**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN HỌC PHẦN**

TÊN HỌC PHẦN: **Kho dữ liệu (DWH)**

MÃ SỐ LỚP HP: **DAWH430784\_02**

Tên đề tài Nhóm: **PHÁT TRIỂN MÔ HÌNH KHO DỮ LIỆU VỀ DỊCH VỤ THƯ VIỆN**

Lớp: **181330B**

**NHÓM :02**

Ngô Phi Lít [5]

**Ngày nộp: 02/06/2022**

**TP.HCM, ngày 02 Tháng 06 năm 2022**

**NHẬN XÉT CỦA GIẢNG VIÊN ĐỐI VỚI NHÓM:**

**………………………………………………………………..**

**………………………………………………………………..**

**……………………………………………………………….**

**Ngày 02/06/2022**

**Giảng viên Ký tên**

**PHÂN CÔNG NHÓM : 02**

Ngô Phi Lít [5]: Thành phần ngôn ngữ

**ĐIỂM ĐÁNH GIÁ NHÓM : 02**

Ngô Phi Lít [5] **:**

**Ngày 02/06/2022**

**Giảng viên Ký tên**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**ĐỀ TÀI ĐỒ ÁN HỌC PHẦN**

TÊN HỌC PHẦN: **Kho dữ liệu (DWH)**

MÃ SỐ LỚP HP: **DAWH430784\_02**

Tên đề tài Nhóm: **Phát triển mô hình kho dữ liệu về dịch vụ Thư viện**

**Giảng viên giảng dạy: VÕ XUÂN THỂ \_ Giảng viên chính**

Lớp: **19110IS1, 19110IS2**

**NHÓM : 02**

**Tên sản phẩm đề tài: 10.G2.05.NgoPhiLit.DAHP.DWH\_ThuVien.rar** (Microsoft SQL Server 2014)

**DB Nguồn:** **10G2NgoPhiLitDichVuThuVien**

**DB DWH:** **10G2NgoPhiLitDichVuThuVien\_DWH**

**Tên máy chủ DB SQL: DESKTOP-5AHDFEC\JHIN**

**Tên đăng nhập DB SQL: sa**

**Mật khẩu (password) đăng nhập SQL: 123456**

**ProjectName BIDS:** **10G2NgoPhiLit DichVuThuVien\_DWH** (Microsoft Visual Sudio 2008)

(sv có thể sử dụng tool khác)

**TP.HCM, ngày 02 Tháng 06 năm 2022**

# LỜI CẢM ƠN

Đầu tiên, em đại diện nhóm xin chân thành cảm ơn Thầy Võ Xuân Thể đã hướng dẫn tụi em tận tình trong môn học Kho dữ liệu và hỗ trợ hết mình trong quá trình hoàn thành đồ án học phần. Trong quá trình học tập cũng như trong đồ án học phần thì vẫn còn những thiếu sót, mong thầy góp ý để tụi em cải thiện hơn trong tương lai ạ!

Em xin chân thành cảm ơn!

# DANH MỤC CÁC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| CSDL hoặc DB | Cơ sở dữ liệu: DataBase |
| LĐQH | Lược đồ quan hệ |
| PTH | Phụ thuộc hàm = Ràng buộc tham chiếu: Reference Constraint |
| PTTK | Phân tích và thiết kế |
| RBTV | Ràng buộc toàn vẹn [dữ liệu] |
|  |  |
| BIDS | Công cụ hỗ trợ tổ chức, quản lý và khai thác kho dữ liệu (Hệ thống tổ chức dữ liệu thông minh cho các doanh nghiệp): Business Intelligence Development Studio của Microsoft |
| DWH | Kho dữ liệu (Nhà kho dữ liệu): DataWareHouse |
| ERD | Sơ đồ thực thể kết hợp: Entity Relationship Diagram |
| MS | Công ty Microsoft |
| NF | Dạng chuẩn của CSDL: Normal Form |
| SQL | Ngôn ngữ vấn tin có cấu trúc: Structured Query Language |

# DANH MỤC CÁC THUẬT NGỮ

|  |  |
| --- | --- |
| Giao diện người dùng  (User-Interface) | Là hệ thống các màn hình giao tiếp cho phép người sử dụng tương tác với các thành phần phần mềm, điều khiển phần mềm hoạt động theo yêu cầu của người dùng - tương ứng các chức năng hiện có của phần mềm. |
| Người dùng (User):  Tài khoản (Account) | Là một quyền làm việc trên hệ thống phần mềm được cấp phát cho một cá nhân thông qua tên tài khoản (username) và mật khẩu (password). |
|  |  |
| BigData | Dữ liệu lớn: là một tập hợp dữ liệu rất lớn và phức tạp, không thể xử lý dữ liệu bằng các phương pháp truyền thống. |

# MỤC LỤC

[LỜI CẢM ƠN 4](#_Toc105170903)

[DANH MỤC CÁC TỪ VIẾT TẮT 5](#_Toc105170904)

[DANH MỤC CÁC THUẬT NGỮ 6](#_Toc105170905)

[Chương 1. GIỚI THIỆU VỀ ĐỒ ÁN HỌC PHẦN 10](#_Toc105170907)

[**1.1.** **Tổng quan về ĐỒ ÁN HỌC PHẦN** 10](#_Toc105170908)

[**1.2.** **Nội dung chuyên môn chính của ĐỒ ÁN HỌC PHẦN** 10](#_Toc105170909)

[**1.3.** **Mục tiêu của ĐỒ ÁN HỌC PHẦN** 10](#_Toc105170910)

[**1.4.** **Công cụ và nền tảng kỹ thuật thực hiện ĐỒ ÁN HỌC PHẦN** 10](#_Toc105170911)

[**1.5.** **Sản phẩm của ĐỒ ÁN HỌC PHẦN** 10](#_Toc105170912)

[**1.6.** **Bố cục của báo cáo** 10](#_Toc105170913)

[Chương 2. CƠ SỞ LÝ THUYẾT CỦA HỌC PHẦN 12](#_Toc105170914)

[**2.1.** **Tổng quan về kho dữ liệu (DWH)** 12](#_Toc105170915)

[2.1.1. Các khái niệm cơ bản về Kho dữ liệu 12](#_Toc105170916)

[2.1.2. Kiến trúc cơ bản của kho dữ liệu: 14](#_Toc105170917)

[2.1.3. Mô hình CSDL đa chiều (Dimensional Modeling) 14](#_Toc105170918)

[**2.2.** **Phân tích và thiết kế Kho dữ liệu** 15](#_Toc105170919)

[2.2.1. Phân tích và thiết kế kho dữ liệu theo hướng mô hình hóa nhiều chiều (Dimensional Modeling) 15](#_Toc105170920)

[2.2.2. Các dạng (phương pháp) tích hợp dữ liệu 15](#_Toc105170921)

[2.2.3. Các bước cơ bản trong tiến trình tích hợp dữ liệu 15](#_Toc105170922)

[2.2.4. Các công cụ tích hợp dữ liệu: 15](#_Toc105170923)

[2.2.5. Lập các luồng công việc tích hợp dữ liệu (Data Integration Workflows) 15](#_Toc105170924)

[2.2.6. Hệ thống OLTP (On-Line Transaction Processing = Xử lý giao dịch trực tuyến) 16](#_Toc105170925)

[**2.3.** **Vấn tin khai thác sử dụng Kho dữ liệu** 16](#_Toc105170926)

[2.3.1. Vấn tin với các phép toán tổng hợp của SQL trên kho dữ liệu 16](#_Toc105170927)

[2.3.2. Vấn tin với các hàm phân tích của SQL trên kho dữ liệu 17](#_Toc105170928)

[2.3.3. Vấn tin dùng Materialized View để phân tích trên kho dữ liệu 17](#_Toc105170929)

[2.3.4. Hệ thống OLAP 18](#_Toc105170930)

[Chương 3. TỔNG QUAN VỀ KHO DỮ LIỆU PHÁT TRIỂN MÔ HÌNH KHO DỮ LIỆU VỀ DỊCH VỤ THƯ VIỆN 19](#_Toc105170931)

[**3.1.** **Ý tưởng hình thành DWH Phát triển mô hình kho dữ liệu về dịch vụ thư viện.** 19](#_Toc105170932)

[**3.2.** **Giới thiệu chung về Kho dữ liệu** 19](#_Toc105170933)

[3.2.1. Mô tả tổng quan về DWH 19](#_Toc105170934)

[3.2.2. Các dữ liệu nguồn hình thành DWH 19](#_Toc105170935)

[**3.3.** **Xác định nhu cần tổ chức và phân tích dữ liệu của toàn bộ DWH** 20](#_Toc105170936)

[**3.4.** **Các thành phần thành phần Kho dữ liệu phân công các cá nhân phụ trách** 20](#_Toc105170937)

[Chương 4. PHÂN TÍCH, THIẾT KẾ VÀ KHAI THÁC SỬ DỤNG CÁC THÀNH PHẦN KHO DỮ LIỆU PHÁT TRIỂN MÔ HÌNH KHO DỮ LIỆU VỀ DỊCH VỤ THƯ VIỆN. 22](#_Toc105170938)

[**4.1.** **10\_ Ngô Phi Lít\_THÀNH PHẦN NGÔN NGỮ** 22](#_Toc105170939)

[4.1.1. Mô tả tổng quan về thành phần DWH Ngôn ngữ. 22](#_Toc105170940)

[4.1.2. Xác định nguồn dữ liệu tích hợp thành phần Kho dữ liệu Ngôn ngữ. 22](#_Toc105170941)

[4.1.3. Thiết lập CSDL nguồn của thành phần của Kho dữ liệu Ngôn ngữ. 23](#_Toc105170942)

[4.1.4. Mô tả các Dim bằng tiêu chí phân tích (đánh giá) của thành phần của Kho dữ liệu Ngôn ngữ. 23](#_Toc105170943)

[4.1.5. Mô tả các Facts bằng số liệu phân tích của thành phần của Kho dữ liệu Ngôn ngữ. 24](#_Toc105170944)

[4.1.6. Thiết lập các Views tính toán cần thiết để phân tích của thành phần của Kho dữ liệu Ngôn ngữ. 24](#_Toc105170945)

[4.1.7. Thiết kế DB mới tổ chức phân tích Kho dữ liệu cho thành phần thể loại ngôn ngữ. 25](#_Toc105170946)

[4.1.8. Thực hiện các thủ tục tích hợp các DB gốc vào thành phần DWH. 26](#_Toc105170947)

[4.1.9. Thiết lập các Views tính toán cần thiết để nạp dữ liệu từ DB gốc vào các Factors của thành phần DWH 30](#_Toc105170948)

[4.1.10. Triển khai thành phần DWH lên BIDS 34](#_Toc105170949)

[4.1.11. Khai thác sử dụng thành phần DWH dùng các vấn tin SQL 38](#_Toc105170950)

[Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN CỦA ĐỒ ÁN HỌC PHẦN 40](#_Toc105170951)

[**5.1.** **Kết luận** 40](#_Toc105170952)

[5.1.1. Những kết quả đạt được 40](#_Toc105170953)

[5.1.2. Hạn chế 40](#_Toc105170954)

[**5.2.** **Hướng phát triển** 40](#_Toc105170955)

[5.2.1. Hướng khắc phục các hạn chế 40](#_Toc105170956)

[5.2.2. Hướng mở rộng ĐỒ ÁN HỌC PHẦN 40](#_Toc105170957)

[DANH MỤC TÀI LIỆU THAM KHẢO 41](#_Toc105170958)

[ĐÁNH GIÁ KẾT QUẢ HỌC TẬP HỌC PHẦN 42](#_Toc105170959)

# GIỚI THIỆU VỀ ĐỒ ÁN HỌC PHẦN

## **Tổng quan về ĐỒ ÁN HỌC PHẦN**

Đồ án Học phần “Phát triển mô hình kho dữ liệu về dịch vụ thư viện” được đưa ra khi mà các thư viện đang quản lý quá nhiều thông tin mượn trả tư liệu, loại tư liệu, các thông tin về loại tư liệu…. Lúc ban đầu thì mô hình cơ sở dữ liệu truyền thống hoạt động rất tốt, nhưng càng về sau lượng dữ liệu tăng lên quá nhiều dẫn đến mô hình trên hoạt động không còn tốt và khá tốn tài nguyên. Vì thế mô hình kho dữ liệu để giải quyết vấn đề trên được thực hiện phục vụ cho việc thống kê các số liệu như tổng số lượt mượn theo ngôn ngữ, lĩnh vực, độc giả,….

## **Nội dung chuyên môn chính của ĐỒ ÁN HỌC PHẦN**

* Tìm hiểu và sử dụng được mô hình kho dữ liệu.
* Tìm hiểu và sử dụng một cách thích hợp công cụ tích hợp dữ liệu để nạp dữ liệu từ cơ sở dữ liệu ban đầu vào kho dữ liệu.
* Áp dụng mô hình kho dữ liệu vào chuỗi thư viện trên toàn quốc.

## **Mục tiêu của ĐỒ ÁN HỌC PHẦN**

* Nắm vững và vận dụng được mô hình kho dữ liệu vào thực tế.
* Thiết kế được mô hình kho dữ liệu của chuỗi thư viện trên toàn quốc.
* Nạp dữ liệu bằng SSIS và phân tích bằng BIDS.

## **Công cụ và nền tảng kỹ thuật thực hiện ĐỒ ÁN HỌC PHẦN**

Các công cụ chính sử dụng trong đồ án này là

+ Microsoft SQL Server

+ SQL Server Business Intelligence Development Studio (BIDS)

## **Sản phẩm của ĐỒ ÁN HỌC PHẦN**

* Kho dữ liệu được thiết kế cho chuỗi thư viện trên toàn quốc.
* Sản phẩm BIDS dùng để phân tích các số liệu dựa trên kho đã tạo.

## **Bố cục của báo cáo**

Báo cáo gồm những nội dung như sau:

Chương 1: Giới thiệu đồ án của học phần DWH

Chương 2: Các cơ sở lý thuyết của HP phục vụ việc thực hiện đề tài HP

Chương 3: Giới thiệu tổng quan về Kho dữ liệu (DWH) Kho dữ liệu về dịch vụ thư viện tại các thư viện trên toàn quốc.

Chương 4: Phân tích, thiết kế và tổ chức khai thác sử dụng từng thành phần của Kho dữ liệu

Chương 5: Tổng kết các kết quả đạt được và còn hạn chế của đồ án, đồng thời đề xuất hướng khắc phục hạn chế và phát triển Đồ án.

# CƠ SỞ LÝ THUYẾT CỦA HỌC PHẦN

## **Tổng quan về kho dữ liệu (DWH)**

### Các khái niệm cơ bản về Kho dữ liệu

#### Kho dữ liệu (DWH)?

* Kho dữ liệu là kho lưu trữ dữ liệu lưu trữ bằng thiết bị điện tử của một tổ chức. Các kho dữ liệu được thiết kế để hỗ trợ việc phân tích dữ liệu và lập báo cáo. Mục tiêu của lưu trữ dữ liệu là tạo ra một kho dữ liệu lịch sử có thể được truy xuất và phân tích để cung cấp thông tin chi tiết hữu ích về hoạt động của tổ chức.
* Ba vấn đề chính khi xây dựng nhà kho dữ liệu: hầu hết mọi hệ thống (dự án) công nghệ thông tin, chúng ta đều xem xét dựa trên 3 khía cạnh: cấu trúc (Structure), dữ liệu (Data), tiến trình (Process). Vì vậy khi xây dựng một hệ thống DWH chúng ta cũng phải cân nhắc 3 khía cạnh này. Chúng ta cần phải trả lời được 6 câu hỏiWHAT/WHO/WHERE/WHEN/WHY/HOW

#### Nguyên lý hình thành DWH

* Subject Oriented (hướng chủ đề):

+ Được tổ chức xung quanh các chủ đề chính

+ Tập trung vào việc mô hình hóa và phân tích dữ liệu cho các nhà đưa ra quyết định mà không tập trung vào các hoạt động hay các xử lý giao dịch hàng ngày.

+ Cung cấp một khung nhìn đơn giản và súc tích xung quanh các sự kiện của các chủ đề.

* Integrated (tích hợp):

+ Được xây dựng bằng việc tích hợp dữ liệu từ các nguồn dữ liệu hỗn tạp

+ Các kỹ thuật làm sạch và tích hợp dữ liệu được áp dụng:

* Đảm bảo sự đồng nhất trong các quy ước đặt tên, cấu trúc mã hóa, các đơn vị đo, thuộc tính, … giữa các nguồn khác nhau.
* Khi dữ liệu được chuyển đến kho dữ liệu, nó sẽ được chuyển đổi.
* Non-volatile (bền vững):

+ Là một lưu trữ vật lý của dữ liệu được chuyển đổi từ môi trường tác nghiệp.

+ Cập nhật tác nghiệp của dữ liệu không xuất hiện trong môi trường kho dữ liệu:

* Không yêu cầu các cơ chế xử lý giao dịch, phục hồi và điều khiển tương tranh
* Chỉ yêu cầu hai thao tác truy cập dữ liệu: nạp và truy cập dữ liệu.
* Time-Varying (biến thời gian):

+ Yêu cầu quan trọng cho kho dữ liệu là phạm vi về thời gian dài hơn so với các hệ thống tác nghiệp.

* Cơ sở dữ liệu tác nghiệp: dữ liệu có giá trị hiện thời.
* Dữ liệu của kho dữ liệu: cung cấp thông tin lịch sử

+ Yếu tố thời gian được lưu trữ trong cơ sở dữ liệu.

#### Các đặc điểm cơ bản của DWH

* Định hướng theo chủ đề: không giống như các hệ thống hoạt động, dữ liệu trong kho dữ liệu xoay quanh các đối tượng của doanh nghiệp (chuẩn hóa cơ sở dữ liệu).
* Tích hợp: dữ liệu được tìm thấy trong kho dữ liệu tích hợp. Vì nó đến từ một số hệ thống hoạt động, tất cả các mâu thuẫn phải được loại bỏ. Tính nhất quán bao gồm các quy ước đặt tên, đo lường các biến, cấu trúc mã hóa, các thuộc tính vật lý của dữ liệu, …
* Biến thể thời gian: dữ liệu kho dữ liệu biểu thị dữ liệu trong một khoảng thời gian dài (lên đến 10 năm), có nghĩa là nó lưu trữ dữ liệu lịch sử
* Không biến đổi: dữ liệu trong kho dữ liệu ở chế độ chỉ đọc (không cập nhật, tạo, xóa).

#### Phân biệt [SS] DWH với CSDL truyền thống

* Cơ sở dữ liệu:

+ Là một hệ thống giao dịch theo dõi và cập nhật dữ liệu thời gian thực để có sẵn dữ liệu mới nhất.

+ Phục vụ xử lý transaction, cập nhật

+ Được chuẩn hóa (dạng chuẩn 1, 2, 3, BCK) để khai thác

* Kho dữ liệu

+ Được lập trình để tổng hợp dữ liệu có cấu trúc theo thời gian (tính ổn định).

+ Là một database rất lớn.

+ Thường chỉ đọc, phục vụ cho những nhu cầu báo cáo.

+ Lấy thông tin từ nhiều nguồn khác nhau.

+ Phải phi chuẩn hoá rồi sau đó có thể chuẩn hoá theo start chema trong Data mart, điều này đồng nghĩa với việc DWH sẽ trùng lắp thông tin.

### Kiến trúc cơ bản của kho dữ liệu:

#### Tầng đáy: nạp dữ liệu

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 trung.

#### Tầng giữa: OLAP (OLAP server): xử lý DWH

* 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.
* OLAP (OLAP server) dạng:

+ Relational OLAP

+ Multidimentional OLAP

+ Hybrid OLAP

#### Tầng trên cùng: ứng dụng = khai thác sử dụng DWH

Chứa các câu truy vấn, báo cáo, phân tích, cung cấp thông tin cho người dùng.

### Mô hình CSDL đa chiều (Dimensional Modeling)

* Mô hình CSDL đa chiều (Dimmensional Modeling) là một kỹ thuật cấu trúc dữ liệu được tối ưu hóa để lưu trữ dữ liệu trong kho dữ liệu, giúp biểu diễn dữ liệu logic phù hợp hơn với quản lý dữ liệu OLAP. Mục đích của CSDL đa chiều là tối ưu hóa CSDL để truy xuất nhanh hơn.
* CSDL đa chiều có liên quan đến DWH vì nó tối ưu hóa để lưu trữ trong DWH, DWH được xây dựng theo mô hình dữ liệu đa chiều, hệ thống được dùng để lưu trữ thông tin cho việc phân tích và báo cáo.
* CSDL đa chiều có liên quan đến BigData vì hiện nay với sự bùng nổ của công nghệ cũng dẫn đến lượng dữ liệu ngày càng lớn, do đó việc tổ chức và xử lý dữ liệu cần đến công cụ để quản lý cũng như tránh mất cắp dữ liệu. Mà DWH là một hệ thống để lưu trữ thông tin dữ liệu nên có thể nói CSDL đa chiều cũng liên quan sâu sắc đến BigData.
* CSDL đa chiều có liên quan đến CMCN 4.0 vì có thể thấy bigdata có liên quan rất nhiều đến CMCN 4.0 vì công nghiệp phát triển thì dữ liệu cũng ngày càng lớn thậm chí là rất lớn.

## **Phân tích và thiết kế Kho dữ liệu**

### Phân tích và thiết kế kho dữ liệu theo hướng mô hình hóa nhiều chiều (Dimensional Modeling)

* Lược đồ hình sao: Gồm 1 bảng Fact (bảng sự kiện) nằm ở trung tâm và được bao quanh bởi những bảng Dimension (bảng chiều). Dữ liệu của lược đồ hình sao không được chuẩn hoá. Các câu hỏi nhằm vào bảng Fact và được cấu trúc bởi các bảng Dimension.

+ Ưu điểm: Bảng Fact, Dimension được mô tả rõ ràng, dễ hiểu. Bảng Dim là bảng dữ liệu tĩnh, và bảng Fact là dữ liệu động được nạp bằng các thao tác. Khoá của Fact được tạo bởi khoá của các bảng Dim. Nghĩa là khoá chính của các bảng Dim chính là khoá của bảng Fact.

+ Nhược: Dữ liệu không được chuẩn hoá.

* Lược đồ hình bông tuyết:

+ Là dạng mở rộng của lược đồ hình sao bằng các bổ sung các Dim. Bảng Fact như lược đồ hình sao, bảng Dim được chuẩn hoá. Các chiều được cấu trúc rõ ràng. Bảng Dim được chia thành chiều chính hay chiều phụ.

+ Ưu điểm: Số chiều được phân cấp thể hiện dạng chuẩn của bảng Dim. Nhược: Cấu trúc phi dạng chuẩn của lược đồ hình sao phù hợp hơn cho việc duyệt các chiều.

### Các dạng (phương pháp) tích hợp dữ liệu

### Các bước cơ bản trong tiến trình tích hợp dữ liệu

* Bước 1: Trích xuất dữ liệu (thu thập dữ liệu lớn từ nhiều nguồn/ đa nền tảng)
* Bước 2: Làm sạch (tìm lỗi và sửa Database nguồn)
* Bước 3: So khớp mẫu (lọc và chuyển đổi dữ liệu hay chuyển đổi sang định dạng DWH)
* Bước 4: Lưu trữ dữ liệu theo cấu trúc (sắp xếp, hợp nhất, tổng hợp)

### Các công cụ tích hợp dữ liệu:

* Công cụ có sẵn: SQL server, Oracle, MS Access, XML, …
* Công cụ dạng phần mềm chuyên dụng: MS Excel, …

### Lập các luồng công việc tích hợp dữ liệu (Data Integration Workflows)

* Tích hợp dữ liệu là quá trình kết hợp dữ liệu từ nhiều nguồn khác nhau để trích xuất giá trị bổ sung. Quá trình này thường bao gồm các hành động như tham gia, làm sạch, xác thực và làm phong phú dữ liệu trong quá trình thực hiện.
* Data Integration Workflows cho phép mọi nhóm sử dụng các ứng dụng phù hợp nhất với mục đích, bao gồm cơ sở dữ liệu, kỹ thuật, hoạt động và tài chính, trong khi vẫn giữ mọi thứ được kết nối. ... Dữ liệu được tự động xác nhận và tải vào mô hình dữ liệu của kho để nhóm GIS sử dụng
* Data Integration Workflows của chúng ta sẽ được thực hiện như sau:

+ Thu thập các yêu cầu kinh doanh

+ Xác định nhu cầu về dữ liệu và chất lượng

+ Lập hồ sơ dữ liệu hoặc hiểu các nguồn dữ liệu và chất lượng liên quan cả trong hệ thống nguồn và trên nhiều hệ thống nguồn, nếu có

+ Thực hiện đánh giá chất lượng dữ liệu dựa trên các chỉ số mà doanh nghiệp đã yêu cầu

+ Xác định khoảng cách giữa dữ liệu có sẵn và chất lượng của dữ liệu đó so với những gì doanh nghiệp đã yêu cầu

+ Sửa đổi kỳ vọng kinh doanh hoặc chi phí dự án và xác định giải pháp dữ liệu đã chọn

+ Mô hình hóa các kho dữ liệu cần thiết - các khu vực tổ chức, kho dữ liệu, kho dữ liệu hoạt động và (các) kho dữ liệu - cả từ góc độ logic (để xác nhận các yêu cầu kinh doanh) và góc độ vật lý (để cho phép thực hiện).

+ Xác định dữ liệu và nguồn phù hợp từ các hoạt động trên.

+ Chuẩn bị dữ liệu

+ Nhượng quyền dữ liệu

+ Quản lý dữ liệu và metadata

### Hệ thống OLTP (On-Line Transaction Processing = Xử lý giao dịch trực tuyến)

OLTP viết tắt từ On-line transactional processing có nghĩa là xử lý giao dịch trực tuyến.

+ Dữ liệu phát sinh từ các hoạt động hằng ngày

+ Thu thâp xử lý để phục vụ công việc nghiệp vụ cụ thể của một tổ chức.

+ Thường được gọi là dữ liệu tác vụ và hoạt động thu thập xử lý dữ liệu.

## **Vấn tin khai thác sử dụng Kho dữ liệu**

### Vấn tin với các phép toán tổng hợp của SQL trên kho dữ liệu

Mệnh đề GROUP BY

+ Các mở rộng của các hệ quản trị cơ sở dữ liệu để hỗ trợ cho kho dữ liệu

+ Mệnh đề SQL CUBE

+ Mệnh đề SQL ROLLUP

+ Mệnh đề SQL GROUPING SETS

+ Kết hợp các phép toán SUBTOTAL

### Vấn tin với các hàm phân tích của SQL trên kho dữ liệu

SQL Analytic

+ Mô hình xử lí và cú pháp cơ bản

+ Cú pháp mở rộng và các hàm xếp hạng

+ So sánh cửa sổ (Window Comparisons)

+ Các hàm so sánh tỷ lệ

### Vấn tin dùng Materialized View để phân tích trên kho dữ liệu

* Phân biệt: Traditional View và Materialized View

|  |  |
| --- | --- |
| Vấn tin truyền thống | Vấn tin phân tích |
| * Chủ yếu là hỏi đáp, sử dụng các câu lệnh như: Select..from…where * Vấn tin trên cơ sở dữ liệu | * Là vấn tin trên kho dữ liệu, phân tích, rút ra kết luận nào đó từ kho dữ liệu, từ DWH: SQL Subtotal, SQL Analytic, Materialized View * Phân tích để xác định các qui luật có sẵn trong DWH, chứ không đơn giản chỉ là hỏi đáp thông thường |
| * Ví dụ: Lấy danh sách các mặt hàng thuộc thương hiệu Xmen có trong CSDL của một cữa hàng tiện lợi ( các loại hỏi đáp) | * Ví dụ: từ DWH thông tin khách hàng mua hàng của một tiệm bán điện thoại dùng vấn tin phân tích cho biết qui luật mua hàng của khách hàng * Nếu khách hàng đã mua điện thoại A thì có xu hướng sẽ mua kèm theo loại linh kiện nào và xác suất là bao nhiêu % |

* Viết truy vấn dùng Materialized View

+ Nhắc lại khái niệm khung nhìn (views)

+ Định nghĩa materialized view và cách thức nó được xử lí

+ Các quy tắc khi viết lại truy vấn (query rewrite) dùng materialized view

### Hệ thống OLAP

* Hệ thống OLAP (On-Line Analytical Processing – xử lý phân tích trực tuyến) là kỹ thuật sử dụng các cube (khối – thể hiện đa chiều) nhằm cung cấp khả năng truy xuất nhanh đến dữ liệu của kho dữ liệu. Tạo khối(cube) cho dữ liệu trong các bảng chiều (dimension table) và bảng sự kiện (fact table) và cung cấp khả năng thực hiện các truy vấn tinh vi và phân tích cho các ứng dụng client.
* OLAP là kỹ thuật cho phép ứng dụng client truy xuất hiệu quả dữ liệu này
* OLAP cung cấp nhiều lợi ích cho người phân tích như sau:

+ Cung cấp mô hình dữ liệu đa chiều trực quan cho phép dễ dàng lựa chọn, định hướng và khám phá dữ liệu.

+ Cung cấp một ngôn ngữ truy vấn phân tích, cung cấp sức mạnh để khám phá các mối quan hệ trong dữ liệu kinh doanh phức tạp.

+ Dữ liệu được tính toán trước đối với các truy vấn thường xuyên nhằm làm cho thời gian trả kết quả của các truy vấn đặc biệt trở nên nhanh chóng hơn.

+ Cung cấp các công cụ mạnh mẽ giúp người dùng tạo các khung nhìn mới của dữ liệu dựa trên một tập các hàm tính toán đặc biệt.

* OLAP được đặt ra để xử lý các truy vấn liên quan đến lượng dữ liệu rất lớn mà nếu cho thực thi các truy vấn này trong hệ thống OLTP sẽ không thể cho kết quả hoặc sẽ mất rất nhiều thời gian.

# TỔNG QUAN VỀ KHO DỮ LIỆU PHÁT TRIỂN MÔ HÌNH KHO DỮ LIỆU VỀ DỊCH VỤ THƯ VIỆN

## **Ý tưởng hình thành DWH Phát triển mô hình kho dữ liệu về dịch vụ thư viện.**

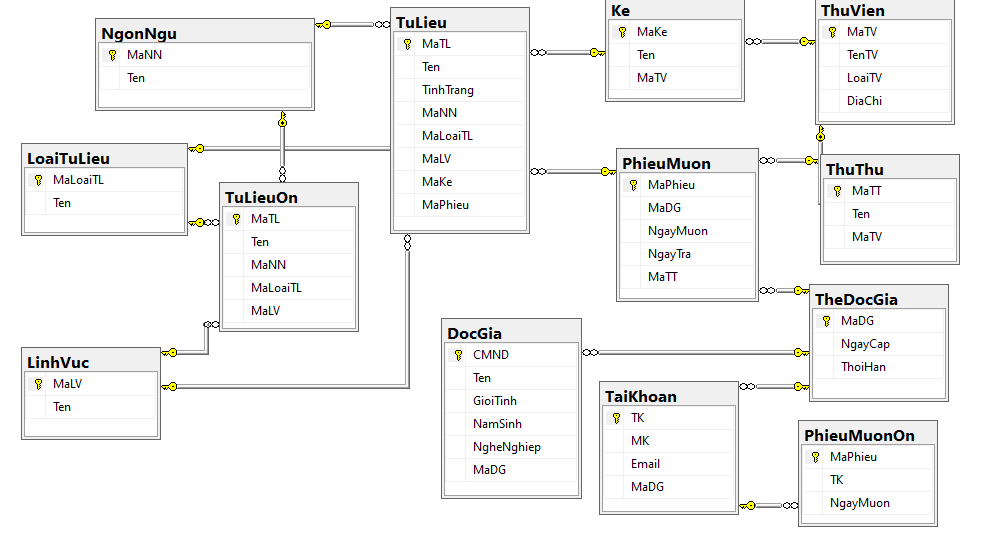
Các thư viện muốn quản lý các thông số thống kê trong quá trình cho thuê mượn sách như là: tổng số lượt mượn trả sách … Nhưng việc mỗi lần tính toán từ rất nhiều bảng thì hết sức phức tạp và dữ liệu ngày càng nhiều nên hiệu quả tính toán không được cao. Và họ muốn một mô hình dữ liệu có thể xử lý hiệu quả trên dữ liệu lớn hơn đồng thời dễ dàng hơn trong việc thống kê. Vì vậy mô hình kho dữ liệu là vô cùng thích hợp cho trường hợp này.

## **Giới thiệu chung về Kho dữ liệu**

### Mô tả tổng quan về DWH

* DWH này tập trung thống kê các số liệu liên quan đến việc đưa ra chiến lược kinh doanh trong thời gian tương lai như là: tổng số lượt mượn trả sách ….
* Các Dim trong DWH này: lĩnh vực(LinhVuc), ngôn ngữ(NgonNgu), loại tư liệu(LoaiTuLieu), độc giả(DocGia), phiếu mượn(PhieuMuon).
* Các Fact trong DWH này: tổng số lượt mượn (ThongKe)

### Các dữ liệu nguồn hình thành DWH



* Bảng NgonNgu: gồm mã ngôn ngữ và tên ngôn ngữ.
* Bảng loaiTuLieu: gồm mã loại tư liệu và tên tư liệu.
* Bảng LinhVuc: gồm mã lĩnh vực và tên lĩnh vực.
* Bảng TuLieuOn: gồm mã tư liệu, tên tư liệu, mã ngôn ngữ, mã loại tư liệu và mã lĩnh vực.
* Bảng TuLieu: gồm mã tư liệu, tên tư liệu, tình trạng, mã ngôn ngữ, mã loại tư liệu và mã lĩnh vực, mã kệ và mã phiếu mượn.
* Bảng Ke: gồm mã kệ, tên kệ và mã thư viện.
* Bảng ThuVien: gồm mã thư viện, tên thư viện, loại thư viện và địa chỉ.
* Bảng ThuThu: gồm mã thủ thư, tên thủ thư và mã thư viện.
* Bảng PhieuMuon: gồm mã phiếu, mã độc giả, ngày mượn, ngày trả, mã thủ thư.
* Bảng DocGia: gồm CMND, tên, giới tính, năm sinh, nghề nghiệp và mã độc giả.
* Bảng TheDocGia: gồm mã độc giả, ngày cấp và thời hạn.
* Bảng TaiKhoan: gồm tài khoản, mật khẩu, email và mã độc giả.
* Bảng PhieuMuonOn: mã phiếu, tài khoản và ngày mượn.

## **Xác định nhu cần tổ chức và phân tích dữ liệu của toàn bộ DWH**

* Tổ chức và phân tích dữ liệu của toàn bộ DWH là một việc làm hết sức cần thiết và quan trọng. Nó giúp chúng ta thấy một bức tranh tổng quan của các số liệu thống kê, nó giúp cho người quản lý dễ dàng sử dụng để có thể thấy được những số liệu thống kê mà mình muốn đồng thời dễ dàng cho việc trực quan hóa. Với vấn đề dữ liệu ngày càng nhiều thì việc tổ chức kho dữ liệu càng sớm càng tốt, nếu để dữ liệu quá nhiều thì sẽ rất tốn kém trong quá trình tích hợp và chuyển đổi dữ liệu.

## **Các thành phần thành phần Kho dữ liệu phân công các cá nhân phụ trách**

17\_Trần Thái Nguyên \_Thành phần khuynh hướng đọc

11\_Phạm Nguyễn Quang Lộc\_Thành phần hình thức đọc

18\_Nguyễn Hữu Thiện Nhân\_Thành phần lứa tuổi bạn đọc

9\_Lê Nguyễn Thế Linh\_Thành phần đối tượng đọc

10\_Ngô Phi Lít\_Thành phần ngôn ngữ

22\_Nguyễn Anh Phú\_Thành phần thể loại tư liệu

G1\_ Ngô Quốc Trung\_Thành phần giới tính bạn đọc

01\_Nguyễn Tấn Cường \_Thành phần khu vực bạn đọc

23\_Nguyễn Hoàn Thai\_Thành phần về loại thư viện

# PHÂN TÍCH, THIẾT KẾ VÀ KHAI THÁC SỬ DỤNG CÁC THÀNH PHẦN KHO DỮ LIỆU PHÁT TRIỂN MÔ HÌNH KHO DỮ LIỆU VỀ DỊCH VỤ THƯ VIỆN.

## **10\_ Ngô Phi Lít\_THÀNH PHẦN NGÔN NGỮ**

### Mô tả tổng quan về thành phần DWH Ngôn ngữ.

Thành phần DWH ngôn ngữ chủ yếu xây dựng và phát triển các Dim phân tích nhà kho dữ liệu về tình hình mượn sách theo ngôn ngữ của độc giả tại các thư viện trên phạm vi toàn Việt Nam.

### Xác định nguồn dữ liệu tích hợp thành phần Kho dữ liệu Ngôn ngữ.

Nguồn dữ liệu tích hợp thành phần Kho dữ liệu Ngôn ngữ là dữ liệu phát sinh theo thời gian khi giao dịch mượn trả tư liệu đối với các trường hợp trực tiếp tại thư viện hoặc tra cứu với các trường hợp trực tuyến.

### Thiết lập CSDL nguồn của thành phần của Kho dữ liệu Ngôn ngữ.



### Mô tả các Dim bằng tiêu chí phân tích (đánh giá) của thành phần của Kho dữ liệu Ngôn ngữ.

**Dim:** tình hình lựa chọn tư liệu theo thể loại ngôn ngữ của độc giả trên địa bàn.

+ Các số liệu mẫu của Dim: thống kê mượn trả sách của các thư viện khoa học tổng hợp của các tỉnh/thành phố Việt Nam.

+ Nơi khai thác sử dụng: thư viện các cấp.

+ Mục đích sử dụng: giúp các thư viện có thể phát triển các thể loại ngôn ngữ và triển khai các loại dịch vụ thư viện phù hợp với mỗi địa bàn.

+ Các **Dim** gồm:

**[1]** Số lượng độc giả mượn tư liệu theo từng ngôn ngữ và tính biến động(tăng hay giảm so với cùng kỳ trước: tháng, quý, năm).

### Mô tả các Facts bằng số liệu phân tích của thành phần của Kho dữ liệu Ngôn ngữ.

**[1]** Số lượng độc giả mượn tư liệu theo từng ngôn ngữ và tính biến động(tăng hay giảm so với cùng kỳ trước: tháng, quý, năm).

=> Dùng hàm COUNT để tính toán dựa trên các fields gốc của CSDL nguồn.

### Thiết lập các Views tính toán cần thiết để phân tích của thành phần của Kho dữ liệu Ngôn ngữ.

**View\_COUNTLuotMuon: Tổng lượt tư liệu của độc giả theo từng thể loại ngôn ngữ và theo năm.**

SELECT dbo.NgonNgu.Ten AS [Tên Ngôn Ngữ], YEAR(dbo.PhieuMuon.NgayMuon) AS Năm, COUNT(dbo.TuLieu.MaTL) AS [Tổng lượt mượn tư liệu]

FROM dbo.TuLieu INNER JOIN

dbo.NgonNgu ON dbo.TuLieu.MaNN = dbo.NgonNgu.MaNN INNER JOIN

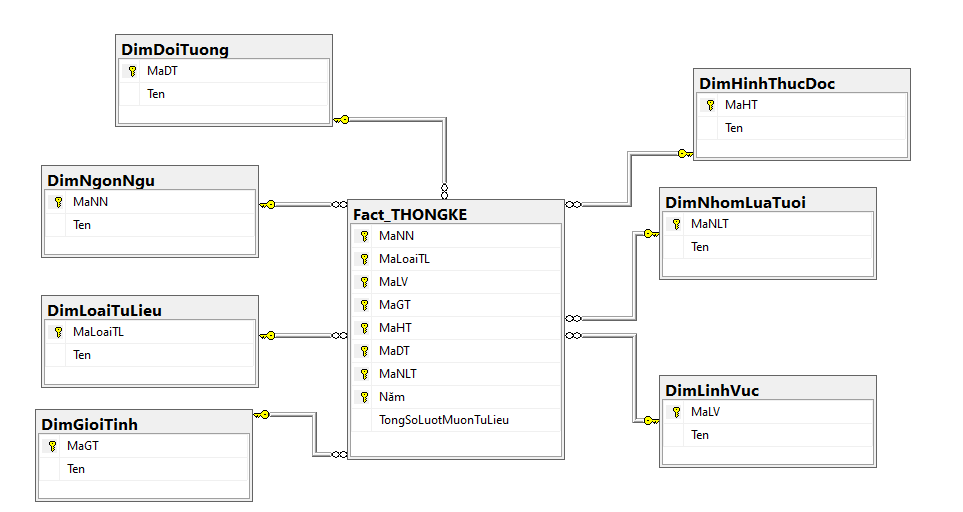
dbo.PhieuMuon ON dbo.TuLieu.MaPhieu = dbo.PhieuMuon.MaPhieu

GROUP BY CUBE (dbo.NgonNgu.Ten), YEAR(dbo.PhieuMuon.NgayMuon)

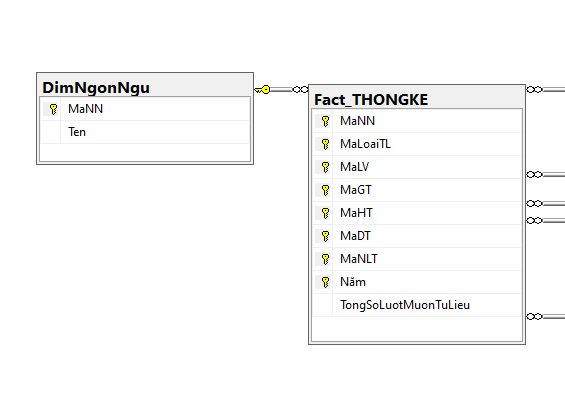


### Thiết kế DB mới tổ chức phân tích Kho dữ liệu cho thành phần thể loại ngôn ngữ.

DWH chung.

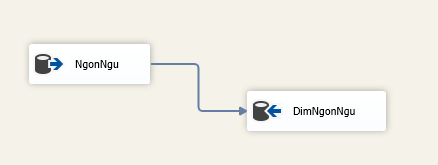


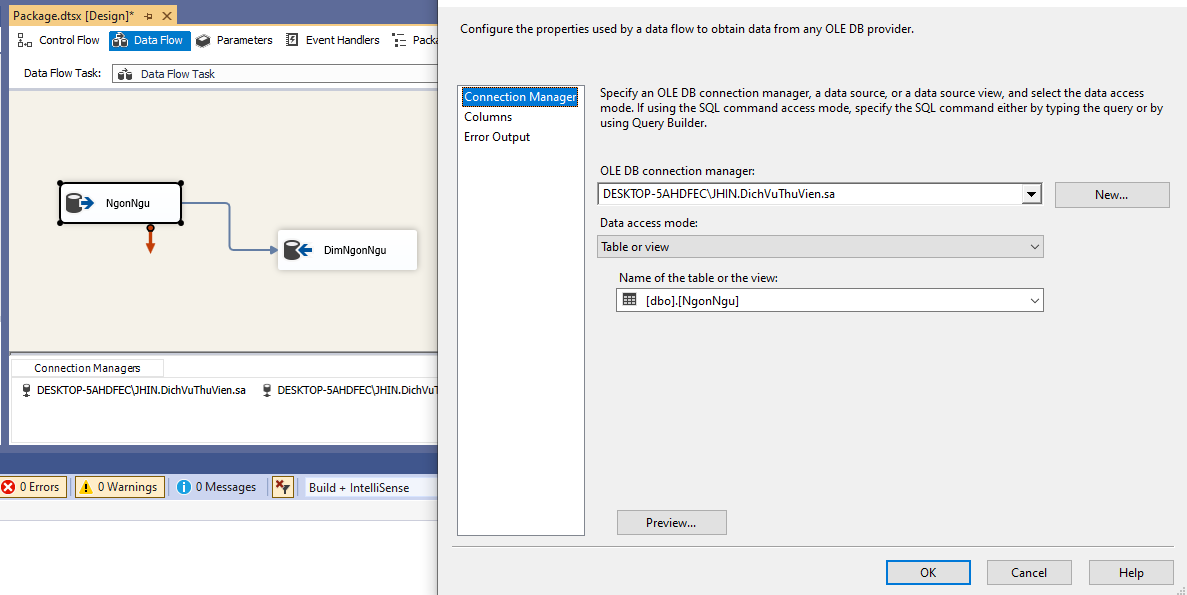
DWH phân hệ theo thể loại ngôn ngữ.

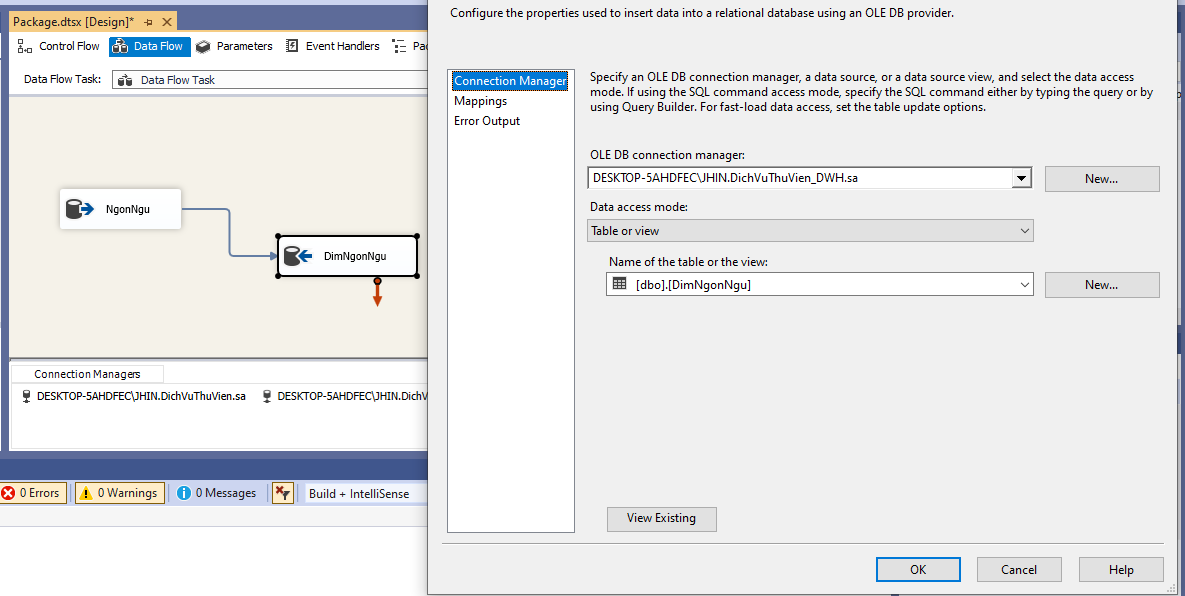
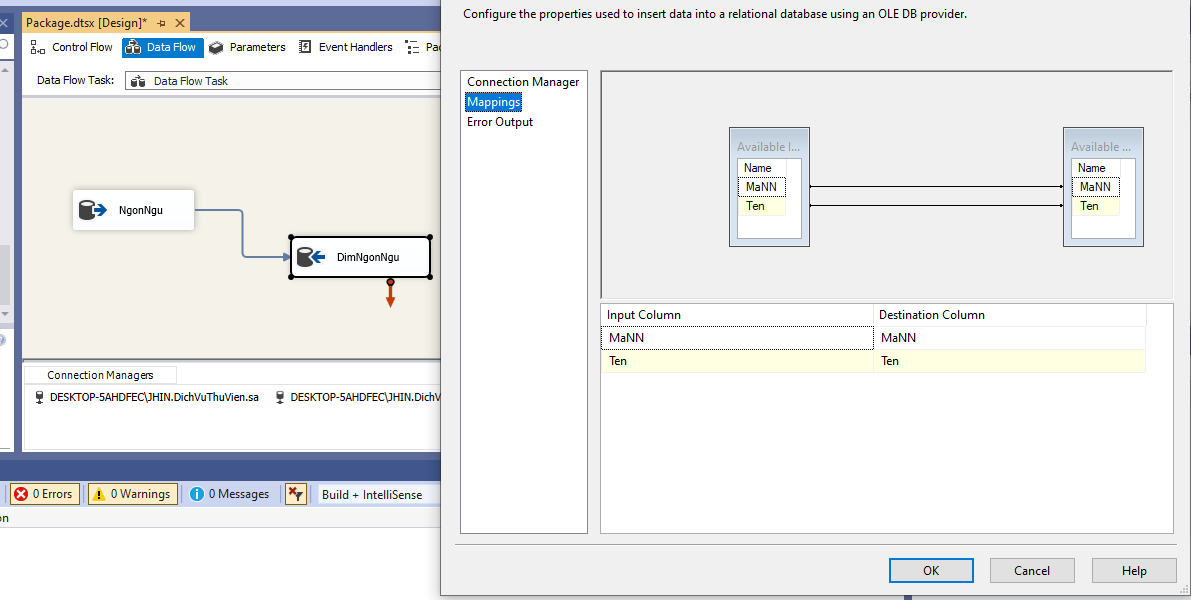


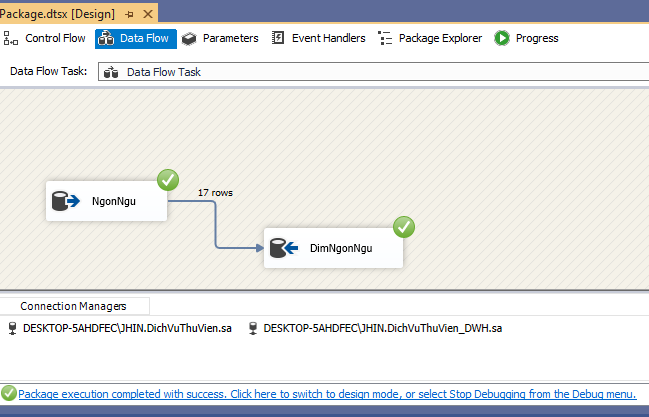
### Thực hiện các thủ tục tích hợp các DB gốc vào thành phần DWH.

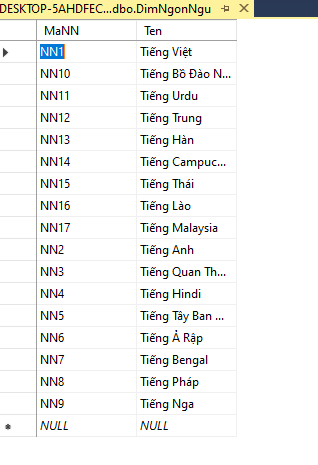
* Thực hiện đổ dữ liệu vào DimNgonNgu:









### Thiết lập các Views tính toán cần thiết để nạp dữ liệu từ DB gốc vào các Factors của thành phần DWH

* Tạo view tính tổng số lượt mượn theo ngôn ngữ, lĩnh vực, loại tư liệu, giới tính, hình thức đọc, nhóm lứa tuổi và đối tượng để nạp vào Fact.

SELECT dbo.TuLieu.MaNN, dbo.TuLieu.MaLoaiTL, dbo.TuLieu.MaLV, dbo.DocGia.GioiTinh AS MaGT, CAST(View\_ChuanbiMaHT\_1.MaHT AS nvarchar(20)) AS MaHT, dbo.DocGia.NgheNghiep AS MaDT,

CAST(dbo.View\_LayLuaTuoi.LuaTuoi AS nvarchar(20)) AS MaNLT, YEAR(dbo.PhieuMuon.NgayMuon) AS Năm, COUNT(dbo.TuLieu.MaTL) AS TongSoLuotMuonTuLieu

FROM dbo.PhieuMuon INNER JOIN

dbo.TuLieu ON dbo.PhieuMuon.MaPhieu = dbo.TuLieu.MaPhieu INNER JOIN

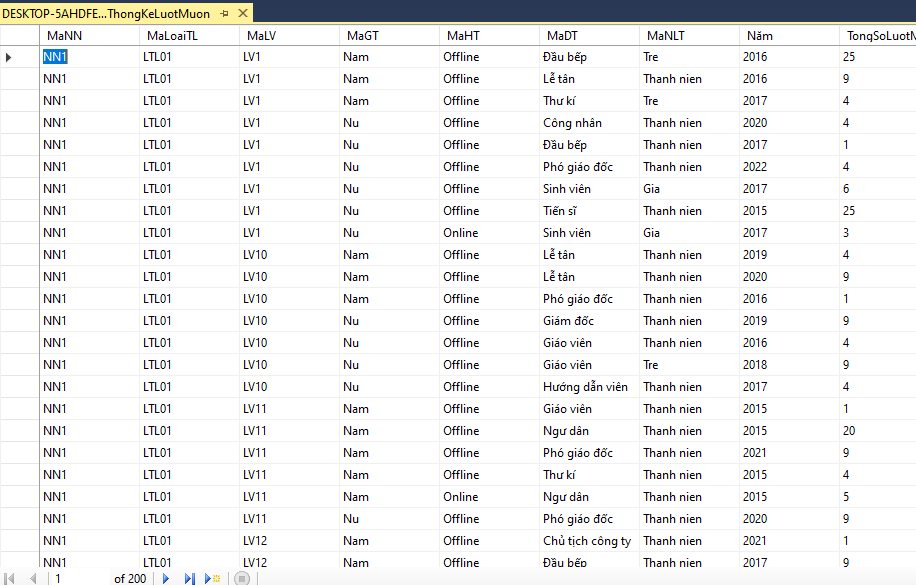
dbo.View\_ChuanbiMaHT ON dbo.PhieuMuon.MaPhieu = dbo.View\_ChuanbiMaHT.MaPhieu INNER JOIN

dbo.View\_LayLuaTuoi ON dbo.PhieuMuon.MaDG = dbo.View\_LayLuaTuoi.MaDG INNER JOIN

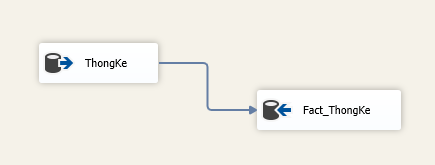
dbo.DocGia ON dbo.PhieuMuon.MaDG = dbo.DocGia.MaDG INNER JOIN

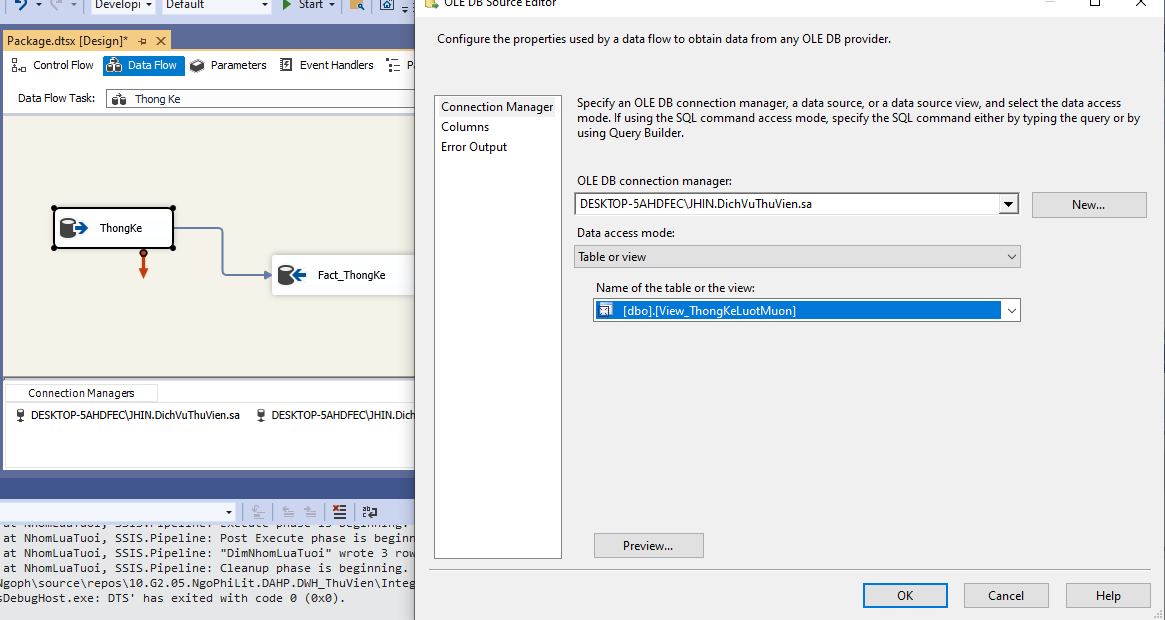
dbo.View\_ChuanbiMaHT AS View\_ChuanbiMaHT\_1 ON dbo.PhieuMuon.MaPhieu = View\_ChuanbiMaHT\_1.MaPhieu

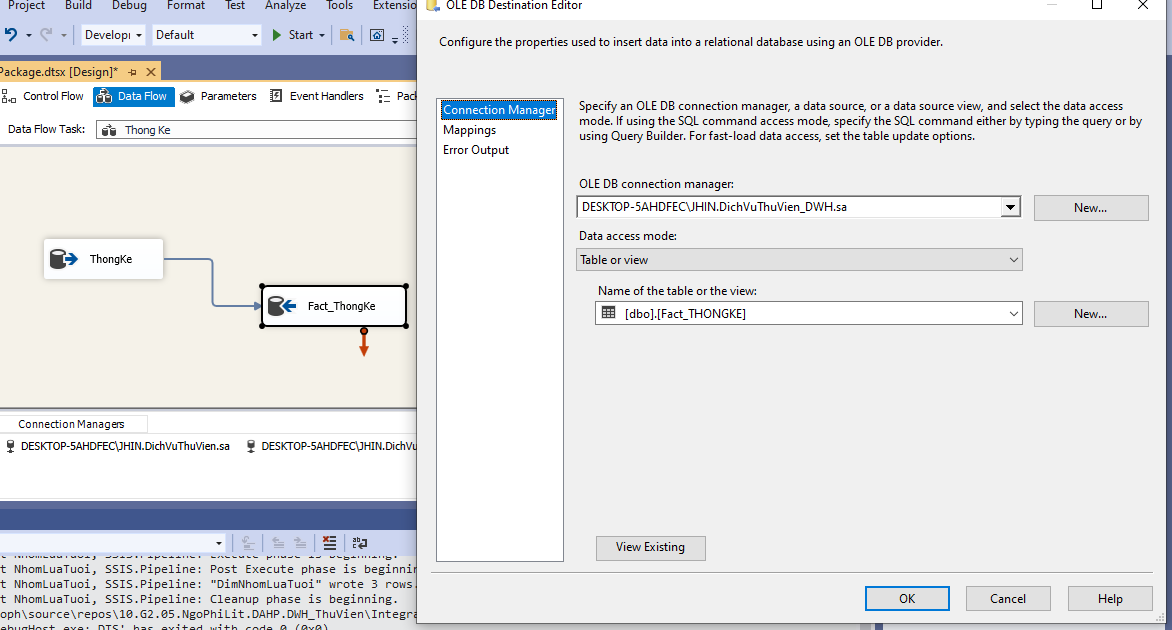
GROUP BY dbo.TuLieu.MaNN, dbo.TuLieu.MaLoaiTL, dbo.TuLieu.MaLV, dbo.DocGia.GioiTinh, View\_ChuanbiMaHT\_1.MaHT, dbo.DocGia.NgheNghiep, dbo.View\_LayLuaTuoi.LuaTuoi, YEAR(dbo.PhieuMuon.NgayMuon)

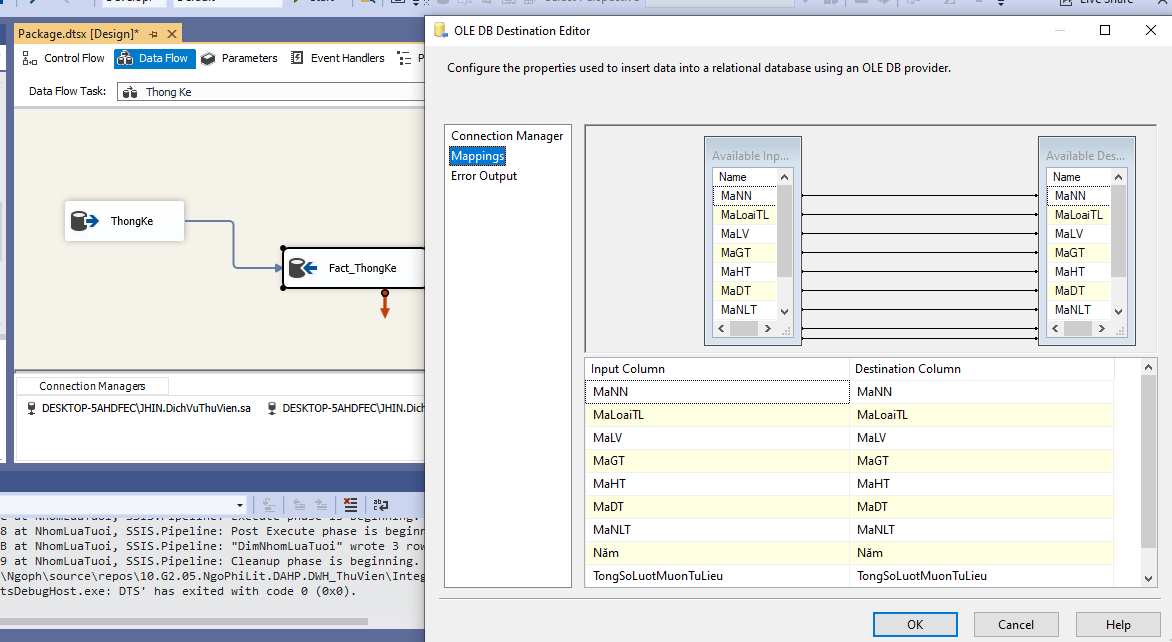


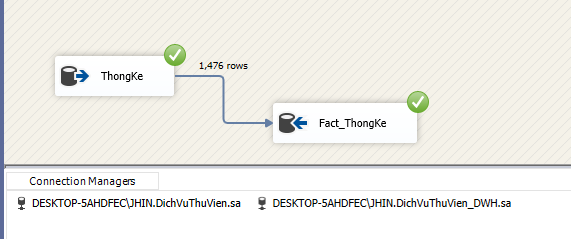
* Dùng SSIS đổ dữ liệu vào Fact.







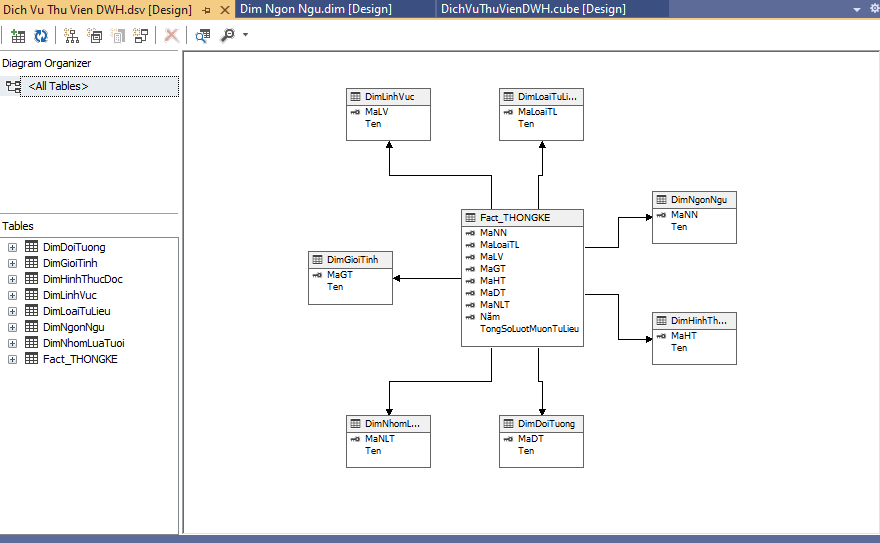




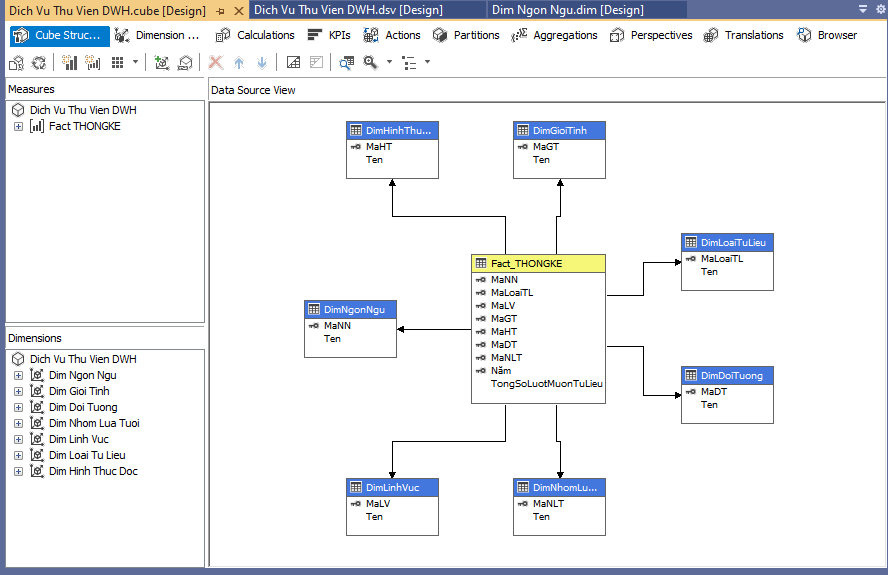


### Triển khai thành phần DWH lên BIDS

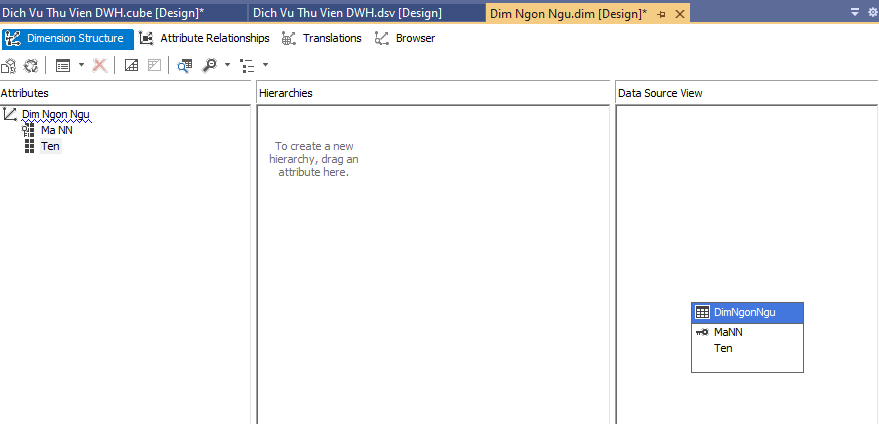
* Provider=SQLNCLI11.1;Data Source=DESKTOP-5AHDFEC\JHIN;Integrated Security=SSPI;Initial Catalog=DichVuThuVien\_DWH\
* Tạo data source view mới.



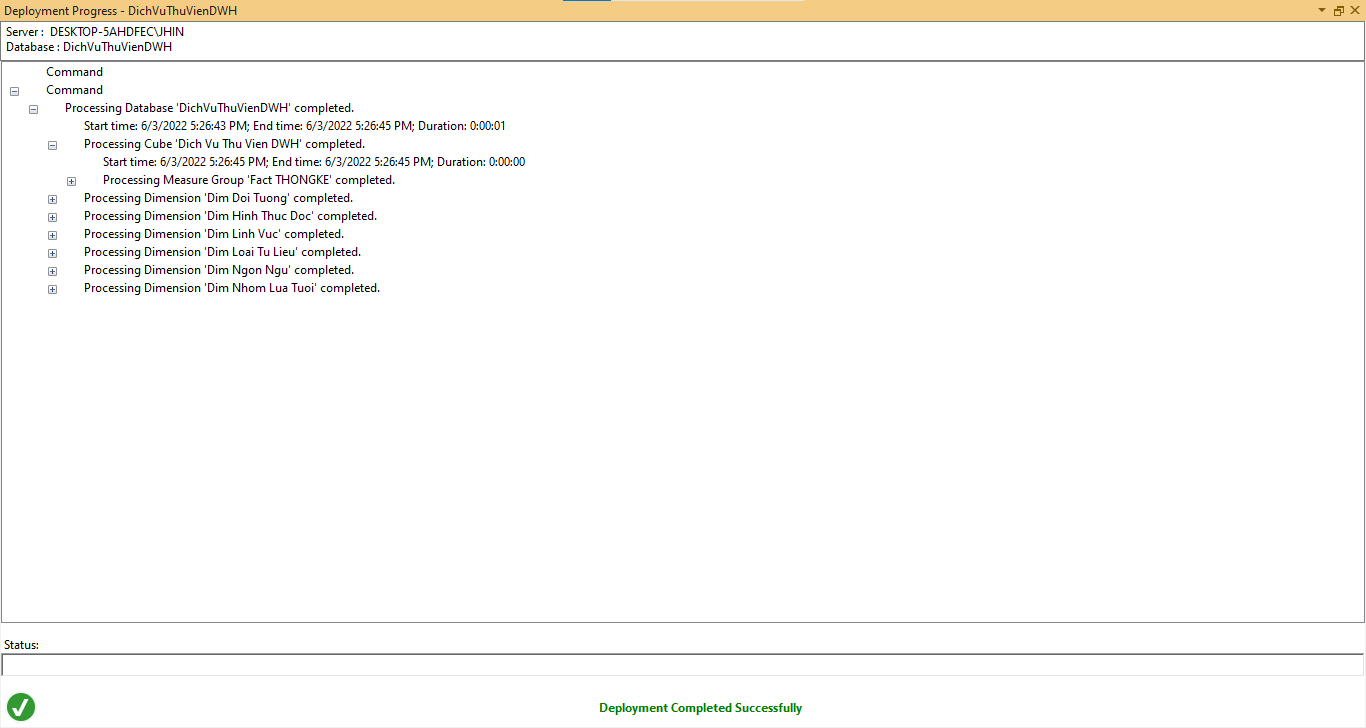
* Thiết lập Cube mới và tạo các Dim trên dữ liệu có sẵn trên bảng

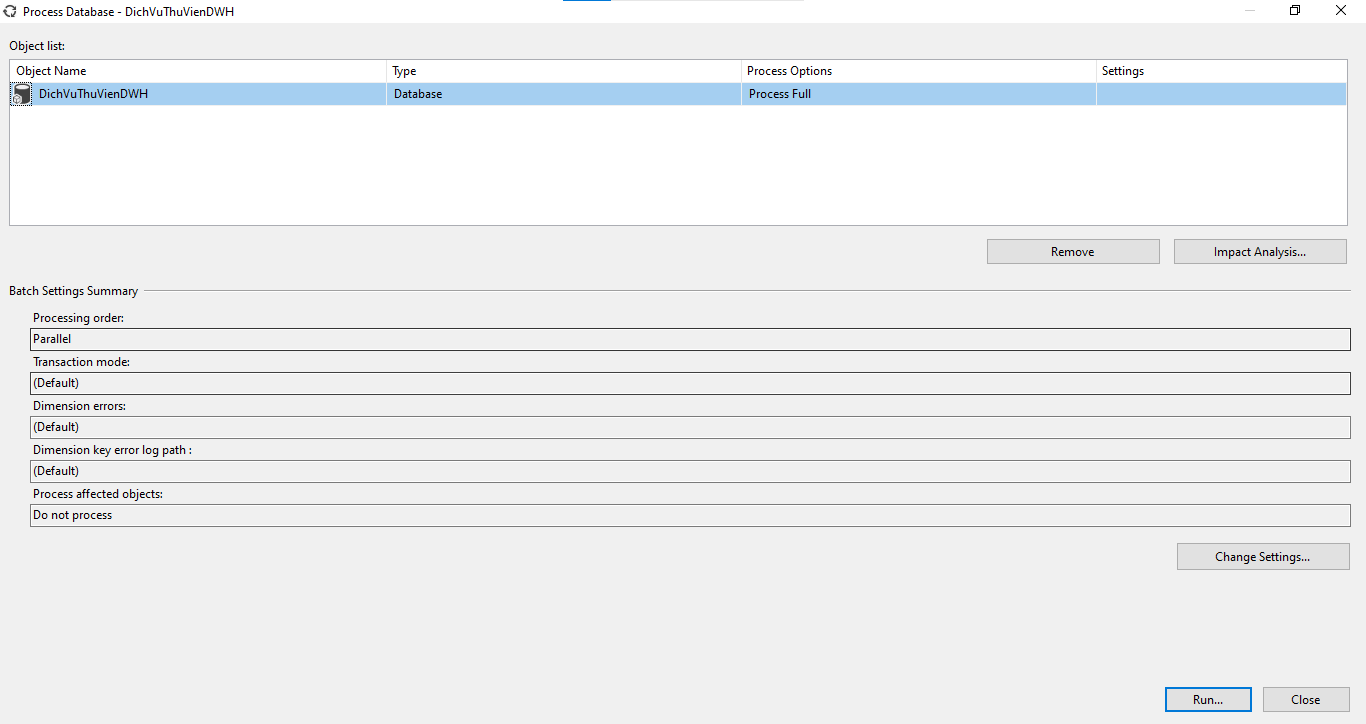


* Chọn ra các thuộc tính từ bảng dữ liệu để sử dụng làm Dim hoặc làm giá trị Measure
* DimNgonNgu

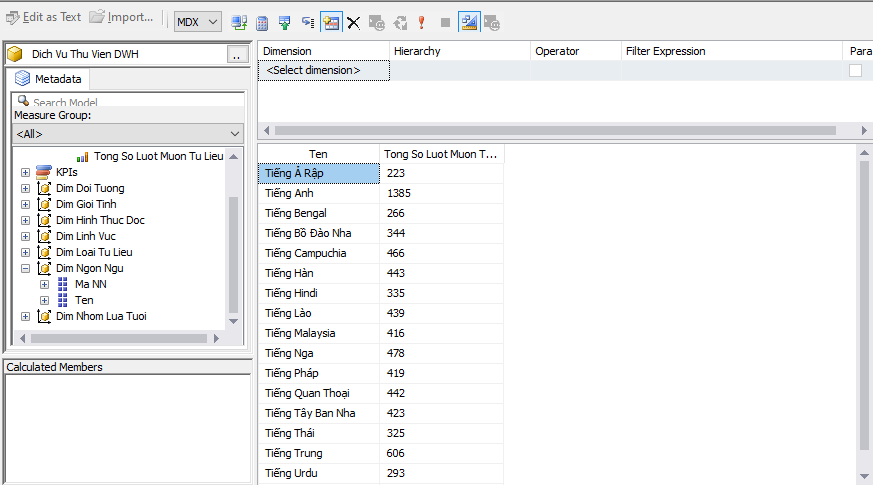


* Triển khai cube.





* Theo dõi và phân tích:



Như ta thấy thì số lượng sách tiếng Việt vẫn chiếm ưu thế rõ rệt. Bên cạnh đó thì sách tiếng Anh cũng được mượn khá nhiều. Các tiếng còn lại thì chiếm số ít nhưng vẫn có độc giả mượn.

### Khai thác sử dụng thành phần DWH dùng các vấn tin SQL

SELECT dbo.NgonNgu.Ten, YEAR(dbo.PhieuMuon.NgayMuon) AS Năm, COUNT(dbo.TuLieu.MaTL) AS SoLuotMuonTuLieu

FROM dbo.TuLieu INNER JOIN

dbo.NgonNgu ON dbo.TuLieu.MaNN = dbo.NgonNgu.MaNN INNER JOIN

dbo.PhieuMuon ON dbo.TuLieu.MaPhieu = dbo.PhieuMuon.MaPhieu

GROUP BY CUBE (dbo.NgonNgu.Ten), YEAR(dbo.PhieuMuon.NgayMuon)

**View\_THONGKE\_NGONNGU:** Thống kê sô lượt mượn sách theo từng thể loại ngôn ngữ qua từng ngày, đồng thời gom nhóm từng năm để biết tổng số lượt mượn trong năm đó.

****

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN CỦA ĐỒ ÁN HỌC PHẦN

## **Kết luận**

### Những kết quả đạt được

* Đưa ra được tổng số lượt mượn theo ngôn ngữ.
* Phân tích tổng số lượt mượn của từng ngôn ngữ (cao thấp).
* Phân tích lý do tại sao tổng số lượt mượn của ngôn ngữ đó lại cao/thấp.
* Đưa ra hướng khắc phục chung cho các loại ngôn ngữ.

### Hạn chế

* Chỉ đưa ra được hướng khắc phục chung mà chưa đưa ra được hướng khắc phục cho từng loại ngôn ngữ.

## **Hướng phát triển**

### Hướng khắc phục các hạn chế

* Khi nắm được lý do tại sao tổng số lượt mượn của từng loại ngôn ngữ cao/thấp thì đưa ra hướng khắc phục cho từng loại ngôn ngữ.

### Hướng mở rộng ĐỒ ÁN HỌC PHẦN

* Đưa các số liệu của đồ án lên web để thấy và phân tích rõ hơn (có thể vẽ chart hay nhiều cách khác nhau).

# DANH MỤC TÀI LIỆU THAM KHẢO

1. Võ Xuân Thể (2021), *Bài giảng học phần Kho dữ liệu*, TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM (lưu hành nội bộ).
2. Website: <https://en.wikipedia.org/wiki/Data_warehouse>
3. Website: <https://tedu.com.vn/co-so-du-lieu/su-khac-nhau-giua-he-thong-olap-va-oltp-147.html>
4. Website: <https://ierp.vn/data-warehouse-la-gi/>
5. Website: <https://www.investopedia.com/terms/d/data-warehousing.asp>
6. Website: <https://vietnambiz.vn/kho-du-lieu-data-warehousing-la-gi-phan-biet-kho-du-lieu-va-co-so-du-lieu-20200416135805599.htm>

# ĐÁNH GIÁ KẾT QUẢ HỌC TẬP HỌC PHẦN

ĐỒ ÁN HP THAY CHO BÀI THI KẾT THÚC HỌC PHẦN

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Hình thức KT** | **Nội dung** | **Chuẩn đầu ra đánh giá** | **Trình độ năng lực** | **Phương pháp đánh giá** | **Công cụ đánh giá** | **Tỉ lệ (%)** |
| **BÀI TẬP LẬP TRÌNH** | | | | | |  |
| BL#1 | Bài tập lập trình theo từng Buổi thực hành = tính bình quân  Kiểm tra viết | G1.1 | 3 | Bài tập  Kiểm tra | Bài tập  Kiểm tra | 50 |
| G2.1 | 3 |
| G3.1  G3.2 | 4  3 |
| G4.1 | 5 |
| **ĐỒ ÁN HỌC PHẦN (Project)** | | | | | |  |
| ĐA#1 | Phát triển một hệ thống KHO DỮ LIỆU đơn giản:  + Đề tài Nhóm có phân công thành phần cụ thể cho từng thành viên thuộc nhóm  + Xây dựng và thực nghiệm Hệ thống Kho dữ liệu theo nhóm và thành phần cá nhân  => báo cáo vào ngày thi. | G1.1 | 3 | Báo cáo, Demo và thuyết trình | Rubric | 50 |
| G2.1 | 3 |
| G3.1 | 4 |
| G3.2  G4.1 | 3  5 |
| G2.1 | 4 |
| G3.1 | 4 |
| G3.2  G4.1 | 3  5 |