*Tạp chí Tin học và Điều khiển học, T. 15, S. 2 (1999), 27–32*

# MỘT SỐ NGUYÊN LÝ HOẠT ĐỘNG CỦA KHO DỮ LIỆU (DATA WAREHOUSE)

VŨ ĐỨC THI, LÊ HẢI KHÔI

**Abstract.** Data warehousing is the biggest trend in information management today. It is the technology which may finally deliver on a dream pursued by management theorists since the 1960s. In this paper we present the general definition of data warehouse, differences between data warehouse and database management system and some essential principles of data warehouse.

Cùng với sự phát triển mạnh mẽ của công nghệ thông tin, lĩnh vực thiết kế và xây dựng các hệ thống quản lí thông tin lớn đã có những thay đổi đáng kể. Một trong những xu thế chính trong trào lưu này là thiết kế và xây dựng các kho dữ liệu (data warehouse). Công nghệ thiết kế kho dữ liệu là một hướng phát triển mạnh trong những năm 90, đáp ứng những ý tưởng về mặt lý thuyết hình thành từ những năm 60 của thế kỷ này.

Hiện nay, rất nhiều doanh nghiệp lớn cũng như các cơ quan chính phủ ở nhiều nước công nghiệp phát triển đã và đang sử dụng công nghệ kho dữ liệu cho việc quản lí các hệ thống thông tin. Hãng máy tính IBM, các hãng phần mềm Oracle, Microsoft, Informix,... đã đưa công nghệ này vào các sản phẩm của mình.

Bài này của chúng tôi nhằm mục đích giới thiệu một số nguyên lý hoạt động và kiến trúc cơ bản của kho dữ liệu.

Trước hết, xin nhắc lại định nghĩa mô tả khái quát kho dữ liệu do W. H. Inman đề xướng. Kho dữ liệu được hiểu là một tập hợp các dữ liệu tương đối ổn định (không dễ thay đổi), cập nhật theo thời gian, được tích hợp theo hướng chủ đề nhằm hỗ trợ quá trình tạo quyết định về mặt quản lí.

Mục tiêu chính của kho dữ liệu là nhằm giải quyết những vấn đề cơ bản sau:

- Tích hợp dữ liệu và siêu dữ liệu từ những nguồn khác nhau.
- 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 hóa các nguồn dữ liệu với kho dữ liệu.
- Phân định và đồng nhất các hệ quản trị cơ sở dữ liệu tác nghiệp (dạng quan hệ hoặc phi quan hệ) như là các công cụ phục vụ cho kho dữ liệu.
- Quản lí các siêu dữ liệu.

Ở đây siêu dữ liệu được hiểu là những định nghĩa logic các bảng, các thuộc tính của kho dữ liệu, những việc định tên các nguồn dữ liệu tác nghiệp, những định nghĩa vật lý các bảng, các cột cùng những đặc trưng của chúng, định nghĩa việc cục bộ hóa và miêu tả các cơ sở dữ liệu, tên, sự miêu tả việc tổng hợp dữ liệu trong kho dữ liệu...

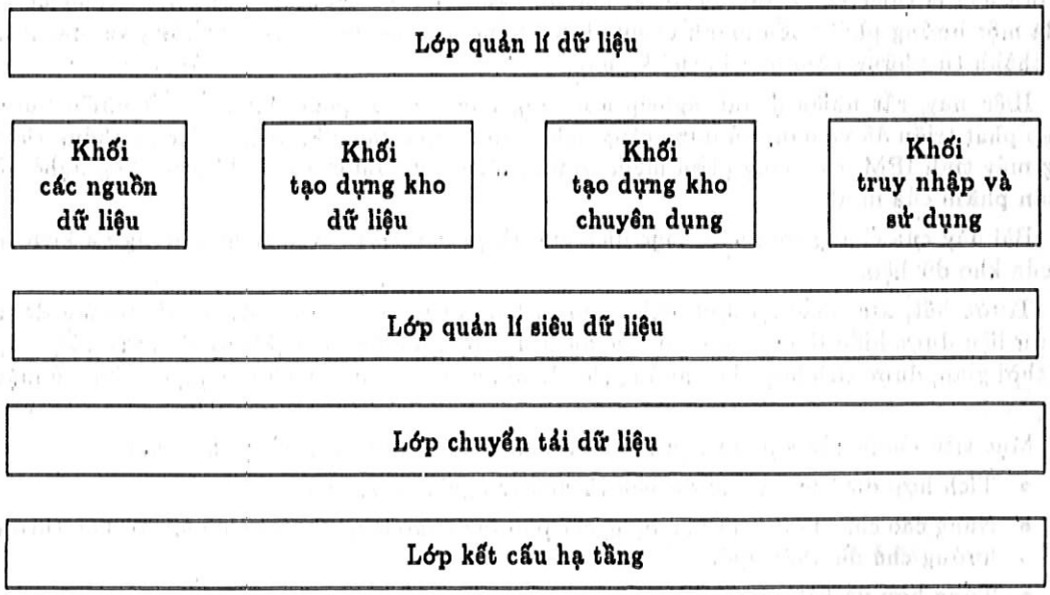
Trên cơ sở các đặc trưng của kho dữ liệu, cần phân biệt kho dữ liệu (KDL) và các hệ quản trị cơ sở dữ liệu (CSDL) tác nghiệp thông thường hiện có bán trên thị trường:

- KDL phải xác định hướng chủ đề. Nó được tổ chức và thực hiện theo ý đồ của người xây dựng cuối, trong khi các hệ quản trị CSDL tác nghiệp dùng để phục vụ ý đồ áp dụng chung.
- KDL quản lí khối lượng lớn thông tin, lưu trữ thông tin trên nhiều phương tiện lưu trữ và xử lí khác nhau. (Tất nhiên, các hệ quản trị CSDL thông thường không phải là không quản

lí những khối lượng thông tin lớn, nhưng điều cơ bản là chúng có thể quản lí cả những lượng thông tin nhỏ và vừa. Trong khi đó KDL chủ yếu là quản lí những khối lượng thông tin rất lớn và đó cũng là một đặc thù chính của KDL).

- KDL nối ghép các loại phiên bản của các loại cấu trúc CSDL. KDL tổng hợp thông tin để thể hiện chúng dưới những hình thức, những dạng biểu diễn để người dùng có thể hiểu được.
- KDL tích hợp và kết nối các thông tin từ các nguồn dữ liệu khác nhau trên nhiều loại phương tiện lưu trữ và xử lí thông tin nhằm phục vụ cho các ứng dụng xử lí tác nghiệp trực tuyến.
- KDL có thể lưu trữ các thông tin tổng hợp được tổ chức theo các chủ thể nghiệp vụ nào đó nhằm tạo ra các thông tin phục vụ hiệu quả việc phân tích.
- KDL thông thường chứa các dữ liệu lịch sử kết nối nhiều năm của các thông tin tác nghiệp được tổ chức lưu trữ có hiệu quả và hiệu chỉnh dễ dàng.

Tiếp theo, xin trình bày một kiến trúc của KDL. Trong kiến trúc này sẽ thấy rõ các thành phần cơ bản của một KDL.




Trong kiến trúc này kho dữ liệu có: khối các nguồn dữ liệu, khối tạo dựng KDL, khối tạo dựng kho chuyên dụng (data mart) và khối truy nhập và sử dụng. Đồng thời KDL cũng chứa các lớp như lớp quản lí dữ liệu, lớp quản lí siêu dữ liệu, lớp chuyển tải dữ liệu và lớp kết cấu hạ tầng.

Một điều cần lưu ý là trong kiến trúc trên các khối liên quan đến chức năng của KDL, còn các lớp thể hiện môi trường cần thiết để thực hiện các khối.

Những lớp quản lí dữ liệu và siêu dữ liệu tương ứng với các hoạt động liên quan đến thu thập, tu chỉnh, cập nhật và làm tươi dữ liệu để bảo đảm việc cung cấp thông tin cho hai khối tạo dựng kho dữ kiệu và kho chuyên dụng.

Ngoài ra, các lớp khác như lớp kết cấu hạ tầng, lớp chuyển tải dữ liệu cũng rất quan trọng cho hai khối tạo dựng KDL và kho chuyên dụng. Các lớp này sẽ cung cấp các phương pháp và các công cụ công nghệ chuẩn để triển khai hoạt động của KDL.

Khi thiết kế kho dữ liệu, các khối và các lớp quản lí dữ liệu và quản lí siêu dữ liệu là các đầu tư mới, còn các lớp chuyển tải dữ liệu và kết cấu hạ tầng là các đầu tư đã có.

Dưới đây sẽ trình bày chi tiết hơn về từng khối và từng lớp của một KDL.

### 1. Khối các nguồn dữ liệu

Bao gồm các thành phần sau:

- Dữ liệu sản phẩm: đó là các dữ liệu được chắt lọc từ các phần mềm ứng dụng và các hệ CSDL tác nghiệp đủ loại.
- Dữ liệu kế thừa: về cơ bản loại dữ liệu này có tính lịch sử. Chúng phục vụ cho quá trình phân tích dữ liệu. Mặt khác, các phương pháp khai thác dữ liệu (data mining) cũng thường xử lí trên các dữ liệu này.
- Các hệ thống dữ liệu bên trong.
- Các hệ thống dữ liệu bên ngoài
- Hệ quản lí siêu dữ liệu cho khối này.

### 2. Khối tạo dựng kho dữ liệu

Bao gồm các khối con:

*Khối con tinh chế:* liên quan đến việc nâng cao chất lượng của dữ liệu. Nó có các chức năng chính sau đối với các dữ liệu:

- Chuẩn hóa.
- Làm sạch.
- Sàng lọc.
- Tương hợp.
- Phân định thời gian cho các thông tin nguồn.
- Có cơ chế quản lí siêu dữ liệu cho khối con này.

*Khối con gia công lại:* có các chức năng chính sau:

- Tích hợp các loại dữ liệu khác từ các hệ thống để tạo ra dữ liệu mới.
- Phân dữ liệu thành các loại cho dễ xử lí.
- Tính toán sơ bộ, tổng hợp và kết xuất dữ liệu theo yêu cầu của người sử dụng.
- Chuyển đổi và hình thành lại các dữ liệu từ các nguồn khác nhau để có thể kết hợp trong cùng một dạng.
- Biến đổi và gia công lại dữ liệu lưu trữ các nguồn dữ liệu gốc.
- Có cơ chế quản lí các siêu dữ liệu.

*Khối con KDL:* có các chức năng chính sau:

- Mô hình hóa, tổng hợp và kết nối ở mức độ cao các loại dữ liệu.
- Tăng chất lượng, giá trị của dữ liệu.
- Tạo ra các dung hòa các loại dữ liệu trong KDL.
- Mô tả các loại cơ sở dữ liệu.
- Xây dựng các từ điển thuật ngữ tác nghiệp...

Về cơ bản các dữ liệu xử lí ở đây được lấy trực tiếp từ khối các nguồn dữ liệu.

### 3. Khối tạo dựng kho chuyên dụng

Dùng để tạo ra kho chuyên dụng từ các nội dung của KDL. Cũng giống như khối tạo dựng KDL, trong khối này cũng có những chức năng như khối trên nhưng thường ở mức cao hơn và có hướng chủ đề rõ ràng.

Các chức năng chính có trong khối này là:

- Tinh chế và gia công lại như khối tạo dựng KDL bằng các phương pháp sau:

- Sàng lọc các dữ liệu đã chắt lọc từ khối tạo dựng KDL.
- Tích hợp dữ liệu vào các lĩnh vực có chủ đề cụ thể.
- Tạo ra các dữ liệu tổng hợp...

• Kiến tạo các kho dữ liệu chuyên dụng bằng các phương pháp mô hình hóa, tổng hợp, kết nối, dung hòa và nâng cao giá trị chất lượng dữ liệu.

• Có cơ chế quản lí các siêu dữ liệu dùng trong khối này.

### 4. Khối truy nhập và sử dụng

Khối này bao gồm hai khối con chính là khối con truy nhập và khối con phân tích và tạo báo cáo.

Khối con truy nhập có những chức năng chính sau:

• Truy nhập trực tiếp vào khối tạo dựng KDL.

• Truy nhập vào các kho chuyên dụng.

• Gia công lại và biến đổi dữ liệu thành các loại dữ liệu có cấu trúc phức tạp hơn.

Khối con phân tích và tạo báo cáo có các chức năng chính sau:

• Tạo ra các công cụ chuẩn để tạo báo cáo, phân tích dữ liệu, mô hình hóa tác nghiệp.

• Tạo ra các phần mềm trợ giúp ra quyết định, các phần mềm khai thác dữ liệu.

Cả hai khối con này đều có cơ chế quản lí các siêu dữ liệu của chúng.

### 5. Lớp quản lí dữ liệu

Bản thân KDL là một hệ thống thông tin lớn cho nên cũng giống như các hệ quản trị CSDL tác nghiệp thông thường, việc quản lí dữ liệu đóng một vai trò rất quan trọng, nhất là phải quản lí một khối lượng rất lớn các dữ liệu lịch sử và hiện tại, mà các loại dữ liệu này bao gồm nhiều kiểu loại khác nhau, rất phong phú và đa dạng, được lưu trữ trong nhiều loại hình mang thông tin. Việc quản lí dữ liệu này tạo môi trường hoạt động cho chính các khối chức năng. Có thể thấy rằng những chức năng như nạp vào, nạp lại, trích đoạn dữ liệu, tuân thủ an toàn, lưu trữ, khôi phục dữ liệu có trong KDL là nhờ lớp quản lí dữ liệu.

Các chức năng chính ở lớp quản lí dữ liệu là:

• Sao lại các dữ liệu thích hợp từ nguồn dữ liệu đã chọn phục vụ cho việc tinh chế và gia công lại dữ liệu trong KDL.

• Giám sát và đáp ứng các đòi hỏi cho các dữ liệu mới rút từ các nguồn dữ liệu khác nhau.

• Bảo quản các dữ liệu trong các nguồn dữ liệu tác nghiệp và nạp lại hoặc cập nhật và làm sạch dữ liệu.

Mặt khác, có thể thấy lớp quản lí dữ liệu sẽ thống nhất các phương pháp quản lí dữ liệu, các thủ tục, các phép toán phục vụ cho việc an toàn, phân quyền truy nhập, lưu trữ và khôi phục dữ liệu. Việc thực hiện các xử lí song song các chất vấn (queries) và phục hồi việc sử dụng các xử lí song song cho việc truy nhập dữ liệu cũng được quản lí trong lớp này.

Với các phân tích trên, có thể thấy lớp quản lí dữ liệu có những chức năng quản lí mới khác với các chức năng của các hệ quản trị CSDL thông thường.

### 6. Lớp quản lí siêu dữ liệu

Do tính đa dạng của các kiểu loại dữ liệu và các phương pháp quản lí dữ liệu mới khác so với các hệ quản trị CSDL tác nghiệp, việc sử dụng các dữ liệu để định nghĩa và xác định các loại dữ liệu, các phương pháp xử lí, các phương pháp quản lí dữ liệu, các biểu bảng... trong KDL tăng lên rất lớn, cho nên phải tính đến việc quản lí loại dữ liệu này. Vì thế trong KDL cần phải hình thành lớp quản lí siêu dữ liệu phục vụ cho công việc lưu trữ, xử lí các dữ liệu này.

Trong việc thiết kế các KDL, các siêu dữ liệu có mặt ở khắp nơi. Các nguồn dữ liệu được đặc trưng bởi định nghĩa của các dữ liệu nhập vào, việc bổ sung các nhãn thời gian đòi hỏi phải định nghĩa các nhãn thời gian dùng trong siêu dữ liệu... Lớp quản lí siêu dữ liệu nhằm quản lí các dữ liệu dùng để mô tả đầy đủ và hoàn chỉnh các dữ liệu được lưu trữ trong KDL.

Các chức năng chính của lớp này là sao chép, tạo mới, lưu trữ, phục hồi, làm sạch và cập nhật các siêu dữ liệu sau đây:

- Các mô hình dữ liệu vật lí và logic của kho dữ liệu, kho chuyên dụng và các sơ đồ tương ứng cũng như các bảng chú giải về kỹ thuật và nghiệp vụ được lưu và quản lí trong chúng.
- Các định nghĩa dữ liệu chuẩn (bao gồm cả định nghĩa kỹ thuật và miêu tả nghiệp vụ) của các dữ liệu lưu trữ trong KDL.
- Các siêu dữ liệu được bảo quản và tạo ra trong các khối tinh chế và gia công lại.
- Các siêu dữ liệu có trong các quá trình phân đoạn, kết nối, tổng hợp...
- Các siêu dữ liệu để mô tả các báo cáo và các chất vấn.
- Các siêu dữ liệu mô tả các chỉ số, các chú giải dùng để truy nhập dữ liệu.
- Các siêu dữ liệu mô tả các luật xác định thời gian sao chép, cập nhật và nạp lại dữ liệu...

### 7. Lớp chuyển tải dữ liệu

Nhiệm vụ chuyển tải dữ liệu giữa các khối do lớp này thực hiện. Lớp này sử dụng sự nạp, sao chép, chuyển tải dữ liệu và các hệ thống mạng, các phần mềm lớp trung gian (middleware tools). Nó bảo đảm tính an toàn và phân quyền cho các nhu cầu chuyển tải dữ liệu.

Lớp chuyển tải dữ liệu xác định các cầu nối truyền thông cần thiết giữa các trang thiết bị phần cứng và phần mềm của kho dữ liệu. Lớp này có thành phần chuyển tải dữ liệu và mạng bao gồm các loại hệ thống sau:

- Các giao tác mạng, ví dụ như TCP/IP (đó là các quy định chung cho trao đổi dữ liệu).
- Các cơ chế quản lí mạng, ví dụ IBM Net View, Sunsoft's Sunnet manager.
- Các hệ điều hành mạng, ví dụ như Unix, Windows NT server.
- Các loại mạng, ví dụ như Ethernet, Tokenring,...

Lớp chuyển tải dữ liệu chứa các loại trang thiết bị sau:

- Các cổng kết nối cơ sở dữ liệu (database gateways), các thiết bị chuyển tải giữa các giao thức.
- Các phần mềm lớp trung gian hướng tới các thông báo (message oriented middleware), ví dụ như IBM MOSeries,...
- Các hệ sao chép và truyền bá, ví dụ như hệ sao chép đối xứng của hãng Oracle OSR, hệ truyền bá dữ liệu quan hệ của hãng IBM DPropR.

Các yêu cầu về an toàn dữ liệu và phân quyền truy nhập cũng được thực hiện ở trong lớp này.

### 8. Lớp kết cấu hạ tầng

Bao gồm các thành phần sau:

a) Thành phần quản lí các hệ thống: cung cấp các khả năng tìm kiếm, quản lí, xác định các phần mềm chuẩn cũng như các phần mềm ứng dụng cho người thiết kế hệ thống và người sử dụng nghiệp vụ.

b) Thành phần thứ hai của lớp này trợ giúp cho quá trình tích hợp, quản lí các phần mềm chuẩn, các phần mềm ứng dụng và hoạt động khác để sao chép, cập nhật, kết nối, tổng hợp dữ liệu... trong các khối tạo dựng KDL, tạo dựng kho chuyên dụng nhằm nâng cao hiệu quả và năng suất làm việc cho người thiết kế hệ thống cũng như người sử dụng nghiệp vụ.

c) Thành phần tiếp theo phục vụ cho công việc lưu trữ. Nó cũng cung cấp các dịch vụ quản lí cho khối các nguồn dữ liệu, các khối tạo dựng kho dữ liệu, tạo dựng kho chuyên dụng, các lưu trữ cục bộ và nhiều chiều cho khối truy nhập và sử dụng.

d) Thành phần cuối cùng của lớp này bao gồm các hệ thống xử lí. Chúng tạo ra các môi trường làm việc cho các khối chính: các nguồn dữ liệu, tạo dựng KDL, tạo dựng kho chuyên dụng.

Ngoài ra, các lớp kết cấu hạ tầng còn bao gồm các hệ thống sau:

- Các hệ quản lí cấu hình trang thiết bị.
- Các hệ quản lí việc lưu trữ.
- Các hệ quản lí an toàn dữ liệu.
- Các hệ quản lí phân phối các phần mềm.
- Các hệ quản lí các bản quyền (licence).

## Kết luận

Hiện nay một số sản phẩm phần mềm về kho dữ liệu đã được các hãng Oracle, Microsoft, Informix, IBM chào bán tại thị trường tin họcViệt Nam. Một số cơ quan đã tiến hành cài đặt các kho dữ liệu của các hãng nói trên. Mặt khác, một số đơn vị nghiên cứu, giảng dạy và đào tạo đang tiến hành nghiên cứu và thiết kế kho dữ liệu. Hy vọng rằng qua bài báo này bạn đọc sẽ có được các tri thức cơ bản về thiết kế kiến trúc và những nguyên lí hoạt động cơ bản của kho dữ liệu, phục vụ cho công việc của mình.

## TÀI LIỆU THAM KHẢO

[1] Harjinder S. Gill and Prakash C. Rao, *The Official Client/Server Computing Guide to Data Warehousing*, Que Corporation, 1996.
[2] M. Fahay, *Mining for Data Gold*, RS/Magazin, January 1995.
[3] Barry Devlin, *Data Warehouse: From Architecture to Implementation*, Addison Wesley, 1997.

*Nhận bài ngày 2 - 4 - 1998*
*Nhận lại sau khi sửa ngày 20 - 6 - 1999*

*Viện Công nghệ thông tin, Trung tâm KHTN và CNQG.*