Kho dữ liệu

Mục tiêu:

- Hiểu được khái niệm và mục đích của kho dữ liệu
- So sánh OLAP với OLTP
- So sánh OLAP với Data mining
- Một số ví dụ về kho dữ liệu

1

Đặt vấn đề

- CSDL có vai trò quan trọng trong các hệ thống thông tin quản lý:
  - Xử lý nghiệp vụ
  - Xử lý giao dịch
- Một CSDL được thiết kế cho nhu cầu lưu trữ và xử lý thông tin của tổ chức, doanh nghiệp, gồm:
  - Một tập hợp dữ liệu có cấu trúc
  - Một bản thông tin mô tả dữ liệu có cấu trúc đó (metadata)

2

Đặt vấn đề

- Từ những năm 80 của thế kỷ XX doanh nghiệp đã nhận ra sự cần thiết sử dụng dữ liệu trong quá khứ để phân tích nhằm hỗ trợ ra quyết định, tạo lợi thế cạnh tranh
- Thuật ngữ "kho dữ liệu" được sử dụng vào năm 1988 trong bài báo kỹ thuật của IBM: *An architecture for a business and information system*

Kho dữ liệu

- Kho dữ liệu là dữ liệu lịch sử được trích ra định kỳ từ các nguồn dữ liệu khác nhau (chủ yếu từ CSDL tác nghiệp) và được chuyển đổi tới một CSDL có thiết kế đặc biệt để xử lý thông tin, xử lý phân tích
- Theo W. H. Inmon thì kho dữ liệu là tập hợp dữ liệu hướng chủ thể, tích hợp, thay đổi theo thời gian và có tính ổn định, với mục đích hỗ trợ ra quyết định

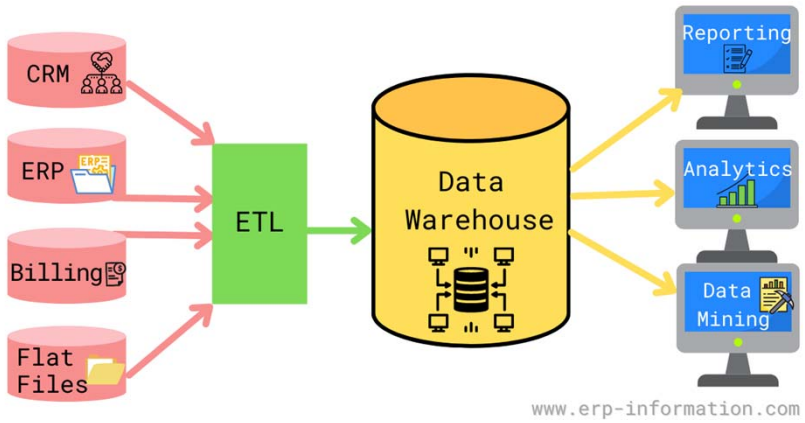
Kho dữ liệu

- Hướng chủ đề
  - Được tổ chức xung quanh chủ đề chính của doanh nghiệp, chẳng hạn như Bán hàng, Khách hàng, Nhà cung cấp, Sản phẩm…
- Tích hợp, thay đổi theo thời gian, ổn định
  - Thu thập dữ liệu ổn định
  - Làm sạch
  - Tải dữ liệu
  - Phát sinh

Kho dữ liệu

- Có nhiều công ty lớn đầu tư vào công nghệ kho dữ liệu và công cụ liên quan:
  - Microsoft
  - Oracle
  - IBM Infosphere
  - Amazon Redshift
  - …

Kiến trúc của kho dữ liệu



7

Mục đích của kho dữ liệu

- Xử lý thông tin - Dữ liệu được xử lý bằng các truy vấn, phân tích thống kê cơ bản, báo cáo bằng cách sử dụng bảng biểu, biểu đồ, đồ thị
- Xử lý phân tích - Dữ liệu có thể được phân tích bằng các thao tác OLAP cơ bản, như roll up, drill down, drill up, pivot…
- Khai phá dữ liệu - Tìm các mô hình và các mối liên kết ẩn, thực hiện phân loại, gom nhóm và dự đoán.

8

OLAP là gì?

- Kho dữ liệu cho phép người dùng ở mức quản lý, ra quyết định thực hiện các phép phân tích tương tác với data bằng hệ thống xử lý phân tích trực tuyến (online analytical processing – OLAP).
- Ngoài ra kho dữ liệu cũng được dùng cho báo cáo, data mining và phân tích thống kê.
- Database và kho dữ liệu, do đó chỉ khác nhau về mặt khái niệm, một cơ sở dữ liệu nếu dùng riêng cho các mục đích trên cũng được coi là kho dữ liệu.

9

OLAP là gì?

- Xử lý phân tích trực tuyến (On-Line Analytical Processing – OLAP)
  - Xử lý phân tích dữ liệu (thời gian thực) để xác định xu hướng, hỗ trợ quyết định chiến lược kinh doanh
  - Xây dựng báo cáo về tài chính, bán hàng, tiếp thị, quản trị, dự báo…
- So sánh OLAP với OLTP (On-Line Transaction Processing)

10

OLTP và OLAP

OLTP	OLAP
▪ Phần lớn là cập nhật	▪ Phần lớn là đọc dữ liệu
▪ Truy vấn đơn giản	▪ Truy vấn dài, phức tạp
▪ Kích thước dữ liệu MB-GB	▪ Kích thước dữ liệu GB-TB
▪ Dữ liệu thô	▪ Dữ liệu tổng hợp
▪ Người dùng đa dạng, nhiều, truy cập đồng thời	▪ Người dùng: lãnh đạo, người quản lý, chuyên gia
▪ Lượng giao dịch cao	▪ Lượng giao dịch trung bình, thấp
▪ Định hướng xử lý	▪ Định hướng chủ đề

OLAP và Data mining

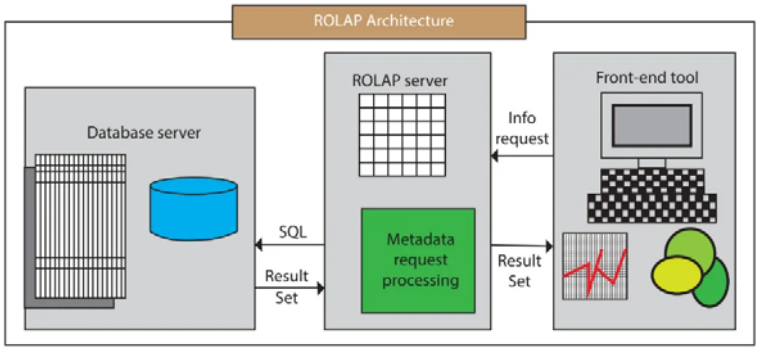
OLAP	Data mining
▪ OLAP là một công nghệ truy cập dữ liệu có cấu trúc đa chiều	▪ Là một lĩnh vực trích rút tri thức (xu hướng, mô hình) từ lượng dữ liệu khổng lồ
▪ Hướng truy vấn	▪ Hướng khai phá
▪ Sử dụng dữ liệu quá khứ để phân tích	▪ Sử dụng dữ liệu quá khứ cho dự báo tương lai
▪ Số chiều dữ liệu ít, vừa phải	▪ Có số chiều dữ liệu lớn, rất lớn
▪ Cách tiếp cận Top-down	▪ Cách tiếp cận Bottom-Up

OLAP

- Dữ liệu trong Kho dữ liệu được tổ chức dưới dạng các khối dữ liệu đa chiều (Multi Dimensional Cube) và OLAP dùng để phân tích trên các khối dữ liệu đó.
- OLAP cho phép người dùng phân tích dữ liệu qua việc cắt lát (slice) dữ liệu theo nhiều khía cạnh khác nhau, khoan xuống (drill down) mức chi tiết hơn hay là cuộn lên (roll up) mức tổng hợp hơn của dữ liệu Ứng dụng của OLAP

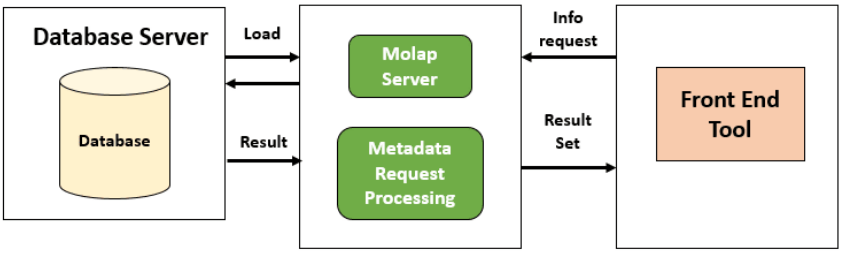
ROLAP và MOLAP

- ROLAP - Relational On-Line Analytical Processing



ROLAP và MOLAP

- MOLAP - Multi-Dimensional On-Line Analytical Processing



15

Các ví dụ về kho dữ liệu

- Kho dữ liệu về phân tích bán hàng
  - Giúp doanh nghiệp phân tích hiệu suất bán hàng, hành vi của khách hàng và xu hướng thị trường.
  - Đưa ra quyết định sáng suốt về giá cả, khuyến mãi và quản lý hàng tồn kho.
- Kho dữ liệu phân tích tài chính
  - Cung cấp cái nhìn tổng quan về hiệu quả tài chính của họ, bao gồm doanh thu, chi phí và dòng tiền.
  - Dữ liệu này có thể được sử dụng để tạo báo cáo tài chính và đưa ra quyết định sáng suốt về đầu tư và chi phí

16

Các ví dụ về kho dữ liệu

- Kho dữ liệu quản lý chuỗi cung ứng
  - Giúp doanh nghiệp quản lý chuỗi cung ứng của mình hiệu quả hơn về: tồn kho, lưu kho, thời gian giao hàng
  - Dữ liệu này có thể sử dụng để tối ưu hóa việc quản lý hàng tồn kho, giảm chi phí và cải thiện sự hài lòng của khách hàng.
- Kho dữ liệu về khách hàng
  - xây dựng báo cáo về khách hàng, loại khách hàng, xu hướng…
  - Hiểu biết khách hàng, cải thiện dịch vụ