Bộ Lao Động Thương Binh Xã Hội
TỔNG CỤC DẠY NGHỀ
Dự án giáo dục kĩ thuật và dạy nghề

GIÁO TRÌNH

Môn Học : QUẢN LÝ KHO DỮ LIỆU
Mã số : ITPRG3_11

NGHỀ: LẬP TRÌNH MÁY TÍNH

TRÌNH ĐỘ: BẬC CAO

Đà lạt -2007

Địa chỉ liên hệ:
Dự án giáo dục kỹ thuật và nghề nghiệp
Tiểu Ban Phát triển Chương trình Học liệu
……
……
.....

LỜI TỰA

Đây là tài liệu được xây dựng theo chương trình của dự án giáo dục kỹ thuật và dạy nghề, để có đươc giáo trình này dự án đã tiến hành theo hai giai đoạn.

Giai đoạn 1 : Xây dựng chương trình theo phương pháp DACUM, kết quả của gian đoạn này là bộ khung chương trình gồm 230 trang cấp độ 2 và 170 trang cấp độ 3.

Giai đoạn 2 : 29 giáo trình và 29 tài liệu hướng dẫn giáo viên cho nghề lập trình máy tính 2 cấp độ.

Để có được khung chương trình chúng tôi đã mời các giáo viên, các chuyên gia đang làm việc trong lĩnh vực công nghệ thông tin cùng xây dựng chương trình.

Trong giai đoạn viết giáo trình chúng tôi cũng đã có những sự điều chỉnh để giáo trình có tính thiết thực và phù hợp hơn với sự phát triển của lĩnh vực công nghệ thông tin.

Kho dữ liệu là một tập hợp các CSDL rất lớn tới hàng trăm GB hay thậm chí hàng Tera byte dữ liệu từ nhiều phân hệ của hệ thống, lưu trữ và phân tích phục vụ cho việc cung cấp các dịch vụ thông tin liên quan tới nghiệp vụ một tổ chức, cơ quan hay xí nghiệp. Thông thường dữ liệu phát sinh từ các hoạt động hàng ngày và được thu thập xử lý để phục vụ công việc nghiệp vụ cụ thể của một tổ chức vì vậy thường được gọi là dữ liệu tác nghiệp và hoạt động thu thập xử lý loại dữ liệu này được gọi là xử lý giao dịch trực tuyến OLTP. Kho dữ liệu trái lại phục vụ cho việc phân tích với kết quả mang tính thông tin cao. Các hệ thống thông tin thu thập xử lý dữ liệu loại này còn gọi là hệ xử lý phân tích trực tuyến OLAP..

Kho dữ liệu được xây dựng để tiện lợi cho việc truy cập theo nhiều nguồn, nhiều kiểu dữ liệu khác nhau sao cho có thể kết hợp được cả những ứng dụng của các công nghệ hiện đại và kế thừa được từ những hệ thống đã có sẵn từ trước

Trong quá trình biên soạn, mặc dù đã cố gắng tham khảo nhiều tài liệu và giáo trình khác nhưng tác giả không khỏi tránh được những thiếu sót và hạn chế. Tác giả chân thành mong đợi những nhận xét, đánh giá và góp ý để cuốn giáo trình ngày một hoàn thiện hơn.

Tài liệu này được thiết kế theo từng mô đun/ môn học thuộc hệ thống mô đun/môn học của một chương trình, để đào tạo hoàn chỉnh *nghề Lập trình máy tính ở cấp trình độ bậc cao* và được dùng làm Giáo trình cho học viên trong các khoá đào tạo, cũng có thể được sử dụng cho đào tạo ngắn hạn hoặc cho các công nhân kỹ thuật, các nhà quản lý và người sử dụng nhân lực tham khảo.

Đây là tài liệu thử nghiệm sẽ được hoàn chỉnh để trở thành giáo trình chính thức trong hệ thống dạy nghề.

Đà lạt tháng 10 năm 2007

MỤC LỤC

GIỚI THIỆU VỀ MÔN HỌC
KHO DỮ LIỆU

Vị trí, ý nghĩa, vai trò của môn học:

Kho dữ liệu là một tập hợp các CSDL rất lớn tới hàng trăm GB hay thậm chí hàng Tera byte dữ liệu từ nhiều phân hệ của hệ thống, lưu trữ và phân tích phục vụ cho việc cung cấp các dịch vụ thông tin liên quan tới nghiệp vụ một tổ chức, cơ quan hay xí nghiệp. Thông thường dữ liệu phát sinh từ các hoạt động hàng ngày và được thu thập xử lý để phục vụ công việc nghiệp vụ cụ thể của một tổ chức vì vậy thường được gọi là dữ liệu tác nghiệp và hoạt động thu thập xử lý loại dữ liệu này được gọi là xử lý giao dịch trực tuyến OLTP. Kho dữ liệu trái lại phục vụ cho việc phân tích với kết quả mang tính thông tin cao. Các hệ thống thông tin thu thập xử lý dữ liệu loại này còn gọi là hệ xử lý phân tích trực tuyến OLAP..

Kho dữ liệu được xây dựng để tiện lợi cho việc truy cập theo nhiều nguồn, nhiều kiểu dữ liệu khác nhau sao cho có thể kết hợp được cả những ứng dụng của các công nghệ hiện đại và kế thừa được từ những hệ thống đã có sẵn từ trước

Mục tiêu của môn học:

Sau khi học xong môn học này học viên có khả năng:

Xây dựng từ những thông tin tác nghiệp hằng ngày của doanh nghiệp hoặc cơ quan thành kho dữ liệu ở mức Data Mart/ Data Warehouse, biết xây dựng kiến trúc hiệu quả cho kho dữ liệu và biết khai thác một cách hiệu quả thông tin từ các những kho dữ liệu phục vụ việc ra quyết định (DSS, Decision support system) và hỗ trợ cung cấp các thông tin điều hành (Executive support system)

Mục tiêu thực hiện của môn học:

Học xong môn học này học viên có khả năng:

- Nắm nguyên tắc và phương pháp từ các hệ thống thông tin của doanh nghiệp hoặc cơ quan xây dựng thành thành kho dữ liệu ở hai quy mô Data Mart/ Data Warehouse
- Nhận diện, quy hoạch được những dữ liệu từ các cơ sở dữ liệu tác nghiệp đang vận hành (Legacy Database).
- Xây dựng quy trình phát triển kho dữ liệu.
- Khai thác thành thạo, đáp ứng hiệu quả các thông tin có thể tập hợp từ các kho dữ liệu.
- Làm tốt công tác bảo trì kho dữ liệu.

Nội dung chính của môn học:

1. CÁC KHÁI NIỆM VỀ KHO DỮ LIỆU

Chủ đề chính:

- Các khál niệm chính về kho dữ liệu, kiến trúc của kho dữ liệu, phân biệt mức Data mart và mức Data Warehouse.

Kỹ năng thực hành :

- Phân tích kiến trúc và các đối tượng cơ sở dữ liệu tác nghiệp hiện hành để nêu hướng tập hợp thông tin vào các kho dữ liệu. Nguyên cứu và tập hợp các yêu cầu khai thác thông tin ở mức ra quyết định và mức điều hành.

Thái độ học viên:

- Thận trọng, tỉ mỉ và sáng tạo.

2. XÂY DỰNG VÀ PHÁT TRIỂN KHO DỮ LIỆU

Chủ đề chính :

- Lập kế hoạch triển khai, chiến lược triển khai, tập hợp các yêu cầu khai thác thông tin. Phân tích, quy hoạch, thiết kế kho dữ liệu. Càl đặt vật lý kho dữ liệu , quy trình bảo trì và phát triển kho dữ liệu.

Kỹ năng thực hành :

- Lập kế hoạch triển khai, phân tích, quy hoạch và thiết kế kho dữ liệu. áp dụng phương pháp càl đặt kho dữ liệu một cách vật lý và áp dụng quy trình bảo trì. Vận dụng hiệu quả việc trích rút, chuyển dạng và nạp thông tin vào kho dữ liệu vật lý (ETL: Extract, Transform, Loading).

Thái độ học viên :

- Làm việc đúng kế hoạch, có đủ tài liệu hiện trạng và kế hoạch, luôn tuân thủ các quy chế về kho dữ liệu.

3. KHAI THÁC KHO DỮ LIỆU

Chủ đề chính :

- Xử lý thông tin từ kho dữ liệu, tập hợp các quy trình khai thác và tích lũy các yêu cầu, sắp xếp phân loại và quy trình hoá việc khai thác kho dữ liệu. Làm hiệu quả việc phân tích xử lý trực tuyến thông tin (OLAP, Online Analytical Processing)

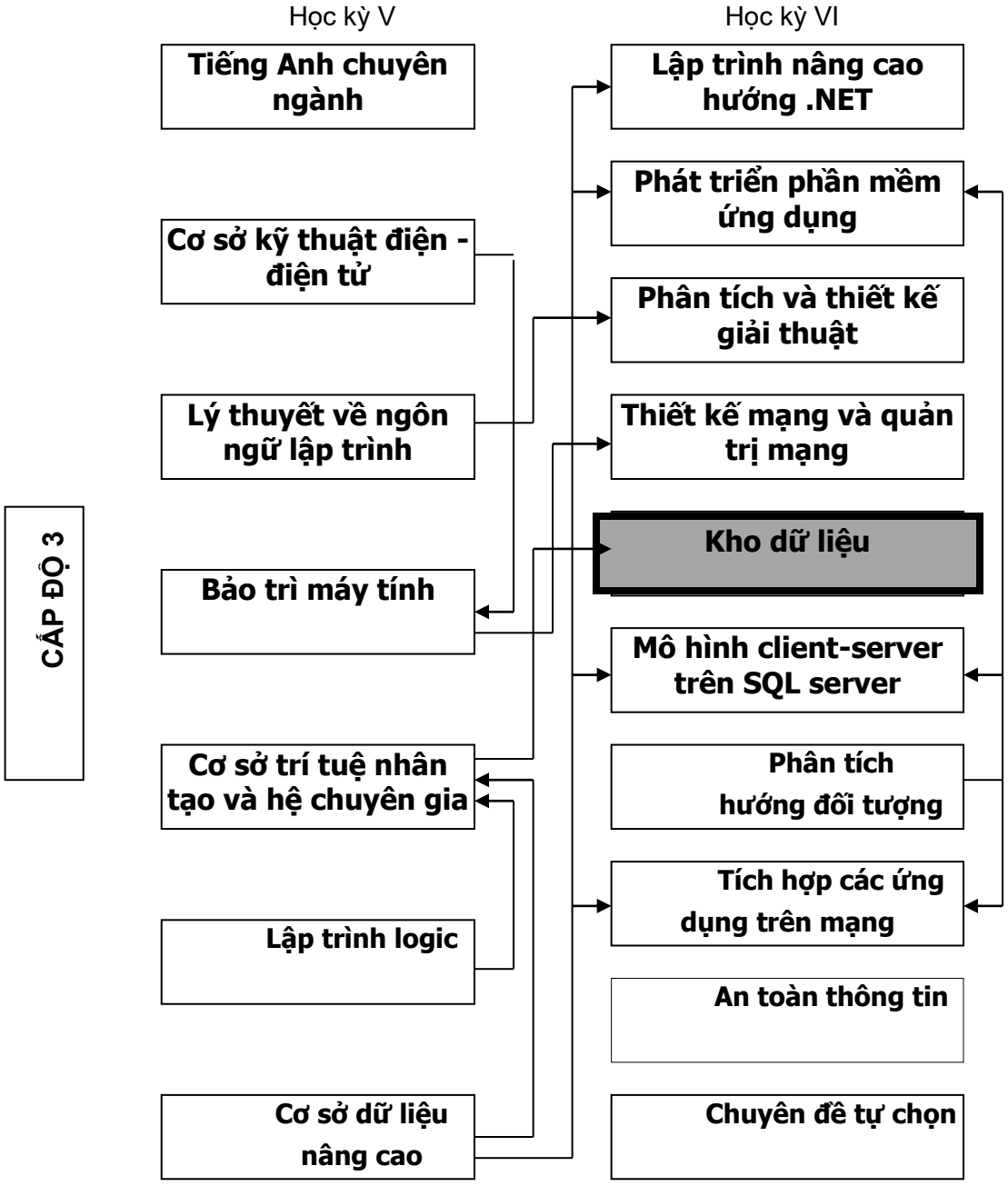
Kỹ năng thực hành :

- Tập hợp thông tin kết xuất đúng cách và tích hợp thông tin đúng nguồn.

Thái độ học viên :

PHÂN TÍCH, KHAI THÁC HIỆU QUẢ, KỊP THỜI KHO DỮ LIỆU.

SƠ ĐỒ QUAN HỆ THEO TRÌNH TỰ HỌC NGHỀ



BÀI 1

TÊN BÀI: CÁC KHÁI NIỆM VỀ KHO DỮ LIỆU

MÃ BÀI: ITPRG3_11.1

1.1. Nhập môn về kho dữ liệu

1.1.1. Kho dữ liệu – Data Warehouse

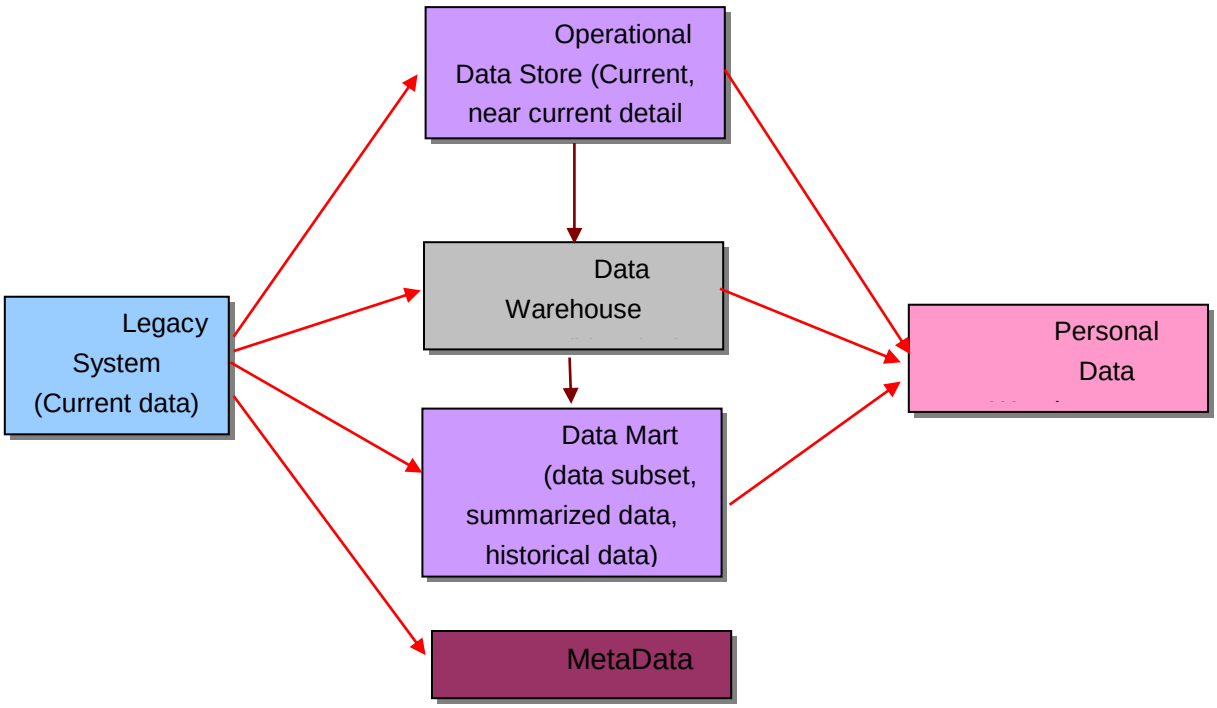
Định nghĩa: Kho dữ liệu (Data Warehouse - DW) là tuyển tập các CSDL tích hợp, hướng chủ đề, được thiết kế để hỗ trợ cho chức năng trợ giúp quyết định, mà mỗi đơn vị dữ liệu đều liên quan tới một khoảng thời gian cụ thể.

Kho dữ liệu thường rất lớn tới hàng trăm GB hay thậm chí hàng Terabyte.

Kho dữ liệu được xây dựng để tiện lợi cho việc truy cập theo nhiều nguồn, nhiều kiểu dữ liệu khác nhau sao cho có thể kết hợp được cả những ứng dụng của các công nghệ hiện đại và kế thừa được từ những hệ thống đã có sẵn từ trước. Dữ liệu phát sinh từ các hoạt động hàng ngày và được thu thập xử lý để phục vụ công việc nghiệp vụ cụ thể của một tổ chức thường được gọi là dữ liệu tác nghiệp (operational data) và hoạt động thu thập xử lý loại dữ liệu này được gọi là xử lý giao dịch trực tuyến (On_line Transaction Processing - OLTP). Kho dữ liệu trái lại phục vụ cho việc phân tích với kết quả mang tính thông tin cao. Các hệ thống thông tin thu thập xử lý dữ liệu loại này còn gọi là hệ xử lý phân tích trực tuyến (On_online Analytical Processing - OLAP).

Nói cách khác, kho dữ liệu là một tập hợp các CSDL rất lớn tới hàng trăm GB hay thậm chí hàng Tera byte dữ liệu từ nhiều phân hệ của hệ thống, lưu trữ và phân tích phục vụ cho việc cung cấp các dịch vụ thông tin liên quan tới nghiệp vụ một tổ chức, cơ quan hay xí nghiệp. Thông thường dữ liệu phát sinh từ các hoạt động hàng ngày và được thu thập xử lý để phục vụ công việc nghiệp vụ cụ thể của một tổ chức vì vậy thường được gọi là dữ liệu tác nghiệp và hoạt động thu thập xử lý loại dữ liệu này được gọi là xử lý giao dịch trực tuyến OLTP. Kho dữ liệu trái lại phục vụ cho việc phân tích với kết quả mang tính thông tin cao. Các hệ thống thông tin thu thập xử lý dữ liệu loại này còn gọi là hệ xử lý phân tích trực tuyến OLAP.

Dòng dữ liệu trong một tổ chức (cơ quan, xí nghiệp, công ty, v.v.) có thể mô tả khái quát như sau:



Hình 1.1. Luồng dữ liệu trong một tổ chức

Dữ liệu cá nhân (Personal Data) không thuộc phạm vi quản lý của hệ quản trị kho dữ liệu. Nó chứa các thông tin được trích xuất ra từ các hệ thống dữ liệu tác nghiệp, kho dữ liệu và từ những kho dữ liệu cục bộ của những chủ đề liên quan bằng các phép gộp, tổng hợp hay xử lý bằng một cách nào đó.

1.1.2. Mục đích của kho dữ liệu

Mục tiêu chính của kho dữ liệu là nhằm đáp ứng các tiêu chuẩn cơ bản:

- Phải có khả năng đáp ứng mọi yêu cầu về thông tin của NSD
- Hỗ trợ để các nhân viên của tổ chức thực hiện tốt, hiệu quả công việc của mình, như có những quyết định hợp lý, nhanh và bán được nhiều hàng hơn, thu được lợi nhuận cao hơn, v.v.
- Giúp cho tổ chức, xác định, quản lý và điều hành các dự án, các nghiệp vụ một cách hiệu quả và chính xác.
- Tích hợp dữ liệu và các siêu dữ liệu từ nhiều nguồn khác nhau.

Muốn đạt được những yêu cầu trên thì DW phải:

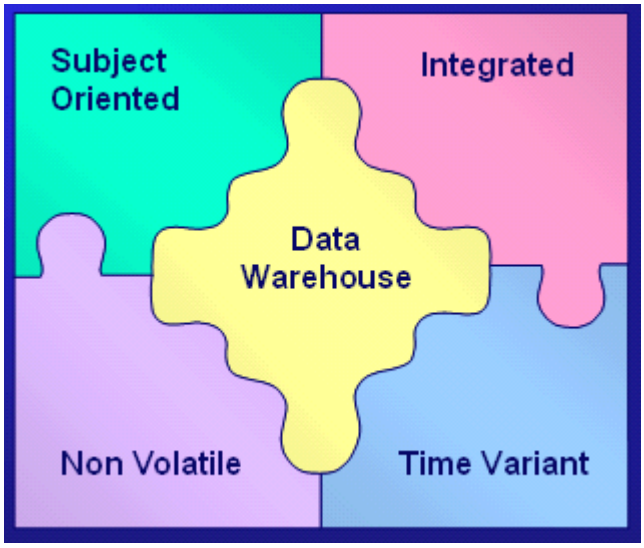
- Nâng cao chất lượng dữ liệu bằng các phương pháp làm sạch và tinh lọc dữ liệu theo những hướng chủ đề nhất định

- Tổng hợp và kết nối dữ liệu
- Đồng bộ hoá các nguồn dữ liệu với DW
- Phân định và đồng nhất các hệ quản trị cơ sở dữ liệu tác nghiệp như là các công cụ chuẩn để phục vụ cho DW.
- Quản lý siêu dữ liệu
- Cung cấp thông tin được tích hợp, tóm tắt hoặc được liên kết, tổ chức theo các chủ đề
- Dùng trong các hệ thống hỗ trợ quyết định (Decision support system - DSS), các hệ thống thông tin tác nghiệp hoặc hỗ trợ cho các truy vấn đặc biệt.

1.1.3. Đặc tính dữ liệu trong kho dữ liệu

Những đặc điểm cơ bản của Kho dữ liệu (DW) là một tập hợp dữ liệu có tính chất sau:

- Hướng chủ đề (Subject Oriented)
- Tích hợp (Integrated)
- Tính ổn định (Non Volatile)
- Có tính lịch sử (Time Variant hoặc Historical)



Hình 1.2. Đặc điểm của DW

1.1.3.1.Hướng chủ đề (Subject Oriented)

Dữ liệu trong DW được tổ chức theo các chủ đề phục vụ cho những tổ chức dễ dàng xác định được những thông tin cần thiết trong từng hoạt động của mình. Ví dụ, trong hệ thống quản lý tài chính cũ có thể dữ liệu được tổ chức theo chức năng: cho vay, quản lý tín dụng, quản lý ngân sách, v.v. Ngược lại, trong DW về tài chính, dữ liệu được tổ chức theo chủ điểm dựa chủ yếu theo các đối tượng: khách hàng, sản phẩm, các xí nghiệp, v.v. Sự

khác nhau của hai cách tiếp cận trên dẫn đến sự khác nhau về nội dung dữ liệu được lưu trữ trong hệ thống:

- DW không lưu trữ dữ liệu chi tiết, chỉ cần lưu trữ những dữ liệu có tính tổng hợp phục vụ chủ yếu cho quá trình phân tích để trợ giúp quyết định.
- Các hệ thống ứng dụng tác nghiệp (Operational Application System- OAS), CSDL tác nghiệp lại cần những dữ liệu chi tiết, phục vụ trực tiếp cho những yêu cầu xử lý theo các chức năng của lĩnh vực ứng dụng hiện thời. Do vậy mối quan hệ của dữ liệu trong những hệ thống này cũng khác, đòi hỏi phải có tính chính xác, có tính thời sự, v.v.
- Dữ liệu gắn thời gian và có tính lịch sử

Một kho chứa dữ liệu bao hàm một khối lượng lớn dữ liệu lịch sử. Dữ liệu được lưu trữ thành một loạt các snapshot (ảnh chụp dữ liệu), mỗi bản ghi phản ánh những giá trị của dữ liệu tại một thời điểm nhất định thể hiện một khung nhìn của một chủ điểm trong một giai đoạn. Do vậy cho phép khôi phục lại lịch sử và so sánh một cách chính xác các giai đoạn khác nhau. Yếu tố thời gian đóng vai trò như một phần của khoá để bảo đảm tính đơn nhất của mỗi hàng và cung cấp đặc trưng về thời gian cho dữ liệu.

Dữ liệu trong OLTP cần phải chính xác ở chính thời điểm truy cập, còn ơ DW chỉ cần có hiệu lực trong khoảng thời gian nào đó, trong khoảng 5 đến 10 năm hoặc lâu hơn. Dữ liệu của CSDL tác nghiệp thường sau một khoảng thời gian nhất định thì sẽ trở thành dữ liệu lịch sử và chúng sẽ được chuyển thành kho dữ liệu. Đó chính là những dữ liệu hợp lý về những chủ điểm cần lưu trữ.

CSDL OLTP	**DW**
+ Dữ liệu nghiệp vụ:	+ ảnh chụp dữ liệu:
+ Thời gian ngắn 30-60 ngày	+ Thời gian dài: 5 đến 10 năm
+ Có thể có yếu tố thời gian hoặc không	+ Luồng có yếu tố thời gian
+ Dữ liệu có thể cập nhật	+ Khi dữ liệu được chụp lại thì không cập nhật được

Hình 1.3. Tính thời gian của dữ liệu

1.1.3.2. Tính tích hợp (Integrated)

Dữ liệu trong DW được tổ chức theo nhiều cách khác nhau sao cho phù hợp với các qui ước đặt tên, thống nhất về số đo, cơ cấu mã hoá và cấu trúc vật lý của dữ liệu, v.v. Một DW là một khung nhìn thông tin mức toàn xí nghiệp, thống nhất các khung nhìn khác nhau thành một khung nhìn theo một chủ điểm nào đó. Ví dụ hệ thống OLTP (xử lý giao dịch trực tuyến) truyền thống được xây dựng trên một vùng nghiệp vụ. Một hệ thống bán hàng và một hệ thống marketing có thể có chung một dạng thông tin về khách hàng, nhưng các vấn đề về tài chính cần một khung nhìn khác cho thông tin về khách hàng. Một DW sẽ có một

khung nhìn toàn thể về một khách hàng. Khung nhìn đó bao gồm các phần dữ liệu khác nhau từ tài chính và marketing.

Tính tích hợp thể hiện ở chỗ: Dữ liệu tập hợp trong kho dữ liệu được thu thập từ nhiều nguồn và trộn ghép với nhau tạo thành một thể thống nhất.

1.1.3.3.Tính ổn định (Non volatile)

Dữ liệu trong DW là dữ liệu chỉ đọc và chỉ có thể được kiểm tra, không được sửa đổi bởi người sử dụng đầu cuối. Nó chỉ cho phép thực hiện hai thao tác cơ bản:

- Nạp dữ liệu vào kho,
- Truy cập vào các vùng trong DW.
- Dữ liệu không biến động

Thông tin trong DW được tải vào sau khi dữ liệu trong hệ thống điều hành được cho là quá cũ. Tính không biến động thể hiện ở chỗ: Dữ liệu được lưu trữ lâu dài trong kho dữ liệu. Mặc dù có thêm dữ liệu mới nhập vào nhưng dữ liệu cũ trong kho vẫn không bị xoá, điều đó cho phép cung cấp thông tin về một khoảng thời gian dài, cung cấp đủ số liệu cần thiết cho các mô hình nghiệp vụ phân tích, dự báo, từ đó có được những quyết định hợp lý, phù hợp với các qui luật tiến hoá của tự nhiên. Tuy nhiên trong thức tế nếu các bảng dữ liệu có kích thước qua lớn thì ta cũng phải đặt ra kế hoạch để lưu trữ bớt các dữ liệu trong quá khứ, thời gian có thể là sau 3-10 năm tuỳ theo yêu cầu nghiệp vụ báo cáo liên quan. Sau khi lưu trữ dữ liệu cũ thì có thể xoá đi để giảm bớt dung lượng cần thiết để lưu trữ và tăng tốc độ truy cập.

1.1.3.4.Tính lịch sử (Time Variant)

DW thường chứa một kkối lượng lớn dữ liệu lịch sử. Dữ liệu được lưu trữ thành hàng loạt các bản chup ảnh (snapshot), mỗi bản ghi phản ánh giá trị của dữ liệu tại một thời điểm nhất định, thể hiện một khung nhìn cho một giao độan nhất định. Do đó cho phép ngươiừ sử dụng có thể lấy lại dữ liệu lich sử và so sánh dữu liệu cho các giai đoạn khác nhau. Yếu tố thời gian đóng vai trò như một phần của khoá để đảm bảo tính duy nhất của một hàng và cung cấp đặc trưng về thời gian cho dữ liệu.

1.1.4. Phân biệt DW với OLTP

Trên cơ sở các đặc trưng của DW, ta phân biệt DW với những hệ quản trị cơ sở dữ liệu tác nghiệp truyền thống:

Kho dữ liệu phải được xác định theo hướng chủ đề. Nó được thực hiện theo ý đồ của người sử dụng đầu cuối trong khi các hệ CSDL tác nghiệp dùng để phục vụ các mục đích áp dụng chung.

DW quản lý một khối lượng lớn thông tin được lưu trữ trên nhiều phương tiện lưu trữ và xử lý khác nhau. Những hệ CSDL thông thường không phải quản lý những lượng thông

tin lớn mà quản lý những lượng thông tin vừa và nhỏ. Trong khi đó thì DW phải quản lý những lượng thông tin rất lớn và đó cũng chính là đặc thù của kho dữ liệu.

DW có thể ghép nối các version khác nhau của các loại cấu trúc CSDL. DW tổng hợp thông tin để thể hiện chúng dưới những hình thức dễ hiểu đối với người sử dụng.

DW tích hợp và kết nối thông tin từ những nguồn khác nhau trên nhiều loại phương tiện lưu trữ và xử lý thông tin nhằm phục vụ cho những ứng dụng xử lý tác nghiệp trực tuyến.

DW có thể lưu trữ các thông tin tổng hợp theo một chủ đề nghiệp vụ nào đó sao cho tạo ra các thông tin phục vụ hiệu quả cho việc phân tích của người sử dụng.

DW thông thường chứa các dữ liệu lịch sử kết nối nhiều năm trước của các thông tin tác nghiệp được tổ chức lưu trữ có hiệu quả và có thể hiệu chỉnh lại dễ dàng. Dữ liệu trong CSDL tác nghiệp thường là mới, có tính thời sự trong khoảng thời gian ngắn.

Dữ liệu từ CSDL tác nghiệp được chắt lọc và tổng hợp lại để chuyển sang môi trường DW. Rất nhiều dữ liệu khác không được chuyển về DW, chỉ những dữ liệu cần thiết cho công tác quản lý hay trợ giúp quyết định mới được chuyển sang DW.

Nói một cách tổng quát, DW làm nhiệm vụ phân phát dữ liệu cho nhiều đối tượng (khách hàng) xử lý thông tin dưới nhiều dạng như: CSDL, SQL query, Reports, v.v.

1.1.5. Một số loại kho dữ liệu cơ bản khác

1.1.5.1. Kho dữ liệu cục bộ - Datamart

Kho dữ liệu cục bộ (Datamart – DM) là CSDL có những đặc điểm giống với kho dữ liệu nhưng với quy mô nhỏ hơn và lưu trữ dữ liệu về một lĩnh vực, một chuyên ngành. Datamart là kho dữ liệu hướng chủ đề. Các Datamart có thể được hình thành từ một tập con dữ liệu của kho dữ liệu hoặc cũng có thể được xây dựng độc lập và sau khi xây dựng xong, các datamart có thể được kết nối tích hợp lại với nhau tạo thành kho dữ liệu. Vì vậy có thể xây dựng kho dữ liệu bắt đầu bằng việc xây dựng các Datamart hay ngược lại xây dựng kho dữ liệu trước sau đó tạo ra các Datamart.

Datamart (DM) là một kho dữ liệu thứ cấp các dữ liệu tích hợp của DW. Datamart được hướng tới một phần của dữ liệu thường được gọi là một vùng chủ đề (Subject Area - SA) được tạo ra và giành cho một nhóm người sử dụng. Dữ liệu trong Datamart cho thông tin về một chủ đề xác định, không phải về toàn bộ các hoạt động nghiệp vụ đang diễn ra trong một tổ chức. Thể hiện thường xuyên nhất của datamart là một kho dữ liệu riêng rẽ theo phương diện vật lý, thường được lưu trữ trên một server riêng, trong một mạng cục bộ phục vụ cho một nhóm người nhất định. Đôi khi datamart một cách đơn giản với công nghệ OLAP tạo ra các quan hệ theo dạng hình sao đặc biệt hoặc những siêu khối (hypercube) dữ liệu cho việc phân tích của một nhóm người có cùng mối quan tâm trên một phạm vi dữ liệu.

Có thể chia ra làm 2 loại: Datamart độc lập và Datamart phụ thuộc

- **Datamart phụ thuộc:** chứa những dữ liệu được lấy từ DW và những dữ liệu này sẽ được trích lọc và tinh chế, tích hợp lại ở mức cao hơn để phục vụ một chủ đề nhất định của Datamart.
- **Datamart độc lập:** không giống như Datamart phụ thuộc, DM loại này được xây dựng trước DW và dữ liệu được trực tiếp lấy từ các nguồn khác nhau. Phương pháp này đơn giản hơn và chi phí thấp hơn nhưng đổi lại có những điểm yếu. Mỗi DM độc lập có cách tích hợp riêng, do đó dữ liệu từ nhiều DM khó đồng nhất với nhau. DM thể hiện hai vấn đề: thứ nhất là tính ổn định trong các tình huống từ một DM nhỏ ban đầu lớn lên nhanh chóng theo nhiều chiều và thứ hai là sự tích hợp dữ liệu. Vì vậy khi thiết kế DM phải chú ý kỹ tới tính ổn định của hệ thống, sự đồng nhất của dữ liệu và vấn đề về khả năng quản lý.

Xây dựng kho dữ liệu (Data Warehousing) không phải là một sản phẩm mà là một quá trình kỹ thuật thu thập, quản lý và khai thác dữ liệu một cách hợp lý từ nhiều nguồn khác nhau, để thiết lập một kho dữ liệu là tập hợp các dữ liệu hợp nhất phản ánh chi tiết một phần hay toàn bộ công tác nghiệp vụ của một tổ chức hay nói cách khác, đây là quá trình xác lập cách nhìn, lập kế hoạch, xây dựng, sử dụng, quản trị, bảo trì và nâng cấp Kho dữ liệu và Datamart. Không phụ thuộc vào việc xây dựng một kho dữ liệu hay một datamart, quá trình rất phức tạp và luôn luôn tiếp diễn với trọng tâm là các nhu cầu nghiệp vụ đối với kiến thức lấy dữ liệu làm căn cứ.

1.1.5.2. Kho Siêu dữ liệu (Metadata)

Metadata là dữ liệu về dữ liệu được sử dụng trong DW (hay gọi là siêu dữ liệu) trả lời các câu hỏi ai, cái gì, khi nào, tại sao, như thế nào về dữ liệu. Nó được sử dụng cho việc xây dựng, duy trì, quản lý và sử dụng DW.

Metadata được chia thành 3 loại: siêu dữ liệu nghiệp vụ, kỹ thuật và tác nghiệp (thao tác)

- **Siêu dữ liệu nghiệp vụ (Business Metadata):** chứa đựng những thông tin khiến cho người sử dụng dễ dàng hiểu được khung cảnh của thông tin được lưu trữ trong DW. Nó chứa đựng những thông tin cho tất cả những người sử dụng đầu cuối và về:
    - Các các vùng chủ đề (Subject Area - SA) và các loại đối tượng thông tin bao gồm các câu truy vấn, các báo cáo, các hình ảnh, video và các audio clip
    - Các trang chủ trên Internet.
    - Các thông tin khác để hỗ trợ cho tất cả các thành phần cấu thành DW. Chẳng hạn như các thông tin liên quan tới các hệ thống phân phối thông tin bao gồm thông tin về lịch làm việc, những chi tiết về nơi phân phối và các đối tượng truy vấn như những truy vấn, báo cáo và các phân tích được xác định trước.

- Các thông tin tác nghiệp của DW như lịch sử của dữ liệu (các snapshot, các version), quyền sở hữu, theo dõi sổ sách, sử dụng dữ liệu.
- Miêu tả các thuộc tính DW bằng cách xác định tên của công việc, các định nghĩa, các bảng mô tả và các bí danh.

- **Siêu dữ liệu kỹ thuật (Technical Metadata):** chứa đựng những thông tin về dữ liệu trong DW của những người thiết kế và quản trị khi tiến hành công việc phát triển và quản lý. Nó bao gồm:
  - Thông tin về các nguồn dữ liệu kể cả những nguồn tác nghiệp và những hệ thống nguồn bên ngoài môi trường DW về vị trí, tên các file, kiểu file, tên các trường và các đặc tính, bí danh, thông tin về phiên bản, những mối quan hệ, độ lớn, tính dễ biến động, người chủ dữ liệu và những người sử dụng có quyền truy nhập.
  - Những mô tả về sự chuyển đổi ví dụ như cách thức ánh xạ từ cơ sở dữ liệu tác nghiệp lên DW và các thuật toán được sử dụng để biến đổi và cải thiện hay chuyển đổi dữ liệu.
  - Những định nghĩa cấu trúc dữ liệu và đối tượng trong môi trường Warehouse cho dữ liệu đích.
  - Những luật dùng để làm sạch và cải thiện dữ liệu.
  - Những phép toán ánh xạ dữ liệu khi lấy dữ liệu từ các hệ thống nguồn và đưa chúng vào cơ sở dữ liệu đích.
  - Quyền truy nhập, lịch sử về backup, về sự lưu trữ, về sự phân phối thông tin, về sự thu nhận dữ liệu, về sự truy nhập dữ liệu, v.v..
- **Siêu dữ liệu tác nghiệp (Operational Metadata - OM)**
  - OM giúp trong việc duy trì và triển khai DW.
  - OM mô tả thông tin chứa đựng trong các bảng đích.
  - Mô tả cốt lõi, khả năng tạo cơ sở dữ liệu đích (tạo ra bảng và thông tin dưới dạng liệt kê), thông tin được lưu trữ hay trực tuyến, ngày làm tươi mới (refresh) dữ liệu, số lượng các bản ghi, lịch thực hiện các công việc và những người sử dụng có khả năng truy nhập vào data.
  - Metadata cung cấp cho người sử dụng sự truy nhập tương tác để giúp cho họ có thể hiểu được nội dung và tìm thấy được dữ liệu cần thiết. Một vấn đề là trong thực tế khả năng kết hợp của công cụ trích lọc dữ liệu và Metadata còn khá thô. Do đó cần phải tạo ra những giao diện dùng Meatadata cho người sử dụng .

Tất cả các thành phần của DW đều cần và có thể lấy dữ liệu từ Metadata. Metadata được lưu trữ ở khu vực trung tâm.

1.1.5.3. Kho dữ liệu tác nghiệp (ODS)

Kho dữ liệu tác nghiệp (Operational Database Store - ODS) là hệ thống tác nghiệp tích hợp căn bản dùng cho mục đích thực hiện công việc trợ giúp quyết định và phân tích trên dữ liệu giao dịch tác nghiệp. Nói một cách khác, ODS là một khái niệm có kiến trúc để hỗ trợ cho việc tạo quyết định tác nghiệp hàng ngày lưu trữ những dữ liệu có giá trị hiện thời được chuyển đến từ các ứng dụng tác nghiệp. Điều đó khiến cho dữ liệu lưu trữ trong ODS biến động thường xuyên khi những dữ liệu liên quan trong các hệ thống tác nghiệp có sự thay đổi. ODS cung cấp một sự lựa chọn cho các ứng dụng trợ giúp quyết định tác nghiệp, truy nhập dữ liệu một cách trực tiếp từ các hệ thống xử lý các giao dịch trực tuyến.

Đôi khi cũng có những sự nhập nhằng giữa ODS với DW, nên cần phải phân biệt chúng với nhau. Trong tất cả các trường hợp, ODS cần phải được xây dựng riêng biệt và là một phần của DW.

Một trong những sự khác nhau cơ bản và quan trọng nhất là ở nội dung và các cấu trúc dữ liệu được lưu trữ. ODS chứa những dữ liệu có giá trị hiện thời hoặc gần với dữ liệu hiện thời, còn DW chứa những dữ liệu lịch sử, có giá trị trong một quá khứ gần. ODS có thể cập nhật còn DW không cập nhật được.

Nói chung dữ liệu trong DW thường là rất lớn, nhiều hơn ở ODS, nghĩa là chúng khác nhau về số lượng, phạm vi lưu trữ dữ liệu.

ODS chỉ tập trung lưu trữ những dữ liệu thuần nhất và có giá trị hiện thời còn DW có thể chứa rất nhiều dữ liệu ở nhiều mức độ khác nhau, những dữ liệu không thuần nhất.

Một sự khác nhau nữa là công nghệ hỗ trợ cho hai hệ thống đó. ODS đòi hỏi phải là môi trường được phép cập nhật, ghi, thay đổi được những dữ liệu cần thiết để cho phù hợp với nghiệp vụ và nhanh chóng trả lời được các yêu cầu của NSD, DW thì ngược lại, chỉ yêu cầu đơn giản là Load-and-Access.

Có thể dùng tần suất cập nhật (thời gian thực hoặc gần thời gian thực, định kỳ hay qua mùa) để phân loại các ODS. Mặc dù kiến trúc của ODS khác nhiều so với DW nhưng hai loại cuối của ODS khá giống với DW. Đó là lý do nhiều yêu cầu ứng dụng của ODS được thực hiện thông qua việc truy nhập trực tiếp tới kho dữ liệu tác nghiệp và cải thiện những xử lý tinh chế để tạo ra DW. Như vậy có thể có trường hợp dữ liệu từ các nguồn không được tinh chế chuyển đổi và tải trực tiếp vào DW mà trước hết được tải và chuyển đổi vào ODS rồi mới được xử lý tinh chế, làm sạch cho vào DW hoặc DM. Tuy nhiên có một số khó khăn chính của ODS vẫn còn tồn tại. Trong đó có những vấn đề cần giải quyết sau:

- Vị trí nguồn dữ liệu thích hợp.
- Việc chuyển đổi nguồn dữ liệu để đáp ứng được nhu cầu của mô hình dữ liệu ODS.
- Sự phức tạp của việc chuyển tải những thay đổi từ các hệ thống tác nghiệp tới ODS.
- Một hệ quản trị cơ sở dữ liệu kết hợp những xử lý truy vấn hiệu quả với khả năng xử lý những giao dịch bảo đảm những thuộc tính giao dịch ACID.

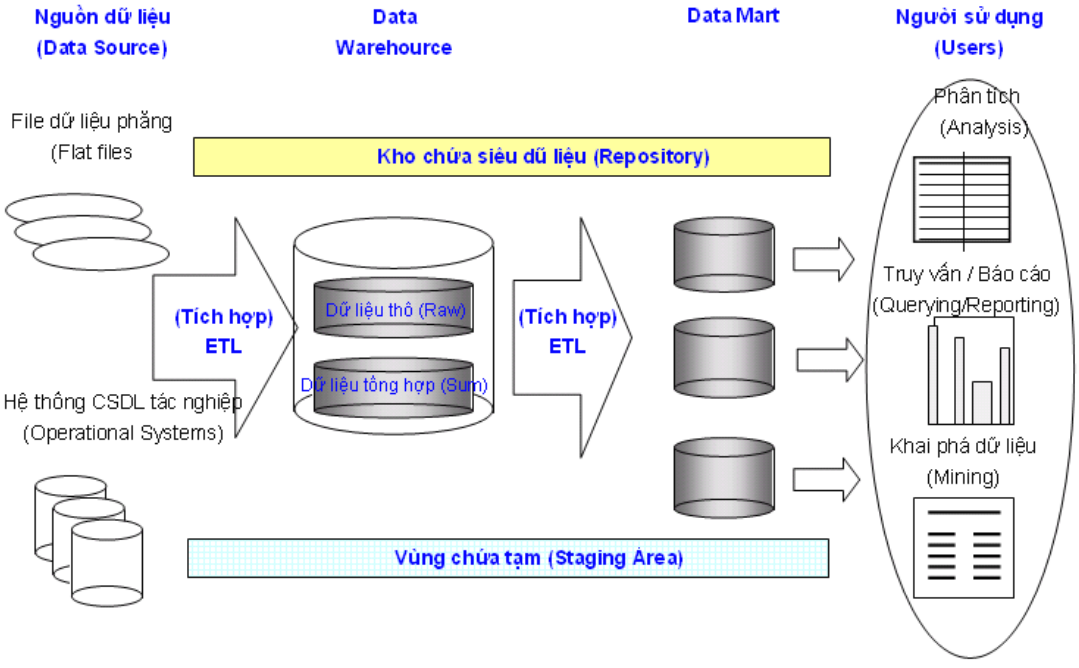
Cách thiết kế cơ sở dữ liệu tối ưu, một mặt hỗ trợ cho các hoạt động của hệ trợ giúp quyết định khắt khe nhất và đồng thời giảm được số lượng chỉ số để giảm thiểu ảnh hưởng tới việc cập nhật.

Về mặt chức năng, ODS cung cấp một khung nhìn tập trung về dữ liệu gần với thời gian thực từ các hệ thống tác nghiệp. Mặc dầu hầu hết các ODS được làm mới lại hàng ngày (đối với những DW có chu kỳ là một ngày) trong những trường hợp nhất định cần thiết có một sự phân tích nhanh để quản lý công việc và nếu dữ liệu tồn tại trong những file riêng rẽ, một ODS là thích hợp nhất với sự phân tích này. Thêm nữa, ODS có thể là vật thay thế cho một bản ghi những thay đổi được dùng cho việc làm mới lại những file DSS khác trong công ty.

Trong mối quan hệ với DW, ODS có thể được sử dụng như kho dữ liệu dùng cho việc tập hợp dữ liệu từ các nguồn khác nhau. Ngược lại ODS không hoạt động như là một kho dữ liệu trung gian cho DW đặc biệt trong trường hợp DW cần dữ liệu từ những nguồn bên ngoài, không nằm trong ODS. Trong trường hợp đó DW có thể lấy dữ liệu một cách riêng rẽ từ ODS hoặc một nguồn dữ liệu bên ngoài được thêm vào thành phần tinh chế dữ liệu của DW.

1.2. Kiến trúc kho dữ liệu

Có rất nhiều loại kiến trúc kho dữ liệu – Data Warehouse khác nhau tuỳ theo mục đích yêu cầu của từng tổ chức. Ngày nay nói đến kiến trúc của Data Warehouse thì phải hiểu là một kiến trúc tổng thể cho hệ thống nghiệp vụ thông minh (BI-Business Intelligence), kiến trúc phổ biến nhất như sau:



Hình 1.4. Kiến trúc DW

1.2.1. Nguồn Dữ liệu (Data Source)

Nguồn dữ liệu cho Kho dữ liệu – Data Warehouse có thể một trong các dạng sau:

- Cơ sở dữ liệu của các phần mềm ứng dụng hoặc của các hệ thống tác nghiệp được lưu trữ bởi một hệ quản trị CSDL như Oracle, SQL Server, Access, DB2....
- Các file phẳng (Flat file), file log....

1.2.2. Thành phần Kho dữ liệu (Data Warehouse)

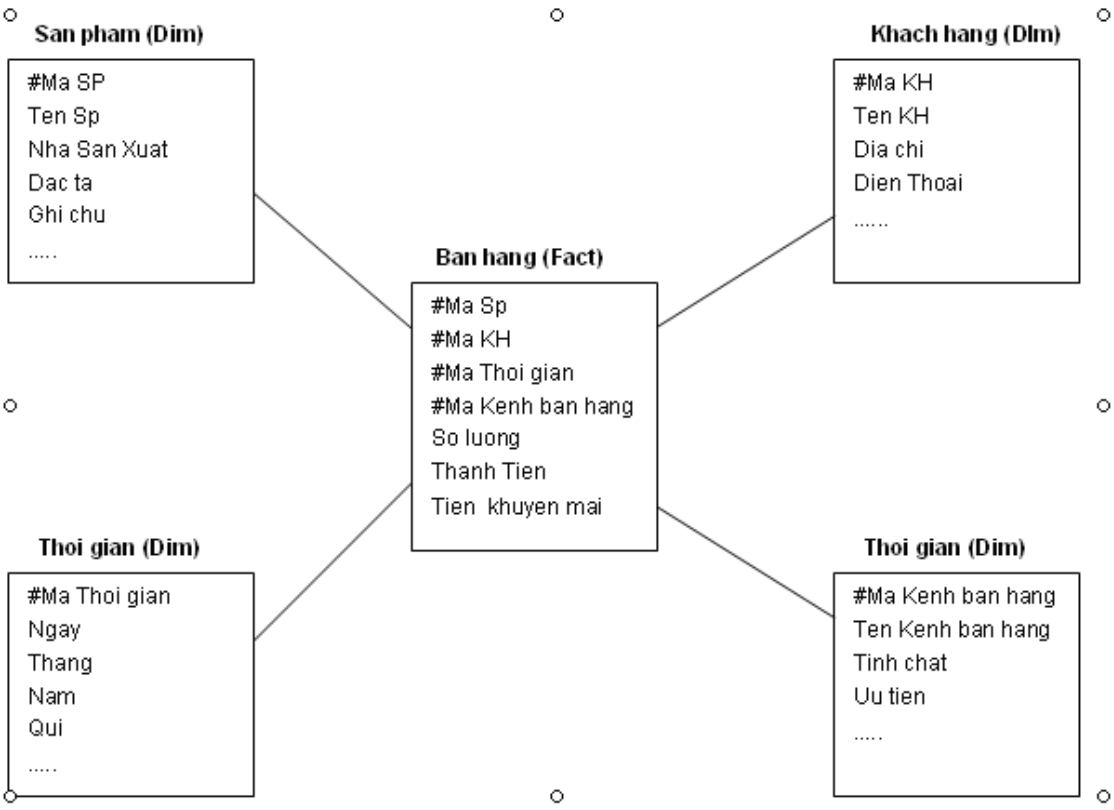
Là thành phần chứa dữ liệu lịch sử (chứa đầy đủ cả dữ liệu trong quá khứ) của nhiều chủ đề khác nhau của tổ chức, nó bao gồm cả dữ liệu thô và dữ liệu đã được tổng hợp ở một mức thấp. Cụ thể như sau:

- **Dữ liệu thô (Raw Data)**: đây là phần chứa toàn bộ dữ liệu ở mức chi tiết nhất (atomic data) được lấy từ dữ liệu nguồn (sau khi đã loại bỏ những dữ liệu không cần thiết và biến đổi chúng), trong phần này dữ liệu thường vẫn được tổ chức tuân theo chuẩn 3NF.
- **Dữ liệu tổng hợp (Summary Data):** đây là phần chứa dữ liệu ở mức tổng hợp hơn (được nhóm theo một số chiều nhất định) tùy theo mục đích cụ thể của từng tổ chức mà ta tạo ra các bảng tổng hợp khác nhau. Các bảng dữ liệu tổng hợp này thường tổ chức phi chuẩn và được dùng cho mục đích phân tích báo cáo cũng như đầu vào dữ liệu cho việc xây dựng Data Mart.

1.2.3. Thành Phần Data Mart

Là thành phần chứa dữ liệu tổng hợp theo một chủ đề nào đó (ví dụ bán hàng, tiền lương, khuyến mãi, thu nợ,....) nhằm phục vụ cho việc truy vấn, báo cáo và phân tích dữ liệu một cách dễ dàng và nhanh chóng có kết quả. Trong thành phần này mô hình dữ liệu thường được tổ chức dưới dạng mô hình hình sao (Star Schema) bao gồm bảng dữ liệu thống kê nằm ở trung tâm gọi là fact và nhiều chiều thống kê gọi là Dimension nằm ở xung quanh.

Ví dụ một mô hình hình sao về bán hàng như sau:



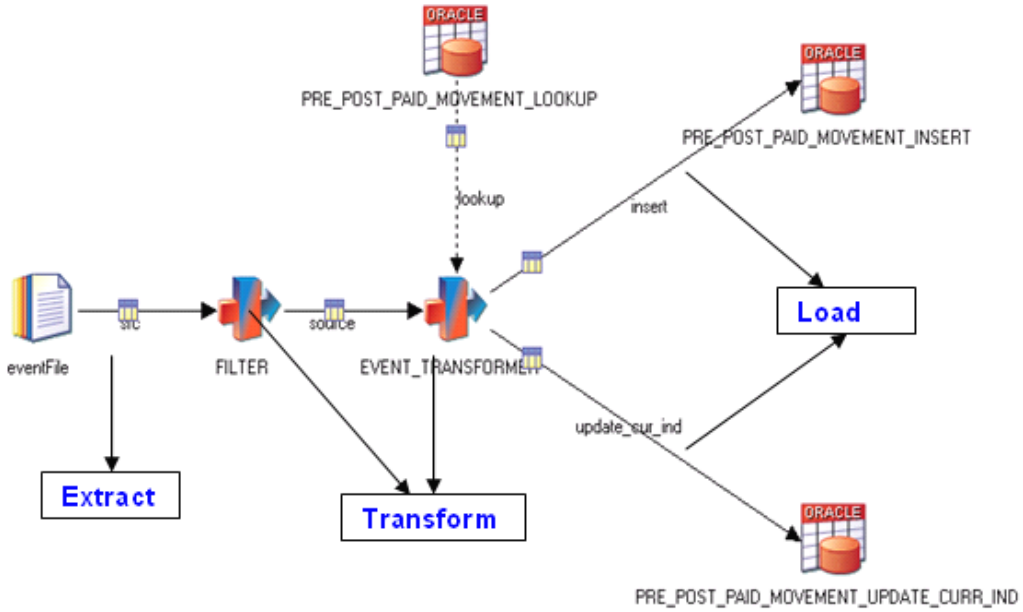
Hình 1.5. Mô hình hình sao (Star Schema)

1.2.4. Thành Phần tích hợp (ETL)

ETL chính là quá trình thực hiện tích hợp dữ liệu, quá trình này thông thường gồm 03 bước sau:

- **Extract (Rút trích, tách trích, lấy...):** là bước đọc dữ liệu nguồn
- **Transform (Biến đổi):** là bước thực hiện các phép gộp, tách, chuẩn hóa dữ liệu ....
- **Load (tải vào):** là bước thực hiện ghi dữ liệu vào bảng đích

Ví dụ: một công việc ETL (một Job)



Hình 1.6. ETL

Với kiến trúc kho dữ liệu Data Warehouse như trên thì ETL được dùng với 02 mục đích là tích hợp dữ liệu cho thành phần Data Warehouse và thành phần Data Mart.

ETL có thể là một chương trình được viết bàng ngôn ngữ cấp cao như C, C+, C++, VB, VB.Net, Java, SÁ... hoặc có thể là bàng số ngôn ngữ đi theo như PL/SQL, TSQL... hoặc cũng có thể dùng những công cụ tích hợp sẵn có trên thị trường như BO Data Integrator, IBM Data Stage, Oracle Warehouse Builder...

Thông thường công cụ ETL phải có các 03 mô đun sau: mô đun tạo lập các Job để thực hiện việc tích hợp, mô đun thiết lập thời gian chạy cho các Job và mô đun theo dõi quá trình chạy cũng nhưng kiểm soát lỗi.

1.2.5. Thành Phần kho chức siêu dữ liệu (Repository)

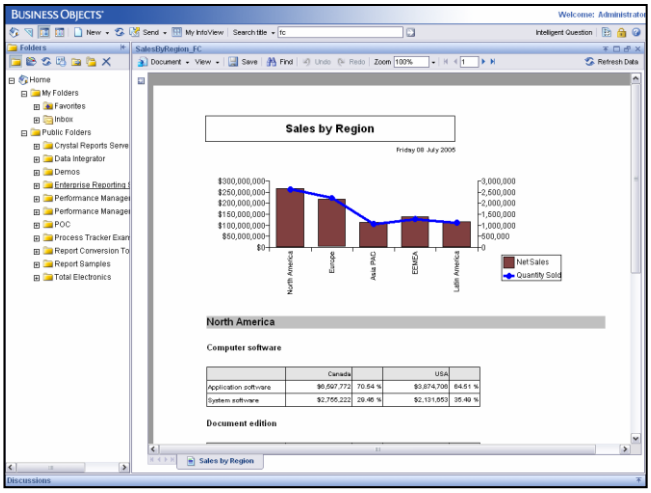
Là thành phần chứa dữ liệu định nghĩa về cấu trúc kho dữ liệu (cả thành phần Data Warehouse và Data mart), định nghĩa về các Job cho quá trình ETL, chứa các dữ liệu về người sử dụng và quyền hạn.... nó có thể được lưu trữ trong một cơ sở dữ liệu quan hệ hoặc dưới dang hệ thống file có tổ chức (theo một thuật toán nào đấy)

1.2.6. Thành phần vùng chứa tạm (Staging Area)

Là thành phần chứa dữ liệu trung gian phục vụ cho quá trình tích hợp dữ liệu được hiệu quả hơn, nó không có ý nghĩa với người dùng đầu cuối. Vùng dữ liệu tạm có thể được lưu trữ trong một cơ sở dữ liệu quan hệ hoặc dưới dang hệ thống file phẳng.

1.2.7. Thành phần khai thác cho người sử dụng (Users)

Đây chính là các công cụ để khai thác kho dữ liệu có thể chạy trên nền web hoặc Desktop và thường có các chức năng truy vấn dữ liệu, tạo báo cáo và phân tích số liệu. Ví dụ: SQL Server Analyze, Oracle Discovery, BO Web Intelligence, Cristal report....



BÀI TẬP:

1. Thế nào về kho dữ liệu ? Quản lý kho dữ liệu ? – để cho học viên phát biểu trước khi đưa ra các khái niệm chính thức
2. Nêu lý do mà bạn cho là các kho dữ liệu thường rất phức tạp ? Phương án để đơn giản hoá các chiều của kho dữ liệu
3. Các công việc của một người quản lý kho dữ liệu là gì? Tiêu chuẩn của người quản lý kho dữ liệu tốt là gì?

BÀI 2

TÊN BÀI: XÂY DỰNG VÀ PHÁT TRIỂN KHO DỮ LIỆU (DW)

MÃ BÀI: ITPRG3_11.2

2.1. Tổng quan về xây dựng và phát triển DW

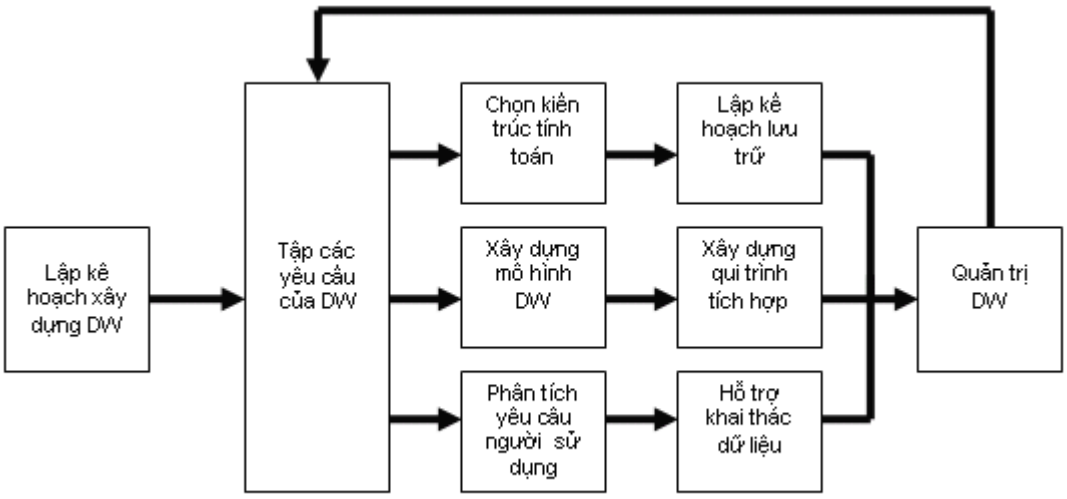
Việc xây dựng và phát triển kho dữ liệu là một việc không hề đơn giản, nó đòi hỏi phải có một phương pháp tiếp cận thích hợp nhất. Hiện nay có nhiều cách tiếp cận khác nhau nhưng đều dựa trên ba kiểu cơ bản sau:

Tiếp cận theo hướng tăng dần (Incremental Approach)

Tiếp cận theo hướng từ trên xuống (To-Down Approach)

Tiếp cận theo hướng từ dưới lên (Bottom-up Approach)

Mỗi cách tiếp cận đều có các ưu và nhược điểm riêng, nên trong giáo trình này ta chỉ tập trung nghiên cứu một phương pháp tiếp cận phổ biến nhất như sau:



Hình 2.1. Qui trình xây dựng và phát triển DW

Như vậy, để xây dựng và phát triển kho dữ liệu cần phải thực hiện các công việc cơ bản sau:

- **Lập kế hoạch xây dựng DW:** kết quả là phải đưa ra được tập các yêu cầu cho việc xây dựng DW và kế hoặc thực hiện.
- **Phân tích yêu cầu người sử dụng:** để tập hợp được tất cả các yêu cầu của người sử dụng.
- **Xây dựng Mô hình DW:** định nghĩa mô hình nghiệp vụ (Concept Model), tạo mô hình logic và tạo mô hình dữ liệu vật lý.

- **Lập kế hoạch cài đặt vật lý:** bao gồm việc lựa chọn kiến trúc tính toán cho hệ thống máy chủ phục vụ cho DW và lập kế hoặc lưu trữ dữ liệu
- **Xây dựng qui trình tích hợp cho DW:** xây dựng ra các công việc (Job) để thực hiện việc tích hợp dữ liệu từ nguồn vào DW
- **Quản trị DW:** cài đặt đưa vào sử dụng DW, quản lý khai thác và cập nhật dư x liệu liên tục cho DW....

2.2. Lập kế hoạch xây dựng và phát triển DW

Việc lập kế hoạch xây dựng và phát triển kho dữ liệu bao gồm các nội dung cơ bản sau:

- Lập kế hoạch tài chính
- Lập kế hoạch về nghiệp vụ
- Lập kế hoạch về kỹ thuật

2.2.1. Lập kế hoạch tài chính

Phải ước lượng được tổng chi phí cần đầu tư cho việc xây dựng và phất triển DW cũng như ai là người cung cấp và quản lý chi phí. Chi phí được xác định dựa trên các khoản chi phí sau:

- Chi phí mua sắm phần cứng (máy móc, thiết bị), mua phần mềm (hệ QTCSDL và các phần mềm khác),
- Chi phí cho nhân lực phát triển
- Chi phí cho nhân lực quản lý
- Chi phí mua mua thông tin từ bên ngoài nếu cần thiết
- Chi phí bảo trì và phát triển thêm sau khi được vào sử dụng...

2.2.2. Lập kế hoạch về nghiệp vụ

Phải định nghĩa được các mục đích nghiệp vụ mà DW sẽ mang lại, định nghĩa các chủ đề mà DW sẽ hướng đến. Cụ thể là khi ập kế hoạch nghiệp vụ phải đề cập đến các nôi dung sau:

- Xác định phạm vi công việc cần thực hiện cho việc xây dựng và phát triển DW
- Phân tích lợi ích của việc xây dựng DW
- Tập hợp tất cả các thông tin yêu cầu của người dùng nghiệp vụ đầu cuối
- Cân nhắc hiệu quả đặt được cho từng đối tượng nghiệp vụ

Thông thường việc lập kế hoạch nghiệp vụ cho dự án xây dựng DW tiến hành theo tuần tự các bước sau:

- Xác định mục đích xây dựng DW
- Xác định phạm vi, xác định cái gì thuộc về công việc trong dự án xây dựng DW, cái gì bên ngoài không tuộc dự án.

- Xác định các đối tượng mà DW hướng đến và mục tiêu cho từng đối tượng
- Xác định vai trò trách nhiệm của các thành phần tham gia dự án
- Xác định các yêu cầu về huấn luyện và đào tạo
- Xác định các yếu tố rủi ro
- Xác định các phương án dự phòng trong qua trình triển khai dự án.

2.2.3. Lập kế hoạch về kỹ thuật

Phải xác định được yêu cầu kỹ thuật để đáp ứng cho DW, bao gồm:

- Bản thiết kế kiến trúc tổng thể của DW
- Phải mô tả các chức năng của từng thành phần cấu thành nên DW, cũng như sự liên quan của các thành phần với nhau.
- Các yêu cầu cụ thể về phần mềm, phần cứng và các tài nguyên mạng, phải có sự ước lượng về hiệu năng và kích cỡ của chúng.
- Phải xác định chiến lược để theo dõi và quản lý kỹ thuật cho toàn bộ dự án

2.3. Xác định các yêu cầu khai thác thông tin từ DW

Mục đích chính của việc xây dựng và phát triển DW là phục vụ cho vai trò khai thác thông tin của người dùng, vì vậy việc xác định được yêu cầu khai khác thông tin của người sử dụng là rất quan trọng và nó là căn cứ để đánh giá mức độ thành công của DW. Công việc này bao gồm các nội dung cơ bản sau:

- Xác định các dạng người sử dụng của DW
- Xác định cách thức để tập hợp các yêu cầu của người sử dụng
- Xác định yêu cầu khai thác thông tin của người sử dụng
- Xác định các công việc quản lý việc khai thác của người sử dụng

2.3.1. Các dạng người sử dụng DW

Có 04 dạng người sử dụng DW như sau:

- **Người sử dụng mức điều hành (Executives)**: những người này cần một một bảng giao diện mang đầy đủ thông tin họ cần, các dữ liệu họ cần thường là các dữ liệu ở mức tổng hợp cao, và dữ liệu hiển thị phải mang tính trực quan. Những người này thường chỉ khai thác các báo cáo đã được xây dựng sẵn.
- **Người sử dụng ở mức quản lý phòng ban bộ phận (Managers):** ở mức này người sử dụng cần bức tranh tổng thể cho bộ phận của họ và cũng cần thêm ở mức chi tiết hơn để làm rõ cho bức tranh tổng thể đó, các dữ liệu họ cần thường là cũng ở mức tổng hợp, họ có thể dùng các báo cáo đã xây dựng sẵn và có thể trực tiếp khai thác dữ liệu qua các công cụ OLAP cho báo cáo riêng của họ.
- **Người sử dụng ở mức phân tích (Business Analysts):** người này có nhiệm vụ phân tích dữ liệu kinh doanh để định hướng và xác định biện pháp thực

hiện. Công cụ khai thác của những người phân tích có thể là công cụ OLAP, các công cụ truy vấn đặc biệt, hoặc là các báo cáo được xây dựng sẵn. Phạm vi dữ liệu truy cập thì ở tất cả các mức cảu DW (cả mức chi tiết và tổng hợp)

- **Người sử dụng kỹ thuật (Technical Users):** thường là những người có kiến thức tốt về IT và đặc biệt là kỹ thuật về DW, công việc của họ liên quan đến tình trạng hoạt động và quản lý khai thác DW. Họ có trách nhiệm là hỗ trợ cho các dạng người sử dung trên (03 dạng) biết cách thực hiện khai thác dữ liệu một cách hiệu quả nhất, họ cũng có thể phải xây dựng ra các báo cáo chuẩn (ít thay đổi và có yêu cầu rỏ ràng).

2.3.2. Tập hợp yêu cầu của người sử dụng

Để có thể tập hợp được yêu cầu của người sử dụng thì phải tiếp cận tập trung vào các chủ đề, cụ thể là nên làm rỏ các vấn đề sau:

- Người sử dụng cần các nghiệp vụ nào và họ cần phải đưa ra được cái gì cho mỗi nghiệp vụ đấy.
- Những thuộc tính (trường) nào mà người dùng cần
- Sự phân cấp nghiệp vụ như thế nào
- những dữ liệu nào họ phải cần có và những dữ liệu nào mà họ thích có
- Công cụ truy cập đầu cuối (Front-End) là gì
- Người sử dụng muốn hiển thị kết quả khai thác như thế nào...

Trong quá trình tập hợp yêu cầu của người sử dụng có thể gặp phải các trở ngại sau:

- Mục đích nghiệp vụ của DW không được định nghĩa rỏ ràng
- Phạm vi của DW quá rộng
- Chưa phân biệt rõ chức đâu là chức năng của DW và đâu là chức của hệ thống tác nghiệp.

2.3.3. Yêu cầu khai thác thông tin của người sử dụng

Yêu cầu cho việc truy cập và khai thác thông tin từ DW là rất đa dạng và ở các mức độ sử dụng khác nhau:

- Mức báo cáo đơn giản
- Mức phân tích theo chiều hướng phức tạp
- Mức phân tích tổng hợp
- Mức phân tích đa chiều
- Báo cáo ngoại lệ
- Báo cáo theo kiểu dự đoán
- Báo cáo động cho phép nhập nhiều tham số
- Khai phá dữ liệu (Data Mining)

Yêu cầu công cụ khai thác thông tin từ DW cũng có thể dựa trên các cách thức và công cụ khác nhau:

- Phân tích báo cáo dựa trên môi trường Web hay Desktop hay cả hai....
- Báo cáo được liên kết với phần mềm văn phòng (bộ Office của Microsoft)
- Báo cáo tự động gửi qua Email, điện thoại di động, và các thiết bị cầm tay khác...

2.3.4. Quản lý khai thác thông tin của người sử dụng

Yêu cầu cho việc truy cập và khai thác thông tin của người sử dụng phải được quản lý cho vừa đảm bảo được tính dễ dàng cho người dùng nhưng vẫn vừa đảm bảo được tính an toàn và bảo mật của hệ thống. Không nên cho người sử dụng nhìn thấy và khai thác tất cả mọi thứ, mà nên phân quyền truy cập theo các chủ đề, sự phân quyền phải có hệ thống sao cho có thể dễ dàng kiểm soát được hoạt động khai thác thông tin từ DW.

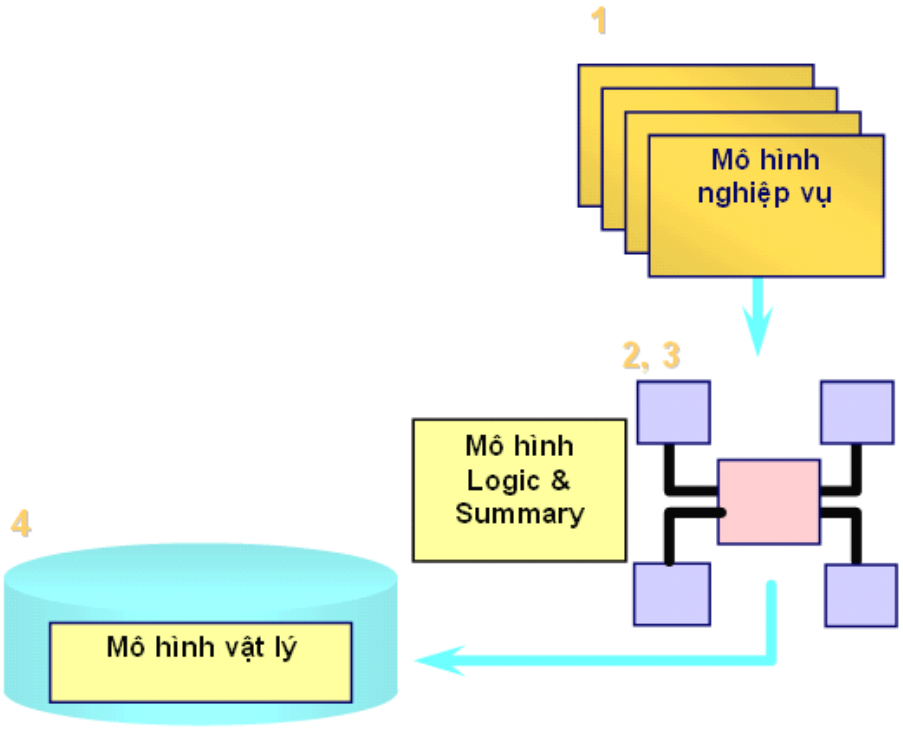
Việc quản lý quyền truy cập và khai thác thông tin có thể theo các mức sau:

- Quản lý truy cập theo từng chủ đề
- Quản lý truy cập theo từng phòng ban
- Quản lý truy cập theo từng người sử dụng
- Quản lý truy cập theo từng đối tượng (từng báo cáo)

2.4. Xây dựng mô hình DW

Việc xây dựng mô hình DW sẽ được thực hiện qua 04 bước cơ bản sau:

- Xác định mô hình nghiệp vụ (Conceptual Model)
- Tạo mô hình Logic (Logical Model)
- Tạo mô hình mức tổng hợp (Summary Model)
- Tạo mô hình vật lý (Phisical Model)



Hình 2.2. Xây mô hình DW

2.4.1. Xác định mô hình nghiệp vụ (Conceptual Model)

Việc xác định mô hình nghiệp vụ dựa trên các bước sau:

- Xác định các yêu cầu nghiệp vụ
- Xác định các đại lượng tính toán (measure) như số lượng, thành tiền, khuyến mãi...
- Xác định các chiều dữ liệu (Dimensions) như hàng hoá, khách hàng, kênh bán hàng, vùng miền, thời gian...
- Xác định các định nghĩa nghiệp vụ và các qui tắc nghiệp vụ
- Xác định các nguồn dữ liệu
- Nguồn dữ liệu nghiệp vụ chính có liên quan đến các nghieepj vụ cần cho DW
- Nguồn dữ liệu khác: dữ liệu từ bên ngoài, dữ liệu không phải dạng CSDL quan hệ...

2.4.2. Tạo mô hình logic (Logical Model)

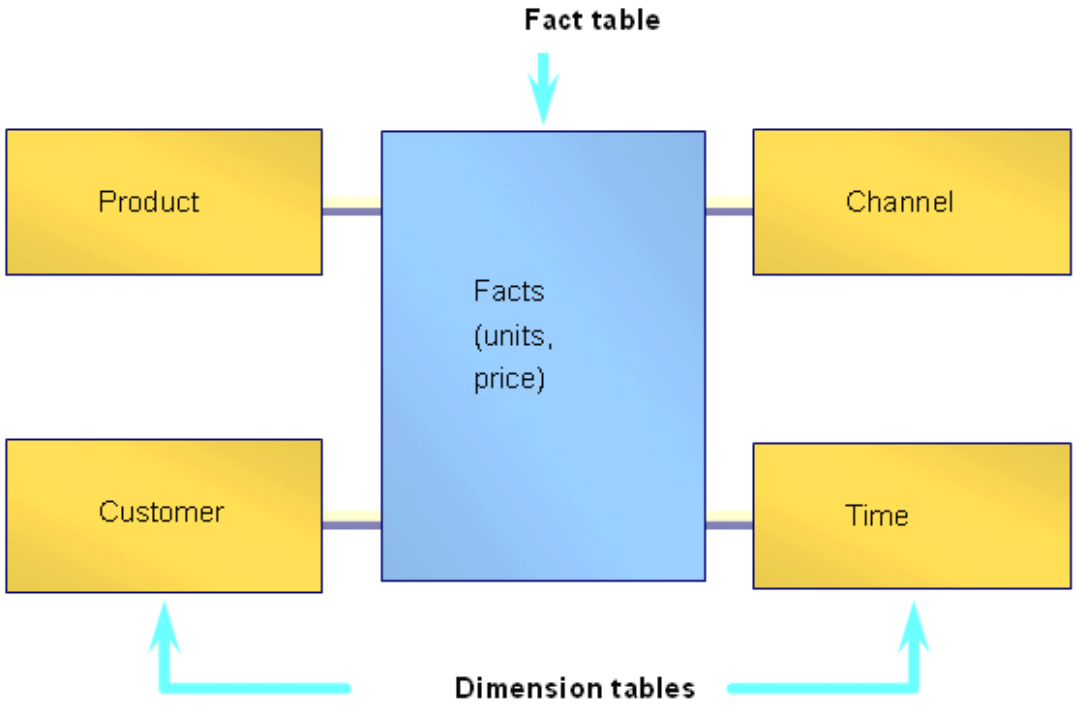
Sau khi xác định được các chủ đề cho DW thì cần xác định mô hình của DW. Có 03 loại kiểu mô hình thường dùng cho DW là: mô hình chuẩn hoá, mô hình phi chuẩn hoá và mô hình hình sao.

- **Mô hình chuẩn hoá:** tức là dữ liệu được tổ chức theo chuẩn, thường là theo chuẩn 3NF giống như khi tổ chức dữ liệu trong OLTP.

- **Mô hình phi chuẩn:** dữ liệu được lưu trữ đầy trong các bảng. Mục đích là phi chuẩn hoá các bảng để thêm hầu hết các cột được truy nhập vào một bảng chung để tránh việc kết hợp các bảng lại với nhau nhằm tăng tốc độ query và dễ dàng thực hiện query.
- **Mô hình hình sao**: dữ liệu được tổ chức thành các sơ đồ hình sao gồm có một bảng Fact nằm ở trung tâm và các bảng Dimension nằm ở xung quanh. Bảng fact chứa các đại lượng tính toán và các trường tham chiếu tới các bảng Dimension.

Đối với một DW tuỳ thuộc vào bản chất và khối lượng dữ liệu mà ta chọn các loại mô hình thích hợp. Một DW có thể sử dụng tất cả hoặc chỉ một loại trong 03 kiểu mô hình ở trên, nhưng thông thường đa số các DW (đặc biệt là ở thành phần Data Mart) ta đều sử dụng mô hình hình sao (Star Schema) để tổ chức dữ liệu. Theo mô hình hình sao thì các bước để tạo mô hình sẽ như sau:

- **Xác định bảng Fact:** bảng fact thường có các đặc tính sau:
  - Chứa các thuộc tính nghiệp vụ có dạng số (Metric)
  - Có thể chứa dữ liệu tổng hợp (Aggregated)
  - Có thể chứa dữ liệu ngày tháng
  - Các thuộc tính thượng có tính cộng được
  - Chứa các trường khoá ngoại đề tham chiếu đến khoá chính trong các bảng Dimension
  - Tổ hợp các trường các trường khoá ngoại này tạo nên khoá chính cho chính bảng Fact này.
- Xác định bảng Dimensions: bảng Dimension thường có các thuộc tính sau:
  - Là thuộc tính nghiệp vụ có dạng chuỗi
  - Là thuộc tính quan đến chiều thống kê
  - Liên kết với và các bảng fact
- Xác định liên kết giữ bảng fact và bảng Dimension
- Tạo ra các khung nhìn (view) cho người sử dụng



Hình 2.3. Mô hình hình sao (Star Schema)

Việc tổ chức dữ liệu theo mô hình hình sao thu được các ưu điểm sau:

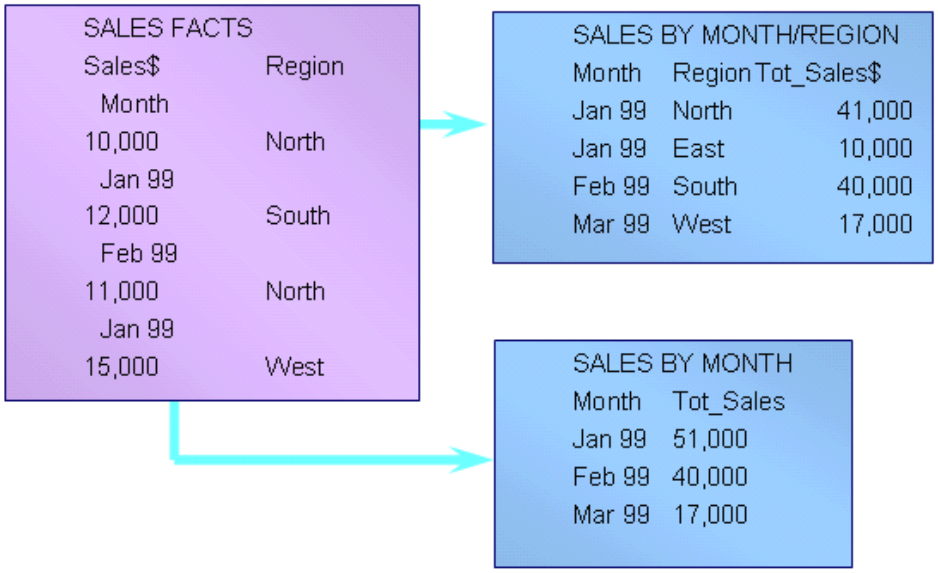
- Dễ hiểu cho người sử dụng
- Khả năng đáp ứng yêu cầu truy vấn (Query) nhanh
- Kho siêu dữ liệu đơn giản hơn
- Được hỗ trợ bởi nhiều công cụ đầu cuối (Front-End)

2.4.3. Tạo mô hình mức tổng hợp (Summary Model)

Mô hình này chứa các loại bảng dữ liệu sau:

- **Bảng tổng hợp (Summary table):** chứa dữ liệu tổng hợp ở mức cao (cao hơn bảng fact) thường là tổng hợp dữ liệu từ bảng fact theo một hoặc vài chiều.
- **Bảng tính trước (Pre-Calculated tables):** chứa dữ liệu đã được tính toán sẵn nhằm phục vụ cho mục đích khai thác (truy vấn, bán cáo hoặc phân tích) nào đó, thường dữ liệu cho các bảng bày không cho phép lên mức cao hơn nữa và để có được dữ liệu này phải thực hiện việc tính toán phức tạp (theo một công thức nghiệp vụ nào đấy).
- **Bảng kiểu chụp ảnh (Snapshot Tables):** các bảng dữ liệu này chứa dữ liệu gắn chặt với yếu tố thời gian, giống như việc chụp ảnh, tại các thời điểm khác nhau thì tập dữ liệu cũng khác nhau. Các bảng này thường chứa các tập dữ liệu được lặp lại theo các chu kỳ khác nhau như ngày, tuần, tháng năm....Ví

dụ: bảng chứa số dư của từng loại tiền theo ngày, bảng chứa số lượng khách hàng đang ở trạng thái Active theo từng vùng và theo từng ngày.....các bảng này được sinh ra tuỳ theo yêu cầu khai thác dữ liệu.



Hình 2.4. Ví dụ về bảng tổng hợp (Summary Table)

Mô hình mức tổng hợp cũng mang lại các ưu điểm sau:

- Cho phép truy cập nhanh các dữ liệu đã được tính trước (tổng hợp trước)
- Giảm yêu cầu về lưu lượng vào ra (I/O), tốc độ CPU và bộ nhớ

2.4.4. Tạo mô vật lý (Phisical Model)

Đây chính là bước chuyển đổi từ mô hình logic sang mô hình vật lý, tức là thực hiện cài đặt các bảng dữ liệu lên một cơ sở dữ liệu cụ thể. Các công việc phải làm trong bước này bao gồm:

- Định nghĩa qui ước đặt tên và các chuẩn qui định chung cho DW. Ví dụ: tên bảng Dimension thì bắt đầu bằng tiền tố DIM_ , Tên bảng fact thì bắt đầu bằng FACT_, Tên index thì bắt đầu bằng IDX_...
- Thực hiện chuyển đổi từ mô hình logic sang mô hình vật lý, bằng việc đặt tên theo qui ước cho các thuộc tính, xác định kiểu dữ liệu cụ thể, chiều dài của các trường... việc chuyển đổi này tuỳ thuộc vào việc DW được cài đặt trên hệ quản trị CDSL cụ thể nào.
- Thiết lập các Index: cho mục thực hiện query được nhanh hơn
- Thiết lập các Partition: cho mục đích query được nhanh và bảo trì DW sau này.

- Cấu hình tối ưu cho DW (Tuning): bằng cách thiết lập các tham số cho cho CSDL để tăng hiệu năng (Performance) thực hiện truy vấn và đảm bảo an toàn cho DW.
- Cấu hình cho DW chạy ở chế độ song song (parallel)...

2.5. Lập kế hoạch cài đặt vật lý

Việc cài đặt vật lý cho DW cũng cần được xem xét một cách kỹ lưỡng khi tiến hành xây dựng DW, nó bao gồm nhiều công việc khác nhau như chuẩn bị cơ sở hạ tầng mạng máy tính, máy chủ, thiết bị lưu trữ, thiết bị backup, thiết bị bảo mật... nhưng ta sẽ tập trung vào hai yếu tố quan trọng khi cài đặt vật lý:

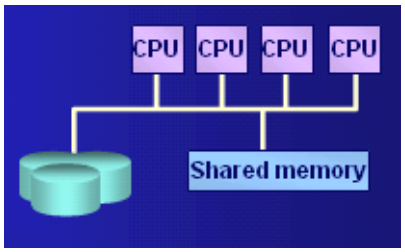
- Chon lựa kiến trúc tính toán cho DW
- Lập giải pháp lưu trữ cho DW

2.5.1. Chọn lựa kiến trúc tính toán (Phisical Model)

Việc chọn lựa kiến trúc tính toán sẽ tập trung vào việc:

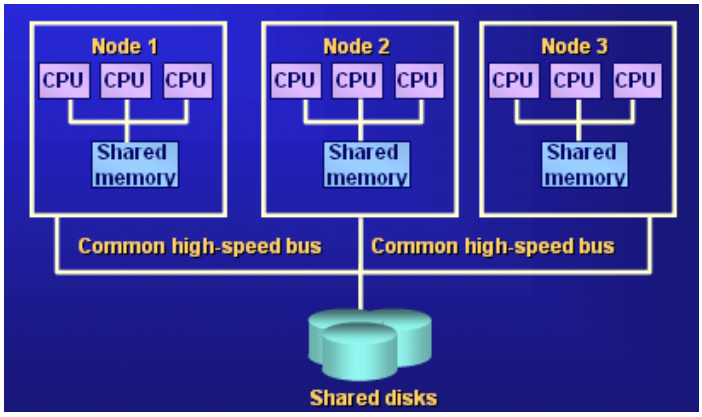
Xác định kiến trúc tính toán theo một trong các mô hình:

Mô hình SMP (Symmetric multiprocessing): tức là kiến trúc mà nhiều CPU trên cùng một máy chủ cùng chia sẽ một bộ nhớ và hệ thống đĩa.



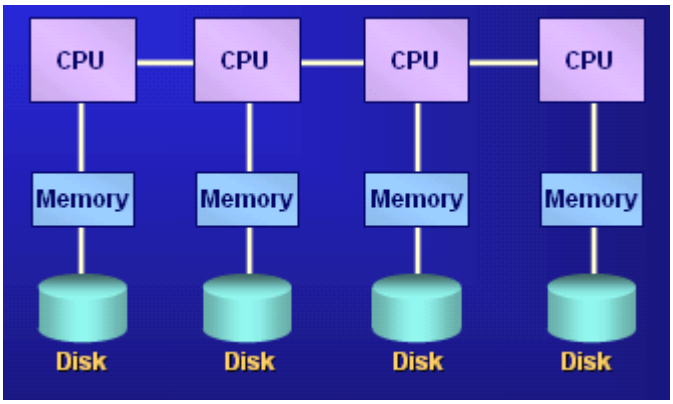
Hình 2.5. Mô hình SMP

Mô hình Cluster: là mô hình mà gồm có nhiều máy chủ được nối với nhau thành một khối thống nhất và cùng chia sẽ chung hệ thống đĩa ngoài, cùng tham gia vào xử lý các công việc với nhau. Ở góc độ người dùng có thể xem các máy chủ này tương đương như một máy chủ. Các máy chủ này thường được nối với nhau thông qua một thiết bị hỗ trợ băng thông rộng và nhanh (High-Speed). Ngày nay công nghệ Cluster đã phát triển rất mạnh và phát sinh ra các chuẩn cluster mới như HA (high-Availability Cluster), Load Balancing Cluster, HPC (High Performance Computing), Grid Computing.... mô hình này rất phù hợp cho DW vì có hệ thống đĩa dùng chung và việc tính toán được chia tải cho các máy chủ.



Hình 2.6. Mô hình Cluster

Mô hình MMP (Massively Parallel Processor (MPP): là mô hình mà gồm rất nhiều máy chủ nối lại với nhau không có hệ thống đĩa dùng chung, nhằm tạo ra một siêu máy tính tính toán song song rất mạnh để thực hiện các phép tính phức tạp. Mục đích chính của kiến trúc này là vẫn có thể tạo ra siêu máy tính có khả năng tính toán mạnh nhưng giá thành lại thấp hơn so với các hệ thống siêu máy tính khác.



Hình 2.7. Mô hình MPP

Xác định hệ điều hành:

- Windows: WinNT, Win2K Server, Win2K3 Server
- Unix:: Unixware, Linux, HP Unix, ýalỏic ....

Xác định hệ quản trị CSDL:

- Oracle
- DB2
- Infomix
- Sysbase
- Teradata

- MS SQL Server....

2.5.2. Lập giải pháp lưu trữ (Warehouse Storage)

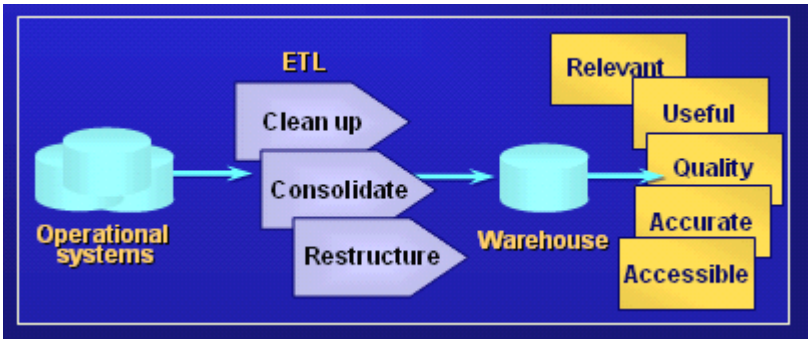
Thiết lập giải pháp lưu trữ bao gồm các công việc sau:

- **Lập kế hoạch tạo Partition cho các bảng có kích thước lớn:** việc tạo Partition phải cân nhắc nên tạo theo dọc (Vertical) hay ngang (Horizontal) cũng như theo kiểu phạm vi (Range), Băm (Hash), liệt kê (List) hoặc là kết hợp.
- **Lập kế hoạch tạo index:** xem xét nên lập Index cho các bảng nào, trường nào, đối với một index phải cân nhắc là nên chọn kiểu index là Btree-Index hay Bitmap Index. Đối với các bảng Fact thì thông thường hay sử dụng Bitmap-Index.
- **Lập kế hoạch an toàn cho hệ thống đĩa (RAID):** chọn mức RAID hợp lý cho hẹe thống đĩa (mức 0-5)
- **Lập kế hoạch lưu dự phòng (Backup) dữ liệu DW:** có nhiều giải pháp backup khác nhau như Offline hoặc Online, backup đầy đủ (full), backup một phần (một số bảng, một số Tablyspace, Schema, ...), backup chỉ những gì thay đổi (increamental)...

2.6 Xây dựng qui trình tích hợp dữ liệu cho DW

Sau khi có được các mô hình vật lý của DW thì ta sẽ tiến hành xây dựng qui trình tích hợp dữ liệu cho DW, nhiệm vụ chính của bước này là phải lấy dữ liệu nguồn, biến đổi dữ liệu nguồn thành dữ liệu có giá trị và tải nó vào dự liệu đích (DW). Dữ liệu được tải vào DW phải là dữ liệu phải đảm bảo các tính chất:

- Có liên quan (Relevant)
- Hữu dụng (Useful)
- Chất lượng (quality)
- Chính xác (Accurate)
- Sử dụng được (Accessible)



Hình 2.7. Qui trình tích hợp

Qui trình tích hợp được thực hiện tuần từ qua các bước sau:

- **Trích dữ liệu (Extract):** tiến hành đọc các dữ liệu nguồn một cách có chọn lọc, dữ liệu ở đây có thể là dữ liệu đang sử dụng cho tác nghiệp (Productive), dữ liệu đang được lưu trữ (Archive), dữ liệu từ bên ngoài tổ chức...
- **Biến đổi dữ liệu (Transform):** quá trình biến đổi dữ liệu có thể đơn giản hoặc phức tạp tuỳ thuộc và dữ liệu nguồn và dữ liệu đích. Nhưng thông thường ở bước này có thể chia ra thành các loại biến đổi như sau:
    - ***Làm sạch dữ liệu (Clean):*** tiến hành việc kiểm tra và sửa chữa các lỗi có thể có của dữ liệu để đảm bảo tính đúng đắn. Công việc này bao gồm các thao tác dọn dẹp, thay đổi và tính toán lại dữ liệu. Làm sạch dữ liệu liên quan đến các tác vụ sau: kiểm tra tất cả các trường đơn lẻ hoặc các trường liên kết chéo nhau, đưa ra và hợp nhất các bản ghi trùng nhau, sắp xếp lại các bản ghi....
    - ***Chuyển đổi dữ liệu (Transform) :*** do mô hình dữ liệu đích khác mô hình dữ liệu nguồn nên việc chuyển đổi phải qua các bước ánh xạ kiểu dữ liệu nguồn sang đích, chuẩn hoá, định dạng lại các trường dữ liệu, các phép biên đổi dự trên qui tắc nào đấy, phân tách một trường thành nhiều trường, tích hợp nhiều trường thành một trường...
    - ***Tích hợp (Integrate):*** Khi có nhiều nguồn dữ liệu thì cần phải được tích hợp lại để hợp nhất và tổ chức lại thông tin. Tiến trình tích hợp có thể là sự phối hợp các thao tác sau đây: sắp xếp - hợp nhất, chia cắt, giải quyết các vi phạm liên quan đến tính nguyên vẹn của dữ liệu, sinh ra các khoa tổng hợp...
- **Tải dữ liệu (Load):** tiến hành thêm mới hoặc nhập nhật dữ liệu đã được biến đổi vào các bảng trong kho dữ liệu đích. Quá trình tải dữ liệu cỏ thức hiện theo từng hàng (row) hoặc theo từng khối (Bulk)

2.6.1. Trích dữ liệu (Extract)

Đề trích dữ liệu cần quan tâm đến các bước sau:

- **Xác định dữ liệu nguồn để trích:** nguồn dữ liệu cho DW có thể ở các dạng sau:
    - ***Dữ liệu tác nghiệp (Production):*** tức là các dữ liệu hiện tại đang sử dụng từ hệ thống OLTP, đang dùng cho các phần mềm ứng dụng như CRM, ERP, SCM... đây là nguồn dữ liệu mà sau này theo chu kỳ ngày, tháng, năm... qui trình tích hợp sẽ trích dữ liệu thường xuyên, đây chính là nguồn dữ liệu để cập nhật mới nhất cho DW.
    - ***Dữ liệu lưu trữ (Archive):*** tức là các dữ liệu tác nghiệp trong quá khứ đã được lưu trữ lại, dữ liệu này sẽ được tải vào DW lần đầu tiên chạy

quá trình tích hợp (First Load). Việc tải dữ liệu này sẽ giúp cho DW mang tích lịch sử tốt hơn (dài hơn).

- ***Dữ liệu bên trong (Internal):*** tức là các dữ liệu bên trong tổ chức nhưng có tính rời rạc như các bảng tính Excel hay các văn bản...
- ***Dữ liệu bên ngoài (External):*** tức là các dữ liệu bên ngoài tổ chức nhưng có liên quan và cần thiết cho DW, nguồn dữ liệu này có thể có được quan việc trao đổi, mua bán, tìm kiếm...

- **Xác định cách thức tích hợp:** việc thực hiện tích hợp có thể theo các cách sau:
  - Sử dụng các ngôn ngũ lập trình cấp cao như C, C+, java, VB, Cobol... để viết ra các phần mềm tích hợp riêng cho tổ chức.
  - Sử dụng các các tiện ích đi kèm theo hệ QTCSDL như PL/SQL, T_SQL, Trigger, Sql Loader...
  - Mua các công cụ tích hợp có sẵn trên thị trường như Data Stage của IBM, Power Builder của Infomatica, Warehouse builder của Oracle, Data Integrator của Business Object...

2.6.2. Biến đổi dữ liệu (Transform)

Đề biến đổi dữ liệu liệu cần quan tâm đến các vấn đề sau:

- Trình tự thực hiện các bước chuyển đổi
- Làm sạch dữ liệu
- Loại bỏ dữ liệu lõi, dữ liệu dư thừa
- Thêm/tách các phần tử
- Trộn dữ liệu
- Tích hợp dữ liệu

Các vấn đề khi gặp phải và giải pháp trong quá trình biến đổi:

- **Khoá phức hợp:** tức là trong hệ thống tác nghiệp là sử dụng khoá là sự kết hợp của một số trường, khi đó ta phải tách mã này ra thành các thành phần cơ bản. Ví dụ: mã của một chi nhánh ngân hàng được đánh mã như sau xxxyynnnn trong đó:
  - xxx: mã hệ thống ngân hàng
  - yy: mã tĩnh thành
  - nnnn: là mã của chi nhánh

  => Nên tách thành 03 trường.
- **Nhiều cách mã hoá:** tức cùng một thuộc tính nhưng ở các nguồn khác nhau có cách mã hoá khác nhau. Ví dụ: cùng một trường về giới tính nhưng có các cách biểu diễn sau:
  - 1,0
  - M, F

- Male, Female

=>Nên qui tất cả về dạng M, F

- **Nhiều chuẩn khác nhau**: tức có sự khác nhau thì các chuẩn về đơn vị đo, về ngày tháng... Ví dụ về đơn vị đo chiều dài thì có thể là Inch hoặc Cm, dạng ngày tháng có thể là DD/MM/YYYY hoặc MM/DD/YYYY.

=> Nên qui các chuẩn này về một dạng duy nhất

2.6.3. Tải dữ liệu (Load)

Việc tải dữ liệu (load) vào DW chính là bước cập nhật nội dung của DW. Các vấn đề cần xem xét khi tải dữ liệu cho DW là:

- **Phương thức chuyển tải dữ liệu vào DW:** có 03 phương thức
  - Phương thức làm tươi (Refresh): không quan tâm đến dữ liệu cũ mà coi như xoá toàn bộ dữ liệu cũ và thêm dữ liệu mới nhất vào. Phương thức này phù hợp cho các bảng chứa dữ liệu nhỏ và không cần báo cáo lịch sử trên bảng này.
  - Phương thức bổ sung (Incremental): vẫn giữ nguyên tất cả dữ liệu cũ và thêm dữ liệu mới phát sinh vào, thường sử dụng thêm yếu tố thời gian vào khoá chính của các bảng để đảm bảo không bao giờ trung khoá. Ví dụ bảng chứa số dư tài khoản cuối ngày.
  - Phương thức kết hợp : tức vừa thêm dữ liệu mới nếu không trùng khoá vừa có thể cập nhật những dữ liệu cũ.
- **Lần tải dữ liệu:** thường có 02 loại tải dữ liệu
  - Tải dữ liệu lần đầu tiên (First-Load): thường chạy bằng tay và chỉ 01 lần đầu tiên khi bắt đầu đưa DW vào sử dụng.
  - Tải dữ liệu theo định kỳ: sau khi đã tải dữ liệu lần đầu tiên thì cần phải thiết lập quá trình tải dữ liệu theo định kỳ, tuỳ theo dữ liệu mà chu kỳ có thể là ngày, tháng, hoặc năm...việc tải dữ liệu theo định kỳ thường được thực hiện tự động theo lịch đặt trước.
- **Thời gian tải dữ liệu:** vì DW là kho dữ liệu rất lớn nên việc tải dữ liệu cũng cần phải cân nhắc liêu tốn hết bao thời gian để hoàn thành các tác vụ của nó. Cụ thể như sau:
  - Đối với tải việc tải lần đầu tiên thường thời gian yêu cầu dài nên phải được tính toán
  - Đối với tải định kỳ thì phải cân nhắc thời gian tải cho một định kỳ (Load Window) vì nó sẽ bị giới hạn trong một khoảng thời gian nhất định. Ví dụ: dữ liệu của ngân hàng cần được báo cáo vào lúc 7h sáng cho dữ liệu dịch ngày hôm trước và 9h tối là giờ đóng sổ thì thời gian được phép tải vào DW là sau 9h tối đến trước 7h sáng hôm sau. Nếu sau 7h sáng mà dữ liệu vẫn chưa tải hết vào DW thì báo cáo sẽ bị sai.

- **Tật tự tải dữ liệu cho các bảng:** tật tự tải các bảng cũng quan trọng và cần phải được thiết lập một cách rỏ ràng để dễ theo dõi và quản lý quá trình tải. Tật tự tải các loại bảng như sau:
    - Tải dữ liệu cho các bảng Dimension
    - Tải dữ liệu cho các bảng Fact
    - Tải dữ liệu cho các bảng Summary
    - Tải dữ liệu cho các bảng Snapshot

2.7. Quản trị DW

Sau khi DW được đưa vào sử dụng thì yêu cầu rất quan trọng là nó phải được quản lý và theo dõi thường xuyên, sao cho đảm bảo thông suốt cho người dùng khai thác thông tin hiệu quả nhất. Việc quản trị DW bao gồm các tác vụ sau:

- Quản lý về an toàn, bảo mật và độ ưu tiên
- Quản lý sự truy cập từ nhiều người khác nhau
- Kiểm tra chất lượng dữ liệu thường xuyên
- Kiểm tra quá trình tích hợp thường xuyên
- Quản lý và cập nhật kho siêu dữ liệu (Metadata)
- Giám sát và lập các báo cáo về tình hình sử dụng và trạng thái của DW như thời gian sử dụng, số người khai thác, thời gian đáp ứng các yêu cầu....
- Quản lý việc phân tán dữ liệu từ DW cho các mục đích bên ngoài.
- Quản lý qui trình lưu trữ dự phòng (Backup)
- Lập kế hoạch sẵn sàng phục hồi DW khi có sự cố
- Lập kế hoạch để nâng cấp và mở rộng (Hệ thống đĩa, RAM, băng thông...) cho sự gia tăng kích cỡ của DW theo thời gian.
- Lập kế hoạch lưu trữ bớt các dữ liệu cũ (không cần thiết cho việc khai thác) ra các thiết bị lưu trữ ngoài DW
- Lập kế hoạch mở rộng phạm vi dữ liệu của DW khi yêu cầu nghiệp vụ thay đổi.

BÀI TẬP

1. Hiểu thế nào về kho dữ liệu ? Quản lý kho dữ liệu ? – để cho học viên phát biểu trước khi đưa ra các khái niệm chính thức.(1 -2 học viên trả lời)
2. Nêu lý do mà bạn cho là các kho dữ liệu thường rất phức tạp ? Phương án để đơn giản hoá các chiều của kho dữ liệu ? (1-2học viên trả lời)
3. Các công việc của một người quản lý kho dữ liệu là gì? Tiêu chuẩn của người quản lý kho dữ liệu tốt là gì? (1-2 học viên trả lời)

BÀI 3

TÊN BÀI: KHAI THÁC KHO DỮ LIỆU

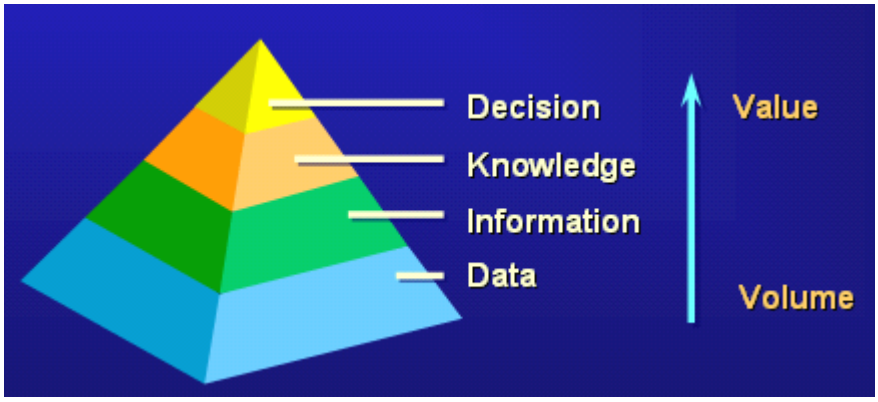
MÃ BÀI: ITPRG3_11.3

3.1. Tổng quan về khai thác thông tin từ DW

3.1.1. Mục đích của việc khai thác dữ liệu từ DW

Cái đích chính của việc xây dựng và phát triển DW là cung cấp thông tin cho các nhà quản lý tạo ra các quyết định chiến lược hiệu quả hơn. Ngày nay việc khai thác thông tin từ DW phục vụ cho mục đích hỗ trợ ra quyết định còn gọi là nghiệp vụ thông minh (BI – Business Intelligence).

Mục đích của việc khai thác dữ liệu từ DW là chuyển đổi một khối lượng lớn dữ liệu thành thông tin, các thông tin liên kết với nhau và biến thành kiến thức hỗ trợ cho việc tao quyết định. Quá trình chuyển đổi được thể hiện theo sơ đồ sau:



Hình 3.1. Sơ đồ chuyển đổi khối dượng dữ liệu thành "giá trị"

Quá trình chuyển đổi khối lượng dữ liệu thành giá trị phải chuyển qua các bước: Dữ liệu (Data) -> Thông tin (Information) -> Kiến thức (Knowledge) -> Quyết định (Decusion).

3.1.2. Các kỹ thuật khai thác DW

Kỹ thuật khai thác DW thường dựa trên kỹ thuật truy vấn đa chiều (Multidimension Query Techniques) bao gồm các kỹ thuật phân tích cơ bản như sau:

- **Slicing (cắt lát):** chính là việc giới hạn khung nhìn dữ liệu theo một nhiều từ các chiều trong số các chiều sẵn có. Ví dụ dữ liệu bán hàng trong bảng fact phụ thuộc vào 03 chiều là sản phẩm, vng và thi gian thì việc lấy dữ liệu theo chiều vùng cho tất cả sản phẩm và thời gian là một "lát cắt"..

- **Dicing (thái nhỏ)**: chính là việc cắt lát theo nhiều chiều khác nhau, và cũng có thể trong mỗi chiều lại bị giới hạn bởi các điều kiện.
- **Drilling (phân tích):** thực hiện phân tích dữ liệu theo nhiều hướng khác nhau kết quả có thể là tập dữ liệu tổng hợp tập dữ liệu chi tiết hơn tập dữ liệu đang xem xét. Việc phân tích cũng chia làm 03 loại.
    - ***Drilling Down (Phân tích chi tiết):*** cho phép xem xét dữ liệu ở mức chi tiết hơn so với mức hiện hành. Ví dụ: đang xem dữ liệu bán hàng theo từng tháng mà chọn Drill Down hì dữ liệu chi tiết đến từng ngày bán hàng cho tháng vừa chọn sẽ hiện ra. Chú ý: thứ tự phân cấp về thời gian <Năm -> Tháng -> ngày> phải được định nghĩa trước.
    - ***Drilling Up (Phân tích tổng hợp):*** cho phép xem xét dữ liệu ở mức tổng hợp hơn so với mức hiện hành. Ví dụ: đang xem dữ liệu bán hàng theo từng tháng mà chọn Drill Up thì dữ liệu sẽ tổng hợp đến từng năm bán hàng.
    - ***Drilling Across (Phân tích chéo):*** tức là đang phân tích dữ liệu theo một cây phân cấp (Hierarchy) lại chuyển sang kết hợp theo một thuộc tính thuộc cây phân cấp khác. Ví dụ: đang xem dữ liệu bán hàng theo mức tháng (theo cây phân cấp theo yếu tố thời gian) thì ta có thể chọn thêm chiều tỉnh thành (Theo cây phân cấp địa lý Vùng -> Tỉnh -> Huyện) để xem dữ liệu liệt kế
- **Pivoting (Xoay chiều):** là kỹ thuật thay đổi trục theo dữ liệu, cho phép ta thay đổi các hàng và cột cho nhau trong một báo cáo dạng bảng (Tabular), nó cho phép người sử dụng có thể nhìn theo nhiều chiều khác nhau mà không cần phải chạy lại truy vấn dữ liệu (requering) cho nó.

3.2. Công cụ khai thác dữ liệu DW

Để khai thác dữ liệu DW thì có thể sử dụng các loại công cụ khai thác dữ liệu sau:

- Công cụ báo cáo (Reporting tools)
- Công cụ truy vấn (Query tools)
- Công cụ phân tích báo cáo trực tuyến (OLAP)
- Bộ công cụ phân tích (Analytical suites)
- Khai phá dữ liệu (Data mining)
- Các ứng dựng phân tích (Analytical application)

3.2.1. Công cụ báo cáo

Công cụ báo cáo là công cụ cho phép người sử dụng tạo tạo ra các báo cáo theo nhiều dạng khác nhau như bảng ngang, bảng dọc, đồ thị, và Pivot. Công cụ báo có thể được kết hợp với một ngôn ngũ lập trình cấp cao như VB, Java, Cobol...để đưa ra các báo cáo tác nghiệp mà đòi hỏi xử lý tính toán phức tạp và theo khối lượng lớn. Công cụ báo cáo

cũng có thể được dùng trực tiếp bởi người dùng đầu cuối như Crystal report, Dynamic report... những công cụ này cho phép người sử dụng đầu cuối có thể tự thiết kế và tạo báo cáo cho họ mà không cần sự hỗ trợ của cán bộ tin học (đương nhiên là họ đã được đào tạo về cách sử dụng công cụ trước đấy). nhuwngx công cụ báo cáo thường có giao diện đồ hoạ hỗ trợ nhiều dạng báo cáo khác nhau, nhiều kiểu định dạng khác nhau và cho phép kết nối đến nhiều lại cơ sở dữ liệu khác nhau như Oracle, Informix, SQL Server... ngày nay công cụ báo cáo không chỉ dừng lại ở mức ứng dụng nên Desktop mà còn phổ biến cả trên nền Web và là một phần bắt buộc của bộ công cụ phân tích hoặc bộ sản phẩm OLAP.

3.2.2. Công cụ truy vấn

Đây là công cụ cho phép người sử dụng truy cập DW lấy ra các thông tin cần thiết để trả lời cho các câu hỏi đột xuất (Ad hoc query). Bản chất của các công cụ này là đều sinh ra ngôn ngữ SQL để truy cập dữ liệu, những công cụ này thường làm đơn giản hoá việc truy vấn cho người sử dụng bằng việc sử dụng lớp ngữ nghĩa (semantic layer) là trung gian giữa người sử dụng đầu cuối và cơ sở dữ liệu.

Lớp ngữ nghĩa chính là tập hợp các đối tượng hướng nghiệp vụ được định nghĩa theo từng chủ để nghiệp vụ, nó sử dụng các thuật ngữ nghiệp vụ đặt tên có các thuộc tính của đối tượng và tích hợp nhiều thuộc tính của liên quan với nhau vào một đối tượng, một đói tượng có thể ánh xạ dên nhiều bảng dữ liệu trong DW.

Khi sử dung công cụ truy vấn để tạo truy vấn người sử dụng chỉ việc chỉ ra đối tượng (theo chủ đề) cần lấy thông tiên và sau đó thực hiện việc kéo và thả các thuộc tính thì sẽ thu được kết quả như mong muốn, công cụ truy vấn sẽ tự biên dịch và sinh ra câu lệnh SQL tương ứng.

3.2.3. Công cụ phân tích trực tuyến (OLAP)

Bản chất của OLAP là dữ liệu được lấy ra từ DW sẽ được chuyển thành mô hình đa chiều và được lưu trữ trong một kho dữ liệu đa chiều (dữ liệu được lưu trữ theo mảng thay vì như mô hình quan hệ) giúp cho việc khai thác thông tin được nhanh hơn rất nhiều. Do trong DW chủ yếu dữ liệu dành cho khai thác được tổ chức theo mô hình hình sao, mô hình đã mang tính nhiều chiều nên rất thuận lợi cho việc cài đặt OLAP.

OLAP có thể xem là một chức năng thông minh, làm cho các thông tin trong công ty có thể hiểu được, giúp cho người dùng đầu cuối có thể hiểu được bản chất bên trong thông qua việc truy cập nhanh và tương tác với khung nhìn theo nhiều dạng khác nhau.

3.2.4. Bộ công cụ phân tích

Bộ công cụ phân tích là một bộ công cụ truy vấn, báo cáo và phân tích chạy trên một máy chủ ứng dụng mạnh và trên mô hình Web. Hay nói cách khác là bộ công cụ phân tích phải tích hợp các công cụ truy vấn, báo cáo và phân tích vào thanh một công cụ.

Thông thường với một bộ công cụ phân tích ngoài các chức năng truy vấn , báo cáo và phân tích nó còn có một mô đun cổng nghiệp vụ (Business Portals), nơi để đưa các báo cáo, các kết quả truy vấn, các kết quả phân tích cho nhiều người cùng sử dụng. Cũng là nơi để tất cả người sử dụng truy cập vào ẩntước khi có thể thực hiện các tác vụ khác như tạo truy vấn, tạo báo cáo, phân tích...

Bộ công cụ phân tích thường phải có một siêu dữ liệu (Repository) để chứa các thông tin mà người dùng định nghĩa ra, và luôn có mô đun quản lý kho siêu dữ liệu để người quản trị dễ dàng theo dõi người sử dụng, phân phối hay xoa bor các đối tượng báo cáo...

3.2.5. Khai phá dữ liệu (Data Mining)

Data Mining là công cụ xác định các hình mẫu và mối quan hệ của dữ liệu có lợi cho việc xây dựng mô hình hỗ trợ ra quyết định. khai phá dư liệu được xem là việc khám phá tri thức trong các cơ sở dữ liệu, là một quá trình trích xuất những thông tin ẩn, trước đây chưa biết và có khả năng hữu ích, dưới dạng các qui luật, ràng buộc, qui tắc trong cơ sở dữ liệu. Nói tóm lại, khai phá dữ liệu là một quá trình học tri thức mới từ những dữ liệu đã thu.

Một quá trình khai phá dư liệu bao gồm năm giai đoạn chính sau:

- Tìm hiểu nghiệp vụ và dữ liệu
- Chuẩn bị dữ liệu
- Mô hình hóa dữ liệu
- Hậu xử lý và đánh giá mô hình
- Triển khai tri thức

Quá trình này có thể được lặp lại nhiều lần một hay nhiều giai đoạn dựa trên phản hồi từ kết quả của các giai đoạn sau.

3.2.6. Ứng dụng phân tích (Analytical Application)

Có hai loại ứng dụng phân tích sau:

Ứng dụng phân tích đóng gói (Packaged): đó là các ứng dụng mà các quá trình trích và biến đổi cho dữ liệu nguồn là được định nghĩa trước, mô hình dữ liệu cũng đã được tạo sẵn, cung cấp sản các mẫu báo cáo, và một giao diện đầu cuối có thể tuỳ biến được.

Ứng dụng phân tích tuỳ biến (custom Analytic Application): ứng dụng này cho phép nhà phát triển có thể dễ dàng tạo dược một ứng dụng phân tích riêng thông qua việc chọn lựa các thành phần với nhau. Các thành phần bao gồm các thành phần giao diện sử dụng, thành phần truy cập dữ liệu, thành phần phân tích và tập các mẫu báo cáo.

3.3. Xử lý phân tích trực tuyến (OLAP)

3.3.1. Tại sao phải xử lý phân tích trực tuyến

Trong các kho dữ liệu lớn vá đa chiều thường chứa nhiều thông tin ẩn mà công cụ truyền thống như sử dụng SQL rất kho phát hiện được. Ví dụ: lãnh đạo một công ty nghiên

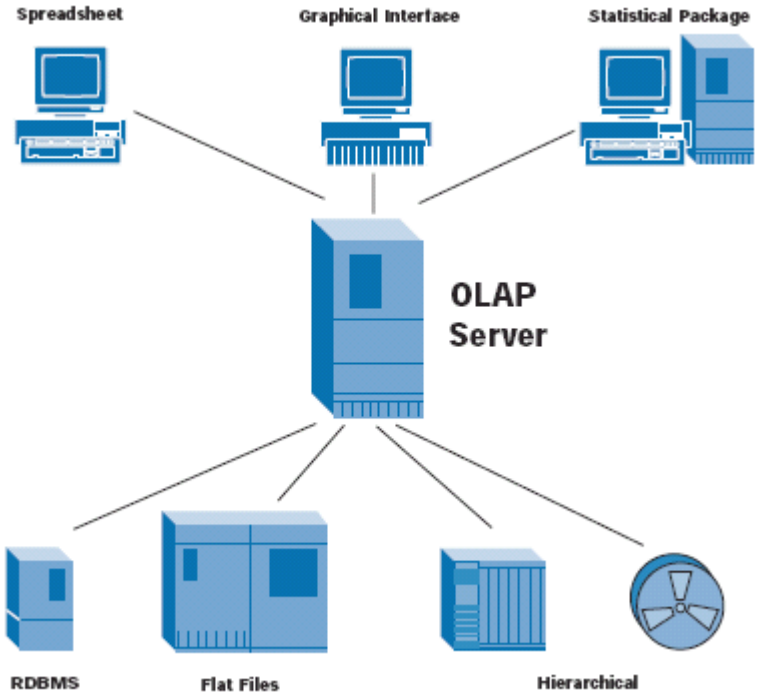
cứu về thị trường muốn biết mặt hàng nào bán chạy nhất trong tháng 12 của vùng núi tây bắc theo các lứa tuổi khác nhau" . Đây là câu hỏi có 04 chiều liên quan là mặt hàng, vùng, lứa tuổi và thời gian. Không hề dễ dàng để có được câu trả lới cho những câu hỏi nhiều chiều như trên nếu như sử dụng ký thuật truy vấn trực tiếp vào nguồn dữ liệu bằng các lệnh SQL. Hơn thế nữa, do yêu cầu người lãnh đạo lại đưa ra các yêu cầu thống kê theo các tiêu chí khác nhau khi thì theo lứa tuổi, khi thì theo vùng, khi thì theo tháng... hoặc là kết hợp giữa hai hay ba chiều lại với nhau... điều này sẽ rất vất vả để người trả lới câu hỏi và mất rất nhiều thời gian để có được kết quả nếu cũng chỉ dùng truy vấn trúc tiếp đến nguồn dữ liệu. Do đó để đáo ứng yêu cầu phân tích số liệu trên các kho dữ liệu lớn như DW thì cần phải áp dụng kỹ thuật xử lý trực tuyến nhằm đáp ứng được yêu cầu về tốc độ trả lời câu hỏi và báo cáo thống kê.

3.3.2. Phân biệt kho dữ liệu quan hệ và kho dữ liệu đa chiều

Kho dữ liệu quan hệ: là kho dữ liệu mà lưu trữ dữ liệu như các bản ghi có khoá và dữ liệu được truy cập bởi ngôn ngữ truy vấn dữ liệu SQL.

Kho dữ liệu đa chiều: là kho dữ liệu mà dữ liệu được lưu trữ trong các mảng (chứa dữ liệu cùng kiểu). Vì vậy mà không có chuẩn chung về mô hình đa chiều, không có phương pháp chuẩn để truy cập được dữ liệu từ kho dữ liệu đa chiều. Một vài sản phẩm hổ trợ các giao diện lập trình (API) hay thiết bị bảng tính đầu cuối để truy cập dữ liệu của kho đa chiều.

3.3.3. Định nghĩa OLAP



Hình 3.2. Mô hình tổng thể của OLAP

OLAP là một công nghệ xử lý trực tuyến các thông tin mới được tạo ra từ những dữ liệu đang tồn tại, thông qua một tập các chuyển đổi và tính toán số. Về bản chất, một hệ OLAP là hệ thống lưu giữ những thông tin tổng hợp và cho phép thể hiện thông tin tổng hợp đó dưới dạng bảng hai chiều.

OLAP là công nghệ phân tích dữ liệu thực hiện những công việc sau:

- Đưa ra một khung nhìn logic, nhiều chiều của dữ liệu trong DW, khung nhìn này hoàn toàn không phụ thuộc vào dữ liệu được lưu trữ thế nào (nó có thể được lưu trữ trong một kho dữ liệu nhiều chiều hay một kho dữ liệu quan hệ)
- Thường liên quan đến những truy vấn phân tích tương tác dữ liệu. Sự tương tác thường là phức tạp yêu cầu phân tích dữ liệu xuống mức chi tiết hợp (Drill Down) hoặc tổng hợp dữ liệu lên mức cao hơn (Drill Up).
- Cung cấp khả năng thiết lập mô hình phân tích bao gồm một mô tơ tính toán cho việc tính tỉ lệ biến đổi liên quan đến những đại lượng số hoặc dữ liệu dạng số qua nhiều chiều.
- Tạo ra sự tổng hợp và kết hợp, phân cấp và dùng những mức tổng hợp, kết hợp cho mỗi phép giao của các bảng theo chiều.
- Hỗ trợ mô hình chức năng cho việc dự báo, phân tích các xu hướng và phân tích thông kê.
- Lấy và hiển thị dữ liệu theo những bảng 2 hay 3 chiều, theo biểu đồ hay đồ thị, dễ dàng xoay đổi các chiều cho nhau. Khả năng xoay là quan trọng vì mỗi người sử dụng cần phân tích dữ liệu từ các cách nhìn khác nhau và sự phân tích theo mỗi cách nhìn sẽ dẫn đến một câu hỏi khác, câu hỏi này sẽ được kiểm tra tính đúng đắn dựa trên một cách nhìn khác về dữ liệu đó.
- Đáp ứng các câu trả lời nhanh, vì vậy quá trình phân tích không bị cắt ngang và thông tin không bị cũ.
- Sử dụng một mô tơ kho dữ liệu đa chiều, lưu trữ dữ liệu theo các mảng (lưu ý là măng lưu trữ những phần tử cùng kiểu khác với bản ghi là các phanàv tử có kiểu khác nhau). Những mảng này là sự biểu diễn logic của các chiều công việc.

Thuật ngữ OLAP và cơ sở dữ liệu đa chiều hay được đồng nhất, gây nên sự mập mờ xung quanh hai khái niệm này. Bản chất của cơ sở dữ liệu đa chiều là một kiến trúc cơ sở dữ liệu lưu trữ thông tin tổng hợp bao gồm tất cả các mục dữ liệu chính (hay còn gọi là các chiều) tham chiếu lẫn nhau. Trong khi đó OLAP là một thể hiện ra bên ngoài cho người sử dụng lựa chọn các chiều và các sự kiện tham chiếu lẫn nhau. Các nguồn dữ liệu cho một ứng dụng OLAP bao gồm cơ sở dữ liệu quan hệ, các bảng tính và cả cơ sở dữ liệu đa chiều.

3.3.4. Kiến trúc của OLAP

Kiến trúc của OLAP được xem xét trên 02 khía cạnh logic và vật lý:

3.3.4.1. Kiến trúc Logic của OLAP

Kiến trúc logic của OLAP gồm có 02 thành phần:

- **Khung nhìn của OLAP:** là sự biểu thị logic và đa đa chiều của dữ liệu đối với người sử dụng, không liên quan đến việc dữ liệu được lưu trữ như thế nào và ở đâu.
- **Kỹ thuật lưu trữ dữ liệu**: là cách lựa chọn lưu trữ dữ liệu như thế nào và lưu trữ dữ liệu ở đâu. Có hai cách thông dụng nhật là lưu trữ trong kho dữ liệu đa chiều và kho dữ liệu quan hệ.

Nếu xét về chức năng của các thành phần cấu thành nên OLAP thì có thể chia làm 03 thành phần:

- Các dịch vụ lưu trữ dữ liệu
- Các dịch vụ bên trong của OLAP
- Các dịch vụ hỗ trợ cho người dùng đầu cuối

Chú ý: Người sử dụng chỉ quan tâm tới khung nhìn dữ liệu đa chiều và một mức thể hiện chấp nhận được. Còn những người cung cấp thông tin thì quan tâm đến việc dữ liệu được lưu trữ ở đâu, lưu trữ thế nào, tốc độ truy cập có chấp nhận được không, và khả năng quản lý nó.

3.3.4.2. Kiến trúc vật lý của OLAP

Kiến trúc vật lý của OLAP phân thành 02 loại cơ bản dựa trên kỹ thuật lưu trữ dữ liệu của OLAP server là trên kho dữ liệu đa chiều hay kho dữ liệu quan hệ.

- **Dựa trên kho kho dữ liệu đa chiều:** kho dữ liệu nằm trên server OLAP, tách biệt với kho dữ liệu DW. Loài này được chia làm 02 loại nhỏ sau:
  - ***Loại thứ nhất:*** Kho dữ liệu đa chiều được lưu trữ trên máy trạm Client do đó thường xảy ra tình trạng tắt nghẽn (nút cổ chai) trên mạng khi dữ liệu được tải vào các máy trạm. Một ảnh hưởng không tốt nũa là vấn đề hiệu suất và an toàn dữ liệu.
  - ***Loại thứ hai:*** Kho dữ liệu đa chiều và các dịch vụ OLAP được thiết kế kết hợp với nhau trên một máy chủ, hoặc kho dữ liệu đa chiều đạt tại một nơi khác với server OLAP khi kho dữ liệu đa chiều này có kích thước lớn.
- **Dựa trên kho dữ liệu quan hệ:** tổ chức lưu trữ dữ liệu OLAP nằm luôn trong kho dữ liệu DW (nhưng đã sử dụng thêm các công nghệ cho phép cache, tính toán trước và thực hiện truy vấn tối ưu) và máy chủ OLAP nằm riêng.

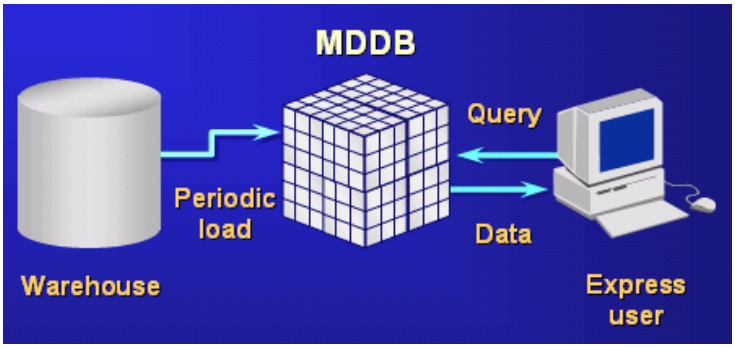
3.3.5. Phân loại OLAP

Dựa trên kiến trúc vật lý của OLAP ta có thể phân biệt OLAP thành 03 loại sau:

- MOLAP (Multidimensional OLAP): OLAP dựa trên cơ sở dữ liệu đa chiều.
- ROLAP (Relational OLAP): OLAP dựa trên cơ sở dữ liệu quan hệ
- HOLAP (Hybrid OLAP): OLAP kết hợp của MOLAP và ROLAP

3.3.5.1 MOLAP

Với kiến trúc này thì kho dữ liệu đa chiều và các dịch vụ của OLAP trên cùng một Server và dữ liệu da chiều của MOLAP được lấy từ DW.



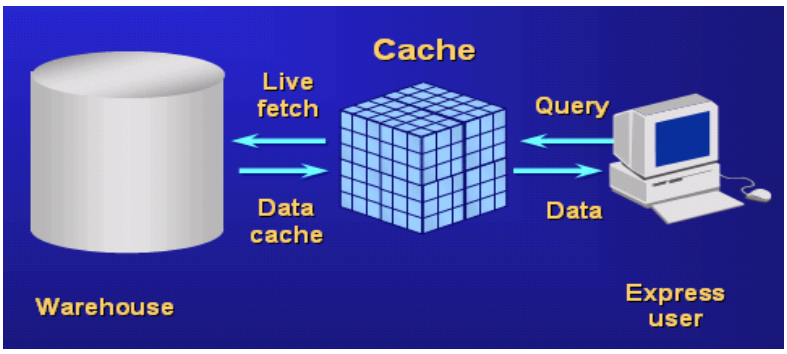
Hình 3.3. MOLAP

MOLAP thường được sử dụng cho các ứng dụng có các đặc điểm sau:

- Yêu cầu tốc độ truy vấn cao
- Yêu cầu phân tích phức tạp
- Yêu cầu tính dễ sử dụng cho người sử dụng chỉ cần qua tâm đến các dữ liệu tổng hợp hoặc tính toán trước theo nhiều chiều
- Chỉ yêu cầu phân tích trên các dữ liệu tổng hợp hoặc dữ liệu đã được tính trước.

3.3.5.2. ROLAP

Với kiến trúc này thì Server OLAP chỉ chứa các dịch vụ của OLAP và cung cấp một mô tơ truy vấn cực kỳ linh động kết hợp với công nghệ bộ đệm (Cache) tất cả các dữ liệu tạo điều kiện cho người dùng đầu cuối dễ dàng trích và tổng hợp dữ liệu theo yêu cầu.



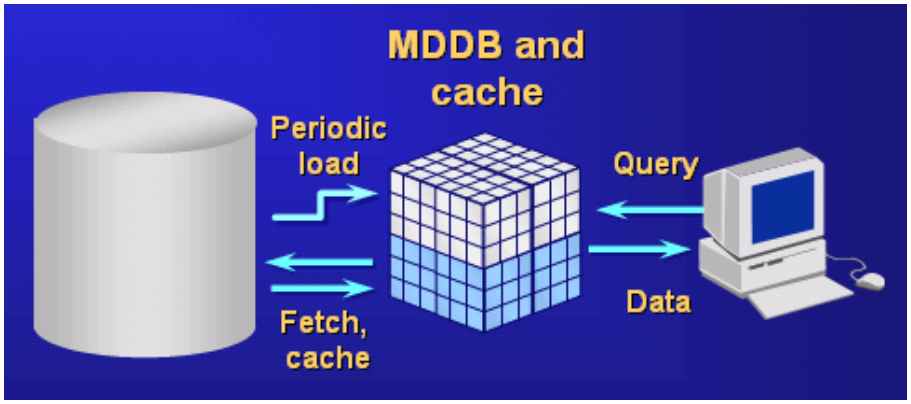
Hình 3.4. ROLAP

ROLAP thường được sử dụng cho các ứng dụng có các đặc điểm sau:

- Dữ liệu thường xuyên thay đổi về cấu trúc (thay đổi về số chiều).
- Khối lượng dữ liệu lớn (có thể lên đến hàng terabyte).
- Các dạng truy vấn thường không được xác định trước.

3.3.5.3. HOLAP

Với kiến trúc này là sự kết hợp giữa MOLAP và ROLAP.



Hình 3.5. HOLAP

Bài tập:

1. Công cụ khai thác dữ liệu DW
2. Xử lý phân tích trực tuyến (OLAP)

- . Marx Gómez, C. Rautenstrauch, P. Cissek, B. Grahlher: *Einführung in SAP Business Information Warehouse.* Springer, Berlin, März 2006, ISBN 3-540-31124-6
- William H. Inmon, Richard D. Hackathorn: *Using the Data Warehouse*, John Wiley & Sons, ISBN 0-471-05966-8
- Andreas Bauer, Holger Günzel: *Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung*, dpunkt, ISBN 3-898642-51-8
- Christian Mehrwald: *Datawarehousing mit SAP BW 3.5 - Architektur, Implementierung, Optimierung*, dpunkt, ISBN 3-89864-331-X
- Reinhard Jung, Robert Winter: *Data Warehousing Strategie*, Springer, ISBN 3-540-67308-3
- Thomas Zeh: *Data Warehousing als Organisationskonzept des Datenmanagements. Eine kritische Betrachtung der Data-Warehouse-Definition von Inmon.* In: *Informatik. Forschung und Entwicklung.*, Band 18, Heft 1, Aug. 2003
- Ralph Kimball, Mary Ross: *The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling.*, John Wiley & Sons, ISBN 0-471-20024-7
- Barry Devlin: *Data Warehouse. From Architecture to Implementation.*, Addison-Wesley, ISBN 0-201-96425-2
- Wolfgang Lehner: *Datenbanktechnologie für Data-Warehouse-Systeme. Konzepte und Methoden.*, dpunkt, ISBN 3-89864-177-5
- Alex Schweizer: *Data Mining, Data Warehousing. Datenschutzrechtliche Orientierungshilfen für Privatunternehmen.*, Orell Füssli, ISBN 3-280-02540-0
- Jan Holthuis: *Der Aufbau von Warehouse-Systemen, Konzept - Datenmodellierung - Vorgehen*, Deutscher-Universitäts-Verlag, ISBN 3-8244-6959-6
- Markus Lusti: *Data Warehousing and Data Mining: Eine Einführung in entscheidungsunterstützende Systeme*, Springer, ISBN 3-540-42677-9
- Eitel von Maur, Robert Winter: *Data Warehouse Management: Das St. Galler Konzept zur ganzheitlichen Gestaltung der Informationslogistik. Metadaten, Datenqualität, Datenschutz, Datensicherheit*, Springer, ISBN 3-540-00585-4
- Caroline Wilmes, Helmut M. Dietl, Remco van der Velden: *Die strategische Ressource "Data Warehouse": Eine ressourcentheoretisch empirische Analyse*, Deutscher Universitätsverlag, ISBN 3-8244-8046-8
- Heiko D. Schinzer, Carsten Bange, Holger Mertens: *Data Warehouse und Data Mining: Marktführende Produkte im Vergleich*, Vahlen, ISBN 3-8006-2466-4
- Reinhard Schütte: *Data Warehouse Managementhandbuch: Konzepte, Software, Erfahrungen*, Springer, ISBN 3-540-67561-2
- Gunnar Auth: *Prozessorientierte Organisation des Metadatenmanagements für Data-Warehouse-Systeme*, Books on Demand, ISBN 3-8334-1926-1
- Katharina Wirtz: *Der Data-Warehouse-Rahmenplan: Entwicklung eines konzeptionellen Schemas*, Deutscher Universitätsverlag, ISBN 3-8244-7621-5
- Michael Böhnlein: *Konstruktion semantischer Data-Warehouse-Schemata*, Deutscher Universitätsverlag, ISBN 3-8244-2148-8
- Eitel von Maur, Robert Winter: *Vom Data Warehouse zum Corporate Knowledge Center*, Physica-Verlag, ISBN 3-7908-1536-5
- J.-H. Wieken: *Der Weg zum Data Warehouse*,Addison-Wesley, ISBN 9-783827-315601