Công nghệ kho dữ liệu
và tiếp cận kho dữ liệu của Tổng cục Thống kê

Nguyễn Văn Đoàn (*)

Công nghệ kho dữ liệu đã xuất hiện từ đầu những năm 90 của Thế kỷ XX, người đầu tiên khởi xướng công nghệ kho dữ liệu là B.Inmon[1]. Theo B.Inmon, kho dữ liệu là một sự kết hợp của một số giải pháp kỹ thuật và được đặt tên là Data Warehoushing - kỹ thuật xây dựng kho dữ liệu. Đến nay, thuật ngữ "kho dữ liệu" đã khá phổ biến trong nhiều lĩnh vực. Ở nước ngoài, kho dữ liệu (Data warehouse) đã được phát hành ở dạng ấn phẩm, như "The Microsoft Data Warehouse Toolkit"; "The Data Warehouse Toolkit"; và "Building the Data Warehouse". Mức độ phổ biến của công nghệ kho dữ liệu được minh chứng bằng 2,14 triệu kết quả tìm kiếm trong 0,21 giây của Google[2] trên Internet. Một số trong các kết quả tìm kiếm cho thấy, kho dữ liệu của Google đạt 6 tỷ thư mục, kho dữ liệu miễn phí của Microsoft có dung lượng lên 5GB, kho dữ liệu nhận dạng an ninh của Cục điều tra liên bang Mỹ - FBI; kho dữ liệu về nguồn nước (Resource Data Warehouse)... Vậy, kho dữ liệu là gì? Tại sao kho dữ liệu trở lên phổ biến như hiện nay? Tổng cục Thống kê đã tiếp cận với công nghệ kho dữ liệu này như thế nào? v.v. Để phần nào trả lời các câu hỏi nói trên, bài viết này sẽ đề cập đến 3 vấn đề chủ yếu sau:

(1). Khái niệm, mục đích, yêu cầu và đặc trưng kho dữ liệu:

Kho dữ liệu là kho chứa các dữ liệu điện tử của một tổ chức. Kho dữ liệu được thiết kế để thuận tiện cho việc lập báo cáo và phân tích [1].

Một cách tiếp cận khác, theo John Ladley[3] "Kho dữ liệu là tập các phương pháp, kỹ thuật và các công cụ có thể kết hợp, hỗ trợ nhau để cung cấp thông tin cho người sử dụng trên cơ sở tích hợp từ nhiều nguồn dữ liệu, nhiều môi trường khác nhau". Kho dữ liệu thường rất lớn tới hàng trăm Ghi byte (GB) hay thậm chí hàng Terabyte (TB). Kho dữ liệu được xây dựng để tiện lợi cho việc truy cập theo nhiều nguồn, nhiều kiểu dữ liệu khác nhau sao cho có thể kết hợp được cả những ứng dụng của các công nghệ hiện đại và kế thừa được từ những hệ thống đã có sẵn từ trước [2].

Theo Bách khoa toàn thư "Kho dữ liệu - Data warehouse, một thuật ngữ mới được sử dụng để chỉ những hệ thống thông tin và dữ liệu có tính tích hợp, hướng tới chủ thể quản lí, nhằm trợ giúp cho quá trình làm quyết định của quản lí.

** Viện trưởng Viện Khoa học Thống kê*

[1] Bill Inmon (William Harvey Inmon) (b. July 20, 1945, in San Diego, California) is recognized by many as the "father of data warehousing... As an author, Mr. Inmon has written about a variety of topics on the building, usage, and maintenance of the data warehouse".

[2] Google là công cụ tìm kiếm trên Internet. Bạn chỉ cần gõ "Data Warehouse" trên thanh công cụ tìm kiếm của google sẽ có hơn 2 triệu kết quả trong 0,24 giây.

[3] John Ladley John Ladley is a Business Technology Thought Leader with 30 years experience in information asset management and successful implementation of information systems.
He is widely-known as a data warehouse pioneer and is a recognized authority on enterprise information management, collaborative applications, business intelligence architectures, data architectures, data governance, risk management and assessment of data.

Khác với các cơ sở dữ liệu tác nghiệp đã có từ trước, các nhà kho dữ liệu thường quản trị một lượng thông tin rất lớn, được lưu trữ dưới dạng đa phương tiện, gồm cả thông tin có cấu trúc và không có cấu trúc, thông tin từ nhiều nguồn, thông tin dưới dạng gộp hoặc đã qua tổng hợp, đặc biệt dưới dạng tri thức đã được khai phá và phát hiện từ dữ liệu... nhằm hướng tới chủ thể quản lý để trợ giúp quyết định" [3].

Qua những hình thức khác nhau thể hiện khái niệm kho dữ liệu, chung qui lại, kho dữ liệu thực chất là hệ thống các cơ sở dữ liệu (hay cơ sở dữ liệu cực lớn) có khả năng tích hợp được với nhau và hướng tới chủ thể nhất định, nhằm trợ giúp cho việc ra các quyết định hoặc tạo ra các dữ liệu theo yêu cầu của chính những người khai thác dữ liệu trong kho dữ liệu.

Xây dựng kho dữ liệu nhằm đạt 4 mục đích cơ bản: (i) Tích hợp dữ liệu và các siêu dữ liệu từ nhiều nguồn khác nhau của một đơn vị; (ii) Có khả năng đáp ứng mọi yêu cầu về thông tin của người sử dụng (Người sử dụng có thể khai thác tối đa thông tin sẵn có của đơn vị); (iii) Hỗ trợ thực hiện có hiệu quả các công việc của mỗi thành viên trong đơn vị; (iv) Xác định, quản lý và điều hành các nghiệp vụ một cách hiệu quả và chính xác nhất của người quản lý đơn vị.

Để đạt được các mục đính nói trên, kho dữ liệu cần đảm bảo 7 yêu cầu: (i) Nâng cao chất lượng dữ liệu bằng các phương pháp làm sạch và tinh lọc dữ liệu theo những hướng chủ đề nhất định; (ii) Tổng hợp và kết nối dữ liệu; (iii) Đồng bộ hoá các nguồn dữ liệu với kho dữ liệu; (iv) Phân định và đồng nhất các hệ quản trị cơ sở dữ liệu tác nghiệp như là các công cụ chuẩn để phục vụ cho kho dữ liệu; (v) Quản lí siêu dữ liệu; (vi) Cung cấp thông tin được tích hợp, tóm tắt hoặc được liên kết, tổ chức theo các chủ đề; (vii) Dùng trong các hệ thống hỗ trợ quyết định (Decision Suport System - DSS), các hệ thống thông tin tác nghiệp hoặc hỗ trợ cho các truy vấn đặc biệt. Các yêu cầu nói trên, phải được đảm bảo đồng bộ và ngang nhau, nếu thiếu một trong các yêu cầu, kho dữ liệu sẽ không phát huy được hiệu quả đích thực của nó.

Kho dữ liệu có các đặc tính: Tính tích hợp; Tính chủ đề; Tính lịch sử; Tính ổn định; và tính tổng hợp.

Tính tích hợp: Dữ liệu từ nhiều nguồn khác nhau được tập trung lại theo hướng chủ đề, thống nhất theo một định dạng và tương thích với nhau. Chẳng hạn, Tổng cục Thống kê có nhiều nguồn dữ liệu khác nhau, như: Dữ liệu từ các cuộc điều tra khác nhau, dữ liệu từ các chế độ báo cáo khác nhau. Kho dữ liệu của Tổng cục Thống kê phải làm sao thu gom được các nguồn dữ liệu khác nhau nói trên theo hướng chủ đề, theo một định dạng thống nhất và tương thích được với nhau.

Tính chủ đề: Dữ liệu trong kho dữ liệu được tổ chức theo các chủ đề phục vụ cho đơn vị dễ dàng xác định được những thông tin cần thiết trong từng hoạt động của mình.

Tính lịch sử: Kho dữ liệu lưu trữ khối lượng dữ liệu rất lớn, dữ liệu được lưu trữ theo khoảng thời gian xuyên suốt từ quá khứ đến hiện tại để phục vụ cho việc phát hiện xu hướng nghiệp vụ và nhu cầu phân tích dữ liệu. Trong kho dữ liệu, dữ liệu của hàng chục năm được lưu trữ nhằm phát hiện sự liên hệ của các yếu tố có thể ảnh hưởng đến những chỉ tiêu cần quan tâm trong một thời gian dài.

Tính ổn định: Dữ liệu trong kho dữ liệu không bị thay đổi khi cập nhật dữ liệu mới hay trong quá trình khai thác dữ liệu. Tính ổn định của dữ liệu trong kho dữ liệu cho phép cung cấp thông tin về một khoảng thời gian dài phục vụ cho phân tích, dự báo.

Tính tổng hợp: Đặc tính này cho thấy, chỉ

có những dữ liệu nào có thể tổng hợp được mới đưa vào kho dữ liệu, những dữ liệu nào không thể tổng hợp được, thì không nên đưa vào kho dữ liệu. Chẳng hạn, những thông tin, như họ và tên, địa chỉ của từng người trong dữ liệu tổng điều tra dân số sẽ không được đưa vào kho dữ liệu, vì những thông tin này không thể tổng hợp được.

Trong quá trình thiết kế, xây dựng kho dữ liệu cần đặc biệt chú ý đến những đặc tính nói trên nhằm tối ưu hóa dữ liệu trong kho dữ liệu.

(2) Cấu trúc và lược đồ tổng quát về kho dữ liệu:

Kho dữ liệu, gồm 3 bộ phận cấu thành là: Công nghệ; Dữ liệu; và Siêu dữ liệu. Công nghệ ở đây được hiểu là công nghệ thông tin, là một trong 3 bộ phận quan trọng cấu thành lên kho dữ liệu. Công nghệ của kho dữ liệu, gồm phần cứng, phần mềm và các thiết bị ngoại vi. Phần cứng (máy chủ và các thiết bị công nghệ thông tin khác) để "chứa, lưu trữ" và "vận chuyển" dữ liệu, siêu dữ liệu và phần mềm của kho dữ liệu. Phần mềm để quản lý, khai thác, sắp xếp, phân loại dữ liệu, siêu dữ liệu, như hệ quản trị cơ sở dữ liệu (Oracle, SQL...) và các công cụ để xử lý, phân tích và truy vấn dữ liệu. Chẳng hạn, công cụ phân tích trực tuyến OLAP.

Dữ liệu trong kho dữ liệu là những dữ liệu từ nhiều nguồn dữ liệu khác nhau và ở các dạng khuôn mẫu khác nhau, nhằm đáp ứng không chỉ với các câu hỏi cho trước mà cho cả các câu hỏi chưa được xác định. Dữ liệu trong kho dữ liệu có thể ở các mức độ chi tiết khác nhau. Dữ liệu có thể ở mức chi tiết nhất (hay còn gọi là dữ liệu vi mô) là những dữ liệu cá nhân của một người hoặc một tổ chức. Ví dụ: Các dữ liệu vi mô về dân số, như họ và tên, giới tính, năm sinh, tôn giáo, địa chỉ thường trú... Dữ liệu tổng hợp (hay gọi là dữ liệu vĩ mô) là những dữ liệu đã được tổng hợp từ các dữ liệu vi mô. Ví dụ: Tổng dân số và dân số theo giới tính là những dữ liệu tổng hợp về dân số. Dữ liệu tổng hợp cũng có thể ở các mức độ khác nhau, như dữ liệu tổng hợp ở mức độ thấp, dữ liệu tổng hợp ở mức độ cao. Ví dụ: Doanh thu của doanh nghiệp được theo ngành, theo tháng (dữ liệu tổng hợp ở mức độ thấp), doanh thu của doanh nghiệp theo quí (dữ liệu tổng hợp ở mức độ vừa), doanh thu của doanh nghiệp theo ngành kinh tế, theo năm (dữ liệu tổng hợp ở mức độ cao).

Siêu dữ liệu trong kho dữ liệu, theo Trần Thị Thúy Nga Siêu dữ liệu (matadata) là dữ liệu về dữ liệu được sử dụng trong kho dữ liệu để trả lời các câu hỏi ai, cái gì, khi nào, tại sao, như thế nào về dữ liệu [4]. Nó được sử dụng cho việc xây dựng, duy trì, quản lý và sử dụng kho dữ liệu. Siêu dữ liệu được chia thành 2 loại: Siêu dữ liệu nghiệp vụ, siêu dữ liệu kỹ thuật. Siêu dữ liệu nghiệp vụ là những thông tin giúp cho người sử dụng dễ dàng hiểu được khung cảnh của thông tin được lưu trữ trong kho dữ liệu. Ví dụ: khái niệm và phương pháp tính giá trị sản xuất, giá trị tăng thêm; danh mục ngành kinh tế Việt Nam là những siêu dữ liệu nghiệp vụ. Siêu dữ liệu loại này sẽ giúp các đối tượng hiểu (nhìn) được dữ liệu trong kho dữ liệu. Siêu dữ liệu kỹ thuật là những thông tin về thiết kế và xây dựng kho dữ liệu (giống như bản vẽ kiến trúc tòa nhà), phục vụ các nhà thiết kế, quản lý và phát triển kho dữ liệu.

Một cách tiếp cận khác, cấu trúc kho dữ liệu, gồm có các khối và các lớp.

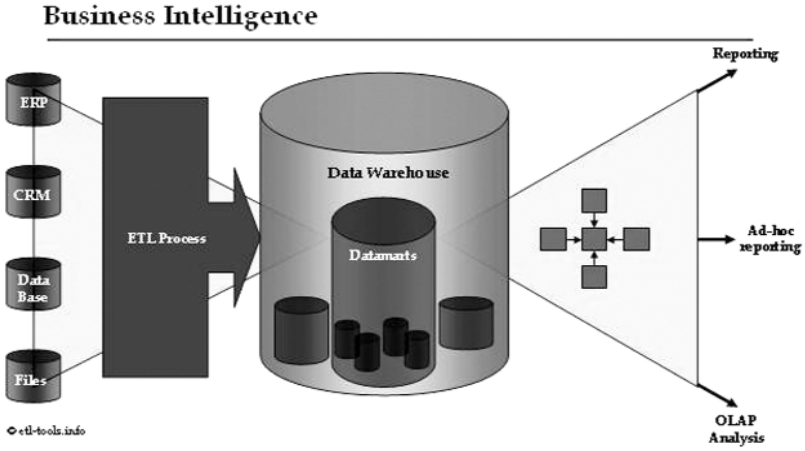
Các khối, gồm: Khối các nguồn dữ liệu; khối tạo dựng kho dữ liệu; khối tạo dựng kho dữ liệu cục bộ; khối truy nhập và sử dụng. Khối các nguồn dữ liệu, bao gồm, dữ liệu sản phẩm (dữ liệu được chắt lọc từ các phần mềm ứng dụng và các hệ CSDL tác nghiệp); dữ liệu kế thừa (có tính lịch sử, phục vụ phân tích); hệ thống dữ liệu bên trong; các hệ thống dữ liệu bên ngoài; hệ

quản lý siêu dữ liệu cho khối này. Khối tạo dựng kho dữ liệu, bao gồm, 3 khối con là: Tinh chế (chuẩn hóa, làm sạch, sàng lọc, tương hợp dữ liệu, phân định thời gian cho các thông tin nguồn); gia công lại (tích hợp các dữ liệu từ các hệ thống để tạo ra dữ liệu mới, phân loại dữ liệu để xử lý, tính toán, chuyển đổi và kiết xuất, chuyển đổi và hình thành lại các dữ liệu từ các nguồn khác nhau, biến đổi và gia công dữ liệu); kho dữ liệu (mô hình hóa, tổng hợp, kết nối mức độ cao các dữ liệu, tạo sự trung hòa các loại dữ liệu, mô tả các loại cơ sở dữ liệu, xây dựng các từ điển, thuật ngữ tác nghiệp). Khối tạo dựng kho dữ liệu cục bộ (Data Mart) theo từng chủ đề. Về nguyên lý và chức năng hoạt động cũng như kho dữ liệu, nhưng ở phạm vi hẹp theo hướng chủ đề nhất định. Tuy theo nhu cầu quản lý và sử dụng, có thể có nhiều kho dữ liệu cục bộ trong kho dữ liệu. Khối truy nhập và sử dụng, bao gồm 2 khối con: Truy nhập; phân tích và tạo báo cáo.

Các lớp, gồm: Lớp quản lý dữ liệu; lớp quản lý siêu dữ liệu; lớp chuyển tải dữ liệu; lớp kết cấu hạ tầng. Lớp quản lý dữ liệu thực hiện các chức năng chính là sao dữ liệu tích hợp từ các nguồn dữ liệu đã chọn phục vụ cho việc tinh chế và gia công lại dữ liệu; giám sát và đáp ứng yêu cầu cho các loại dữ liệu mới rút ra từ các nguồn dữ liệu khác nhau; làm sạch, cập nhật và bảo quản dữ liệu. Lớp quản lý siêu dữ liệu thực hiện chức năng quản lý các siêu dữ liệu trong kho dữ liệu. Lớp chuyển tải dữ liệu thực hiện chức năng chuyển tải dữ liệu giữa các khối trong hệ thống kho dữ liệu, chuyển tải từ kho dữ liệu đến hệ thống mạng... và phân quyền cho các nhu cầu chuyển tải dữ liệu. Lớp kết cấu hạ tầng thực hiện các chức năng chính: quản lý các hệ thống (tìm kiếm, các phần mềm chuẩn, các phần mềm ứng dụng); trợ giúp quá trình tích hợp và các hoạt động khác để sao chép, cập nhật, kết nối, tổng hợp, lưu trữ dữ liệu); hệ thống xử lý (tạo ra môi trường làm việc cho các khối chính).

Những nội dung đã được trình bầy ở trên là những kiến thức tổng quan về kho dữ liệu, thiết nghĩ nó khá trìu tượng và khó hình dung kho dữ liệu có hình hài như thế nào. Lược đồ như Hình 1 dưới đây sẽ thể hiện một cách trực quan hơn về kho dữ liệu.

Hình 1: Lược đồ tổng quát về kho dữ liệu



Nguồn: http://images.google.com

Những hình trụ nhỏ ở bên tay trái thể hiện các nguồn dữ liệu, có thể là các cơ sở dữ liệu riêng của từng lĩnh vực và các dữ liệu tổng hợp theo một số khuôn dạng khác nhau. Thứ đến, là mũi tên mầu xám (ETL Process) thể hiện quá trình triết xuất dữ liệu (Extract), chuyển đổi dữ liệu (Transform) và tải dữ liệu (Load) vào kho dữ liệu. Tiếp đến, hình trụ to ở vị trí giữa (Data Warehouse) thể hiện kho chứa các dữ liệu, siêu dữ liệu (các hình trụ con (Datamarts) trong hình trụ lớn thể hiện các kho dữ liệu cục bộ hay kho dữ liệu chủ đề và các siêu dữ liệu. Các hình còn lại (phía tay phải) thể hiện các sản phẩm đầu ra

của kho dữ liệu (reporting: tạo lập báo cáo; Adhoc reporting: tái tạo báo cáo; OLAP Analysis: phân tích trực tuyến).

Lược đồ và mô tả lược đồ như trên dễ hình dung hơn về kho dữ liệu

(3) Tổng cục Thống kê tiếp cận với công nghệ kho dữ liệu:

Tổng cục Thống kê, hiện tại và tương lai sẽ có rất nhiều loại dữ liệu đến từ các nguồn khác nhau và các dữ liệu này càng ngày, càng nhiều. Nhưng hiện nay, Tổng cục Thống kê chưa có chiến lược tiếp nhận, lưu trữ, quản lý, khai thác và sử dụng các nguồn dữ liệu to lớn này. Ngoài việc xây dựng được một số cơ sở dữ liệu đơn lẻ, như cơ sở dữ liệu Tổng điều tra dân số, nhà ở năm 1999; Tổng điều tra cơ sở kinh tế, hành chính, sự nghiệp năm 2002, 2007; Tổng điều tra nông thôn, nông nghiệp năm 2001, 2006; cơ sở dữ liệu khảo sát mức sống dân cư năm năm 2004, 2006; cơ sở dữ liệu điều tra doanh nghiệp năm 2002, 2003, 2004, 2005, 2006, 2007 và 2008; cơ sở dữ liệu cập nhật doanh nghiệp (BDS)... Các cơ sở dữ liệu nói trên không có bất kỳ một sự liên kết nào với nhau, kể cả các bảng danh mục chuẩn, như Danh mục các đơn vị hành chính, mỗi cơ sở dữ liệu sử dụng một bản sao (coppy) danh mục nói trên. Ngay cả cơ sở dữ liệu của một lĩnh vực nào đó cũng không có sự liên kết với nhau giữa các năm. Chẳng hạn, cơ sở dữ liệu điều tra doanh nghiệp vẫn riêng rẽ từng năm. Các nguồn dữ liệu khác đều được tiếp nhận, lưu trữ ở dạng giấy (hard coppy) hoặc bằng các file dữ liệu rời rạc.

Quy trình tổng hợp và biên soạn số liệu trong Tổng cục Thống kê chưa hợp lý, còn trùng chéo giữa trung ương và địa phương gây lãng phí thời gian, nhân lực và tài chính. Hơn nữa, còn là một trong những nguyên nhân, thậm chí là nguyên nhân quan trọng của hiện tượng chênh lệch số liệu giữa trung ương và địa phương. Các khái niệm, định nghĩa, phương pháp tính, danh mục phân loại (metadata) chưa có sự nhất quán và quản lý tập trung, thống nhất. Chưa có các công cụ cần thiết để người tổng hợp, biên soạn số liệu có thể sử dụng để khai thác các nguồn dữ liệu có sẵn một cách nhanh chóng, thống nhất và hiệu quả.

Để khắc phục được những hạn chế nói trên, Tổng cục Thống kê sẽ phải sử dụng đồng thời nhiều giải pháp. Xây dựng kho dữ liệu thống kê, sẽ là một giải pháp hữu hiệu để quản lý các nguồn dữ liệu rời rạc, phân tán ở hầu hết các vụ nghiệp vụ trong Tổng cục. Nếu kho dữ liệu được hình thành và vận hành tốt sẽ không chỉ khai thác triệt để, có hiệu quả các nguồn dữ liệu hiện có của Tổng cục Thống kê, mà còn tiến tới tích hợp các dữ liệu thống kê của các bộ và dữ liệu thống kê của các tổ chức khác. Nhận thức được ích lợi của công nghệ kho dữ liệu đối với Tổng cục Thống kê, trong hiện tại và tương lai, Viện Nghiên cứu Khoa học Thống kê đang triển khai nghiên cứu đề tài khoa học "Nghiên cứu xây dựng kho dữ liệu đầu vào của Tổng cục Thống kê". Mặc dù công nghệ kho dữ liệu đã xuất hiện từ đầu những năm 1990 của Thế kỷ trước, nhưng khá mới mẻ với ngành Thống kê nói chung và Viện Nghiên cứu Khoa học Thống kê nói riêng. Rất mong nhận được các ý kiến chia sẻ của độc giả Tạp chí Thông tin khoa học thống kê. ■

Tài liệu tham khảo

[1] Wikipedia, the free encyclopedia, Data warehouse, truy cập ngày 28/4/2009, từ http://en.wikipedia.org/wiki/Data_warehouse

[2] John Ladley, truy cập ngày 26/4/2009, từ http://www.dataqualitypro.com/john-ladley-data-quality/

[3] Bách khoa toàn thư, nhà kho dữ liệu, truy cập ngày 9/11/2008, từ http://www.bachkhoatoanthu.gov.vn

[4] Trần Thúy Nga, Luận văn thạc sĩ (2007).