

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: BẠC CAO



Đà lạt -2007

Tuyên bố bản quyền :

Tài liệu này thuộc loại sách giáo trình
Cho nên các nguồn thông tin có thể
được

phép dùng nguyên bản hoặc trích
dùng cho các mục đích về đào tạo và tham
khảo .

Mọi mục đích khác có ý đồ lệch lạc
hoặc

sử dụng với mục đích kinh doanh
thiếu lành

mạnh sẽ bị nghiêm cấm.

Tổng Cục Dạy nghề sẽ làm mọi cách
để bảo vệ bản quyền của mình.

Tổng Cục Dạy Nghề cảm ơn và hoan
nghênh các thông tin giúp cho việc tu sửa
và hoàn thiện tốt hơn tài liệu này.

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 được 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

TÊN BÀI	TRANG
1. CÁC KHÁI NIỆM VỀ KHO DỮ LIỆU	9
1.1. Nhập môn về kho dữ liệu	9
1.1.1. kho dữ liệu – Data warehouse	9
1.1.2. mục đích của kho dữ liệu	10
1.1.3. đặc tính dữ liệu trong kho dữ liệu	11
1.1.3.1. hướng chủ đề (Subject oriented)	11
1.1.3.2. tích tích1 hợp(Integrated)	12
1.1.3.3. tính ổn định (Non volatile)	13
1.1.3.4. tính lịch sử (time Variant)	13
1.1.4. phân biệt DW với OLTP	13
1.1.5. một số loại kho dữ liệu khác	14
1.1.5.1. Kho dữ liệu cục bộ - Datamart	14
1.1.5.2. kho siêu dữ liệu (metadata)	15
1.1.5.3. Kho dữ liệu tác nghiệp (ODS)	17
1.2. Kiến trúc kho dữ liệu	18
1.2.1. Nguồn dữ liệu (Data Source)	19
1.2.2. Thành phần kho dữ liệu (Data Qarehouse)	19
1.2.3. Thành phần Data mart	19
1.2.4. Thành phần tích hợp (ETL)	20
1.2.5. Thành phần kho chức siêu dữ liệu(Repository)	21
1.2.6. Thành phần vùng chứa tạm (Staging Area)	21
1.2.7. Thành phần khai thác cho người sử dụng (users)	22
2. XÂY DỰNG VÀ PHÁT TRIỂN KHO DỮ LIỆU (DW)	23
2.1. Tổng quan về xây dựng phát triển DW	23
2.2. Lập kế hoạch xây dựng và phát triển DW	24
2.2.1. Lập kế hoạch tài chính	24
2.2.2. lập kế hoạch nghiệp vụ	24
2.2.3. lập kế hoạch về kĩ thuật	25
2.3. xác định các yêu cầu khai thác thông tin từ DW	25
2.3.1. các dạng người sử dụng Dw	25
2.3.2. Tập hợp yêu cầu người sử dụng	26
2.3.3. yêu cầu khai tahc1 thông tin của người sử dụng	26
2.3.4. quản lý khai thác thông tin của người sử dụng	27
2.4. Xây dựng mô hình DW	27
2.4.1. Xác định mô hình nghiệp vụ (Conceptual Model)	28
2.4.2. Tạo mô hình logic (logical model)	28
2.4.3. tạo mô hình mức tổng hợp (Summany Model)	30
2.4.4. tạo mô hình vật lý (Phisical Model)	31
2.5. Lập kế hoạch cài đặt vật ly'	32
2.5.1. Chọn lựa kiến trúc itnh1 toán (Phisical Model)	32
2.5.2. Lập giải pháp lưu trữ (Warehouse Stoorage)	34
2.6. xây dựng quy trình tích hợp dữ liệu cho DW	34
2.6.1. trích dữ liệu (Extract)	35
2.6.2. Biến đổi dữ liệu (Transform)	36

2.6.3.	Tải dữ liệu (Load)	37
2.7.	Quản trị DW	38
3.	KHAI THÁC KHO DỮ LIỆU	39
3.1.	Tổng quan về khai thác thông tin từ DW	39
3.1.1.	Mục đích của việc khai thác dữ liệu từ DW	39
3.1.2.	Các kĩ thuật khai thác dữ liệu từ DW	39
3.2.	Công nghệ khai thác dữ liệu DW	40
3.2.1.	Công cụ báo cáo	40
3.2.2.	Công cụ truy vấn	41
3.2.3.	Công cụ phân tích trực tuyến (OLAP)	41
3.2.4.	Bộ công cụ phân tích	41
3.2.5.	Khai phá dữ liệu (Data Mining)	42
3.2.6.	Ứng dụng phân tích (Analytical Application)	42
3.3.	Xử lý phân tích trực tuyến (OLAP)	42
3.3.1.	Tại sao phải xử lý phân tích trực tuyến	42
3.3.2.	Phân biệt kho dữ liệu quan hệ và kho dữ liệu đa chiều	43
3.3.3.	Định nghĩa OLAP	43
3.3.4.	Kiến trúc của OLAP	45
3.3.4.1.	Kiến trúc Logic của OLAP	45
3.3.4.2.	Kiến trúc vật lý của OLAP	45
3.3.5.	Phân loại OLAP	47
3.3.5.1.	MOLAP	47
3.3.5.2.	ROLAP	47
3.3.5.3.	HOLAP	48

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 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 khái 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. Cài 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 cài 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)

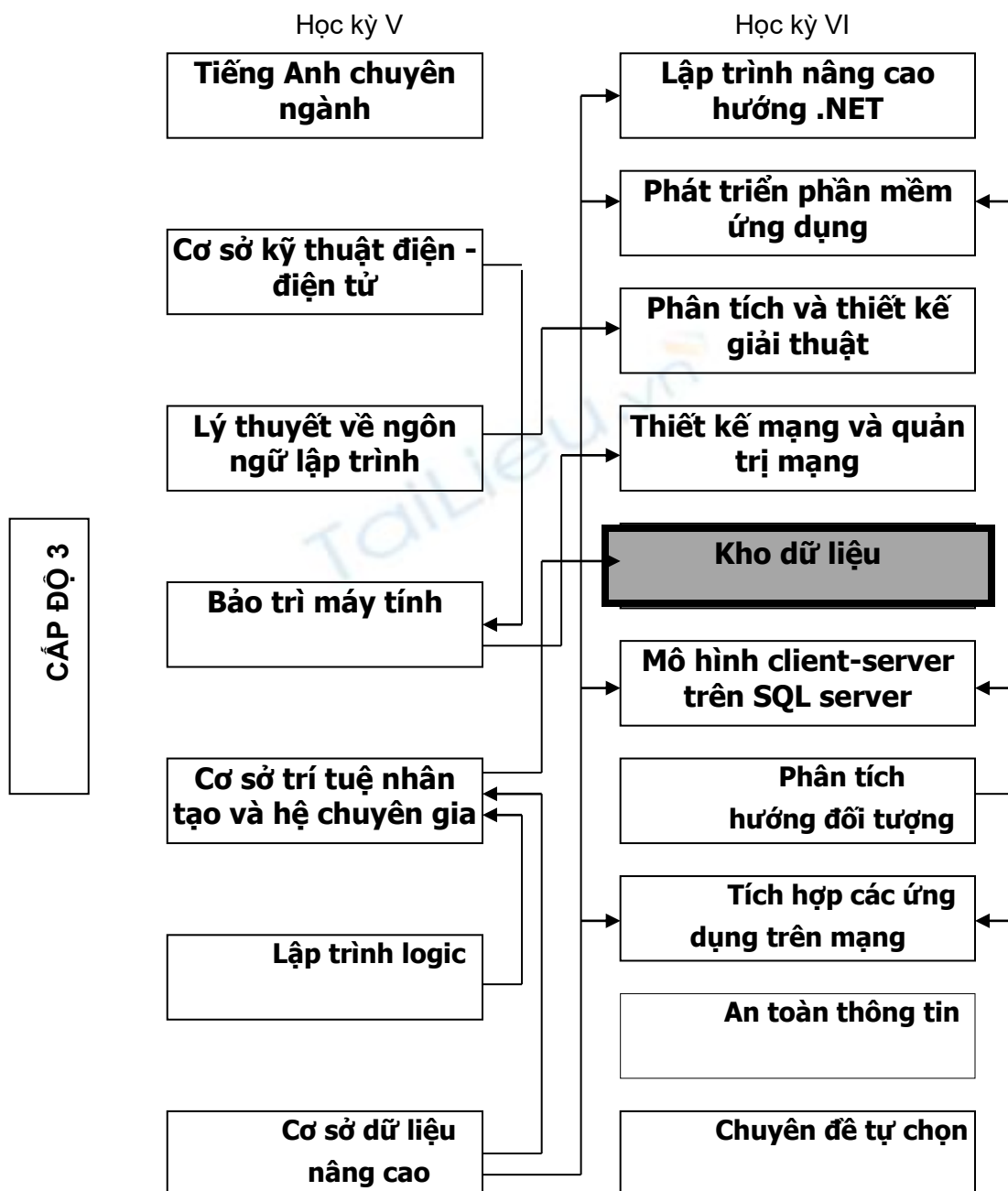
Kỹ năng thực hành :

- 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

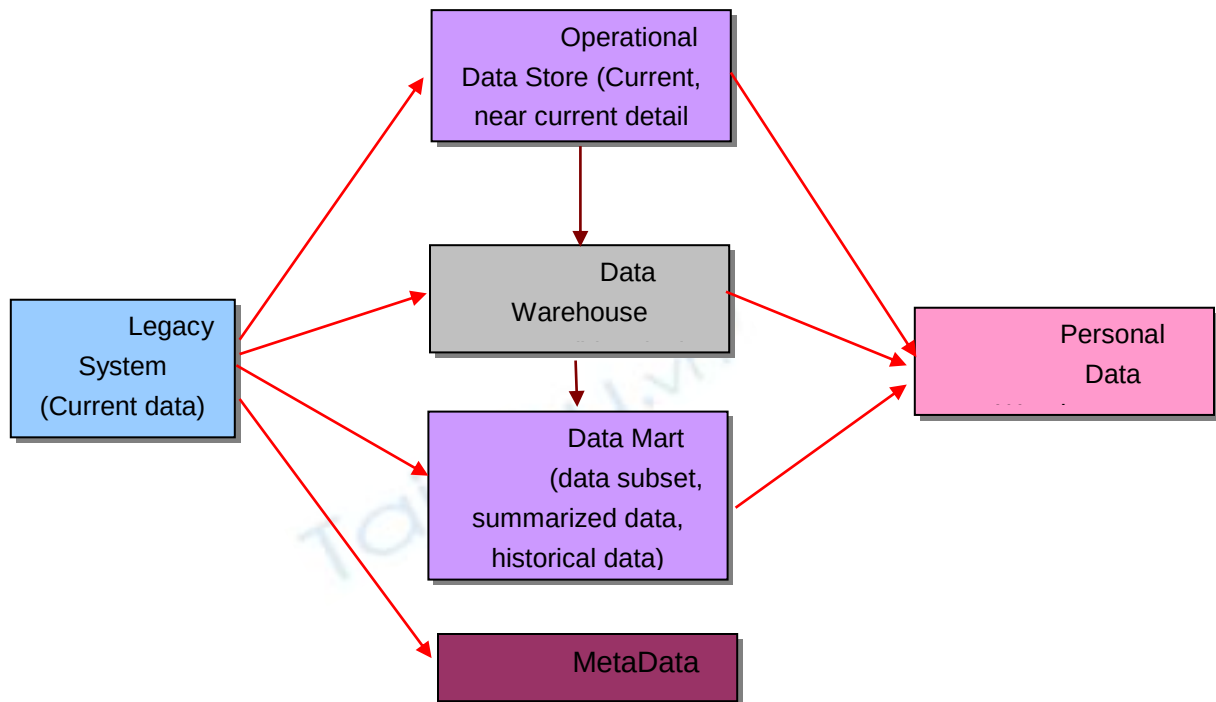
Định nghĩa: Kho dữ liệu (Data Warehouse - DW) là tuyển tập các CSDL tích hợp, hướng chủ đề, được thiết kế để hỗ trợ cho chức năng trợ giúp quyết định, mà mỗi đơn vị dữ liệu đều liên quan tới một khoảng thời gian cụ thể.

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