**TRƯỜNG ĐẠI HỌC MỎ ĐỊA CHẤT**

**Khoa Công Nghệ Thông Tin**



**Bài kiểm tra môn : Kho Dữ Liệu**

**Đề tài : Tìm hiểu về đặc điểm , lợi ích của kho dữ liệu**

Họ và tên : Trần Minh Tiến

Mã SV :1821050398

Mã MH : 7080231 – Nhóm 100

**1. Định nghĩa Kho dữ liệu**

**Kho dữ liệu** ban đầu được định nghĩa là một tập dữ liệu được dùng để hỗ trợ quy trình ra quyết định. Theo quá trình phát triển, **Kho dữ liệu** được định nghĩa là một **môi trường thông tin** (information environment) (từ môi trường thông tin thể hiện nó không phải là một sản phẩm (product)), có các chức năng:

**Cung cấp một góc nhìn toàn diện về doanh nghiệp:**

Cho dù doanh nghiệp có nhiều mảng kinh doanh được quản lý bởi các hệ thống tác nghiệp khác nhau, nhưng **Kho dữ liệu** là nơi đã tập hợp (tích hợp) được mọi thông tin về các mảng nghiệp vụ khác nhau để cung cấp một góc nhìn toàn diện. Việc tích hợp này còn cung cấp khả năng đánh giá chéo các mảng nghiệp vụ khác nhau để đánh giá sự tương quan giữa chúng.

**Cung cấp đầy đủ thông tin hiện tại và lịch sử của doanh nghiệp, và sẵn sàng cho việc khai thác, sử dụng cho việc hỗ trợ ra quyết định chiến lược.**

Từ thông tin ở đây thể hiện dữ liệu ở trong kho không chỉ đơn thuần là dữ liệu thô lấy từ các hệ thống tác nghiệp mà nó đã được tổng hợp, tính toán thành các độ đo có ý nghĩa phân tích.

**Có khả năng cung cấp dữ liệu chi tiết theo nhu cầu mà không phải truy xuất các hệ thống tác nghiệp:**

Điều này thể hiện trong một số trường hợp có thể nhu cầu phân tích dữ liệu ở mức giao dịch, thì nó cũng sẽ được lưu sẵn ở Kho dữ liệu.

**Đảm bảo thông tin trong Kho dữ liệu có tính nhất quán:**

Ví dụ doanh nghiệp có nhiều nghiệp vụ được quản lý với các hệ thống tác nghiệp khác nhau, nhưng cần đảm bảo ánh xạ được một khách hàng xuất hiện ở nhiều hệ thống về 1 người duy nhất. Điều này áp dụng cho các đối tượng dữ liệu khác. Ngoài ra một trường dữ liệu có thể được biểu diễn bằng nhiều cách khác nhau, ví dụ Tên khách hàng có hệ thống tách họ và tên thành 2 trường khác nhau, có hệ thống chỉ lưu trong 1 trường. Khi tổng hợp về Kho dữ liệu thì sẽ được chuyển về 1 dạng thống nhất.

**Là nguồn thông tin chiến lược mềm dẻo và có tính tương tác:**

Chữ mềm dẻo và tương tác ở đây thể hiện người dùng có thể lấy các thông tin khác nhau của cùng một đối tượng. Từ tương tác ở đây thể hiện có thể thực hiện được nhiều thao tác với các đối tượng dữ liệu thay vì trả lại một danh sách tĩnh. Ví dụ có thể lấy dữ liệu tổng hợp theo ngày, theo tuần, theo tháng của cùng 1 đối tượng dữ liệu. Tính tương tác thể hiện người dùng có thể áp dụng các thao tác phân tích dữ liệu, một ví dụ là nền tảng phân tích dữ liệu SQL Server Analysis Services (SSAS) cho phép người dùng có thể tương tác để phân tích dữ liệu.

**2. Đặc điểm của Kho dữ liệu**

Dữ liệu được lưu trong **Kho dữ liệu** không được tạo ra trực tiếp từ người dùng mà được lấy từ các nguồn dữ liệu sẵn có và mục đích là phục vụ tạo ra các báo cáo quản trị do đó nó có các tính chất sau:

**Hướng chủ đề (subject-oriented):**

Mục đích của **Kho dữ liệu** là phục vụ các yêu cầu phân tích, hoặc khai phá cụ thể được gọi là chủ đề. Ví dụ với chủ đề phân tích nhân sự thì có thể bao gồm các độ đo về doanh thu của từng người, số ngày nghỉ trong tháng, số dự án tham gia trong tháng, theo các chiều phân tích: thời gian, chi nhánh, sản phẩm, …

Một sự so sánh dễ hiểu, giống như chẩn đoán một bệnh ví dụ bệnh liên quan đến tim, thì bác sỹ cần quan tâm không chỉ một mà một vài chỉ số như các chỉ số liên quan đến máu, chỉ số về huyết áp, nhịp tim, điện tâm đồ. Ngoài ra còn cần theo dõi theo thời gian (có thể là hàng ngày) để xem xét sự thay đổi mà có phương pháp điều trị kịp thời. Trong trường hợp này thời gian được gọi là chiều phân tích. Để chẩn đoán được chính xác thì cần đầy đủ các thông tin về các chỉ số trên, và cũng không cần các chỉ số khác lẫn vào làm nhiễu quá trình chẩn đoán và cũng không cần thiết. Việc tổ chức dữ liệu theo chủ đề này sẽ dẫn đến nhu cầu tổ chức lưu trữ dữ liệu khác với các cơ sở dữ liệu tác nghiệp.

**Được tích hợp (integrated):**

Tại một bệnh viện, các phòng khác nhau sẽ thực hiện các xét nghiệm khác nhau, do đó để có được đầy đủ thông tin phục vụ chẩn đoán thì cần thu thập được kết quả từ nhiêu nguồn. Điều này hoàn toàn tương tự như tại doanh nghiệp, dữ liệu cần để phân tích có thể nằm rải rác ở nhiều hệ thống tác nghiệp khác nhau, và vì vậy cần tích hợp lại. Quá trình tích hợp này sẽ được thực hiện trong quá trình ETL như đã trình bày ở bài trước. Việc tổng hợp dữ liệu từ nhiều nguồn vào một kho dữ liệu cho phép chúng ta có thể xem đồng thời nhiều nhóm chỉ tiêu khác nhau (từ nhiều hệ thống nghiệp vụ khác nhau), ví dụ ta có thể xem chỉ tiêu doanh thu ở nhiều mảng nghiệp vụ khác nhau để có thể so sánh được sự tương quan giữa các mảng nghiệp vụ này. Điều này cũng giống trong chuẩn đoán bệnh ta có thể cần nhiều xét nghiệm (thử máu, thử nước tiếu, siêu âm, …) và kiểm tra khác nhau để có thể đưa ra kết luận chính xác.

**Có gán nhãn thời gian (time variant):**

Như đã đề cập, với các chỉ số thay đổi liên tục (như huyết áp, nhịp tim) việc chẩn đoán bệnh sẽ cần dữ liệu của các hôm trước để so sánh phục vụ quá trình điều trị. Do đó hàng ngày cần phải lưu lại giá trị của các chỉ số này. Hay nói cách khác các chỉ số này khi lưu sẽ được gán 1 nhãn thời gian tương ứng. Tương tự như vậy, dữ liệu lịch sử có tầm quan trọng đặc biệt trong phân tích dữ liệu, cùng một độ đo sẽ có nhiều giá trị khác nhau trong lịch sử có thể dùng để so sánh với nhau để biết được sự thay đổi là tốt hay xấu.

Ví dụ, độ đo doanh thu của một mặt hàng của tháng hiện tại, nếu đem so sánh với doanh thu của mặt hàng đó trong tháng trước, tháng này năm trước thì sẽ có nhiều thông tin hơn để đánh giá doanh thu của mặt hàng đó là tốt hay không, trên cơ sở đó sẽ có các quyết định phù hợp. Ngoài ra, dữ liệu lịch sử còn cho phép dự báo được tương lai khi ứng dụng khai phá dữ liệu.

**Bất biến (non-volatile):**

Khác với các **cơ sở dữ liệu** (CSDL) giao dịch, nơi thông tin của một đối tượng có thể được cập nhật thay đổi hàng ngày, dữ liệu trong Kho dữ liệu có chức năng báo cáo lại các chỉ số về hoạt động kinh doanh thực tế đã xảy ra. Do đó dữ liệu trong **Kho dữ liệu** không thể cập nhật, thay đổi vì nó sẽ không phản ánh đúng thực tế. Do đó với kho dữ liệu chỉ có 2 thao tác chính là tải dữ liệu vào kho và truy cập (đọc) dữ liệu từ kho.

Trường hợp sau khi tổng hợp dữ liệu mà dữ liệu ở trong nguồn bị thay đổi, khi đó một giải pháp xử lý là thực hiện lại quá trình ETL để tải lại dữ liệu từ nguồn vào Kho dữ liệu chứ không cho phép sửa đổi dữ liệu ở trong Kho. Một ý nghĩa khác của tính chất này là dữ liệu lịch sử vẫn được bảo tồn, vẫn có ý nghĩa chứ không như dữ liệu tác nghiệp ở các cơ sở dữ liệu, dữ liệu cũ (năm trước, hoặc tháng trước) không có ý nghĩa phục vụ hoạt động hằng ngày.

Như vậy, nếu như CSDL tác nghiệp được ví như cái tủ sách cá nhân, nơi người ta thường xuyên tra cứu, cập nhật, hiệu đính, ghi chú vào lề, thêm mới hoặc chuyển sách đi. Thì Kho dữ liệu lại được so sánh với thư viện quốc gia, nơi các tài liệu kinh điển được đưa đến liên tục để lưu trữ và tham khảo, không ai sửa chữa hoặc chuyển chúng qua chỗ nào khác cả.