LỜI CẢM ƠN

Trước hết, em xin chân thành gửi lời cảm ơn sâu sắc đến cô giáo ***Ths.Nguyễn Thị Xuân Hương,*** *người đã tận tình hướng dẫn và tạo mọi điều kiện cho em trong quá trình làm tốt nghiệp.*

Em xin chân thành cảm ơn các thầy cô giáo trong khoa Công Nghệ Thông Tin Trường Đại Học Dân Lập Hải Phòng đã truyền đạt những kiến thức quý báu và giúp đỡ em trong suốt bốn năm học và trong quá trình làm tốt nghiệp vừa qua.

Em xin trân trọng cảm ơn thầy ***Trần Hữu Nghị*** *- Hiệu trưởng trường Đại Học Dân Lập Hải Phòng đã ủng hộ, động viên, và tạo mọi điều kiện tốt nhất cho chúng em trong thời gian học tập tại trường.*

Cuối cùng tôi xin gửi lời cảm ơn chân thành tới tất cả những người thân cùng bạn bè đã động viên, giúp đỡ và đóng góp nhiều ý kiến quý báu cho tôi trong quá trình học tập cũng như khi làm tốt nghiệp.

Hải Phòng, tháng 7 năm 2010

Sinh viên

Nguyễn Thị Mai Hương

MỤC LỤC

LỜI NÓI ĐẦU

Khi một doanh nghiệp đi vào hoạt động, những nhà quản lý doanh nghiệp sẽ phải đặt các câu hỏi và có nhu cầu muốn biết về tình hình kinh doanh, tốc độ tăng trưởng, lượng giao dịch hàng ngày, hàng tháng, hàng quý, hàng năm, so sánh giữa năm này, năm khác, hoặc phân khúc các khách hàng của doanh nghiệp, hoặc phân tích doanh thu.

Đối với mỗi doanh nghiệp, họ sẽ tự xây dựng cho mình một hệ thống quản lý giao dịch (OLTP – Online Transaction Procesing) hay chính là các ứng dụng (applications), chương trình (software), hệ thống vận hành (system) hàng ngày của doanh nghiệp. Ví dụ như các ngân hàng, các công ty viễn thông (họ thường phải thuê xây dựng hệ thống chuyên biệt). Tuy nhiên các hệ thống này chỉ được thiết kế cho việc nhập dữ liệu hàng ngày hoặc để vận hành hệ thống. Chúng cũng có khả năng cho phép lấy dữ liệu cho một số báo cáo đơn giản.

Tuy nhiên đối với những yêu cầu báo cáo theo nhiều chiều như: loại khách hàng, theo thời gian, đòi hỏi phải tính toán phức tạp thì hầu như các hệ thống này rất khó thực hiện.

Mặt khác các doanh nghiệp lớn như ngân hàng, viễn thông, họ phải có nhiều hệ thống con vận hành song song với nhau. Ví dụ: ngân hàng thì có phân hệ tiền gửi (cá nhân, sổ tiết kiệm), tiền vay, kho quỹ. Viễn thông thì có trả trước, trả sau, bán hàng. Như thế, để thực hiện được việc báo cáo, họ phải tổng hợp dữ liệu từ nhiều hệ thống con khác nhau mới có thể thể thiện được các báo cáo một cách tổng thể.

Xuất phát từ những vấn đề trên, họ phải bắt buộc xây dựng một hệ thống nữa, chính là một cơ sở dữ liệu mới dành cho việc truy vấn và báo cáo ở phạm

vi toàn doanh nghiệp. Hay còn gọi là kho dữ liệu, là nơi tổng hợp dữ liệu từ tất cả các hệ thống con lại, thực hiện việc tính toán trên các dữ liệu này và kết xuất ra các bảng mà dữ liệu của bảng đã được tính toán theo một mục đích nào đó.

Kho dữ liệu là một hướng công nghệ mới được sử dụng phổ biến cho các bài toán lớn hiện nay như: quản trị doanh nghiệp, Y tế, bảo hiểm, ngân hàng, dân số, viễn thông. Bởi vì việc xây dựng kho dữ liệu không những giúp cho doanh nghiệp lưu trữ một lượng thông tin lớn hằng ngày mà còn giúp cho các nhà quản lý doanh nghiệp có thể trích rút nguồn tài nguyên một cách nhanh chóng, chính xác. Đồng thời giúp họ phân tích và đưa ra các báo cáo một cách kịp thời, góp phần thúc đẩy cho việc kinh doanh đạt kết quả tốt nhất.

Đây cũng là kiến thức rất hữu ích và cần thiết để có thể khai thác ngày một hiệu quả các thành tựu tin học. Đó cũng là lý do em chọn đề tài này làm đồ án tốt nghiệp. Đề tài gồm có 4 chương:

Chương 1: Giới thiệu về Kho dữ liệu (Data warehouse),

Chương 2: Các yếu tố cơ bản của Kho dữ liệu,

Chương 3: Giới thiệu kiến trúc logic của Kho dữ liệu,

Chương 4: Giới thiệu về Ngôn ngữ cho kho dữ liệu: trong chương này giới thiệu về OLAP và trình bày một ví dụ xây dựng kho dữ liệu.

Và cuối cùng là phần kết luận.

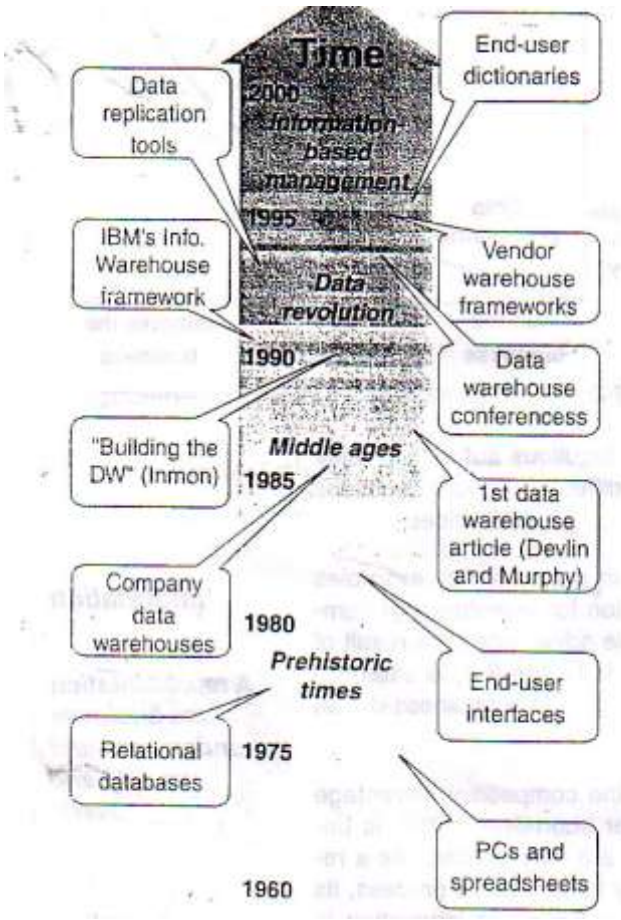
Chương 1. GIỚI THIỆU VỀ KHO DỮ LIỆU

1.1. Lịch sử phát triển của kho dữ liệu

Khái niệm của kho dữ liệu xuất phát từ việc tổng hợp của hai tập nhu cầu:

- Yêu cầu thương mại cho công ty mở rộng về bối cảnh thông tin.
- Sự cần thiết của các hệ thống thông tin trong lĩnh vực quản lý dữ liệu công ty một cách tốt nhất.

Vào những năm 1990, kho dữ liệu trở thành một từ thông dụng của công nghiệp máy tính.



Hình 1:Data warehouse evolution

Các cuộc cách mạng dữ liệu đầu năm 1990:

Phần lớn các kho triển khai trong thời kỳ này đã được khai sáng bởi các tổ chức hệ thống thông tin. Có thể thấy rằng các phương pháp tiếp cận trước đó không đủ mạnh để cung cấp các dữ liệu hỗ trợ cho sự phát triển trong tương lai và khả năng người sử dụng các dữ liệu sẽ bị suy yếu do thiếu điều kiện doanh

nghiệp. Sự thành công của thực hiện này đã thuyết phục của các nhà quản lý hệ thống thông tin, những người bán khái niệm cho doanh nghiệp.

Tiếp cận mới này phụ thuộc vào cộng đồng doanh nghiệp trong sự việc nhận ra sự cần thiết và giá trị của tầm nhìn khái quát về dữ liệu kinh doanh hơn khả năng đã có trước đó.

Đặc biệt, có một chủ đề phổ biến là sử dụng dữ liệu cho việc tiếp thị và tăng cường lợi thế cạnh tranh.

Vào đầu thời kỳ này, nhiều ngành công nghiệp đã bị thay đổi đáng kể trong môi trường kinh doanh. Quốc tế suy thoái cắt giảm lợi nhuận, các chính phủ bãi bỏ các kiểm soát chặt chẽ các ngành công nghiệp, sự gia tăng cạnh tranh trong thị trường hàng hóa, chính phủ thay thế thị trường tập trung bằng kinh tế thị trường nhiều thành phần.

Điều này cho thấy các yêu cầu về kinh doanh dẫn đến cuộc cách mạng về dữ liệu. Công việc kinh doanh cần đến tầm nhìn mới về việc công ty được vận hành như thế nào, nó bao trùm các định hướng phân chia trước đó của công việc kinh doanh.

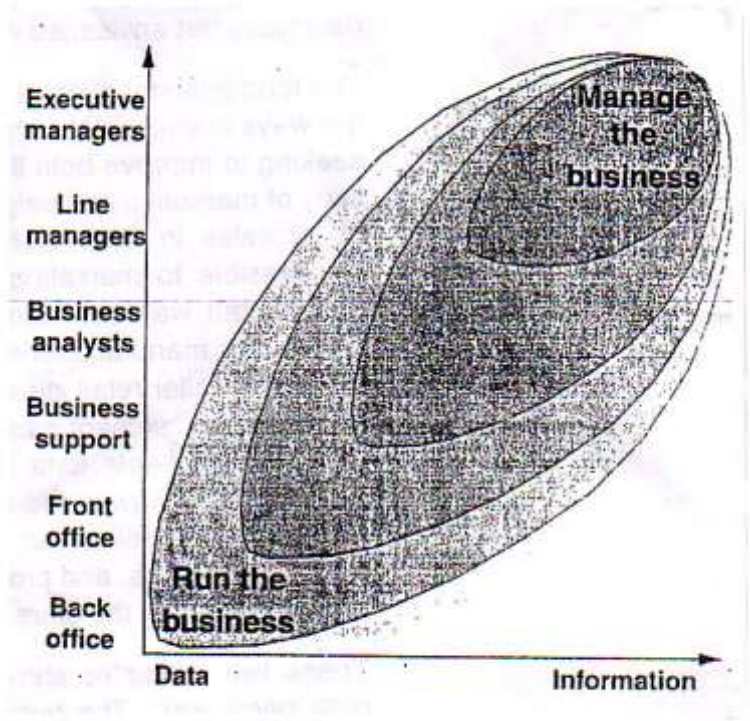
Sự thay đổi tập trung vào kho điều khiển dữ liệu kinh doanh thực hiện tạo điều kiện cho việc đánh giá lại các lợi ích mà kho có thể cung cấp. Đặc tính của kho dữ liệu trong thời kỳ này, khi hệ thống thông tin được điều khiển thực hiện, được giả định là kho là đúng đắn bằng tiết kiệm về giá và hiệu quả được cải thiện. Sự xuất hiện này từ các tiếp cận hệ thống thông tin truyền thống để điều chỉnh chi phí, dựa trên tính vững chắc trong mô hình điều khiển ứng dụng.

Thời đại của thông tin dựa trên quản lý trong thế kỷ 21:

Phân tích về mặt lý thuyết và việc thực hiện của kho dữ liệu đã phát triển mạnh từ những năm 2000 trở về đây. Tuy nhiên, những bí quyết kinh doanh, được hỗ trợ bởi những chỉ dẫn kỹ thuật, đã được định nghĩa trước đây vẫn có thể được xem như là những chỉ dẫn quan trọng ngày nay.

Hiện nay, chúng ta đang sử dụng các dữ liệu nguồn để dự đoán tương lai. Chìa khóa cho việc dự đoán này là công nhận sự sự cần thiết của lợi thế

cạnh tranh là điều khiển hỗ trợ cơ bản cho việc ra quyết định từ dữ liệu hướng đến thông tin, và mở rộng đối tượng hỗ trợ vượt ra ngoài ranh giới của thị trường quản lý truyền thống.



Hình 2:From data to information

Hướng này có thể được đặc trưng bởi thuật ngữ: Quản lý thông tin cơ sở (Information-based management viết tắt là IMB). Là sự chuyển đổi cách hỗ trợ quyết định được giao cho cộng đồng người dùng cuối. Nó có thể được tổng hợp thành năm chủ đề sau đây:

1. Một nguồn thông tin duy nhất:

Các dữ liệu thô mong muốn từ nhiều nguồn khác nhau, gồm dữ liệu trong và dữ liệu ngoài công ty, và tồn tại ở nhiều dạng, từ dữ liệu có cấu trúc truyền thống, dữ liệu phi cấu trúc, loại tài liệu hoặc đa phương tiện,.. Dù nguồn dữ liệu có kiểu hay dữ liệu thô, trước khi được đưa vào môi trường người dùng cuối, nó phải được làm sạch và tương thích để đảm bảo chất lượng và tính toàn vẹn của nó. Thông tin tương thích là duy nhất, là nguồn thông tin cuối cùng cho quản lý thông tin cơ bản.

2. Phân phối thông tin sẵn có:

Quản lý thông tin cơ bản không chỉ duy nhất một chức năng chính, nhưng được đánh giá cao về tổ chức phân bố và vị trí địa lý. Các hoạt động này có thể cần thiết, và thường yêu cầu độc lập, nhưng các kho thông tin kết nối logic để dễ dàng thay đổi, sự thực hiện, tăng cường độ tin cậy.

3. Thông tin trong một bối cảnh kinh doanh:

Người dùng có thể hiểu tốt nhất và xử lý thông tin khi nó được đặt trong bối cảnh hoạt động kinh doanh mà họ tham gia. Các định nghĩa dữ liệu được cung cấp bởi các chuyên gia kinh doanh trở thành chuẩn, và danh mục các thông tin bao gồm các định nghĩa và hướng vào người dùng cuối để trở thành nguồn cho các định nghĩa dữ liệu và hệ thống thông tin doanh nghiệp.

4. Truyền thông tin tự động:

Dữ liệu được chuyển thành thông tin và chuyển thông qua con đường ngày càng phức tạp trong và giữa các tổ chức, cơ chế truyền tự động là cần thiết. Tự động hóa cần thiết không chỉ trong quá trình truyền thực tế mà còn trong việc định nghĩa các chuyển đổi dữ liệu cần thiết và sự di chuyển. Đặc biệt trong lĩnh vực phân phối thông tin, các tiện ích của các cơ chế này tự động phân phối phải được bảo đảm.

5. Chất lượng thông tin và quyền sở hữu (Information quality and ownership)

Thông tin là một sở hữu quan trọng của công ty bất kỳ, và giống như bất kỳ sở hữu khác, đó là phải quản lý và bảo vệ. Chất lượng của nó phải được đảm bảo. Quyền sở hữu của tài liệu và thông tin theo dõi là một điều kiện tiên quyết để nhận thức rõ giá trị của sở hữu này.

Môi trường phát triển ngày nay(Today's development environment)

1. Phát triển ứng dụng phân tán (Fragmented application develop)

Tất cả các công cụ mới và các công nghệ đều được ứng dụng tại các doanh nghiệp. Tuy nhiên, các công cụ mới rất tốn kém và phải được áp dụng ở các khu vực để mang lại lợi tức đầu tư lớn nhất. Các công nghệ mới chưa được thử nghiệm sẽ có nguy cơ thất bại, vì vậy nó phải bao gồm phương pháp

tiếp cận mới được thực hiện trong một loạt các dự án thí điểm. Điều này cũng được áp dụng trong lý dữ liệu.

Các yếu tố này, cùng với sự quản lý có giới hạn của con người dẫn đến thực hiện phân mảnh quá trình xử lý dữ liệu trong tất cả các hoạt động kinh doanh. Doanh nghiệp hoặc đơn vị, địa phương, tổ chức, có các ứng dụng vận hành riêng để thực hiện những phần của doanh nghiệp họ đảm nhận. Phân mảnh này có thể được thấy trong các ví dụ như sau:

- Các ứng dụng đặt hàng khác nhau được sử dụng cho dòng sản phẩm khác nhau trong cùng một công ty.

- Một quá trình hợp lý liên tục từ đặt hàng thông qua đơn để thanh toán được tách ra trên một số ứng dụng độc lập dựa trên trách nhiệm của tổ chức.

Sự phân đoạn này đem lại một số lợi ích. Với các ứng dụng độc lập tập trung vào việc phân chia vùng của chức năn kinh doanh, các dự án có thể nhận được chức năng ứng dụng để xác định nhóm người dùng cuối với các yêu cầu định nghĩa chuẩn.

2. Phát triển ứng dụng vận hành (Operational application development)

Môi trường vận hành được điều khiển bởi các nhu cầu của doanh nghiệp để cung cấp hàng hoá hoặc dịch vụ. Do đó nó được xác định chủ yếu bởi các hoạt động cần thiết hơn là bởi các dữ liệu được sử dụng. Sự cần thiết của người dùng được mô tả trên cơ sở các hoạt động ngắn hạn. Phân tích có thể tập trung vào những gì là cần thiết để nhận một đơn đặt hàng, một lịch trình giao hàng, và tương tự như vậy. Hệ thống thông tin có thể tập trung vào các yếu tố đầu vào và đầu ra cần thiết và các hoạt động xung quanh. Các hoạt động cá nhân có thể dẫn đến các ứng dụng độc lập, mỗi tối ưu hóa cho các nhu cầu của hoạt động liên quan của nó. Yêu cầu người sử dụng ở đây có thể được tổng hợp như "tự động hoá các thủ tục này". Sự thành công của tự động hóa được đánh giá trên các phép đo đơn giản bằng việc thông qua mức tăng hoặc giảm chi phí trong kinh doanh và về tính dễ sử dụng hoặc thời gian phản hồi ở cấp độ của người sử dụng.

Mô hình này đã được sử dụng thành công để xử lý dữ liệu. Hầu hết các tính toán kinh doanh đã được hướng vào các hệ thống hoạt động. Hệ thống thông tin có tầm nhìn hướng ứng dụng. Một ứng dụng đơn giản là một tập các chức năng cho người sử dụng có liên quan và được phát triển trong một số cách tích hợp. Tuy nhiên, Hệ thống thông in tích hợp các chức năng xác định làm thế nào phát triển được phạm vi của dữ liệu trong các ứng dụng.

3. Hỗ trợ quyết định điều khiển ứng dụng (Application – driven decision support):

Từ khi ứng dụng thông tin được sử dụng rộng rãi trên hệ thống máy tính, có một khối lượng lớn dữ liệu được lưu trữ và xử lý trên máy tính. VẤn đề ứng dụng thông tin hiện nay không chỉ là lưu trữ vận hành dữ liệu, mà còn là việc tổ chức các nguồn dữ liệu đó để rút trích thông tin và hỗ trợ ra quyết định. Đây chính là một sự tiến hóa cần thiết cho các hệ thống thông tin.

1.2.Kho dữ liệu là gì (What is the data warehouse)?

Kho dữ liệu (data warehouse), gọi một cách chính xác hơn là kho thông tin (information warehouse), là một cơ sở dữ liệu hướng đối tượng được thiết kế với việc tiếp cận các ý kiến trong mọi lĩnh vực đặc biệt là trong lĩnh vực kinh doanh. Nó cung cấp các công cụ để đáp ứng thông tin cần thiết cho các nhà quản trị kinh doanh tại mọi cấp độ tổ chức - không những là những yêu cầu dữ liệu phức hợp, mà còn là điều kiện thuận tiện nhất để đạt được việc lấy thông tin nhanh, chính xác. Một kho dữ liệu được thiết kế để người sử dụng có thể nhận ra thông tin mà họ muốn có và truy cập đến bằng những công cụ đơn giản.

Một kho dữ liệu là một sự pha trộn của nhiều công nghệ, bao gồm các cơ sở dữ liệu đa chiều và mối quan hệ giữa chúng, kiến trúc chủ khách, giao diện người dùng đồ họa và nhiều nữa. Nguyên nhân chính cho sự phát triển một kho dữ liệu là hoạt động tích hợp dữ liệu từ nhiền nguồn khác nhau vào một kho dữ liệu đơn lẻ và dày đặc mà kho này cung cấp cho việc phân tích và ra quyết định trong công việc kinh doanh, quản lý.

Đối với một số công việc kinh doanh tin rằng thông tin là nguồn tài nguyên có giá trị rất lớn thì một kho dữ liệu tương đối giống như một nhà kho chứa hàng. Hệ điều hành tạo ra những phần dữ liệu và nạp chúng vào kho.

Một số phần được tóm tắt trong thành phần thông tin và được cất vào kho. Người sử dụng kho dữ liệu đưa ra những yêu cầu và được cung cấp sản phẩm được tạo ra từ các thành phần và các phân đoạn được lưu trong kho.

Kho dữ liệu là một hướng công nghệ nóng nhất. Một kho dữ liệu được xác định đúng hướng, hoạt động hiệu quả có thể trở thành một công cụ cạnh tranh có giá trị cao trong kinh doanh.

1.3. Đặc điểm

Trước tiên Data Warehouse là cơ sở dữ liệu rất lớn (very large database-VLDB). Data Warehouse thường chỉ đọc, phục vụ cho những nhu cầu báo cáo, Data Warehouse hướng về tính ổn định.

Data Warehouse sẽ lấy thông tin có thể từ nhiều nguồn khác nhau: DB2, Oracle, SQLserver thậm chí cả File thông thưởng rồi làm sạch chúng và đưa vào cấu trúc của nó-đó là VLDB(very large database).

Data Warehouse rất lớn nên muốn cho từng bộ phận chuyên biệt người sử dụng cuối cùng có thể khai thác thông dễ dàng thì bản thân Data Warehouse phải được chuyên hoá, phân ra thành những chủ đề, do đó những chủ đề chuyên môn hóa đó tạo thành một cơ sở dữ liệu chuyên biệt-đó là Data marts. Có một điểm lưu ý ở đây là có một công cụ hay đúng hơn là một chuẩn công cụ mà mọi hệ quản trị cơ sở dữ liệu hỗ trợ cho việc truy vấn thông tin trong Data marts rồi đưa ra những quyết định, nhận dịnh những thông tin trong Datamart - Đó là OLAP, bộ phân tích trực tuyến (Online Analyze Proceesing).

1.4. Mục đích của kho dữ liệu

Mục đích chính của kho dữ liệu là:

- 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, năng suất cao 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

1.5. Mục tiêu của kho dữ liệu

Một Data Warehouse phải đảm bảo được các mục tiêu sau:

1.5.1. Truy cập dễ dàng

Thông tin lưu trữ trong DW phải trực quan và dễ hiểu với người dùng. Dữ liệu nên được trình bày thông qua các tên gọi quen thuộc và gần gũi với nghiệp vụ của người dùng.

Tốc độ truy cập data warehouse phải nhanh. Do phải xử lý một số lượng bản ghi lớn cùng một lúc nên đây là một trong những yêu cầu cần phải có của một DW

1.5.2. Thông tin nhất quán

Dữ liệu trong một DW thường đến từ nhiều nguồn khác nhau. Do vậy trước khi được đưa vào DW dữ liệu cần phải được làm sạch và đảm bảo về chất lượng. Việc làm sạch sẽ giúp cho việc đồng nhất dữ liệu trở nên dễ dàng.

Một nguyên tắc được đặt ra cho qúa trình này là:

- Nếu dữ liệu có cùng tên thì bắt buộc phải chỉ đến cùng một địa chỉ.
- Nếu dữ liệu chỉ đến các thực thể khác nhau thì phải được đặt tên khác nhau.

1.5.3. Thích nghi với sự thay đổi

DW cần phải được thiết kế để xử lý những thay đổi có thể xảy ra. vì thay đổi là diều không thể tránh khỏi cho bất cứ ứng dụng nào. Nói vậy có nghĩa là khi có thay đổi mới dữ liệu cũ trong DW vẫn phải đảm bảo tính đúng đắn.

1.5.4. Hỗ trợ ra quyết định

Đây là mục tiêu quan trọng nhất của doanh nghiệp khi xây dựng DW. Những người quản lý doanh nghiệp muốn dưa vào thông tin để từ đó đưa ra những chiến lược góp phần đem lại kết quả kinh doanh tốt nhất.

1.5.5. Bảo mật

Dữ liệu trong DW đến từ nhiều nguồn khác nhau. Vì vậy việc đảm bảo thông tin không bị lộ ra ngoài là một điều vô cùng quan trọng.

1.6. Các chức năng chính:

1. Phân hệ tích hợp dữ liệu
2. Phân hệ phân tích dữ liệu
3. Phân hệ giám sát hệ thống
4. Phân hệ sao lưu và phục hồi hệ thống
5. Phân hệ bảo mật dữ liệu

1.7.Lợi ích:

*** Đối với người khai thác:**

- Cung cấp công cụ phân tích, khai thác dữ liệu nhanh gọn, đầy đủ và chính xác, dễ dàng đưa ra các chính sách mới.
- Giúp người sử dụng khai thác dữ liệu theo chủ đề với các nguồn và khoảng thời gian khác nhau
- Dữ liệu được xử lý nhanh chóng
- Dễ dàng tạo ra các báo cáo đơn giản phù hợp với nhiều trình độ khai thác

*** Đối với người quản trị hệ thống:**

- Hỗ trợ xây dựng một kho dữ liệu lớn
- Thiết kế mềm dẻo giúp dễ dàng tích hợp dữ liệu tác nghiệp mới và tạo ra các báo cáo mới theo yêu cầu người khai thác.

1.8. Đặc tính của kho dữ liệu

Kho dữ liệu (DW) là một tập hợp dữ liệu có tính chất sau:

***Tính tích hợp (Integration);**Dữ liệu tập hợp từ nhiều nguồn khác nhau. Điều này sẽ dẫn đến việc quá trình tập hợp phải thực hiện việc làm sạch, xắp xếp, rút gọn dữ liệu.*

***Dữ liệu gắn thời gian và có tính lịch sử.** Các dữ liệu đến từ quá trình kinh doanh của công ty có thể có từ nhiều năm trước.

*****Dữ liệu có tính ổn định (nonvolatility):*: ***Khi một Transaction hoàn chỉnh, dữ liệu không thể tạo thêm hay sửa đổi.***

*****Dữ liệu không biến động

*****Dữ liệu tổng hợp

Dữ liệu tổng hợp nhanh (lightly summarized data) là dấu hiệu xác nhận chất lượng của một kho dữ liệu. Tất cả các yếu tố của công việc kinh doanh (phòng ban, lĩnh vực hoạt động, chức năng hoạt động, …) có những yêu cầu thông tin khác nhau, vì thế

việc thiết kế kho dữ liệu phải có kết quả cung cấp dữ liệu tuỳ biến, tổng hợp nhanh cho mỗi yếu tố doanh nghiệp (xem thêm phần kho dữ liệu thông minh bên dưới). Mỗi yếu tố của công việc kinh doanh có thể có truy cập đến dữ liệu chi tiết và tổng hợp, nhưng sẽ không có nhiều hơn tổng số dữ liệu được lưu trữ trong chi tiết hiện hành.

Dữ liệu tổng hợp chất lượng cao (hightly summarized data) là căn bản cho việc tiến hành công việc kinh doanh. Dữ liệu tổng hợp chất lượng cao có thể đến từ dữ liệu tổng hợp nhanh được dùng cho các yếu tố công việc kinh doanh hoặc từ chi tiết hiện hành. Số lượng dữ liệu ở mức độ này có ít hơn ở các mức độ khác, nó mô tả một tập hợp được chọn lọc cung cấp một sự sự đa dạng rộng lớn cho các nhu cầu và các sự quan tâm. Thêm vào đó để truy cập đến dữ liệu tổng hợp chất lượng cao, việc tiến hành nói chung cũng cần có khả năng tăng mức độ cập nhật chi tiết thông qua tiến trình khoan đi xuống (drill down).

1.9.Cấu trúc dữ liệu cho kho dữ liệu

Vì dữ liệu trong kho dữ liệu rất lớp và không có những thao tác như sửa đổi hay tạo mới nên nó được tối ưu cho việc phân tích và báo cáo.

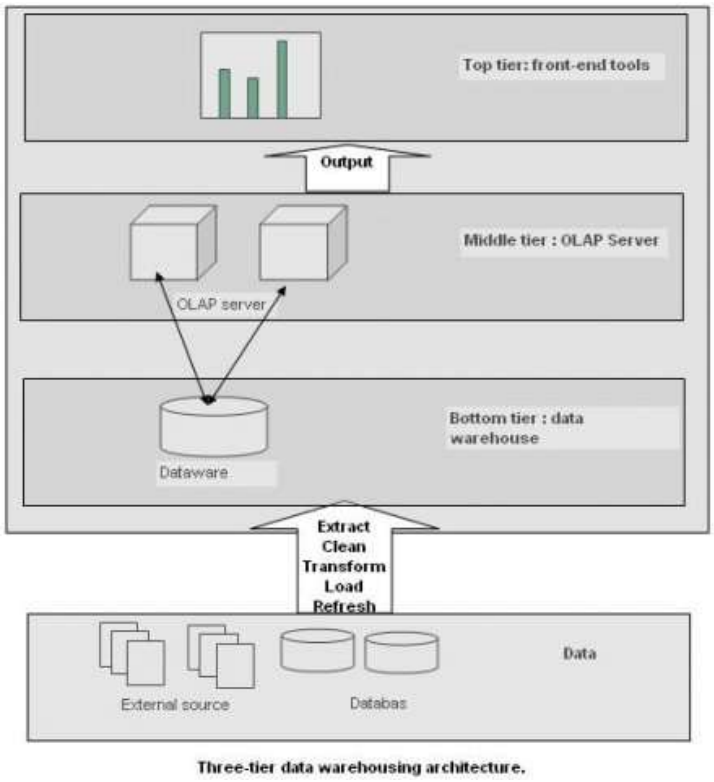
Các thao tác với dữ liệu của kho dữ liệu dựa trên cơ sở là Mô hình dữ liệu đa chiều (multidimensional data model), được mô hình vào đối tượng gọi là data cube.

Data cube là nơi trung tâm của vấn đề cần phân tích, nó bao gồm một hay nhiều tập dữ kiện (fact) và các dữ kiện được tạo ra từ nhiều chiều dữ kiện khác nhau (dimention).

Ví dụ: Một thống kê doanh số bán hàng dựa trên ba yếu tố là: địa điêm, thời gian và chủng loại hàng. Data cube là vấn đề "Thống kê bán hàng" với ba chiều là ba yếu tố: địa điểm, thời gian và chủng loại hàng. Bảng fact là bảng tổng hợp dữ liệu của mối liên quan của doanh số với 3 yếu tố.
trong SQL).

1.10.Kiến trúc của một hệ thống kho dữ liệu

Kiến trúc kho dữ liệu mô tả các cấu kiện, công cụ và dịch vụ của kho dữ liệu, cũng như quan hệ và sự phát triển của chúng. Mục đích của việc chuẩn hoá kiến trúc kho dữ liệu là tích hợp các hệ thông tin cấp dưới để phục vụ các hệ thông tin cấp trên và ngược lại. Kiến trúc này cung cấp một cơ chế tổ chức dữ liệu, cải thiện việc chia sẻ thông tin giữa các cơ quan và về lâu dài có khả năng tái sử dụng dữ liệu cũng như phát triển các dự án kho dữ liệu tiếp theo được nhanh hơn.



Hình 3:**Cấu trúc 3 lớp của kho dữ liệu**

Bao gồm ba tầng :

- Tầng đáy : Là nơi cung cấp dịch vụ lấy dữ liệu từ nhiều nguồn khác sau đó chuẩn hóa, làm sạch và lưu trữ dữ liệu đã tập tung.
- Tầng giữa : cung cấp các dịch vụ để thực hiện các thao tác với kho dữ liệu gọi là dịch vụ OLAP (OLAP server). Có thể cài đặt bằng Relational OLAP, Multidimensional OLAP hay kết hợp cả hai mô hình trên Hybrid OLAP.
- Tầng trên cùng : nơi chứa các câu truy vấn, báo cáo, phân tích.

1.11.Mối quan hệ giữa kho dữ liệu và khai phá dữ liệu

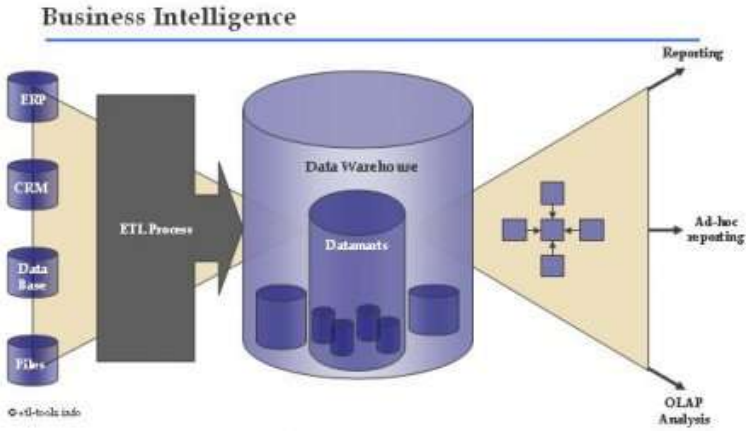
Cả hai đều có thể đứng độc lập với nhau, tuy nhiên khi kết hợp được kho dữ liệu với khai phá dữ liệu thì lợi ích rất lớn vì các lý do như :

- Dữ liệu của kho dữ liệu rất phù hợp cho việc khai phá dữ liệu (Data Mining) do đã được tập hợp và làm sạch.
- Cơ sở hạ tầng của kho dữ liệu hỗ trợ rất tốt cho các việc như xuất, nhập cũng như các thao tác cơ bản trên dữ liệu.
- OLAP cung cấp các tập lệnh rất hữu hiệu trong phân tích dữ liệu.

1.12.Các lĩnh vực ứng dụng

Có thể đưa kho dữ liệu vào ba hướng ứng dụng chính cần đến trí tuệ kinh doanh (Business Intelligence):

- Xử lý thông tin như tạo ra các báo cáo và trả lời các câu hỏi định trước.
- Phân tích và tổng hợp dữ liệu, kết quả được thể hiện bằng các báo cáo và bảng biểu.
- Dùng cho các dự án có mục đích kế hoạch hoá như khai phá dữ liệu.



Hình 4: **Ứng dụng kiểu Business Intelligence**

Các lĩnh vực hiện tại có ứng dụng kho dữ liệu bao gồm:

- Thương mại điện tử.
- Kế hoạch hoá nguồn lực doanh nghiệp (ERP - Enterprise Resource Planning).
- Quản lý quan hệ khách hàng (CRM - Customer Relationship Management)
- Chăm sóc sức khỏe.
- Viễn thông.

Chương 2. CÁC YẾU TỐ CƠ BẢN CỦA KHO DỮ LIỆU

2.1.Kiểu của dữ liệu và cách sử dụng

2.1.1. Kiểu của dữ liệu (Types of data)

2.1.1.1. Ý nghĩa

Dữ liệu cơ bản của máy tính đã được sử dụng từ lâu để vận hành và quản lý một doanh nghiệp. Dữ liệu này được gọi là dữ liệu công việc (thương mại), đặc trưng cho trạng thái của Doanh nghiệp.

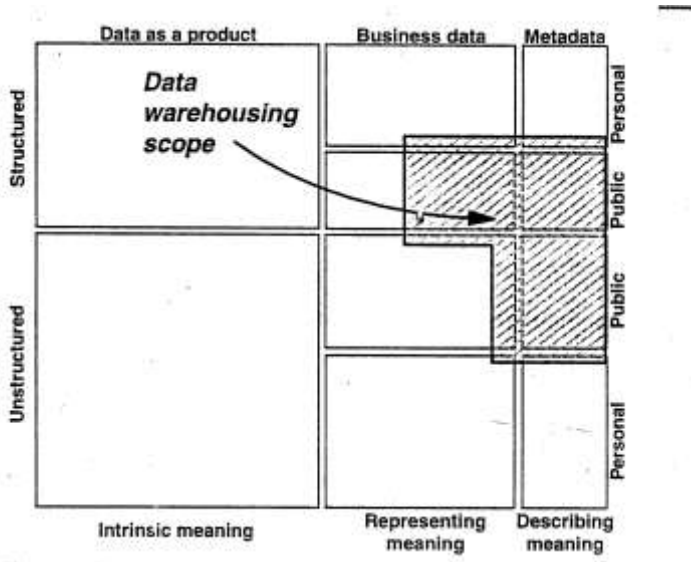
Một kiểu khác của dữ liệu là khái niệm về tầm quan trọng của dữ liệu, giá trị của dữ liệu nằm trong nội dung của nó hơn là giá trị mà nó thể hiện. Kiểu dữ liệu này được gọi dữ liệu một sản phẩm, bởi vì nó đã được sản xuất, được mua, và được bán như bất kì một sản phẩm vật lý nào. Ví dụ như phim ảnh hoặc sách được lưu trữ dạng số.

Ở mức cuối cùng chính là siêu dữ liệu(Metadata), nó dùng để mô tả ý nghĩa của dữ liệu. Siêu dữ liệu này chỉ được định nghĩa hoặc mô tả dữ liệu công việc hoặc dữ liệu như một sản phẩm.

2.1.1.2. Cấu trúc

Dữ liệu có thể có cấu trúc ở mức cao, bao gồm định nghĩa hoàn chỉnh liên quan đến các trường hoặc các bản ghi, hoặc không có cấu trúc, khi mà cẩu trúc nội bộ là rất biến động, hoặc nó có thể nằm ở giữa hai kiểu trên.

2.1.1.3. Phạm vi(Scope)



Hình 5: Types of data and the scope of the warehouse

Dữ liệu có thể là dữ liệu cá nhân, khi đó chủ nhân của nó có thể thay đổi nó theo ý muốn của mình, hoặc công cộng - nơi sử dụng của nó là chia sẻ giữa một số người sử dụng và bất kỳ thay đổi theo yêu cầu phải được quản lý cẩn thận.

2.1.2. Dữ liệu công việc (Business data)

2.1.2.1.Định nghĩa

Dữ liệu công việc là dữ liệu được sử dụng trong công việc kinh doanh và trong công tác quản lý của các doanh nghiệp hoặc tổ chức. Nó thể hiện hoạt động của doanh nghiệp đảm nhận hoặc các đối tương trong thế giới thực như: các khách hàng, các vị trí, các sản phẩm, với các cách giải quyết của nó.

Dữ liệu công việc được tạo ra và sử dụng trong hệ thống xử lý chuyển tiếp và hệ thống hỗ trợ quyết định. (DSS)

2.1.2.2. Tiêu chuẩn cho kiểu của dữ liệu công việc:

Có 4 tiêu chuẩn được sử dụng để xác định các kiểu của dữ liệu công việc. Bao gồm: sử dụng chúng trong Doanh nghiệp, phạm vi của dữ liệu, có hay không đọc/ghi hoặc chỉ đọc dữ liệu, và giá trị của dữ liệu.

❖ Giá trị sử dụng trong Doanh nghiệp

Dữ liệu được sử dụng trong doanh nghiệp nhằm đạt tới hai đối tượng sau:

- Dữ liệu vận hành (Operational Data): được sử dụng để vận hành doanh nghiệp và có quan hệ tới các hoạt động và các quyết định.
- Dữ liệu thông tin được sử dụng để quản lý doanh nghiệp.

❖ Phạm vi của dữ liệu

Dữ liệu có thể thể hiện một thông tin đơn hoặc một giao dịch, hoặc nó có thể tổng kết hiệu quả của tập các thông tin hoặc các giao dịch.

- Dữ liệu chi tiết (detailed data)hoặc dữ liệu nguyên tử (atomic data) là mục tiêu để quản lý doanh nghiệp, nhưng nó cũng sử dụng trong một số nhiệm vụ quản lý doanh nghiệp đơn giản. Nó thường tập trung vào các đối tượng cơ bản hoặc giao dịch cơ bản như các sản phẩm cá nhân, các yêu cầu, các khách hàng.

- Dữ liệu tổng hợp (Summary data) được sử dụng trong quản lý và hiển thị tổng quan các cách vận hành doanh nghiệp.

❖ Lựa chọn Đọc/ghi hay chỉ đọc dữ liệu

- Đọc/ghi dữ liệu yêu ầu thiết kế cẩn thận trong tiến trình cập nhật và phải chắc chắn rằng các luật an toàn cho doanh nghiệp phải được thực hiện.
- Chỉ đọc dữ liệu: thường được thiết kế với việc không yêu cầu ghi lại và cung cấp cơ bản là đọc nhiều lần.

❖ Giá trị của dữ liệu:

- Dữ liệu hiện tại (current data): là một cách nhìn về thương mại trong thời điểm hiện tại. Nó đạt tới mức thứ hai và là đối tượng có thể thay đổi theo thời gian dựa trên các hoạt động thương mại. Nó thể hiện biểu diễn chính xác của sự thực hiện hiện tại của doanh nghiệp.
- Dữ liệu thời điểm (Point-in-time data): là sự ổn định ngắn của dữ liệu công việc tại một thời điểm hiện tại và phản ánh trạng thái của công việc tại thời điểm hiện tại. Dữ liệu công việc hàng ngày và tập dữ liệu hàng tháng, dữ liệu này có thể thể hiện trong quá khứ hoặc dự đoán, thể hiện kế hoạch hoặc các sự kiện dự đoán trong tương lai.
- Dữ liệu định kỳ (periodic data) là lớp dữ liệu tương lai quan trọng. Nó cung cấp bản ghi định nghĩa của công việc như các thay đổi chu kỳ theo thời gian. Các định kỳ của thời gian có rất nhiều chu kỳ, nhưng chu kỳ thời gian bao trùm một số năm được quan tâm trong DW.

2.1.2.3. Ba kiểu của dữ liệu công việc:

❖ *Dữ liệu thời gian thực (Real time data):* là dữ liệu hiện thời hoặc dữ liệu đến mức thứ 2 biểu diễn trạng thái hiện tại của dữ liệu công việc và được sử dụng trong Doanh nghiệp. Nó xuất hiện tại mức chi tiết và được truy cập trong chế độ đọc/ghi.

Dữ liệu thời gian thực là dữ liệu được tạo ra, được vận dụng và sử dụng bởi các thao tác hoặc các ứng dụng sản xuất. Dữ liệu này cơ bản được lấy ra từ các file hoặc cơ sở dữ liệu trong môi trường máy tính lớn. Và được kiểm soát và quản lý bởi bộ phận hệ thống thông tin.

Dữ liệu thời gian thực không bị hạn chế trong máy tính lớn hay các ứng dụng kế thừa. Một mô hình mới của ứng dụng client/server tạo ra dữ liệu thời gian thực trong các trạm làm việc và các máy chủ. Dữ liệu thời gian thực này được phân bố thông qua hoạt động kinh doanh và hiếm khi dưới sự kiểm soát trực tiếp của bộ phận hệ thống thông tin.

Hơn nữa, dữ liệu thời gian thực có nguồn gốc bên ngoài doanh nghiệp. Nó xuất hiện khi xử lý thông tin hoạt động kinh doanh, chẳng hạn như các đơn đặt hàng hoặc các hoá đơn thanh toán, giữa các tổ chức giữa các tổ chức trao đổi dữ liệu điện tử (EDI), và các dữ liệu vào được sử dụng cơ bản cho các hoạt động của công ty nhận được.

Data	Industry	Usage	Technology	Volumes
Customer file	All	Track customer details	Legacy application, flat files, mainframe	Small – medium
Account balance	Finance	Control account activities, e. g., witharawals	Legacy application, hier – archical database, mainframe	Large
Point – of – sale data	Retail	Generate bills manage stock	Client/server, relational database, UNIX system	Very lange
Call record	Telecommu n- ications	Billing	Legacy application, hier – archical database, mainframe	`Very lange
Production record	Manufac- turing	Control production	New application, relational database, AS/400	Medium

Hình.6: Ví dụ của thời gian thực

❖ *Dữ liệu nguồn (Derived data):* Dữ liệu nguồn là dữ liệu đơn giản được tạo ra, thông qua một số xử lý, từ dữ liệu thời gian thực. Nó được sử dụng để quản lý doanh nghiệp, trong chế độ chỉ đọc, hơn là các vận hành hàng ngày của doanh nghiệp. Nó có thể đạt đến mức chi tiết hoặc mức tổng hợp. Bởi vì nó nhận từ dữ liệu thời gian thực, nó thậm chí là thời điểm trong thực tế, thể hiện quan sát của doanh nghiệp tại thời điểm đó, hoặc định kỳ trong thực tế, bảo toàn lịch sử bản ghi của doanh nghiệp qua kỳ thời gian.

Dữ liệu nguồn là tập các dữ liệu truyền thống được sử dụng để hỗ trợ quyết định. Nó được phát hiện thông qua tổ chức ngày nay, từ các cơ sở dữ liệu quan hệ trong các máy tính lớn, cho các gói bảng dữ liệu chuyên dụng trong các máy tính cá nhân, và nhiều thứ trong đó. Mặc dù quan niệm là dữ liệu nguồn có thể được cập nhật tự động, trong một số trường hợp việc xử lý được làm thủ công, với các nội dung của các báo cáo được in ra được gõ lại vào các công cụ quản lý thông tin.

❖ *Dữ liệu điều chỉnh (Reconciled data):*

Dữ liệu điều chỉnh được sinh ra bằng một xử lý thiết kế để đảm bảo tính thống nhất nội bộ của dữ liệu kết quả. Quá trình này được vận hành trong dữ liệu thời gian thực ở mức chi tiết. Hướng thứ hai của xử lý sinh ra là duy trì nó hoặc tạo ra tập lịch sử của dữ liệu. Dữ liệu điều chỉnh được xem như là loại đặc biệt của dữ liệu nguồn.

Trong các môi trường hỗ trợ quyết định truyền thống, dữ liệu điều chỉnh là hiếm khi được xác định rõ ràng. Trong nhiều trường hợp, nó không tồn tại. Trường hợp không tồn tại, nó hiếm khi được lưu trữ vật lý, chỉ là kết quả hợp lý của một số hoạt động diễn ra trong quá trình tính toán. Trong trường hợp khác, nó chỉ tồn tại trong các tập tin tạm thời. Như thế thì không công nhận là có bất kỳ kết quả kinh doanh. Trong thực tế, đối chiếu dữ liệu là yếu tố then chốt của kho dữ liệu. Là một kết quả của việc sử dụng tiếp cận phát triển nguồn ứng dụng, dữ liệu thời gian thực không phải là tự nhất quán trên toàn bộ phạm vi của doanh nghiệp. Điều này tạo ra điều chỉnh dữ liệu là cần thiết.

Vì vậy, bất cứ khi nào dữ liệu từ nhiều nguồn đã được kết hợp, phát triển đầu tiên phải phân tích cấu trúc và nội dung của các nguồn để xác định các quy tắc kết hợp. Sau đó, họ cần phải phát triển một quá trình để thực thi các quy tắc này. Thông thường, quá trình đó bao gồm các chức năng như nối và thao tác của các trường, sự thay đổi của các trường dữ liệu theo các mẫu phù hợp, và trong những tình huống cuối cùng, các loại sửa chữa lỗi.

2.1.3. Siêu dữ liệu(Meta data)

Một trong những phần quan trọng nhất của kho dữ liệu là kho dữ liệu về dữ liệu (metadata)– dữ liệu quản lý dữ liệu.

2.1.3.1.Khái niệm

Metadata là toàn bộ tất cả các mức độ của kho dữ liệu, kể cả các dạng tồn tại và các chức năng ở một chiều khác biệt của kho dữ liệu khác. Hay nói một cách khác thì Meta data là dạng dữ liệu miêu tả dữ liệu.

Trong cơ sở dữ liệu, Metadata là các dạng biểu diễn khác nhau của các đối tượng trong cơ sở dữ liệu

Trong cơ sở dữ liệu quan hệ thì Metadata là các định nghĩa của bảng, cột, view, và nhiều đối tượng khác.

Còn Trong kho dữ liệu Metadata là dạng định nghĩa của dữ liệu như bảng, cột, một báo cáo, các luật doanh nghiệp hay những quy tắc biến đổi. Metadata bao quát tất cả các phương diện của kho dữ liệu.

2.1.3.2. Mục đích

Các chuyên viên phát triển kho dữ liệu sử dụng Metadata để quản trị, điều khiển sự hình thành và duy trì sự tồn tại các kho dữ liệu nằm bên ngoài kho dữ liệu nói trên.

Metadata của người sử dụng kho dữ liệu là một phần của chính kho dữ liệu đó và có thể được dùng để điều khiển sự phân tích và truy cập kho dữ liệu đó.

Đối với người sử dụng kho dữ liệu, Metadata giống như là một tờ mục lục (card catalog) về các chủ đề có trong kho dữ liệu.

2.1.3.3. Metadata phải chứa các thông tin:

- Cấu trúc của dữ liệu
- Thuật toán sử dụng để tổng hợp dữ liệu
- Ánh xạ xác định sự tương ứng dữ liệu từ môi trường tác nghiệp sang kho dữ liệu

2.1.3.4. Tác dụng của metadata

Metadata là dữ liệu để mô tả dữ liệu. vì vậy khi dữ liệu được cung cấp cho người dùng cuối, Metadata sẽ cung cấp những thông tin cho phép người dùng hiểu rõ hơn bản chất dữ liệu mà họ đang có. Những thông tin này sẽ giúp cho người dùng có được những quyết định sử dụng đúng đắn và phù hợp về dữ liệu mà họ đang có.

Tuỳ thuộc vào từng mục đích sử dụng khác nhau, từng loại dữ liệu khác nhau mà cấu trúc và nội dung dữ liệu Metadata có thể có những sự khác biệt. Trong đó bao gồm một số loại thông tin:

- Thông tin mô tả về bản thân dữ liệu Metadata
- Thông tin về dữ liệu mà Metadata mô tả
- Thông tin về cá nhân, tổ chức có liên quan đến dữ liệu Metadata và dữ liệu

2.1.3.5. Tiêu chuẩn cho các kiểu siêu dữ liệu

Tương tự như dữ liệu công việc, metadata được phân lớp theo một số tiêu chuẩn cơ bản. Có hai tiêu chuẩn cơ bản: khi nó sử dụng trong vòng đời ứng dụng và khi nó được sử dụng tích cực hoặc bị động.

a). Mối liên hệ tới vòng đời ứng dụng:

Việc sử dụng siêu dữ liệu trong quá trình xác định và xây dựng ứng dụng doanh nghiệp và cơ sở dữ liệu liên quan của họ khác với việc sử dụng nó trong các ứng dụng và cơ sở dữ liệu trong sản xuất. Nó được phân biệt giữa:

- Siêu dữ liệu thời gian xây dựng (Build- time metadata): thiết kế để thuận lợi cho việc sử dụng, cũng như tái sử dụng cả dữ liệu và chức năng bởi những người thiết kế ứng dụng và cơ sở dữ liệu.

- Siêu dữ liệu thời gian sản xuất (Production - time metadata): Được thiết kế để thuận lợi cho việc tìm kiếm, sự hiểu biết, và sử dụng các dữ liệu cần thiết trong công việc.

b). Sử dụng chủ động hoặc thụ động: Đặc tính này mô tả ký thuật sử dụng tạo ra siêu dữ liệu thời gian sản xuất:

- Siêu dữ liệu được sử dụng để điều khiển hành động hoặc chức năng của một số ứng dụng hoặc phần khác của phần mềm có vai trò tích cực.
- Siêu dữ liệu được sử dụng trong chế độ tìm kiếm, thường là một người, để tìm một số dữ liệu công việc hoặc để hiểu một số đặc tính của dữ liệu công việc đang được sử dụng trong một chế độ thụ động.

2.1.3.6. Ba loại siêu dữ liệu

a). Siêu dữ liệu thời gian sản xuất (Build time metadata):

Nguồn gốc của siêu dữ liệu được sử dụng trong kho là quá trình mà theo đó các ứng dụng kinh doanh và các dữ liệu được mô tả và định nghĩa. Siêu dữ liệu được tạo ra và được sử dụng trong giai đoạn này là siêu dữ liệu thời gian sản xuất.

Theo định nghĩa của phạm vi kho dữ liệu, siêu dữ liệu thời gian sản xuất là ở bên ngoài phạm vi kho. Tuy nhiên, như đối với dữ liệu công việc thời gian thực, siêu dữ liệu thời gian sản xuất không thể bỏ qua bởi vì nó là nguồn gốc của các siêu dữ liệu mà không thuộc phạm vi của kho. Ngày nay, siêu dữ liệu thời gian sản xuất được tạo ra và lưu trong mô hình dữ liệu và các công cụ thiết kế ứng dụng như CASE tools. Theo yêu cầu, các ứng dụng tồn tại, siêu dữ liệu thời gian sản xuất thường tồn tại hoàn toàn chỉ trong cơ sở dũ liệu hoặc các thiết kế file của ứng dụng hoặc trong thiết kế hoặc tài liệu người dùng.

Siêu dữ liệu thời gian sản phẩm là ổn định so với các dữ liệu công việc nó mô tả. Nói chung, siêu dữ liệu thay đổi chỉ khi cấu trúc tổng thể của doanh nghiệp hoặc thực hiện của chúng trong các ứng dụng thay đổi. Siêu dữ liệu đã được định nghĩa trong việc thiết kế của một ứng dụng sẽ không thay đổi từ việc phiên bản đầu tiên của ứng dụng đó cho đến khi một phiên bản cuối cùng, và vẫn tồn tại đến khi phiên bản được nâng cấp.

b). Siêu dữ liệu điều khiển:

Siêu dữ liệu điều khiển được sử dụng tích cực bởi các thành phần kho như một cơ chế để quản lý và kiểm soát hoạt động của các thành phần riêng của nó. Do đó, nó là một phần của siêu dữ liệu thời gian sản xuất. Nó có hai nguồn.

- Thông tin cấu trúc vật lý chi tiết có nguồn gốc từ việc xây dựng siêu dữ liệu thời gian xây dựng. Bởi vì nó được thiết kế để sử dụng cho các thành phần kho, siêu dữ liệu này là không phù hợp cho người dùng cuối.

- Nguồn thứ hai là các thành phần kho của nó. Như siêu dữ liệu mô tả những hoạt động đang xảy ra mà siêu dữ liệu là đối tượng. Siêu dữ lieuj là quan trọng với cả người dùng cuối và người quản trị trong kho dữ liệu. Có hai kiểu:

+ Siêu dữ liệu tiền tệ (currency metadata): siêu dữ liệu tiền tệ mô tả các thông tin thực tế về tiền tệ hoặc tính thời điểm của các dữ liệu công việc. Ví dụ như thời gian cập nhật cuối cùng của một bảng trong một cơ sở dữ liệu, hoặc lần đầu tiên một ứng dụng đặc biệt chạy trên bất cứ ngày nào. Thông tin này có thể được cung cấp chỉ bởi công cụ hay ứng dụng cung cấp cho dữ liệu công việc hoặc chạy một ứng dụng.

+ Siêu dữ liệu tận dụng (Utilization metadata): Siêu dữ liệu tận dụng là liên quan tới an toàn và tính năng cho phép sử dụng để kiểm soát truy cập vào kho. Ngoài ra, siêu dữ này liệu cung cấp điều kiện để truy vết dữ liệu hoặc các chức năng được sử dụng trong kho, và vì thế cho việc đánh giá tính hữu dụng của nó hoặc giá trị cho người dùng cuối.

c). Siêu dữ liệu sử dụng (Usage metadata):

Siêu dữ liệu sử dụng là siêu dữ liệu quan trọng nhất cho người sử dụng dữ liệu công việc, đặc biệt là trong môi trường thông tin. Đây là nơi người dùng cuối đạt được lợi ích kinh doanh và hệ thống thông tin nhân sự đạt được những cải thiện về năng suất.

Siêu dữ liệu sử dụng bắt nguồn từ siêu dữ liệu thời gian sản xuất và tương tự trong nội dung. Sự khác biệt nằm trong cách siêu dữ liệu tại mức này cần được cấu trúc theo khả năng của các người dùng để tìm kiếm hiệu quả và

khai thác nó. Cấu trúc yêu cầu bởi người dùng cuối và tín hiệu khác cần thiết từ những người thiết kế ứng dụng và cơ sở dữ liệu.

Siêu dữ liệu sử dụng mô tả bởi các khía cạnh sau của dữ liệu hoặc ứng dụng:

- Điều kiện của doanh nghiệp: Loại siêu dữ liệu này mô tả hoạt động của doanh nghiệp trong hình thức hoặc cách cấu trúc. Đặc tính này cho phép các người dùng liên kết các phần tử dữ liệu hoặc chức năng của ứng dụng cho mục đích của họ trong kinh doanh.

Khi điều kiện của dữ liệu và ứng dụng được biết, người dùng có thể liên kết chúng lại trong kinh doanh thực, và hệ thống thông tin cá nhân và kết nối người dùng có thể kết nối như nhau.

- Chủ sở hữu và cương vị quản lý:

Chủ sở hữu buộc mối quan hệ giữa dữ liệu hoặc ứng dụng và tổ chức, và chỉ rõ người có trách nhiệm với khía cạnh riêng biệt và duy trì chúng. Chủ sở hữu có thể được phân chia, ví dụ một người có trách nhiệm về độ chính xác của file dữ liệu, trong khi người khác nhận trách nhiệm về tính đa dạng thời gian. Chủ sở hữu dữ liệu có thể phân chia để thực hiện các quyết định công việc. Trong trường hợp này, chức năng phụ trợ của người quản lý dữ liệu được định nghĩa là chỉ ra trách nhiệm thường xuyên với dữ liệu.

Trong môi trường kho, chủ sở hữu dữ liệu là quan trọng hơn chủ sở hữu của chức năng ứng dụng, nhưng chủ sở hữu dữ liệu trái ngược là khó xác định sự phân chia. Khi đó nó được định nghĩa, và lưu vết, người dùng cuối có thể lấy trách nhiệm cho chất lượng của dữ liệu.

- Cấu trúc dữ liệu

Cấu trúc của siêu dữ liệu mô tả kỹ thuật sắp xếp của dữ liệu. Có một số kiểu khác nhau của cấu trúc cần cho việc lưu trữ. Ví dụ, một phần tử dữ liệu có thể được mô tả dưới dạng nới nó lưu trữ vật lý, cái mà cấu trúc dữ liệu được sử dụng, khi nó là ký tự hoặc số, kích thước của nó là bao nhiêu và ứng dụng nào quản lý nó.

- Các khía cạnh ứng dụng

Siêu dữ liệu phải bao gồm mô tả các chức năng của ứng dụng, ngôn ngữ mà nó được viết, dữ liệu mà nó sử dụng và kết quả, và các điều kiện tiên quyết nào, và nếu cần là các yêu cầu khi sử dụng nó. Trong ngữ cảnh này, người dùng cuối có thể sử dụng trực tiếp các ứng dụng hoặc họ chịu trách nhiệm về sự thực hiện của các dữ liệu trong kho.

2.1.4. Dữ liệu vượt quá phạm vi của kho dữ liệu (Data beyond the scope of the Data Warehouse)

*2.1.4.1.*Dữ liệu giống như một sản phẩm(Data as a product)

Một số sưu tầm nhóm, thao tác, hoặc thông tin sản xuất dưới dạng điện tử đang tăng lên nhanh chóng về tầm quan trọng và giá trị nhưng không thuộc phạm vi của kho dữ liệu như đã được định nghĩa, và thực sự nó nằm bên ngoài phạm vi của hệ thống xử lý dữ liệu truyền thống. Dữ liệu là một sản phẩm được tạo ra và được lưu trữ, nó không phải là một phương tiện chạy hoặc quản lý một doanh nghiệp. Nó là một sản phẩm của một hoạt động doanh nghiệp, có thể được mua và bán, và phải được quản lý và kiểm soát như bất kỳ một sản phẩm vật lý. Ví dụ, giá trị của một quyển sách là dữ liệu thông tin của nó. Như một sản phẩm, nó được sản xuất trên giấy. Tuy nhiên, phần lớn các tiến trình sản xuất của nó tồn tại dạng nguyên bản và dữ liệu ảnh nằm trong một máy tính.

Dữ liệu là một sản phẩm nằm ngoài phạm vi của dữ liệu như đã được định nghĩa. Tuy nhiên, các công cụ và kỹ thuật được sử dụng để xây dựng và quản lý một kho dữ liệu cũng có thể được sử dụng trong một cách tương tự để xây dựng và quản lý dữ liệu như là một sản phẩm.

2.1.4.2. Dữ liệu công việc cá nhân và siêu dữ liệu

Dữ liệu cá nhân được định nghĩa đơn giản là dữ liệu nằm dưới sự kiểm soát của một cá nhân duy nhất. Đó là tạo ra, sử dụng, và xóa bằng theo yêu cầu của quá trình kinh doanh mà người đó chịu trách nhiệm. Những dữ liệu này đã luôn luôn tồn tại, từ nhân viên bán hàng viết vội ghi chú về một trật tự các điều hành có chứa tên, địa chỉ, và ngày sinh của địa chỉ liên lạc của khách hàng; từ viết tay của dự báo doanh số bán hàng năm bên cạnh để làm danh

mục các nhiệm vụ vào ngày mai,... Khi sử dụng máy tính lớn, rất nhiều dữ liệu được lưu trữ trong bảng tính, quản lý thông tin cá nhân, vv

Trước năm 1990, dữ liệu cá nhân có tầm quan trọng hạn chế trong hệ thống thông tin. Nó tồn tại trong các hệ thống thông tin của các cửa hàng. Tuy nhiên, khối lượng của nó khá hạn chế, và tương đối cô lập với dòng chính của các dữ liệu công việc. Từ đó đến nay đã có sự thay đổi đáng kể cả hai yếu tố này. Người sử dụng cuối hiện nay lưu trữ dữ liệu trên máy tính cá nhân với hàng trăm GB. Những cải thiện trong mạng LAN và client/Server, mạng Internet, công nghệ đã dẫn đến sự gia tăng lớn sự trao đổi dữ liệu giữa các máy tính và các công ty trong môi trường hệ thống thông tin. Dữ liệu cá nhân được liên kết trong mạng lưới, có thể dễ dàng chia sẻ nó.

2.1.5. Dữ liệu bên trong và bên ngoài (Internal and external data)

Trước đây, phần lớn các dữ liệu có ích cho một tổ chức đều có nguồn gốc trong tổ chức đó. Thậm chí khi dữ liệu nằm bên ngoài, số lượng của các nguồn đã đủ nhỏ, khối lượng của dữ liệu đã đủ ít mà ảnh hưởng của dữ liệu bên ngoài vào kiến trúc tổng thể là tương đối quan trọng. Điều này là không còn giá trị. Ví dụ, nó được báo cáo rằng hiện nay có hơn 10. 000 người tiêu dùng các nguồn dữ liệu trực tuyến ở Hoa Kỳ, bao gồm 1.500 biến về 150 tỉ người. Sự tăng trưởng bất thường của Internet trong những năm qua cũng đã gây ra một sự tăng trưởng theo hàm mũ trong các khối dữ liệu điện tử vào, ra tất cả các tổ chức.

Trong phạm vi qui định của kho dữ liệu, sự tương tác bên trong hay bên ngoài đều cần phải được xem xét. Trong đó gồm có:

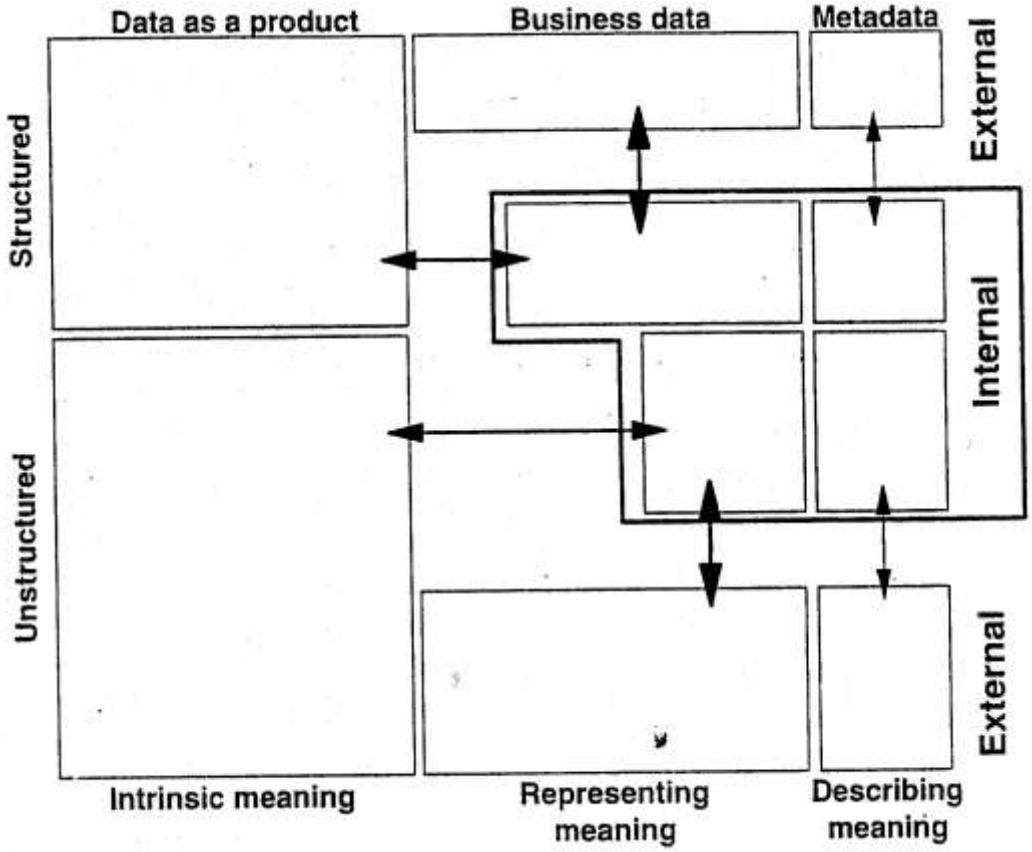
- Dữ liệu công việc có cấu trúc: dễ dàng có thể tổng hợp dữ liệu nội bộ hiện tại, dữ liệu có cấu trúc bên ngoài phải được xử lý thủ công. Dữ liệu phải trải qua một quá trình hợp nhất với các dữ liệu trong để bảo đảm tính thống nhất của nó với dữ liệu nội bộ hiện tại. Điều này ngụ ý rằng các siêu dữ liệu liên quan bên ngoài cũng phải được tạo sẵn cho việc thu nhận vào.

Với dữ liệu công việc ra bên ngoài cấu trúc, các siêu dữ liệu liên quan cũng phải được làm sẵn có. Trong trường hợp này, yêu cầu về trách nhiệm pháp lý có thể phát sinh từ việc cung cấp dữ liệu không chính xác.

- Dữ liệu công việc không có cấu trúc: tương tự áp dụng cho dữ liệu công việc phi cấu trúc. Tuy nhiên, vì có khó khăn hơn để dữ liệu phi cấu trúc tự động nhúng trong quá trình ra quyết định.

- Dữ liệu là một sản phẩm: Dữ liệu bên ngoài như là một sản phẩm vào kho dữ liệu như dữ liệu công việc.

- Siêu dữ liệu: Siêu dữ liệu ít khi loại bỏ hoặc đưa vào tổ chức. Thay vào đó, nó đi kèm với dữ liệu công việc trên ranh giới của tổ chức. Việc này là cần thiết để cho phép các dữ liệu công việc được hiểu và hợp nhất theo yêu cầu.



Hình 7: Relationships between internal and external data

2.1.6. Kết luận:

Rất khó xác định phạm vi của kho dữ liệu. Đặc biệt đúng cho sự phổ biến của các đối tượng và nỗ lực của các nhà cung cấp để mang lại lợi ích bằng cách liên tục mở rộng phạm vi để bao gồm càng nhiều các dòng sản

phẩm của họ càng tốt. Phần này đã trình bày về xác định phạm vi của kho dữ liệu về các loại dữ liệu mà nó hỗ trợ. Tuy nhiên dữ liệu được chia ra, trên cơ sở sử dụng của nó, trong dữ liệu doanh nghiệp và siêu dữ liệu được bao gồm trong các kho và dữ liệu được coi như một sản phẩm.

2.2. Khái niệm kiến trúc dữ liệu(Conceptual data architecture):

Một trong những bước đầu tiên trong việc thiết kế bất kỳ hệ thống xử lý dữ liệu là thiết lập một kiến trúc tổng thể cho hệ thống và để đạt được sự chấp nhận rộng rãi các kiến trúc đó. Việc thiết kế của một kho dữ liệu cũng vậy.

Theo truyền thống, việc thiết kế các hệ thống hoạt động bắt đầu với kiến trúc ứng dụng. Kết quả từ nơi các ứng dụng hoạt động với các chức năng mà người dùng yêu cầu. Cách tiếp cận này được hỗ trợ bởi các phạm vi dữ liệu tương đối hẹp như địa chỉ các ứng dụng. Tuy nhiên, do tầm quan trọng trọng của sự gắn kết dữ liệu trong kho dữ liệu, cả dữ liệu công việc và siêu dữ liệu phải là điểm khởi đầu trong kiến trúc của kho**.**

Ở đây xem xét ba kiến trúc dữ liệu cho dữ liệu công việc. Mỗi kiến trúc đều có lợi thế và bất lợi riêng của nó. Có các tiêu chí quan trọng để đánh giá chúng như: sự linh hoạt mà dữ liệu có thể được truy cập và sử dụng cho người dùng cuối; quản lý chất lượng dữ liệu cho hệ thống thông tin cá nhân và một số yếu tố khác trong các tình huống cụ thể. Tuy nhiên, không có kiến trúc duy nhất đó là phù hợp nhất với mọi tình huống, một tiếp cận riêng có thể sẽ thành công trong phần lớn các trường hợp.

Đối với siêu dữ liệu thì đơn giản hơn. Một kiến trúc dữ liệu duy nhất hỗ trợ cả ba lựa chọn thay thế của kiến trúc dữ liệu công việc.

2.2.1. Các kiến trúc dữ liệu công việc (Business data architectures)

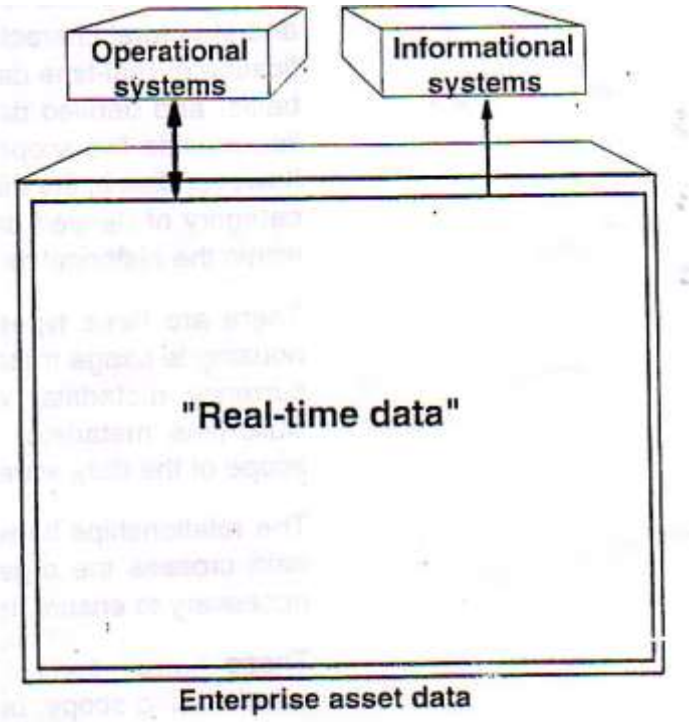
Ba mô hình kiến trúc được mô tả trong các phần sau đây có một điểm chung là đều dựa trên thực tế kinh nghiệm.

Trong ba kiến trúc được đặt tên theo số lớp của dữ liệu bao quanh chúng. Các lớp dữ liệu này là khái niệm hóa hơn là vật lý. Vì vậy, trong bất kỳ thực hiện nào, một lớp có thể được xác định bởi các loại dữ liệu của nó, chứ không phải bởi vị trí vật lý của nó.

2.2.2. Kiến trúc đơn lớp dữ liệu (The single-layer data architecture)

Nguyên tắc cơ bản quan trọng trong kiến trúc đơn lớp là bất kỳ yếu tố dữ liệu nào chỉ được lưu trữ một lần và một lần duy nhất. Trong khi mục tiêu này có khó khăn hoặc không thể đạt được, cấu trúc của kiến trúc này cho phép có thể đạt được mục tiêu này. Trong một kiến trúc đơn lớp, không phân biệt sự tạo ra giữa bất kỳ các loại dữ liệu được mô tả trước, tất cả dữ liệu được coi như nhau.

Mặc dù không có sự mô tả chính xác chặt chẽ, kiến trúc này chủ yếu đề cập đến tất cả dữ liệu thực sự có thể tồn tại trong thời gian thực. Dữ liệu xuất phát có thể tồn tại trong phạm vi kiến trúc này, nhưng nó không được xem xét bất kỳ khác biệt từ các dữ liệu thời gian thực từ nguồn gốc của nó.



H

Hình 8 :The single layer data architecture

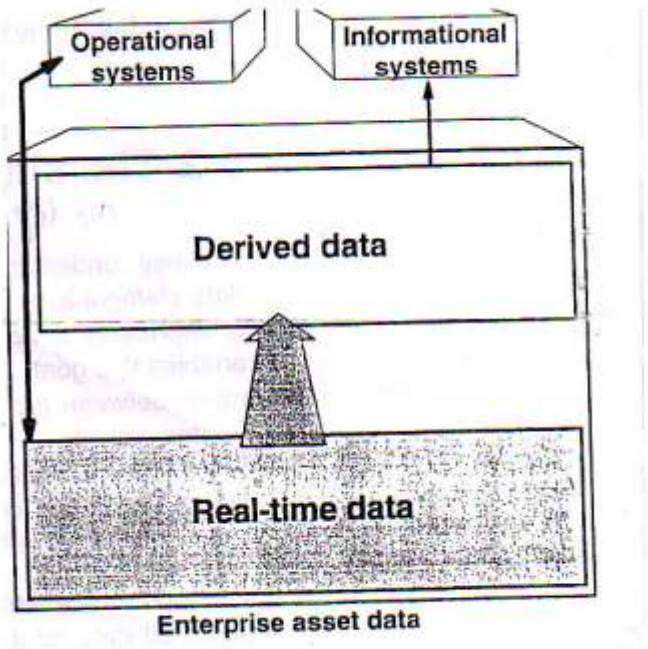
Sức mạnh của kiến trúc đơn xuất phát từ mục tiêu lưu trữ mỗi phần tử dữ liệu. Bởi vì nó tối thiểu các yêu cầu lưu trữ dữ liệu và cho ngăn chặn vấn đề sao chép dữ liệu trong đồng bộ hóa. Điểm yếu của tiếp cận này là sự bất đồng xuất hiện giữa sự vận hành và các ứng dụng thông tin, dẫn đến việc dữ

liệu không sẵn sàng cho các ứng dụng hoặc thơi gian phản hồi chậm cho các thao tác ứng dụng. Điểm yếu nữa là nó không cung cấp sự trợ giúp trong việc làm thế nào dữ liệu được phân loại có thể thực hiện được hoặc làm thé nào người dùng ở các vị trí địa lý khác nhau có thể truy cập được dữ liệu của công ty.

2.2.3. Kiến trúc hai lớp dữ liệu (The two-layer data architecture)

Đây là một cải tiến cho kiến trúc lớp đơn với hai cách sử dụng dữ liệu khác nhau - hoạt động và thông tin, và phân để chia dữ liệu thành hai lớp (trong hình vẽ). Lớp thấp hơn, được sử dụng bởi các ứng dụng vận hành ở chế độ đọc/ghi, đây là dữ liệu thời gian thực. Lớp trên, sử dụng bởi các ứng dụng thông tin, là dữ liệu nguồn. Dữ liệu nguồn có thể đơn giản như một bản sao trực tiếp của các dữ liệu thời gian thực, hoặc nó có thể được bắt nguồn từ dữ liệu thời gian thực bằng một số tính toán.

Cách tiếp cận này ngay lập tức giải quyết một trong những vấn đề chính của kiến trúc lớp đơn - giữa hai loại dữ liệu sử dụng khi vận hành trên một nguồn dữ liệu đơn. Lợi ích thứ hai là những người dùng cuối có địa chỉ rõ ràng cần thiết cho dữ liệu khác nhau được lưu trữ như dữ liệu thời gian thực.



Hình 9 :The two layer data architecture

Tuy nhiên một trong những vấn đề kiến trúc này là mức độ cao của sự nhân đôi dữ liệu, trong lớp dữ liệu nguồn. Việc nhân đôi này dẫn đến sự bùng nổ trong lưu trữ dữ liệu, nhưng quan trọng là vấn đề quản lý dữ liệu và các vấn đề quản trị.

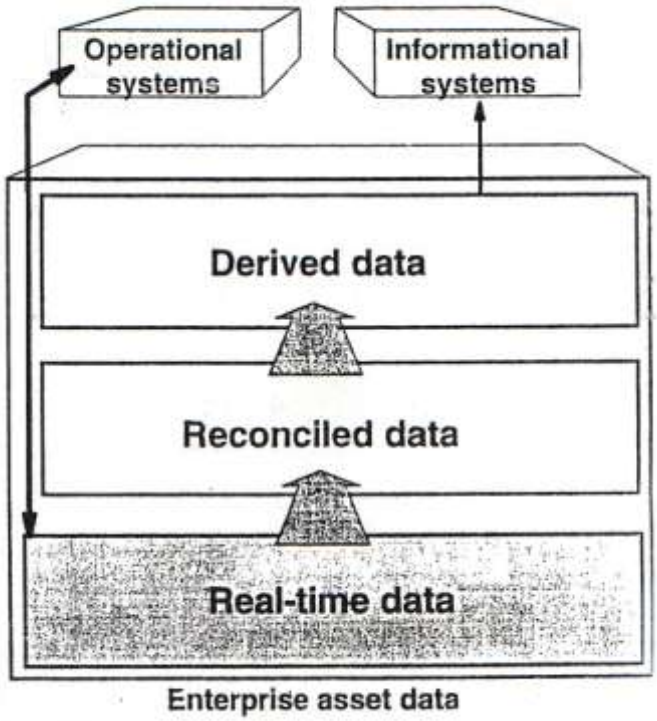
2.2.4. Kiến trúc ba lớp dữ liệu (The three-layer data architecture)

Kiến trúc ba tầng là sự chuyển đổi của dữ liệu thời gian thực và dữ liệu nguồn thêm một bước so với kiến trúc hai tầng. Nó bao gồm:

1. Điều chỉnh dữ liệu từ các tập hợp dữ liệu đa dạng trong lớp thời gian thực.

2. Nguồn các dữ liệu cần thiết cho người sử dụng từ các dữ liệu được điều chỉnh.

Điều này dẫn đến các kiến trúc được mô tả trong hình

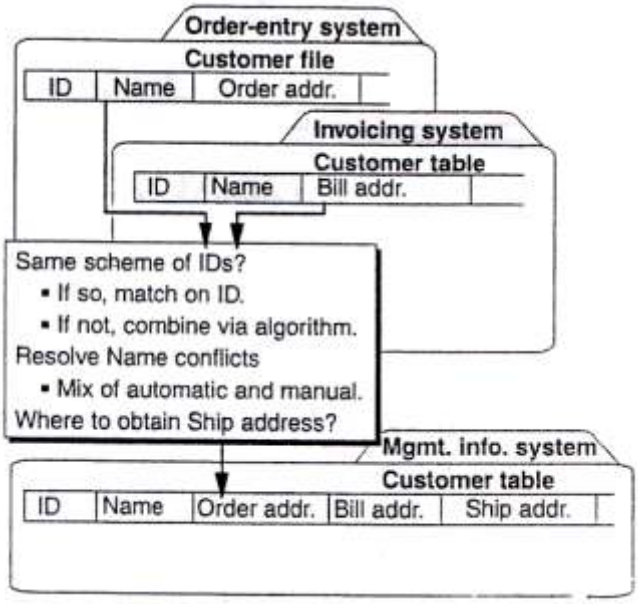


Hình 10: **The three-layer data architecture**

Trong phương pháp này, lớp thấp nhất là dữ liệu thời gian thực, lớp trên cùng là dữ liệu nguồn, và các lớp ở giữa là dữ liệu điều chỉnh. Sự điều chỉnh dữ liệu giữa các tập dữ liệu khác nhau trong các yêu cầu dữ liệu thời gian thực giữa các bộ khác nhau của dữ liệu trong thời gian thực yêu cầu sự

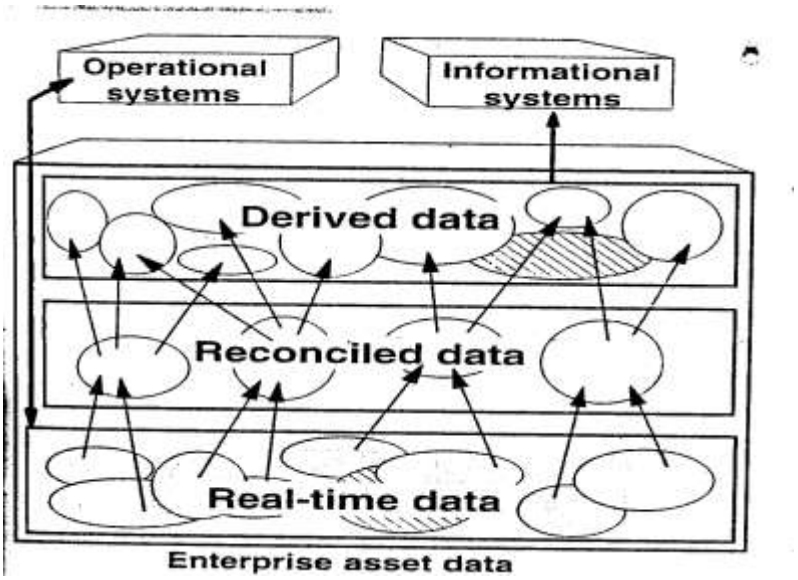
hiểu biết về cách các tập hợp dữ liệu liên quan đến nhau, và vai trò của chúng trong công việc. Trong thực tế, sự hiểu biết này được xác định thông qua quá trình mô hình hóa dữ liệu. Mối quan hệ giữa các lớp dữ liệu điều chỉnh và mô hình dữ liệu doanh nghiệp là quan trọng để nắm được các công việc của kiến trúc ba lớp.

Chúng ta có thể hiểu khái niệm bằng cách xem xét làm thế nào người ta có thể hợp lý hoá các dữ liệu từ bất kỳ hai ứng dụng hiện có và những kết quả sẽ được. Ví dụ về sự điều chỉnh, giả định rằng một ứng dụng quản lý đơn hàng quản lý một cơ sở dữ liệu bao gồm một tập tin khách hàng, tập tin một sản phẩm, và bảng một đơn đặt hàng và bảng một hoá đơn. Một ứng dụng quản lý một cơ sở dữ liệu hoá đơn có chứa một bảng khách hàng và bảng một hoá đơn. Khi dữ liệu từ hai hệ thống được yêu cầu trong lĩnh vực quản lý thông tin, các phần của dữ liệu này phải được tổng hợp và hợp lý hóa. Các tập tin của khách hàng từ hệ thống yêu cầu và bảng khách hàng từ hệ thống lập hoá đơn phải được kết hợp để tạo thành một bàng khách hàng duy nhất trong kho. Vì vậy, một thực thể khách hàng tổng quát hơn phải được xác định, đáp ứng nhu cầu của cả hai lĩnh vực kinh doanh.



Hình 11:An example of reconciliation

Hơn nữa, trong môi trường thông tin quản lý, dữ liệu từ các vùng ứng dụng này phải được liên kết với các dữ liệu khác theo dự tính ban đầu trong các ứng dụng vận hành. Ví dụ, có thể cần phải phân tích làm thế nào hoá đơn liên quan đến các đơn đặt hàng của khách hàng ban đầu để tìm thấy những gì tỷ lệ phần trăm đơn đặt hàng trong một chuyến.



Hình 12: reconciliation and derivation in the ther layers

Chương 3.

GIỚI THIỆU KIẾN TRÚC LOGIC KHO DỮ LIỆU

3.1. Dữ liệu công việc trong kho dữ liệu (Business data in the data warehouse)

3.1.1. Các hệ thống vận hành (Operational systems)

Các hệ thống vận hành là các ứng dụng được sử dụng để thực hiện các các công việc, và các dữ liệu mà họ sử dụng, trong các tập tin và cơ sở dữ liệu là các dữ liệu thời gian thực. Ngày nay, các ứng dụng như vậy tồn tại với nhiều định dạng và vị trí, chúng ở dạng hỗn tạp và được phân bố theo kiểu nào đó. Các ứng dụng mới được xây dựng được thực hiện trong các môi trường client/Server.

Các hệ thống hoạt động thường được kế thừa, nhưng chúng khác nhau một khía cạnh quan trọng. Các hệ thống kế thừa thường gồm các chức năng báo cáo, được sử dụng để quản lý công việc. Đây chỉ là một phần nhỏ của ứng dụng được kế thừa, được phân biệt với các chức năng vận hành. Vị trí thích hợp của nó là lớp nguồn.

Vì các hệ thống vận hành tương tác với nhau, thông qua dữ liệu và sửa đổi nó khi cần thiết, nó luôn luôn cần thiết để xác định chính xác và sớm nhất một cách có thể nguồn gốc chính xác của bất kỳ mục dữ liệu cụ thể trong kho. Mô hình dữ liệu đặc biệt, sự phân tích các dữ liệu tồn tại trong bối cảnh của mô hình dữ liệu doanh nghiệp có một vai trò quan trọng ở đây.

3.1.2. Kho dữ liệu công việc (The business data warehouse)

Kho dữ liệu công việc (BDW) là sự thực hiện vật lý của lớp dữ liệu điều chỉnh. Các đặc tính của lớp dữ liệu điều chỉnh được mô tả gồm:

Chi tiết (Detailed)

Lịch sử (historical)

Phù hợp (consistent)

Mô hình hóa (modeled)

Chuẩn hóa (normalized)

Một BDW được thực hiện trong môi trường quan hệ, là môi trường tốt nhất để mô hình hóa và chuẩn hóa tự nhiên. Trong lý thuyết này, BDW có thể được phân chia, các yêu cầu xử lý điều chỉnh cho một lượng lớn dữ liệu được nối và được quan tâm, tiến trình đó thích hợp với thực hiện không phân chia.

Các khía cạnh tổ chức cũng điều khiển BDW hướng tới thực hiện tập trung. Bởi vì BDW được dự kiến là một điểm điều khiển, nơi mà chất lượng và tính an toàn của dữ liệu được đảm bảo trước sự tạo ra khả năng mở rộng các thành viên người dùng cuối của nó.

Tính an toàn của BDW là khía cạnh quan trọng, vì nó bao gồm tất cả các mẫu dữ liệu được tích hợp. An toàn vật lý cũng đảm bảo một tiếp cận để lưu trữ tập trung sở hữu của công ty.

Việc đưa ra kích thước lớn của BDW cũng là kết quả của lịch sử tự nhiên của nó – chỉ các thành phần của nó, có thể trực tuyến vật lý tại bất kỳ thời điểm nào.

BDW là rất thông thường, được sử dụng trực tiếp bởi người dùng. Hơn nữa, nó là nguồn cho tất cả các dữ liệu trong các kho thông tin công việc. Do đó, dự thực hiện đó để cho BDW tập trung xung quanh một lượng lớn không trực tuyến hoặc tiến trình xử lý theo khối của sự bố trí của nó từ hệ thống vận hành và sự trích rút dữ liệu từ dòng sử dụng.

3.1.3. Các kho thông tin công việc (Business information warehouses -BIW)

Một kho thông tin công việc là tên gọi thông thường cho bất kỳ hệ thống sử dụng báo cáo, phân tích hoặc dự đoán công việc. Nó bao gồm báo cáo thông tin quản lý, hỗ trợ quyết định, và các hệ thống thông tin thực hiện tương tự như các hệ thống phân tích tiếp thị, các ứng dụng khai phá dữ liệu,..

Môi trường này được phân chia ở mức cao, có thể thấy trong mô hình client/server và các thực hiện dựa trên workstation. Trong khi mô hình này có thể tiếp tục được phân chia ở mức cao, nó kém đồng nhất hơn lớp dữ liệu thời gian thực. Phần lớn các BIW tồn tại tron cấu trúc quan hệ dựa trên hàng và cột. Môi trường quan hệ bao gồm cơ sở dữ liệu truyền thống như bảng tính và các công cụ phân tích đa chiều.

Các BIW bao gồm dữ liệu nguồn, được định nghĩa để hỗ trợ cho các yêu cầu doanh nghiệp và người dùng cuối. Chúng có thể bao gồm dữ liệu chi tiết hoặc dữ liệu tổng hợp ở mức cao, dữ liệu dự đoán qua lịch sử thời gian, hoặc ngắn hạn. Cấu trúc của các BIW là phù hợp cho các thực hiện truy vấn trực tuyến, thậm chí không dự tính trước hoặc định nghĩa trước.

Có hai kiểu của BIW là staging BIWs là tác giả ban đầu của BIW và user BIWs (không phải là tác giả). Staging BIWs yêu cầu quản lý đặc biệt để chắc chắn thỏa mãn tính ổn định và toàn vẹn của dữ liệu lưu trữ trong đó.

3.2.Các vấn đề khác của dữ liệu công việc (Business data - other considerations)

3.2.1 Các nhu cầu dữ liệu đặc biệt (Special data needs)

- Các sửa chữa (Corrections): Khi người dùng cuối cùng phát hiện ra sai sót của thực tế trong kho thông tin công việc, họ thường sử dụng dữ liệu riêng của họ và mong muốn các sửa chữa này có kết quả trở lại vào dữ liệu nguồn để đảm bảo một cái nhìn nhất quán của công việc. Các sửa chữa là cần thiết trong các hệ thống vận hành, kho dữ liệu công việc, và các kho dữ liệu thông tin công việc.

- Các điều chỉnh (Adjustments): Tương tự như hiệu lực để sửa chữa, các điều chỉnh phản ánh một sự thay đổi trong phân loại của dữ liệu trong công việc do hoàn cảnh thay đổi. Các dữ liệu được làm chính xác ban đầu, nhưng sau đó người dùng cần phải sử dụng hoặc phân tích nó một cách khác nhau. Điều này dẫn đến sự cần thiết phải thay đổi dữ liệu trong kho dữ liệu công việc và có thể đôi khi cũng ảnh hưởng đến các hệ thống hoạt động.

- Tái sử dụng dữ liệu (Data reuse): Dữ liệu nguồn ban đầu có thể trở thành dữ liệu vào cho quá trình vận hành. Ví dụ, trong phân tích các mẫu khách mua hàng, người dùng cuối cùng (như các quản lý bán hàng) có thể yêu cầu tổng hợp các phân lớp khách hàng cơ bản. Các loại này mới được tạo ra như là một phần của quá trình nguồn, và được lưu trữ trong các kho thông tin cồn việc. Các loại dữ liệu như được sử dụng làm cơ sở cho một hệ thống

nhiệm vụ mới cho lực lượng bán hàng. Đây là quá trình vận hành yêu cầu dữ liệu từ các kho thông tin công việc.

- Dữ liệu dự đoán (Predictive data): Dữ liệu được sử dụng để dự báo xu hướng và thiết lập các trạng thái vận hành trong tương lai bắt đầu từ một kho thông tin công việc và sử dụng để thiết lập dữ liệu trong các hệ thống vận hành. Ví dụ, một phân tích về giá vật liệu thô trong lớp dữ liệu nguồn cho phép tính toán ra giá bán mới, nó có thể là đầu vào cho các hệ thống vận hành.

3.2.2. Nhân tố cơ bản cho luồng dữ liệu duy nhất (The rationate for uniditrecional data flow)

Nhân tố cơ bản cho một luồng dữ liệu duy nhất dựa trên định nghĩa cơ bản của các loại dữ liệu và bắt nguồn từ nguyên tắc quản lý cơ sở dữ liệu. Nó được công nhận rộng rãi rằng dữ liệu phải được tạo ra và duy trì trong môi trường kiểm soát và quản lý cẩn thận, để nó có thể được xác minh và xác nhận trong dữ liệu vào thông qua một tập được định nghĩa thống nhất về thủ tục kiểm tra đầu vào. Hoạt động hệ thống phải đáp ứng điều kiện đó.

3.2.3. Hỗ trợ "đối chiếu" các luồng dữ liệu (Supporting " reverse " data flows):

Giải pháp cho mỗi sự cần thiết dựa trên sự được thừa nhận, trong mỗi trường hợp, dữ liệu mới đang được tạo ra. Thực tế dữ liệu này mới được gần với dữ liệu hiện có. Nguyên tắc là dữ liệu mới được tạo ra và duy trì trong lớp dữ liệu thời gian thực bằng các hệ thống vận hành. Và trách nhiệm quan trọng của các hệ thống vận hành là để xác minh và xác nhận các dữ liệu mà họ nhận được từ bất cứ nguồn nào.

3. 2. 4. Dữ liệu cá nhân (Personal data)

Dữ liệu cá nhân phần lớn rơi bên ngoài phạm vi của kho dữ liệu. Đây là một kết quả của mức độ kiểm soát và quản lý có thể được thực hiện trên các dữ liệu đó trong việc so sánh với các dữ liệu chung. Tuy nhiên, khi dữ liệu cá nhân không thuộc phạm vi của các kho dữ liệu, vị trí của nó trong kiến trúc phải được xác định.

Kiến trúc ba lớp cho phép dữ liệu cá nhân tồn tại trong các lớp dữ liệu thời gian thực và cả lớp dữ liệu nguồn. Ở cấp độ khái niệm, không có sự phân biệt giữa các dữ liệu cá nhân và dữ liệu chung trong hai lớp này. dữ liệu cá nhân có thể được tập trung hay phân tán. Nó có thể được tương thích với các dữ liệu chung hoặc bắt nguồn từ nó. Dữ liệu cá nhân không tồn tại trong các lớp dữ liệu tương thích, bởi vì lớp này là đại diện duy nhất hợp lý của mô hình dữ liệu doanh nghiệp (EDM), và do đó là đối lập với dữ liệu cá nhân.

Ở cấp độ logic, sự khác biệt giữa dữ liệu chung và dữ liệu cá nhân là sự cần thiết trong lớp dữ liệu nguồn.

3.3. Dữ liệu bên ngoài.

3.3.1. Thông tin quản lý bên ngoài(Exteral management information):

Phân tích và nắm được sự thực hiện của một công ty yêu cầu truy cập đến các dữ liệu vận hành tổng hợp của công ty đó một cách có cấu trúc. Tuy nhiên, để hiểu triển vọng của công ty trên thị trường và để lên kế hoạch thành công cho tương lai, việc truy cập dữ liệu từ thị trường nói chung là một yêu cầu mạnh mẽ. Kết quả là, các nhà hoạch định chiến lược và các nhà quản lý hành thường cần một số lượng đáng kể dữ liệu bên ngoài. Trước đây, dữ liệu ngoài này được thu thập không theo mẫu và chưa tự động. Vì vậy, hiệu quả bị bỏ qua bởi các hệ thống thông tin.

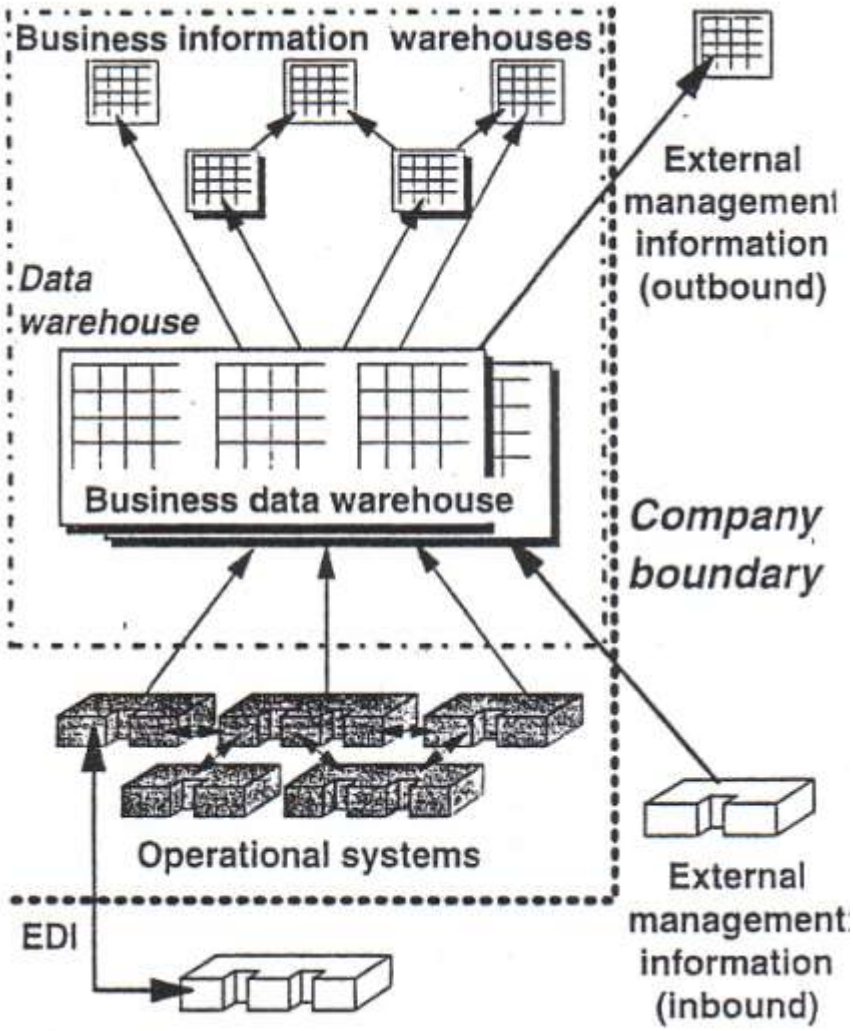
Với sự quản lý sử dụng máy tính và mạng Internet ngày càng tăng của các giám đốc điều hành, và sự có sẵn rộng rãi các dữ liệu ngoài, dữ liệu bên ngoài đã trở thành một xem xét quan trọng trong kho. Với sự phổ biến của mạng Internet ngày nay đang gây ra một sự tăng trưởng bùng nổ trong số lượng và loại thông tin.

Hơn nữa, sự liên quan tài chính trong các quyết định trong các tổ chức có ý nghĩa quan trọng. Khi quyết định được dựa trên thông tin ngoài, dữ liệu ngoài có thể được đưa vào trong công việc để được quản lý và điều khiển một cách cẩn thận như dữ liệu trong.

3.3.2. Trao đổi dữ liệu điện tử (Electronic data interchange - EDI):

Sự gia tăng chuyển giao dữ liệu giữa các phương tiện thông thường khác là trao đổi dữ liệu điện tử (EDI). EDI chủ yếu là một quá trình hoạt động và là phương tiện để các ứng dụng vận hành trong hai công ty trao đổi thông tin. Các loại dữ liệu có liên quan là dữ liệu thời gian thực. Hình 7.6

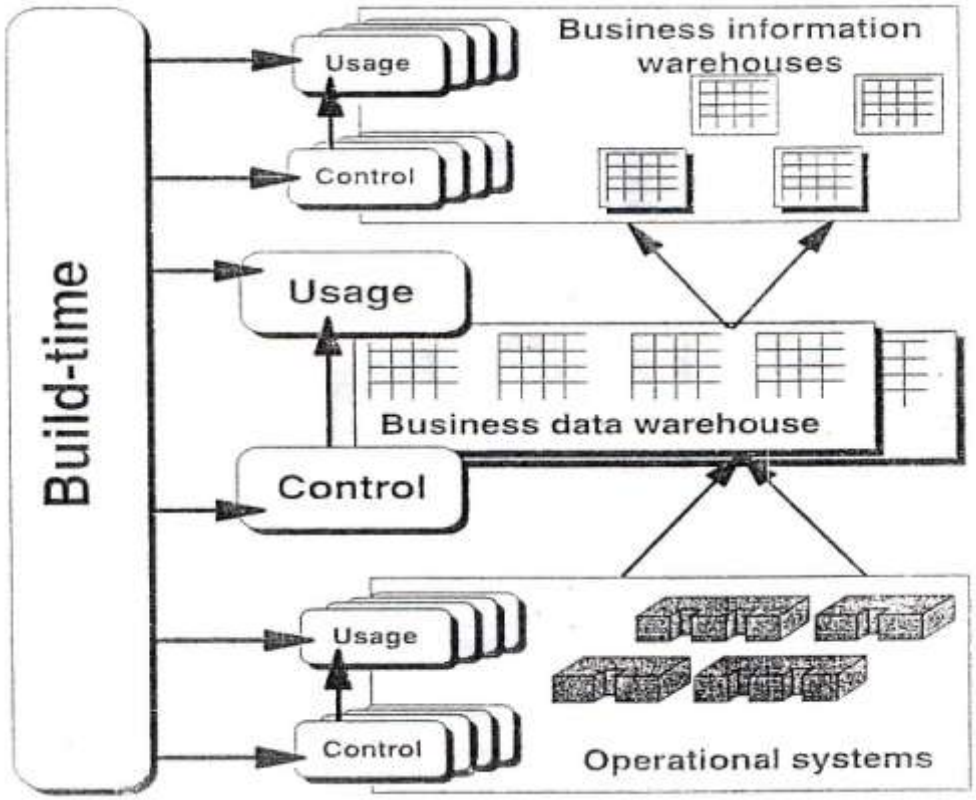
Như với bất kỳ dữ liệu vào khác trong các ứng dụng cận hành, trao đổi dữ liệu dữ liệu điện tử là đối tượng được thẩm tra, vàc các kiểm tra khác như là một phần của tiến trình bởi ứng dụng vận hành chấp nhận nó. Kết quả là, ở thời điểm môi trường thông tin nhìn thấy dữ liệu này, nó đã được đồng hóa vào dữ liệu thời gian thực nội bộ. Như trong **hình 13**, do đó, trao đổi dữ liệu dữ liệu điện tử có tương tác với kho dữ liệu.



Hình 13: The data warehouse and external data

3.4.Siêu dữ liệu trong kho dữ liệu (Metadata in the Data warehouse)

Siêu dữ liệu được yêu cầu có kiến trúc 3 lớp. Tuy nhiên không phải tất cả các kho dữ liệu đều yêu cầu tất cả các lớp này.



Hình 14: **The placement of metadata of the three layer architecture**

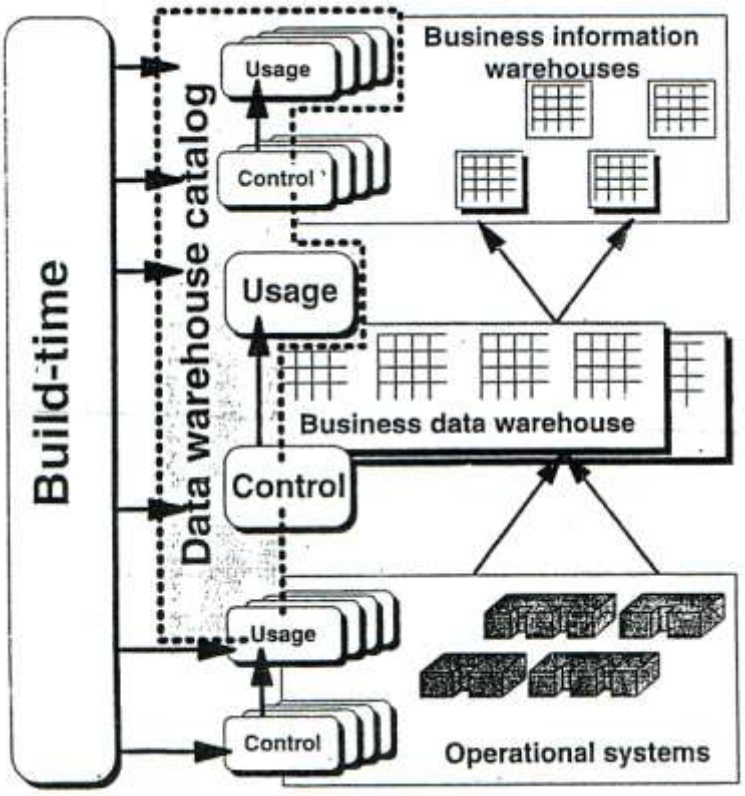
Hình 14 là các yêu kiến trúc thông thường của việc xây dựng siêu dữ liệu thời gian, bao gồm định nghĩa 3 lớp để giải thích mối quan hệ giữa chúng. Đây là khả năng sử dụng các công cụ mô hình khác nhau cho các môi trường khác nhau, nhưng siêu dữ liệu định nghĩa phải được thống nhất.

3.5. Danh mục kho dữ liệu (The data warehouse catalog -DWC):

Trong tập các siêu dữ liệu được định nghĩa, có thể để xác định một tập con cụ thể để sử dụng và quản lý của kho dữ liệu. Tập con này được gọi bằng nhiều tên, ví dụ như "thư mục dữ liệu công việc", "Thư mục thông tin công việc", "thư mục thông tin". Một số những thuật ngữ này chỉ là một phần trong việc sử dụng tập các siêu dữ liệu được đưa vào kho dữ liệu

Chúng ta tập trung vào các nội dung của các siêu dữ liệu, và sử dụng "Danh mục kho dữ liệu - DWC" để mô tả này tập con này. DWC chứa tất cả các siêu dữ liệu cần thiết để sử dụng và quản lý các kho dữ liệu. Như vậy, bao

gồm tất cả các siêu dữ liệu sử dụng và một phần của siêu dữ liệu điều khiển liên kết với các kho dữ liệu công việc và kho thông tin công việc, cũng như một phần của siêu dữ liệu sử dụng liên kết với các hệ thống vận hành như **hình 15**.



Hình 15: The data warehouse catalog

Siêu dữ liệu thời gian xây dựng không bao gồm trong DWC vì quá trình xây dựng các kho là phân chia logic từ quá trình sử dụng và quản lý nó. Tuy nhiên, phần lớn thời gian xây dựng siêu dữ liệu được nhân đôi trong sự kiểm soát và các thành phần sử dụng. Một số siêu dữ liệu điều khiển trong môi trường thông tin cũng được loại trừ khỏi DWC bởi vì siêu dữ liệu này tồn tại chỉ để hỗ trợ các thành phần cơ bản. Các phần của siêu dữ liệu điều khiển bao gồm liên quan đến việc lập kế hoạch và tiền tệ của dữ liệu. DWC cũng bao gồm một phần của siêu dữ liệu sử dụng của các hệ thống vận hành. phần này mô tả việc sử dụng cụ thể của dữ liệu trong môi trường hoạt động có thể khác biệt với trong môi trường thông tin, nhưng có giá trị cho người sử dụng hiểu nguồn gốc cuối cùng của dữ liệu của họ.

DWC và các phương tiện mà người dùng cuối truy cập và sử dụng nó là thành phần quan trọng trong thực hiện bất kỳ kho dữ liệu. Nó đều cung cấp cho người sử dụng khả năng sử dụng hiệu quả các dữ liệu công việc được lưu trữ trong kho.

3.6. Các hệ thống vận hành (Operational systems)

Mặc dù nằm bên ngoài của kho dữ liệu, các hệ thống vận hành là nguồn chủ yếu của kho dữ liệu**.** Cấu trúc và kiến trúc của các hệ thống vận hành là nhân tố chính trong việc xác định độ phức tạp của việc thực hiện một kho dữ liệu.

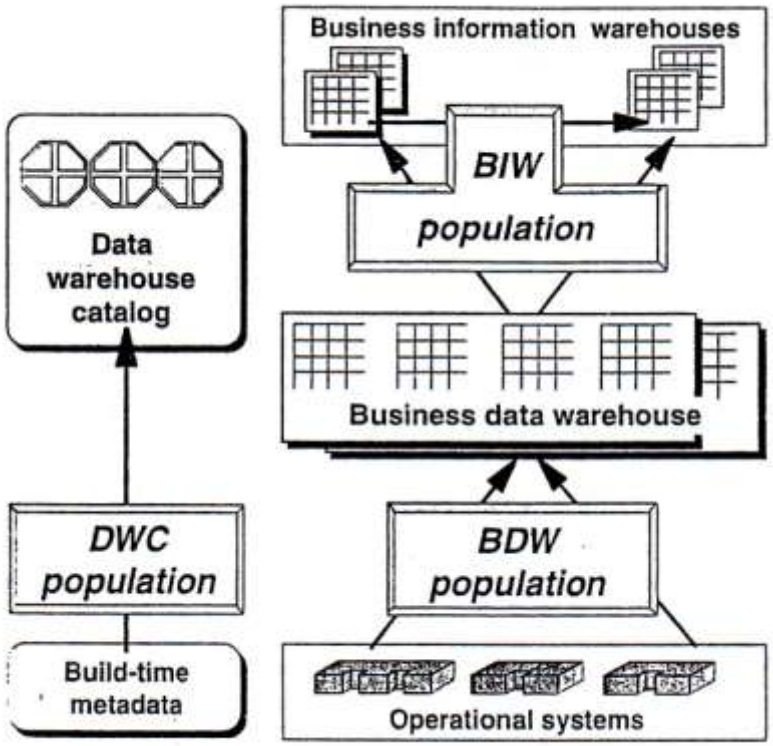
Một bước cơ bản của kho dữ liệu là các hệ thống vận hành không yêu cầu thiết kế lại bất kỳ một quy mô nào theo việc xây dựng kho dữ liệu. Hướng kiến trúc của các hệ thống vận hành thường bắt nguồn từ thiết kế kho dữ liệu của nó.

3.7.Chức năng kho dữ liệu (Data warehouse functionality):

Trong nghiên cứu về kiến trúc logic, chúng ta tập trung vào các khía cạnh liên quan đến dữ liệu, do tầm quan trọng của sự gắn kết, nhất quán, và tích hợp của dữ liệu trong kho. Mức độ quan trọng của chức năng cần thiết để hỗ trợ kiến trúc dữ liệu như mô tả. Phần này giới thiệu và xác định vị trí các chức năng này.

Hình 7.8. thể hiện kiến trúc 3 lớp cho kho dữ liệu công việc, được mở rộng bao gồm siêu dữ liệu. Nó được đơn giản hóa để làm nổi bật sự rõ ràng của kiến trúc.

Có những điểm giống nhau cơ bản giữa các quy trình của sự phổ biến các mục tiêu khác nhau, và sử dụng một tập các công cụ tái tạo dữ liệu. Tuy nhiên, cũng có sự khác biệt đáng kể giữa các loại khác nhau của vị trí. Ví dụ, vị trí kho dữ liệu công việc đòi hỏi phải tăng cường đáng kể độ phức tạp của dữ liệu trong giai đoạn tái tạo hơn so với vị trí của kho thông tin công việc. Tương tự như vậy vị trí của danh mục kho dữ liệu (DWC) ít đòi hỏi về thời gian hơn so với vị trí của kho dữ liệu công việc, kho thông tin kinh công việc. Điều này đưa đến sự khác biệt về chức năng giữa vị trí của BDW, BIW, và DWC như trong **hình 16.**

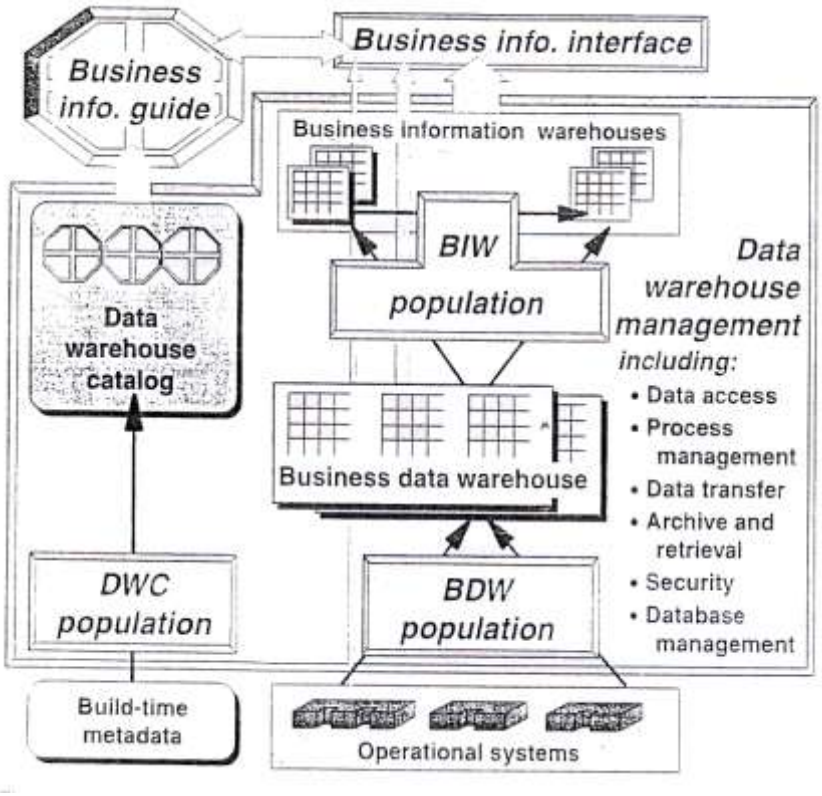


Hình 16: The population functionality of the warehouse

Sự mở rộng thứ hai của các chức năng cung cấp cho việc truy cập và sử dụng các dữ liệu công việc và siêu dữ liệu trong kho. Những người sử dụng cuối sử dụng kho dữ liệu công việc và siêu dữ liệu theo những cách khác nhau. Trong khi dữ liệu công việc được tìm kiếm và phân tích, siêu dữ liệu được khám phá (nhưng không phân tích), từ đó để nắm được các dữ liệu công việc. Những sử dụng khác nhau dẫn đến hai thành phần chức năng. Giao diện thông tin công việc (BII) cung cấp chức năng cần thiết cho dữ liệu công việc. trong khi các hướng dẫn thông tin công việc (BIG) cung cấp chức năng cần thiết cho siêu dữ liệu.

BII (Business information interface) là giao diện để truy cập đến dữ liệu công việc.

BIG (Business information guide) cung cấp các chức năng cần thiết để sử dụng danh mục các kho dữ liệu trong một số cách để tìm dữ liệu công việc liên quan, để nắm được độ quan trọng và lợi ích từ việc sử dụng nó. Chức năng này yêu cầu các truy cập phức tạp hơn đến Danh mục kho dữ liệu (DWC).



Hình 17: The complete logical architecture of the warehouse

Quản lý kho dữ liệu (Data warehouse management) bao gồm một số các chức năng để vận hành và quản lý toàn bộ môi trường kho dữ liệu và các thành phần cơ bản đã được định nghĩa. Bao gồm:

- Truy cập dữ liệu (Data access): Một số định dạng vật lý và vị trí trong dữ liệu có thể yêu cầu các thành phần truy cập dữ liệu.

- Quản lý tiến trình (process management): là cần thiết trong các hoạt động phối hợp, thường vận hành trong các nền khác nhau.

- Vận chuyển dữ liệu (Data transfer) Chức năng vận chuyển dữ liệu là yêu cầu di chuyển dữ liệu vật lý vào trong và bên trong phạm vi kho dữ liệu. Nó cung cấp lớp vận chuyển cần thiết cho chức năng xác định vị trí, hỗ trợ cả về số lượng lớn cà vận chuyển các mức.

- An toàn (Security) Kho dữ liệu bao gồm sở hữu dữ liệu toàn vẹn của tổ chức, an toàn là yêu cầu để điều khiển truy cập và sử dụng dữ liệu trong đó.

- Quản lý cơ sở dữ liệu (Data management) Vì kho dữ liệu được mô tả vật lý như một tập các dữ liệu cơ bản, tập trung và được phân loại nên chức năng quản lý cơ sở dữ liệu là bắt buộc phải có.

Chương 4. NGÔN NGỮ CHO KHO DỮ LIỆU

4.1. Khái niệm

OLAP là "On Line Analytical Processing". là Hệ chuyên xử lý phân tích dữ liệu trực tuyến. Và Data warehouse chính là đầu vào cho quá trình xử lý phân tích trực tuyến.

Do nhu cầu phân tích dữ liệu trước đây hoặc các dữ liệu hiện tại nhằm hỗ trợ cho việc ra quyết định thật chính xác, đúng lúc, giảm rủi ro. Đây cũng là nhu cầu lớn nhất ở mỗi doanh nghiệp nhằm phục vụ các quyết định chiến lược cho công ty. Nhất là các công ty sản xuất lớn với khối lượng dữ liệu lớn.

4.2. Bản chất của OLAP

Bản chất cốt lõi của OLAP là dữ liệu được lấy ra từ Kho dữ liệu hoặc từ dữ liệu chủ đề (Datamart) sau đó được chuyển thành mô hình đa chiều và được lưu trữ trong một kho dữ liệu đa chiều.

4.3. OLAP tập trung vào các câu lệnh sau:

Thu nhỏ (roll-up): ví dụ: nhóm dữ liệu theo năm thay vì theo quý.

Mở rộng (drill-down): ví dụ: mở rộng dữ liệu, nhìn theo tháng thay vì theo quý.

Cắt lát (slice): nhìn theo từng lớp một. Ví dụ: từ danh mục bán hàng của Q1, Q2, Q3, Q4 chỉ xem của Q1

Thu nhỏ (dice): bỏ bớt một phần của dữ liệu (tương ứng thêm điều kiện vào câu lệnh WHERE trong SQL).

4.4. Đối tượng chính của OLAP

Đối tượng chính của OLAP là khối, một sự biểu diễn đa chiều của dữ liệu chi tiết và tổng thể.

Một khối bao gồm một bảng sự kiện (Fact), một hoặc nhiều bảng chiều (Dimensions), các đơn vị đo (Measures) và các phân hoạch (Partitions).

4.4.1. Khối (Cube)

Khối là phần tử chính trong xử lý phân tích trực tuyến, là tập con dữ liệu từ kho dữ liệu, được tổ chức và tổng hợp trong các cấu trúc đa chiều.

Để xác định một khối, ta chọn một bảng Fact và các đơn vị đo lường đồng nhất (các cột số theo sự quan tâm của người dùng khối) trong bảng Fact. Sau đó chọn các chiều, mỗi chiều gồm một hay nhiều cột từ bảng liên quan khác. Các chiều cung cấp mô tả rõ ràng bởi các đơn vị đo lường được chia ra của người dùng khối.

Mỗi chiều có thể chứa một hệ thống các cấp độ để chỉ sự phân chia rõ ràng của người dùng. Mỗi cấp độ trong chiều lại chi tiết hơn mức cha của nó. Ví dụ: lục địa chứa các quốc gia, các bảng hay các tỉnh chứa các thành phố. Tương tự, hệ thống chiều thời gian có thể gồm có các cấp độ năm, quý, tháng và ngày.

4.4.2. Chiều (Dimension)

Các chiều là cách mô tả chủng loại mà theo đó các dữ liệu số trong khối được phân chia để phân tích. Khi xác định một chiều, chọn một hoặc nhiều cột của một trong các bảng liên kết (bảng chiều). Nếu ta chọn các cột phức tạp thì tất cả cần có quan hệ với nhau, chẳng hạn các giá trị của chúng có thể được tổ chức theo hệ thống phân cấp đơn. Để xác định hệ thống phân cấp, sắp xếp các cột từ chung nhất tới cụ thể nhất. Ví dụ: một chiều thời gian (Time) được tạo ra từ các cột Năm, Qúy, Tháng, Ngày (Year, Quarter, Month và Day).

Mỗi cột trong chiều góp phần vào một cấp độ cho chiều. Các cấp độ được sắp đặt theo nét riêng biệt và được tổ chức trong hệ thống cấp bậc mà nó thừa nhận các con đường hợp logic cho việc đào sâu (drill_down). Ví dụ: chiều thời gian được miêu tả ở trên cho phép người dùng khối đào sâu (drill_down) từ Năm tới Qúy, từ Qúy tới Tháng và từ Tháng tới Ngày. Mỗi drill_down cung cấp nét đặc trưng hơn.

❖ Chiều có phân cấp*:* Phân cấp là cột sống của việc gộp dữ liệu hay nói một cách khác là dựa vào các phân cấp mà việc gộp dữ liệu mới có thể thực hiện được. Phần lớn các chiều đều có một cấu trúc đa mức hay phân cấp. Nếu chúng ta làm những quyết định về giá sản phẩm để tối đa doanh thu thì chúng ta cần quan sát ở những dữ liệu về doanh thu sản phẩm được gộp theo giá sản phẩm, tức là chúng ta đã thực hiện một cách gộp. Khi cần làm những

quyết định khác thì chúng ta cần thực hiện những phép gộp tương ứng khác. Như vậy có thể có quá nhiều tiến trình gộp. Thế nên các tiến trình gộp này cần phải được thực hiện một cách rất dễ dàng, linh hoạt để có thể hỗ trợ những phân tích không hoạch định trước. Điều này có thể được giải quyết trên cơ sở có sự trợ giúp của những phân cấp rộng và sâu.

❖ Roll_up và Drill_down dựa trên phân cấp chiều*:* Dựa trên phân cấp theo chiều, từ một mức dưới, chúng ta có thể cuộn lên (Roll_up) các mức trên, thực hiện một phép gộp, để có được kết qủa tổng hợp hơn. Và từ một mức trên, có thể khoan sâu xuống (Drill_down) các mức dưới, để có các kết quả chi tiết hơn.

4.4.3. Các đơn vị đo lường (Measures)

Các đơn vị đo của khối là các cột trong bảng Fact. Các đơn vị đo lường xác định những giá trị số từ bảng Fact mà được tổng hợp phân tích như định giá, trị giá, hoặc số lượng bán.

4.4.4. Các phân hoạch (Partitions)

Tất cả các khối đều có tối thiểu một phân hoạch để chứa dữ liệu của nó; một phân hoạch đơn được tự động tạo ra khi khối được định nghĩa. Khi ta tạo một phân hoạch mới cho một khối, phân hoạch mới này được thêm vào trong tập hợp các phân hoạch đã tồn tại đối với khối. Khối phản ánh dữ liệu đã được kết nối có trong tất cả các phân hoạch của nó. Một bảng phân hoạch của khối là vô hình đối với người dùng.

Các phân hoạch tiêu biểu cho một công cụ mạnh, mềm dẻo cho việc quản trị các khối OLAP, đặc biệt các khối lớn

4.4.5. Một ví dụ vè tổ chức kho dữ liệu trong hệ thống giáo dục

Trong phần này trình bày về . Theo truyền thống, các tổ chức, cơ quan giáo dục không tập trung vào tổng thu nhập và lợi ích, nhưng lại quan tâm nhiều đến giá trị gia tăng và mối quan hệ cạnh tranh về chất lượng giáo dục trong sự thu hút và duy trì chất lượng sinh viên. Trên thực tế, mối quan tâm mạnh mẽ đến sự hiểu biết và mối quan hệ không thuộc phạm vi giáo dục. Nhưng cũng có một bao quát cần thiết để hiểu các khách hàng sinh viên của

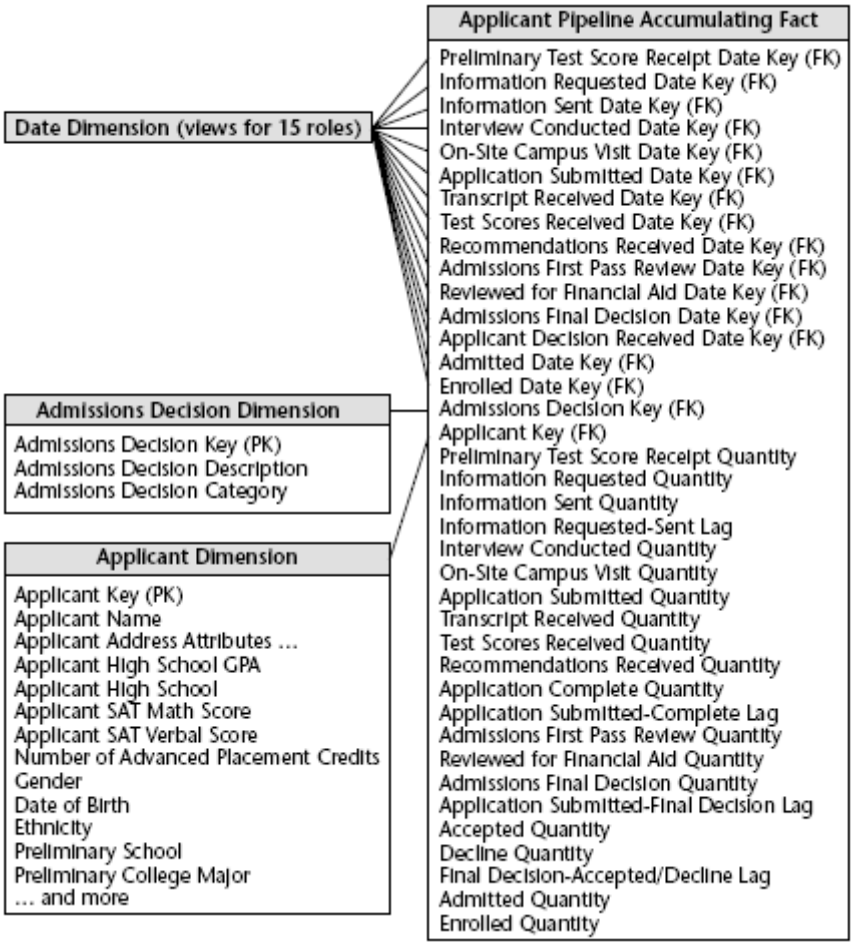
chúng ta là ai, đã mua những khóa học nào. Cuối cùng, chúng ta sẽ có tầm nhìn cao cho việc sử dụng các điều kiện thuận lợi nhất của một trường đại học.

Sau đây là các đặc tính riêng biệt của một bảng fact:

- Mỗi hàng thể hiện lịch sử hoàn thành của một thông tin.
- Một bảng fact là thích hợp nhất cho quá trình tồn tại trong thời gian ngắn, như các yêu cầu hoặc hóa đơn.
- Các tập không giới hạn của các bảng fact tích lũy các đơn vị đo quan tâm.
- Mỗi hàng được duyệt lại hoặc thay đổi khi có một sự kiện xảy ra.
- Cả khóa ngoài và các bảng fact tích lũy có thể thay đổi trong quá trình duyệt.

Trong quá trìn theo dõi đơn xin việc, các sinh viên tương lai xúc tiến thông qua một tập chuẩn hàng trăm, hàng nghìn hồ sơ. Có thể chúng ta quan tâm đến phạm vi hoạt động xung quanh các khóa thời gian như: receipt of preliminary admissions test scores, nformation requested (via Web or otherwise), information sent, interview conducted, on-site campus visit, application received, transcript received, test scores received, recommendations received, first pass review by admissions, review for financial aid, final decision from admissions, accepted, admitted, and enrolled.. Tại bất kỳ thời điểm nào, mọi người được thừa nhận và được kết nạp vùng quản lý có quan tâm đến việc có bao nhiêu đơn xin việc tại mỗi giai đoạn trong quá trình. Những người được phép cũng có thể phân tích sự thiếu đơn xin việc bằng rất nhiều các đặc tính.

Khuynh hướng của sự tích lũy nhanh để lưu vết vòng đời của đơn xin việc tại một hàng cho một sinh viên tương lai. Thể hiện này ở mức thấp nhất của chi tiết được nắm giữ khi các triển vọng vào sắp xảy ra. Rất nhiều thông tin được thu thập trong tiến tới ứng dụng, sự chấp nhận và cho phép, chúng ta tiếp tục duyệt lại và cập nhật các trạng thái triển vọng trong hàng của bảng fact. Hình sau:



Hình 18: Studen applicant pipeline as an accumulating snapshot

Có rất nhiều chiều thời gian trong bảng fact tương ứng với các giai đoạn quan trọng xử lý chuẩn. Chúng ta muốn phân tích sự tiến tới triển vọng bằng thời gian để xác định bước di chuyển thông qua kênh cung cấp, và chúng ta cũng muốn phát hiện ra những con đường hẹp. Điều này đặc biệt quan trọng nếu chúng ta thấy độ trễ quan trọng liên quan đến ứng cử mà chúng ta quan tâm thu hút. Mỗi một thời gian này được xem xét như một role-playing dimention, sử dụng các khóa đại diện để nắm được những thời gian không xảy ra khi dòng đầu tiên được xem đến.

Chiều của đơn xin việc bao gồm một số thuộc tính quan tâm bao gồm các sinh viên khả năng. Các phân tích cho phép được quan tâm trong các lát cắt, khối nhỏ của các đặc tính đơn xin việc bởi vị trí địa lý, khả năng xuất phát, giới tính, ngày sinh, dân tộc, và sơ khảo chính. Phân tích các đặc tính này tại một số giai đoạn của kênh cung cấp sẽ giúp điều chỉnh cá nhân được

phép điều chỉnh các chiến lược của họ để động viên nhiều sinh viên đạt được điểm thi đua tiếp theo.

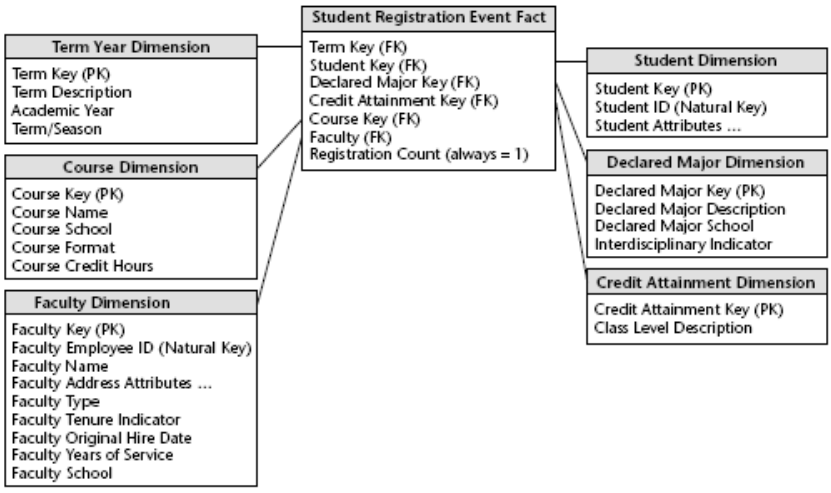
Các bảng fact thực tế (Factless Fact Tables)

Chúng ta thiết kế các bảng fact với một số cấu trúc đặc tính. Mỗi bảng thường có ba đến khoảng 15-20 cột khóa, tiếp theo bởi một hoặc nhiều chữ số, các giá trị tiếp theo, tốt nhất là thêm các sự kiện. Các dữ kiện có thể được coi là phép đo tại sự giao nhau của của các giá trị khóa chiều. Từ quan điểm này, các sự kiện chứng minh cho bảng fact, và các giá trị khóa là cấu trúc điều khiển quản lý để xác định các sự kiện.

Các sự kiện cho sinh viên đăng ký

Có nhiều tình huống trong đó các sự kiện cần phải được ghi lại, đồng thời gắn liền với nhau của một số chiều xác định. Ví dụ, chúng ta có thể theo dõi học sinh đăng ký theo một thời hạn. Khuynh hướng của bảng fact sẽ là một hàng cho mỗi khóa học đăng ký của sinh viên và thời hạn. Như minh họa trong hình-12.2, bảng thực tế đa chiều gồm: thời hạn, sinh viên, chuyên ngành của sinh viên, khóa học, và giảng viên.

Chúng ta đang làm việc với dữ liệu thực tế ở mức độ giới hạn hơn là theo lịch, ngày, tuần, hoặc tháng. Thời hạn là mức thấp nhất có sẵn cho các sự kiện đăng ký. Chiều của thời hạn phải phù hợp đến chiều ngày trong lịch. Nói cách khác, mỗi ngày trong lịch hàng ngày của chúng ta, và giới hạn mùa của năm học.



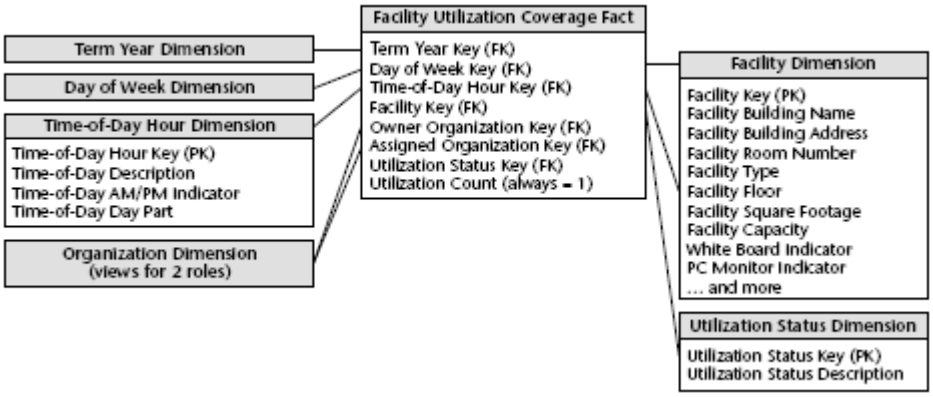
Hình 19:Student registration events as a factless fact table.

Bao trùm sự tận dụng cơ sở vật chất

Kiểu thứ hai của bảng fact thực tế nhất được đưa ra trong bảng sự kiện. Chúng ta đưa ra chuỗi sự kiện phân chia với quản lý cơ sở vật chất để phục vụ cho một minh họa. Các trường đại học dành một lượng vốn lớn trong các dự án cố định và cơ sở vật chất. Nó có thể dễ hiểu khi cơ sở vật chất được sử dụng cho mục đích nào trong suốt thời gian đó. Ví dụ, cơ sở vật chất được sử dụng nhiều nhất là gì? Tỷ lệ sở hữu trung bình của cơ sở vật chất trong chức năng thời gian là bao nhiêu? Sự giảm giá đáng kể vào thứ 6 khi không có ai đến dạy tại các lớp học là bao nhiêu?

Bảng fact kém thực tế có thể bị giải phóng. Trường hợp này bao gồm các hàng trong bảng fact mà mỗi cơ sở vật chất cho khối thời gian chuẩn trong mỗi ngày của mỗi tuần không được dùng tới khi cơ sở vật chất được dùng hoặc không. Minh họa trong hình 20

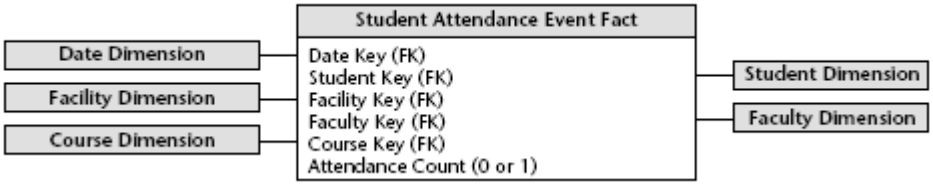
Chiều cơ sở vật chất bao gồm tất cả các kiểu của thuộc tính mô tả về cơ sở vật chất, như toàn nhà, kiểu cơ sở vật chất (VD như phòng học, phòng lab hoặc văn phòng), số m^2, khả năng chứa, và tiện nghi (máy chiếu, bảng trắng..). Chiều của trạng thái tận dụng trước đó bao gồm dòng môt tả với giá trị " Có khả năng" (available) hoặc "được tận dụng" (Utilized). Rất nhiều tổ chức có thể liên quan đến sự tận dụng cơ sở vật chất. Có thể như: một tổ chức sở hữu cơ sở vật chất trong khối thời gian, khi mà một tổ chức khác cũng đăng ký người dùng cơ sở vật chất.



Hình 20: Facilities utilization as a coverage factless fact table

Các sự kiện có mặt của sinh viên:

Chúng ta có thể tưởng tượng giản đồ ghi vết sự có mặt của sinh viên trong một khóa học. Trong trường hợp này một thành phần có thể là một hàng cho mỗi sinh viên đi học qua các phòng học theo khóa mỗi ngày. Bảng fact sự kiện yếu này có thể chia sẻ các chiều giống nhau chúng ta đã thảo luận với khía cạnh các sự kiện đăng ký. Sự khác nhau cơ bản của mỗi thành phần là theo ngày lịch hơn là theo mùa. Mô hình chiều này, được minh họa trong hình sau, cho phép chúng ta trả lời câu hỏi là khóa học nào có sinh viên học đông nhất? Những sinh viên nào đăng ký vào các khóa học nào? Những giáo viên nào dạy phần lớn các sinh viên?



Hình 21: Bảng sự kiện có mặt của sinh viên (Student attendance fact table)

Một số lĩnh vực phân tích đáng quan tâm

Một số xử lý phân tích khác có thể thực hiện trong ví dụ này như: các tài nguyên con người và sự thu nhận, là các khả năng áp dụng trước đó cho môi trường giáo dục đại học đưa ra mong muốn để chi phí điều hành và quản lý tốt hơn. Khi chúng ta tập trung vào cách tính thu nhập, các hỗ trợ cho nghiên cứu, vấn đề nghiên cứu, sự nghiên cứu của giảng viên, và thu nhập từ học phí,..

KẾT LUẬN

Trong thời gian thực hiện đề tài, em đã tìm hiểu và đã trình bày được các vấn đề:

1. Tổng quan về kho dữ liệu: như khái niệm, đặc điểm, lợi ích, mục tiêu, tính chất , thành phần…của kho dữ liệu
2. Các khái niệm cơ bản trong kho dữ liệu
3. Kiến trúc logic kho dữ liệu
4. Ngôn ngữ cho kho dữ liệu và một minh họa cho việc tổ chức kho dữ liệu.

Đồ án bước đầu đã giới thiệu những kiến thức cơ bản về kho dữ liệu, giúp người đọc có cái nhìn tổng quan và căn bản nhất về kho dữ liệu và các khái niệm liên quan. Tuy nhiên do hạn chế về điều kiện thời gian và kiến thức, đồ án không thể tránh khỏi những thiếu xót.Vì vậy em mong nhận được những ý kiến đóng góp của các thầy cô giáo cùng toàn thể các bạn.

Em xin chân thành cảm ơn!

TÀI LIỆU THAM KHẢO

1. Barry Devin, *"Data Warehouse*", Addison Wesley, 1997.
2. Ralph Kimball, Margy Ross," *The Data Warehouse Toolkit", pp 1-65, 243-254*, John Wiley & Sons, Inc, 2002.
3. *http://vi.wikipedia.org/wiki/Kho_d%E1%BB%AF_li%E1%BB%87u*
4. W. H. Inmon, *OLAP and Data Warehouse,* 2000.