

NGHIÊN CỨU XÂY DỰNG, KHAI THÁC KHO DỮ LIỆU

Lê Tỷ Khánh*

Tóm tắt

Việc xây dựng một cơ sở dữ liệu đã được nghiên cứu từ lâu, đã có nhiều công cụ và giải pháp của các hãng hỗ trợ phát triển, nhưng xây dựng kho dữ liệu và khai phá dữ liệu là một khái niệm khá mới, mặc dù cũng dựa trên nền tảng cơ sở dữ liệu nhưng mang những đặc thù riêng biệt, được nghiên cứu trong những năm gần đây khi khối lượng dữ liệu lưu trữ trên máy tính, Internet ngày càng khổng lồ, có nơi đã bắt đầu lên đến hàng terabyte. Các hệ quản trị cơ sở dữ liệu lớn như Oracle 1xi, SQL server 201x,... đã và đang xây dựng các công cụ hỗ trợ cho việc xây dựng kho dữ liệu và khai phá dữ liệu. Trong các lĩnh vực: bán hàng, ngân hàng, viễn thông ... việc xây dựng kho dữ liệu và khai phá dữ liệu đã trở nên phổ biến và mang lại những hiệu quả rất lớn. Bài viết này phân tích và cung cấp cái nhìn sâu hơn về thực hiện xây dựng kho dữ liệu và khai phá dữ liệu.

Từ khóa: Kho dữ liệu, khai phá dữ liệu, tài dữ liệu, trích xuất dữ liệu, chuyển đổi dữ liệu.

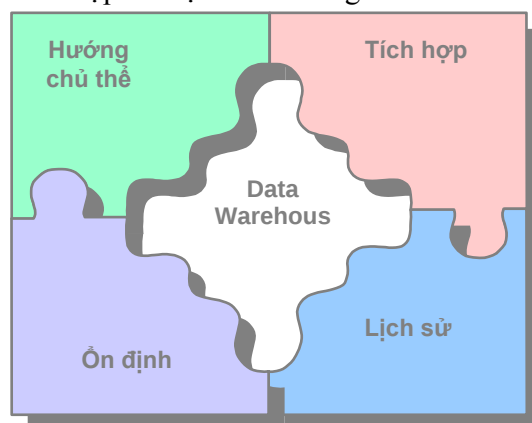
1. Khái niệm kho dữ liệu

Kho dữ liệu (**Data Warehouse - DWH**) là gì? Thông thường, một DWH được xem như một cơ sở dữ liệu quan hệ được thiết kế phục vụ cho truy vấn và phân tích hỗ trợ ra quyết định, được duy trì tách biệt từ những cơ sở dữ liệu tác nghiệp (Operational Database) của tổ chức. Nó cung cấp một nền tảng đồng nhất cho việc hợp nhất, lịch sử hoá dữ liệu. Có nhiều định nghĩa về DWH, **theo W.H.Inmon [5]** DWH được định nghĩa như một “*tập hợp dữ liệu hướng chủ thể (Subject Oriented), tích hợp (Integrated), có tính ổn định (Non Volatile), tính lịch sử (Time Variant) hỗ trợ cho xử lý thực hiện quyết định quản trị*”.

Theo Paul Lucas – IBM [2], là nơi lưu trữ dữ liệu đầy đủ và nhất quán, được tổng hợp về từ nhiều nguồn, được xây dựng sẵn cho người dùng cuối, dễ hiểu. **Theo John Laddy [6]**, Công nghệ DWH (Data Warehouse Technology) là tập các phương pháp, kỹ thuật và các công cụ có thể kết

hợp, hỗ trợ nhau để cung cấp thông tin cho người sử dụng trên cơ sở tích hợp từ nhiều nguồn dữ liệu, nhiều môi trường khác nhau.

Như vậy, mục tiêu chính của DWH đáp ứng các tiêu chí cơ bản sau: Đáp ứng yêu cầu thông tin của người sử dụng; Hỗ trợ lãnh đạo, nhân viên của tổ chức thực hiện hiệu quả nhiệm vụ của mình, có những quyết định hợp lý, nâng cao năng suất làm việc, thu được nhiều lợi nhuận...; Giúp tổ chức, cơ quan quản lý điều hành các dự án, nghiệp vụ một cách hiệu quả, tiết kiệm; Tích hợp dữ liệu từ nhiều nguồn khác nhau.



Hình 1: Minh họa định nghĩa W.H.Inmon

*ThS, Sở Thông tin và Truyền thông Phú Yên

2. So sánh DWH và các hệ cơ sở dữ liệu tác nghiệp (Online Transaction Processing - OLTP)

Xử lý giao dịch trực tuyến OLTP: Tác vụ chính của các hệ quản trị CSDL quan hệ truyền thống; Các thao tác hàng ngày: nhập, sửa, báo cáo,...

Xử lý phân tích trực tuyến OLAP (Online Analysis Processing): Tác vụ chính của hệ thống kho dữ liệu, phân tích dữ liệu và đưa ra quyết định.

Các đặc điểm khác nhau (OLTP và OLAP) [2]

Xử lý giao dịch trực tuyến (OLTP)	Kho dữ liệu, xử lý phân tích trực tuyến (OLAP)
• Công nghệ: CSDL quan hệ • Hướng toàn tác • Chuẩn hóa, không dư thừa • Tập trung vào dữ liệu hiện tại • Trả lời các truy vấn đơn • Tính toàn vẹn, bảo mật, đồng thời, khóa • Xử giao dịch trực tuyến	• CSDL quan hệ, CSDL đa chiều • Chấp nhận dư thừa • Tiền tính toán tổng hợp • Dữ liệu lịch sử • Phân tích rất phức tạp • Tích hợp dữ liệu từ đa nguồn • Dữ liệu rất lớn • Các câu hỏi phức tạp

Tổng quát, DWH làm nhiệm vụ phân phát dữ liệu cho nhiều đối tượng

- Theo quan điểm người dùng: Khách hàng so với thị trường.

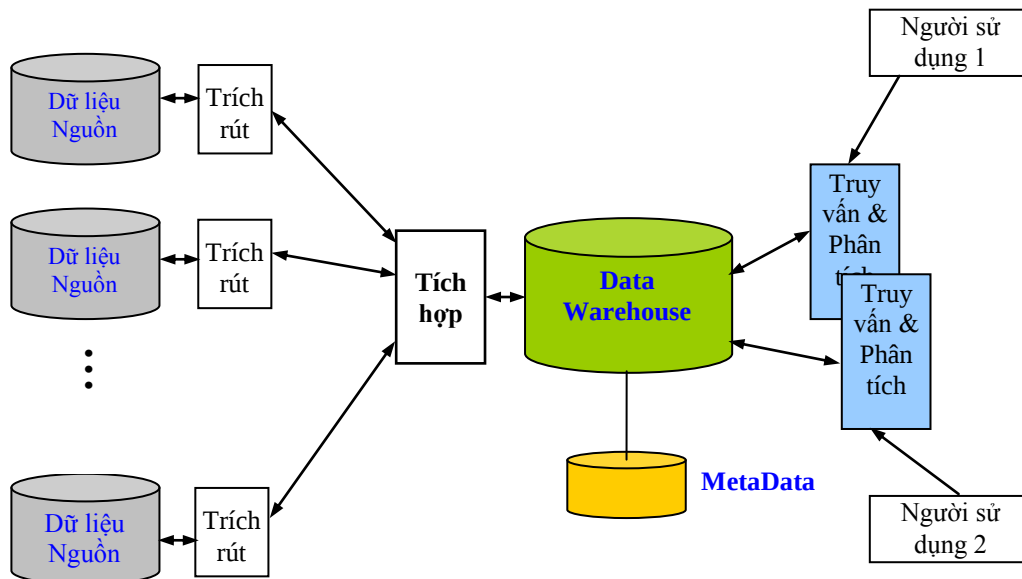
- Các nội dung dữ liệu: Hiện tại, chi tiết so với lịch sử và tóm lược.

- Thiết kế CSDL: ER + ứng dụng so với hình sao và hướng chủ thể.

- Khung nhìn: Hiện tại, cục bộ so với lịch sử và tích hợp.

- Các mẫu truy cập: Cập nhật so với các truy vấn chỉ đọc.

(người sử dụng), xử lý thông tin dưới nhiều dạng: CSDL, SQL Query, Report,...



Hình 2: Mô hình hoạt động tổng quát DWH

3. Các kiểu dữ liệu trong DWH

- **Dữ liệu nghiệp vụ (Business data):** Là dữ liệu dùng để vận hành và quản lý một đơn vị tổ chức. Nó phản ánh những hoạt động của doanh nghiệp và những đối tượng

trong thế giới thực như khách hàng, kho hàng, sản phẩm, hoá đơn bán hàng,... Nó được tạo ra và sử dụng bởi các hệ thống xử lý giao tác cũng như hệ thống hỗ trợ ra quyết định.

Các đặc điểm để phân loại dữ liệu nghiệp vụ: Phương thức sử dụng dữ liệu, phạm vi dữ liệu, dữ liệu đọc, ghi hay chỉ đọc, dữ liệu theo thời gian.

Phương thức sử dụng trong nghiệp vụ: thực hiện 2 mục tiêu:

- Dữ liệu tác nghiệp: Xử lý công việc và liên quan đến các hoạt động hay quyết định mang tính ngắn hạn. Đây là dữ liệu gốc của một tổ chức, nó là nguồn của dữ liệu mang tính thông tin;

- Dữ liệu mang tính thông tin: Xử lý và khai thác dữ liệu trong thời gian dài.

- Cả dữ liệu tác nghiệp và dữ liệu mang thông tin đều được cấu trúc theo nhu cầu truy cập và sử dụng của tổ chức xây dựng ra nó.

Phạm vi dữ liệu: Dữ liệu có thể phản ánh một mẫu tin hoặc một giao dịch, nó cũng có thể là một tổng hợp các mẫu tin hoặc các giao dịch. Dữ liệu có thể là:

- Dữ liệu chi tiết hay dữ liệu nguyên tử: Thường là các đối tượng hay giao dịch cơ sở như: sản phẩm, đơn đặt hàng hay khách hàng,...

- Dữ liệu tổng hợp: Trong quản lý doanh nghiệp thể hiện ở mức độ bao quát hoạt động nghiệp vụ.

Dữ liệu đọc, ghi hay chỉ đọc:

- Dữ liệu đọc, ghi: Yêu cầu được thiết kế cẩn thận các tiến trình cập nhật nhằm bảo đảm các qui tắc, các luật nghiệp vụ. Xây dựng cấu trúc tối ưu cho việc đọc và ghi dữ liệu vào CSDL.

- Dữ liệu chỉ đọc: Thiết kế phục vụ truy cập thông tin. Cấu trúc xây dựng tối ưu cho việc truy vấn nhiều lần.

Dữ liệu theo thời gian: xét theo khía cạnh thời gian, gồm:

- Dữ liệu hiện tại: Là khung nhìn của công việc nghiệp vụ vào thời điểm hiện tại. Có thể hiểu nó là dữ liệu tức thời và như vậy sẽ thay đổi theo thời gian trên các hoạt động nghiệp vụ.

- Dữ liệu tại thời điểm: là một ảnh chụp dữ liệu nghiệp vụ trong một thời điểm nhất định, phản ánh trạng thái công ty tại thời điểm đó. Dữ liệu thể hiện khung nhìn quá khứ, có thể dùng để định kế hoạch hay dự đoán.

- Dữ liệu định kỳ: Đây là lớp dữ liệu mở rộng quan trọng. Nó thể hiện sự thay đổi của công việc nghiệp vụ trong mỗi giai đoạn thời gian.

- Các khái niệm này là cơ sở cho việc xử lý các dữ liệu lịch sử (bao gồm dữ liệu định kỳ và các ảnh chụp dữ liệu quá khứ) của DWH.

Từ các đặc điểm phân tích trên, có thể xác định 3 loại dữ liệu nghiệp vụ. Việc phân loại này dựa trên dữ liệu có cấu trúc bởi các lý do, thứ nhất, dữ liệu có cấu trúc bao giờ cũng được cài đặt vào kho chứa đầu tiên; thứ 2, với dữ liệu có cấu trúc phân biệt sự khác biệt 3 loại dữ liệu sẽ rõ ràng hơn.

- Dữ liệu thời gian thực (real-time data): Dữ liệu chi tiết, tức thời dùng trong vận hành công việc và được truy xuất theo chế độ đọc, ghi thông qua các giao dịch đã được xác định trước. Dữ liệu thời gian thực được tạo lập, thao tác và sử dụng trong các ứng dụng thao tác hay sản xuất. Chúng có thể tổ chức thành các tập tin hay CSDL.

- Dữ liệu dẫn xuất (derived data): Dữ liệu xác thực theo thời điểm hoặc dữ liệu định kỳ ở mức chi tiết hoặc tổng hợp, thuộc chế độ chỉ đọc, nhận được từ việc xử lý dữ liệu thời gian thực và dùng để quản lý nghiệp vụ. Dữ liệu dẫn xuất là tập dữ liệu thường được sử dụng để trợ giúp quyết định.

- Dữ liệu tương hợp, hoà hợp (reconciled data): là loại dữ liệu dẫn xuất đặc biệt, sinh ra bởi 1 tiến trình được thiết kế nhằm bảo đảm sự vững chắc nội tại của dữ liệu kết quả, tiến trình này thực hiện dựa vào dữ liệu thời gian thực tại mức chi tiết, duy trì hoặc tạo ra các dữ liệu lịch sử.

- **Siêu dữ liệu (Metadata):** Là dữ liệu về dữ liệu, được sử dụng trong DWH, trả lời các câu hỏi ai? Cái gì? Khi nào? Tại sao? Như thế nào về dữ liệu? Các thuộc tính này được sử dụng cho việc xây dựng, duy trì, quản lý và sử dụng DWH. Metadata là một trong những dữ liệu quan trọng nhất của DWH.

Ở mức tối thiểu, Metadata phải mô tả được về dữ liệu chứa trong DWH, bao gồm: Vị trí, mô tả về DWH và các thành phần dữ liệu (các đối tượng của DWH); các tên gọi, định nghĩa, cấu trúc và nội dung của DWH cùng với quan sát của người sử dụng; xác nhận căn cứ của các nguồn cung cấp dữ liệu; các quy tắc chuyển đổi và tích hợp dữ liệu được sử dụng trong DWH, trong đó có cả những phép ánh xạ các CSDL tác nghiệp sang DWH, kể cả những thuật toán chuyển đổi; các quy tắc chuyển đổi và tích hợp dữ liệu được sử dụng để cung cấp dữ liệu đến người sử dụng; những thông tin mô tả về hệ thống thông tin cấp phát; những thông tin thao tác trong DWH, bao gồm lịch sử quá trình cập nhật DWH, quá trình làm tươi, sao chụp dữ liệu,...; các hệ số đo (metric) được sử dụng để phân tích hiệu suất sử dụng và hiệu quả của DWH; sự đảm bảo về an toàn dữ liệu và danh sách quản lý quyền truy cập.

Như vậy, mô hình thích hợp cho Metadata chính là mô hình quan hệ thực thể hay biểu đồ lớp trong UML. Trong các mô hình này có các thực thể (entity), thuộc tính (attribute), mối quan hệ (relationship).

- Đối tượng vào/ra (Input - Output Object): Mô tả các đối tượng dữ liệu vào/ra DWH.
- Các phần tử dữ liệu (Data Element): mô tả các đơn vị cơ sở của các sự kiện truy nhập được như các cột trong các CSDL.
- Các thành phần quan hệ (Relationship

Member): Mô tả về sự tham gia của các thực thể trong mỗi quan hệ xác định.

4. Kiến trúc DWH [7]

Thông thường kiến trúc DWH gồm 3 tầng (layer): Tầng thể hiện (Presentation layer), Tầng xử lý phân tích trực tuyến (OLAP layer), Tầng chiết (Extraction layer)

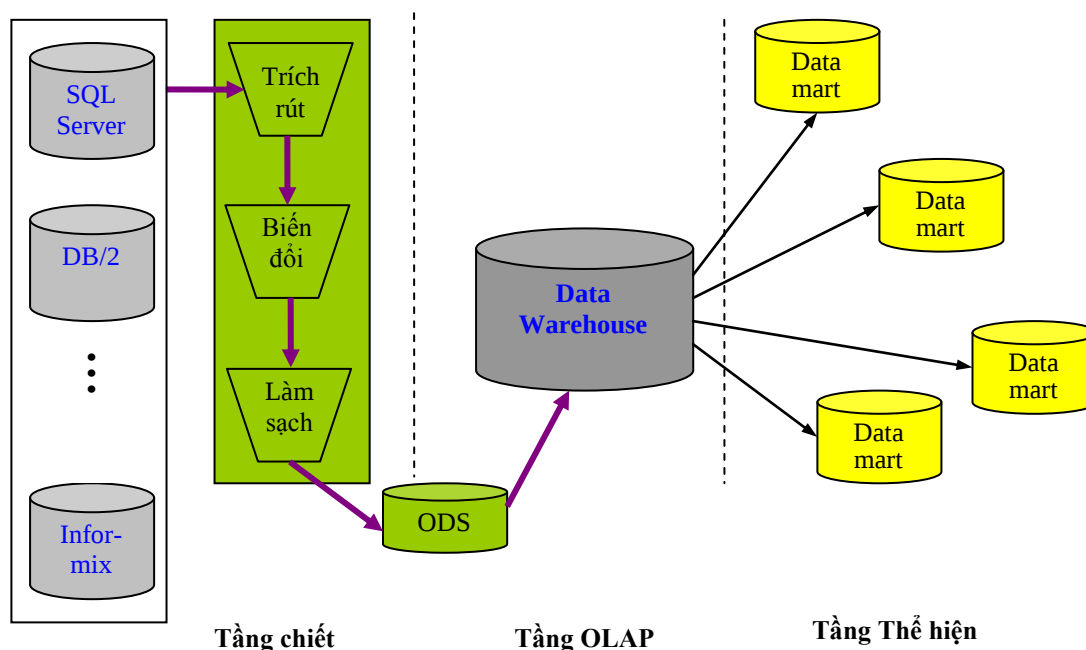
- **Tầng thể hiện:** Được biết đến như là một phần của DWH; Cung cấp các báo cáo, gồm: báo cáo thường kỳ, thống kê, báo cáo chi tiết (so sánh dựa trên các yêu cầu hiện thời), báo cáo ngoại lệ được đưa ra cho một số yêu cầu đặc biệt trong việc phân tích dữ liệu, phân tích tiền tính toán, các đồ thị, biểu đồ; Cho phép phân tích sâu trực tuyến dữ liệu; Thăm dò sự tương tác dữ liệu; Cung cấp các giao diện người dùng phức tạp; Cung cấp khả năng phân tích dữ liệu phức tạp bằng phương thức đơn giản; Kiến trúc khách chủ; Giao diện cho DWH trung tâm và các mảng dữ liệu cục bộ (data mart); Nhiều loại giao diện cho nhiều cấp người dùng: báo cáo và phân tích; Các báo cáo kinh doanh, ước lượng, dự báo, phân loại,...; Các kỹ thuật biểu diễn dữ liệu; Công nghệ khai phá dữ liệu. Tầng này trả lời các câu hỏi do người sử dụng đặt ra. Chẳng hạn trong một công ty kinh doanh có thể có các câu hỏi sau: khách hàng mua mặt hàng nào nhiều nhất, mặt hàng nào ít nhất đến thời điểm hiện tại? mặt hàng nào bán chạy nhất trong các kỳ lễ? So sánh mức chi tiêu trung bình của khách hàng thanh niên và khách hàng lớn tuổi?

- **Tầng xử lý phân tích trực tuyến:** Ở tầng này ứng dụng kỹ thuật phân tích đa chiều của nhiều tập dữ liệu lớn; Dữ liệu được trình bày theo mô hình đa chiều; Công nghệ OLAP cho phép truy cập nhanh tới các toán tử hỗ trợ đặt biệt như rolling-up, drilling down,...; Cung cấp giao diện cho DWH và các báo cáo để tương tác với dữ liệu; Tối ưu hoá câu truy vấn; Quản lý

và phân tích các truy vấn thường xuyên ;An toàn: cấp quyền và quản lý truy cập.

- **Tầng chiết (hay còn gọi tầng Thu thập):** Tầng này chứa các loại dữ liệu nguồn và các thủ tục, chương trình để trích, rút, sàng lọc các dữ liệu cần thiết từ dữ liệu nguồn. Dữ liệu nguồn gồm: 1) Dữ liệu từ các hệ thống tác nghiệp. Các hệ thống này chứa dữ liệu chi tiết và hiện tại, sử dụng cho các giao dịch hàng ngày, đây là nguồn

dữ liệu chính được tích hợp để xây dựng DWH; 2) Hệ thống kế thừa, đó là các sưu tập dữ liệu cũ không được dùng cho các mục đích hoạt động; 3) Các nguồn dữ liệu bên ngoài, chẳng hạn dữ liệu phân tích thị trường, dữ liệu báo cáo thời tiết, dữ liệu thuế,...không phải là dữ liệu được tạo ra trong công ty và có thể có cấu trúc và mã hóa hoàn toàn khác nhau phụ thuộc vào nhà cung cấp;



Hình 3: Kiến trúc 3 tầng của DWH

5. Quy trình xây dựng DWH [5]

Khi quyết định xây dựng một DWH cần thực hiện: Thu thập các yêu cầu chức năng và phi chức năng để đảm bảo chắc chắn hệ thống xây dựng sẽ hỗ trợ người sử dụng đạt được những mục tiêu đề ra; định nghĩa các yêu cầu chức năng như: Hệ thống DWH sẽ làm gì? Các câu hỏi hoặc các vấn đề mà hệ thống sẽ trả lời, dữ liệu gì sẽ lưu trữ trong DWH? Phân tích những gì người sử dụng sẽ khai thác; Các yêu cầu phi chức năng như bảo mật, tính sẵn sàng, sự thực thi. Chẳng hạn về tính sẵn sàng: mong muốn thời gian dừng hệ thống nhỏ hơn 1 giờ/tháng; Nghiên cứu sâu kỹ về hệ thống:

Tính khả thi của dữ liệu truy xuất, những rủi ro tiềm tàng gây khó khăn khi xây dựng và vận hành hệ thống. Từ đó, tạo thiết kế logic và thiết kế vật lý cho DWH

Thiết kế logic mang tính khái quát và trừu tượng hơn thiết kế vật lý. Trong thiết kế logic xác định những mối quan hệ logic trong số những đối tượng. Trong thiết kế vật lý, chú trọng về hiệu quả cách lưu trữ và khôi phục những đối tượng cũng như xử lý việc chuyển đổi và sao lưu/ khôi phục.

- Thiết kế logic: Một kỹ thuật có thể sử dụng để mô hình hoá những yêu cầu thông tin logic là mô hình ER (Entity Relationship). Mô hình ER bao gồm xác định những thực

thể, những thuộc tính của thực thể, và những mối quan hệ giữa các thực thể.

Quá trình thiết kế logic bao gồm sắp đặt dữ liệu vào trong những mối quan hệ logic gọi là những thực thể và những thuộc tính. Một thực thể đại diện cho một đơn vị thông tin. Trong những cơ sở dữ liệu quan hệ, một thực thể thường ánh xạ tới một bảng. Một thuộc tính là một thành phần của một thực thể, giúp định nghĩa tính duy nhất của thực thể. Trong những cơ sở dữ liệu quan hệ, một thuộc tính ánh xạ tới một cột. Trong khi sơ đồ ER theo truyền thống được kết hợp với những mô hình được tiêu chuẩn hóa cao chẳng hạn như những ứng dụng OLTP, kỹ thuật đó hữu ích cho thiết kế kho hàng dữ liệu trong định dạng mô hình chiều. Trong mô hình chiều, thay vì việc tìm kiếm khám phá những đơn vị nguyên tử của thông tin (như những thực thể và những thuộc tính) và tất cả những mối quan hệ giữa chúng, sẽ nhận dạng thông tin nào thuộc về một bảng sự kiện trung tâm và thông tin nào thuộc về những bảng chiều có liên hệ với bảng sự kiện trung tâm. Xác định những chủ đề kinh doanh hay những lĩnh vực dữ liệu, định nghĩa những mối quan hệ giữa chúng, đặt tên những thuộc tính..

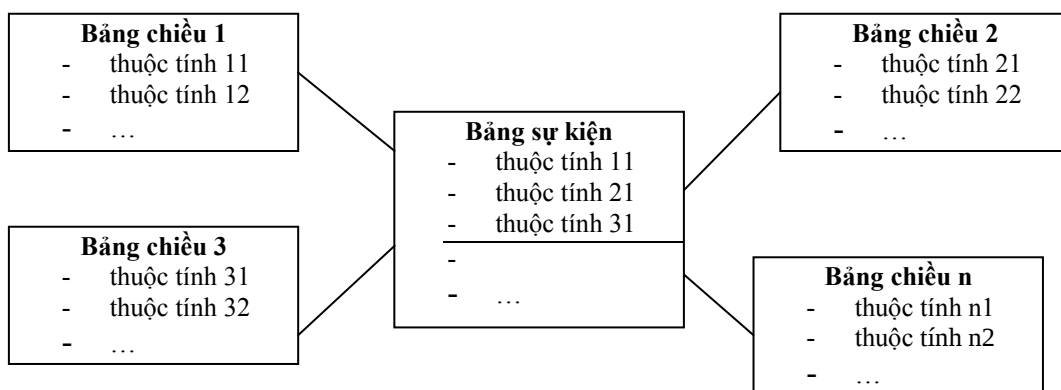
Kết quả thiết kế logic là một tập hợp những thực thể và những thuộc tính tương ứng tới những bảng sự kiện, những bảng chiều và một mô hình gồm: dữ liệu nguồn cung cấp cho DWH.

Lược đồ DWH:

Lược đồ là Một tập hợp những đối tượng cơ sở dữ liệu, gồm các table, view, index, và vấn đề khác. Có nhiều cách sắp xếp các đối tượng của lược đồ trong mô hình thiết kế lược đồ trong một DWH. Đa số những kho dữ liệu hiện nay sử dụng mô hình chiều. Mô hình dữ liệu nguồn và những yêu cầu những người sử dụng là cơ sở để thiết kế lược đồ DWH.

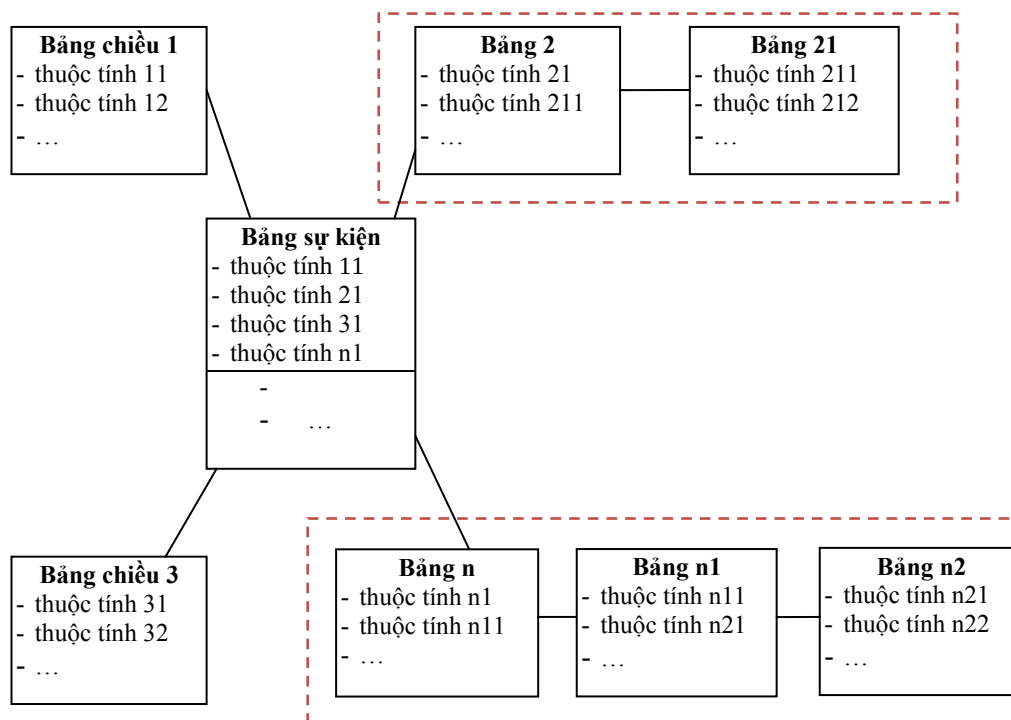
Lược đồ hình sao: là mô hình DWH đơn giản nhất. Được gọi là lược đồ hình sao bởi vì lược đồ giống với một ngôi sao, nhiều điểm được kết nối từ một trung tâm. Trung tâm hình sao gồm một hoặc nhiều bảng sự kiện, những điểm của ngôi sao là những bảng chiều.

Cách tự nhiên nhất để xây dựng một kho dữ liệu có lược đồ hình ngôi sao: Xây dựng 1 bảng sự kiện và các bảng chiều chung quanh có quan hệ đến bảng sự kiện. Một mô hình ngôi sao tối ưu hóa sự thực hiện bởi việc giữ cho những truy vấn đơn giản và cung cấp sự đáp ứng nhanh nhất.



Hình 4: Lược đồ hình Sao

Một số Lược đồ khác: **Lược đồ bông tuyết:** Căn cứ vào lược đồ hình sao, Bản sự kiện giống như lược đồ hình sao, Các chiều được chuẩn hoá (chuẩn 3), Các chiều được cấu trúc rõ ràng



Hình 5: Minh họa lược đồ hình Bông tuyết

Những đối tượng trong lược đồ DWH

Bảng sự kiện: Là những bảng lớn, chính trong lược đồ DWH, nơi cất giữ những khối lượng lớn sự kiện hoạt động của đơn vị và những khóa ngoại liên quan đến các bảng chiều. Những bảng sự kiện biểu diễn dữ liệu, thông thường là số, cho phép thêm vào và có thể được phân tích và kiểm tra.

Bảng chiều: Có thể hiểu như những bảng tham chiếu hay tra cứu, chứa đựng dữ liệu tính tương đối trong DWH. Những bảng chiều lưu trữ thông tin dùng chứa đựng những câu hỏi. Những bảng chiều thông thường là văn bản và miêu tả, có thể sử dụng chúng như những dòng đầu của tập hợp kết quả.

Chiều là Một cấu trúc, thường bao gồm một hoặc nhiều sự phân cấp. Những thuộc tính chiều giúp để mô tả giá trị chiều. Chúng có tính miêu tả. Có vài chiều phân biệt rõ ràng, kết hợp với những các sự kiện, cho phép trả lời những truy vấn. Những tích lũy hay những sự tổng hợp tự nhiên

này bên trong một bảng chiều được gọi là những sự phân cấp.

Phân cấp: Sự phân cấp là cấu trúc logic mà việc sử dụng những mức theo thứ tự như một phương tiện tổ chức dữ liệu. Một sự phân cấp có thể được dùng để tổng hợp dữ liệu. Ví dụ, trong chiều *Thời gian (Time)*, có thể tập hợp dữ liệu từ mức *Tháng (Month)*, tới mức *Quý (Quarter)*, tới mức *Năm (Year)*. Một sự phân cấp cũng có thể được dùng để định nghĩa thiết lập một cấu trúc phả hệ. Bên trong sự phân cấp, mỗi mức được nối một cách logic tới những mức ở trên và ở dưới. những giá trị dữ liệu tại các mức thấp hơn tập hợp thành những giá trị dữ liệu tại những mức cao hơn. Một chiều có thể bao gồm nhiều hơn 1 phân cấp.

Phân cấp chiều cũng cho phép nhóm các mức từ tổng quan tới hạt (phần tử nhỏ nhất). Những công cụ truy vấn sử dụng sự phân cấp cho phép khoan sâu vào trong dữ liệu. Đây là một trong số những lợi ích chính một kho dữ liệu, những mối quan hệ phả hệ này cho phép những người phân tích truy nhập dữ liệu nhanh.

Mức: Một mức đại diện cho một vị trí trong sự phân cấp. Các mức sắp xếp từ tổng quan đến cụ thể, mức gốc là mức cao nhất hay chung nhất. Những mức trong một chiều được tổ chức thành một hoặc nhiều sự phân cấp.

Những mối quan hệ mức: Chỉ rõ thứ tự từ đỉnh tới đáy, từ mức gốc đến thông tin chi tiết nhất. Chúng định nghĩa mối quan hệ cha - con giữa các mức trong một hệ thống phân cấp.

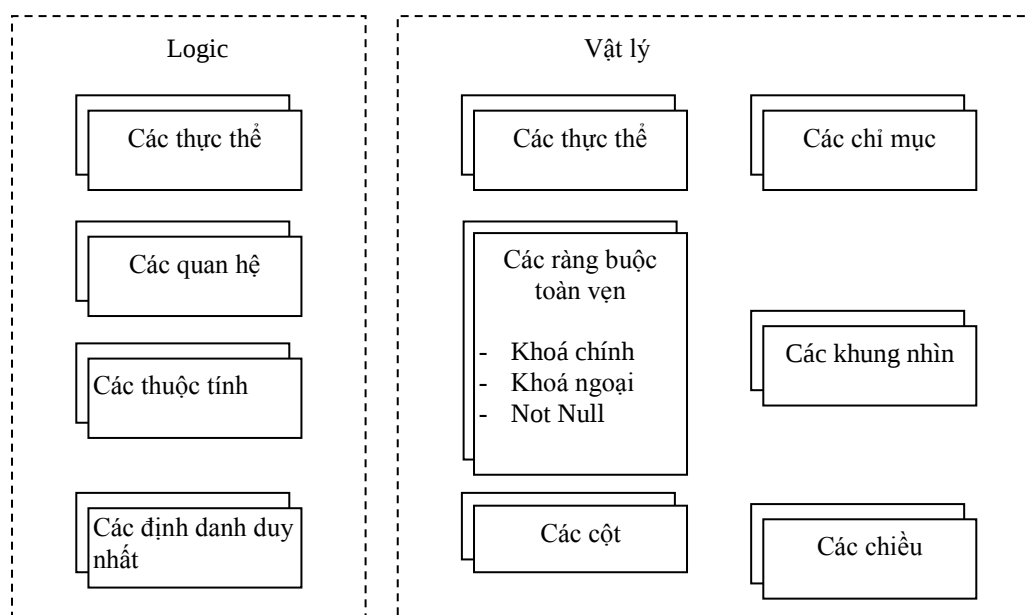
Quan hệ: Những mối quan hệ bảo đảm sự toàn vẹn dữ liệu. Thiết kế một mối quan hệ giữa những bảng chiều tuân theo những quy tắc qui định trong cơ sở dữ liệu.

Các khoá: Khóa chính nằm trên các bảng chiều, Khóa ngoại nằm trên bảng sự

kiện, Tạo lập các index để tăng tốc độ, Khóa được xác định trong giai đoạn thiết kế, Các khóa kết hợp có thể được áp dụng.

- Thiết kế vật lý trong DWH, Thiết kế vật lý là sự tạo ra CSDL từ mô hình thiết kế logic bằng các công cụ ngôn ngữ khác nhau. Ở đây ta dùng các câu lệnh của SQL. Trong quá trình thiết kế vật lý là quá trình chuyển đổi dữ liệu được thu thập trong thiết kế logic thành mô tả cấu trúc CSDL vật lý.

Trong thiết kế logic đã định nghĩa sẵn một mô hình DWH, những thực thể, thuộc tính, và những mối quan hệ. Những thực thể được liên kết với nhau sử dụng những mối quan hệ. Những thuộc tính được dùng để mô tả những thực thể. Định danh duy nhất để phân biệt giữa các thực thể.



Hình 6: So sánh giữa thiết kế logic và thiết kế vật lý

Quá trình thiết kế vật lý chuyển đổi, ánh xạ: Các thực thể thành các bảng, Các quan hệ thành các khoá ngoại, Các thuộc tính thành các cột, Các định danh duy nhất thành các ràng buộc khoá chính, Các định danh duy nhất thành các ràng buộc khoá duy nhất

Những cấu trúc thiết kế vật lý

Những cấu trúc sau được tạo ra khi thiết kế vật lý: Không gian chứa bảng (tablespaces),

Bảng và các bảng được phân chia, Khung nhìn, Ràng buộc toàn vẹn, Chiều. Đồng thời, cấu trúc sau có thể được tạo ra để cải thiện sự thực thi: Chỉ mục và chỉ mục được phân chia, Các khung nhìn cụ thể hoá

Không gian chứa bảng

Không gian chứa bảng (tablespace) gồm chứa một hoặc nhiều tập tin dữ liệu (datafile). Một datafile chỉ được kết hợp

đến 1 tablespace. Từ quan điểm thiết kế, tablespace là nơi chứa những cấu trúc thiết kế vật lý. Những tablespace cần được tách rời ra. Chẳng hạn, những bảng sẽ được tách ra từ những chỉ mục của chúng và những bảng nhỏ sẽ được tách ra từ những bảng lớn.

Bảng và bảng phân chia (table and partitioned tables)

Bảng là đơn vị cơ sở để lưu trữ dữ liệu. Chúng chứa dữ liệu trong một DWH. Sử dụng bảng phân chia thay vì bảng không phân chia để giải quyết vấn đề về những khối lượng dữ liệu rất lớn bằng việc cho phép phân tách chúng thành những phần nhỏ hơn và dễ quản lý hơn. Tiêu chí thiết kế chính cho việc phân chia là tính dễ điều khiển, tuy vậy, cũng nhìn thấy những lợi ích thực thi trong đa số những trường hợp, bởi vì sự phân chia hay xử lý song song thông minh.

Khung nhìn (view)

Một view là một sự biểu diễn của dữ liệu chứa trong một hoặc nhiều bảng hay các view khác. Một view xuất ra từ một truy vấn và xem như một bảng. view không yêu cầu bất kỳ không gian nào trong CSDL.

Ràng buộc toàn vẹn (Integrity Constraint)

Những sự ràng buộc toàn vẹn được dùng để giám sát những quy tắc kết hợp CSDL và để ngăn ngừa thông tin sai trong bảng. Những ràng buộc toàn vẹn trong môi trường DWH không giống ràng buộc trong môi trường OLTP. Trong môi trường OLTP, chúng chủ yếu ngăn ngừa sự chèn dữ liệu sai vào trong một bản ghi, nhưng điều này không là một vấn đề lớn trong môi trường xây dựng DWH, vì độ chính xác đã được bảo đảm. Trong môi trường DWH, những sự ràng buộc chỉ được sử dụng cho truy vấn ghi lại.

Chỉ mục và những chỉ mục phân chia (Index and Partitioned Indexes)

Những chỉ mục là những cấu trúc tùy chọn liên kết với những bảng hay những

nhóm bảng. Ngoài chỉ mục cây nhị phân (B-tree) cổ điển, những chỉ mục ánh xạ bit (bitmap index) rất phổ biến trong môi trường DWH. Chỉ mục ánh xạ bit tối ưu hóa cấu trúc chỉ mục hỗ trợ quá trình thực thi. Đồng thời, chúng cần thiết cho những phương pháp truy nhập dữ liệu được tối ưu hóa. Chỉ mục cũng có thể phân chia giống như bảng, mặc dù chiến lược phân chia không phụ thuộc cấu trúc bảng. Việc phân chia chỉ mục giúp quản lý DWH dễ dàng hơn trong việc làm tươi và cải thiện truy vấn.

Khung nhìn cụ thể hoá (Materialized View)

Khung nhìn cụ thể hoá là những kết quả truy vấn đã được cất giữ từ trước, vì thế, tránh mất thời gian vào những tính toán khi thực hiện những câu lệnh SQL. Từ quan điểm thiết kế vật lý của khung nhìn, khung nhìn cụ thể hoá giống như bảng hay bảng được phân chia.

Chiều (Dimension)

Chiều là một đối tượng của lược đồ, định nghĩa những mối quan hệ có thứ bậc giữa những cột hay những tập hợp cột. Một mối quan hệ có thứ bậc là một phụ thuộc chức năng từ mức của một phân cấp đến mức kế tiếp. Một chiều tiêu biểu là xã (hoặc phường), huyện (hoặc thành phố) tỉnh, và quốc gia.

- Trích xuất, chuyển đổi và nạp dữ liệu (ETL - Extraction, transformation, loading) vào DWH [6]

Trích xuất, Trích xuất dữ liệu là một phép xử lý để lấy dữ liệu đã được xác định trước ra khỏi các hệ thống tác nghiệp và các nguồn dữ liệu ngoài. Có vài nguyên lý cơ bản để hiểu khi rút dữ liệu từ một hệ thống nguồn là cơ sở dữ liệu tác nghiệp cho mục đích tích hợp vào kho dữ liệu. Khối lượng dữ liệu đang được rút ra lớn, có thể hàng trăm hoặc hàng ngàn megabyte. Một hệ thống OLTP thiết kế sao cho dữ liệu được rút ra từng những mảnh nhỏ, không

phải khối lớn như nguyên gốc, vì vậy, phải cẩn thận để không làm chậm lại hệ thống nguồn xuống quá nhiều. Điều này cũng ảnh hưởng đến quyết định tần số và thời gian trích xuất, lập ra kế hoạch đối với từng hệ thống nguồn và loại dữ liệu được trích khác nhau. Sau khi trích xuất dữ liệu, nạp vào trong DWH càng sớm càng tốt. Cần áp dụng một số sự biến đổi vào dữ liệu lấy từ hệ thống nguồn để thỏa mãn khuôn dạng và cấu trúc thiết kế được chọn. Các loại dữ liệu và nơi chứa dữ liệu trích xuất: CSDL gốc, CSDL ảnh, Truy cập từ đĩa hoặc băng từ.

Chuyển đổi, Bước chuyển đổi gồm 2 bước: lọc và làm sạch dữ liệu, thẩm định và chuyển đổi dữ liệu

Lọc và làm sạch dữ liệu, Quá trình lọc và làm sạch dữ liệu là kiểm tra và sửa chữa lỗi có thể có của dữ liệu để đảm bảo tính đúng đắn của dữ liệu. Công việc này bao gồm dọn dẹp, thay đổi và tính toán lại. làm sạch dữ liệu liên quan đến một số hoặc tất cả các tác vụ sau: kiểm tra tất cả các trường đơn lẻ, lọc ra và hợp nhất các bản ghi trùng lặp,...

Thẩm định và chuyển đổi dữ liệu, Dữ liệu sau đó phải được kiểm tra, thẩm định để đảm bảo chất lượng nhằm đáp ứng các yêu cầu phân tích phục vụ và hỗ trợ ra quyết định. Các công cụ hỗ trợ để thực hiện những công việc nêu trên dựa vào tập các thông số đã được xác định trước. Trước khi

chuyển đổi, cần thiết lập hệ thống đo lường và chuẩn hoá các luật nghiệp vụ.

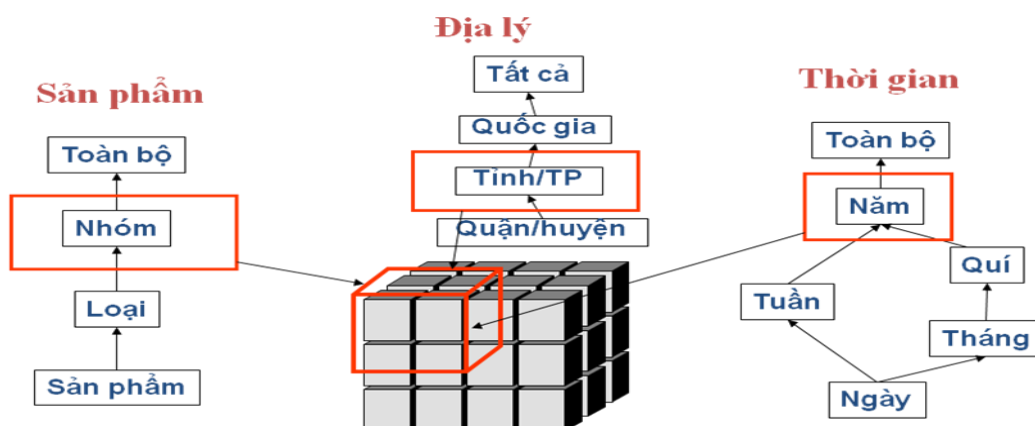
Mục đích của việc chuyển đổi và tích hợp là chuyển dữ liệu thành thông tin có thể hiểu được và hữu ích đối với người sử dụng.

Tải

Việc tải dữ liệu vào DWH có thể thực hiện: Làm tươi lại dữ liệu (Refresh), Bổ sung (incremental) để tạo thêm các dữ liệu ảnh chụp (snapshot) vào bảng dữ liệu; Gắn thêm giá trị thời gian vào khoá, Cập nhật, Đọc trước và tải dữ liệu (Preload and load); Sửa chữa và đánh giá (Repair and Evaluate): Trong quá trình làm sạch, chuyển đổi và tích hợp dữ liệu có thể xảy ra lỗi, do vậy, trong môi trường chứa dữ liệu nguồn phải có chức năng làm nhiệm vụ phát hiện và sửa lỗi này, Quá trình có thể làm bằng thay hoặc bằng các thuật toán tùy thuộc phương thức và công cụ phát triển; Xây dựng chu kỳ cho việc nạp dữ liệu vào kho (giờ cố định trong ngày, hàng ngày, tháng, quý, năm), thời gian cho một lần nạp.

- Các bảng tổng hợp

Mục đích: trả lời nhanh các câu hỏi thường gặp. Các bước: Nhận diện các câu hỏi thường gặp, Nhận diện các chiều và các tổng hợp tương ứng, Định nghĩa phân cấp của các tổng hợp, Tạo một cách có chọn lựa bản sự kiện tổng hợp, Tạo các bảng chiều tổng hợp.



Hình 7: Minh họa tổ chức tổng hợp báo cáo theo 3 chiều: thời gian, sản phẩm, vùng địa lý.

Tóm lại: Với những nghiên cứu phân tích về DWH, việc áp dụng xây dựng và khai thác kho dữ liệu phục vụ trong các ngành kinh doanh hiện nay đang bắt đầu phổ biến, và bước đầu cũng đã mang lại hiệu quả rất tốt. Đặc biệt, trong cải cách hành chính tại địa phương, việc xây dựng

các DWH về kinh tế xã hội, dữ liệu chuyên ngành Y tế, Giáo dục, Văn hóa, Du lịch,... là rất khả quan và tiềm năng. Với những bài viết sau, chúng tôi sẽ đi sâu nghiên cứu ứng dụng DWH vào các kho dữ liệu chuyên ngành, đặc biệt là DWH về kinh tế xã hội tỉnh □

TÀI LIỆU THAM KHẢO

- [1] Đoàn Văn Ban (1997), *Phương pháp thiết kế và khai thác kho dữ liệu*, Đề tài nghiên cứu cấp trung tâm KHTN&CNQG.
- [2] Nguyễn Thanh Bình (2007), *Bài giảng Data warehouse*, TT CNTT, Đại học Huế
- [3] Nguyễn Văn Chức (2007), *Nghiên cứu và ứng dụng mô hình phân tán kho dữ liệu – luận văn thạc sỹ*, Đại học Đà Nẵng.
- [4] Kim Ball Group (2006), *The Microsoft Data Warehouse Toolkit With SQL Server 2005 and the Microsoft Business Intelligence Toolset*.
- [5] William H. Inmon (2005), *Building the Data Warehouse*, Fourth Edition.
- [6] Ralph Kimball, Joe Caserta (2004), *The Data Warehouse ETL Toolkit*, Wiley Publishing
- [7] Vincent Rainardi (2008), *Building a Data Warehouse*.

Abstract

Research on building a data warehouse and data mining

The building of a database has been studied for a long time, and there have been many tools and solutions proposed by some development and support companies. However, the work of building a data warehouse and data mining is a relatively new concept, even though they are also based on the database, they have their own distinct characteristics, and been studied in the recent years when more and more volumes of data have been stored on the computer and the Internet, and in some places, the volumes are up to terabytes. The big databases management systems such as Oracle 1xi, SQL Server 201x,... have currently been building their tools to support the data warehouse development and data mining. In the fields of business, banking, telecommunications... the building of data warehouse and data mining has become popular and brought about enormous efficiency. This article analyses and provides some deeper insight into the construction of a data warehouse and data mining.

Keywords: Data warehouse, data mining, data loading, data extraction, data transformation.