# THÔNG TIN NHÓM

|  |  |  |
| --- | --- | --- |
| **Tên đề tài** | **XÂY DỰNG HỆ THỐNG RÚT TRÍCH TIN TỨC TỪ INTERNET VÀ THỐNG KÊ CÁC TỪ KHÓA CHÍNH XUẤT HIỆN** | **Đóng góp của các thành viên** |
| **Họ và tên thành viên 1** | Trần Minh Hợp - 1050080180 | Đóng góp: Lý thuyết + Word  Điểm tự đánh giá quá trình: 9  Điểm tự đánh giá cuối kì: 6 |
| **Họ và tên thành viên 2** | Nguyễn Đức Quyền - 1050080195 | Đóng góp: Lý thuyết + Word  Điểm tự đánh giá quá trình: 6  Điểm tự đánh giá cuối kì: 5 |
| **Họ và tên thành viên 3** | Nguyễn Khánh Duy - 1050080173 | Đóng góp: Lý thuyết +Word  Điểm tự đánh giá quá trình: 7  Điểm tự đánh giá cuối kì: 6 |
| **Họ và tên thành viên 4** | Nguyễn Thảo Trọng - 1050080207 | Đóng góp: 0  Điểm tự đánh giá quá trình: 7  Điểm tự đánh giá cuối kì: 4 |
| **Họ và tên thành viên 5** | Hoàng Văn Hiếu - 1050080179 | Đóng góp: 0  Điểm tự đánh giá quá trình:  Điểm tự đánh giá cuối kì: |

**BÁO CÁO TÓM TẮT**

**Lý do chọn đề tài:**

Việc chọn đề tài xây dựng hệ thống rút trích tin tức từ internet và thống kê các từ khóa chính có nhiều lý do hấp dẫn. Đầu tiên, nó có thể xuất phát từ sự quan tâm cá nhân, với mục tiêu theo dõi thông tin từ internet, các chủ đề yêu thích hoặc cập nhật về những sự kiện quan trọng. Hệ thống này cũng có ứng dụng rộng rãi trong nghiên cứu và kinh doanh, giúp thu thập dữ liệu quan trọng cho phân tích và định hướng chiến lược. Bên cạnh đó, việc xây dựng nó mang lại cơ hội phát triển kỹ năng kỹ thuật và ứng dụng Trí tuệ Nhân tạo trong thực tế. Cuối cùng, nó còn hỗ trợ trong việc kiểm tra và đảm bảo tính chính xác của thông tin từ các nguồn tin tức trực tuyến.

**Tầm quan trọng:**

Xây dựng hệ thống rút trích tin tức từ internet và thống kê từ khóa chính là một dự án quan trọng với tầm quan trọng rất cao. Dự án này giúp theo dõi thông tin mới nhất, nghiên cứu, phân tích dữ liệu, và đảm bảo tính chính xác của thông tin. Nó cũng cung cấp cơ hội cho việc phát triển kỹ năng kỹ thuật và ứng dụng Trí tuệ Nhân tạo. Bên cạnh đó, trong lĩnh vực thương mại, hệ thống này đóng vai trò quan trọng trong việc theo dõi thị trường và định hình chiến lược kinh doanh. Tóm lại, việc xây dựng hệ thống này có tầm quan trọng đa dạng và quyết định đối với nhiều mục tiêu và lĩnh vực ứng dụng

**Các nội dung và kết quả:**

Thu thập dữ liệu

* Lựa chọn nguồn tin tức
* Crawl dữ liệu

Tiền xử lý dữ liệu

* Loại bỏ HTML tags
* Loại bỏ stop words
* Tokenization
* Làm sạch dữ liệu

Phân tích văn bản

* Tính toán TF-IDF
* Phát hiện từ khóa

Thống kê và trực quan hóa

* Đếm từ khóa
* Sắp xếp theo tần suất
* Trực quan hóa

Giao diện người dùng (tùy chọn)

* Phát triển giao diện để người dùng có thể tùy chỉnh tìm kiếm hoặc nhập URL.

Triển khai hệ thống

* Triển khai hệ thống trên máy chủ web hoặc nền tảng đám mây.

Kiểm tra và điều chỉnh

* Kiểm tra để đảm bảo hoạt động đúng và điều chỉnh nếu cần.

**BÁO CÁO CHI TIẾT**

# Tổng quan về Big Data



## Nguồn gốc

Thập kỷ 1960 và 1970: Thoạt đầu, nguồn gốc của Big Data có thể được tìm thấy trong sự phát triển của cơ sở hạ tầng máy tính và lưu trữ dữ liệu. Thời đó, thế giới dữ liệu chỉ mới bắt đầu với sự ra đời của các trung tâm dữ liệu đầu tiên. Hệ thống quản lý cơ sở dữ liệu SQL cũng bắt đầu xuất hiện và phát triển.

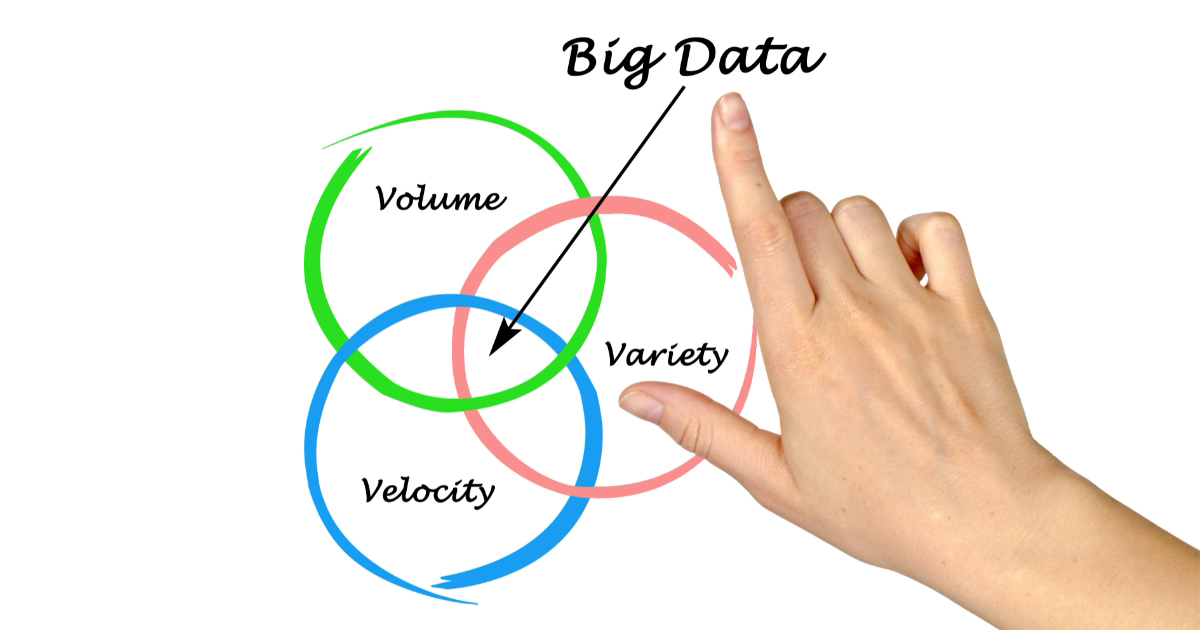
Năm 1984: Teradata đã giới thiệu hệ thống xử lý dữ liệu song song DBC 1012. Đây là một trong những hệ thống đầu tiên có khả năng phân tích và lưu trữ 1 terabyte dữ liệu. Từ đó, nó đã phát triển để hỗ trợ hàng chục cơ sở dữ liệu với dung lượng lên đến hàng petabyte, với một số trường hợp có lượng dữ liệu lớn nhất vượt qua ngưỡng 50 petabytes.

Năm 2000: Trong thời kỳ này, internet và sự phát triển của các dịch vụ trực tuyến như YouTube, Facebook và các nền tảng xã hội khác đã dẫn đến việc tạo ra lượng lớn dữ liệu từ người dùng trực tuyến. NoSQL cũng trở nên phổ biến hơn và đã hỗ trợ sự phát triển của các framework như Hadoop, giúp thúc đẩy Big Data.

Thập kỷ 2010 và sau đó: Big Data đã thấy sự gia tăng nhanh chóng về khối lượng dữ liệu, đặc biệt sau khi Internet of Things (IoT) bắt đầu phát triển. IoT mang lại lượng lớn dữ liệu từ các thiết bị kết nối mạng, và máy móc cũng tạo ra dữ liệu tự động. Các công nghệ như Hadoop, Apache Spark và các công cụ phân tích dữ liệu khác đã trở nên quan trọng trong việc xử lý và trích xuất giá trị từ dữ liệu lớn này.

*Tóm lại, Big Data không có một nguồn gốc duy nhất mà nó phát triển từ nhiều yếu tố, bao gồm sự tiến bộ trong công nghệ máy tính, lưu trữ dữ liệu, internet, sự phát triển của các dịch vụ trực tuyến và IoT. Nó đã trở thành một phần quan trọng trong cuộc sống và doanh nghiệp hiện đại, giúp chúng ta hiểu rõ hơn về dữ liệu và đưa ra quyết định dựa trên thông tin từ dữ liệu lớn*

## Khái niệm và đặc trưng



*Big Data là một thuật ngữ phổ biến được sử dụng để mô tả các tập hợp dữ liệu lớn và phức tạp mà công cụ và ứng dụng xử lý truyền thống không thể quản lý, xử lý và thu thập hiệu quả. Các tập hợp dữ liệu lớn này có thể bao gồm dữ liệu có cấu trúc, không có cấu trúc và bán cấu trúc, và chúng thường có những đặc điểm quan trọng sau:*

**Khối lượng lớn:** Big Data thường bao gồm rất nhiều dữ liệu, thường trong khoảng từ terabytes đến exabytes hoặc hơn nữa. Điều này đòi hỏi cơ sở hạ tầng mạng và lưu trữ mạnh mẽ để quản lý và lưu trữ dữ liệu này.

Dữ liệu xã hội: Mạng xã hội như Facebook và Twitter tạo ra hàng tỷ dữ liệu mới mỗi ngày.

Dữ liệu khoa học: Dự án khoa học lớn như thám hiểm vũ trụ hoặc nghiên cứu dịch tễ học có thể tạo ra hàng petabyte dữ liệu.

Dữ liệu IoT: Thiết bị IoT gửi dữ liệu liên tục, ví dụ: cảm biến trên xe hơi thông báo vị trí và tình trạng của xe

**Đa dạng về loại dữ liệu:** Big Data thường bao gồm nhiều loại dữ liệu khác nhau, bao gồm dữ liệu có cấu trúc (ví dụ: trong cơ sở dữ liệu), dữ liệu không có cấu trúc (ví dụ: tài liệu văn bản, email) và dữ liệu bán cấu trúc (ví dụ: dữ liệu JSON). Nó cũng có thể kết hợp dữ liệu từ nhiều nguồn khác nhau.

**Tốc độ nhanh (Velocity):** Big Data thường được tạo ra hoặc cập nhật với tốc độ nhanh.

Ví dụ:

Dữ liệu mạng xã hội: Mạng xã hội như Twitter tạo ra hàng triệu tweet mỗi giây.

Dữ liệu IoT: Dữ liệu từ các cảm biến trên máy móc hoặc thiết bị IoT cập nhật theo thời gian thực.

Dữ liệu tài chính: Giao dịch trên thị trường chứng khoán xảy ra trong khoảng thời gian ngắn và cần được xử lý một cách nhanh chóng

Bên cạnh đó, Big Data còn có nhiều **đặc trưng** cụ thể **khác** như:

**Tính xác thực (Veracity):** Đây ám chỉ mức độ tin cậy của dữ liệu.

Ví dụ:

Dữ liệu từ cảm biến: Cảm biến có thể gây ra lỗi hoặc nhiễu, làm cho dữ liệu không chắc chắn.

Dữ liệu xã hội: Thông tin trên mạng xã hội có thể không đáng tin cậy, chứa tin đồn hoặc thông tin giả mạo.

**Giá trị (Value):** Giá trị của Big Data nằm trong khả năng trích xuất thông tin hữu ích để hỗ trợ quyết định hoặc tạo ra lợi ích.

Ví dụ:

Dự đoán thị trường: Phân tích dữ liệu để dự đoán xu hướng thị trường và điều chỉnh chiến lược kinh doanh.

Y tế và chẩn đoán bệnh: Sử dụng dữ liệu y tế để hỗ trợ quyết định trong việc chuẩn đoán bệnh.

Tiếp thị và tương tác khách hàng: Theo dõi dữ liệu mạng xã hội để đánh giá tương tác của khách hàng và hiệu suất chiến dịch tiếp thị.

**Biến đổi (Variability):** Tính biến đổi đề cập đến sự thay đổi trong dữ liệu theo thời gian.

Ví dụ:

Dữ liệu thời tiết: Dữ liệu thời tiết có sự biến đổi theo mùa và thời gian trong ngày.

Dữ liệu thương mại điện tử: Dữ liệu từ trang web thương mại điện tử có thể thay đổi theo các chương trình khuyến mãi và sự kiện

Các nguồn tạo ra Big Data có thể bao gồm trang web, mạng xã hội, ứng dụng di động, máy tính, các thiết bị IoT, các dự án nghiên cứu và thử nghiệm khoa học kỹ thuật, và nhiều nguồn khác. Big Data đã trở thành một phần quan trọng trong nhiều lĩnh vực, cho phép các tổ chức và cá nhân hiểu rõ hơn về dữ liệu và đưa ra quyết định dựa trên thông tin phân tích từ dữ liệu lớn này.

***Các ví dụ về Big Data***

Các dữ liệu trên mạng xã hội: Mạng xã hội như là facebook với hàng tỷ người dùng trên khắp thế giới. Tất cả hình ảnh, video, chữ, số, tổng lượt like, share hay thả icon và comment của người dùng đều được hệ thống lưu lại. Hệ thống lưu trữ dữ liệu của facebook là một hệ thống Big data.

Nền tảng Google và các nền tảng tìm kiếm khác: Những bài viết trên công cụ tìm kiếm google đều được lưu trữ vào hệ thống. Khi người dùng search tìm kiếm, dữ liệu sẽ được lấy từ hệ thống dữ liệu này. Hệ thống lưu trữ dữ liệu tìm kiếm của google đó chính là Big data.

## Vai trò

### Tài chính và Ngân hàng:

Xác định tệp khách hàng mục tiêu: Big Data giúp doanh nghiệp hiểu rõ hành vi, sở thích, và nhu cầu của khách hàng bằng cách thu thập dữ liệu từ nhiều nguồn, chẳng hạn như lịch sử trình duyệt web và thói quen sử dụng mạng xã hội. Thông qua phân tích dữ liệu, doanh nghiệp có thể lựa chọn đúng tệp khách hàng phù hợp với sản phẩm hoặc dịch vụ của họ

Tối ưu hóa năng suất và định hướng công việc: Big Data cung cấp thông tin chi tiết về xu hướng tiêu dùng mới nhất thông qua dữ liệu từ thiết bị thông minh và người dùng. Điều này giúp doanh nghiệp định hướng tối ưu hóa năng suất và lập kế hoạch cho tương lai dựa trên thông tin phân tích

Nâng cao bảo mật và giảm rủi ro: Big Data hỗ trợ xây dựng hệ thống thám đo và ngăn chặn các rủi ro, như gian lận, xâm nhập hệ thống và đánh cắp thông tin mật. Các ngân hàng và công ty thẻ tín dụng, ví dụ, sử dụng Big Data để ngăn chặn các giao dịch gian lận

Tối ưu hóa giá cả: Big Data giúp doanh nghiệp nghiên cứu mức giá phù hợp cho sản phẩm hoặc dịch vụ của họ bằng cách thu thập và phân tích dữ liệu nhanh chóng và chính xác

Nắm bắt các giao dịch tài chính: Big Data cung cấp thông tin từ giao dịch tài chính trong thương mại điện tử, giúp doanh nghiệp hiểu rõ hành vi mua sắm của khách hàng và đưa ra gợi ý để tăng tỷ lệ mua sắm

### Y tế:

Phân tích dữ liệu lâm sàng: Dữ liệu lâm sàng lớn được thu thập từ bệnh viện, phòng khám, và hồ sơ y tế của bệnh nhân. Big Data giúp phân tích dữ liệu này để tìm ra các mô hình và thông tin quan trọng về bệnh lý, thuốc, và phản ứng của bệnh nhân với các loại điều trị.

Phát hiện dấu hiệu bệnh: Big Data có khả năng phát hiện sớm dấu hiệu của các bệnh lý. Bằng cách theo dõi dữ liệu y tế cá nhân, hệ thống có thể cảnh báo về các biểu hiện sớm của bệnh lý, giúp bệnh nhân và bác sĩ can thiệp kịp thời.

Tối ưu hóa quản lý bệnh viện: Big Data giúp quản lý bệnh viện cải thiện hiệu suất và tối ưu hóa tài nguyên. Bằng cách theo dõi luồng bệnh nhân, thời gian chờ đợi, và tình hình tài chính, bệnh viện có thể hoạt động hiệu quả hơn và cung cấp dịch vụ tốt hơn cho bệnh nhân.

Nghiên cứu và phát triển: Big Data hỗ trợ trong việc nghiên cứu y tế và phát triển các phương pháp điều trị mới. Dữ liệu lớn từ nhiều nguồn có thể giúp nhà nghiên cứu xác định các loại bệnh lý, dự đoán kết quả điều trị, và tạo ra các công cụ và thuốc mới.

Đảm bảo chất lượng dịch vụ: Big Data giúp theo dõi chất lượng dịch vụ y tế bằng cách thu thập phản hồi từ bệnh nhân và đo lường các tiêu chuẩn chất lượng. Điều này giúp cải thiện trải nghiệm của bệnh nhân và đảm bảo chất lượng chăm sóc y tế.

Big Data đã thay đổi cách lĩnh vực y tế hoạt động và đã tạo ra nhiều cơ hội để cải thiện dịch vụ y tế và quản lý bệnh lý.

### Chuỗi cung ứng và Quản lý kho:

Tối ưu hóa lịch trình giao hàng: Big Data cho phép theo dõi và phân tích thông tin về các yếu tố ảnh hưởng đến lịch trình giao hàng. Điều này bao gồm thông tin về giao thông, thời tiết, sự cố hậu cần và hơn thế nữa. Nhờ đó, các doanh nghiệp có thể tối ưu hóa lịch trình giao hàng, giảm thiểu thời gian chờ đợi và chi phí giao hàng.

Dự đoán cầu đầu vào: Big Data giúp dự đoán mức cầu đầu vào cho các sản phẩm hoặc dịch vụ cụ thể. Bằng cách phân tích dữ liệu về đặt hàng trước, hành vi mua sắm của khách hàng và các yếu tố thị trường, doanh nghiệp có thể tạo ra các dự đoán chính xác về nhu cầu của thị trường.

Tối ưu hóa lượng tồn kho: Big Data giúp quản lý lượng tồn kho hiệu quả hơn bằng cách cung cấp thông tin về mức tồn kho, tốc độ bán hàng, và dự đoán về việc tái cung cấp hàng hóa. Điều này giúp giảm lượng tồn kho không cần thiết, tiết kiệm chi phí và tối ưu hóa quản lý kho.

Quản lý hiệu suất chuỗi cung ứng: Big Data cho phép theo dõi hiệu suất của các phần tử trong chuỗi cung ứng. Việc này bao gồm việc theo dõi thời gian chế biến, hiệu suất nhà cung cấp và các chỉ số khác. Khi có dữ liệu chính xác về hiệu suất, các tổ chức có thể cải thiện quy trình và tối ưu hóa hiệu suất chuỗi cung ứng.

Tăng cường tích hợp trong chuỗi cung ứng: Big Data cho phép các đối tác trong chuỗi cung ứng chia sẻ dữ liệu và tương tác dễ dàng hơn. Điều này giúp tạo ra tích hợp mạnh mẽ hơn giữa các phần tử trong chuỗi, giảm thiểu sự cố và đảm bảo là hàng hóa di chuyển một cách trơn tru trong toàn bộ chuỗi cung ứng

### Lĩnh vực Môi trường:

Theo dõi và dự đoán biến đổi khí hậu: Dữ liệu về khí hậu được thu thập từ nhiều nguồn, chẳng hạn như các trạm quan trắc khí tượng, máy bay không người lái, vệ tinh, và các thiết bị cảm biến khác. Dữ liệu này sau đó được phân tích bằng Big Data để theo dõi và dự đoán biến đổi khí hậu, biến đổi thời tiết, và các hiện tượng khí hậu đặc biệt. Điều này có thể giúp chính phủ và các tổ chức môi trường đưa ra các biện pháp bảo vệ môi trường hiệu quả hơn.

Quản lý tài nguyên thiên nhiên: Big Data có thể được sử dụng để theo dõi và quản lý tài nguyên thiên nhiên như rừng, đại dương, và nguồn nước. Dữ liệu từ các cảm biến trên mặt đất, trong nước, và dưới nước giúp hiểu rõ tình trạng của tài nguyên và đưa ra quyết định quản lý bền vững.

Bảo vệ và giám sát môi trường: Big Data giúp trong việc giám sát môi trường và phát hiện các sự cố, như sự rò rỉ chất độc hại hay ô nhiễm. Các hệ thống cảm biến có thể gửi dữ liệu liên tục, và Big Data có thể phân tích dữ liệu này để cảnh báo về các vấn đề môi trường.

Phân tích dữ liệu về sự sống hoang dã: Big Data cũng được sử dụng để theo dõi và nghiên cứu động, thực vật, và các loài động vật hoang dã. Dữ liệu về di chuyển, số lượng, và hành vi của các loài này có thể giúp trong việc bảo vệ và quản lý động thực vật hoang dã.

Sử dụng Big Data trong lĩnh vực môi trường có thể giúp theo dõi và đối phó với các vấn đề quan trọng liên quan đến biến đổi khí hậu, bảo vệ tài nguyên thiên nhiên và duy trì môi trường sống lành mạnh cho con người và các hệ sinh thái

### Lĩnh vực Xã hội và Chính trị:

Big Data có thể phân tích dữ liệu truyền thông xã hội để đo lường tác động của chính trị và xã hội, dự đoán kết quả bầu cử, và phát hiện xu hướng xã hội.

Dự đoán kết quả bầu cử: Big Data được sử dụng để theo dõi và phân tích dữ liệu từ truyền thông xã hội, các cuộc thăm dò dư luận, và các dữ liệu chính trị để dự đoán kết quả bầu cử. Điều này có thể giúp các ứng cử viên và các chính trị gia điều chỉnh chiến dịch của họ.

### Giáo dục:

Big Data giúp trường học và trường đại học theo dõi tiến bộ học tập của học sinh và sinh viên, cung cấp giảng dạy tùy chỉnh, và cải thiện quản lý học tập.

Theo dõi tiến bộ học tập: Big Data trong lĩnh vực giáo dục giúp trường học và trường đại học theo dõi tiến bộ học tập của học sinh và sinh viên. Dựa trên dữ liệu này, họ có thể cung cấp hỗ trợ tùy chỉnh và các khóa học đặc biệt cho những người có nhu cầu riêng biệt

### Công nghiệp sản xuất:

Big Data được sử dụng để theo dõi hiệu suất máy móc và thiết bị, dự đoán thời gian hỏng hóc và cải thiện quá trình sản xuất.

### Vận tải và Giao thông:

Big Data giúp quản lý lưu lượng giao thông, cải thiện lên lịch tuyến đường, và dự đoán nhu cầu vận tải.

### **Nghiên cứu khoa học và Phát triển sản phẩm:**

Phân tích dữ liệu từ thí nghiệm khoa học: Big Data cho phép các nhà nghiên cứu thu thập, lưu trữ, và phân tích lượng lớn dữ liệu từ các thí nghiệm khoa học. Điều này giúp họ tìm hiểu về hiệu suất thí nghiệm, kết quả, và xu hướng quan trọng từ dữ liệu số liệu kỹ thuật.

Tạo ra sản phẩm và dịch vụ mới: Dữ liệu lớn có thể cung cấp thông tin để phát triển sản phẩm và dịch vụ mới. Phân tích dữ liệu có thể giúp xác định nhu cầu của thị trường và phát triển sản phẩm hoặc dịch vụ tùy chỉnh dựa trên thông tin này.

Cải thiện quy trình nghiên cứu và phát triển: Big Data có thể được sử dụng để tối ưu hóa quy trình nghiên cứu và phát triển. Bằng cách theo dõi hiệu suất dự án, dự báo rủi ro, và tối ưu hóa nguồn lực, tổ chức nghiên cứu có thể làm việc hiệu quả hơn và giảm thiểu thất thoát.

Thăm dò dữ liệu từ nguồn mở và thư viện dữ liệu: Big Data cho phép nghiên cứu viên truy cập vào các nguồn dữ liệu mở và thư viện dữ liệu để thực hiện các nghiên cứu cụ thể. Điều này giúp tạo điều kiện thuận lợi cho việc nghiên cứu và tiến hành nghiên cứu cắt giảm chi phí và thời gian.

Nghiên cứu y khoa và chăm sóc sức khỏe: Trong lĩnh vực y khoa, Big Data giúp nghiên cứu các mô hình bệnh lý, cải thiện quản lý bệnh viện và dự đoán tình hình sức khỏe cộng đồng

## Quy trình hoạt động

### Xây dựng Chiến lược Big Data:

**Xác định mục tiêu chính:** Bạn cần xác định rõ mục tiêu hoặc vấn đề mà bạn muốn giải quyết bằng Big Data. Các mục tiêu có thể liên quan đến tăng cường doanh số bán hàng, tối ưu hóa quá trình sản xuất, hoặc nâng cao trải nghiệm khách hàng.

**Xác định nguồn tài nguyên:** Để triển khai chiến lược, bạn cần xác định nguồn tài nguyên như ngân sách, nhân lực và công nghệ mà bạn có sẵn.

**Tạo lịch trình:** Lên kế hoạch cho việc triển khai chiến lược. Xác định các giai đoạn và thời gian để đảm bảo rằng chiến lược được triển khai một cách hiệu quả.

### Xác định Các Nguồn Dữ liệu:

**Dữ liệu từ mạng xã hội:** Dữ liệu từ các mạng xã hội như Facebook, Twitter, và Instagram có thể cung cấp thông tin về sở thích và tương tác của khách hàng.

**Dữ liệu công bố:** Dữ liệu từ các nguồn công cộng như dữ liệu chính phủ hoặc tổ chức có thể cung cấp thông tin về thị trường, dân số, hoặc các sự kiện quan trọng.

**Dữ liệu từ thiết bị thông minh:** Dữ liệu từ các thiết bị IoT (Internet of Things) như cảm biến có thể cung cấp thông tin về hiệu suất máy móc, môi trường, hoặc lưu lượng giao thông.

### Truy cập, Quản lý và Lưu trữ Big Data:

**Truy cập dữ liệu:** Đảm bảo rằng bạn có khả năng truy cập vào các nguồn dữ liệu cần thiết. Điều này có thể đòi hỏi sự tích hợp với các hệ thống nguồn dữ liệu khác nhau.

**Quản lý dữ liệu:** Đảm bảo rằng dữ liệu của bạn được quản lý một cách an toàn và tuân theo các quy định về bảo mật dữ liệu.

**Lưu trữ dữ liệu:** Chọn giải pháp lưu trữ phù hợp với quy mô và yêu cầu của bạn. Đám mây có thể cung cấp mô hình lưu trữ linh hoạt và có thể mở rộng.

### Phân Tích Dữ liệu:

Sử dụng công cụ phân tích: Sử dụng các công cụ phân tích dữ liệu như Hadoop, Spark, và các mô hình học máy để khám phá thông tin và tạo dự đoán từ dữ liệu lớn.

### Đưa Ra Quyết Định Dựa Trên Dữ liệu:

Đưa ra quyết định: Dựa trên phân tích dữ liệu, đảm bảo rằng bạn có quyết định chiến lược được dựa trên thông tin cụ thể và chính xác.

Điều chỉnh chiến lược: Liên tục đánh giá và điều chỉnh chiến lược Big Data của bạn để đảm bảo rằng nó phản ánh tình hình thực tế và đáp ứng các mục tiêu kinh doanh của bạn

# Các công nghệ Big Data hiện nay

## Hadoop

### Giới thiệu về Hadoop

#### Lịch sử phát triển

Hadoop xuất phát từ một dự án nghiên cứu của Doug Cutting và Mike Cafarella vào năm 2002. Nó được tạo ra với mục tiêu xử lý và lưu trữ dữ liệu lớn với chi phí thấp. Tên "Hadoop" thậm chí xuất phát từ cái tên của con voi ngựa mà Doug Cutting's con trai đặt tên.

Năm 2005: Hadoop trở thành một dự án mã nguồn mở của Apache và bắt đầu nhận sự quan tâm rộng rãi của cộng đồng.

Năm 2006: Yahoo! bắt đầu sử dụng Hadoop cho công việc của họ và đóng góp cho dự án.

Năm 2008: Hadoop đã phát triển nhanh chóng và nhiều công ty lớn như Facebook và IBM đã tham gia dự án. Điều này đã giúp Hadoop trở thành một nền tảng phổ biến cho việc xử lý dữ liệu lớn.

Năm 2009: Apache Hadoop 0.20, phiên bản quan trọng đầu tiên, được phát hành.

Năm 2012: Hadoop 1.0 ra đời với nhiều cải tiến và ổn định hơn.

Năm 2013: Phiên bản Hadoop 2.0 với YARN (Yet Another Resource Negotiator) cho phép xử lý dữ liệu đa dạng hơn, từ Hadoop MapReduce truyền thống đến các ứng dụng khác.

Năm 2014: Hadoop trở thành một công nghệ không thể thiếu cho việc xử lý dữ liệu lớn và được triển khai rộng rãi trong nhiều lĩnh vực.

Hiện tại: Hadoop tiếp tục phát triển và có nhiều dự án liên quan khác nhau như HBase, Hive, Pig, và nhiều công cụ hỗ trợ khác

#### Tổng quan về Hadoop

Hadoop gồm 2 tầng chính:

Tầng xử lý và tinh toán **Hadoop MapReduce (MapReduce)**: MapReduce là mô hình lập trình cho xử lý dữ liệu phân tán trên Hadoop. Nó chia tác vụ xử lý thành hai phần: bước Map, nơi dữ liệu được tạo ra từ các nguồn khác nhau và sau đó bước Reduce, nơi dữ liệu được tổng hợp và xử lý cuối cùng. MapReduce cho phép xử lý dữ liệu lớn theo cách phân tán và có thể tự động chia tác vụ xử lý trên nhiều nút máy tính, giúp tăng tốc độ xử lý MapReduce là một mô hình lập trình song song hóa để xử lý dữ liệu lớn trên một cụm gồm nhiều các máy tính thương mại (commodity hardware)

Tầng lưu trữ **Hadoop Distributed File System (HDFS)**: HDFS là hệ thống tệp phân tán dùng để lưu trữ dữ liệu trên nhiều máy tính. Dữ liệu được chia thành các khối (block) và lưu trữ trên nhiều nút máy tính. HDFS thiết kế để đảm bảo độ tin cậy và khả năng mở rộng ngang, cho phép lưu trữ và quản lý dữ liệu lớn một cách hiệu quả

Ngoài 2 thành phần được đề cập ở trên thì Hadoop framework cũng gồm **2 module** sau:

**Hadoop Common:** là các thư viện, tiện ích viết bằng ngôn ngữ Java

**Hadoop Yarn:** lập lịch và quản lý các tài nguyên

### Cơ cấu và nguyên tắc hoạt động của Hadoop

#### HDFS và cách nó lưu trữ dữ liệu phân tán

Hệ thống tệp phân tán Hadoop (HDFS) là một hệ thống tệp phân tán dành riêng cho việc lưu trữ và quản lý dữ liệu lớn. Điểm khác biệt quan trọng của HDFS so với các hệ thống tệp phân tán khác bao gồm:

Khả năng chịu lỗi cao: HDFS được thiết kế để đảm bảo độ tin cậy và khả năng phục hồi từ lỗi. Dữ liệu được sao chép và phân tán trên nhiều máy tính, vì vậy nếu một máy tính gặp sự cố, dữ liệu vẫn có thể truy cập thông qua các bản sao dự phòng.

Thiết kế trên phần cứng giá rẻ: HDFS được phát triển để hoạt động trên phần cứng thông thường, chẳng hạn như các máy chủ x86 giá rẻ. Điều này giúp giảm chi phí triển khai hệ thống lưu trữ dữ liệu lớn.

Khả năng mở rộng: HDFS có khả năng mở rộng ngang, cho phép bạn thêm nhiều nút máy tính vào cụm Hadoop để tăng dung lượng lưu trữ và xử lý dữ liệu.

Quyền truy cập thông lượng cao: HDFS được thiết kế để hỗ trợ truy cập thông lượng cao vào dữ liệu ứng dụng. Điều này làm cho nó phù hợp với các ứng dụng có tập dữ liệu lớn cần xử lý trực tiếp trên dữ liệu lưu trữ.

Thay đổi một số yêu cầu POSIX: HDFS chấp nhận một số thay đổi so với giao diện tệp POSIX truyền thống để đảm bảo hiệu suất và độ tin cậy. Chẳng hạn, nó không hỗ trợ ghi và đọc ngẫu nhiên dữ liệu, mà thay vào đó tập trung vào việc đọc và ghi lớn theo cách tuần tự.

HDFS ban đầu được xây dựng làm cơ sở hạ tầng cho dự án công cụ tìm kiếm web Apache Nutch. HDFS là một phần của dự án Apache Hadoop Core. URL dự án là <http://hadoop.apache.org/>

#### Cơ chế của MapReduce và cách nó xử lý dữ liệu

Cơ chế của MapReduce là một mô hình lập trình để xử lý các tập dữ liệu lớn trên một cụm máy tính phân tán. Cơ chế của MapReduce có thể được mô tả qua các bước sau:

Đọc dữ liệu đầu vào: MapReduce sử dụng HDFS (Hadoop Distributed File System) để lưu trữ và đọc dữ liệu đầu vào. HDFS chia dữ liệu thành các khối có kích thước nhất định (mặc định là 64 MB) và phân bổ chúng trên các máy tính trong cụm.

Xử lý dữ liệu đầu vào: MapReduce gán một hàm Map cho mỗi khối dữ liệu, và thực thi hàm Map trên máy tính nơi khối dữ liệu được lưu trữ. Hàm Map xử lý dữ liệu trong khối và xuất ra các cặp khóa/giá trị trung gian.

Sắp xếp và trộn các kết quả: MapReduce sắp xếp các cặp khóa/giá trị trung gian theo khóa, và gom nhóm các giá trị có cùng khóa lại với nhau. Sau đó, + + MapReduce phân phối các nhóm này cho các máy tính khác nhau để thực hiện hàm Reduce.

Tổng hợp các kết quả trung gian: MapReduce thực thi một hàm Reduce cho mỗi nhóm khóa/giá trị trung gian, và xuất ra các cặp khóa/giá trị cuối cùng. Hàm Reduce có thể thực hiện các phép toán như tổng, trung bình, min, max, hoặc bất kỳ hàm nào do người dùng định nghĩa.

+ Đưa ra kết quả cuối cùng: MapReduce ghi kết quả cuối cùng vào HDFS hoặc bất kỳ nguồn nào do người dùng chỉ định.

### Các thành phần của Hadoop

Các thành phần của Hadoop là những module hoặc phần mềm hỗ trợ cho việc lưu trữ và xử lý dữ liệu lớn trên một cụm máy tính phân tán. Theo kết quả tìm kiếm trên web, có thể liệt kê các thành phần chính của Hadoop như sau:

**Hadoop Common**: Đây là một bộ thư viện tiện ích cung cấp các chức năng cơ bản cho các module khác của Hadoop. Hadoop Common chứa các mã lệnh Java để khởi động và chạy Hadoop, cũng như các file và lớp OS trừu tượng.

**Hadoop Distributed File System (HDFS):** Đây là một hệ thống file phân tán cho phép lưu trữ và truy cập dữ liệu lớn với thông lượng cao. HDFS chia dữ liệu thành các khối có kích thước nhất định (mặc định là 64 MB) và phân bổ chúng trên các máy tính trong cụm. HDFS sử dụng kiến trúc master/slave, trong đó master gồm một NameNode để quản lý metadata và một hay nhiều slave DataNodes để lưu trữ dữ liệu thực tế.

**Hadoop MapReduce:** Đây là một framework dựa trên YARN để xử lý song song các tập dữ liệu lớn. MapReduce có hai hàm chính là Map và Reduce, được định nghĩa bởi người dùng. Hàm Map nhận đầu vào là các cặp khóa/giá trị và xuất ra các cặp khóa/giá trị trung gian. Hàm Reduce nhận các khóa trung gian và các giá trị liên quan, và xuất ra các cặp khóa/giá trị cuối cùng.

**Hadoop YARN:** Đây là một framework để quản lý tài nguyên và lập lịch cho các tiến trình của Hadoop. YARN viết tắt của Yet Another Resource Negotiator, là một nền tảng cho việc tính toán phân tán. YARN gồm hai thành phần chính là ResourceManager để quản lý các nguồn lực của cụm và ApplicationMaster để quản lý vòng đời của các ứng dụng

### Ứng dụng

Ứng dụng của Hadoop là rất đa dạng và phong phú, bởi Hadoop là một công nghệ cho phép lưu trữ và xử lý các tập dữ liệu lớn trên một cụm máy tính phân tán. Một số ứng dụng tiêu biểu của Hadoop có thể kể đến như sau:

* **Phân tích dữ liệu**: Hadoop có thể được sử dụng để phân tích các loại dữ liệu khác nhau, như dữ liệu văn bản, hình ảnh, video, âm thanh, địa lý, mạng xã hội, v.v. Hadoop có thể kết hợp với các công cụ khác như Apache Spark, Apache Hive, Apache Pig, để thực hiện các phân tích như phân loại, gom nhóm, khai phá quy luật, khuyến nghị, v.v. [Một số ví dụ về ứng dụng phân tích dữ liệu với Hadoop là: Facebook sử dụng Hadoop để phân tích hành vi người dùng và tạo ra các quảng cáo và nội dung phù hợp](https://tenten.vn/tin-tuc/hadoop-la-gi/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Netflix sử dụng Hadoop để phân tích lịch sử xem của người dùng và đưa ra các gợi ý phim và chương trình2](https://topdev.vn/blog/hadoop-la-gi/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Amazon sử dụng Hadoop để phân tích lịch sử mua hàng của người dùng và đề xuất các sản phẩm liên quan](https://tenten.vn/tin-tuc/hadoop-la-gi/).
* **Tìm kiếm thông tin**: Hadoop có thể được sử dụng để tạo ra các công cụ tìm kiếm thông tin hiệu quả và nhanh chóng. Hadoop có thể xử lý các tập dữ liệu lớn và đa dạng, cũng như áp dụng các thuật toán xếp hạng và tối ưu hoá để trả về các kết quả tìm kiếm chính xác và phù hợp. [Một số ví dụ về ứng dụng tìm kiếm thông tin với Hadoop là: Google sử dụng Hadoop để xây dựng Google File System (GFS) và Google MapReduce, hai thành phần cốt lõi của công cụ tìm kiếm Google](https://thuthuat.taimienphi.vn/hadoop-la-gi-61162n.aspx" \t "https://edgeservices.bing.com/edgesvc/_blank). Yahoo sử dụng Hadoop để xây dựng Yahoo Search Webmap, một ứng dụng tạo ra một bản đồ toàn cầu của web và cung cấp các kết quả tìm kiếm liên quan. Bing sử dụng Hadoop để xây dựng Bing Search Platform (BSP), một nền tảng cho phép Bing cập nhật liên tục các chỉ số tìm kiếm và cải thiện chất lượng kết quả.
* **Máy học**: Hadoop có thể được sử dụng để thực hiện các ứng dụng máy học trên các tập dữ liệu lớn. Hadoop có thể kết hợp với các thư viện máy học như Apache Mahout, Apache MLlib, TensorFlow, để huấn luyện và triển khai các mô hình máy học như hồi quy, phân loại, gom nhóm, mạng nơ-ron, v.v. Một số ví dụ về ứng dụng máy học với Hadoop là: LinkedIn sử dụng Hadoop để huấn luyện và triển khai mô hình People You May Know (PYMK), một ứng dụng đề xuất người bạn có thể biết trên mạng xã hội. Twitter sử dụng Hadoop để huấn luyện và triển khai mô hình Who To Follow (WTF), một ứng dụng đề xuất người dùng có thể theo dõi trên mạng xã hội. IBM sử dụng Hadoop để huấn luyện và triển khai mô hình Watson, một ứng dụng trí tuệ nhân tạo có thể trả lời các câu hỏi phức tạp.

Đây là một số ứng dụng của Hadoop mà tôi có thể tìm thấy. Tất nhiên, còn rất nhiều ứng dụng khác của Hadoop trong các lĩnh vực như y tế, giáo dục, bảo mật, v.v. Hadoop là một công nghệ có tiềm năng lớn và đang được phát triển và cải tiến liên tục.

### Lợi ích và thách thức

**Lợi ích của Hadoop:**

Hadoop cho phép lưu trữ và xử lý các tập dữ liệu lớn trên một cụm máy tính phân tán, với chi phí thấp và hiệu suất cao. Hadoop có thể mở rộng linh hoạt từ vài nút đến hàng ngàn nút, và có khả năng chịu lỗi cao nhờ tính năng sao chép dữ liệu.

Hadoop cung cấp một framework MapReduce để xử lý song song các tập dữ liệu lớn, với mô hình lập trình đơn giản và dễ hiểu. MapReduce cho phép các nhà phát triển viết các chương trình có thể xử lý song song một lượng lớn dữ liệu trên một số lượng lớn các nút.

Hadoop hỗ trợ nhiều loại dữ liệu khác nhau, từ dữ liệu cấu trúc, bán cấu trúc đến không cấu trúc. Hadoop có thể xử lý các dữ liệu văn bản, hình ảnh, video, âm thanh, địa lý, mạng xã hội, v.v. Hadoop không yêu cầu người dùng định nghĩa cấu trúc của dữ liệu trước khi xử lý.

Hadoop có một cộng đồng mã nguồn mở rất lớn và sôi động, với sự đóng góp của nhiều tổ chức và cá nhân. Hadoop được phát triển và cải tiến liên tục để đáp ứng nhu cầu của người dùng. Hadoop cũng có nhiều công cụ và thư viện hỗ trợ cho việc phân tích, máy học, biểu đồ, v.v.

**Thách thức của Hadoop:**

Hadoop yêu cầu người dùng có kiến thức về Java, web, và các framework liên quan để sử dụng hiệu quả. Hadoop cũng yêu cầu người dùng có kỹ năng quản lý và vận hành các cụm máy tính phân tán, bao gồm việc giám sát, bảo trì, sao lưu, khắc phục sự cố, v.v.

Hadoop không phù hợp cho các tác vụ xử lý dữ liệu theo thời gian thực hoặc tương tác. Hadoop được thiết kế để xử lý các tập dữ liệu lớn theo kiểu hàng loạt (batch), với thông lượng cao nhưng độ trễ cao. Hadoop không hỗ trợ các hoạt động ghi ngẫu nhiên hoặc chỉnh sửa tệp.

Hadoop gặp khó khăn trong việc bảo mật và tuân thủ các quy định về dữ liệu. Hadoop không có một giải pháp bảo mật toàn diện cho việc xác thực, ủy quyền, mã hóa và kiểm soát truy cập. Hadoop cũng không có một giải pháp quản lý chất lượng dữ liệu toàn diện để đảm bảo tính nhất quán, chính xác và đầy đủ của dữ liệu.

### Tương lai của Hadoop

Tương lai của Hadoop là một chủ đề thú vị và có nhiều ý kiến khác nhau. Theo kết quả tìm kiếm trên web, có thể nói rằng Hadoop là một công nghệ có tiềm năng lớn và đang được phát triển và cải tiến liên tục để đáp ứng nhu cầu của người dùng trong việc lưu trữ và xử lý các tập dữ liệu lớn trên một cụm máy tính phân tán. Một số xu hướng và hướng phát triển của Hadoop có thể kể đến như sau:

Hadoop sẽ hỗ trợ nhiều loại dữ liệu khác nhau, không chỉ là dữ liệu cấu trúc và bán cấu trúc, mà còn là dữ liệu không cấu trúc như hình ảnh, video, âm thanh, v.v. Hadoop sẽ kết hợp với các công cụ và thư viện khác như Apache Spark, Apache Hive, Apache Pig, TensorFlow, để thực hiện các phân tích, máy học, biểu đồ, v.v. trên các loại dữ liệu này.

Hadoop sẽ cải thiện khả năng xử lý dữ liệu theo thời gian thực hoặc tương tác. Hadoop hiện tại được thiết kế để xử lý các tập dữ liệu lớn theo kiểu hàng loạt (batch), với thông lượng cao nhưng độ trễ cao. Hadoop sẽ phát triển các giải pháp cho phép xử lý dữ liệu liên tục (streaming), với độ trễ thấp và khả năng phản hồi nhanh.

Hadoop sẽ tăng cường khả năng bảo mật và tuân thủ các quy định về dữ liệu. Hadoop hiện tại không có một giải pháp bảo mật toàn diện cho việc xác thực, ủy quyền, mã hóa và kiểm soát truy cập. Hadoop cũng không có một giải pháp quản lý chất lượng dữ liệu toàn diện để đảm bảo tính nhất quán, chính xác và đầy đủ của dữ liệu. Hadoop sẽ tích hợp với các công nghệ và tiêu chuẩn bảo mật khác như Kerberos, SSL/TLS, LDAP, OAuth, v.v. để bảo vệ dữ liệu khỏi các cuộc tấn công và rò rỉ.

## Apache Spark

### Giới thiệu

Apache Spark là một hệ thống xử lý dữ liệu mã nguồn mở và phân tán, được thiết kế để xử lý các tác vụ xử lý dữ liệu lớn và phức tạp. Nó ra đời để đáp ứng các thách thức liên quan đến xử lý và phân tích dữ liệu lớn, đặc biệt là trong môi trường phân tán. Dự án Apache Spark được phát triển bởi Apache Software Foundation và đã nhanh chóng trở thành một công cụ quan trọng trong lĩnh vực xử lý dữ liệu lớn và trí tuệ nhân tạo.

Apache Spark™ là một công cụ thống nhất cho phân tích dữ liệu quy mô lớn. Nó cung cấp các API cấp cao trong Java, Scala, Python và R, và một engine được tối ưu hóa hỗ trợ đồ thị thực thi chung. Dưới đây là một số tính năng chính của Apache Spark™:

Dữ liệu theo lô/streaming: Hợp nhất việc xử lý dữ liệu của bạn theo lô và streaming thời gian thực, sử dụng ngôn ngữ ưa thích của bạn: Python, SQL, Scala, Java hoặc R.

Phân tích SQL: Thực hiện các truy vấn SQL ANSI nhanh, phân tán cho việc tạo bảng điều khiển và báo cáo ad-hoc. Chạy nhanh hơn hầu hết các kho dữ liệu.

Khoa học dữ liệu ở quy mô lớn: Thực hiện Phân tích Dữ liệu Khám phá (EDA) trên dữ liệu quy mô petabyte mà không phải recurr đến downsampling.

Học máy: Đào tạo các thuật toán học máy trên máy tính xách tay và sử dụng cùng một đoạn mã để mở rộng đến chế độ chịu lỗi.

### Kiến trúc của Apache Spark

[Kiến trúc của Apache Spark dựa trên hai trừu tượng chính](https://www.edureka.co/blog/spark-architecture/" \t "https://edgeservices.bing.com/edgesvc/_blank):

Resilient Distributed Dataset (RDD): RDD là một tập hợp phân tán mà có thể chịu được sự cố. [Nó là một tập hợp các phần tử phân tán trên nhiều nút của cụm để thực hiện song song hoặc để chịu được lỗi](https://www.javatpoint.com/apache-spark-architecture" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Directed Acyclic Graph (DAG): DAG là một đồ thị hướng không chu trình, được sử dụng để lưu trữ tất cả các chi tiết về các toán tử RDD1](https://www.edureka.co/blog/spark-architecture/" \t "https://edgeservices.bing.com/edgesvc/_blank).

Kiến trúc của Spark tuân theo kiến trúc master-slave. [Cụm của nó bao gồm một master duy nhất và nhiều slave](https://www.javatpoint.com/apache-spark-architecture" \t "https://edgeservices.bing.com/edgesvc/_blank). [Kiến trúc Spark bao gồm bốn thành phần, bao gồm driver spark, executor, quản lý cụm và nút worker](https://www.interviewbit.com/blog/apache-spark-architecture/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Nó sử dụng Dataset và data frames như cơ chế lưu trữ dữ liệu cơ bản để tối ưu hóa quá trình Spark và tính toán big data](https://www.interviewbit.com/blog/apache-spark-architecture/" \t "https://edgeservices.bing.com/edgesvc/_blank).

### Xử lý dữ liệu

[Apache Spark là một công cụ xử lý dữ liệu mã nguồn mở cho các tập dữ liệu lớn](https://www.bing.com/aclick?ld=e8FJ8KtFl8wWapq3EIuM7KiTVUCUwjMEVSsQ0LZI86evWHmqccugiZOSN9dBArX-PZXPdLfZ1DFRLNMBWdpa9NayjDicgSuF22f45jf6UgEp4U6Zu-Pyt-ne32Z3uKLMTNJ4kpaoh7wZps9owOlSoG9UMit-NQKaec6NFR9dl3zNjemaDH&u=&rlid=caf4980a47fc178679bafd3439af8ad4" \t "https://edgeservices.bing.com/edgesvc/_blank). [Nó được thiết kế để cung cấp tốc độ tính toán, khả năng mở rộng và khả năng lập trình cần thiết cho Big Data - cụ thể là cho dữ liệu luồng, dữ liệu đồ thị, học máy và ứng dụng trí tuệ nhân tạo (AI)](https://www.bing.com/aclick?ld=e8FJ8KtFl8wWapq3EIuM7KiTVUCUwjMEVSsQ0LZI86evWHmqccugiZOSN9dBArX-PZXPdLfZ1DFRLNMBWdpa9NayjDicgSuF22f45jf6UgEp4U6Zu-Pyt-ne32Z3uKLMTNJ4kpaoh7wZps9owOlSoG9UMit-NQKaec6NFR9dl3zNjemaDH&u=&rlid=caf4980a47fc178679bafd3439af8ad4" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Apache Spark cung cấp các API cấp cao trong Java, Scala, Python và R, và một engine được tối ưu hóa hỗ trợ đồ thị thực thi chung](https://www.ibm.com/topics/apache-spark). [Nó cũng hỗ trợ một tập hợp phong phú các công cụ cấp cao bao gồm Spark SQL cho SQL và xử lý dữ liệu có cấu trúc, API pandas trên Spark cho các công việc pandas, MLlib cho học máy, GraphX cho xử lý đồ thị và Structured Streaming cho tính toán tăng dần và xử lý luồng](https://spark.apache.org/docs/latest/" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Spark có thể chạy trên Apache Hadoop, Apache Mesos, Kubernetes, độc lập, trong đám mây và đối với các nguồn dữ liệu đa dạng](https://cloud.google.com/learn/what-is-apache-spark" \t "https://edgeservices.bing.com/edgesvc/_blank). [Nó được sử dụng rộng rãi trong các doanh nghiệp trên nhiều ngành công nghiệp](https://www.databricks.com/spark/about" \t "https://edgeservices.bing.com/edgesvc/_blank).

### Máy học AI với MLlib

[MLlib là một thư viện của Apache Spark, cung cấp các công cụ học máy (Machine Learning) trên nền tảng phân tán](https://onetech.vn/blog/mot-so-artificial-intelligence-frameworks-va-thu-vien-can-biet-6514" \t "https://edgeservices.bing.com/edgesvc/_blank). [MLlib bao gồm các thuật toán học máy phổ biến như phân loại, hồi quy, gom nhóm và lọc cộng tác, cũng như các công cụ hỗ trợ như đánh giá mô hình và xử lý dữ liệu](https://onetech.vn/blog/mot-so-artificial-intelligence-frameworks-va-thu-vien-can-biet-6514" \t "https://edgeservices.bing.com/edgesvc/_blank).

[MLlib giúp tận dụng sức mạnh của Spark, cho phép xử lý song song và phân tán trên các nút của một cụm máy tính, từ đó tăng tốc độ và khả năng mở rộng của quá trình học máy](https://onetech.vn/blog/mot-so-artificial-intelligence-frameworks-va-thu-vien-can-biet-6514" \t "https://edgeservices.bing.com/edgesvc/_blank). [Điều này đặc biệt quan trọng khi làm việc với các tập dữ liệu lớn, điều mà Spark được thiết kế để xử lý](https://onetech.vn/blog/mot-so-artificial-intelligence-frameworks-va-thu-vien-can-biet-6514" \t "https://edgeservices.bing.com/edgesvc/_blank).

### Ứng dụng

[Apache Spark đang được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau từ kinh doanh, y tế, tài chính, giáo dục đến khoa học và công nghệ](https://vegafintech.vn/vi/apache-spark-la-gi-vai-tro-va-ung-dung-cua-apache-spark" \t "https://edgeservices.bing.com/edgesvc/_blank). Dưới đây là một số ứng dụng cụ thể của Apache Spark:

[Dịch vụ tài chính: Apache Spark được sử dụng để dự đoán tỷ lệ tham gia/huỷ bỏ của khách hàng về sản phẩm của họ](https://vegafintech.vn/vi/apache-spark-la-gi-vai-tro-va-ung-dung-cua-apache-spark).

[Ngành y tế: Apache Spark được sử dụng để xây dựng dịch vụ chăm sóc bệnh nhân toàn diện, cho phép tương tác và trao đổi thông tin trực tuyến](https://vegafintech.vn/vi/apache-spark-la-gi-vai-tro-va-ung-dung-cua-apache-spark).

[Phân tích thời gian thực: Apache Spark có thể xử lý hàng triệu sự kiện, thông tin mỗi giây từ các luồng sự kiện thời gian thực như Data Twitter hoặc những lượt chia sẻ, bài đăng trên Facebook](https://vegafintech.vn/vi/apache-spark-la-gi-vai-tro-va-ung-dung-cua-apache-spark" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Phát hiện gian lận: Tất cả các khoản thanh toán đều được thực hiện trong thời gian thực và Apache Spark có thể giúp ngăn chặn phần nào các giao dịch gian lận trong khi quá trình thanh toán đang diễn ra](https://vegafintech.vn/vi/apache-spark-la-gi-vai-tro-va-ung-dung-cua-apache-spark" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Xử lý dữ liệu có tốc độ nhanh: Apache Spark còn có thể được sử dụng để xử lý dữ liệu có tốc độ nhanh và tương tác, xử lý đồ thị, công việc lặp đi lặp lại, xử lý thời gian thực, joining Dataset hay là Machine Learning](https://vegafintech.vn/vi/apache-spark-la-gi-vai-tro-va-ung-dung-cua-apache-spark" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Các tổ chức như Yelp, Zillow, Crowdstrike, Hearst, Bigfinite và Gumgum đã chọn Apache Spark cho các ứng dụng của họ](https://bkhost.vn/blog/apache-spark/" \t "https://edgeservices.bing.com/edgesvc/_blank).

### Thách thức và tương lai

[Apache Spark đã trở thành một công cụ quan trọng trong việc xử lý dữ liệu quy mô lớn và học máy](https://meeyland.com/chuyen-doi-so/apache-spark-la-gi-nhung-uu-diem-noi-bat-cua-apache-spark" \t "https://edgeservices.bing.com/edgesvc/_blank). Tuy nhiên, như bất kỳ công nghệ nào khác, Apache Spark cũng đối mặt với một số thách thức:

[Độ trễ: Mặc dù Spark có thể xử lý dữ liệu theo thời gian thực, tuy nhiên, độ trễ vẫn còn tồn tại và có thể ảnh hưởng đến hiệu suất của các ứng dụng yêu cầu thời gian thực](https://meeyland.com/chuyen-doi-so/apache-spark-la-gi-nhung-uu-diem-noi-bat-cua-apache-spark" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Quản lý bộ nhớ: Spark sử dụng RAM để lưu trữ dữ liệu, điều này có thể dẫn đến vấn đề về bộ nhớ khi xử lý các tập dữ liệu lớn](https://meeyland.com/chuyen-doi-so/apache-spark-la-gi-nhung-uu-diem-noi-bat-cua-apache-spark).

[Khả năng mở rộng: Dù Spark có khả năng mở rộng tốt, nhưng việc mở rộng quy mô lớn vẫn còn là một thách thức](https://meeyland.com/chuyen-doi-so/apache-spark-la-gi-nhung-uu-diem-noi-bat-cua-apache-spark" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Về tương lai của Apache Spark, nó sẽ tiếp tục phát triển và trở thành công cụ không thể thiếu trong việc xử lý dữ liệu quy mô lớn và học máy](https://meeyland.com/chuyen-doi-so/apache-spark-la-gi-nhung-uu-diem-noi-bat-cua-apache-spark). [Các nhà phát triển sẽ tiếp tục cải tiến Spark để giải quyet các thách thức hiện tại và đáp ứng nhu cầu ngày càng tăng của người dùng](https://meeyland.com/chuyen-doi-so/apache-spark-la-gi-nhung-uu-diem-noi-bat-cua-apache-spark" \t "https://edgeservices.bing.com/edgesvc/_blank).

## NoSQL databases

### Giới thiệu

[NoSQL Databases là viết tắt của “Not Only SQL” hoặc “Not SQL”](https://meeyland.com/chuyen-doi-so/tim-hieu-nhung-thong-tin-co-ban-ve-nosql-databases/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Đây là một hệ cơ sở dữ liệu không sử dụng mô hình dữ liệu quan hệ, đặc biệt nhấn mạnh đến mô hình lưu trữ cặp giá trị - khóa và hệ thống lưu trữ phân tán](https://meeyland.com/chuyen-doi-so/tim-hieu-nhung-thong-tin-co-ban-ve-nosql-databases/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Mục đích chính của việc sử dụng NoSQL Database là dành cho các kho dữ liệu phân tán với nhu cầu lưu trữ dữ liệu lớn](https://meeyland.com/chuyen-doi-so/tim-hieu-nhung-thong-tin-co-ban-ve-nosql-databases/" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Các ưu điểm của NoSQL Databases bao gồm khả năng mở rộng và tính sẵn sàng cao](https://meeyland.com/chuyen-doi-so/tim-hieu-nhung-thong-tin-co-ban-ve-nosql-databases/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Cơ sở dữ liệu NoSQL sử dụng sharding để chia tỷ lệ ngang, thêm nhiều máy hơn để xử lý dữ liệu thay vì phải bổ sung thêm những máy chủ lớn hơn để tải nhiều dữ liệu hơn](https://meeyland.com/chuyen-doi-so/tim-hieu-nhung-thong-tin-co-ban-ve-nosql-databases/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Nếu có bất kỳ lỗi nào, dữ liệu sẽ tự động sao chép về trạng thái nhất quán trước đó](https://meeyland.com/chuyen-doi-so/tim-hieu-nhung-thong-tin-co-ban-ve-nosql-databases/).

Tuy nhiên, NoSQL Databases cũng có một số hạn chế. Quản lý dữ liệu trong NoSQL phức tạp hơn nhiều so với SQL. [Đặc biệt, NoSQL khó cài đặt và tốn nhiều thời gian để quản lý hàng ngày](https://meeyland.com/chuyen-doi-so/tim-hieu-nhung-thong-tin-co-ban-ve-nosql-databases/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Một số cơ sở dữ liệu NoSQL như MongoDB không có cách tiếp cận để sao lưu dữ liệu một cách nhất quán](https://meeyland.com/chuyen-doi-so/tim-hieu-nhung-thong-tin-co-ban-ve-nosql-databases/" \t "https://edgeservices.bing.com/edgesvc/_blank).

[NoSQL thích hợp cho những người dùng muốn truy cập nhanh vào dữ liệu, mục tiêu là quan tâm nhiều hơn đến tốc độ, sự đơn giản của việc truy cập hơn là sự đáng tin cậy hoặc tính nhất quán của các giao dịch](https://meeyland.com/chuyen-doi-so/tim-hieu-nhung-thong-tin-co-ban-ve-nosql-databases/).

### Phân loại

[Cơ sở dữ liệu NoSQL có bốn loại chính](https://aws.amazon.com/vi/nosql/" \t "https://edgeservices.bing.com/edgesvc/_blank):

Document databases: Đây là loại cơ sở dữ liệu tài liệu lưu trữ và truy xuất dữ liệu dưới dạng một cặp khóa giá trị. [Các giá trị đa dạng chẳng hạn như số, chuỗi, đối tượng…](https://bing.com/search?q=Ph%c3%a2n+lo%e1%ba%a1i+NoSQL+databases) *[Ví dụ:](https://bing.com/search?q=Ph%c3%a2n+lo%e1%ba%a1i+NoSQL+databases)* [Apache CouchDB, Clusterpoint, Couchbase, DocumentDB, HyperDex, Lotus Notes, MarkLogic, MongoDB, OrientDB, Qizx, RethinkDB](https://bing.com/search?q=Ph%c3%a2n+lo%e1%ba%a1i+NoSQL+databases).

[Key – Value databases: NoSQL loại này khá đơn giản vì mỗi mục của nó chỉ chứa key và value](https://bing.com/search?q=Ph%c3%a2n+lo%e1%ba%a1i+NoSQL+databases). *[Ví dụ:](https://bing.com/search?q=Ph%c3%a2n+lo%e1%ba%a1i+NoSQL+databases)* [Aerospike, CouchDB, Dynamo, FairCom c-treeACE, FoundationDB, HyperDex, MemcacheDB, MUMPS, Oracle NoSQL Database, OrientDB, Redis, Riak, Berkeley DB](https://bing.com/search?q=Ph%c3%a2n+lo%e1%ba%a1i+NoSQL+databases).

[Column: Accumulo, Cassandra, Druid, HBase, Vertica](https://bing.com/search?q=Ph%c3%a2n+lo%e1%ba%a1i+NoSQL+databases)[1](https://bing.com/search?q=Ph%C3%A2n+lo%E1%BA%A1i+NoSQL+databases" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Graph databases: Dữ liệu ở Graph databases được lưu trữ trong các nút và các cạnh](https://aws.amazon.com/vi/nosql/" \t "https://edgeservices.bing.com/edgesvc/_blank).

### Kiến trúc và cách thức hoạt động

[Cơ sở dữ liệu NoSQL lưu trữ dữ liệu của mình theo dạng cặp giá trị “key – value”](https://bing.com/search?q=Ki%e1%ba%bfn+tr%c3%bac+v%c3%a0+c%c3%a1ch+th%e1%bb%a9c+ho%e1%ba%a1t+%c4%91%e1%bb%99ng+c%e1%bb%a7a+NoSQL+databases). [Nó sử dụng số lượng lớn các node để lưu trữ thông tin](https://bing.com/search?q=Ki%e1%ba%bfn+tr%c3%bac+v%c3%a0+c%c3%a1ch+th%e1%bb%a9c+ho%e1%ba%a1t+%c4%91%e1%bb%99ng+c%e1%bb%a7a+NoSQL+databases)[1](https://bing.com/search?q=Ki%E1%BA%BFn+tr%C3%BAc+v%C3%A0+c%C3%A1ch+th%E1%BB%A9c+ho%E1%BA%A1t+%C4%91%E1%BB%99ng+c%E1%BB%A7a+NoSQL+databases" \t "https://edgeservices.bing.com/edgesvc/_blank). [Mô hình phân tán dưới sự kiểm soát phần mềm](https://bing.com/search?q=Ki%e1%ba%bfn+tr%c3%bac+v%c3%a0+c%c3%a1ch+th%e1%bb%a9c+ho%e1%ba%a1t+%c4%91%e1%bb%99ng+c%e1%bb%a7a+NoSQL+databases). [Chấp nhận dữ liệu bị trùng lặp do một số node sẽ lưu cùng thông tin giống nhau](https://bing.com/search?q=Ki%e1%ba%bfn+tr%c3%bac+v%c3%a0+c%c3%a1ch+th%e1%bb%a9c+ho%e1%ba%a1t+%c4%91%e1%bb%99ng+c%e1%bb%a7a+NoSQL+databases).

[Các cơ sở dữ liệu NoSQL thường đơn giản, kích thước nhỏ, lưu phân tán tại nhiều nơi và lưu ở trong RAM nhằm tăng tốc độ xử lý](https://viblo.asia/p/tong-quan-ve-database-sql-va-nosql-WR5JRmKr4Gv" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Các cơ sở dữ liệu NoSQL có bốn loại chính](https://bkhost.vn/blog/nosql/" \t "https://edgeservices.bing.com/edgesvc/_blank):

Document databases: Đây là loại cơ sở dữ liệu tài liệu lưu trữ và truy xuất dữ liệu dưới dạng một cặp khóa giá trị. [Các giá trị đa dạng chẳng hạn như số, chuỗi, đối tượng…](https://bkhost.vn/blog/nosql/" \t "https://edgeservices.bing.com/edgesvc/_blank)

[Key – Value databases: NoSQL loại này khá đơn giản vì mỗi mục của nó chỉ chứa key và value](https://bkhost.vn/blog/nosql/" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Wide-column stores: Dữ liệu ở NoSQL loại này được lưu trữ trong bảng, cột động và hàng](https://bkhost.vn/blog/nosql/" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Graph databases: Dữ liệu ở Graph databases được lưu trữ trong các nút và các cạnh](https://bkhost.vn/blog/nosql/" \t "https://edgeservices.bing.com/edgesvc/_blank).

### Ứng dụng

[Cơ sở dữ liệu NoSQL được ứng dụng rộng rãi trong nhiều lĩnh vực](https://bkhost.vn/blog/nosql/" \t "https://edgeservices.bing.com/edgesvc/_blank). Dưới đây là một số ví dụ:

[Phát triển Agile hiện đại: Khi nhà phát triển mong muốn phát triển Agile hiện đại với tốc độ nhanh](https://bkhost.vn/blog/nosql/" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Lưu trữ dữ liệu có cấu trúc, bán cấu trúc: Khi muốn lưu trữ dữ liệu có cấu trúc, bán cấu trúc](https://bkhost.vn/blog/nosql/" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Lượng dữ liệu lớn: Khi lượng dữ liệu cần lưu trữ quá lớn, vượt quá mức cho phép của cơ sở dữ liệu quan hệ](https://bkhost.vn/blog/nosql/" \t "https://edgeservices.bing.com/edgesvc/_blank).

Các công ty, tập đoàn lớn: NoSQL được sử dụng ở rất nhiều công ty, tập đoàn lớn. Ví dụ như FaceBook sử dụng Cassandra do FaceBook phát triển. [Google phát triển và sử dụng BigTable](https://memart.vn/tin-tuc/blog/tim-hieu-nosql-database-la-gi-va-nhung-ung-dung-cua-no-trong-linh-vuc-big-data-vi-cb.html" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Ứng dụng trong điện toán đám mây: Hiện tại nhiều tổ chức sử dụng NoSQL để cung cấp khả năng mở rộng quy mô phân phối dữ liệu trên điện toán đám mây](https://bkhost.vn/blog/nosql/" \t "https://edgeservices.bing.com/edgesvc/_blank).

### Thách thức và tương lai

[Cơ sở dữ liệu NoSQL đang đối mặt với một số thách thức](https://bkhost.vn/blog/nosql/" \t "https://edgeservices.bing.com/edgesvc/_blank):

[Khả năng mở rộng: Mặc dù NoSQL có khả năng mở rộng tốt, nhưng việc mở rộng quy mô lớn vẫn còn là một thách thức](https://bkhost.vn/blog/nosql/).

Quản lý dữ liệu: Quản lý dữ liệu trong NoSQL phức tạp hơn nhiều so với SQL. [Đặc biệt, NoSQL khó cài đặt và tốn nhiều thời gian để quản lý hàng ngày](https://bkhost.vn/blog/nosql/).

[Độ trễ: Mặc dù NoSQL có thể xử lý dữ liệu theo thời gian thực, tuy nhiên, độ trễ vẫn còn tồn tại và có thể ảnh hưởng đến hiệu suất của các ứng dụng yêu cầu thời gian thực](https://bkhost.vn/blog/nosql/).

[Về tương lai của NoSQL, nó sẽ tiếp tục phát triển và trở thành công cụ không thể thiếu trong việc xử lý dữ liệu quy mô lớn và học máy](https://bkhost.vn/blog/nosql/)[3](https://m.cpc.vn/vi-vn/Tin-tuc-su-kien/Tin-tuc-chi-tiet/articleId/54352" \t "https://edgeservices.bing.com/edgesvc/_blank). [Các nhà phát triển sẽ tiếp tục cải tiến NoSQL để giải quyet các thách thức hiện tại và đáp ứng nhu cầu ngày càng tăng của người dùng](https://bkhost.vn/blog/nosql/).

## SQL - on - Hadoop

### Giới thiệu

[SQL-on-Hadoop là một loạt các công cụ được thiết kế để thực hiện các truy vấn SQL trực tiếp trên dữ liệu lưu trữ trong Apache Hadoop](https://topdev.vn/blog/hadoop-la-gi/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Công nghệ này cung cấp một cách tiếp cận cho việc phân tích dữ liệu quy mô lớn, và là một phần của xu hướng bao gồm SQL và NoSQL, thay vì chọn một trong hai](https://topdev.vn/blog/hadoop-la-gi/).

[Hadoop là một framework mã nguồn mở cho phép phát triển các ứng dụng phân tán để lưu trữ và quản lý các tập dữ liệu lớn](https://topdev.vn/blog/hadoop-la-gi/). [Hadoop hiện thực mô hình MapReduce, mô hình mà ứng dụng sẽ được chia nhỏ ra thành nhiều phân đoạn khác nhau được chạy song song trên nhiều node khác nhau](https://topdev.vn/blog/hadoop-la-gi/" \t "https://edgeservices.bing.com/edgesvc/_blank).

[SQL-on-Hadoop giúp kết hợp sức mạnh của SQL, ngôn ngữ truy vấn dữ liệu phổ biến nhất, với Hadoop, hệ thống lưu trữ và xử lý dữ liệu quy mô lớn](https://topdev.vn/blog/hadoop-la-gi/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Điều này giúp cho việc phân tích dữ liệu trở nên dễ dàng và hiệu quả hơn](https://topdev.vn/blog/hadoop-la-gi/" \t "https://edgeservices.bing.com/edgesvc/_blank).

### Các công cụ SQL-on-Hadoop

Có một số công cụ SQL-on-Hadoop phổ biến mà bạn có thể muốn xem xét:

1. [Interbase: Đây là một cơ sở dữ liệu SQL siêu nhanh, có thể mở rộng, có thể nhúng với bảo mật dữ liệu cấp thương mại, khôi phục hư hỏng và đồng bộ hóa thay đổi](https://gitiho.com/blog/huong-dan-tim-hieu-20-cong-cu-quan-ly-sql-tot-nhat-nam-2020-phan-1.html" \t "https://edgeservices.bing.com/edgesvc/_blank).
2. [Studio for SQL Server: dbForge Studio for SQL Server là một IDE tích hợp mạnh mẽ, cho phép người dùng quản lý, quản trị, phát triển, báo cáo dữ liệu, phân tích dữ liệu](https://gitiho.com/blog/huong-dan-tim-hieu-20-cong-cu-quan-ly-sql-tot-nhat-nam-2020-phan-1.html" \t "https://edgeservices.bing.com/edgesvc/_blank).
3. [dbWatch: dbWatch là một giải pháp quản lý và giám sát cơ sở dữ liệu hoàn chỉnh cho SQL Server, Oracle, PostgreSQL, Sybase, MySQL và Azure](https://gitiho.com/blog/huong-dan-tim-hieu-20-cong-cu-quan-ly-sql-tot-nhat-nam-2020-phan-1.html" \t "https://edgeservices.bing.com/edgesvc/_blank).
4. [Visual Expert: Visual Expert là một bộ các công cụ phân tích mã mạnh mẽ dành cho cơ sở dữ liệu Oracle và SQL Server](https://gitiho.com/blog/huong-dan-tim-hieu-20-cong-cu-quan-ly-sql-tot-nhat-nam-2020-phan-1.html" \t "https://edgeservices.bing.com/edgesvc/_blank).
5. [Mỗi công cụ có những tính năng và ưu điểm riêng, do đó việc lựa chọn công cụ phù hợp sẽ phụ thuộc vào yêu cầu và nhu cầu cụ thể của bạn](https://gitiho.com/blog/huong-dan-tim-hieu-20-cong-cu-quan-ly-sql-tot-nhat-nam-2020-phan-1.html" \t "https://edgeservices.bing.com/edgesvc/_blank).

### Lợi ích và ứng dụng

[SQL-on-Hadoop mang lại nhiều lợi ích và ứng dụng trong thực tế](https://tenten.vn/tin-tuc/hadoop-la-gi/" \t "https://edgeservices.bing.com/edgesvc/_blank):

[Phát triển Agile hiện đại: SQL-on-Hadoop hỗ trợ phát triển Agile hiện đại với tốc độ nhanh](https://tenten.vn/tin-tuc/hadoop-la-gi/" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Lưu trữ dữ liệu có cấu trúc, bán cấu trúc: SQL-on-Hadoop cho phép lưu trữ dữ liệu có cấu trúc, bán cấu trúc](https://tenten.vn/tin-tuc/hadoop-la-gi/).

[Lượng dữ liệu lớn: SQL-on-Hadoop có khả năng xử lý một lượng lớn dữ liệu](https://tenten.vn/tin-tuc/hadoop-la-gi/).

[Các công ty, tập đoàn lớn: Nhiều công ty, tập đoàn lớn đã sử dụng SQL-on-Hadoop](https://tenten.vn/tin-tuc/hadoop-la-gi/).

[Ứng dụng trong điện toán đám mây: Hiện tại nhiều tổ chức sử dụng SQL-on-Hadoop để cung cấp khả năng mở rộng quy mô phân phối dữ liệu trên điện toán đám mây](https://tenten.vn/tin-tuc/hadoop-la-gi/" \t "https://edgeservices.bing.com/edgesvc/_blank).

[SQL-on-Hadoop giúp kết hợp sức mạnh của SQL, ngôn ngữ truy vấn dữ liệu phổ biến nhất, với Hadoop, hệ thống lưu trữ và xử lý dữ liệu quy mô lớn](https://tenten.vn/tin-tuc/hadoop-la-gi/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Điều này giúp cho việc phân tích dữ liệu trở nên dễ dàng và hiệu quả hơn](https://tenten.vn/tin-tuc/hadoop-la-gi/" \t "https://edgeservices.bing.com/edgesvc/_blank).

## Data Streaming

### Giới thiệu

[Dữ liệu truyền phát, hay còn gọi là Data Streaming, là dữ liệu được phát ở khối lượng lớn theo cách tăng dần, liên tục nhằm mục đích xử lý với độ trễ thấp](https://aws.amazon.com/vi/what-is/streaming-data/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Các tổ chức có hàng nghìn nguồn dữ liệu thường phát ra đồng thời các thông báo, bản ghi hoặc dữ liệu có kích thước từ vài byte đến vài megabyte (MB)](https://aws.amazon.com/vi/what-is/streaming-data/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Dữ liệu truyền bao gồm dữ liệu vị trí, sự kiện và cảm biến mà các công ty sử dụng để quan sát và phân tích theo thời gian thực nhiều khía cạnh kinh doanh của họ](https://aws.amazon.com/vi/what-is/streaming-data/" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Dữ liệu truyền phát có những đặc điểm sau](https://aws.amazon.com/vi/what-is/streaming-data/):

[Có tính quan trọng theo thứ tự thời gian: Từng thành phần trong luồng dữ liệu đều có dấu thời gian](https://aws.amazon.com/vi/what-is/streaming-data/).

[Truyền liên tục: Luồng dữ liệu không có bắt đầu cũng như kết thúc](https://aws.amazon.com/vi/what-is/streaming-data/).

[Duy nhất: Việc lặp lại quá trình truyền của luồng dữ liệu đầy thách thức vì tính nhạy cảm về mặt thời gian](https://aws.amazon.com/vi/what-is/streaming-data/).

[Không đồng nhất: Một số nguồn có thể truyền dữ liệu ở nhiều định dạng có cấu trúc như JSON, Avro và các giá trị được phân tách bằng dấu phẩy (CSV) với các loại dữ liệu bao gồm chuỗi, số, ngày và loại nhị phân](https://aws.amazon.com/vi/what-is/streaming-data/).

[Không hoàn hảo: Những lỗi tạm thời tại nguồn có thể khiến các thành phần trong dữ liệu được truyền bị hỏng hoặc thiếu](https://aws.amazon.com/vi/what-is/streaming-data/).

[Dữ liệu truyền phát rất quan trọng trong việc thu nạp và xây dựng cấu trúc dữ liệu trước khi phân tích](https://aws.amazon.com/vi/what-is/streaming-data/). [Trong những năm gần đây, tính chất của dữ liệu doanh nghiệp và các hệ thống xử lý dữ liệu cơ bản đã thay đổi đáng kể](https://aws.amazon.com/vi/what-is/streaming-data/). [Khối lượng dữ liệu được tạo từ các nguồn luồng dữ liệu có thể rất lớn, do đó việc phân tích theo thời gian thực để quản lý tính toàn vẹn (việc xác thực), cấu trúc (sự phát triển) hay vận tốc (thông lượng và độ trễ) của dữ liệu truyền trở thành thách thức](https://aws.amazon.com/vi/what-is/streaming-data/" \t "https://edgeservices.bing.com/edgesvc/_blank).

### Các hệ thống của Data Streaming phổ biến

Dưới đây là một số hệ thống Data Streaming phổ biến:

1. [Kafka: Đây là công nghệ phổ biến nhất trong lĩnh vực Data Streaming](https://insights.magestore.com/posts/kien-truc-du-lieu-lon).
2. [RabbitMQ: Đây là một hệ thống Data Streaming khác](https://insights.magestore.com/posts/kien-truc-du-lieu-lon" \t "https://edgeservices.bing.com/edgesvc/_blank).
3. [ActiveMQ: Đây cũng là một hệ thống Data Streaming](https://insights.magestore.com/posts/kien-truc-du-lieu-lon" \t "https://edgeservices.bing.com/edgesvc/_blank).
4. [PostgreSQL: PostgreSQL cũng hỗ trợ sao chép dữ liệu dạng luồng (Streaming replication)](https://viblo.asia/p/tim-hieu-he-quan-tri-co-so-du-lieu-postgresql-m68Z0eLdlkG" \t "https://edgeservices.bing.com/edgesvc/_blank).

### Lợi ích và ứng dụng

[Dữ liệu truyền phát mang lại nhiều lợi ích và ứng dụng trong thực tế](https://aws.amazon.com/vi/what-is/streaming-data/" \t "https://edgeservices.bing.com/edgesvc/_blank):

[Phân tích theo thời gian thực: Dữ liệu truyền phát cho phép các doanh nghiệp theo dõi thay đổi trong quan điểm của công chúng đối với thương hiệu và sản phẩm của họ bằng cách liên tục phân tích đường dẫn nhấp chuột và bài đăng của khách hàng từ các luồng dữ liệu mạng xã hội, sau đó phản hồi kịp thời khi cần](https://aws.amazon.com/vi/what-is/streaming-data/" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Quan trọng theo thứ tự thời gian: Từng thành phần trong luồng dữ liệu đều có dấu thời gian](https://aws.amazon.com/vi/what-is/streaming-data/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Bản thân luồng dữ liệu có thể chỉ hữu ích trong thời gian nhất định, với tầm quan trọng giảm sau một khoảng thời gian cụ thể](https://aws.amazon.com/vi/what-is/streaming-data/" \t "https://edgeservices.bing.com/edgesvc/_blank).

[Truyền liên tục: Luồng dữ liệu không có bắt đầu cũng như kết thúc](https://aws.amazon.com/vi/what-is/streaming-data/" \t "https://edgeservices.bing.com/edgesvc/_blank). [Luồng sẽ thu thập dữ liệu liên tục và không ngừng chừng nào còn được yêu cầu](https://aws.amazon.com/vi/what-is/streaming-data/).

[Không đồng nhất: Một số nguồn có thể truyền dữ liệu ở nhiều định dạng có cấu trúc như JSON, Avro và các giá trị được phân tách bằng dấu phẩy (CSV) với các loại dữ liệu bao gồm chuỗi, số, ngày và loại nhị phân](https://aws.amazon.com/vi/what-is/streaming-data/).

[Không hoàn hảo: Những lỗi tạm thời tại nguồn có thể khiến các thành phần trong dữ liệu được truyền bị hỏng hoặc thiếu](https://aws.amazon.com/vi/what-is/streaming-data/). [Việc đảm bảo tính nhất quán của dữ liệu có thể là thách thức do tính chất liên tục của luồng](https://aws.amazon.com/vi/what-is/streaming-data/).

[Dữ liệu truyền phát rất quan trọng trong việc thu nạp và xây dựng cấu trúc dữ liệu trước khi phân tích](https://aws.amazon.com/vi/what-is/streaming-data/). [Trong những năm gần đây, tính chất của dữ liệu doanh nghiệp và các hệ thống xử lý dữ liệu cơ bản đã thay đổi đáng kể](https://aws.amazon.com/vi/what-is/streaming-data/). [Khối lượng dữ liệu được tạo từ các nguồn luồng dữ liệu có thể rất lớn, do đó việc phân tích theo thời gian thực để quản lý tính toàn vẹn (việc xác thực), cấu trúc (sự phát triển) hay vận tốc (thông lượng và độ trễ) của dữ liệu truyền trở thành thách thức](https://aws.amazon.com/vi/what-is/streaming-data/).

# Mô hình đề xuất

# Mô tả dữ liệu

# Kết quả

<https://itnavi.com.vn/blog/big-data/>

<https://vdigital.vn/vai-tro-cua-big-data/>

<https://hadoop.apache.org/>

<https://topdev.vn/blog/hadoop-la-gi/>

<https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html>

<https://cwiki.apache.org/confluence/display/hadoop>

<https://demanejar.github.io/posts/hadoop-introduction/>

<https://sieumarketing.com/mapreduce-la-gi-khai-quat-ve-mo-hinh-lap-trinh/>

<https://itnavi.com.vn/blog/cau-truc-va-thanh-phan-cua-cong-nghe-hadoop>

<https://thuthuat.taimienphi.vn/hadoop-la-gi-61162n.aspx>

<https://appmaster.io/vi/blog/kien-truc-du-lieu-lon-hadoop-apache>

<https://spark.apache.org/docs/latest/>

<https://www.interviewbit.com/blog/apache-spark-architecture/>

<https://cloud.google.com/learn/what-is-apache-spark>

<https://onetech.vn/blog/mot-so-artificial-intelligence-frameworks-va-thu-vien-can-biet-6514>

<https://aws.amazon.com/vi/compare/the-difference-between-artificial-intelligence-and-machine-learning/>

<https://vegafintech.vn/vi/apache-spark-la-gi-vai-tro-va-ung-dung-cua-apache-spark>

<https://meeyland.com/tin-tuc/apache-spark-la-gi-nhung-uu-diem-noi-bat-cua-apache-spark-378158745>

<https://aws.amazon.com/vi/nosql/>

<https://bkhost.vn/blog/nosql/>

<https://topdev.vn/blog/hadoop-la-gi/>

<https://www.toponseek.com/blogs/sql-la-gi/>