ĐỀ TÀI KHOA HỌC SỐ 2.1.2-B08-09

NGHIÊN CỨU XÂY DỰNG KHO DỮ LIỆU ĐẦU VÀO CỦA TỔNG CỤC THỐNG KÊ

Cấp đề tài	Tổng cục
Thời gian nghiên cứu	2008-2009
Đơn vị thực hiện	Trung tâm Tin học thống kê
Chủ nhiệm đề tài	ThS. Nguyễn Văn Đoàn

MỞ ĐẦU

Công nghệ kho dữ liệu (Data Warehouse Technology) là tập hợp 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 [1]. Như vậy, kho dữ liệu vừa cung cấp dữ liệu, đồng thời cung cấp công cụ giúp người sử dụng có thể truy cập, khai thác và sử dụng được các dữ liệu từ nhiều nguồn khác nhau, phục vụ theo yêu cầu của chính người sử dụng. Đó cũng là điểm khác cơ bản giữa kho dữ liệu với cơ sở dữ liệu.

Tổng cục Thống kê (TCTK) có nhiệm vụ biên soạn số liệu thống kê kinh tế xã hội từ 03 nguồn: (i) Kết quả các cuộc tổng điều tra và điều tra do TCTK trực tiếp thực hiện; (ii) Báo cáo tổng hợp của các Bộ, ngành; (iii) Báo cáo tổng hợp của các Cục Thống kê tỉnh, thành phố trực thuộc trung ương. Các nguồn dữ liệu nói trên khá lớn và được bổ sung hàng năm, nhưng hiện nay, các nguồn dữ liệu này còn phân tán, chưa được quản lý tập trung, tính tương thích khá thấp. Cơ sở dữ liệu (CSDL) được xây dựng riêng rẽ, không tương thích với nhau, không có khả năng liên kết giữa các năm trong một CSDL và giữa các CSDL với nhau để tính các chỉ tiêu thống kê tổng hợp theo những chương trình thống nhất. CSDL siêu dữ liệu (metadata), chưa được xây dựng… Để khắc phục những bất cập nói trên, cần tổ chức lại dữ liệu đầu vào ở TCTK.

Kinh nghiệm của cơ quan thống kê một số nước và tổ chức quốc tế như Hàn Quốc, OECD đã xây dựng được kho dữ liệu phục vụ cho việc biên soạn các số liệu thống kê. TCTK cần xây dựng kho dữ liệu mới có thể giải quyết triệt để những bất cập của các nguồn dữ liệu và qui trình xử lý, tổng hợp, biên

soạn số liệu thống kê nói trên và khai thác triệt để, có hiệu quả các nguồn dữ liệu sẵn có. Do vậy, thực hiện nghiên cứu đề tài "Nghiên cứu xây dựng kho dữ liệu đầu vào của Tổng cục Thống kê" là rất cần thiết và cấp bách vừa có tính thời sự, vừa có tính lâu dài đối với TCTK.

Mục tiêu đề tài nhằm: (i) Nghiên cứu về mặt lý luận và thực tiễn việc xây dựng và ứng dụng kho dữ liệu đầu vào của TCTK; (ii) Đánh giá thực trạng nguồn dữ liệu đầu vào và hạ tầng CNTT tại TCTK; (iii) Thiết kế lý thuyết mô hình kho dữ liệu đầu vào và thử nghiệm mô hình kho dữ liệu trong một lĩnh vực cụ thể.

Báo cáo này trình bầy tóm tắt các kết quả nghiên cứu, ngoài phần mở đầu, kết luận và kiến nghị, Báo cáo được kết cấu thành 3 chương: *Chương I* - Cơ sở lý luận và thực tiễn xây dựng kho dữ liệu phục vụ công tác thống kê; *Chương II* - Thực trạng nguồn dữ liệu, siêu dữ liệu và hạ tầng CNTT tại TCTK; *Chương III* - Đề xuất thiết kế mô hình kho dữ liệu đầu vào.

CHƯƠNG I
CƠ SỞ LÝ LUẬN VÀ THỰC TIỄN XÂY DỰNG KHO DỮ LIỆU PHỤC VỤ CÔNG TÁC THỐNG KÊ

I. Tổng quan nghiên cứu về kho dữ liệu

BCN đề tài đã nghiên cứu nhiều tài liệu về kho dữ liệu và công nghệ xây dựng kho dữ liệu, hầu hết những tài liệu này cung cấp các kiến thức cơ bản về kho dữ liệu và công nghệ xây dựng kho dữ liệu nói chung. Trong đó, có 02 tài liệu đã tiếp cận gần hơn với kho dữ liệu thống kê là "Báo cáo khảo sát của Dự án 00040722 thiết kế và thực hiện kho dữ liệu thống kê của TCTK" và "Báo cáo kết quả đánh giá tình hình thông tin và CNTT và xác định yêu cầu, nội dung phát triển kho dữ liệu của TCTK". Tuy nhiên, 02 tài liệu này chủ yếu trình bày kết quả khảo sát hiện trạng ứng dụng CNTT ở TCTK và đưa ra một số kiến nghị mang tính định hướng cho việc phát triển kho dữ liệu của TCTK, chưa giải quyết được vấn đề làm thế nào để thiết kế được kho dữ liệu của TCTK.

II. Những vấn đề cơ bản về kho dữ liệu

1. Khái niệm, định nghĩa, nội dung và tính hiệu quả của kho dữ liệu

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[16]. 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, nhất là trong lĩnh vực CNTT. Các tài liệu (đề tài đã nghiên cứu) đã trình bày khái niệm kho dữ liệu dưới các gốc độ khác nhau, đề tài trích dẫn 2 khái niệm về kho dữ liệu như sau:

– Kho dữ liệu là tuyển tập các cơ sở dữ liệu tích hợp, hướng chủ đề, được thiết kế để hỗ trợ cho chức năng trợ giúp quyết định. Theo John Ladley, Công nghệ kho dữ liệu (Data Warehouse Technology) là tập các phương pháp, kỹ thuật và các công cụ có thể kết hợp, hỗ trợ nhau để cung cấp thông tin cho người sử dụng trên cơ sở tích hợp từ nhiều nguồn dữ liệu, nhiều môi trường khác nhau [1].

– "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. Khác với các cơ sở dữ liệu tác nghiệp đã có từ trước, các kho dữ liệu thường quản trị 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 lí để trợ giúp quyết định" [4].

Như vậy, có nhiều cách thể hiện khái niệm kho dữ liệu, nhưng thực chất kho dữ liệu cũng là một cơ sở dữ liệu có qui mô rất lớn, các hệ quản trị cơ sở dữ liệu quản lý và lưu trữ nó như một cơ sở dữ liệu thông thường, tuy nhiên có hỗ trợ thêm về quản lý dữ liệu lớn và truy vấn. Kho dữ liệu, thường gồm: Thiết bị CNTT (nhà kho) để chứa, bảo quản dữ liệu; phần mềm (công cụ) để nhập, khai thác dữ liệu; dữ liệu và siêu dữ liệu (số liệu, hình ảnh, âm thanh, chữ viết...) là đối tượng cần lưu trữ, quản lý của kho dữ liệu.

[16] **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. Mục đích, yêu cầu, đặc điểm kho dữ liệu

– Mục đích của kho dữ liệu: Kho dữ liệu nhằm đáp ứng 4 mục đích: (1) Tích hợp dữ liệu và các siêu dữ liệu từ nhiều nguồn khác nhau; (2) Có khả năng đáp ứng mọi yêu cầu về thông tin của người sử dụng; (3) Hỗ trợ thực hiện tốt, hiệu quả công việc của mỗi thành viên trong tổ chức; (4) Giúp cho tổ chức, xác định, quản lý và điều hành các nghiệp vụ một cách hiệu quả và chính xác nhất.

– Yêu cầu của kho dữ liệu: Kho dữ liệu cần đảm bảo được 7 yêu cầu: (1) Nâng cao chất lượng dữ liệu bằng các phương pháp làm sạch và tinh lọc dữ liệu theo những hướng chủ đề nhất định; (2) Tổng hợp và kết nối dữ liệu; (3) Đồng bộ hoá các nguồn dữ liệu với kho dữ liệu; (4) Phân định và đồng nhất các hệ quản trị cơ sở dữ liệu tác nghiệp như các công cụ chuẩn để phục vụ kho dữ liệu; (5) Quản lí siêu dữ liệu; (6) Cung cấp thông tin được tích hợp, tóm tắt hoặc được liên kết, tổ chức theo các chủ đề; (7) Dùng trong các hệ thống hỗ trợ quyết định.

– Đặc tính của kho dữ liệu: Kho dữ liệu là một tập hợp dữ liệu có 5 đặc tính: (1) Tính tích hợp; (2) Hướng chủ đề; (3) Tính lịch sử; (4) Tính ổn định (nonvolatility); (5) Dữ liệu tổng hợp.

3. Thành phần, lược đồ, kiến trúc kho dữ liệu

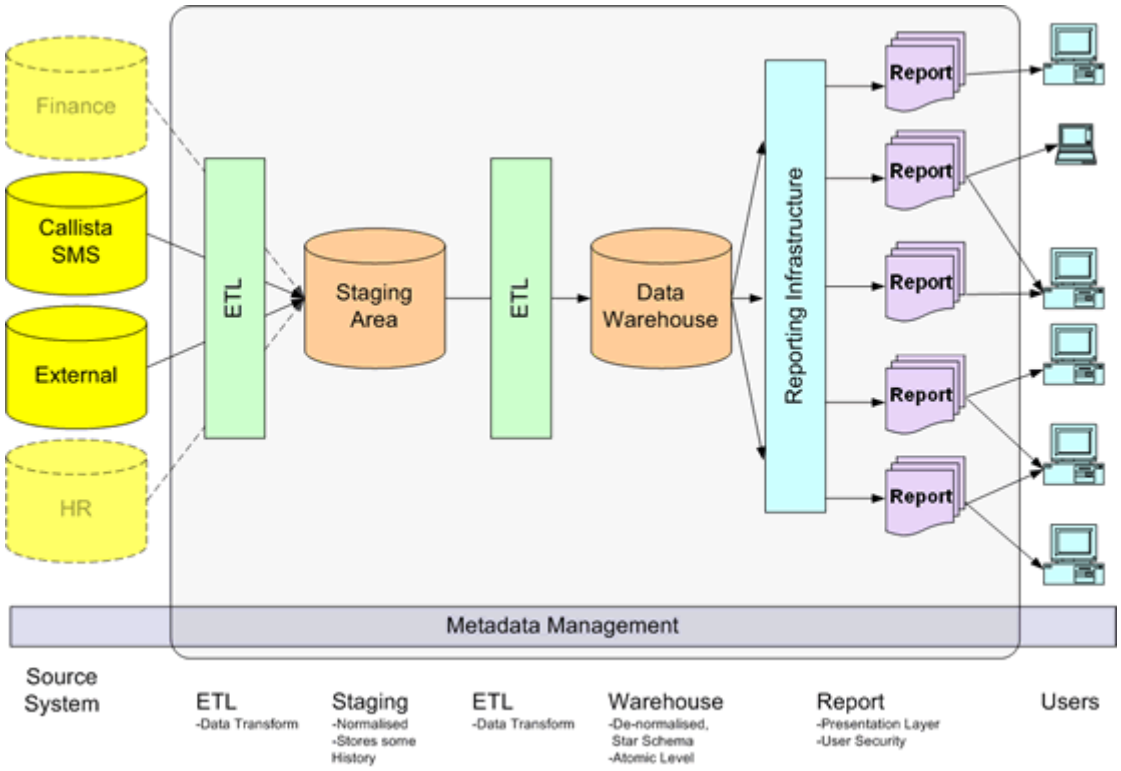
Kho dữ liệu, gồm 3 thành phần: Công nghệ, dữ liệu, và siêu dữ liệu.

– Công nghệ ở đây được hiểu là CNTT là một trong 3 bộ phận quan trọng cấu thành nên kho dữ liệu. Công nghệ của kho dữ liệu, gồm phần cứng, phần mềm máy tính 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" phần mềm, dữ liệu và siêu dữ liệu 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, DB2,…) và công cụ xử lý, phân tích trực tuyến OLAP.

– Dữ liệu trong kho dữ liệu: Dữ liệu đổ vào kho dữ liệu từ nhiều nguồn khác nhau và ở một số dạng khuôn mẫu khác nhau, nhằm đáp ứng không chỉ với các câu hỏi cho trước mà cho cả các câu hỏi chưa xác định. Kho dữ liệu chứa các dữ liệu ở nhiều mức. Mức vi mô (dữ liệu vi mô) là những dữ liệu chi tiết (từng bản ghi) của cá nhân hoặc tổ chức. Mức vĩ mô là những dữ liệu đã được tổng hợp từ các dữ liệu vi mô và được tổng hợp ở các cấp độ khác nhau.

– Siêu dữ liệu trong kho dữ liệu, theo Trần Thị Thúy Nga (2007) Siêu dữ liệu (matadata) là 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 [5]

Mô hình tổng quát về kho dữ liệu được thể hiện như Hình 1 dưới đây.



Hình 1: Sơ đồ Tổng quát kho dữ liệu

Kho dữ liệu được thể hiện theo nhiều dạng, nhưng đều có chung một kiến trúc tham chiếu điển hình được trình bày ở Mục 4 dưới đây.

Kiến trúc tham chiếu kho dữ liệu

Kiến trúc tham chiếu điển hình của kho dữ liệu, bao gồm: 4 khối và 4 lớp.

Bốn khối của kho dữ liệu, gồm: Khối các nguồn dữ liệu; Khối tạo dựng kho dữ liệu; Khối tạo dựng kho dữ liệu cục bộ theo từng chủ đề; Khối truy nhập và sử dụng.

– Bốn lớp của kho dữ liệu, gồm: Lớp quản lý dữ liệu; Lớp quản lý siêu 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 quản lý các hệ thống kho dữ liệu.

III. Kinh nghiệm xây dựng kho dữ liệu

Đề tài đã nghiên cứu kinh nghiệm xây dựng kho dữ liệu của Cơ quan Thống kê quốc gia Hàn Quốc (KNSO), OECD, Macedonia và kho dữ liệu của Ngân hàng Nhà nước và rút ra một số kinh nghiệm sau:

Thứ nhất, TCTK hoàn toàn có thể xây dựng được kho dữ liệu phục vụ việc lập báo cáo thống kê, phân tích và dự báo. Cơ quan thống kê nước nào cũng có chức năng chính là thu thập, xử lý, tổng hợp và phổ biến số liệu. Thống kê các nước và tổ chức OECD nói trên đã xây dựng được kho dữ liệu thống kê của riêng mình, về nguyên tắc, TCTK cũng sẽ xây dựng được kho dữ liệu của TCTK trên cơ sở các nguồn dữ liệu từ điều tra và báo cáo thống kê.

Thứ hai, cần có kế hoạch tổng thể xây dựng kho dữ liệu, kế hoạch tổng thể này sẽ được chia thành nhiều giai đoạn khác nhau (KNSO có 4 giai đoạn, NHNN có 2 giai đoạn).

Thứ ba, xây dựng kho dữ liệu theo chủ đề/lĩnh vực. Dữ liệu thống kê, bao gồm tất cả các lĩnh vực kinh tế, xã hội. Cần lựa chọn một hoặc một số lĩnh vực để xây dựng kho dữ liệu (gọi là kho dữ liệu theo chủ đề).

CHƯƠNG II

THỰC TRẠNG NGUỒN DỮ LIỆU, SIÊU DỮ LIỆU VÀ HẠ TẦNG CÔNG NGHỆ THÔNG TIN TẠI TỔNG CỤC THỐNG KÊ

I. Thực trạng nguồn dữ liệu

1. Nguồn dữ liệu vi mô

Trước năm 2000, dữ liệu vi mô không được xây dựng thành các CSDL mà được lưu trữ thành các tệp (file) thường dưới dạng text hoặc tệp Foxpro đơn giản dưới dạng nén. Thiết bị lưu trữ phổ dụng nhất khi đó là đĩa mềm hoặc trên băng từ. Vì vậy, rất nhiều dữ liệu điều tra trước năm 2000, đến nay không thể sử dụng được.

Sau năm 2000, đã hình thành được một CSDL vi mô được lưu trữ trên các máy chủ sẽ là những điểm thuận lợi cơ bản khi chuyển các dữ liệu này vào kho dữ liệu đầu vào. Tuy nhiên, nguồn dữ liệu vi mô cũng có một số hạn chế, như: Một bộ phận nguồn dữ liệu vi mô chưa được xây dựng thành các CSDL (lưu giữ trên giấy hoặc trong máy tính với các định dạng khác nhau) và phân tán ở nhiều nơi trong Tổng cục; các CSDL vi mô chưa liên kết được với nhau;

không theo chuẩn chung về tên CSDL, tên bảng, tên trường; không có sự thống nhất về loại trường, độ rộng trường cho trường dữ liệu có nội dung giống nhau trong các CSDL khác nhau. Ví dụ như tên bảng danh mục sản phẩm trong CSDL điều tra doanh nghiệp 2003 là DMSP, trong khi đó CSDL Tổng điều tra cơ sở kinh tế, hành chính, sự nghiệp năm 2002 là dmsp. Nguyên nhân của tình trạng này là do không có qui định chuẩn về định dạng kiểu dữ liệu, nên người thiết kế CSDL là người quyết định kiểu dữ liệu của CSDL; Các CSDL vi mô về một lĩnh vực nào đó theo các năm, nhưng mức độ tương thích giữa các năm cũng rất thấp, người sử dụng không thể liên kết CSDL vi mô giữa các năm lại với nhau để tạo dãy số liệu theo năm. Ví dụ, người sử dụng không thể truy cập vào CSDL điều tra doanh nghiệp để trích xuất ra bảng số liệu số doanh nghiệp qua 9 năm (2000 – 2008). Dưới đây sẽ phân tích sâu hơn những hạn chế của CSDL điều tra doanh nghiệp.

Dữ liệu vi mô điều tra doanh nghiệp được lưu trữ theo từng năm, mỗi năm được lưu trữ dưới dạng các tệp (file) có đuôi mở rộng là dbf, số lượng file dữ liệu mỗi năm khác nhau. Cụ thể: năm 2000 có 16 file, năm 2001 có 11 file, năm 2002 có 14 file, năm 2003 có 13 file và năm 2004 có 19 file. Nghiên cứu phiếu điều tra trong các Phương án điều tra doanh nghiệp từng năm, cho thấy, nội dung thông tin cần điều tra rất khác nhau qua các năm. Cụ thể có 16/35 mục tin không được thu thập liên tục trong các năm từ 2000 đến 2005. Nếu phân tích các mục tin ở các mẫu phiếu khác sử dụng trong điều tra doanh nghiệp qua các năm sẽ còn có sự khác biệt rất lớn. *Kết quả phân tích này rất đáng lưu tâm khi xây dựng kho dữ liệu đầu vào của Tổng cục.*

Tiếp theo, chúng ta chọn và phân tích 01 file chứa thông tin chung của doanh nghiệp qua dữ liệu của các năm 2000, 2001, 2002, 2003, 2004 có tên file tương ứng là: DN2000, DN2001, DN2002, DN2003 và DN2004. Kết quả phân tích cho thấy, số lượng trường dữ liệu rất khác nhau giữa các file dữ liệu nói trên. Cụ thể, file DN2000 có số trường ít nhất là 20 trường, file DN2003 có số trường nhiều nhất là 206 trường. Tên các trường dữ liệu cũng rất khác nhau. Ngay đối với những trường dữ liệu có tên giống nhau (dù là rất ít), nhưng vị trí trường trong nguồn dữ liệu cũng khác nhau. Các trường dữ liệu này cũng không được mô tả là trường dữ liệu nào, kiểu trường (bigint, tinyint, numeric...), đơn vị tính (đồng, triệu đồng...), các giá trị của trường... Sự khác biệt đối với file thông tin chung về doanh nghiệp qua các năm sẽ gây rất nhiều khó khăn cho người sử dụng, khai thác dữ liệu doanh nghiệp. *Kết quả phân*

tích trên khẳng định việc lựa chọn và chuẩn hóa dữ liệu điều tra doanh nghiệp sẽ là công việc hết sức nặng nề.

2. Nguồn dữ liệu vĩ mô

Như phần đầu Chương 2 (báo cáo tổng hợp) đã đề cập chi tiết, dữ liệu vĩ mô (hay dữ liệu tổng hợp) có tại TCTK được hình thành từ các dữ liệu vi mô và từ chế độ báo cáo tổng hợp hay khai thác từ hồ sơ hành chính. Dữ liệu dạng này chủ yếu được lưu trữ dưới dạng ấn phẩm (sách, đĩa CD ROM) hoặc tệp dữ liệu không cấu trúc. Vài năm gần đây một số dữ liệu vĩ mô đã xây dựng thành các CSDL với giao diện web, do đó, có sự đồng nhất cao giữa các năm và có thể liên kết dữ liệu nhiều năm với nhau trong một ấn phẩm. Đây sẽ là những điểm thuận lợi khi chuyển dữ liệu này vào kho dữ liệu.

II. Thực trạng siêu dữ liệu

Siêu dữ liệu là một trong 3 thành phần cực kỳ quan trọng của kho dữ liệu. Siêu dữ liệu, bao gồm: Hệ thống bảng danh mục, phân loại; khái niệm, định nghĩa, phương pháp tính chỉ tiêu thống kê; mô tả về dữ liệu, chương trình xử lý, tổng hợp.

1. Hệ thống bảng danh mục sử dụng cho hoạt động thống kê

Hiện nay, hệ thống bảng danh mục (phân loại) đang được sử dụng cho các hoạt động thống kê, gồm các bảng danh mục do Thủ tướng Chính phủ ban hành (ví dụ, Hệ thống ngành kinh tế Việt Nam); các bảng danh mục do Tổng cục trưởng TCTK ban hành (ví dụ, danh mục các đơn vị cơ sở kinh tế). Các bảng danh mục nói trên hầu như chưa được quản lý thống nhất, chưa có sự liên kết, tương thích với nhau, phần lớn trong số đó được lưu trữ ở dạng giấy hoặc tệp phẳng, chưa được tin học hóa. Một số danh mục không được cập nhật trong các CSDL, ví dụ như Bảng danh mục đơn vị hành chính Việt Nam được sửa đổi hàng năm, nhưng các CSDL hiện có không được cập nhật sự thay đổi này. Mỗi CSDL đều chứa một bản sao bảng danh mục riêng… Hiện trạng trên cho thấy, trước hết cần phải tiến hành chuẩn hóa các bảng danh mục và tin học hóa chúng trước khi xây dựng kho dữ liệu

2. Khái niệm, định nghĩa, phương pháp tính các chỉ tiêu thống kê

Phần lớn các khái niệm, định nghĩa, phương pháp tính các chỉ tiêu thống kê đã được thể hiện trong các tài liệu hướng dẫn điều tra của từng cuộc điều tra hoặc trong các chế độ báo cáo thống kê. Nhưng siêu dữ liệu loại này không

được hệ thống hóa theo tiêu chuẩn nhất định, mà chủ yếu ở dạng in ra giấy hoặc file phẳng; không lưu trữ tập trung, mà nằm rải rác ở các Vụ nghiệp vụ. Thậm chí, các mô tả dữ liệu của nhiều năm về trước bị thất lạc, không thể phục hồi. Các CSDL vi mô, vĩ mô được đề cập ở trên cũng không có siêu dữ liệu mô tả dữ liệu trong các CSDL này. Người sử dụng, kể cả những người trực tiếp tạo ra dữ liệu cũng gặp rất nhiều khó khăn khi khai thác các CSDL này.

3. Các mô tả dữ liệu và hồ sơ thiết kế CSDL

Cho đến thời điểm hiện nay, chưa có hệ thông tin mô tả dữ liệu trong các CSDL. Tuy nhiên, trong một vài CSDL đã có một số thông tin như tên gọi bảng dữ liệu, nội dung thông tin của bảng, tên gọi các trường dữ liệu, nội dung thông tin của trường dữ liệu, loại dữ liệu của trường dữ liệu (loại số nguyên, số thập phân, nhị phân, ký tự...) và độ dài của trường dữ liệu. Những thông tin này hiện đang lưu giữ ngay trong CSDL. Về lâu dài, những thông tin dạng mô tả về dữ liệu cần lưu trong cơ sở dữ liệu siêu dữ liệu.

Như vậy, TCTK đã có một số siêu dữ liệu, tuy nhiên, còn quá nhiều hạn chế, như: chưa được tin học hóa; lưu trữ phân tán; tính thống nhất, tương thích của một số khái niêm, định nghĩa giữa các chuyên ngành còn hạn chế. Thực trạng này, sẽ là khó khăn lớn trong quá trình tin học hóa nói chung và xây dựng kho dữ liệu nói riêng.

III. Thực trạng hạ tầng công nghệ thông tin

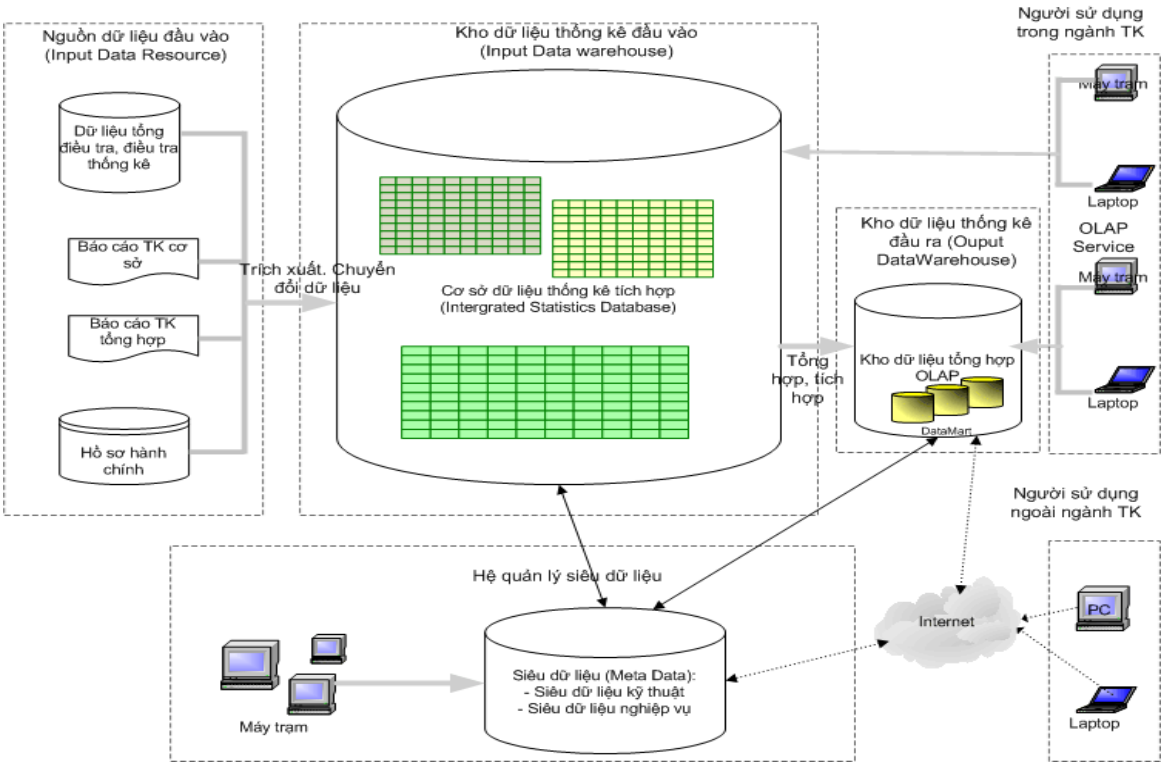
Cơ sở hạ tầng CNTT của TCTK trong những năm gần đây đã được đầu tư, trang bị khá đồng bộ về phần cứng, phần mềm, mạng diện rộng và Internet. Điều đó, cho phép xây dựng kho dữ liệu trên cơ sở hạ tầng CNTT sẵn có của Tổng cục. Tuy nhiên, trong quá trình xây dựng kho dữ liệu, cần thay thế các thiết bị cũ đã hết thời hạn sử dụng và bổ sung thiết bị với dung lượng lớn, thiết bị an ninh, thiết bị giám sát hệ thống, phần mềm chuyển đổi dữ liệu…

CHƯƠNG III

ĐỀ XUẤT THIẾT KẾ MÔ HÌNH KHO DỮ LIỆU ĐẦU VÀO

I. Mô hình kho dữ liệu

1. Mô hình tổng quát kho dữ liệu thống kê



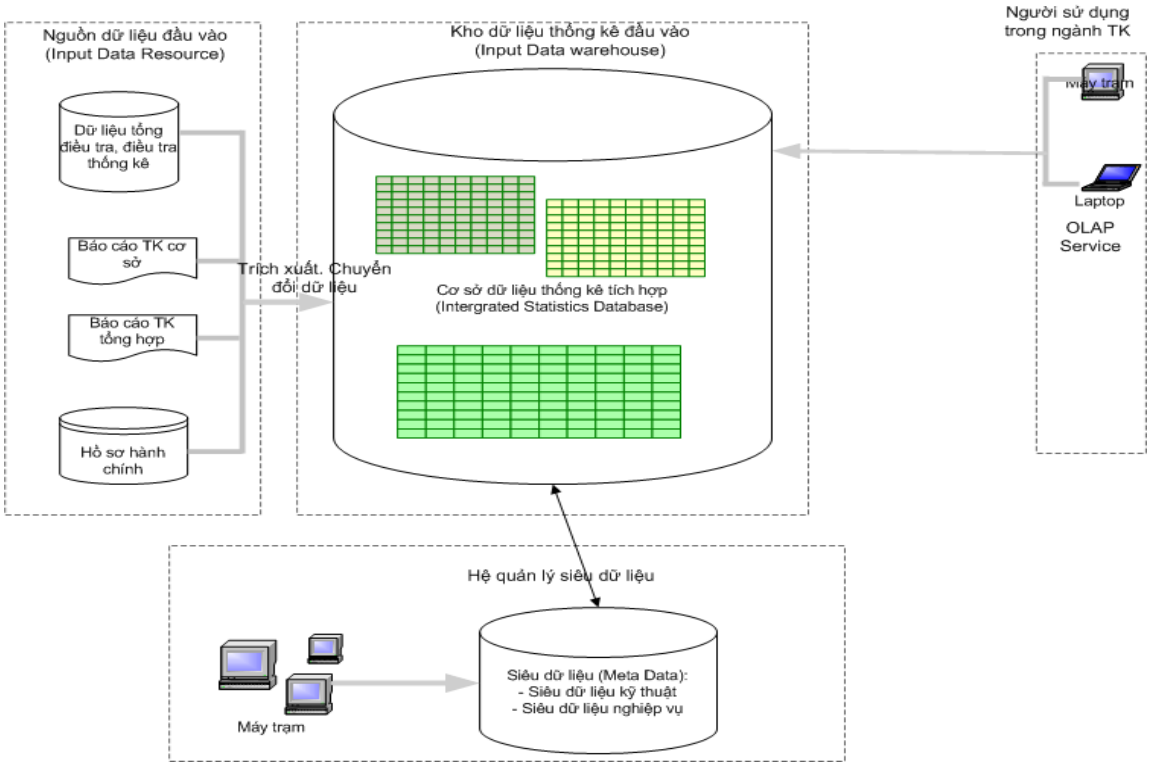
Hình 2: Mô hình tổng quát kho dữ liệu thống kê

2. Mô hình kho dữ liệu đầu vào

Kho dữ liệu đầu vào lưu giữ dữ liệu từ các nguồn khác nhau như tổng điều tra, điều tra thống kê, báo cáo thống kê cơ sở và từ dữ liệu hành chính của các bộ, ngành,... Những dữ liệu cần phải chuyển đổi về dạng thống nhất (tên gọi, đơn vị tính, kiểu dữ liệu, loại dữ liệu,...) và phân tổ theo các bảng danh mục, bảng phân loại dùng chung trước khi được tích hợp, lưu giữ và quản lý tập trung trong kho dữ liệu đầu vào. Dữ liệu trong kho dữ liệu này được liên kết với nhau theo các chiều phân tổ, theo chiều thời gian với cấu trúc dữ liệu thống nhất. Mô hình kho dữ liệu đầu vào đề xuất trong mục 1 "*Mô hình tổng quát kho dữ liệu thống kê*". Tuy nhiên, có thể xây dựng kho dữ liệu đầu vào theo kiểu mô hình ***kho dữ liệu đầu vào tập trung*** hoặc theo mô hình ***kho dữ liệu đầu vào theo chủ đề (lĩnh vực)***.

2.1. Mô hình kho dữ liệu đầu vào tập trung

Kho dữ liệu đầu vào tập trung, tất cả dữ liệu được lưu giữ trong một cơ sở dữ liệu duy nhất. Do các bảng danh mục đã có trong CSDL siêu dữ liệu nên dữ liệu trong kho này là các chữ số. Những số này được chứa trong các bảng (Tables) 2 chiều của cơ sở dữ liệu quan hệ. Số lượng các bảng dữ liệu tùy thuộc vào thiết kế kho dữ liệu, nhưng thông thường người ta thường chia thành nhiều bảng. Mỗi chủ đề hoặc lĩnh vực có thể có một số bảng dữ liệu. Mô hình kho dữ liệu tập trung có dạng như Hình 3 dưới đây.



Hình 3: Mô hình kho dữ liệu đầu vào tập trung

Mô hình kho dữ liệu tập trung có ưu, nhược điểm sau:

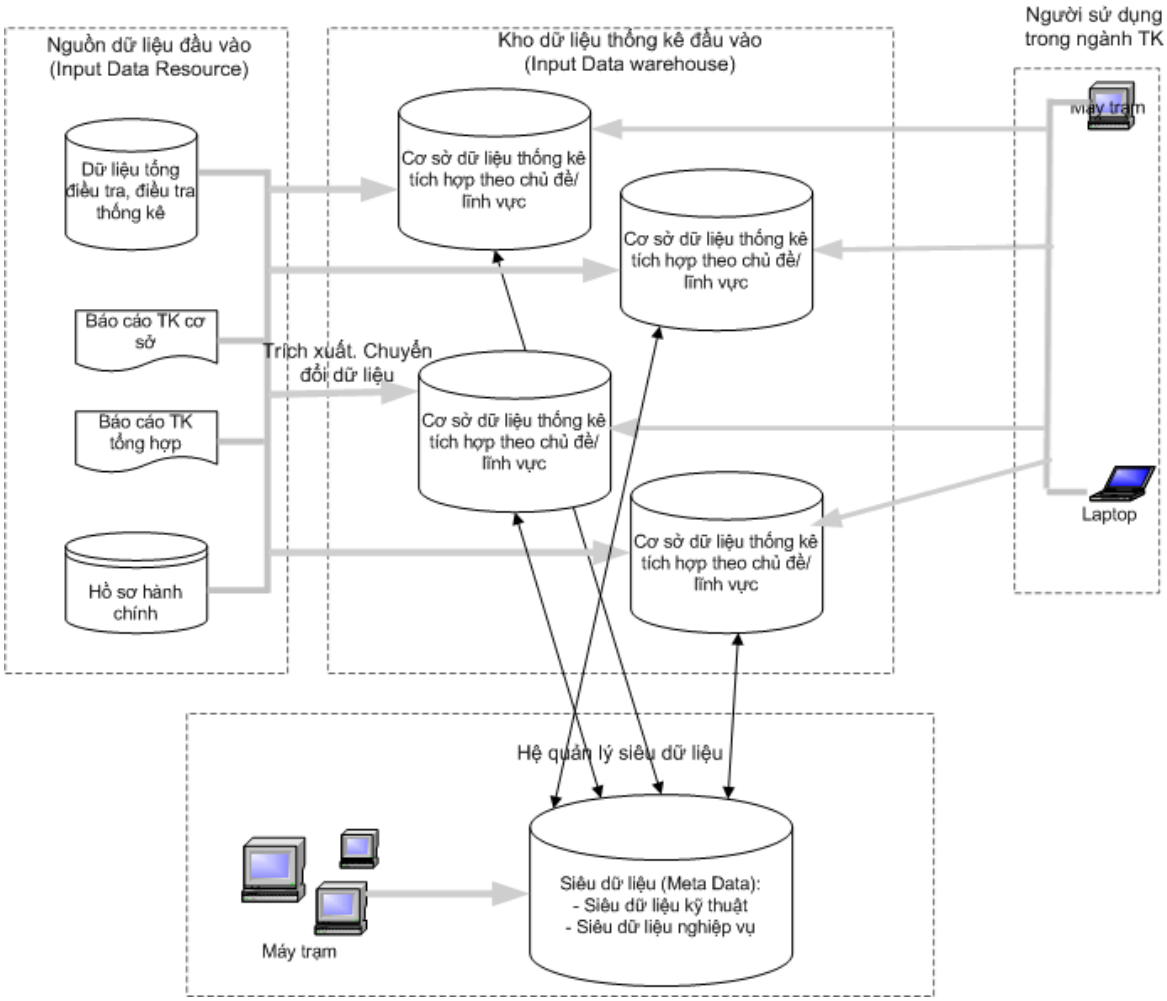
+ ***Ưu điểm:*** Dữ liệu lưu giữ và quản lý tập trung. Việc quản lý, khai thác dữ liệu đơn giản hơn kho dữ liệu xây dựng theo mô hình kho dữ liệu đầu vào theo lĩnh vực hoặc chuyên đề. Việc tìm hiểu về kho dữ liệu đối với người sử dụng nói chung là đơn giản.

+ ***Nhược điểm:*** Do khối lượng dữ liệu rất lớn nên yêu cầu cấu hình máy chủ phải rất cao mới đảm bảo khả năng lưu giữ và đảm bảo yêu cầu về thời gian truy nhập. Đặc biệt, để đảm bảo yêu cầu về thời gian khai thác dữ liệu khi có quá nhiều người truy nhập cùng một lúc đòi hỏi máy chủ phải có nhiều CPU

và bộ nhớ trong (RAM) có dung lượng lớn đến vài chục GB. Do vậy chi phí để đầu tư cao.

2.2. Mô hình kho dữ liệu đầu vào theo chủ đề hoặc theo lĩnh vực

Kho dữ liệu đầu vào theo chủ đề hoặc theo lĩnh vực, dữ liệu được phân theo chủ đề/lĩnh vực và lưu giữ trong cơ sở dữ liệu thống kê tích hợp theo chủ đề (ví dụ, kho dữ liệu chủ đề doanh nghiệp). Do vậy, về thực chất kho dữ liệu đầu vào xây dựng theo mô hình này là một hệ cơ sở dữ liệu riêng biệt, nhưng vẫn có cùng kiểu kiến trúc dữ liệu. Các cơ sở dữ liệu này có thể đặt trên một hoặc trên một số máy chủ khác nhau tại Trung tâm tích hợp dữ liệu. Mô hình kho dữ liệu theo chủ đề như Hình 4 dưới đây.



Hình 4: Mô hình kho dữ liệu đầu vào theo chủ đề

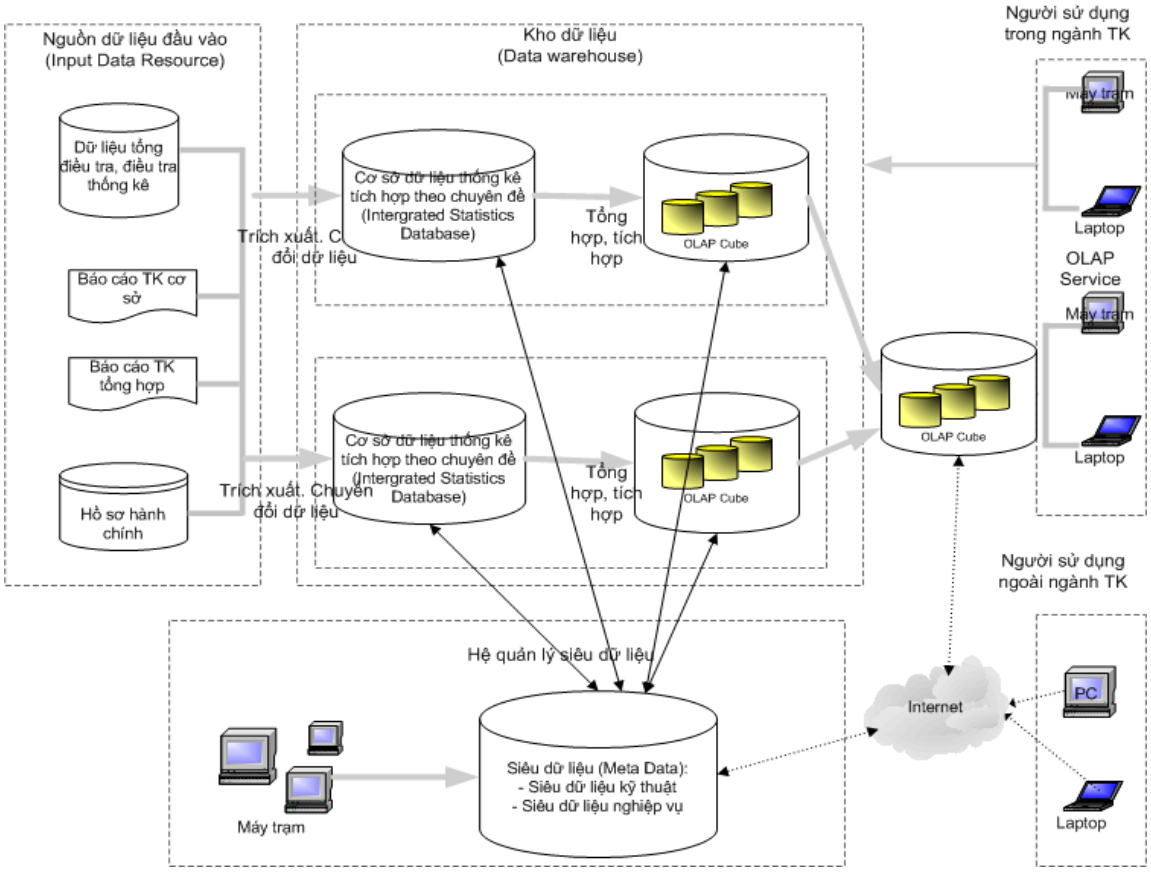
Mô hình kho dữ liệu đầu vào theo chủ đề có ưu, nhược điểm sau:

+ ***Ưu điểm:*** Dữ liệu có thể được lưu giữ trên các máy chủ khác nhau nên khối lượng dữ liệu trong mỗi cơ sở dữ liệu không quá lớn. Do vậy không đòi

hỏi cấu hình máy chủ cao như kiểu kho dữ liệu đầu vào tập trung và tốc độ truy nhập, khai thác có thể tốt hơn.

+ ***Nhược điểm:*** Dữ liệu không tập trung một cách triệt để nên khi cần số liệu của các lĩnh vực khác nhau, phải truy nhập tới các cơ sở dữ liệu khác nhau.

Căn cứ vào thực trạng nguồn dữ liệu, cơ sở hạ tầng CNTT và nguồn lực khác của TCTK, xây dựng kho dữ liệu theo chủ đề là hướng đi phù hợp. Mô hình tổng quát của kho dữ liệu theo chủ đề như Hình 5 dưới đây.



Hình 5: Mô hình kho dữ liệu theo chủ đề

Trong mô hình này, mỗi kho dữ liệu ngoài cơ sở dữ liệu thống kê tích hợp theo chuyên đề đều có cơ sở dữ liệu nhiều chiều (OLAP Cube) phục vụ công tác nghiên cứu, phân tích số liệu thống kê dành cho cán bộ nghiệp vụ thống kê của TCTK. Số liệu trong cơ sở dữ liệu nhiều chiều này rất chi tiết và chỉ cho phép truy nhập theo thẩm quyền. Chỉ những số liệu đã kiểm tra, được phép công bố rộng rãi cho người dùng tin thì mới được chuyển từ cơ sở dữ liệu nhiều chiều này vào cơ sở dữ liệu nhiều chiều (OLAP Cube) ngoài mạng LAN để cung cấp cho người sử dụng qua Internet.

Tóm lại: Qua nghiên cứu, phân tích ưu và nhược điểm của mô hình kho dữ liệu đầu vào tập trung và mô hình kho dữ liệu đầu vào theo chủ đề đồng thời căn cứ vào nguồn lực và vật lực hiện nay của TCTK đề tài đề xuất chọn theo mô hình kho dữ liệu đầu vào theo chủ đề.

II. Thử nghiệm thiết kế kho dữ liệu đầu vào

1. Lựa chọn công nghệ

Hệ quản trị cơ sở dữ liệu Microsoft SQL server phiên bản 2008 là lựa chọn phù hợp nhất để xây dựng kho dữ liệu của TCTK. Vì các CSDL (nguồn dữ liệu) hiện có của TCTK chủ yếu xây dựng trên hệ quản trị Microsoft SQL Server.

2. Phân tích, thiết kế, trích xuất, chuyển đổi dữ liệu

Phân tích, thiết kế là công việc rất quan trọng để đảm bảo kho dữ liệu đáp ứng được yêu cầu đã được đặt ra trong phần mục đích xây dựng kho dữ liệu thống kê. Thiết kế kho dữ liệu cần đảm bảo khả năng tích hợp khối lượng lớn dữ liệu mà vẫn đảm bảo thời gian truy nhập, trích xuất dữ liệu; đảm bảo tính ổn định (không thay đổi theo thời gian); có khả năng mở rộng mà không thay đổi cấu trúc của hệ thống...

Trích xuất, chuyển dữ liệu từ các nguồn dữ liệu khác nhau vào kho dữ liệu: Đây là công việc không thể thiếu khi xây dựng cơ sở dữ liệu. Với cơ sở dữ liệu thông thường, việc chuyển dữ liệu vào trong cơ sở dữ liệu thông qua chương trình ứng dụng còn đối với cơ sở dữ liệu thống kê dạng này là các công việc trích xuất, chuyển đổi và lưu dữ liệu vào trong các bảng dữ liệu. Công việc này được thực hiện bằng chương trình ứng dụng hoặc bằng công cụ có sẵn của hệ thống nào đó, ví dụ như SQL Server Integration Services của Microsoft SQL Server...

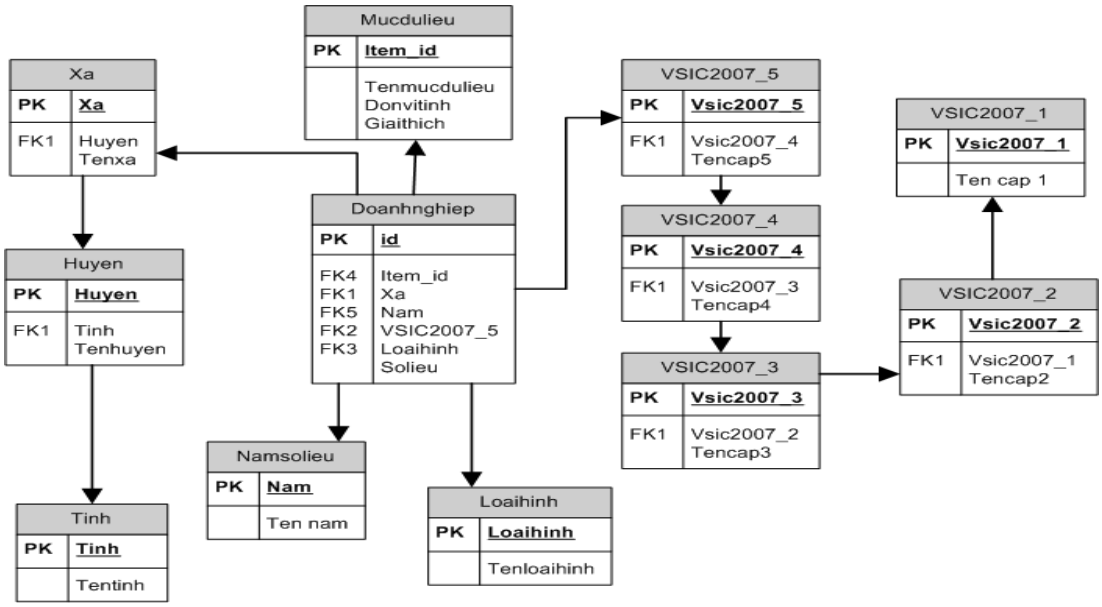
3. Quản lý, khai thác kho dữ liệu đầu vào

Quản lý và khai thác kho dữ liệu đầu vào là một công việc không kém phần quan trọng. Những dữ liệu nào đã được tích hợp vào kho dữ liệu? những dữ liệu nào đã được tích hợp nhưng đã được hiệu chỉnh? quản lý người sử dụng và cấp phát thẩm quyền truy nhập,... là những công việc không thể thiếu của quá trình xây dựng cơ sở dữ liệu nói chung và xây dựng kho dữ liệu đầu vào nói riêng. Thông thường những công việc này do một hoặc một nhóm người

thực hiện bằng các chương trình ứng dụng đã được phát triển khi xây dựng kho dữ liệu.

4. Kết quả thử nghiệm xây dựng kho dữ liệu

Đề tài xây dựng thử nghiệm kho dữ liệu điều tra doanh nghiệp năm 2006, 2007 và 2008 với 5 chỉ tiêu: Số doanh nghiệp, số lao động cuối năm, doanh thu, doanh thu thuần và tổng số tài sản của doanh nghiệp. Kho dữ liệu này được xây dựng thử nghiệm bằng sơ đồ bông tuyết với cấu trúc một cột số liệu. Trung tâm của sơ đồ này là bảng "Doanhnghiep" như Hình 6 dưới đây.



Hình 6: Cấu trúc kho dữ liệu đầu vào doanh nghiệp

Trong bảng "Doanhnghiep", số liệu của các chỉ tiêu nói trên đều lưu giữ trong cột "Solieu". Các cột còn lại (trừ cột id) chứa các mã dùng để liên kết với các bảng chiều phân tổ: Bảng danh mục đơn vị hành chính, bảng loại hình doanh nghiệp, bảng phân ngành kinh tế quốc dân. Các bảng Tinh, Huyen, Xa, Loaihinh, VSIC2007_1, VSIC2007_2, VSIC2007_3, VSIC2007_4, VSIC2007_5 tương ứng với danh mục tỉnh, danh mục huyện, danh mục xã, bảng mã loại hình doanh nghiệp, bảng phân ngành kinh tế quốc dân 2007 cấp 1, cấp 2, cấp 3, cấp 4, cấp 5. Bảng "Mucdulieu" lưu giữ danh mục dữ liệu (Item data). Mỗi mục dữ liệu được gán một mã định danh dùng để kết nối với bảng "Doanhnghiep". Danh mục các năm được lưu giữ trong bảng "Namsolieu". Một điều đáng lưu ý là dữ liệu trong bảng "Doanhnghiệp" trong

kho dữ liệu đầu vào là số liệu điều tra của doanh nghiệp năm 2006, 2007 và 2008.

Trước khi chuyển dữ liệu từ các cơ sở dữ liệu điều tra doanh nghiệp vào kho dữ liệu đầu vào, phải thực hiện các công việc trích xuất, chuyển đổi dữ liệu. Trích xuất dữ liệu thực hiện lấy dữ liệu từ các bảng trong cơ sở dữ liệu điều tra doanh nghiệp, chủ yếu lấy từ Bảng TTCB (thông tin cơ bản) của các năm, riêng năm 2008 thông tin tổng doanh thu và doanh thu thuần được lấy từ bảng kq_sxkd (kết quả sản xuất kinh doanh). Sau khi trích xuất, đã thực hiện kiểm tra những sai sót về mã ngành kinh tế, mã danh mục đơn vị hành chính. Bước cuối cùng là thực hiện chuyển dữ liệu đã chuẩn hóa sang bảng dữ liệu "Doanhnghiệp".

(Xem đề mô kho dữ liệu điều tra doanh nghiệp)

Tóm lại: Kết quả thử nghiệm mô hình kho dữ liệu đầu vào theo chủ đề "Điều tra doanh nghiệp" rất phù hợp với mô hình kho dữ liệu theo chủ đề do Đề tài đề xuất, một lần nữa khẳng định TCTK sớm xây dựng mô hình kho dữ liệu đầu vào theo chủ đề sẽ có tính khả thi.

III. Giải pháp chuyển đổi dữ liệu để đưa vào kho dữ liệu

1. Giải pháp chuyển đổi dữ liệu có định dạng mô hình quan hệ, text, SPSS, Stata, Word, Excel

Hiện tại, TCTK có các nguồn dữ liệu được tạo ra với nhiều định dạng khác nhau, như: Dữ liệu dạng mô hình quan hệ, dữ liệu dang text, dữ liệu dạng SPSS, Statat và dữ liệu dạng Word, Excel. Đối với các dạng dữ liệu nói trên, cần phải có công cụ để chuyển đổi cho đưa vào kho dữ liệu đầu vào. Hiện nay, phần mềm CSPro được sử dụng khá thông dụng cho các cuộc điều tra thống kê của TCTK (kể cả một số cuộc điều tra thống kê không sử dụng CSPro cho xử lý cũng vẫn chuyển dữ liệu vào CSPro cho việc phân phát dữ liệu như một công cụ khai thác dữ liệu). Thêm nữa việc chuyển đổi dữ liệu giữa CSPro với các phần mềm phân tích thống kê SPSS, Stata hay một vài phần mềm khác như IMPS khá dễ dàng. Chính vì thế, rất nhiều cuộc điều tra đã có sẵn hoặc dễ dàng tạo ra bằng một vài thao tác đơn giản file mô tả dữ liệu (từ điển dữ liệu) dưới dạng một file text.

Việc xây dựng một công cụ chuyển đổi dùng chung cho nhiều loại dữ liệu cần được nghiên cứu, đầu tư sau khi có thiết kế chi tiết của kho dữ liệu, như vậy sẽ tiết kiệm được chi phí cho việc xây dựng những chương trình chuyển

đổi dữ liệu riêng cho từng cuộc điều tra, từng loại dữ liệu khác cần đưa vào kho dữ liệu. Tuy nhiên, khó có thể xây dựng được công cụ chuyển đổi được mọi dữ liệu vi mô vào kho dữ liệu, nhất là đáp ứng được những đòi hỏi khác nhau về yêu cầu kiểm tra, làm sạch, biến đổi dữ liệu đặc thù riêng cho mỗi loại dữ liệu, mỗi cuộc điều tra. Một số dữ liệu có thể sẽ phải thực hiện việc kiểm tra, biến đổi riêng biệt trước khi sử dụng công cụ chung này để nạp dữ liệu vào kho dữ liệu.

2. Giải pháp chuyển đổi dữ liệu dạng khác vào kho dữ liệu

Đối với dữ liệu vi mô của một số cuộc điều tra đã được xây dựng thành CSDL SQL Server có thể thực hiện được bằng các công cụ của hệ quản trị CSDL, không nhất thiết phải phát triển các chương trình, công cụ bổ sung.

Đối với dữ liệu chưa được tin học hoá, giải pháp hợp lý nhất phải là tin học hoá trước, tiếp theo mới là việc chuyển đổi vào kho dữ liệu. Tuy nhiên, nếu có những dữ liệu nhất thiết phải đưa vào kho dữ liệu ngay trước khi có hệ thông tin học hoá thì cần xây dựng công cụ nhập liệu vào kho dữ liệu, nhập liệu thủ công từ bàn phím, hoặc có tính năng nhập liệu bằng cách "sao" (copy) từ Excel và "dán" (paste) vào bảng trong CSDL.

3. Giải pháp sử dụng các công cụ chuyển đổi dữ liệu

Việc chuyển đổi dữ liệu vào kho dữ liệu có thể thực hiện bằng các công cụ khác nhau, trong đó công cụ chuyển đổi dữ liệu SQL Server Integration Services (SSIS) của Microsoft sẽ là lựa chọn phù hợp nhất với hạ tầng và các ứng dụng CNTT của TCTK.

4. Yêu cầu chuyển đổi dữ liệu vào kho dữ liệu đầu vào

Chuyển đổi dữ liệu bao gồm các công việc: Rút trích, biến đổi và nạp dữ liệu vào kho dữ liệu (ETL – Extraction, Transformation and Loading).

4.1. Rút trích dữ liệu (Extraction)

Rút trích dữ liệu (extraction) là lựa chọn và cắt ra một tập dữ liệu con tại một thời điểm hay một khoảng thời gian từ cơ sở dữ liệu tác nghiệp để chuẩn bị chuyển đổi vào kho dữ liệu. Rút trích bao gồm cả việc loại bỏ đi những dữ liệu không cần thiết hoặc quá chuyên dụng. Dữ liệu được chọn lọc, rút trích được đưa vào khu vực "chuẩn bị dữ liệu" để chuyển sang bước xử lý làm sạch, biến đổi. Đối với các nguồn dữ liệu đầu vào của TCTK, rút trích dữ liệu chủ yếu là việc lựa chọn những chỉ tiêu, thông tin để đưa vào kho dữ liệu. Về bản chất

những số liệu thống kê đã thu thập được đều là những thông tin hữu ích cần thiết cho những phân tích theo nhiều chiều khác nhau. Việc loại bỏ những thông tin không đưa vào kho dữ liệu trước hết là những thông tin bị đánh giá là không đạt mức độ tin cậy cần thiết do sai số mẫu hay phi mẫu lớn, các thông tin định danh (tên, địa chỉ...), những thông tin này có thể cần thiết cho việc lập dàn mẫu chứ không phục vụ cho mục đích của kho dữ liệu là phân tích dữ liệu thống kê, sản xuất các thông tin đầu ra.

4.2. Biến đổi dữ liệu (Transformation)

Biến đổi dữ liệu, bao gồm, làm sạch dữ liệu và biến đổi dữ liệu. Đây là một quá trình xử lý những dữ liệu đã rút trích trước đó cả về nội dung lẫn hình thức cho phù hợp với các yêu cầu của kho dữ liệu. Làm sạch dữ liệu trước hết phải xử lý, sửa chữa hết những lỗi có thể có trong dữ liệu như: lỗi chính tả, thiếu dữ liệu, sai hoặc không nhất quán khuôn dạng ngày tháng, dữ liệu không nhất quán sai logic, trùng lặp dữ liệu. Sau khi làm sạch dữ liệu, có thể phải thay đổi lại giá trị các mã nếu chưa thống nhất với danh mục trong siêu dữ liệu (dữ liệu đặc tả) trong kho dữ liệu, chuyển đổi khuôn dạng cấu trúc, gán thêm biến thời gian, tạo khoá cho các bản ghi, v.v… Trong rất nhiều trường hợp biến đổi dữ liệu còn bao gồm cả việc tạo thêm những dữ liệu tổng hợp từ dữ liệu ban đầu (tính tổng, trung bình,…), tích hợp dữ liệu.

Biến đổi dữ liệu về mặt cấu trúc dữ liệu còn phải thực hiện nếu trong kho dữ liệu thiết kế trong cùng một bảng dữ liệu vi mô các kỳ điều tra khác nhau có sự khác nhau. Việc ghép dữ liệu vi mô các kỳ điều tra vào cùng các bảng dữ liệu cho phép tổng hợp các Cube (OLAP CUBE) có một chiều là chiều thời gian thuận lợi cho việc phân tích số liệu so sánh các năm. Tuy nhiên việc ghép số liệu của những kỳ điều tra nào với nhau và ghép như thế nào sẽ là những cân nhắc không đơn giản. Nếu chỉ ghép vào bảng những câu thống nhất thì mỗi câu lại có thể chỉ thống nhất chung cho những khoảng thời gian khác nhau. Nếu ghép mọi dữ liệu vào một data mart chung, những câu thống nhất trong cùng trường (field), những trường khác có thể chứa dữ liệu của một, hai, hay nhiều hơn số kỳ điều tra. Cách này làm rộng ra về chiều ngang và người dùng cũng khó khăn hơn với số trường quá lớn mà các trường lại khác biệt về việc có những kỳ điều tra nào có dữ liệu trong mỗi trường.

Như vậy, việc biến đổi dữ liệu phụ thuộc vào thiết kế cấu trúc của kho dữ liệu và các CSDL bên trong kho dữ liệu dựa trên những yêu cầu chung nhất là đảm bảo sự thống nhất giữa các số liệu trong kho dữ liệu, tính so sánh giữa các

số liệu, các thời kỳ, tính minh bạch, dễ hiểu, dễ sử dụng số liệu cho người khai thác kho dữ liệu.

4.3. Nạp dữ liệu (Loading)

Nạp dữ liệu là bước cuối trong quá trình chuyển đổi dữ liệu, thực hiện việc đưa các dữ liệu đã biến đổi vào kho dữ liệu và tạo các chỉ mục. Việc nạp dữ liệu có thể tiến hành bằng cách tạo thêm hay ghi lại (rewrite) một hay nhiều bảng của kho dữ liệu, hoặc chỉ cập nhật những thay đổi vào những bảng này. Kho dữ liệu của TCTK hầu như không có nhu cầu phải cập nhật dữ liệu thay đổi vào những dữ liệu đã có sẵn trong kho mà luôn phải bổ sung dữ liệu mới của các điều tra mới, kỳ điều tra mới với những cấu trúc mới. Như vậy, công cụ thực hiện việc nạp dữ liệu cho kho dữ liệu của TCTK phải là một công cụ mềm dẻo, không cứng nhắc cho một loại dữ liệu nào mà phải tùy biến cho nhiều loại kiểu dữ liệu khác nhau.

KẾT LUẬN VÀ KIẾN NGHỊ

Kết luận

Kho dữ liệu đã khá phổ biến trên thế giới, nhất là trong lĩnh vực tài chính, ngân hàng, hàng không, an ninh, quốc phòng. Đối với lĩnh vực thống kê, dữ liệu khá lớn và gồm các dữ liệu ở các lĩnh vực kinh tế, xã hội, nên việc xây dựng kho dữ liệu để lưu trữ, chia sẻ, cung cấp và sử dụng hiệu quả các nguồn dữ liệu sẵn có là một yêu cầu cần thiết không chỉ đối với cơ quan thống kê, mà còn có ý nghĩa rất lớn đối với xã hội. Theo nghiên cứu của đề tài, đến nay đã có một số cơ quan thống kê quốc gia của một số nước đã xây dựng được kho dữ liệu thống kê của quốc gia mình, như: Hàn quốc, Úc, Newzealand, Canada, Macedonia… Kinh nghiệm xây dựng kho dữ liệu thống kê của các quốc gia nói trên sẽ là cơ sở tham khảo tốt cho việc xây dựng kho dữ liệu đầu vào của TCTK.

Đề tài đã xem xét và đánh giá thực trạng nguồn dữ liệu (dữ liệu vi mô và vĩ mô), siêu dữ liệu (danh mục, khái niệm, định nghĩa các chỉ tiêu thống kê, mô tả dữ liệu và hồ sơ thiết kế CSDL...) và thực trạng hạ tầng CNTT (là các thành phần cơ bản của kho dữ liệu đầu vào), cho thấy, nguồn dữ liệu, siêu dữ liệu và hạ tầng CNTT của TCTK có những thuận lợi nhất định cho việc xây dựng kho dữ liệu đầu vào, như đã hình thành được các CSDL của các cuộc điều tra, tổng điều tra; hệ thống các bảng danh mục chuẩn; hệ thống máy tính (máy chủ, máy

trạm) và mạng nội bộ, mạng diện rộng và Internet thông suốt toàn ngành. Tuy nhiên, còn nhiều hạn chế, bất cập, đòi hỏi phải dành nhiều nguồn lực để gia công, bổ sung và chuẩn hóa một cách đồng bộ mới có thể hình thành được kho dữ liệu đầu vào của TCTK.

Trên cơ sở nghiên cứu tài liệu về kho dữ liệu, công nghệ xây dựng kho dữ liệu, kinh nghiệm xây dựng kho dữ liệu thống kê của một số nước và thực trạng dữ liệu, hạ tầng CNTT của TCTK, đề tài đã thiết kế mô hình tổng quát kho dữ liệu thống kê bao gồm kho dữ liệu đầu vào, kho dữ liệu đầu ra và CSDL siêu dữ liệu, đồng thời thiết kê mô hình, cấu trúc kho dữ liệu đầu vào của TCTK. Đề tài đã tiến hành thử nghiệm mô hình, cấu trúc kho dữ liệu nêu trên trong phạm vi hẹp: xây dựng kho dữ liệu đầu vào từ 5 chỉ tiêu (Số doanh nghiệp, số lao động cuối năm, doanh thu, doanh thu thuần, tổng số tài sản) của doanh nghiệp trong điều tra doanh nghiệp các năm 2006, 2007, 2008. Trên cơ sở kho dữ liệu đầu vào đã xây dựng thử nghiệm kho dữ liệu đầu ra từ dữ liệu trong kho dữ liệu đầu vào và từ một biểu kết quả của TĐTDS&NO năm 1999, đồng thời nghiên cứu và thử nghiệm các công cụ khai thác kho dữ liệu thống kê. Kết quả thử nghiệm của đề tài đã đưa mô hình lý thuyết của kho dữ liệu trở thành kho dữ liệu thực, mặc dù với qui mô rất nhỏ. Kết quả thử nghiệm cho thấy, không nên tách biệt kho dữ liệu đầu vào với kho dữ liệu đầu ra, mà cần thiết kế kho dữ liệu thống kê thống nhất từ các dữ liệu đầu vào đến các sản phẩm đầu ra từ kho dữ liệu. Hệ quản trị CSDL SQL server phiên bản 2008 thích hợp nhất để xây dựng kho dữ liệu của TCTK, và SQL Server Integration Services (SSIS) là công cụ thích hợp nhất để chuyển đổi dữ liệu vào kho dữ liệu. Kết quả thử nghiệm cũng cho thấy, việc chuẩn hóa dữ liệu là một trong khâu khó khăn nhất trong quá trình xây dựng kho dữ liệu.

Với kết quả nghiên cứu của đề tài được trình bày trong báo cáo này và các sản phẩm khác của đề tài, cho phép kết luận: Trong điều kiện hiện tại, TCTK hoàn toàn có thể xây dựng được kho dữ liệu thống kê với qui mô thích hợp, nhằm đáp ứng các yêu cầu sản xuất dữ liệu đầu ra của TCTK.

Kiến nghị

1. TCTK cần triển khai xây dựng kho dữ liệu ngay sau khi đề tài này kết thúc. Nếu không, các kiến thức về xây dựng kho dữ liệu được tích lũy trong quá trình nghiên cứu đề tài sẽ bị mai một, thậm trí ra đi cùng với một số thành viên nghiên cứu của đề tài. Hơn nữa, nguồn dữ liệu của TCTK rất lớn, đa dạng và đã bộc lộ nhiều bất cập trong việc lưu trữ, quản lý, chia sẽ, khai thác sử

dụng, nếu không bắt tay ngay vào việc xây dựng kho dữ liệu đầu vào, thì TCTK không những không hạn chế được bất cập, mà còn làm bất cập tăng lên đến mức sẽ không thể kiểm soát được. Hơn thế nữa, giai đoạn này đang triển khai thực hiện Đề án đổi mới đồng bộ các hệ thống chỉ tiêu thống kê, trong đó, có việc hoàn thiện siêu dữ liệu và chương trình phát triển và ứng dụng CNTT sẽ là cơ sở vững chắc cho việc phát triển kho dữ liệu của TCTK.

2. Kho dữ liệu của TCTK được phát triển theo hướng kho dữ liệu chủ đề (data mark), sau đó sẽ tích hợp các kho dữ liệu chủ đề thành kho dữ liệu thống kê chung. Kho dữ liệu chủ đề về dữ liệu điều tra doanh nghiệp hàng năm và kho dữ liệu Khảo sát mức sống hộ gia đình sẽ là 02 kho dữ liệu chủ đề được lựa chọn xây dựng đầu tiên. Trên cơ sở xây dựng được 02 kho dữ liệu chủ đề này sẽ tiếp tục mở rộng xây dựng kho dữ liệu chủ đề cho một số lĩnh vực khác và tiến tới hình thành kho dữ liệu chung của TCTK.

3. Đồng thời với việc triển khai xây dựng 02 kho dữ liệu chủ đề nói trên, cần tiến hành xây dựng CSDL siêu dữ liệu. Trước hết, tiến hành kiểm kê, tập hợp và phân loại các loại tài liệu về phương pháp chế độ thống kê, trên cơ sở đó tiến hành chuẩn hóa các tài liệu này và xây dựng CSDL siêu dữ liệu phục vụ cho việc thu thập, khai thác, biên soạn dữ liệu. Xây dựng CSDL siêu dữ liệu vào thời gian này là rất thuận lợi, vì toàn ngành đang triển khai thực hiện giải thích các chỉ tiêu thống kê quốc gia, chỉ tiêu thống kê bộ và hoàn thiện các bảng phân loại.

4. Xây dựng kho dữ liệu của TCTK sẽ dựa trên nền tảng CNTT sẵn có của TCTK, đó là công nghệ khách/chủ. Hệ quản trị CSDL thích hợp nhất với công nghệ khách/chủ là Microsoft SQL server phiên bản 2008. Chuyển đổi dữ liệu sẽ sử dụng công cụ có sẵn là SQL Server Integration Services của Microsoft SQL server. Vần đề an ninh, an toàn mạng và bảo mật dữ liệu (không thuộc phạm vi nghiên cứu của đề tài) là vấn đề quan trọng, cần được đầu tư đồng bộ với công nghệ khách/chủ, hệ quản trị CSDL. Việc nâng cấp cơ sở hạ tầng CNTT để đáp ứng yêu cầu phát triển kho dữ liệu sẽ rất thuận tiện, vì TCTK đang thực hiện dự án rất lớn, đó là Dự án hiện đại hóa TCTK do WB tài trợ.

5. TCTK giao cho TTTHTK là đơn vị chủ trì phối hợp Viện KHTK giúp Tổng cục xây dựng kho dữ liệu của TCTK. Lực lượng xây dựng kho dữ liệu sẽ gồm các kỹ sư tin học của TTTHTK, một số cán bộ Viện KHTK là các thành

viên tham gia nghiên cứu đề tài này và một số cán bộ nghiệp vụ thống kê của các Vụ trong Tổng cục.

6. Trung tâm Tin học thống kê cần hình thành đơn vị chuyên trách phát triển kho dữ liệu trên cơ sở Đơn vị "Tích hợp dữ liệu" hiện nay của Trung tâm. Đơn vị này sẽ có trách nhiệm tiếp quản kết quả nghiên cứu của đề tài này và dự thảo kế hoạch tổng thể (hay đề án) phát triển kho dữ liệu của TCTK, trình lãnh đạo Tổng cục phê duyệt và tổ chức thực hiện kế hoạch tổng thể này sau khi lãnh đạo Tổng cục đã phê duyệt.

7. Mời chuyên gia hỗ trợ kỹ thuật xây dựng kho dữ liệu: Xây dựng kho dữ liệu của TCTK là công việc rất lớn và mới của ngành Thống kê, do đó, cần có sự hỗ trợ kỹ thuật của chuyên gia có kinh nghiệm xây dựng kho dữ liệu thống kê. Với kinh nghiệp xây dựng kho dữ liệu thống kê của KNSO, quan hệ hợp tác song phương giữa TCTK với KNSO và với chính sách "Hàn Quốc là bạn với các nước", TCTK nên mời một số chuyên gia của Hàn Quốc giúp xây dựng kho dữ liệu của TCTK.

TÀI LIỆU THAM KHẢO

1. Wikipedia, trang http://en.wikipedia.org/wiki/Data_warehouse.
2. James Wightman, "Pro SQL Server 2005 Integration Services".
3. Designing a Data Warehouse**,** Creating and Using Data Warehouses (SQL Server 2000), Website msdn.microsoft.com.
4. ThS. Nguyễn Thế Quyền, "Giới thiệu về kiến trúc của OLAP", http://www.tapchibcvt.gov.vn/news.
5. Ralph Kimball, Data Warehouse Architect.
6. What is data warehouse, truy cập ngày 5/10/2008, tại trang web: http://www.deakin.edu.au.com.
7. Building the datawarehouse, http://www.ebookee.com/Building-a-Data-Warehouse-With-Examples-in-SQL-Server_154787.html.
8. The datawarehouse Toolkit
9. Báo cáo kết quả của Đoàn khảo sát tại Cơ quan Thống kê quốc gia Hàn Quốc. Tổng cục Thống kê, 2008.
10. Tổng cục Thống kê (2008), "Nghiên cứu xây dựng hệ thống SMIS tại ABS- Australia".

11. Designing and Implementing a DataWarehouse Using Microsoft SQL Server 7.0. Microsoft training and certification, 1998.

12. SQL Server Analysis Services 2005 with MDX. Wiley Publishing, Inc, 2006.

13. TCTK, Phương án điều tra doanh nghiệp năm 2005, 2006, 2007, 2008.

14. Công ty Minh Việt (2007), Báo cáo khảo sát của Dự án 00040722 thiết kế và thực hiện kho dữ liệu thống kê của Tổng cục Thống kê.

15. Lê Mạnh Hùng (2007), Báo cáo kết quả đánh giá tình hình thông tin và công nghệ thông tin và xác định yêu cầu, nội dung phát triển kho dữ liệu của Tổng cục Thống kê.