

1. Big data là gì?

Big Data là các tập dữ liệu có khối lượng lớn và phức tạp. Độ lớn đến mức các phần mềm xử lý dữ liệu truyền thống không có khả năng thu thập, quản lý và xử lý dữ liệu trong một khoảng thời gian hợp lý.

Những tập dữ liệu lớn này có thể bao gồm các dữ liệu có cấu trúc, không có cấu trúc và bán cấu trúc, mỗi tập có thể được khai thác để tìm hiểu insights.

Bao nhiêu dữ liệu để đủ gọi là ” big ” vẫn còn được tranh luận, nhưng nó có thể là các bội số của petabyte – và các dự án lớn nhất với phạm vi exabytes.

2. Đặc trưng của Big Data

Big data thường đặc trưng với ba V:

Volume: Khối lượng dữ liệu

Variety: Nhiều loại dữ liệu đa dạng

Velocity: Vận tốc mà dữ liệu cần phải được xử lý và phân tích

Dữ liệu tạo thành các kho dữ liệu lớn có thể đến từ các nguồn bao gồm các trang web, mạng xã hội, ứng dụng dành cho máy tính để bàn và ứng dụng trên thiết bị di động, các thí nghiệm khoa học, và các thiết bị cảm biến ngày càng tăng và các thiết bị khác trong internet (IoT).

Khái niệm big data đi kèm với các thành phần có liên quan cho phép các tổ chức đưa dữ liệu vào sử dụng thực tế và giải quyết một số vấn đề kinh doanh. Bao gồm:

Cơ sở hạ tầng IT cần thiết để hỗ trợ big data.

Các phân tích áp dụng với dữ liệu.

Công nghệ cần thiết cho các dự án big data các bộ kỹ năng liên quan.

Và các trường hợp thực tế có ý nghĩa đối với big data.

3. Big data và analytics

Điều thực sự mang lại giá trị từ các tổ chức dữ liệu lớn là phân tích dữ liệu. Nếu không có phân tích, nó chỉ là một tập dữ liệu với việc sử dụng hạn chế trong kinh doanh.

Bằng cách phân tích dữ liệu lớn, các công ty có thể có những lợi ích như tăng doanh thu, dịch vụ khách hàng được cải thiện, hiệu quả cao hơn và tăng khả năng cạnh tranh.

Phân tích dữ liệu liên quan đến việc kiểm tra bộ dữ liệu để thu thập thông tin chi tiết hoặc rút ra kết luận về những gì chúng chứa, chẳng hạn như các xu hướng và dự đoán về hoạt động trong tương lai.

Bằng cách phân tích dữ liệu, các tổ chức có thể đưa ra các quyết định kinh doanh tốt hơn như khi nào và ở đâu nên chạy chiến dịch tiếp thị hoặc giới thiệu một sản phẩm hoặc dịch vụ mới.

Sự phân tích có thể tham khảo các ứng dụng kinh doanh thông minh hay tiên tiến hơn. Phép phân tích dự đoán như ứng dụng được các tổ chức khoa học sử dụng.

Loại phân tích dữ liệu cao cấp nhất là data mining, nơi các nhà phân tích đánh giá các bộ dữ liệu lớn để xác định mối quan hệ, mô hình và xu hướng.



Phân tích dữ liệu có thể bao gồm phân tích dữ liệu thăm dò (để xác định các mẫu và mối quan hệ trong dữ liệu) và phân tích dữ liệu xác nhận (áp dụng các kĩ thuật thống kê để tìm ra giả thiết về một bộ dữ liệu có đúng hay không).

Một mảng khác là phân tích dữ liệu định lượng (hoặc phân tích dữ liệu số có các biến có thể so sánh theo thống kê) so với phân tích dữ liệu định tính (tập trung vào các dữ liệu không phải là dữ liệu cá nhân như video, hình ảnh và văn bản).

4. Cơ sở hạ tầng IT để hỗ trợ big data

Đối với khái niệm big data để làm việc, các tổ chức cần phải có cơ sở hạ tầng để thu thập và chứa dữ liệu, cung cấp quyền truy cập và đảm bảo thông tin trong khi lưu trữ và chuyển tiếp.

Ở cấp độ cao, bao gồm hệ thống lưu trữ và máy chủ được thiết kế cho big data, phần mềm quản lý và tích hợp dữ liệu, thông tin kinh doanh và phần mềm phân tích dữ liệu, và các ứng dụng big data.

Phần lớn các cơ sở hạ tầng này sẽ tập trung một chỗ, vì các công ty muốn tiếp tục tận dụng các khoản đầu tư vào trung tâm dữ liệu của mình. Nhưng ngày càng có nhiều tổ chức dựa vào các dịch vụ điện toán đám mây để xử lý nhiều yêu cầu big data của họ.

Thu thập dữ liệu đòi hỏi phải có nguồn. Rất nhiều trong số những ứng dụng sau đây, như các ứng dụng web, các kênh truyền thông xã hội, ứng dụng di động và lưu trữ email đã được cài sẵn.

Nhưng khi IoT trở nên phổ biến hơn, các công ty có thể cần phải triển khai cảm biến trên tất cả các loại thiết bị, phương tiện và sản phẩm để thu thập dữ liệu, cũng như các ứng dụng mới tạo ra dữ liệu người dùng. Phân tích dữ liệu theo định hướng IoT có các kỹ thuật và công cụ chuyên biệt của nó.

Để lưu trữ tất cả các dữ liệu đến, các tổ chức cần phải có đủ dung lượng lưu trữ tại chỗ. Các tùy chọn lưu trữ bao gồm kho dữ liệu truyền thống, data lakes và lưu trữ trên đám mây.

Các công cụ cơ sở hạ tầng bảo mật có thể bao gồm việc mã hóa dữ liệu, xác thực người dùng và các điều khiển truy cập khác, hệ thống giám sát, tường lửa, quản lý di động của doanh nghiệp và các sản phẩm khác để bảo vệ hệ thống và dữ liệu.

5. Các công nghệ đặc biệt dành cho Big data

Ngoài cơ sở hạ tầng IT được sử dụng cho dữ liệu nói chung, có một số công nghệ cụ thể dành cho big data mà cơ sở hạ tầng IT của bạn nên hỗ trợ.

Hệ sinh thái Hadoop

Hadoop là một trong những công nghệ liên quan chặt chẽ nhất với big data. Dự án Apache Hadoop phát triển phần mềm mã nguồn mở cho máy tính có khả năng mở rộng và phân tán.

Thư viện phần mềm Hadoop là một khuôn mẫu cho phép xử lý phân tán các bộ dữ liệu lớn trên các nhóm máy tính sử dụng các mô hình lập trình đơn giản. Nó được thiết kế để mở rộng từ một máy chủ duy nhất sang hàng ngàn máy khác, mỗi máy cung cấp tính toán và lưu trữ cục bộ.

Dự án bao gồm rất nhiều phần:

Hadoop Common, các tiện ích phổ biến hỗ trợ các phần Hadoop khác



Hadoop Distributed File System, cung cấp khả năng truy cập dữ liệu ứng dụng cao
Hadoop YARN, một khuôn mẫu cho kế hoạch làm việc và quản lý tài nguyên cụm
Hadoop MapReduce, một hệ thống dựa trên YARN để xử lý song song bộ dữ liệu lớn.

Apache Spark

Một phần của hệ sinh thái Hadoop, Apache Spark là một khuôn mẫu tính toán cụm nguồn mở được sử dụng làm công cụ xử lý big data trong Hadoop.

Spark đã trở thành một trong những khuôn mẫu xử lý big data quan trọng, và có thể được triển khai theo nhiều cách khác nhau. Nó cung cấp các phương thức hỗ trợ đối với Java, Scala, Python (đặc biệt là Anaconda Python distro), và ngôn ngữ lập trình R (R đặc biệt phù hợp với big data) và hỗ trợ SQL, streaming data, machine learning và xử lý đồ thị.

Data lakes

Data lakes là các kho lưu trữ chứa khối lượng dữ liệu thô rất lớn ở định dạng gốc của nó cho đến khi những người dùng doanh nghiệp cần dữ liệu.

Các yếu tố giúp tăng trưởng data lakes là những phong trào kỹ thuật số và sự phát triển của IoT. Các data lakes được thiết kế để giúp người dùng dễ dàng truy cập vào một lượng lớn dữ liệu khi có nhu cầu.

NoSQL Databases

Các cơ sở dữ liệu SQL thông thường được thiết kế cho các transaction đáng tin cậy và các truy vấn ngẫu nhiên.

Nhưng chúng có những hạn chế như giãn đồ cứng nhắc làm cho chúng không phù hợp với một số loại ứng dụng. Cơ sở dữ liệu NoSQL nêu ra những hạn chế, và lưu trữ và quản lý dữ liệu theo những cách cho phép tốc độ hoạt động cao và sự linh hoạt tuyệt vời.

Nhiều cơ sở dữ liệu đã được phát triển bởi các công ty để tìm cách tốt hơn để lưu trữ nội dung hoặc xử lý dữ liệu cho các trang web lớn. Không giống như các cơ sở dữ liệu SQL, nhiều cơ sở dữ liệu NoSQL có thể được mở rộng theo chiều ngang trên hàng trăm hoặc hàng ngàn máy chủ.

In-memory databases

Cơ sở dữ liệu trong bộ nhớ (IMDB) là một hệ thống quản lý cơ sở dữ liệu chủ yếu dựa vào bộ nhớ chính (Ram), thay vì HDD, để lưu trữ dữ liệu. Cơ sở dữ liệu trong bộ nhớ nhanh hơn các cơ sở dữ liệu được tối ưu hóa trong đĩa, một điểm quan trọng để sử dụng phân tích big data và tạo ra các kho dữ liệu và các siêu dữ liệu. Đọc thêm Redis là gì?

6. Các kỹ năng Big data

Big data và các nỗ lực phân tích big data yêu cầu kỹ năng cụ thể, dù là từ bên trong tổ chức hay thông qua các chuyên gia bên ngoài.

Nhiều kỹ năng có liên quan đến các thành phần công nghệ dữ liệu quan trọng như Hadoop, Spark, NoSQL, cơ sở dữ liệu trong bộ nhớ và phần mềm phân tích.

Các lĩnh vực khác cụ thể là về các nguyên tắc như khoa học dữ liệu, khai thác dữ liệu, phân tích thống kê và định lượng, hình dung dữ liệu, lập trình mục đích chung, và cấu



trúc dữ liệu và các thuật toán. Ngoài ra cũng cần có những người có kỹ năng quản lý tổng thể để quản lý tiến độ của các dự án big data.

Với độ phổ biến của các dự án phân tích dữ liệu và sự thiếu hụt nhân lực về các kỹ năng trên, việc tìm kiếm các chuyên gia có kinh nghiệm có thể là một trong những thách thức lớn nhất đối với các tổ chức.

7. Các trường hợp sử dụng Big data

Big data và phân tích có thể được áp dụng trong nhiều vấn đề kinh doanh và nhiều trường hợp sử dụng khác nhau. Sau đây là vài ví dụ:

Phân tích khách hàng. Các công ty có thể kiểm tra dữ liệu khách hàng để nâng cao trải nghiệm của khách hàng, cải thiện tỷ lệ chuyển đổi và giữ khách hàng tốt hơn.

Phân tích hoạt động. Nâng cao hiệu quả hoạt động và sử dụng tốt hơn tài sản của công ty là mục tiêu của nhiều công ty. Phân tích big data có thể giúp doanh nghiệp vận hành hiệu quả hơn và cải thiện hiệu suất.

Phòng chống gian lận. Phân tích dữ liệu có thể giúp các tổ chức xác định các hoạt động khả nghi, và các mẫu có thể chỉ ra hành vi gian lận và giúp giảm thiểu rủi ro.

Tối ưu hóa giá cả. Các công ty có thể sử dụng phân tích big data để tối ưu hóa giá đặt cho sản phẩm và dịch vụ, giúp tăng doanh thu.

Mời bạn đọc cùng tham khảo thêm tại mục [hỏi đáp thắc mắc](#) trong mục [tài liệu](#) nhé.

