1. Giới thiệu Big data
2. Các làn sóng quản lý dữ liệu

Mỗi làn sóng quản lý dữ liệu được sinh ra từ nhu cầu cần giải quyết một vấn đề quản lý dữ liệu. Từng làn sóng hay các giai đoạn đều là quan hệ nhân quả. Khi một giải pháp công nghệ mới xuất hiện trên thị trường, nó cần các phương pháp khai thác mới. Khi csdl quan hệ xuất hiện, nó cần một tập các công cụ cho phép nhà quản lý nghiên cứu mối quan hệ giữa các thành phần dữ liệu. Khi các công ty bắt đầu lưu trữ dữ liệu không cấu trúc, họ cần khả năng phân tích mới như các công cụ phân tích bằng ngôn ngữ tự nhiên để có được thông tin cần cho kinh doanh. Nếu bạn là lãnh đạo của một công ty tìm kiếm, bạn nhận ra là bạn cần truy cập vào một lượng dữ liệu khổng lồ để kiếm tiền. Để thu được giá trị của dữ liệu, cần phải có các công cụ và phương pháp mới.

Các làn sóng dữ liệu trong hơn 5 thập kỷ qua đã tạo ra chúng ta hôm nay: sự khởi đầu của kỷ nguyên big data. Do đó, để hiểu được big data, chúng ta phải hiểu được nền tảng của các làn sóng trước. Bạn cũng cần lưu ý là khi chúng tả chuyển từ làn sóng này sang làn sóng khác, chúng ta không bỏ đi các công cụ và công nghệ và kinh nghiệm mà chúng ta sử dụng để giải quyết các vấn đề khác nhau.

Làn sóng 1 – Sự hình thành của các cấu trúc dữ liệu quản lý

Khi điện toán được áp dụng vào thị trường thương mại trong cuối những năm 60, dữ liệu được lưu trong các file không có cấu trúc. Khi các công ty cần hiểu chi tiết hơn về khách hàng, họ phải áp dụng các phương pháp trực tiếp, gồm các mô hình lập trình chi tiết để tạo ra các giá trị. Cuối những năm 70, mọi thứ thay đối với sự ra đời của mô hình dữ liệu quan hệ và hệ thống quản trị csdl quan hệ (RDBMS), với sự giới thiệu về cấu trúc và phương pháp cải thiện hiệu suất hoạt động. Quan trọng nhất là mô hình quan hệ thêm vào mức trừu tượng (truy vấn cấu trúc SQL, xuất báo cáo, và các công cụ quản lý) để cho phép các lập trình viên đáp ứng được yêu cầu nghiệp vụ đang tăng trưởng nhằm rút giá trị từ dữ liệu một cách dễ hơn.

Mô hình quan hệ tạo ra một hệ sinh thái các công cụ từ một nhóm lớn các công ty phần mềm. Nó đáp ứng được nhu cầu tăng lên giúp các công ty quản lý quản lý dữ liệu tốt hơn và có thể so sánh các giao dịch ở nhiều nơi khác nhau. Hơn nữa, nó giúp các nhà quản lý kinh doanh cần quản lý nhà kho và so sánh với thông tin đơn hàng để ra quyết định tốt hơn. Nhưng một vấn đề phát sinh cần giải quyết là: Việc lưu trữ lượng lớn dữ liệu càng lúc càng tăng này rất tốn kém và tốc độ truy cập thấp. Tệ hơn là, có nhiều dữ liệu trùng lặp, và rất khó để đo lường giá trị kinh doanh thực sự.

Vào lúc này, có nhu cầu cấp thiết cần một tập các công nghệ mới hỗ trợ mô hình quan hệ. Mô hình quan hệ thực thể (ER) ra đời, với việc thêm trừu tượng để tăng tính khả dụng của dữ liệu. Trong mô hình này, mỗi mục (item) được định nghĩa độc lập với khả năng sử dụng nó. Do đó, những người điều hành có thể tạo ra các quan hệ mới giữa các nguồn dữ liệu mà không cần lập trình phức tạp. Tại thời điểm đó, đây là là một bước tiến lớn, và nó cho phép các nhà phát triển đẩy lên giới hạn của công nghệ và tạo ra các mô hình phức tạp hơn cần cho công nghệ phức tạp để kết hợp các thực thể với nhau. Thị trường cho csdl quan hệ bùng nổ và vẫn còn cho đến tận ngày nay. Điều này đặc biệt quan trọng đối với việc quản lý dữ liệu giao dịch của các dữ liệu cấu trúc.

Khi khối lượng dữ liệu các tổ chức cần quản lý vượt quá tầm kiểm soát, giải pháp đề ra là data warehouse. Data warehouse cho phép các tổ chức IT lựa chọn một tập con các thực thể dữ liệu lưu trữ giúp cho doanh nghiệp thu được thông tin hiểu biết dễ hơn. Data warehouse dùng để giúp các công ty xử lý khối lượng lớn dữ liệu cấu trúc đang tăng dần lên mà họ cần để có thể phân tích bằng cách giảm khối lượng dữ liệu thành một thứ nhỏ hơn, và tập trung về một lãnh vực kinh doanh cụ thể hơn. Nó đáp ứng nhu cầu tách biệt quá trình xử lý hỗ trợ quyết định và hỗ trợ quyết định – vì lý do hiệu suất. Hơn nữa, các warehouse thường lưu dữ liệu của nhiều năm trước đó để hiểu hơn về hiệu năng tổ chức, xác định xu hướng, và hỗ trợ trong việc tìm ra các hành vi khuôn mẫu. Nó cũng cung cấp nguồn thông tin tích hợp từ nhiều nguồn dữ liệu khác nhau khi dùng cho phân tích. Data warehouse được thương mại hóa vào những năm 90, và ngày nay, các hệ thống quản lý nội dung và data warehouse đều có thể lợi dụng khả năng tăng trưởng của phần cứng, công nghệ ảo hóa, và khả năng tạo ra các hệ thống phần cứng và phần mềm tích hợp, còn được biết với tên gọi thiết bị ứng dụng.

Đôi khi những data warehouse này quá phức tạp và quá lớn và nó không đáp ứng được tốc độ và tính linh hoạt doanh nghiệp cần. Giải pháp là tinh chế dữ liệu sau hơn thông qua data mart. Các data mart này tập trung vào các vấn đề nghiệp vụ cụ thể và được sắp xếp hợp lý hơn, và hỗ trợ các nhu cầu nghiệp vụ cần tốc độ truy vấn tốt hơn so với data ware lớn. Giống như bất kỳ làn sóng quản lý dữ liệu khác, warehouse đã phát triển các công nghệ hỗ trợ như là hệ thống tích hợp và thiết bị sử dụng dữ liệu.

Data warehouse và data mart giải quyết các vấn đề về cách quản lý thống nhất dữ liệu giao dịch lớn cho các công ty. Nhưng đối mặt với một lượng lớn dữ liệu không cấu trúc hoặc bán cấu trúc, data warehouse không thể tiến hóa để đáp ứng sự thay đổi của nhu cầu. Để khiến vấn đề phức tạp hơn, data warehouse được nhập vào theo khối từng kỳ, thường là hàng ngày hoặc hàng tuần. Điều này bình thường đối với việc lên kế hoặc, báo cáo tài chính, và các chiến dịch marketing truyền thống, nhưng quá chậm đối với kinh doanh thời gian thực và môi trường khách hàng.

Các công ty chuyển đổi các phương pháp quản lý dữ liệu truyền thống như thế nào để xử lý các nhân tố dữ liệu không cấu trúc này? Giải pháp không phát sinh ngay được. Khi các công ty bắt đầu lưu các dữ liệu không cấu trúc, các đại lý bắt đầu thêm vào khả năng gọi là BLOBs (đối tượng lớn nhị phân). Thực tế, thành phần dữ liệu không cấu trúc sẽ được lưu trong csdl quan hệ dưới dạng một khối dữ liệu liên tiếp. Đối tượng này được gắn nhãn (chẳng hạn, hóa đơn khách hàng) nhưng bạn không thể xem nội trung bên trong của đối tượng. Rất rõ ràng, điều này không giải quyết được nhu cầu khách hàng và nghiệp vụ.

Hệ thống quản lý csdl đối tượng ra đời. Csdl đối tượng lưu các BLOB thành tập các mẩu dữ liệu để có thể xem được nội dung bên trong. Không như BLOB, một đơn vị độc lập phụ thuộc vào csdl quan hệ truyền thống, csdl đối tượng cung cấp một phương pháp thống nhất để xử lý dữ liệu không cấu trúc. Csdl đối tượng bao gồm một ngôn ngữ lập trình và một cấu trúc dành cho các thành phần dữ liệu để thao tác các đối tượng dữ liệu khác nhau dễ hơn mà không cần lập trình hay kết hợp phức tạp. Csdl đối tượng tạo ra một bước tiến mới dẫn đến làn sóng quản lý dữ liệu lần thứ hai.

Làn sóng 2 – Quản lý nội dung và web

Đa số dữ liệu hiện nay trên thế giới là dữ liệu không cấu trúc. Một nghịch lý là các công ty lại tập trung đầu tư vào các hệ thống dữ liệu cấu trúc liên quan chắt chẽ với doanh thu: các hệ thống giao dịch theo dòng kinh doanh. Các hệ thống quản lý nội dung Doanh nghiệp được phát triển vào những năm 80 nhằm cung cấp khả năng quản lý dữ liệu không cấu trúc tốt hơn, chủ yếu là tài liệu. Vào những năm 90 với sự ra đời của web, các tổ chức muốn phát triển không chỉ với tài liệu mà có thể lưu trữ, quản lý nội dung web, ảnh, âm thanh và video.

Thị trường tiến hóa từ một tập các giải pháp không liên quan thành một mô hình thống nhất hơn, mang các thành phần này lại thành một platform kết hợp với việc quản lý quá trình kinh doanh, version control, nhận dạng thông tin, quản lý văn bản và kết hợp. Các hệ thống thế hệ mới này đưa vào khái niệm meta data (thông tin về tổ chức và các tính chất của thông tin được lưu trữ). Các giải pháp này vẫn giữ vai trò quan trong đối với các công ty cần quản lý tất cả các dữ liệu một cách logic. Nhưng đồng thời, một thế hệ mới các yêu cầu bắt đầu nảy sinh dẫn chúng ta tới làn sóng thứ ba. Những yều cầu này, phần lớn được thúc đẩy bởi sự chuyển đổi của các nhân tố bao gồm web, ảo hóa và điện toán đám mây. Trong làn sóng mới này, các tổ chức bắt đầu hiểu rằng họ cần quản lý các nguồn dữ liệu thế hệ mới với một khối lượng dữ liệu và sự biến đổi không thể lường được và cần được xử lý với một tốc độ chưa từng nghe thấy.

Làn sóng 3 – Quản lý big data

Big data là một sự phát triển mới hay nó là sự tiến hóa trong quản lý dữ liệu? Câu trả lời là cả hai. Tương tự như các làn sóng quản lý dữ liệu khác, big data được xây dựng trên sự tiến hóa của quản lý dữ liệu hơn 5 thập kỷ qua. Cái mới ở đây là lần đầu tiên, chi phí tính toán và lưu trữ đạt tớn điểm cực hạn. Tại sao điều này lại quan trọng? Chỉ vài năm trước đây, các tổ chức thường lưu trữ thông qua snapshot hoặc tập con các thông tin quan trọng vì chi phí lưu trữ và giới hạn xử lý ngăn họ lưu mọi thứ họ muốn phân tích.

Đề mục

Giới thiệu big data

1. Nền tảng của big data
   1. Sự tiến hóa của quản lý dữ liệu

Dù là cách mạng về dữ liệu hay là phát triển, các trạng thái mới hay các làn sóng quản lý dữ liệu đều được xây dựng dựa trên công nghệ trước đó. Quản lý dữ liệu phải bao gồm công nghệ phát triển từ phần cứng, lưu trữ, mạng, và các mô hình tính toán như là ảo hóa và điện toán đám mây.

Big data được định nghĩa là bất kỳ nguồn dữ liệu nào có ít nhất 3 tính chất sau:

* Dữ liệu có dung lượng cực lớn
* Dữ liệu có tốc độ xử lý cực cao
* Dữ liệu rất đa dạng, nhiều chủng loại

Big data cho phép các tổ chức thu thập, lưu trữ, quản lý, và thao tác lượng dữ liệu khổng lồ với tốc độ thích hợp, ở thời điểm thích hợp, để thu được tri thức cần thiết.

* 1. Các làn sóng quản lý data

**Làn song quản lý dữ liệu**: Được sinh ra từ sự cần thiết để giải quyết các loại vấn đề quản lý dữ liệu.

**Làn sóng 1: Cấu trúc dữ liệu có thể quản lý**

Cuối những năm 60, dữ liệu được lưu trữ dưới dạng file đơn giản, không có cấu trúc -> khó có thể tìm hiểu, tổng hợp để thu được tri thức. => Phát mình mô hình dữ liệu quan hệ (cuối những năm 70) và hệ quản trị csdl quan hệ => trích rút giá trị của dữ liệu sử dụng SQL, công cụ sinh báo cáo và quản lý dữ liệu.

* Phát triển mô hình ER (Entity-Relationship): tăng tính khả dụng của dữ liệu, tránh trùng lặp, tăng tốc độ xử lý dựa vào join entities với nhau. Trong mô hình này, mỗi item được định nghĩa độc lập với ứng dụng của nó.

Khi kích thước data vượt quá tầm kiểm soát, phát triển data warehouse: cho phép lựa chọn 1 tập con dữ liệu để xử lý. Xử lý lượng lớn dữ liệu bằng cách giảm dung lượng data cần phân tích thành 1 thứ nhỏ hơn và tập trung vào 1 lĩnh vực cụ thể. (Phát triển vào những năm 90)

* Data mart: phát triển từ data warehouse, tập trung vào các vấn đề liên quan, có tốc độ truy cập nhanh hơn so với data warehouse.

Để mở rộng việc xử lý dữ liệu không cấu trúc, ra đời hệ quản trị csdl đối tượng (ODBMS), lưu trữ BLOB (Binary Large Object) thành tập các mẩu thông tin cho phép người dùng xem được nội dung trong đó.

**Làn sóng 2: Web và quản lý nội dung**

Vào những năm 90, phát triển web, lưu trữ, quản lý nội dung web, hình ảnh, âm thanh, video => tạo ra meta data (thông tin về tổ chức và các thuộc tính của thông tin được lưu).

**Làn sóng 3: Quản lý big data**

Ảo hóa dữ liệu để có thể lưu trữ hiệu quả, tận dụng lưu trữ đám mây, tiết kiệm hơn. Công nghệ: MapReduce, Big Table và Hadoop.

* 1. Định nghĩa big data

Big data không phải là một công nghệ đơn lẻ mà là tập hợp của các công nghệ cũ và mới nhằm thu được tri thức.

Big data có các tính chất sau: V4

* Dung lượng (Volumn)
* Tốc độ xử lý (Velocity)
* Đa dạng (Variety)
* Xác thực (Veracity)
  1. Xây dựng kiến trúc quản lý big data thành công

Yêu cầu chức năng: Thu thập, tổ chức, tích hợp, phân tích, hành động (quyết định).

Đầu tiên, dữ liệu phải được thu thập, tổ chức và tích hợp.

Sau khi hoàn thành giai đoạn này, dữ liệu được phân tích dựa trên vấn đề đưa ra, từ đó đưa ra quyết định.

**Nền tảng xây dựng kiến trúc**: sức mạnh và tốc độ tính toán.

A picture containing text

Description automatically generated

* Interface và feeds: các service API tồn tại giữa các tầng giúp big data giao tiếp, thu thập thông tin
* Hạ tầng vật chất dự phòng: dựa trên mô hình điện toán phân tán (địa điểm lưu trữ khác nhau, thông qua mạng để giao tiếp).
* Hạ tầng bảo mật
* Nguồn dữ liệu hoạt động
  + Hệ thống các bản ghi theo dõi các dữ liệu quan trọng cần cho hoạt động thời gian thực hoặc gần thực.
  + Liên tục cập nhật theo giao dịch xảy ra trong từng đơn vị và từ web.
  + Gồm dữ liệu cấu trúc và không cấu trúc
  + Có thể mở rộng để hỗ trợ hàng ngàn user trên một nền tảng thống nhất.
* Hiệu suất: mô hình điện toán phân tán đẩy nhanh tốc độ tính toán.
  + Các công cụ và dịch vụ tổ chức dữ liệu
  + MapReduce (Google: các hàm thực thi map và reduce trên nền tảng phân tán đám mây), Big Table (hệ thống lưu trữ phân tán của Google) và Hadoop (framework phần mềm Apache dựa trên MapReduce và Big Table, hỗ trợ Yahoo!)
  1. Con đường big data

1. Các kiểu big data
   1. Định nghĩa dữ liệu cấu trúc
   2. Định nghĩa dữ liệu không cấu trúc
   3. Yêu cầu thời gian thực và không phải thời gian thực
   4. Tổng hợp big data