/2)

## | Carrier class server to Commodity class server

- ▶ Thời gian ngừng hoạt động ngoài kế hoạch của các máy chủ có thể thay đổi đáng kể.
- ▶ Thời gian chết càng thấp thì chi phí phần cứng càng đắt.
- ▶ Khi sử dụng các máy chủ hạng phổ thông làm môi trường phân tán, các hệ thống cần có khả năng xử lý mất dữ liệu.

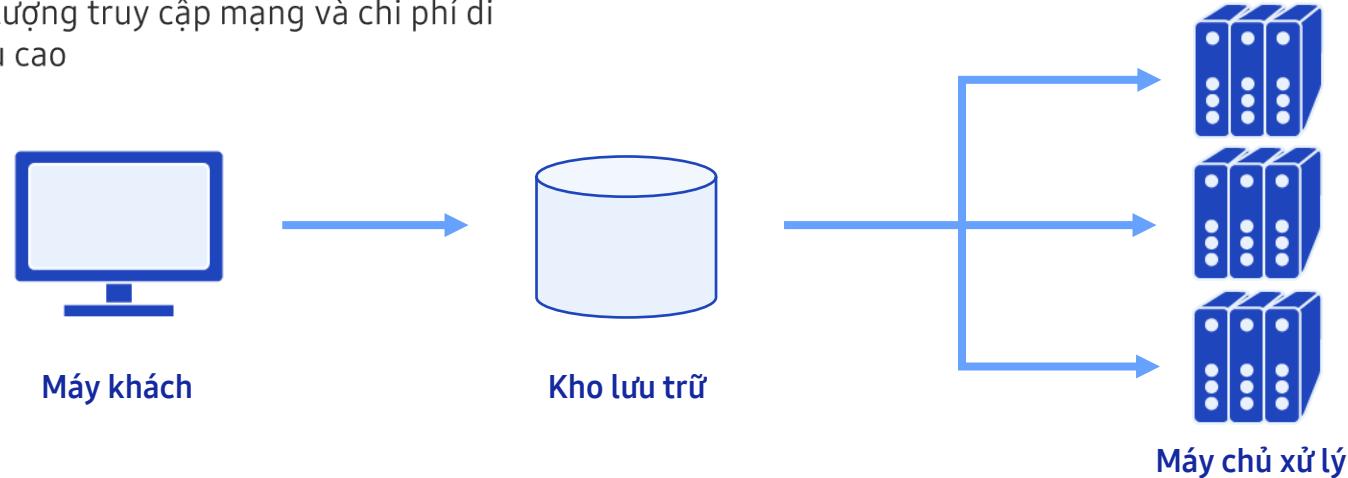
## | Kiến trúc Client-Server sang kiến trúc Parallel Cluster

- ▶ Lưu trữ dữ liệu với bộ xử lý
- ▶ Sao chép dữ liệu để có khả năng chịu lỗi và tăng tính khả dụng

## Mô hình khách-chủ vs. Cụm máy chủ song song (1/3)

### | Kiến trúc truyền thống – Mô hình khách-chủ (Client-server)

- ▶ Một hoặc nhiều máy tính khách được kết nối với máy chủ trung tâm qua mạng hoặc kết nối internet
- ▶ Quản lý dữ liệu tập trung
- ▶ Khách hàng yêu cầu và kéo dữ liệu để xử lý
- ▶ Phát sinh lưu lượng truy cập mạng và chi phí di chuyển dữ liệu cao



## Mô hình khách-chủ vs. Cụm máy chủ song song (2/3)

### Mô hình xử lý điển hình của Client-Server

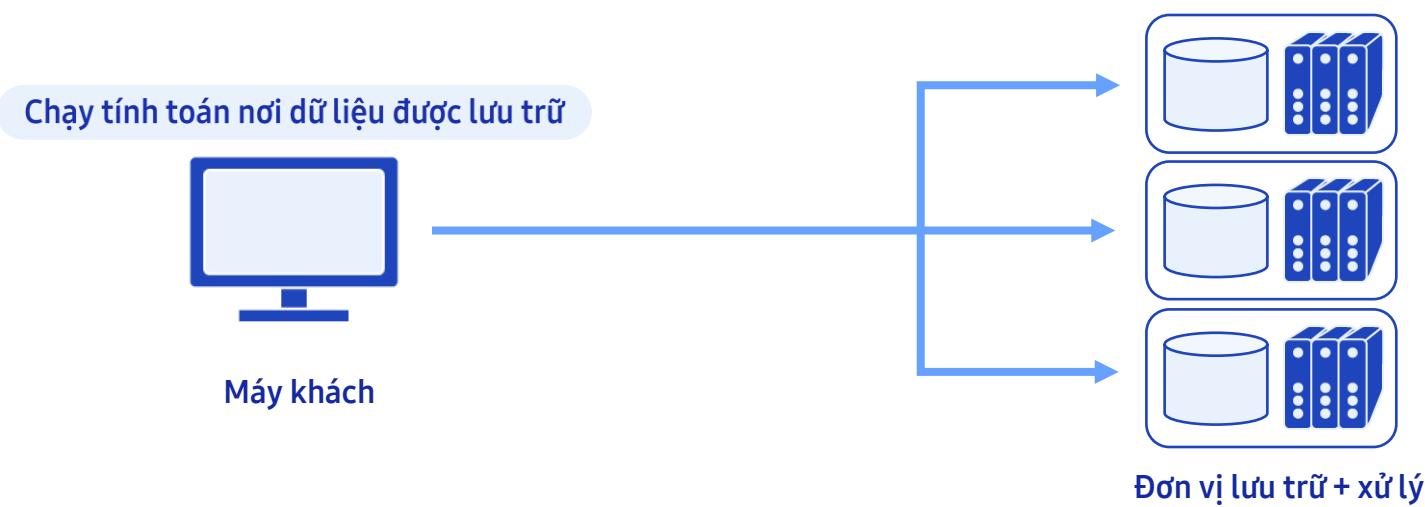
- ▶ Giai đoạn 1: sao chép dữ liệu đầu vào từ đĩa đến máy chủ xử lý
- ▶ Giai đoạn 2: thực hiện xử lý với dữ liệu
- ▶ Giai đoạn 3: sao chép dữ liệu đầu ra trở lại bộ lưu trữ



## Mô hình khách-chủ vs. Cụm máy chủ song song (3/3)

### Cụm máy chủ song song

- ▶ Dữ liệu được sao chép trên nhiều máy chủ
- ▶ Dữ liệu được xử lý tại nơi dữ liệu được lưu trữ
- ▶ Giảm chi phí I/O dữ liệu



# Xử lý song song phân tán

## Tóm lược

- ▶ Tạo kiến trúc cụm xử lý song song phân tán trong đó tất cả các bộ xử lý đều là máy chủ
- ▶ Sử dụng phần cứng hàng hóa tiết kiệm
- ▶ Cài đặt lưu trữ dựa trên đĩa một cách rộng rãi trên tất cả các máy chủ và xử lý dữ liệu nơi chứa dữ liệu
- ▶ Sao chép dữ liệu nhiều lần trong trường hợp đĩa hoặc máy chủ bị lỗi
- ▶ Quản lý khả năng chịu lỗi và tăng hiệu suất bằng các giải pháp dựa trên phần mềm thay vì dựa trên phần cứng

Bài 2.

# Xu hướng hiện tại trong Big Data

Giới thiệu về Big Data

Bài 2.

## Xu hướng hiện tại trong Big Data

- | 2.1. Edge-To-AI : Các công cụ tự động hóa mới nổi
- | 2.2. Vai trò mới nổi của Dịch vụ đám mây công cộng trong Big Data

# Chuyển đổi kỹ thuật số (Digital Transformation)

## I DT (Chuyển đổi kỹ thuật số) là gì?

- ▶ Nó đề cập đến những nỗ lực của một tổ chức để kết hợp các công nghệ, quy trình và văn hóa mới vì một mục đích chung
- ▶ Nó liên quan đến nỗ lực của một tổ chức trong việc kết hợp các công nghệ, quy trình và văn hóa mới vì một mục đích chung
- ▶ Mục tiêu có thể là cải thiện trải nghiệm của khách hàng hoặc tăng tốc đổi mới hoặc tìm cách tồn tại trước sự thay đổi của ngành do gián đoạn kỹ thuật số gây ra

## I Hội tụ tăng tốc Chuyển đổi số

Java AI IoT Hadoop  
MachineLearning Cloud SQL AWS  
**Digital Transformation**  
Spark Intelligence Bigdata DeepLearning  
Azure

# Cốt lõi của cuộc cách mạng công nghiệp lần thứ 4

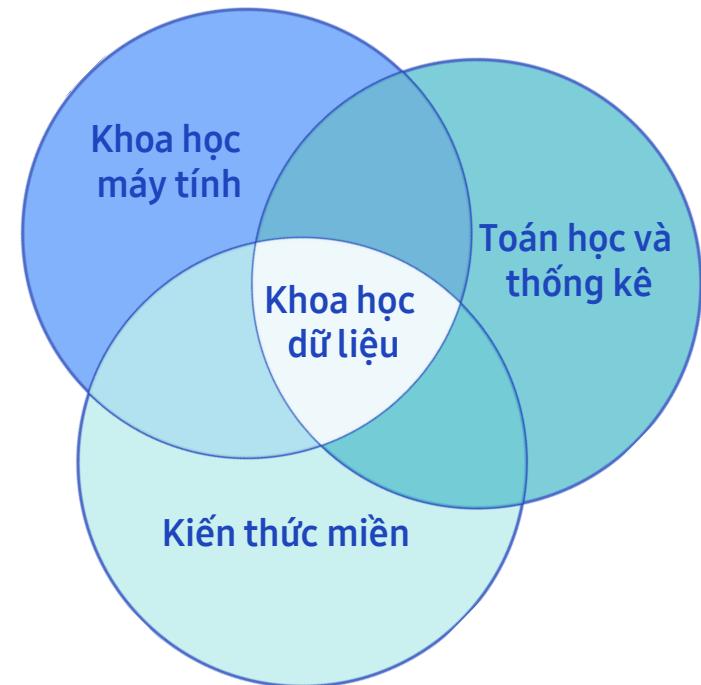
- IoT (Internet of Things): Vạn vật kết nối
  - ▶ Nó đề cập đến một công nghệ hoặc môi trường trong đó các cảm biến được gắn vào các đối tượng để trao đổi dữ liệu trong thời gian thực thông qua Internet, v.v.
- CPS (Cyber-Physical System): Hệ thống vật lý điện tử
  - ▶ Một hệ thống tích hợp các hệ thống vật lý thời gian thực như robot và thiết bị y tế, phần mềm trong không gian mạng và môi trường xung quanh trong thời gian thực
- Big Data: Dữ liệu lớn
  - ▶ Nó đề cập đến các loại dữ liệu khác nhau được tạo trong môi trường kỹ thuật số và nó đề cập đến dữ liệu quy mô lớn với quy mô lớn và chu kỳ tạo nhanh
- AI: Trí tuệ nhân tạo
  - ▶ Một lĩnh vực khoa học máy tính và công nghệ thông tin cho phép máy tính bắt chước các hành vi thông minh cụ thể của con người như suy nghĩ, học tập và phát triển bản thân

# Nhà khoa học dữ liệu vs Kỹ sư dữ liệu

- | Các nhà khoa học dữ liệu tập trung vào việc sử dụng dữ liệu đã được làm sạch để tạo các mô hình AI bằng MLOps
- | Sự tăng cường sử dụng AI trên Edge: sử dụng thuật toán AI để chạy ở rìa mạng
  - ▶ Thực hiện các động khác nhau ở biên nơi dữ liệu được thu thập, thay vì quay lại máy chủ
  - ▶ Khai thác tính năng
  - ▶ Tiền xử lý dữ liệu (loại bỏ ngoại lệ, xác thực dữ liệu)
  - ▶ Ứng dụng mô hình và thuật toán
- | Kỹ sư dữ liệu tập trung vào Đường ống dữ liệu với đám mây hoặc cụm tại chỗ
  - ▶ Cụm đám mây là việc phân phối các dịch vụ đám mây công cộng đến các vị trí thực tế, trong khi hoạt động, quản trị, cập nhật và phát triển của dịch vụ là trách nhiệm của nhà cung cấp đám mây công cộng ban đầu
  - ▶ Tiền xử lý dữ liệu (ETL)
  - ▶ Thu thập một lượng lớn dữ liệu phi cấu trúc và chuyển đổi nó thành một định dạng hữu ích hơn.
  - ▶ Thu thập và khám phá dữ liệu

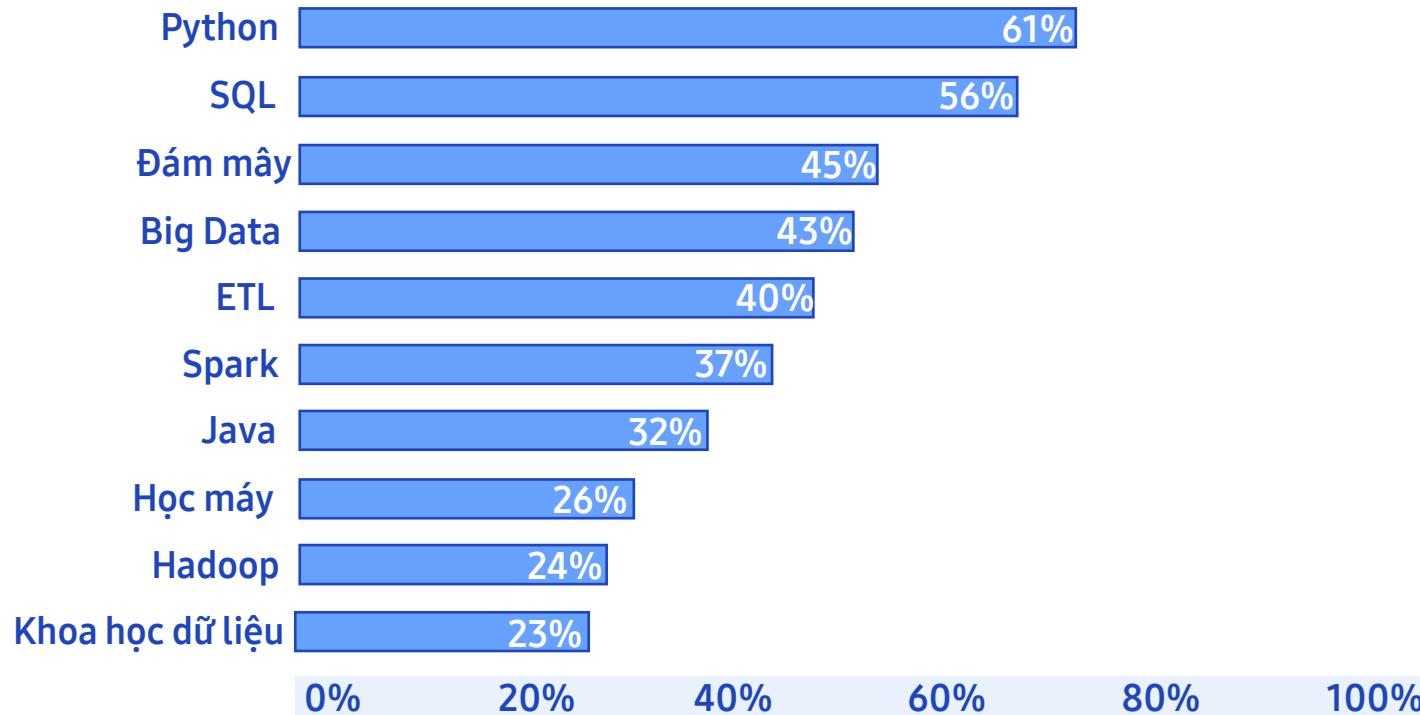
# Kiến thức và kỹ năng dành cho nhà khoa học dữ liệu

- | Bộ kỹ năng của các nhà khoa học dữ liệu
  - ▶ Khoa học máy tính dành cho phân tích
  - ▶ Toán học và thống kê cho Mô hình hóa dữ liệu
  - ▶ Kiến thức miền cho Insight
- | Công việc
  - ▶ Sử dụng các công nghệ dựa trên dữ liệu để giải quyết các vấn đề liên quan đến kinh doanh.
  - ▶ Làm việc với nhiều ngôn ngữ lập trình bao gồm SAS, R và Python.
  - ▶ Có hiểu biết vững chắc về số liệu thống kê, bao gồm xác nhận và phân phối thống kê.
  - ▶ Dựa trên các kỹ thuật phân tích như học máy, học sâu và phân tích văn bản.



# Kiến thức và kỹ năng dành cho kỹ sư dữ liệu

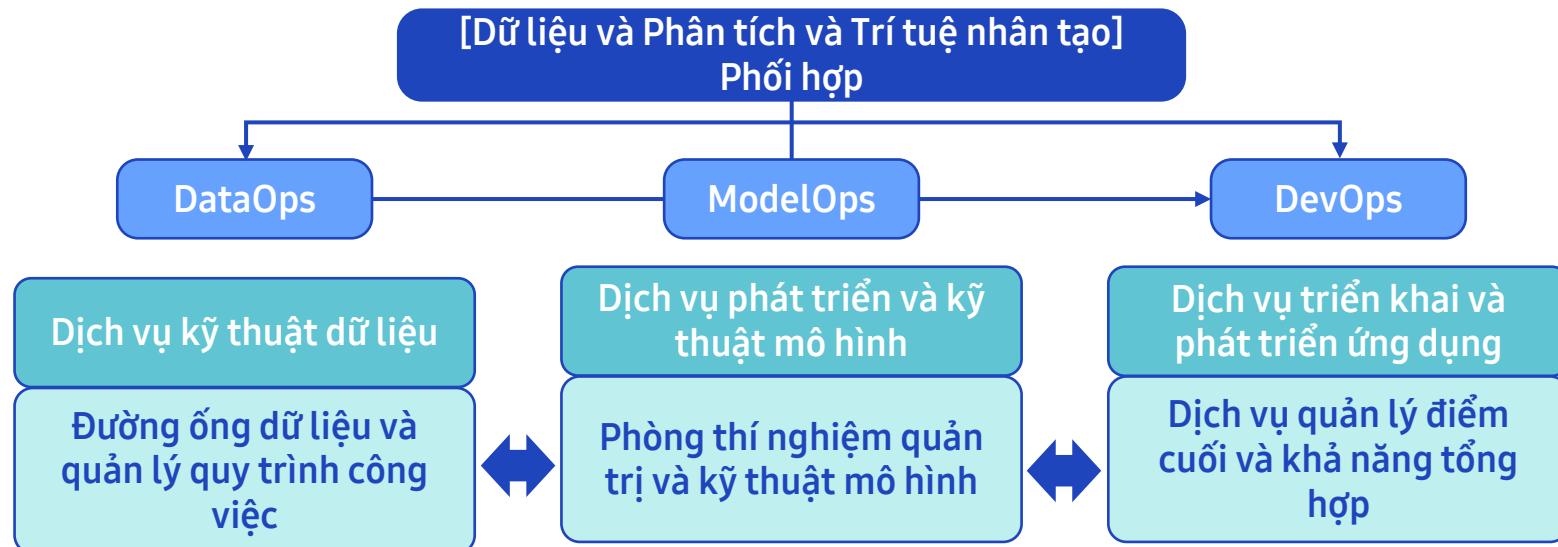
- 10 kỹ năng hàng đầu dành cho kỹ sư dữ liệu năm 2021



# Điện toán biên đối với AI (1/2)

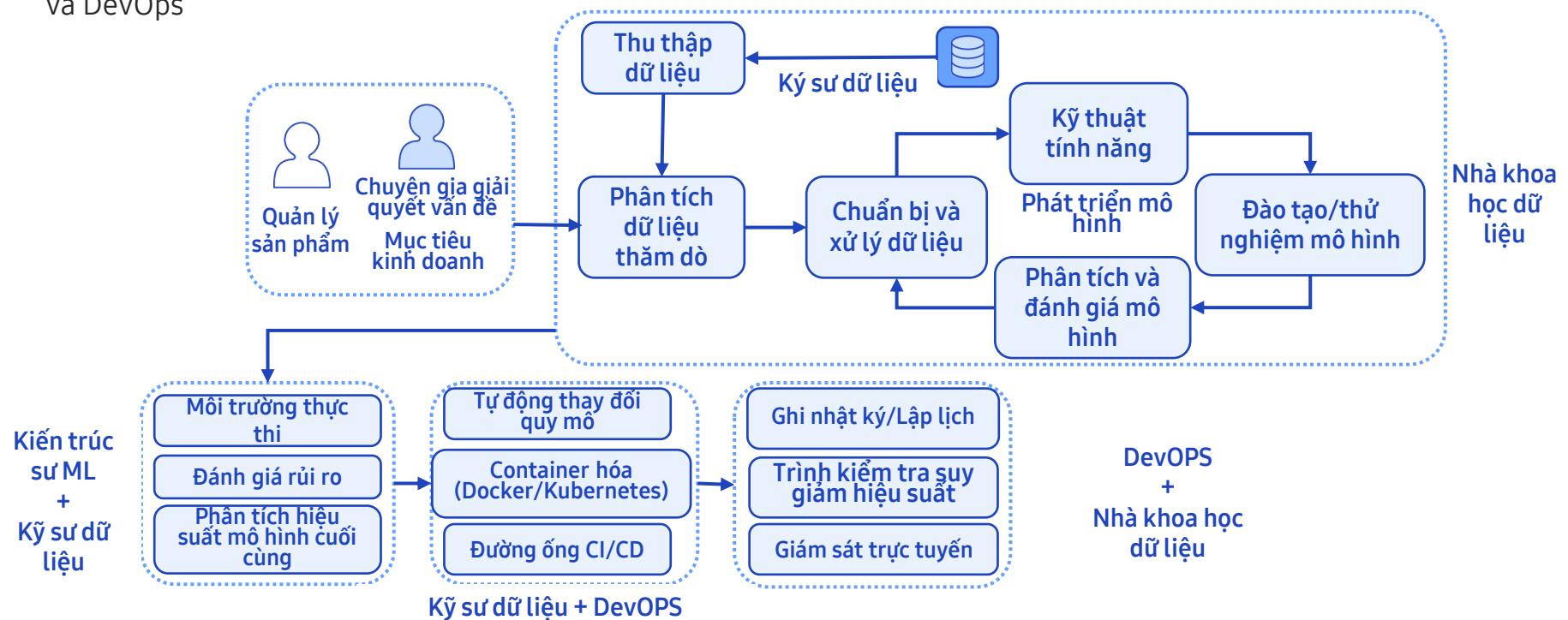
## I XOps là gì?

- Mục tiêu của XOps (dữ liệu, học máy, mô hình, nền tảng) là đạt được hiệu quả và tính kinh tế theo quy mô bằng cách sử dụng các phương pháp hay nhất của DevOps – đồng thời đảm bảo độ tin cậy, khả năng sử dụng lại và khả năng lặp lại đồng thời giảm sự trùng lặp của công nghệ và quy trình cũng như cho phép tự động hóa. - Gartner -



# Điện toán biên đối với AI (2/2)

- Nhà khoa học dữ liệu công dân (CDS) và Kỹ sư dữ liệu công dân (CDE) kết hợp cả hai vai trò
  - Xu hướng là phát triển các công cụ tự động hóa để hỗ trợ và tạo điều kiện thuận lợi cho việc kết hợp DataOps, MLOps và DevOps



Bài 2.

## Xu hướng hiện tại trong Big Data

- | 2.1. Edge-To-AI : Các công cụ tự động hóa mới nổi
- | 2.2. Vai trò mới nổi của Dịch vụ đám mây công cộng trong Big Data

# Điện toán đám mây

## I Điện toán đám mây là gì?

- ▶ Trong kiến trúc này, dữ liệu chủ yếu nằm trên các máy chủ 'ở đâu đó trên Internet' và ứng dụng chạy trên cả 'máy chủ đám mây' và trình duyệt của người dùng. -Eric Schmidt –

## I Tại sao Google lại nghĩ về điều đó?

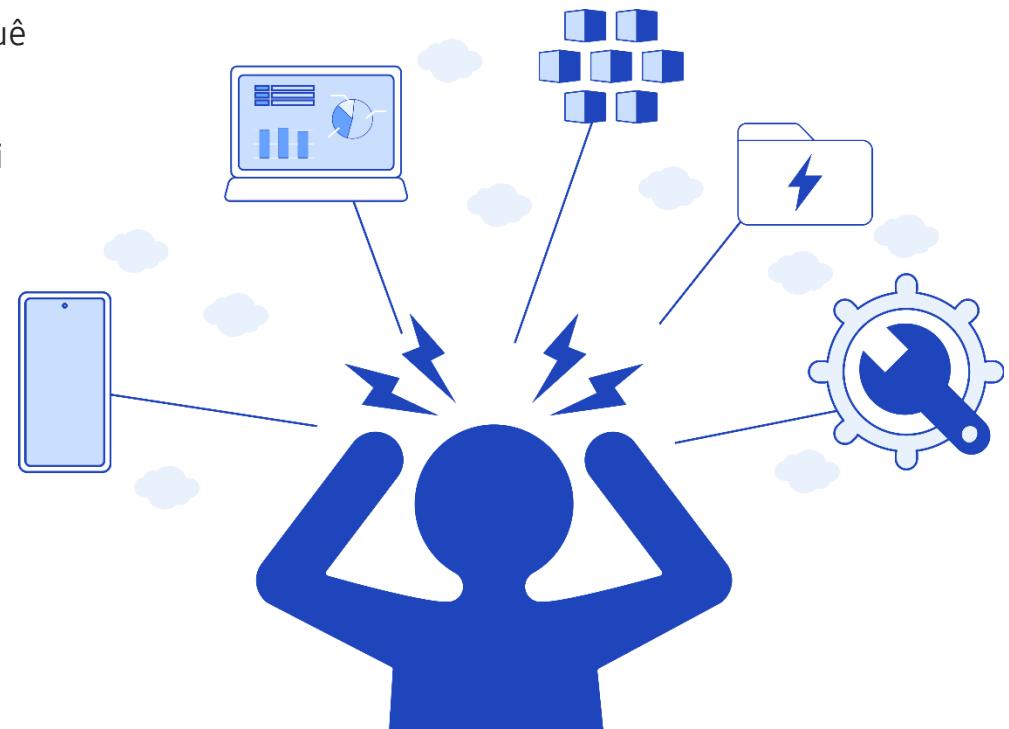
- ▶ Một dịch vụ rất thành công để lưu giữ các ứng dụng và dữ liệu của người dùng máy tính trên web thông qua các dịch vụ web tiên tiến nhất của ngành (Google Docs, Google Notes, v.v.).
- ▶ Thông qua đó, người ta dự đoán rằng tất cả các phần mềm sẽ tồn tại trên Internet trong tương lai.
- ▶ Giới thiệu khái niệm “SaaS”
- ▶ Tích lũy kinh nghiệm của Google để giải quyết khối lượng kết nối toàn cầu

## I Kinh doanh hiện đại trên đám mây là điều bình thường mới.

# Hạn chế của kiến trúc truyền thống

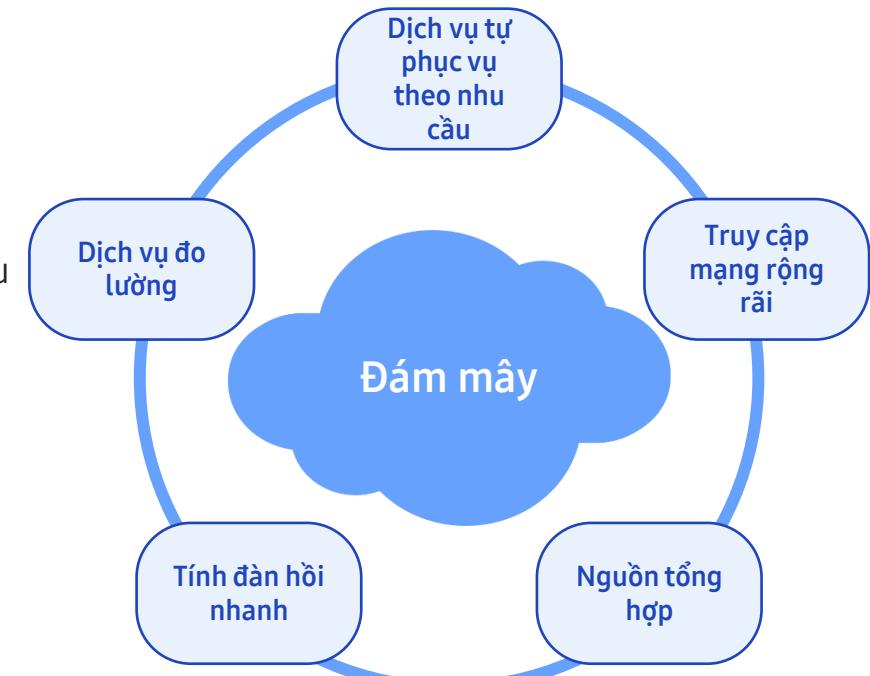
## I Điện toán tại chỗ

- ▶ Sắp xếp thứ tự lưu trữ và tính toán
- ▶ Hoạt động cụm không hiệu quả do nhiều người thuê
- ▶ Phân bổ nguồn lực không hiệu quả
- ▶ Không thể hỗ trợ tự động mở rộng quy mô theo tải
- ▶ Bảo mật cụm dữ liệu vật lý

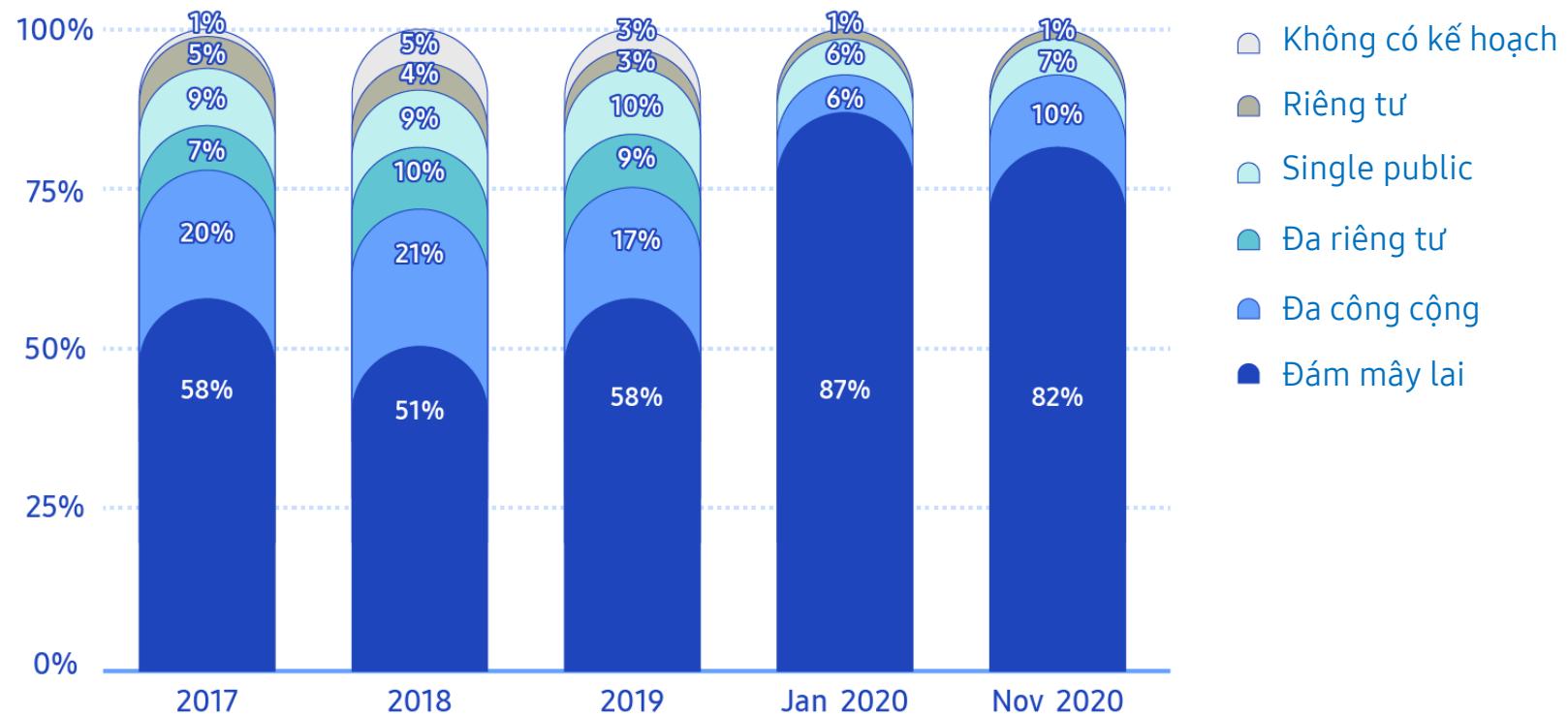


## Các khía cạnh chính của Đám mây công cộng

- █ Có khả năng hỗ trợ tự động mở rộng quy mô theo tải
- █ Mạng tốc độ cao cho phép tách biệt lưu trữ khỏi tài nguyên máy tính
  - ▶ Có thể mở rộng quy mô riêng cho từng yêu cầu
- █ Cửa hàng đối tượng (S3, ADLS) cung cấp một phương pháp hữu ích để lưu trữ và xử lý các loại dữ liệu khác nhau
- █ Công nghệ thùng chứa
- █ Cải tiến bảo mật đáng tin cậy cho các cụm dữ liệu



## Chiến lược đám mây doanh nghiệp trên toàn thế giới



# Sự phát triển của Kiến trúc (1/2)

## I. Nhiều đám mây công cộng

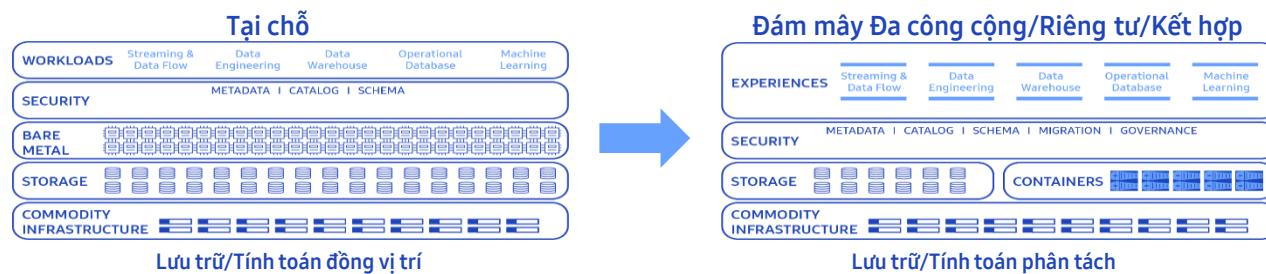
- Nhiều đám mây công cộng đề cập đến việc triển khai hai hoặc nhiều đám mây cùng loại được cung cấp bởi các nhà cung cấp khác nhau.

## II. Đám mây riêng

- Đám mây riêng là môi trường đám mây dành riêng cho người dùng cuối, thường nằm trong tường lửa của người dùng.

## III. Đám mây kết hợp

- Đám mây kết hợp tích hợp một số mức độ tính di động, điều phối và quản lý khối lượng công việc trên hai hoặc nhiều môi trường (công khai và riêng tư)
- Nhiều người thuê, container hóa



# Sự phát triển của Kiến trúc (2/2)

## I. Những lợi thế quan trọng với Đám mây kết hợp

- ▶ Phân tách ngăn xếp phần mềm – lưu trữ, tính toán, bảo mật và quản trị
- ▶ Nhiều tùy chọn hơn để triển khai tài nguyên tính toán và lưu trữ
- ▶ Tùy chỉnh tài nguyên triển khai bằng máy chủ tại chỗ, bộ chứa, máy ảo hoặc tài nguyên đám mây

## II. Cách ly khối lượng công việc

- ▶ Khi triển khai một cụm cho cơ sở hạ tầng đám mây của mình, bạn có thể tạm thời chấm dứt cụm điện toán của mình để tránh các chi phí không cần thiết.
- ▶ Trong khi vẫn giữ dữ liệu từ các ứng dụng khác để tiếp tục sử dụng
- ▶ Các cụm điện toán có thể giúp giải quyết tranh chấp tài nguyên
- ▶ Có thể tách biệt các khối lượng công việc dài hạn hoặc sử dụng nhiều tài nguyên để chạy trong một cụm máy tính chuyên dụng

# Làm thế nào để xây dựng một đám mây kết hợp?

## I. Kiến trúc đám mây kết hợp truyền thống

- ▶ Đám mây kết hợp ban đầu theo đúng nghĩa đen là kết nối môi trường đám mây riêng với môi trường đám mây công cộng như một sự lặp lại của phần mềm trung gian lớn và phức tạp
- ▶ Bạn có thể xây dựng đám mây riêng của mình hoặc sử dụng cơ sở hạ tầng đám mây được đóng gói sẵn, bạn cũng cần một đám mây công cộng như bên dưới

 Alibaba Cloud

 AWS

 Google Cloud

 IBM

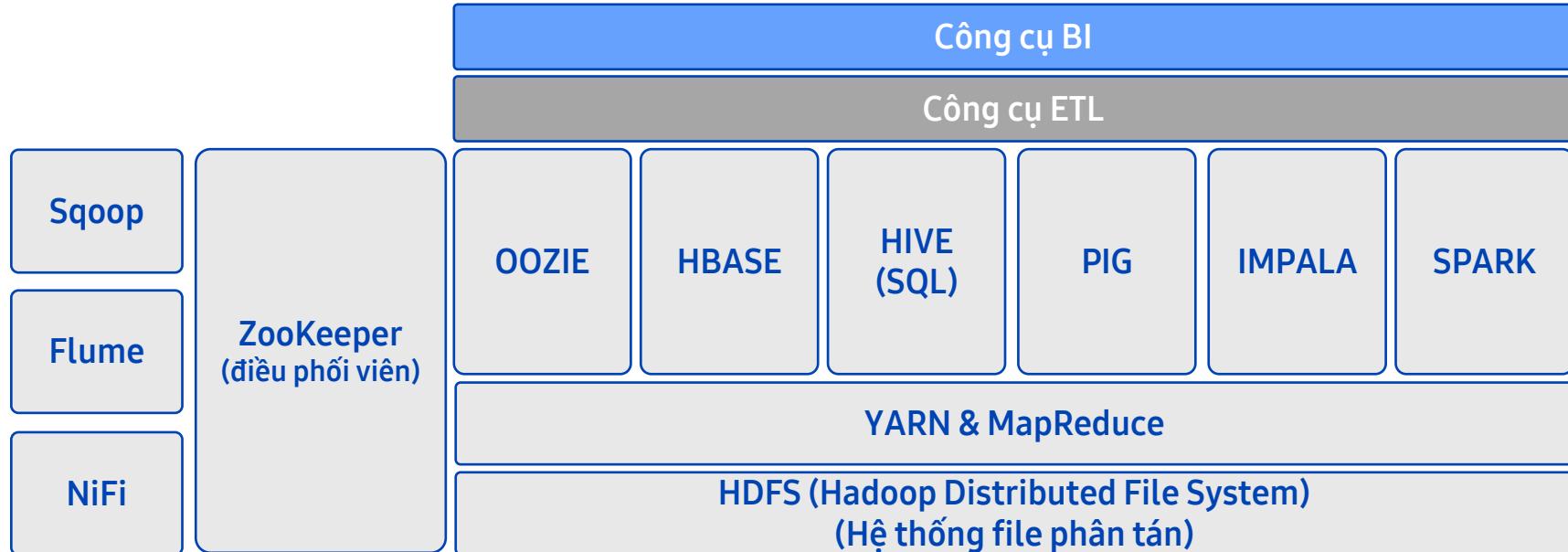
 Azure

## I. Kiến trúc đám mây kết hợp hiện đại

- ▶ Tất cả các môi trường CNTT đều chạy cùng một hệ điều hành và quản lý mọi thứ thông qua một nền tảng hợp nhất, đồng thời cho phép mở rộng tính phổ biến của các ứng dụng sang các môi trường con
- ▶ Tất cả các yêu cầu phần cứng được trừu tượng hóa bằng cách sử dụng cùng một hệ điều hành và nền tảng điều phi trừu tượng hóa tất cả các yêu cầu ứng dụng.

# Tiếp theo là gì?

- Chương 1 là phần giới thiệu về Big Data
  - ▶ Tổng quan và nền tảng của Big Data
  - ▶ Xu hướng hiện tại trong Big Data
- Chương 2 ~ Chương 9 có nội dung kỹ thuật chi tiết hơn





# Together for Tomorrow! Enabling People

Education for Future Generations

©2021 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of book.

This book is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this book other than the curriculum of Samsung innovation Campus or to use the entire or part of this book, you must receive written consent from copyright holder.