1. Một số đặc điểm của Data Warehouse
2. Hướng chủ thể (Subject-Oriented)

* Kho dữ liệu được định hướng theo chủ thể vì nó cung cấp thông tin về một chủ thể thay vì các hoạt động đang diễn ra của công ty. Những chủ thể này có thể là bán hàng, tiếp thị, phân phối, v.v.
* Một kho dữ liệu không bao giờ tập trung vào các hoạt động đang diễn ra. Thay vào đó, nó tập trung vào mô hình hóa và phân tích dữ liệu để **ra quyết định**. Nó cũng cung cấp một cái nhìn đơn giản và ngắn gọn về chủ thể cụ thể bằng cách loại trừ dữ liệu không hữu ích để hỗ trợ quá trình quyết định.

1. Tích hợp (Intergrated)

* Trong Data Warehouse, tích hợp có nghĩa là thiết lập một đơn vị đo lường chung cho tất cả các dữ liệu tương tự từ cơ sở dữ liệu khác nhau. Dữ liệu cũng cần được lưu trữ trong Datawarehouse theo cách phổ biến và được mọi người chấp nhận.
* Kho dữ liệu được phát triển bằng cách tích hợp dữ liệu từ các nguồn khác nhau như máy tính lớn, cơ sở dữ liệu quan hệ, tệp phẳng, v.v. Hơn nữa, nó phải giữ các quy ước đặt tên, định dạng và mã hóa nhất quán.
* Sự tích hợp này giúp phân tích dữ liệu một cách hiệu quả. Phải đảm bảo tính nhất quán trong các quy ước đặt tên, các thước đo thuộc tính, cấu trúc mã hóa, v.v.

1. Biến thể thời gian (Time-Variant)

* Khoảng thời gian cho kho dữ liệu khá rộng so với các hệ thống vận hành. Dữ liệu được thu thập trong kho dữ liệu được ghi nhận với một giai đoạn cụ thể và cung cấp thông tin theo quan điểm lịch sử. Nó chứa một yếu tố thời gian, một cách rõ ràng hoặc hàm ý.
* Một trong những nơi mà phương sai thời gian hiển thị dữ liệu Datawarehouse nằm trong cấu trúc của khóa bản ghi. Mọi khóa chính chứa trong DW phải có một phần tử thời gian ngầm định hoặc rõ ràng. Như ngày, tháng trong tuần, v.v.
* Một khía cạnh khác của phương sai thời gian là khi dữ liệu được đưa vào kho, dữ liệu đó không thể được cập nhật hoặc thay đổi.

1. Không bay hơi (Non-volatile)

* Kho dữ liệu không bay hơi có nghĩa là dữ liệu trước đó không bị xóa khi dữ liệu mới được nhập vào.
* Dữ liệu ở chế độ chỉ đọc và được làm mới định kỳ. Điều này cũng giúp phân tích dữ liệu lịch sử để hiểu điều gì đã xảy ra khi nào. Nó không yêu cầu quy trình giao dịch, cơ chế phục hồi và kiểm soát đồng thời.
* Các hoạt động như delete, update, insert được thực hiện trong môi trường ứng dụng hoạt động bị bỏ qua trong môi trường Kho dữ liệu. Chỉ có hai loại hoạt động dữ liệu được thực hiện trong Kho dữ liệu là:
* Data loading
* Data access
* Sự khác biệt giữa ứng dụng và Data Warehouse:

|  |  |
| --- | --- |
| Ứng dụng | Data Warehouse |
| Chương trình phải được lập trình để đảm bảo rằng các quá trình update dữ liệu duy trì tính toàn vẹn của sản phẩm cuối cùng. | Loại vấn đề này không xảy ra vì cập nhật dữ liệu không được thực hiện. |
| Dữ liệu được đặt ở dạng chuẩn hóa để đảm bảo dư thừa tối thiểu | Dữ liệu không được lưu ở dạng chuẩn hóa |
| Cần có các kĩ thuật hỗ trợ giải quyết vấn đề về transaction; khôi phục, rollback dữ liệu (vì có những deadlock khá phức tạp) | Các kĩ thuật xử lý các vấn đề tương tự khá đơn giản |

1. Một số tool cho data warehouse và ứng dụng của chúng
2. Sourcing, Acquisition, Clean-up and Transformation Tools (ETL)

* Các công cụ tìm nguồn, chuyển đổi và di chuyển dữ liệu được sử dụng để thực hiện tất cả các chuyển đổi, tóm tắt và các thay đổi cần thiết để chuyển đổi dữ liệu thành một định dạng thống nhất trong datawarehouse. Chúng còn được gọi là Extract, Transform and Load tool (ETL).
* Chức năng của chúng bao gồm:
* Ẩn danh dữ liệu theo quy định
* Loại bỏ dữ liệu không mong muốn trong cơ sở dữ liệu hoạt động khi load vào Data warehouse
* Tìm kiếm và thay thế các tên và định nghĩa phổ biến cho dữ liệu đến từ các nguồn khác nhau.
* Tính toán tóm tắt và dữ liệu dẫn xuất
* Trong trường hợp thiếu dữ liệu, điền vào các giá trị mặc định.
* Khử trùng lặp từ nhiều nguồn dữ liệu.

1. Tool truy vấn

* Được chia thành 4 loại:
* Tool truy vấn và báo cáo:
* Tool báo cáo: tạo các báo cáo về hoạt động, phân tích
* Tool quản lý truy vấn: Hỗ trợ người dùng xử lý lỗi trong database, SQL và cấu trúc database bằng cách chèn meta-layer giữa người dùng và database

1. Tool phát triển ứng dụng

* Đôi khi các công cụ đồ họa và phân tích tích hợp sẵn không thỏa mãn nhu cầu phân tích của một tổ chức. Trong những trường hợp như vậy, các báo cáo tùy chỉnh được phát triển bằng cách sử dụng các công cụ phát triển Ứng dụng.

1. Tool khai thác dữ liệu

* Khai thác dữ liệu là một quá trình khám phá mối tương quan, mô hình và xu hướng mới có ý nghĩa bằng cách khai thác dữ liệu số lượng lớn. Các công cụ khai thác dữ liệu được sử dụng để làm cho quá trình này tự động.

1. Tool OLAP

* Các công cụ này dựa trên các khái niệm về cơ sở dữ liệu đa chiều. Nó cho phép người dùng phân tích dữ liệu bằng cách sử dụng các quan điểm đa chiều phức tạp và phức tạp.