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. CData Sync (<https://www.cdata.com/sync>)

* Hợp nhất dữ liệu từ bất kỳ ứng dụng hoặc nguồn dữ liệu nào vào Cơ sở dữ liệu hoặc Kho dữ liệu mà bạn lựa chọn.
* Kết nối dữ liệu hỗ trợ doanh nghiệp với BI, Analytics và Machine Learning.
* Tự động sao chép dữ liệu gia tăng thông minh
* Chuyển đổi dữ liệu ETL / ELT có thể custom

1. BiG EVAL (<https://go.bigeval.com/automated-dwh-testing>)

* Đo lường chất lượng dữ liệu và hỗ trợ giải quyết vấn đề
* Autopilot testing để phát triển agile, trích xuất meta-data từ cơ sở dữ liệu hoặc kho meta-data
* Tóm tắt cho bất kỳ loại dữ liệu nào (RDBMS, API, Flatfiles, ứng dụng doanh nghiệp…).

1. Xplenty (<https://www.xplenty.com>)

* Tập hợp và chuẩn bị dữ liệu cho BI
* Truyền và chuyển đổi dữ liệu giữa các cơ sở dữ liệu nội bộ hoặc kho dữ liệu
* Gửi dữ liệu bổ sung của bên thứ ba tới Heroku Postgres (và sau đó tới Salesforce qua Heroku Connect) hoặc trực tiếp tới Salesforce.
* Trình kết nối API Rest để lấy dữ liệu từ bất kỳ API Rest nào

1. Oracle (<https://www.oracle.com/autonomous-database/autonomous-data-warehouse>)

* Phân phối dữ liệu theo cùng một cách trên các đĩa để mang lại hiệu suất đồng nhất
* Hoạt động với cụm ứng dụng hoặc ứng dụng riêng lẻ
* Test ứng dụng trong môi trường thực
* Kết nối tốc độ cao để chuyển data khối lượng lớn
* Hỗ trợ trực quan hóa
* Cho phép kết nối database, table, view từ xa

1. Informatica (<https://www.informatica.com>)

* Hệ thống ghi lỗi tập trung tạo điều kiện cho các lỗi ghi nhật ký và từ chối dữ liệu vào các bảng quan hệ
* Khả năng mở rộng quy mô tích hợp dữ liệu

1. Google BigQuery (<https://cloud.google.com/bigquery>)

* Cung cấp tính năng Nhập dữ liệu linh hoạt
* Đọc và ghi dữ liệu thông qua Cloud Dataflow, Hadoop, and Spark.
* Dịch vụ chuyển dữ liệu tự động
* Toàn quyền kiểm soát quyền truy cập vào dữ liệu được lưu trữ
* BigQuery cung cấp cơ chế kiểm soát chi phí