**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\_06**

Tên đề tài Nhóm: **Phát triển mô hình kho dữ liệu tuyển sinh tại các trường Đại học Việt Nam khai thác sử dụng tại các cơ sở đào tạo Đại học.**

Lớp: **181330B**

**NHÓM: 9**

DANH SÁCH THÀNH VIÊN NHÓM

{bản mềm: nhóm trưởng để nguyên ds sau = bản in; bản mềm các thành viên: chỉ để tên +… của mình, Xóa các thành viên khác}

1. Phan Thành Đạt [04] \_ Nhóm trưởng
2. Nguyễn Anh Triều [47]
3. Lê Chí Hiếu [09]
4. Đỗ Đình Phùng [33]

**Ngày nộp: 15/06/2021**

**TP.HCM, ngày 15 Tháng 06 năm 2021**

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

**………………………………………………………………..**

**………………………………………………………………..**

**……………………………………………………………….**

**NHÓM : 9**

**ĐIỂM ĐÁNH GIÁ**

{bản mềm: nhóm trưởng để nguyên ds sau; bản mềm các thành viên: chỉ để tên +… của mình[đúng số tt/nhóm], Xóa các thành viên khác: NẾU DS DÀI = Trang riêng: 1 lần :BỎ DS BÌA & BÌA LÓT}

1. Phan Thành Đạt [04] **:………….(…………………)………..**
2. Nguyễn Anh Triều [47] **:………….(…………………)………….**
3. Lê Chí Hiếu [09] **:………….(…………………)………….**
4. Đỗ Đình Phùng [33] **:………….(…………………)………….**

**Ngày 15/06/2021**

**Giảng viên Ký tên**

THÀNH VIÊN:

{bản mềm: nhóm trưởng để nguyên ds sau; bản mềm các thành viên: chỉ để tên +… của mình[đúng số tt/nhóm], Xóa các thành viên khác}

1. Phan Thành Đạt [04] \_ Nhóm trưởng: Thành phần kho dữ liệu ngành học - khu vực.
2. Nguyễn Anh Triều [47]: Thành phần kho dữ liệu chất lượng đoàn tuyển sinh
3. Lê Chí Hiếu [09]: Thành phần kho dữ liệu chất lượng sinh viên - khu vực.
4. Đỗ Đình Phùng [33]: Thành phần kho dữ liệu ngành học - thời gian.

**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\_06**

Tên đề tài Nhóm: **Phát triển mô hình kho dữ liệu tuyển sinh tại các trường Đại học Việt Nam khai thác sử dụng tại các cơ sở đào tạo Đại học.**

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

Lớp: **181330B (Sáng thứ 4)**

**Tên sản phẩm đề tài: 04.N9.01.PTDat.DAHP.DWH.rar** (Microsoft SQL Server 2019)

**DB Nguồn:**  **04N9PTDatTuyenSinh\_Source**

**DB DWH:**  **04N9PTDatTuyenSinh\_DWH**

**Tên máy chủ DB SQL: localhost**

**Tên đăng nhập DB SQL: sa**

**Mật khẩu (password) đăng nhập SQL: 123**

**ProjectName BIDS: sttNxHoTenTuyenSinh\_DWH VD: 05N1VXTheQLKCBBHYT\_DWH**

**NHÓM : 9**

THÀNH VIÊN: 1. Phan Thành Đạt [09]\_NT; 2. Nguyễn Anh Triều [47]; 3. Lê Chí Hiếu [09]; Đỗ Đình Phùng [33].

. . . .

{bản mềm: nhóm trưởng để nguyên ds sau; bản mềm các thành viên: chỉ để tên +… của mình[đúng số tt/nhóm], Xóa các thành viên khác}

**TP.HCM, ngày 15 Tháng 06 năm 2021**

# LỜI CẢM ƠN

Để hoàn thành tốt đề tài và bài báo cáo này, chúng em xin gửi lời cảm ơn chân thành đến giảng viên Võ Xuân Thể, người đã trực tiếp hỗ trợ chúng em trong suốt quá trình làm đề tài. Chúng em cảm thầy đã đưa ra những lời khuyên từ kinh nghiệm thực tiễn của mình để định hướng cho chúng em đi đúng với yêu cầu của đề tài đã chọn, luôn giải đáp thắc mắc và đưa ra những góp ý, chỉnh sửa kịp thời giúp chúng em khắc phục nhược điểm và hoàn thành tốt cũng như đúng thời hạn Khoa đã đề ra. Chúng em cũng xin gửi lời cảm ơn chân thành các quý thầy cô trong khoa Công Nghệ Thông Tin đã tận tình truyền đạt những kiến thức cần thiết giúp chúng em có nền tảng để làm nên đề tài này, đã tạo điều kiện để chúng em có thể tìm hiểu và thực hiện tốt đề tài. Cùng với đó, chúng em xin được gửi cảm ơn đến các bạn cùng khóa đã cung cấp nhiều thông tin và kiến thức hữu ích giúp chúng em có thể hoàn thiện hơn đề tài của mình. Đề tài và bài báo cáo được chúng em thực hiện trong khoảng thời gian ngắn, với những kiến thức còn hạn chế cùng nhiều hạn chế khác về mặt kĩ thuật và kinh nghiệm trong việc thực hiện một dự án phần mềm. Do đó, trong quá trình làm nên đề tài có những thiếu sót là điều không thể tránh khỏi nên chúng em rất mong nhận được những ý kiến đóng góp quý báu của các quý thầy cô để kiến thức của chúng em được hoàn thiện hơn và chúng em có thể làm tốt hơn nữa trong những lần sau.

**TP.HCM, ngày 15 tháng 06 năm 2021**

**Nhóm sinh viên thực hiện**

# 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](#_Toc74843045)

[DANH MỤC CÁC TỪ VIẾT TẮT 5](#_Toc74843046)

[DANH MỤC CÁC THUẬT NGỮ 6](#_Toc74843047)

[MỤC LỤC 7](#_Toc74843048)

[Chương 1. GIỚI THIỆU VỀ ĐỒ ÁN HỌC PHẦN 10](#_Toc74843049)

[1.1. Tổng quan về ĐỒ ÁN HỌC PHẦN 10](#_Toc74843050)

[1.2. Nội dung chuyên môn chính của ĐỒ ÁN HỌC PHẦN 10](#_Toc74843051)

[1.3. Mục tiêu của ĐỒ ÁN HỌC PHẦN 10](#_Toc74843052)

[1.4. Công cụ và nền tảng kỹ thuật thực hiện ĐỒ ÁN HỌC PHẦN 11](#_Toc74843053)

[1.5. Sản phẩm của ĐỒ ÁN HỌC PHẦN 11](#_Toc74843054)

[1.6. Bố cục của báo cáo 11](#_Toc74843055)

[Chương 2. CƠ SỞ LÝ THUYẾT CỦA HỌC PHẦN 11](#_Toc74843056)

[2.1. Tổng quan về kho dữ liệu (DWH) 11](#_Toc74843057)

[2.1.1. Các khái niệm cơ bản về Kho dữ liệu 11](#_Toc74843058)

[2.1.2. Kiến trúc cơ bản của kho dữ liệu: 12](#_Toc74843059)

[2.1.3. Mô hình CSDL đa chiều (Dimensional Modeling) 13](#_Toc74843060)

[2.2. Phân tích và thiết kế Kho dữ liệu 13](#_Toc74843061)

[2.2.1. Phân tích và thiết kế kho dữ liệu theo hướng mô hình hóa nhiều chiều (Dimensional Modeling) 13](#_Toc74843062)

[2.2.2. Các dạng (phương pháp) tích hợp dữ liệu 13](#_Toc74843063)

[2.2.3. Các bước cơ bản trong tiến trình tích hợp dữ liệu 13](#_Toc74843064)

[2.2.4. Các công cụ tích hợp dữ liệu: SQL và phần mềm 13](#_Toc74843065)

[2.2.5. Lập các luồng công việc tích hợp dữ liệu (Data Integration Workflows) 13](#_Toc74843066)

[2.2.6. Hệ thống OLTP (On-Line Transaction Processing = Xử lý giao dịch trực tuyến) 14](#_Toc74843067)

[2.3. Vấn tin khai thác sử dụng Kho dữ liệu 14](#_Toc74843068)

[2.3.1. Vấn tin với các phép toán tổng hợp của SQL trên kho dữ liệu 14](#_Toc74843069)

[2.3.2. Vấn tin với các hàm phân tích của SQL trên kho dữ liệu 14](#_Toc74843070)

[2.3.3. Vấn tin dùng Materialized View để phân tích trên kho dữ liệu 14](#_Toc74843071)

[2.3.4. Hệ thống OLAP 15](#_Toc74843072)

[Chương 3. TỔNG QUAN VỀ KHO DỮ LIỆU PHÁT TRIỂN MÔ HÌNH KHO DỮ LIỆU TUYỂN SINH TẠI CÁC TRƯỜNG ĐẠI HỌC VIỆT NAM KHAI THÁC SỬ DỤNG TẠI CÁC CƠ SỞ ĐÀO TẠO ĐẠI HỌC 15](#_Toc74843073)

[3.1. Ý tưởng hình thành DWH Phát triển mô hình kho dữ liệu tuyển sinh tại các trường Đại học Việt Nam khai thác sử dụng tại các cơ sở đào tạo Đại học. 15](#_Toc74843074)

[3.2. Giới thiệu chung về Kho dữ liệu 15](#_Toc74843075)

[3.2.1. Mô tả tổng quan về DWH 15](#_Toc74843076)

[3.2.2. Các dữ liệu nguồn hình thành DWH 16](#_Toc74843077)

[3.3. Xác định nhu cần tổ chức và phân tích dữ liệu của toàn bộ DWH 16](#_Toc74843078)

[3.4. Các thành phần phân hệ Kho dữ liệu phân công các cá nhân phụ trách 16](#_Toc74843079)

[Chương 4. PHÂN TÍCH, THIẾT KẾ VÀ KHAI THÁC SỬ DỤNG CÁC PHÂN HỆ KHO DỮ LIỆU PHÁT TRIỂN MÔ HÌNH KHO DỮ LIỆU TUYỂN SINH TẠI CÁC TRƯỜNG ĐẠI HỌC VIỆT NAM KHAI THÁC SỬ DỤNG TẠI CÁC CƠ SỞ ĐÀO TẠO ĐẠI HỌC 17](#_Toc74843080)

[4.1. 04\_Phan Thành Đạt\_ Thành phần kho dữ liệu ngành học - khu vực 17](#_Toc74843081)

[4.2. 09\_Lê Chí Hiếu\_ Thành phần kho dữ liệu chất lượng sinh viên - khu vực 17](#_Toc74843082)

[4.3. 33\_Đỗ Đình Phùng \_ Thành phần kho dữ liệu ngành học - thời gian 17](#_Toc74843083)

[4.4. 47\_Nguyễn Anh Triều \_ Thành Phần đánh giá chất lượng của đoàn tuyển sinh 17](#_Toc74843084)

[4.4.1. Mô tả tổng quan về phân hệ DWH đánh giá chất lượng của các đoàn tuyển sinh 17](#_Toc74843085)

[4.4.2. Xác định nguồn dữ liệu tích hợp phân hệ Kho dữ liệu chất lượng của đoàn tuyển sinh 17](#_Toc74843086)

[4.4.3. Xác định ý tưởng hình thành phân hệ của Kho dữ liệu chất lượng của đoàn tuyển sinh 17](#_Toc74843087)

[4.4.4. Mô tả hệ thống các DB gốc liên quan phân hệ DWH đánh giá chất lượng của đoàn tuyển sinh. 17](#_Toc74843088)

[4.4.5. Phần phân tích các DB gốc xác định yêu cầu phân tích DWH 17](#_Toc74843089)

[4.4.6. Thiết kế DB mới tổ chức phân tích Kho dữ liệu cho phân hệ 17](#_Toc74843090)

[4.4.7. Thiết lập các Views tính toán cần thiết để nạp dữ liệu từ DB gốc vào các Factors của phân hệ DWH 17](#_Toc74843091)

[4.4.8. Triển khai phân hệ DWH lên BIDS 17](#_Toc74843092)

[4.4.9. Khai thác sử dụng phân hệ DWH dùng các vân tin SQL 17](#_Toc74843093)

[4.4.10. Sử dụng Tool Tableau để đưa dữ liệu lên các chart 17](#_Toc74843094)

[Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN CỦA ĐỒ ÁN HỌC PHẦN 17](#_Toc74843095)

[5.1. Kết luận 17](#_Toc74843096)

[5.1.1. Những kết quả đạt được 17](#_Toc74843097)

[5.1.2. Hạn chế 18](#_Toc74843098)

[5.2. Hướng phát triển 18](#_Toc74843099)

[5.2.1. Hướng khắc phục các hạn chế 18](#_Toc74843100)

[5.2.2. Hướng mở rộng ĐỒ ÁN HỌC PHẦN 18](#_Toc74843101)

[DANH MỤC TÀI LIỆU THAM KHẢO 18](#_Toc74843102)

[CÁC PHỤ LỤC 19](#_Toc74843103)

[ĐÁNH GIÁ KẾT QUẢ HỌC TẬP HỌC PHẦN 20](#_Toc74843104)

# GIỚI THIỆU VỀ ĐỒ ÁN HỌC PHẦN

## Tổng quan về ĐỒ ÁN HỌC PHẦN

Đồ án Học phần Kho dữ liệu (DWH) Tuyển sinh được xây dựng nhằm giúp các trường Đại Học phân tích, định hướng kế hoạch cụ thể quá trình tuyển sinh.

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

+ Tổng quan về kho dữ liệu:

* Các khái niệm cơ bản về kho dữ liệu
* Kiến trúc cơ bản của kho dữ liệu
* Mô hình CSDL đa chiều (Dimensional Modeling)

+ Kỹ năng về thiết kế kho dữ liệu và sử dụng các công cụ phổ biến để lập các luồng công việc tích hợp dữ liệu (Data Integration Workflows)

+ Kỹ năng khai thác sử dụng kho dữ liệu: các chức năng mở rộng về CSDL của SQL để vấn tin phân tích trên kho dữ liệu (đặc biệt là kho dữ liệu trong kinh doanh):

* SQL Subtotal
* SQL Analytic
* Materialized View.

## Mục tiêu của ĐỒ ÁN HỌC PHẦN

+ Các khái niệm về cơ sở dữ liệu quan hệ cho dữ liệu nhiều chiều

* Khối dữ liệu (data cube)
* Các phép toán trên khối dữ liệu

+ Thiết kế kho dữ liệu theo phương pháp mô hình hóa dữ liệu nhiều chiều (dimensional modeling)

+ Bài tập tình huống

+ Đề xuất project cuối kỳ

## 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)

+ SQL Server Analysis Services (SSAS) thay cho BIDS vì phiên bản của visual studio 2019 hỗ trợ SSAS thay vì BIDS.

## Sản phẩm của ĐỒ ÁN HỌC PHẦN

Phân tích thiết kế được một kho dữ liệu thực số liệu đủ lớn để hình thành DWH. Khai thác phân tích nhà kho này.

## 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 việc thực hiện đề tài HP

Chương 3: Giới thiệu tổng quan về Kho dữ liệu (DWH) Kho dữ liệu ……………tại ………….

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

Chương 5: Tổng kết các kết quản 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)?

Data Warehouse tạm gọi kho dữ liệu là một kỹ thuật thu thập và quản lý dữ liệu từ nhiều nguồn khác nhau để cung cấp những thông tin kinh doanh có ý nghĩa. Data Warehouse là sự pha trộn của các công nghệ và các thành phần hỗ trợ trong việc sử dụng chiến lược dữ liệu.

#### Nguyên lý hình thành DWH

Nguồn dữ liệu của DWH được hình thành từ nhiều loại khác nhau:

* Dữ liệu từ các hệ thống tác nghiệp.
* Hệ thống kế thừa.
* Các nguồn dữ liệu bên ngoài

#### Các đặc điểm cơ bản của DWH

Các đặc điểm cơ bản của DWH

* Tính tích hợp.
* Gắn thời gian.
* Có tính lịch sử.
* Chỉ đọc
* Không biến động
* Dữ liệu tổng hợp và chi tiết

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

DWH về bản chất cũng là một database bình thường, các hệ quản trị cơ sở dữ liệu quản lý và lưu trữ nó như các database thông thường tuy nhiên có hỗ trợ thêm về quản lý dữ liệu lớn và truy vấn. Thực ra nét khác biệt của DWH so với database là ở quan niệm, cách nhìn vấn đề.

### Kiến trúc cơ bản của kho dữ liệu:

#### Tầng đáy: nạp dữ liệu

**Tầng đáy (Data Sources)**: là nơi thu thập, tích hợp dữ liệu từ nhiều nguồn khác nhau sau đó chuẩn hóa về cùng định dạng, làm sạch xử lí dữ liệu để tìm lỗi và sửa và lưu trữ dữ liệu đã tổng hợp.

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

**Tầng giữa (Warehouse)**: cung cấp dịch vụ xử lý phân tích trực tuyến (OLAP – Online Analytical Processing) để thao tác với kho dữ liệu đã được tích hợp. Tầng giữa có thể được cài đặt bằng Relational OLAP, Multidimensional OLAP hay kết hợp cả hai mô hình trên Hybrid OLAP

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

**Tầng trên cùng (Users)**: nơi thực hiện các thao tác truy vấn, báo cáo, phân tích để tìm ra xu hướng, trung bình, tổng hợp…

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

Bản chất đa chiều của các câu hỏi trong nghiệp vụ được phản ánh trong thực tế chẳng hạn như những người quản lí thị trường không được thoả mãn với câu hỏi theo một chiều đơn giản, thay vào đó là những câu hỏi phức tạp. Một cách để quan sát một mô hình dữ liệu nhiều chiều là nhìn nó như một hình khối. Các thành phần chính:

* Các dữ kiện (Facts)
* Các chiều (Dimensions)
* Và các khối đa chiều (Cubes)

## 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)

+ Khối dữ liệu (data cube)

+ Các phép toán trên khối dữ liệu

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

+ Hướng chủ đề (Subject Oriented)

+ Hướng tích hợp (Integrated Oriented)

+ Hướng biến đổi theo thời gian (Time Variant Oriented)

+ Hướng ổn định (Non -Volatile Oriented)

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

B1. Trích xuất dữ liệu = thu thập dữ liệu lớn từ nhiều nguồn

B2. Làm sạch = tìm lỗi và sửa sai

B3. So khớp mẫu = Lọc và chuyển đổi dữ liệu = chuyển đổi sang định dạng kho

B4. Lưu trữ dữ liệu theo cấu trúc = sắp xếp, hợp nhất và tổng hợp, ...

### Các công cụ tích hợp dữ liệu: SQL và phần mềm

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

Tích hợp dữ liệu có 2 hình thức cơ bản đó là:

* Tích hợp để đảm bảo việc trao đổi dữ liệu giữa các hệ thống được thành công (effectiveness), bàn giao hoặc xử lý các công việc tiếp theo theo các luồng công việc định trước.
* Tích hợp các dữ liệu từ các nguồn khác nhau để có cái nhìn tổng thể (gọi là Unified) cho người sử dụng/cấp độ quản lý hoặc chuẩn bị cho các Report cụ thể (dạng Report cho BI).

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

Các chức năng cơ bản của OLTP: cùng với khả năng truy cập và cập nhật các dữ liệu chia sẻ, các hệ thống OLTP còn hỗ trợ các user khả năng truy cập trực tuyến (online), khả năng truy cập tức thời (availability), khả năng phản hồi nhanh chóng (response), và tiết kiệm chi phí đối với từng transaction (low cost).

## Vấn tin khai thác sử dụng Kho dữ liệu

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

**SQL Subtotal**

* Ôn lại 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 lý 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.

+ Viết truy vấn dùng Materialized View:

* Ôn tập lại khái niệm khung nhìn (views)
* Định nghĩa materialized view và cách thức nó được xử lý
* Các quy tắc khi viết lại truy vấn (query rewrite) dùng materialized view
* Ví dụ áp dụng về viết lại truy vấn

### Hệ thống OLAP

\* Cube (khối) Khối dữ liệu + Tạo khối (cube) cho dữ liệu với

* Các bảng đa chiều (dimension table) và
* Bảng sự kiện (fact table)

\* Một số thao tác (lệnh) cơ bản của OLAP:

* Thu nhỏ (roll-up)
* Mở rộng (drill-down)
* Cắt lát (slice)
* Rút ngắn (dice)

# TỔNG QUAN VỀ KHO DỮ LIỆU PHÁT TRIỂN MÔ HÌNH KHO DỮ LIỆU TUYỂN SINH TẠI CÁC TRƯỜNG ĐẠI HỌC VIỆT NAM KHAI THÁC SỬ DỤNG TẠI CÁC CƠ SỞ ĐÀO TẠO ĐẠI HỌC

## Ý tưởng hình thành DWH **Phát triển mô hình kho dữ liệu tuyển sinh tại các trường Đại học Việt Nam khai thác sử dụng tại các cơ sở đào tạo Đại học.**

**-** Việc xây dựng phương án tuyển sinh đóng vai trò rất quan trọng trong quá trình tuyển Sinh tại các cơ sở đào tạo Đại học. Vì thế việc phân tích dữ liệu sinh viên hiện tại cũng như việc đánh giá kết quả tuyển sinh của các đoàn tuyển sinh cũng trở nên quan trọng. Chính vì thế, nhóm chúng em chọn đề tài “*Phát triển mô hình kho dữ liệu tuyển sinh tại các trường Đại học Việt Nam khai thác sử dụng tại các cơ sở đào tạo Đại học*”.

## Giới thiệu chung về Kho dữ liệu

- Kho dữ liệu tuyển sinh được thiết kế theo mô hình “star schema”, bao gồm 1 bảng Fact và 6 bảng Dim thể hiện các vấn đề có liên quan đến việc xây dựng kế hoạch tuyển sinh hợp lý

### Mô tả tổng quan về DWH

+ Bảng Fact trung tâm: Fact\_Statistical (SchoolId, ProvinceId, MajorId, FacultyId, Eyear, CountStudent, AvgEMark, AvgAMark, Cost)

+ Bảng Dim đại diện trường THPT: Dim\_School (Id, Name, IsTop, ProvinceId)

+ Bảng Dim đại diện tỉnh: Dim\_Province (Id, Name)

+ Bảng Dim đại diện ngành: Dim\_Major (Id, Name, FacultyId)

+ Bảng Dim đại diện khoa: Dim\_Faculty (Id, Name)

+ Bảng Dim đại diện nhóm tuyển sinh: Dim\_Consulting\_Group (Id, Name)

+ Bảng Dim đại diện năm: Dim\_Year (Id)

### Các dữ liệu nguồn hình thành DWH

- Dữ liệu nguồn là 1 cơ sở dữ liệu bao gồm 8 bảng:

+ Sinh viên

+ Tình trạng học tập

+ Trường THPT

+ Khoa

+ Ngành

+ Tỉnh

+ Đoàn tuyển sinh

+ Thống kê chi tiêu, thành viên của Đoàn tuyển sinh tại các trường THPT

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

- Các vấn đề cần quan tâm khi tuyển sinh:

+ Xu hướng nghề nghiệp của sinh viên tại các khu vực

+ Chất lượng của sinh viên tại các khu vực

+ Xu hướng nghề nghiệp theo thời đại

+ Chất lượng của các Đoàn tuyển sinh

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

<Đề tài chung của Nhóm: copy từ ds phân công vào>>

<04>\_Phan Thành Đạt\_ Nhóm trưởng: Thành phần kho dữ liệu ngành học - khu vực.

<09>\_Lê Chí Hiếu: Thành phần kho dữ liệu chất lượng sinh viên - khu vực.

<47>\_Nguyễn Anh Triều: Thành phần kho dữ liệu chất lượng đoàn tuyển sinh.

<33>\_Đỗ Đình Phùng: Thành phần kho dữ liệu ngành học - thời gian.

Chú ý: không nhất thiết phân công hết = do số lượng thành viên

+ các nội dung khác (tùy đề tài)

# PHÂN TÍCH, THIẾT KẾ VÀ KHAI THÁC SỬ DỤNG CÁC PHÂN HỆ KHO DỮ LIỆU PHÁT TRIỂN MÔ HÌNH KHO DỮ LIỆU TUYỂN SINH TẠI CÁC TRƯỜNG ĐẠI HỌC VIỆT NAM KHAI THÁC SỬ DỤNG TẠI CÁC CƠ SỞ ĐÀO TẠO ĐẠI HỌC

## 04\_Phan Thành Đạt\_ Thành phần kho dữ liệu ngành học - khu vực

## 09\_Lê Chí Hiếu\_ Thành phần kho dữ liệu chất lượng sinh viên - khu vực

## 33\_Đỗ Đình Phùng \_ Thành phần kho dữ liệu ngành học - thời gian

* 1. **47\_Nguyễn Anh Triều \_ Thành Phần kho dữ liệu đánh giá chất lượng của đoàn tuyển sinh**
     1. **Mô tả tổng quan về phân hệ DWH đánh giá chất lượng của các đoàn tuyển sinh**

Phân hệ nhằm phục vụ cho các cấp trên có cái nhìn tổng quan hơn về các đoàn tuyển sinh qua các năm, cụ thể:

Số lượng sinh viên và tổng tiền qua các năm,các tỉnh mà các được các đoàn tuyển sinh có chênh lệch nhau nhiều không.

Tổng số lượng sinh viên được đoàn tuyển sinh có liên quan gì đến số tiền mà đoàn được cấp không.

Khi có cái nhìn quan hơn từ tập dữ liệu đã được làm sạch và gọn gàng thì nhà trường có được đánh giá tốt để đưa ra các chiến lược tuyển sinh phù hợp cho nhà trường.

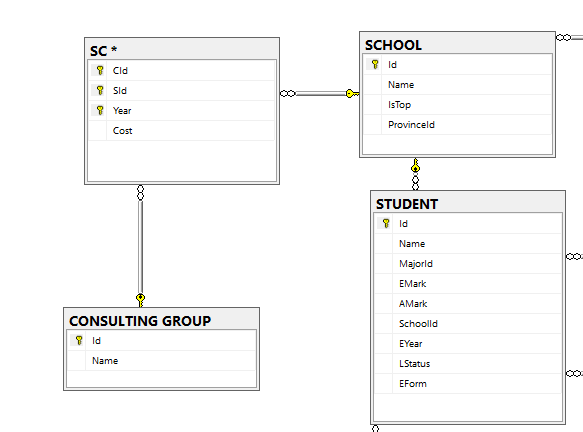
* + 1. **Xác định nguồn dữ liệu tích hợp phân hệ Kho dữ liệu chất lượng của đoàn tuyển sinh**

Dữ liệu liên quan đến phân hệ được trích từ nhiều nguồn. Nguồn các danh sách tất cả các trường từ bảng School, nguồn chi phí và số năm tuyển sinh của đoàn tuyển sinh từ table SC, nguồn danh sách tất cả các đoàn từ table Consulting Group

* + 1. **Xác định ý tưởng hình thành phân hệ của Kho dữ liệu chất lượng của đoàn tuyển sinh**

Ý tưởng: Tham khảo phương thức tư vấn tuyển sinh của trường đại học sư phạm kỹ thuật thành phố Hồ Chí Minh thì nhóm em quyết định chọn đề tài tư vấn tuyển sinh làm đề tài thực hiện cho môn kho dữ liệu (DWH). Phân hệ đánh giá chất lượng của đoàn tuyển sinh giúp nhà trường có kế hoạch cho mỗi đoàn, cách đánh giá tốt về chất lượng của mỗi đoàn, giúp cải thiện việc tuyển sinh cho nhà trường tốt hơn qua các năm.

* + 1. **Mô tả hệ thống các DB gốc liên quan phân hệ DWH đánh giá chất lượng của đoàn tuyển sinh.**



* ***Mô tả các table:***

SCHOOL

**Danh sách của tất cả các trường của các sinh viên theo học tại nhà trường.**

Id: Mã Trường

Name: Tên Trường

IsTop: Trường đó có thuộc top hay không

Provinceid: Mã Tỉnh

CONSULTING GROUP

Danh sách của 30 đoàn tuyển sinh của nhà trường

Id: Mã Đoàn

Name: Tên Đoàn

SC

**Danh sách các nhà trường được đoàn tuyển sinh đi đến qua các năm**

Cid: Mã Đoàn

Sid: Mã Trường

Year: Năm

Cost: Số Tiền

STUDENT

**Danh sách tất cả các sinh viên theo học tại nhà trường**

Id: Mã Sinh Viên

Name: Tên Sinh Viên

Majorid: Mã Tỉnh

Schoolid: Mã Trường

Eyear: Năm

Lstatus: Tình trạng học tập

Eform: Hình thức xét tuyển

* + 1. **Phần phân tích các DB gốc xác định yêu cầu phân tích DWH**

**Factor: Fact\_ThongKe**

+ Cid: Mã đoàn từ Table Consulting Group

+ Sid: Mã trường từ Table School

+ Year: Số năm từ Table Year

+ **Tổng tiền =** AVG(**Cost**) WHERE Dim SC

Mô tả: Tổng tiền của các đoàn được cấp tại mỗi trường.

**+ Tổng sinh viên** = Count(**id**) WHERE Dim STUDENT

Mô tả: Tổng sinh viên tại các trường mà đoàn tư vấn.

**Dim: Dim School, Dim Year, Dim Consulting Group**

+ [Dim School].id = PK, name: Danh sách của tất cả các trường, để lấy được danh sách các trường đoàn đã tuyển sinh tại trường đó.

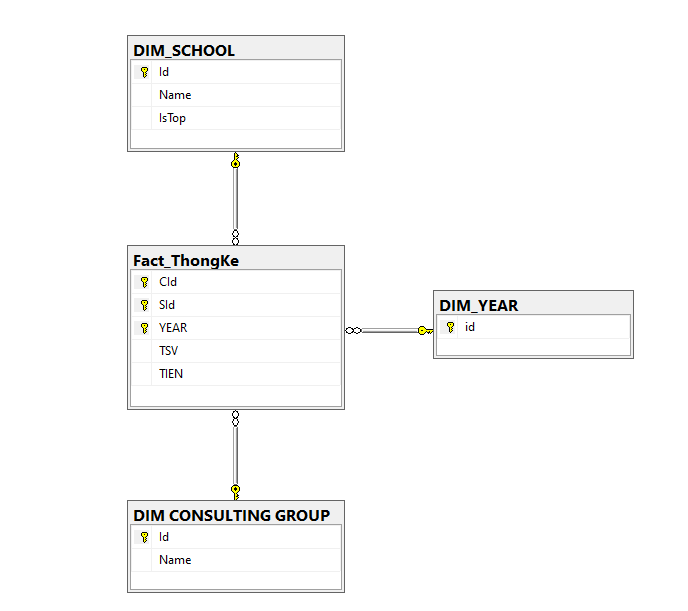
+ [Dim Year].id = PK: Danh sách tất các các năm mà các đoàn tuyển sinh(2010🡪2020)

+ [Dim Consulting Group].id = PK, name: Danh sách của tất cả 30 đoàn để phục vụ cho việc xem dữ liệu số tiền và tổng số sinh viên của các đoàn.

* + 1. **Thiết kế DB mới tổ chức phân tích Kho dữ liệu cho phân hệ**

Bảng Fact\_statis của nhà kho là dữ liệu của tất cả sinh viên được và không được đoàn tuyển sinh gồm: mã trường, mã tỉnh, mà khoa, tổng sinh viên , số tiền, điểm số, mã đoàn, năm.

Bảng Fact\_statis chi tiết nên dữ liệu phục vụ cho việc phân tích các đoàn tại các nhà trường chưa phù hợp, cụ thể khi lấy tổng tiền của đoàn tại nhà trường thì nó sẽ lấy số lượng sinh viên thuộc trường đó để cộng tổng vào nên dữ liệu ta khó xử lý và rườm rà 🡪 Cần thiết thiết kế hệ thống nhà kho riêng cho phân hệ cá nhân.

**Hệ thống nhà kho riêng của cá nhân** 

**Thiết kế:**

DWH gồm 1 Fact và 3 Dim, tổ chức theo mô hình hình sao.

Phân tích:

Bảng Fact: Số tiền và tổng sinh viên tại mỗi nhà trường mà đoàn đã tư vấn qua các năm.

Fact\_ThongKe: (Cid, Sid, Year, TSV, TIEN)

Primary Key: Cid, Sid, Year

Bảng Dim: Phục vụ cung cấp dữ liệu cho bảng Fact.

Dim school(id, Name, is Top)

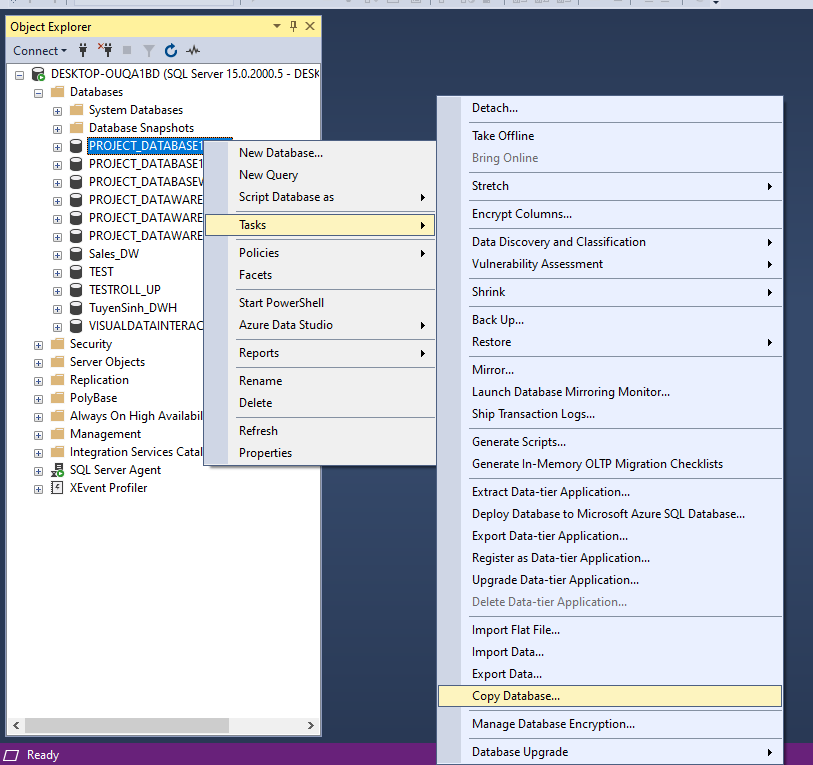
Dim Consulting Group(Id,Name)

Dim year(id)

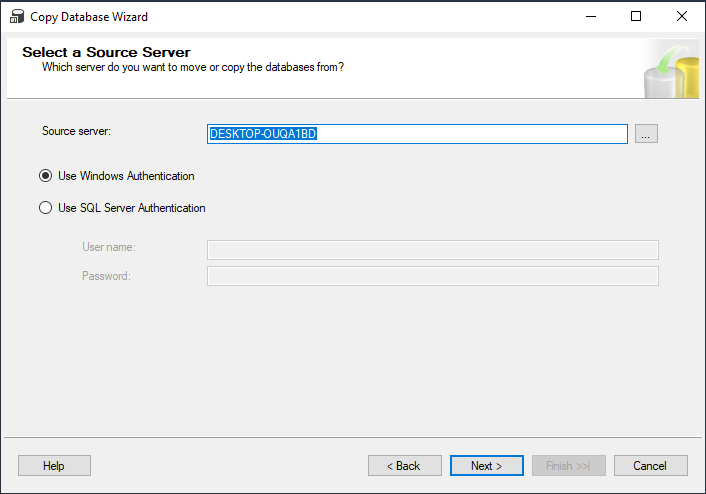
Thực hiện các thủ tục tích hợp các DB gốc vào phân hệ DWH đánh giá chất lượng của đoàn tuyển sinh

Dữ liệu được copy từ Data base gốc(PROJECT\_DATABASE1\_new) 🡪 sang DWH() để thuận lợi cho việc phân tích nhà kho

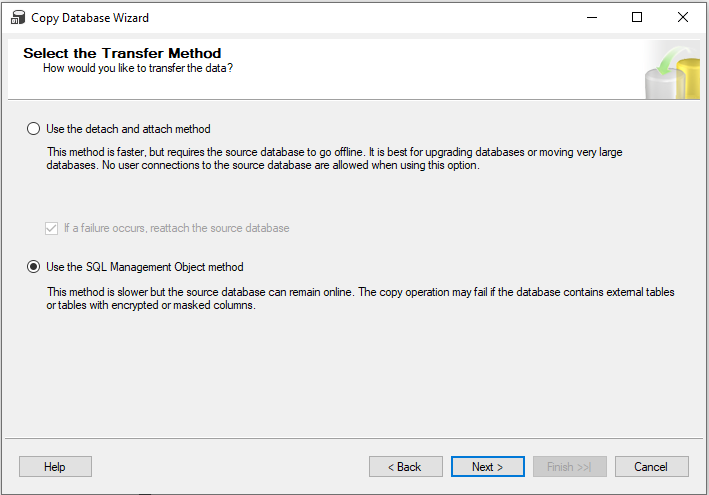
Chọn Data base gốc để copy



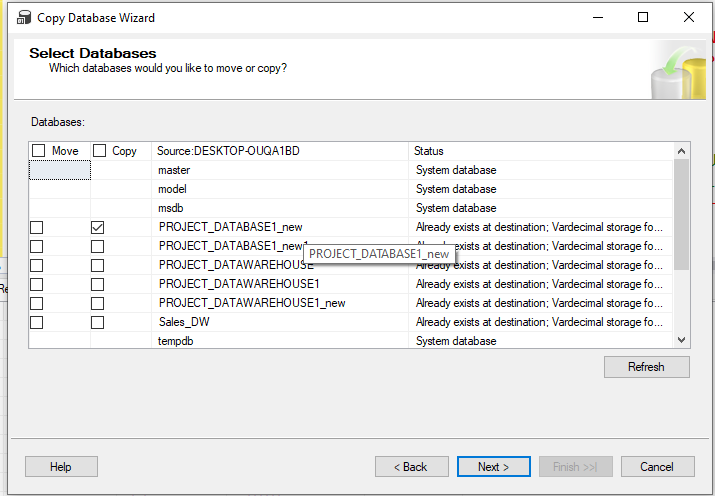
Chọn tên server để kết nối



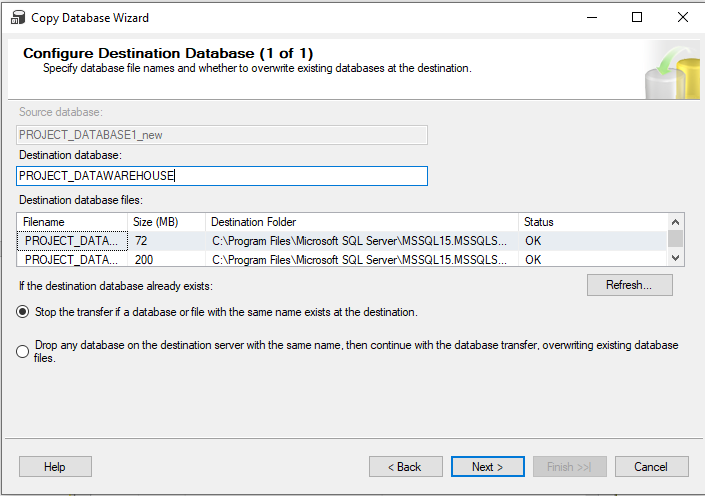
Check in Use the SQL management Object method: có nghĩa là database sau khi copy vẫn chạy một cách độc lập và không bị phụ thuộc vào database gốc và database gốc không bị tình trạng offline.



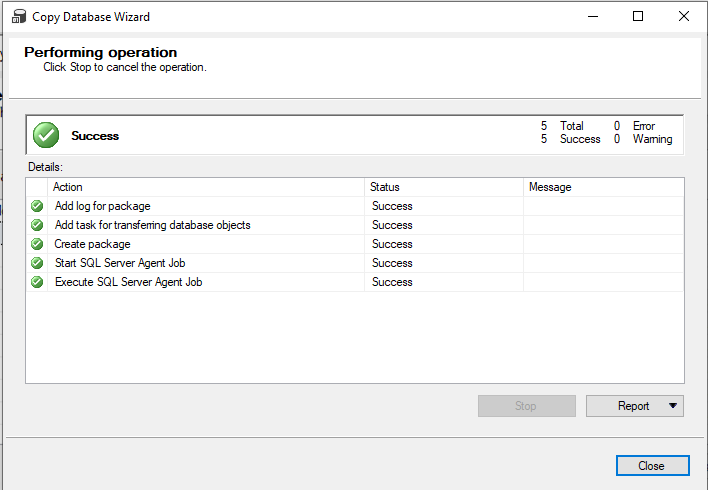
Chọn database cần copy



Đặt tên cho data base mới



Copy data thành công



* + 1. **Thiết lập các Views tính toán cần thiết để nạp dữ liệu từ DB gốc vào các Factors của phân hệ DWH**

Đầu tiên tạo ra một view chung để có được tất cả sinh viên: có mã đoàn , mã trường, năm vào, số tiền, id

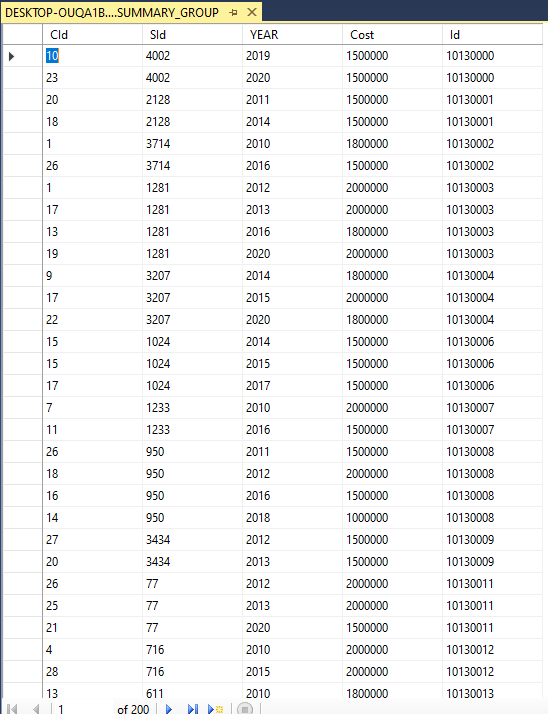
CREATE VIEW [dbo].[SUMMARY\_GROUP] AS

SELECT CId, SId ,YEAR ,Cost , STUDENT.Id

FROM SC, STUDENT

WHERE SC.SId = STUDENT.SchoolId

GROUP BY CId, SId, Year, Cost, STUDENT.Id



Số dòng: 132683

Đồng nghĩa sẽ có 132683 sinh viên.

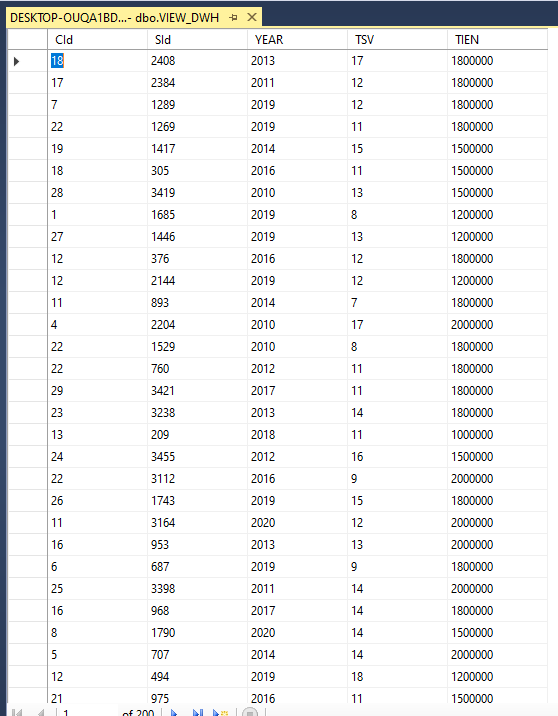
Tiếp theo tạo ra một view để phục vụ cho việc nạp dữ liệu vào bảng Fact

CREATE VIEW [dbo].[VIEW\_DWH] AS

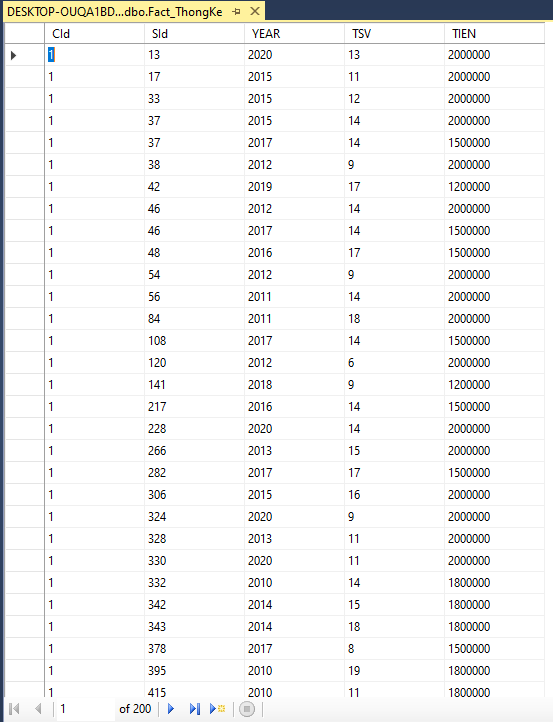
SELECT SUMMARY\_GROUP.CId,SUMMARY\_GROUP.SId, SUMMARY\_GROUP.YEAR, COUNT(SUMMARY\_GROUP.Id) AS TSV, AVG(SUMMARY\_GROUP.Cost) AS TIEN

FROM SUMMARY\_GROUP

GROUP BY SUMMARY\_GROUP.CId,SUMMARY\_GROUP.SId, SUMMARY\_GROUP.YEAR



Dữ liệu được đỗ thành công vào bảng Fact

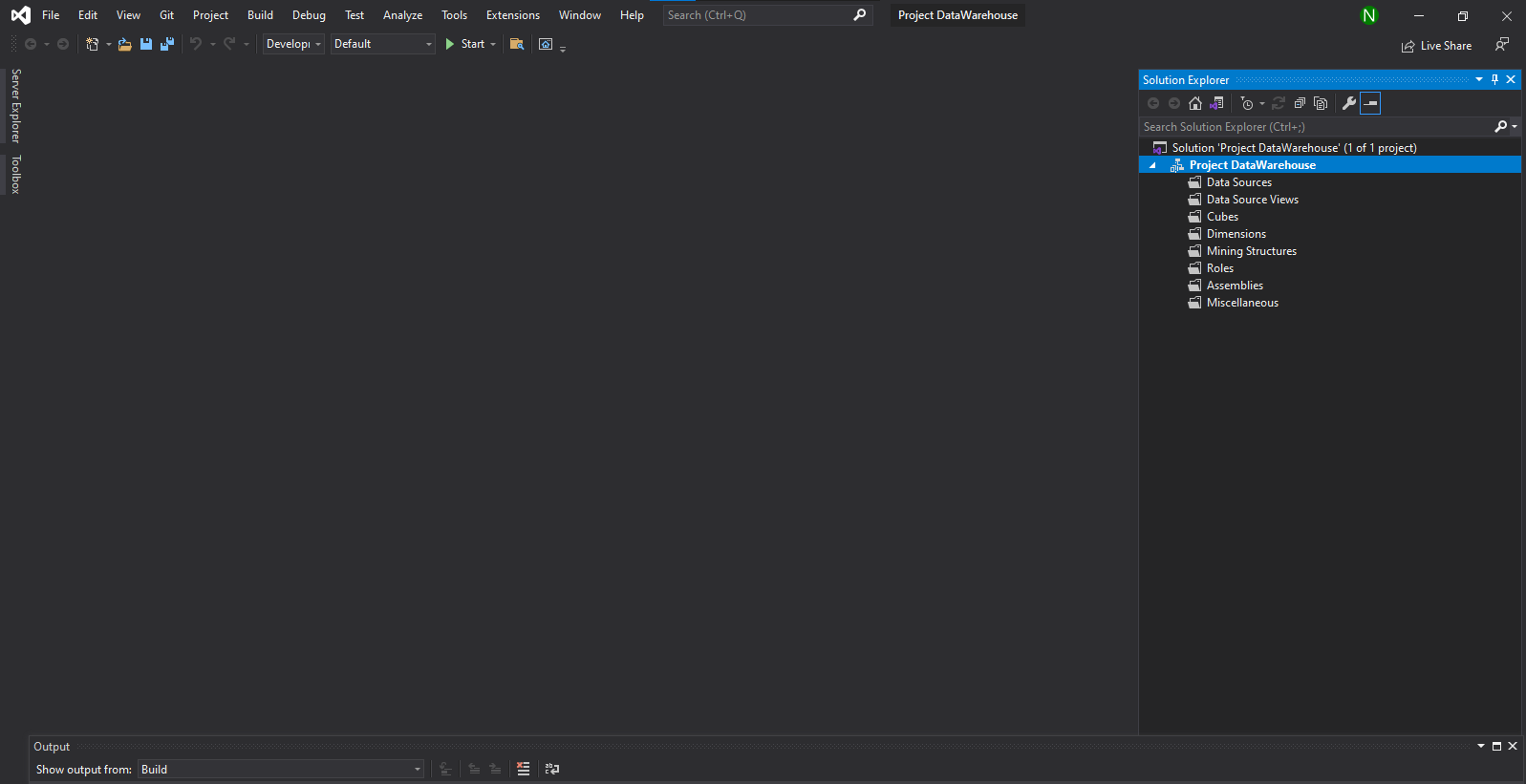


Dữ liệu bảng Fact gồm: 9900 dòng.

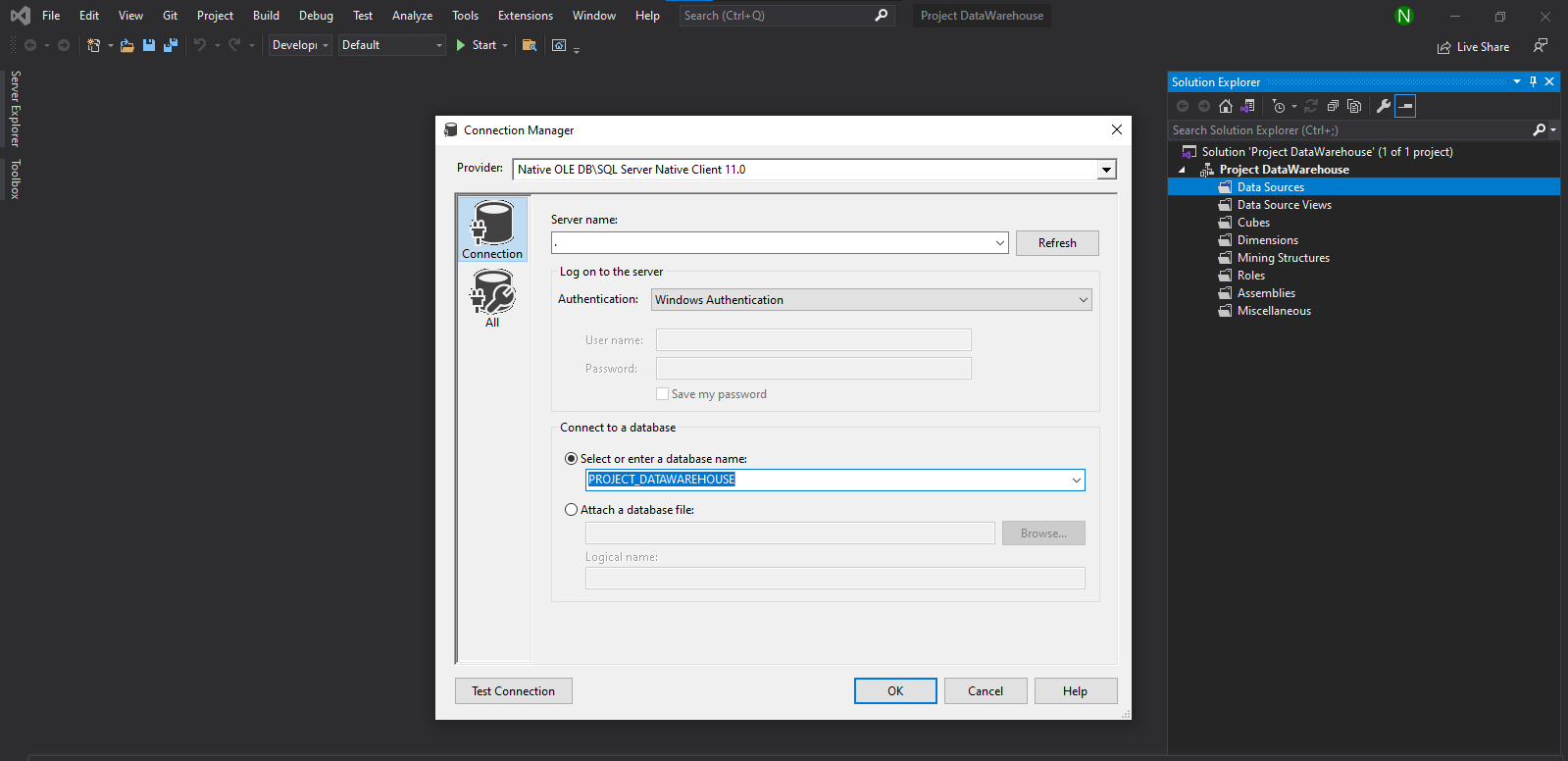
* + 1. **Triển khai phân hệ DWH lên BIDS**

**Tạo project SSAS trên visual studio 2019**

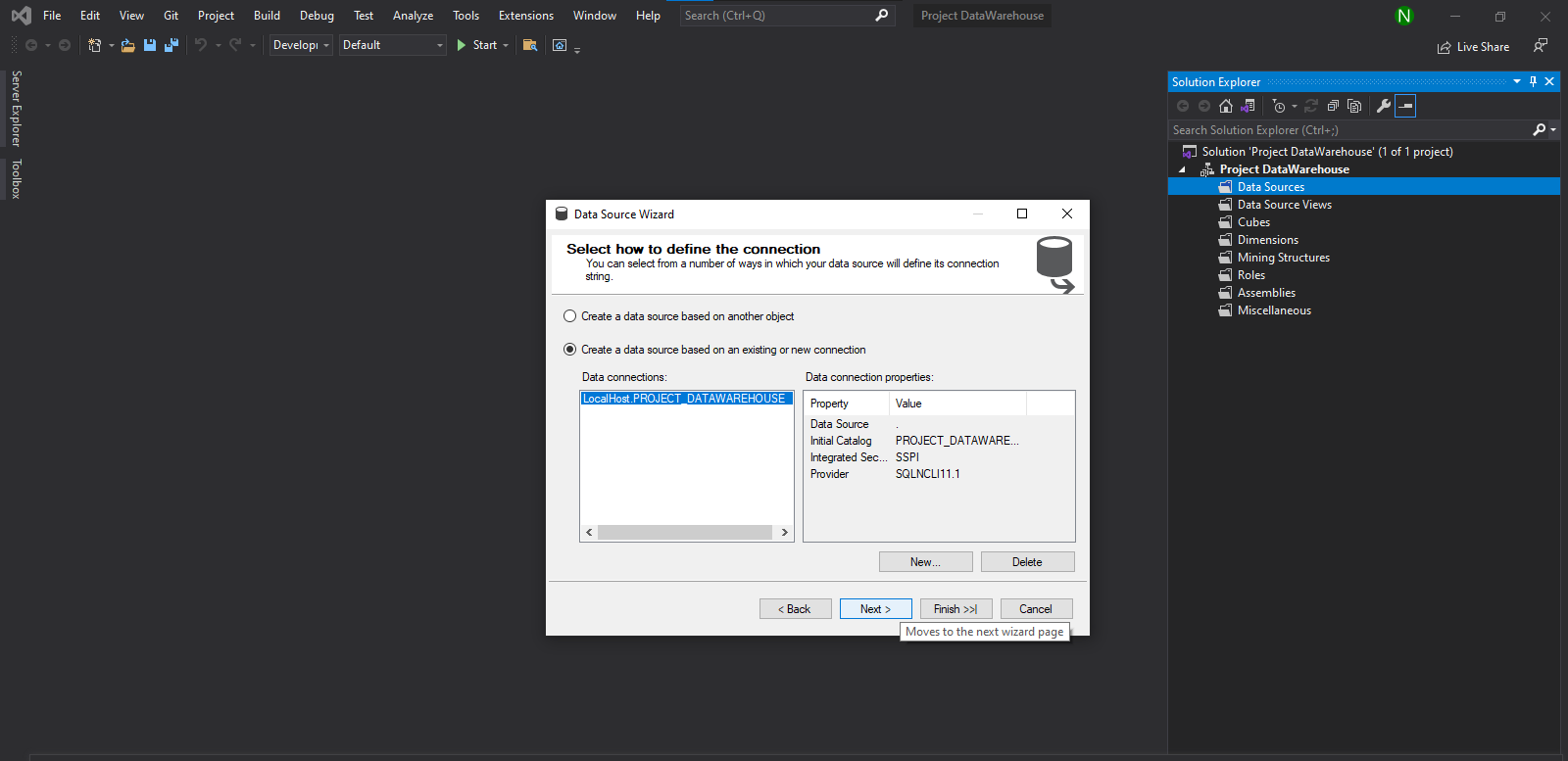
Tạo Project tên Project\_DataWarehouse



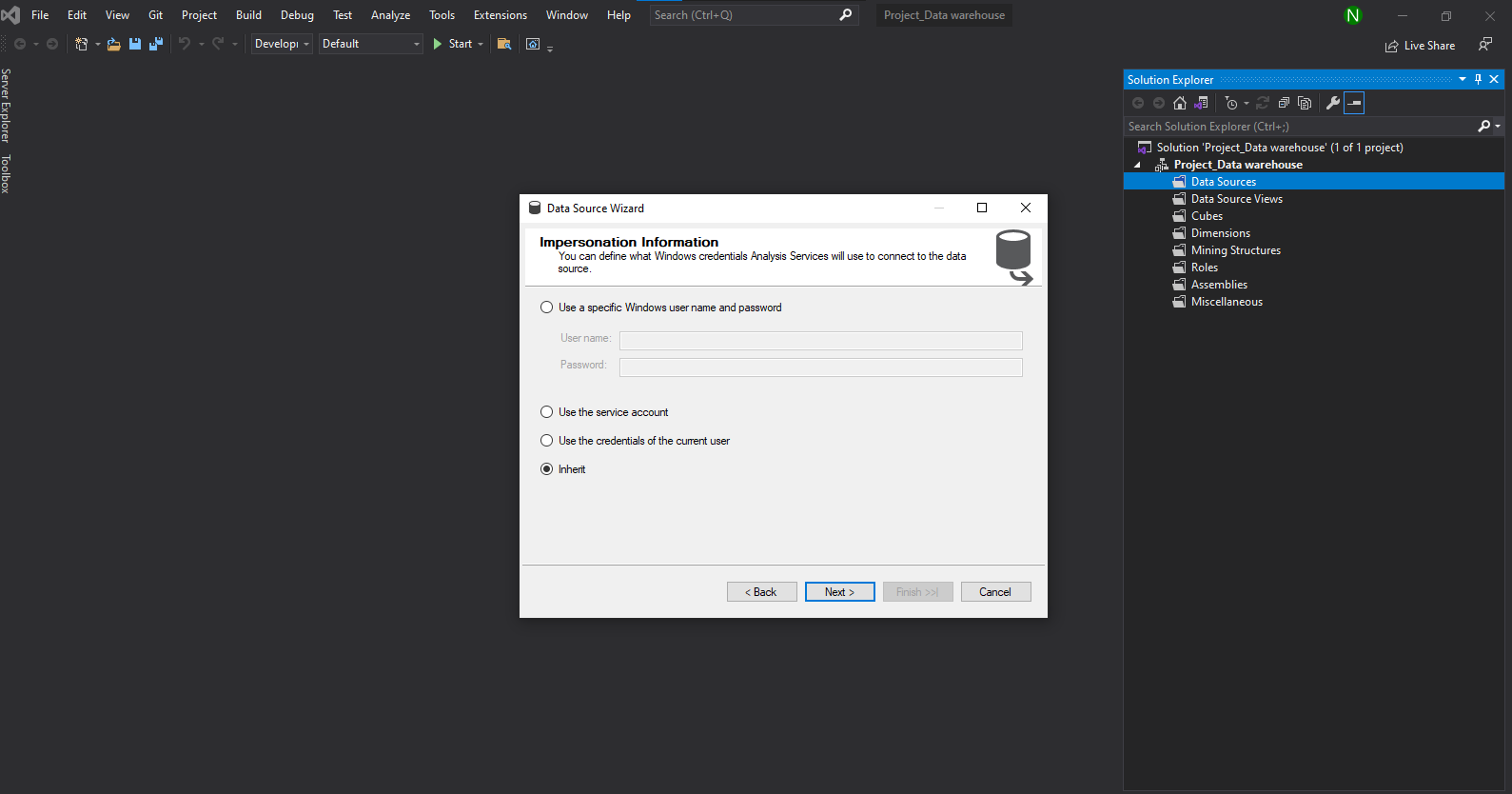
Chọn data base để đưa lên BIDS



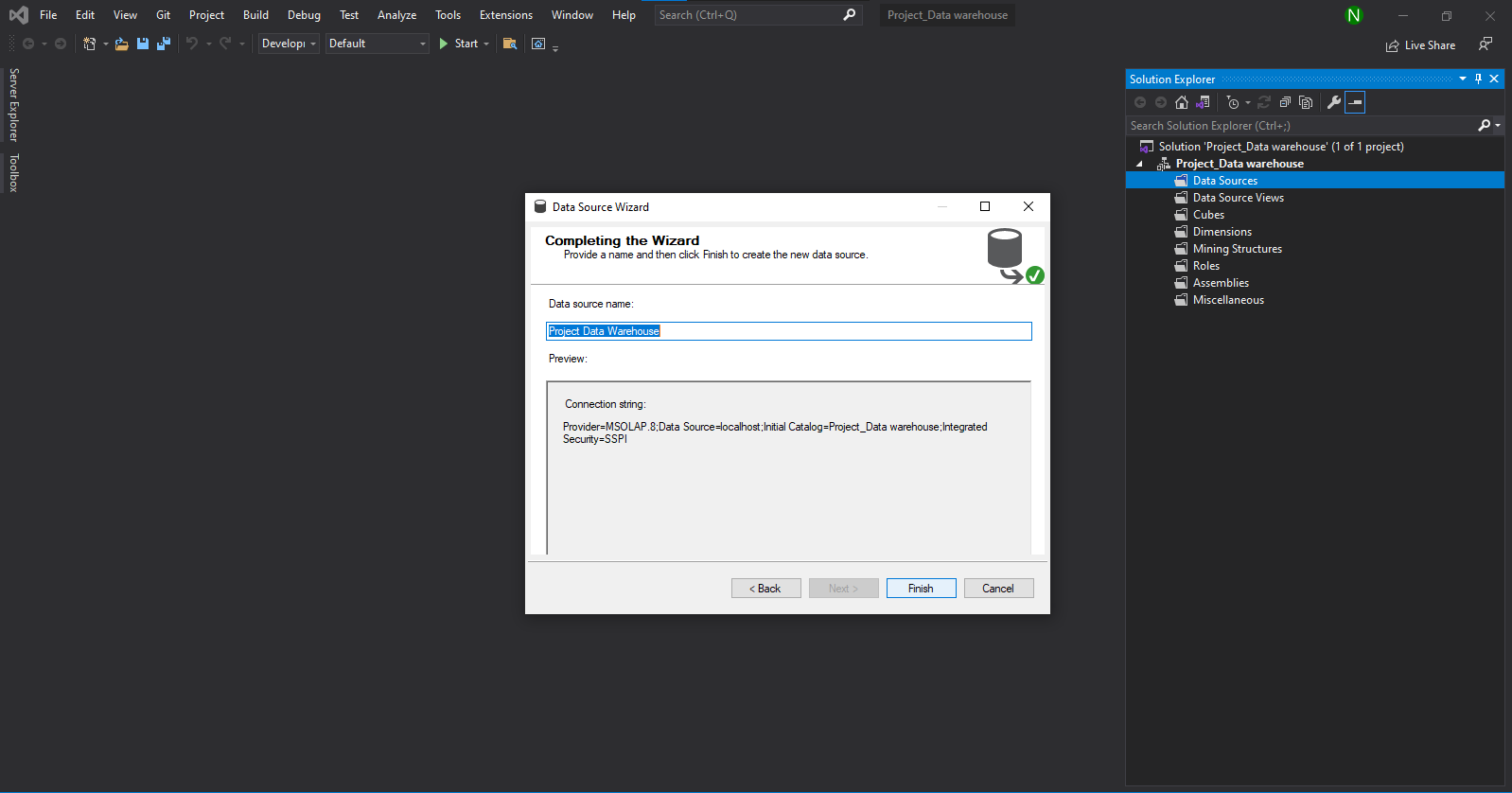
Chọn next



Chọn Inhert

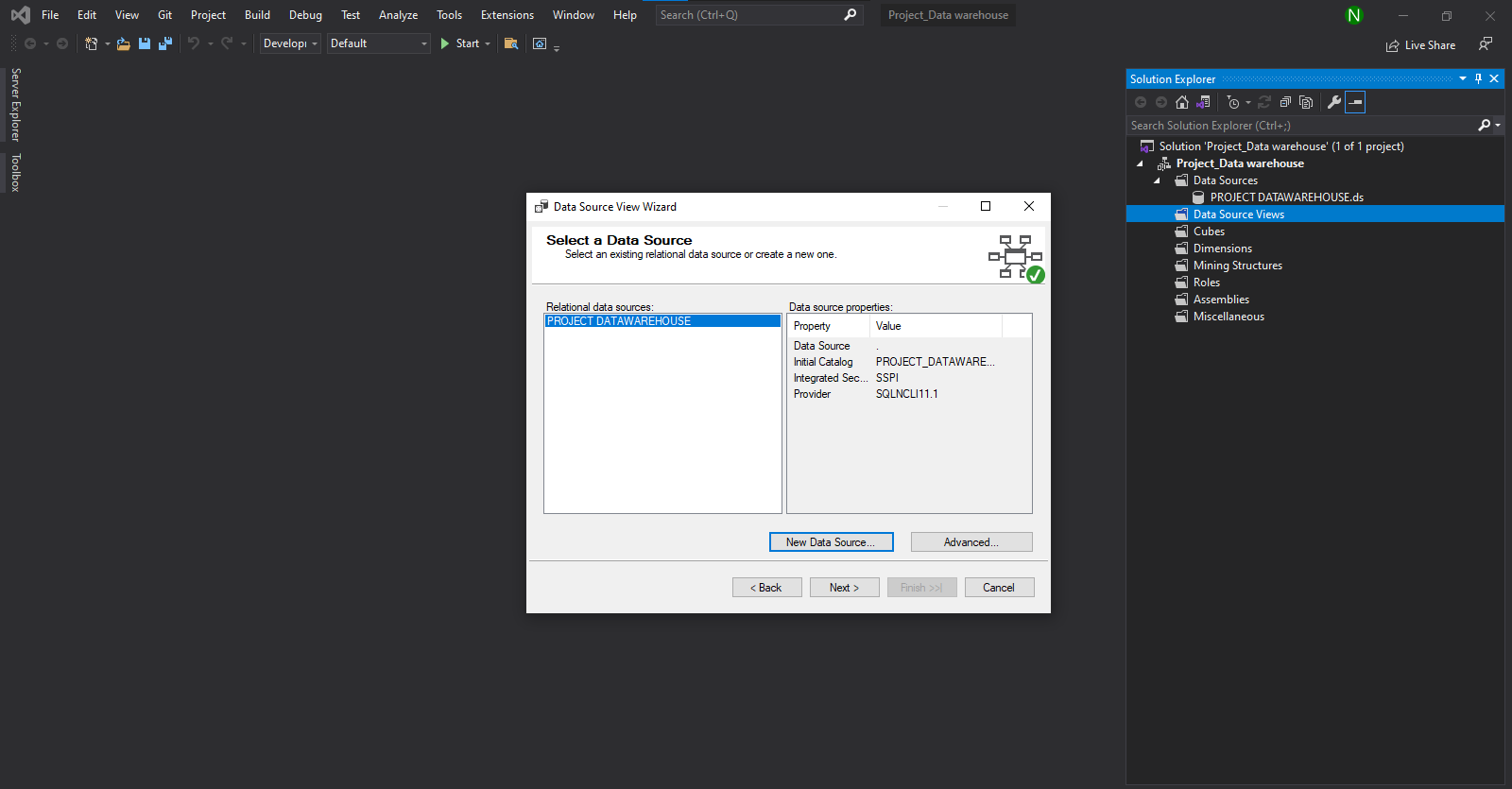


Chọn finish

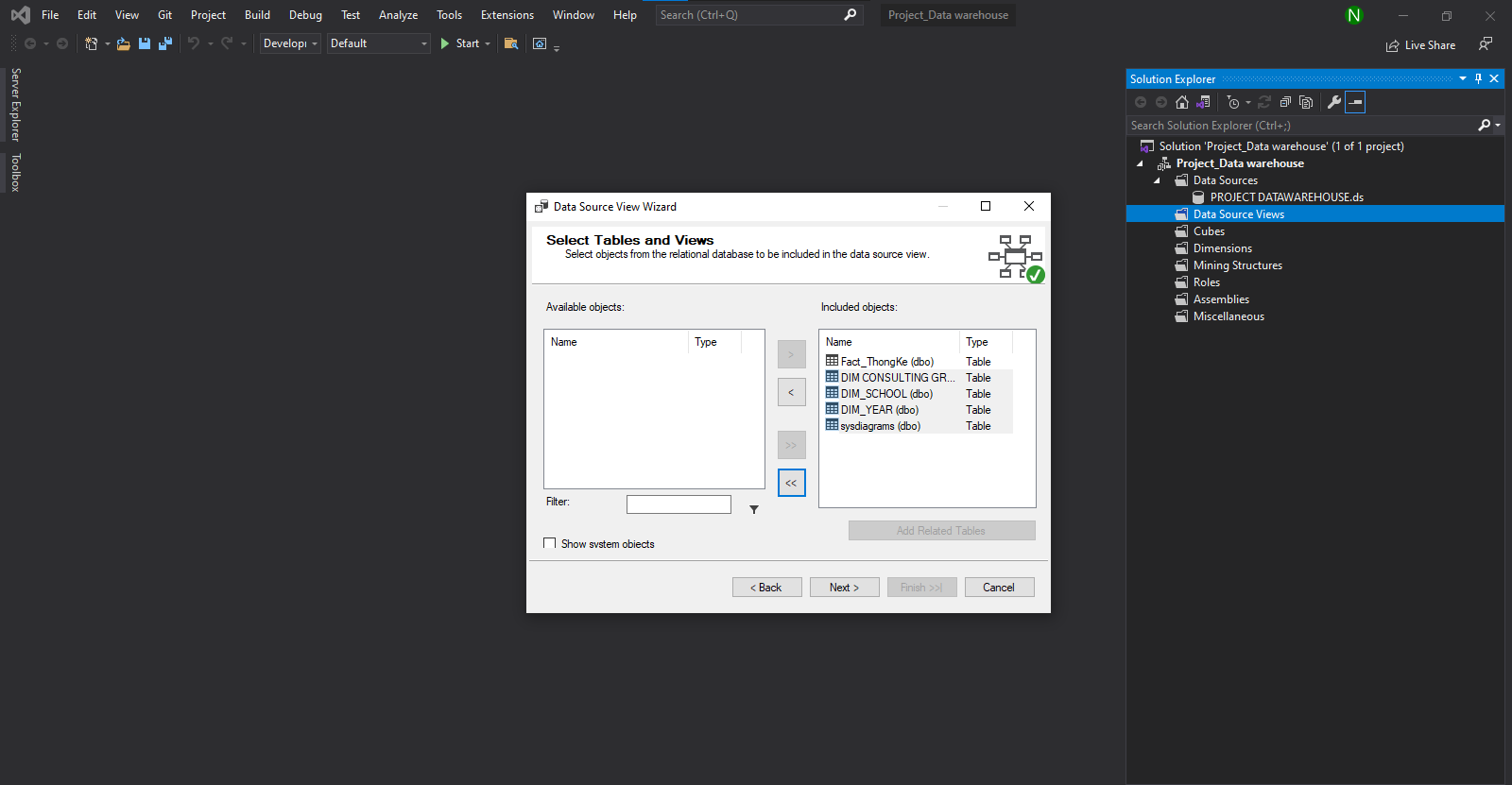


Tạo data source view

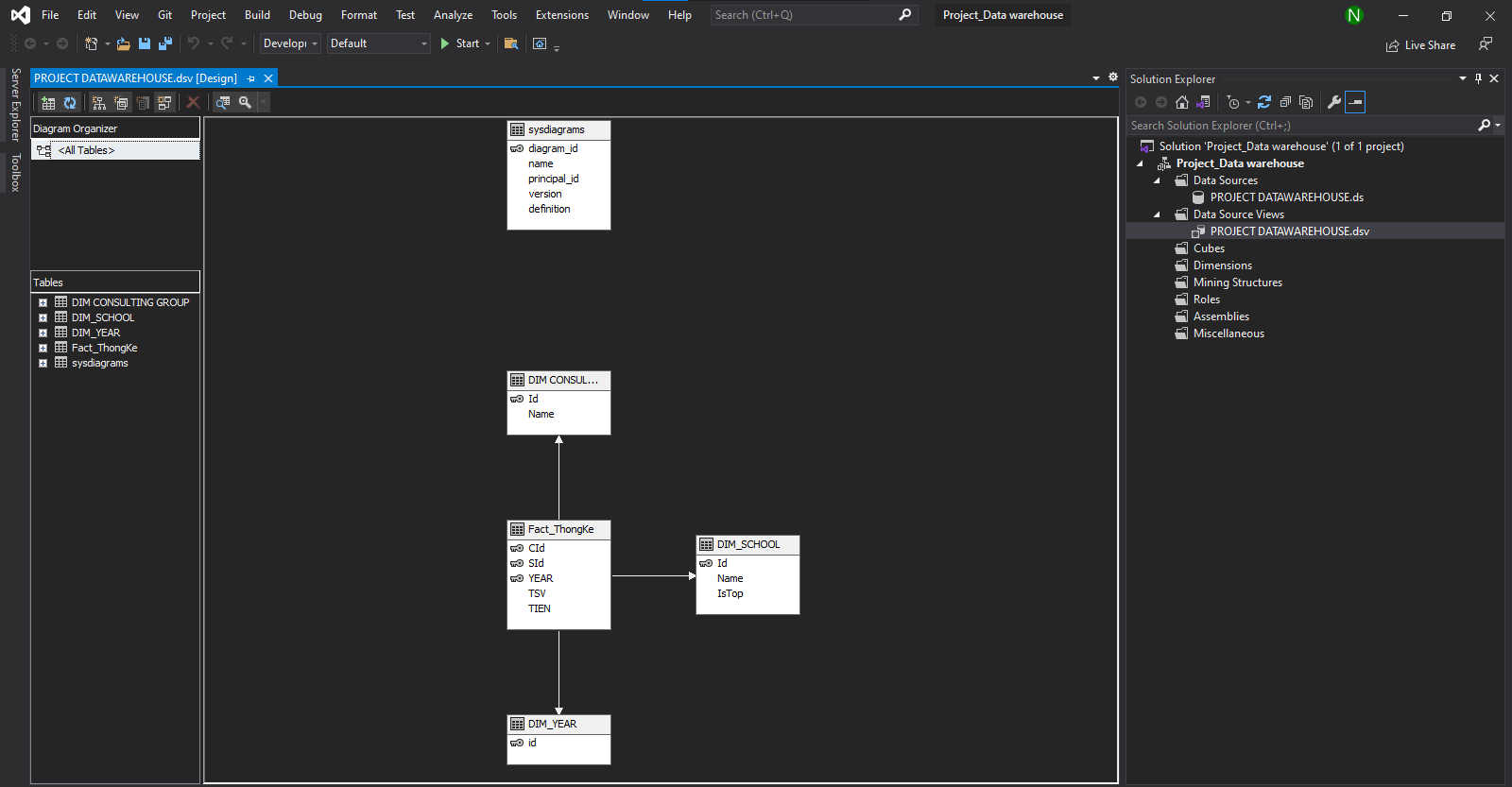
Nhấn next



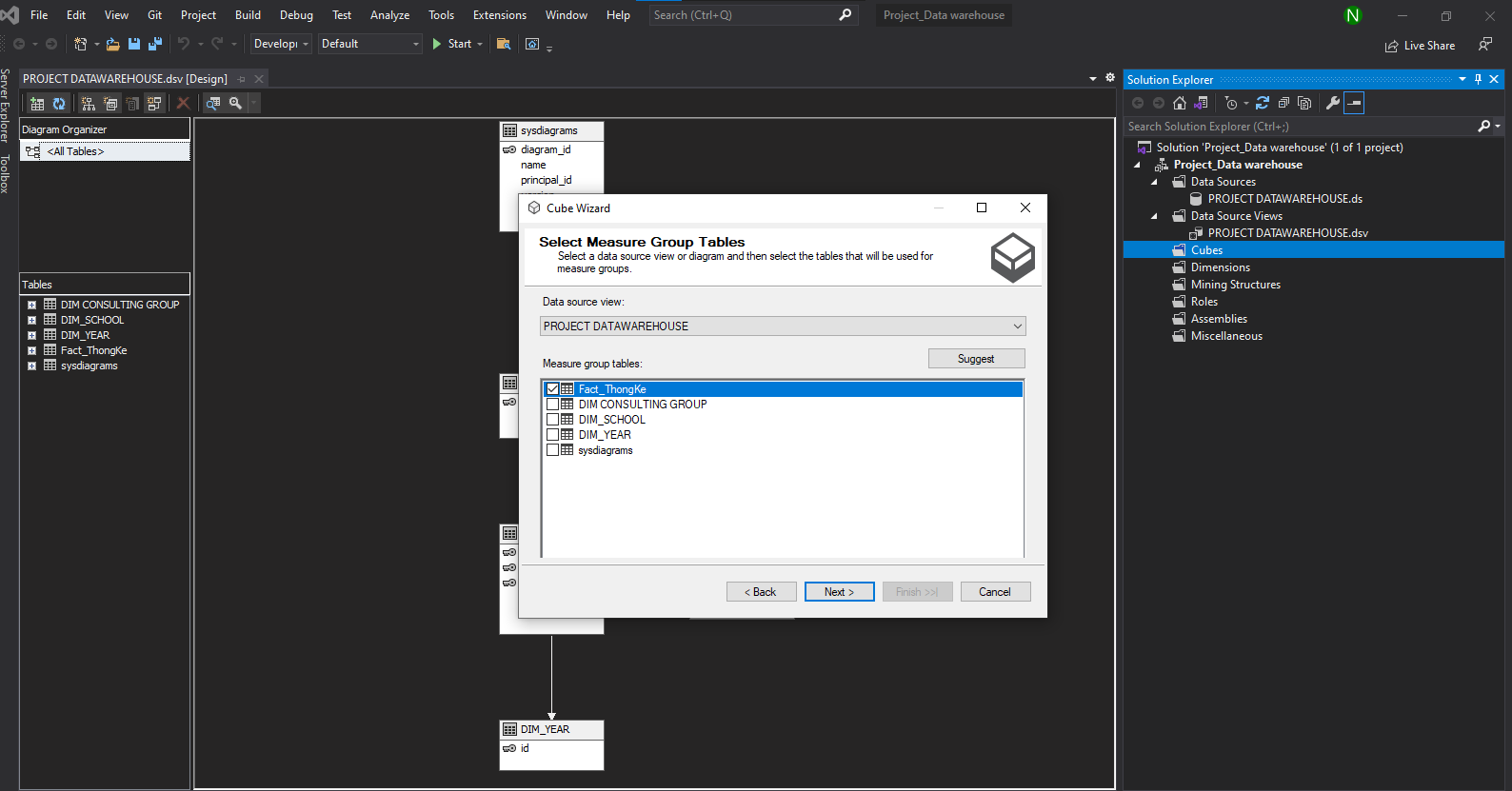
Đưa bảng Fact và Dim vào include object rồi nhấn next



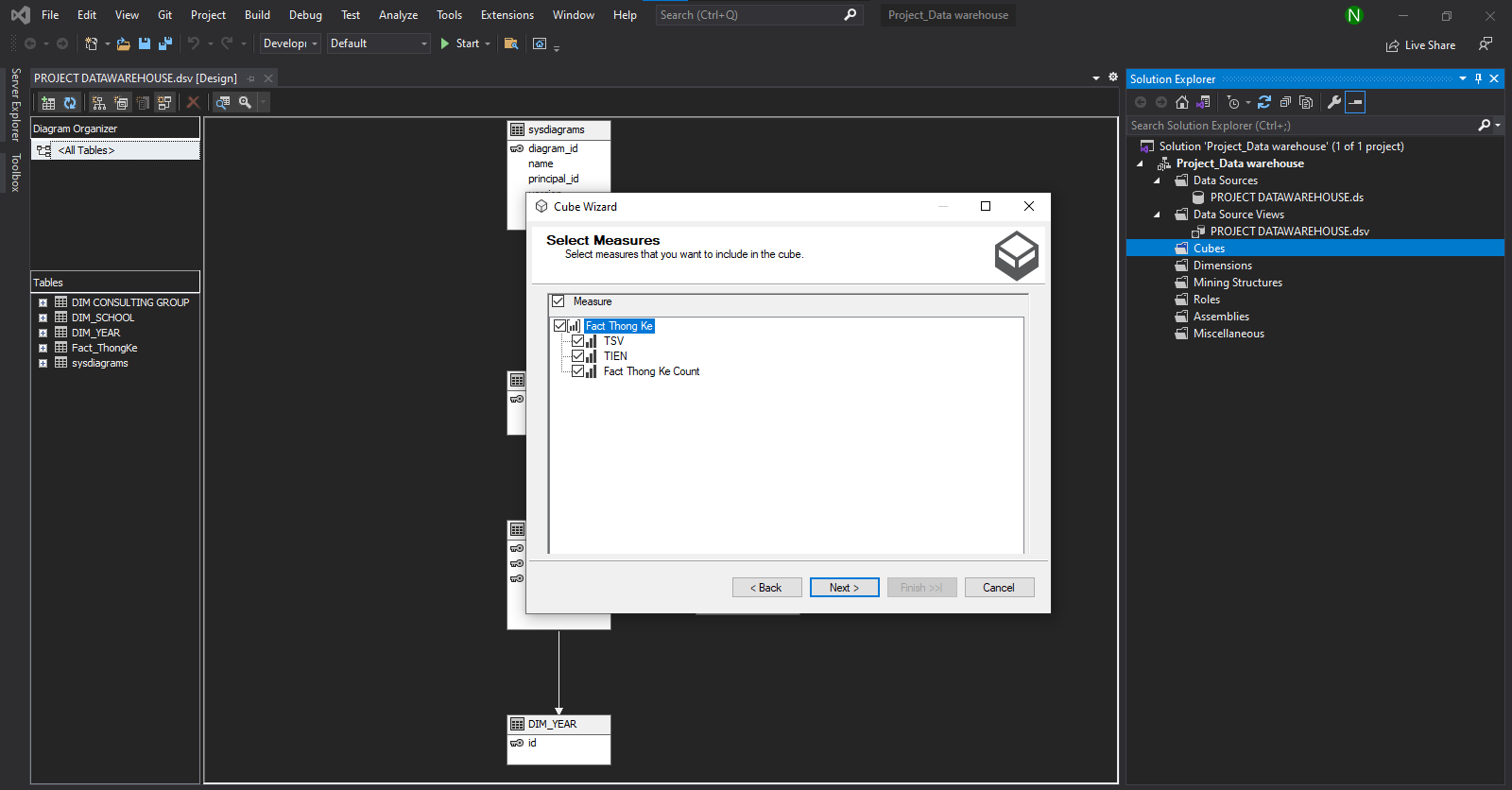
Tạo thành công data source view



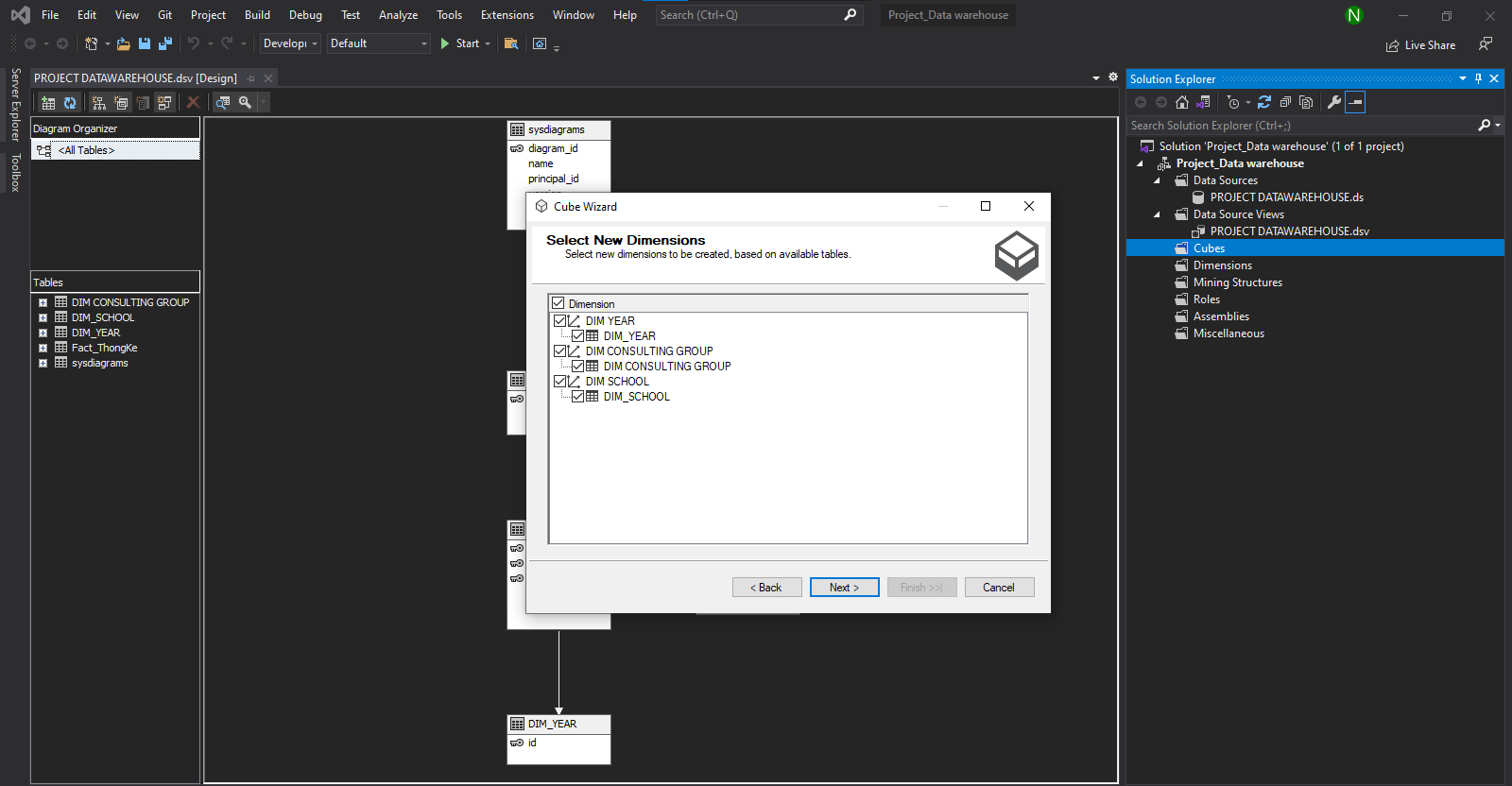
Tạo khối cube chọn measure để phân tích 🡪 Chọn Fact thống kê 🡪 next



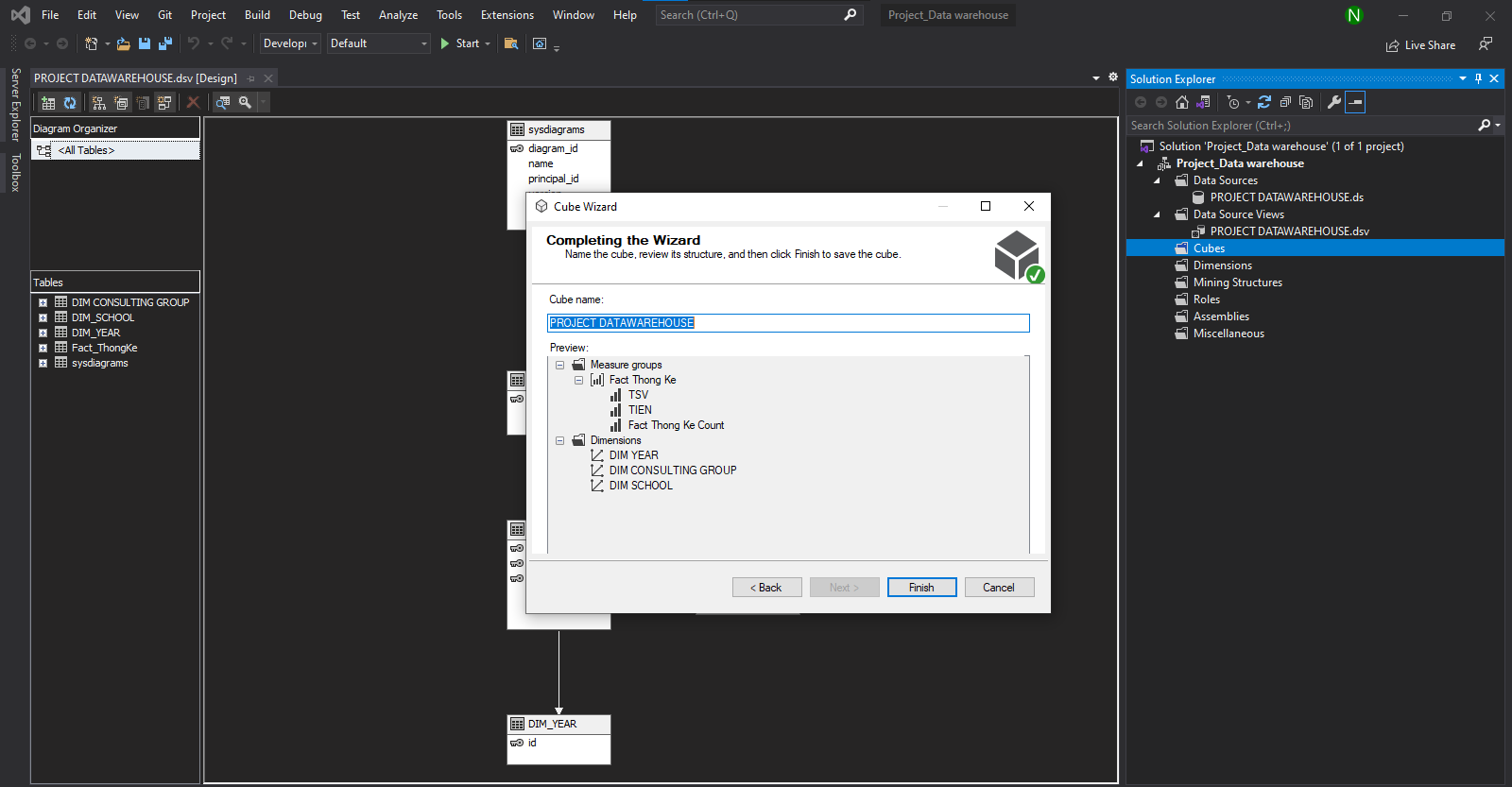
Chọn next



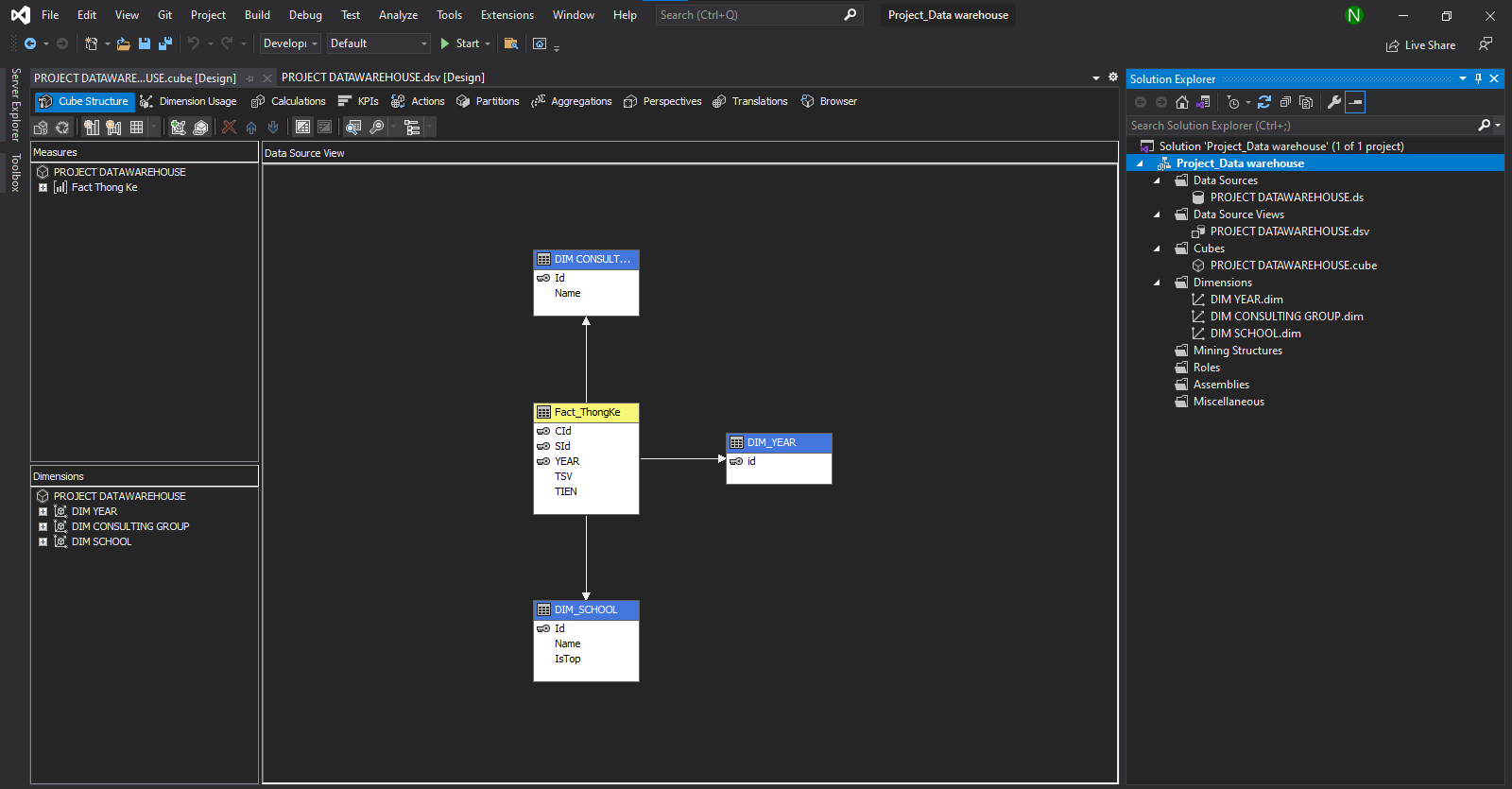
Chọn next



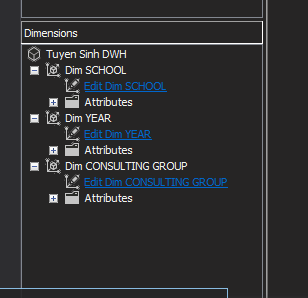
Chọn finish



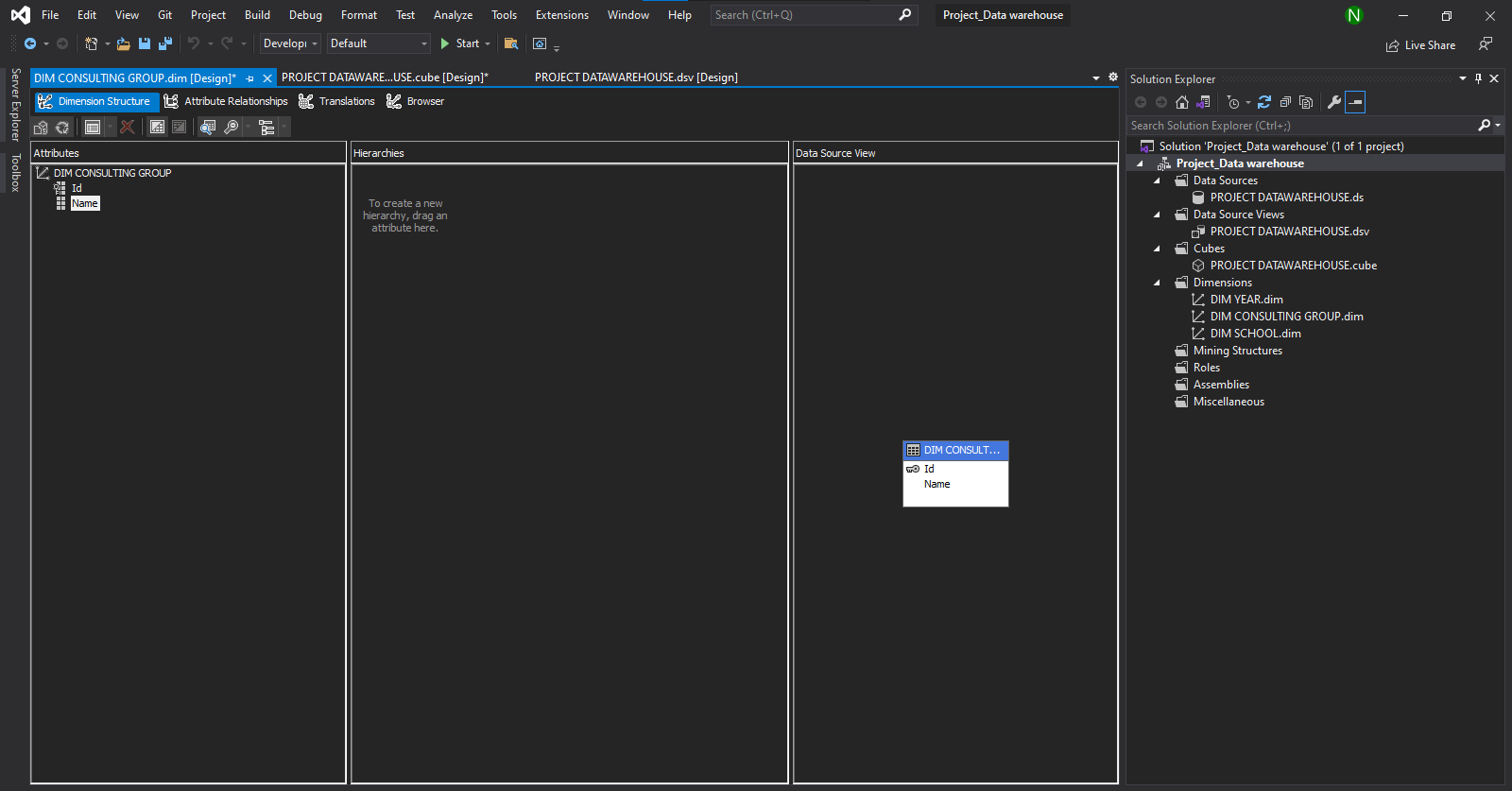
Tạo khối cube thành công



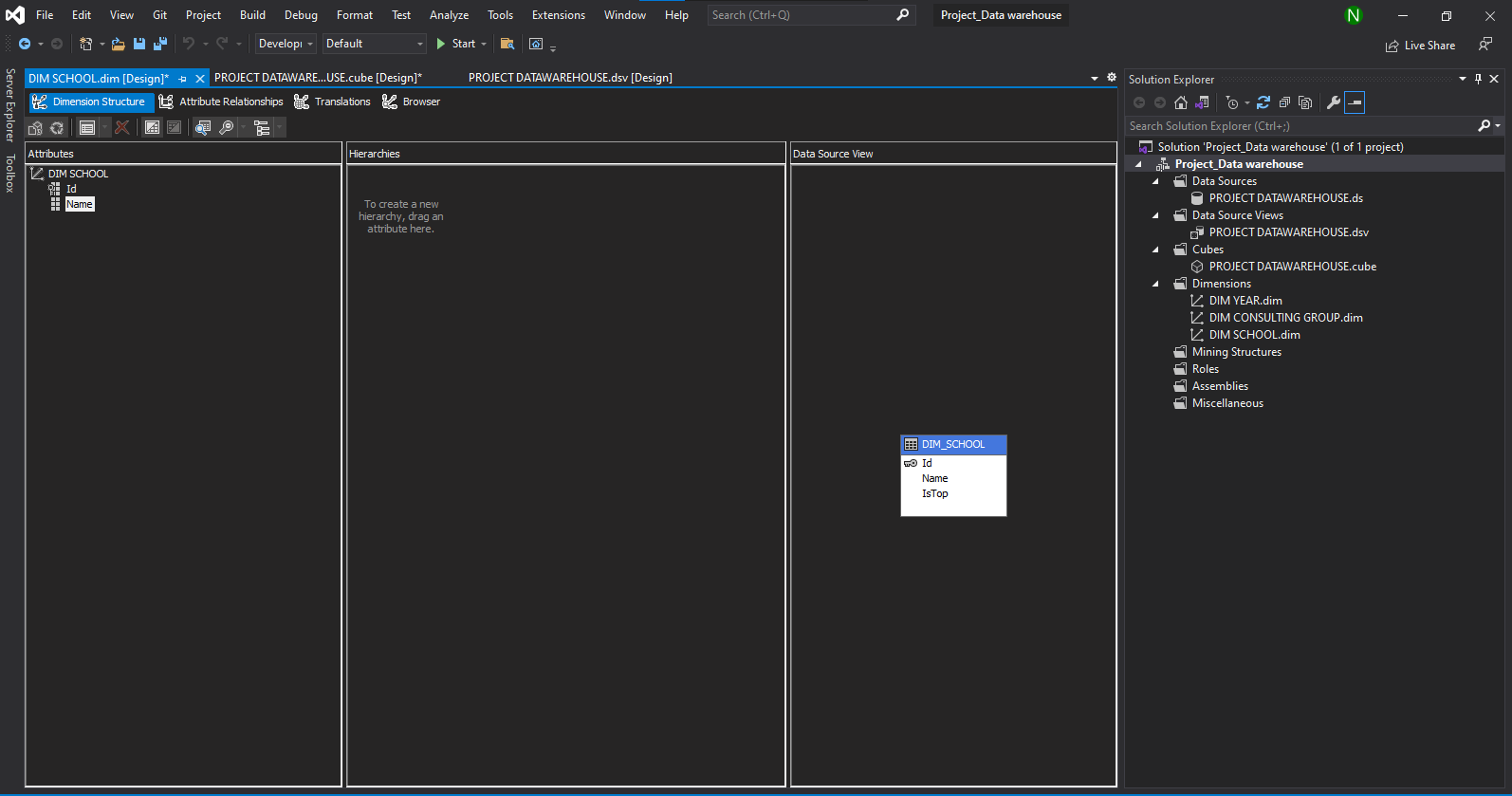
Chọn edit để lấy thêm các cột để phân tích



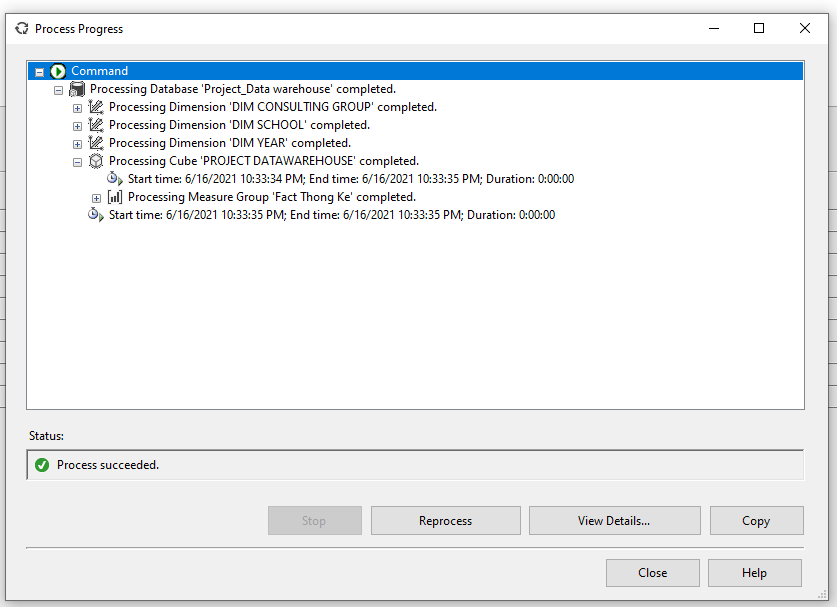
Kéo thả các cột cần phân tích ví dụ là name ở bảng Consulting group



Kéo cột name ở bảng School

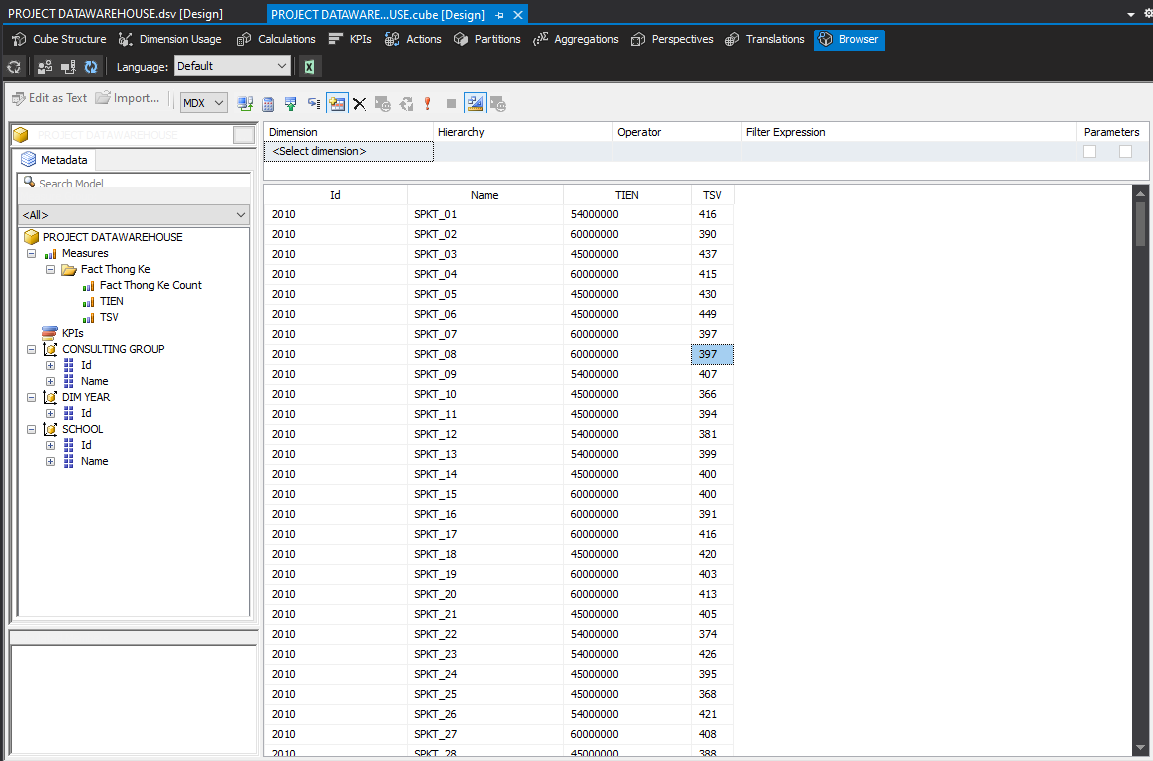


Thực hiện tiến trình process



Dữ liệu được load thành công

Danh sách tổng sinh viên và tổng số tiền của các đoàn qua từng năm



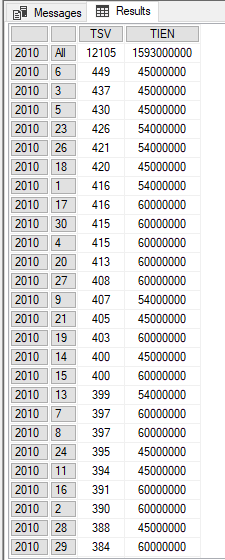
* **Dùng ngôn ngữ MDX trên SQL để truy vấn nhiều chiều:**

Danh sách tổng sinh viên từ cao đến thấp qua từng năm của tất cả các đoàn

SELECT {Measures.[TSV],Tien} on columns,

non empty order ([Dim year].[id].[id]\*[CONSULTING GROUP].[id].members, Measures.[TSV], DESC)on rows

FROM [PROJECT DATAWAREHOUSE]



Dữ liệu được sắp xếp theo chiều giảm dần của tổng số sinh viên để dễ quan sát.

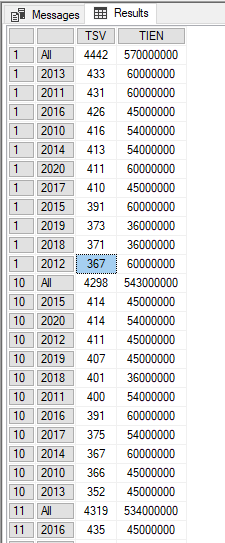
Qua đây ta có thể đưa ra phỏng đoán: Số tiền mà mỗi đoàn được cấp không ảnh hưởng đến mà tổng sinh viên của tất cả các đoàn qua mỗi năm

Danh sách tổng sinh viên từ cao đến thấp của từng đoàn qua tất cả các năm:

SELECT {Measures.[TSV],Tien} on columns,

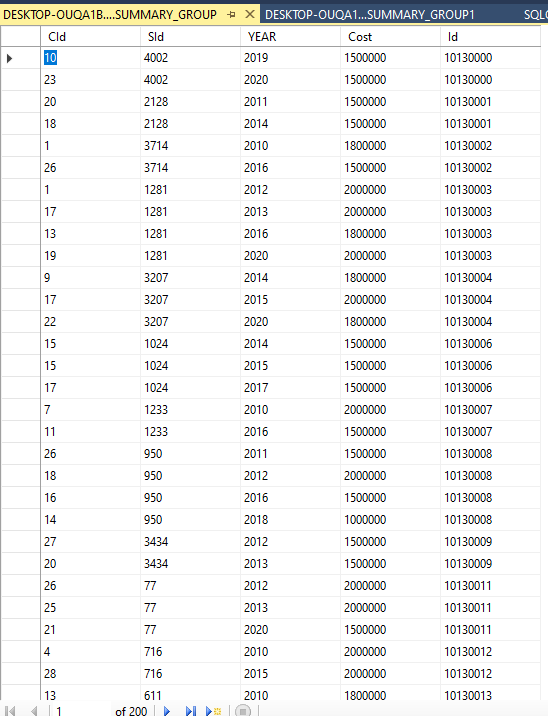
non empty order ([CONSULTING GROUP].[id].[id]\*[Dim year].[id].members, Measures.[TSV], DESC)on rows

FROM [PROJECT DATAWAREHOUSE]



* + 1. **Khai thác sử dụng phân hệ DWH dùng các vấn tin SQL**
* **TẠO CÁC VIEW ĐỂ PHỤC VỤ CHO VIỆC VẤN TIN SQL:**

VIEW SUMMARY\_GROUP: Danh sách của tất cả sinh viên được các đoàn tuyển sinh qua các năm bao gồm(Mã trường, Mã đoàn, Năm, Cost, Id).



VIEW\_SUMMARY\_GROUP1: DANH SÁCH TỔNG SINH VIÊN VÀ SỐ TIỀN CỦA MỖI ĐOÀN TẠI CÁC TRƯỜNG QUA CÁC NĂM.

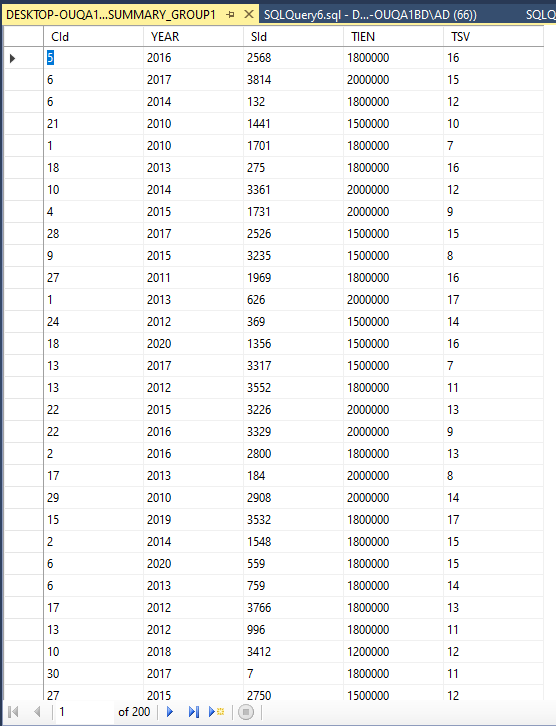
CREATE VIEW [dbo].[SUMMARY\_GROUP1] AS

SELECT CId, YEAR,SId, AVG(SUMMARY\_GROUP.Cost) AS TIEN, COUNT(SUMMARY\_GROUP.Id) AS TSV

FROM SUMMARY\_GROUP

GROUP BY CId, YEAR,SId

GO



* **DÙNG ROOL UP VÀ CUBE VÀ GROUPING SET ĐÁNH GIÁ SƠ LƯỢC ĐỂ CÓ CÁI NHÌN TỐT NHẤT VỀ TẬP DỮ LIỆU**

DANH SÁCH TỔNG TIỀN CỦA TỪNG ĐOÀN QUA TẤT CÁC NĂM (ROOLUP)

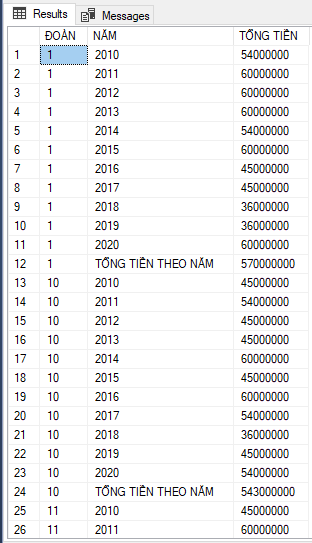
SELECT CASE WHEN GROUPING(SUMMARY\_GROUP1.CId) = 1 THEN N'TỔNG TIỀN THEO ĐOÀN' ELSE SUMMARY\_GROUP1.CId END AS N'ĐOÀN',

CASE WHEN GROUPING(SUMMARY\_GROUP1.YEAR) = 1 THEN N'TỔNG TIỀN THEO NĂM' ELSE SUMMARY\_GROUP1.YEAR END AS N'NĂM',

SUM(SUMMARY\_GROUP1.TIEN) AS N'TỔNG TIỀN'

FROM SUMMARY\_GROUP1

GROUP BY ROLLUP(SUMMARY\_GROUP1.CId, SUMMARY\_GROUP1.YEAR)



DANH SÁCH TỔNG SINH VIÊN CỦA TỪNG ĐOÀN QUA CÁC NĂM(ROLLUP)

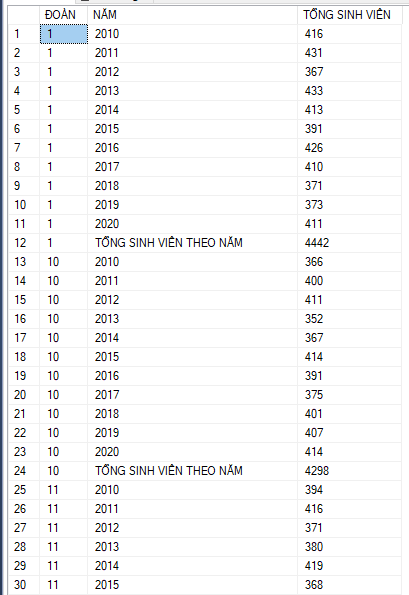
SELECT CASE WHEN GROUPING(SUMMARY\_GROUP.CID) = 1 THEN N'TỔNG TIỀN THEO ĐOÀN' ELSE SUMMARY\_GROUP.CId END AS N'ĐOÀN',

CASE WHEN GROUPING(SUMMARY\_GROUP.YEAR) = 1 THEN N'TỔNG SINH VIÊN THEO NĂM' ELSE SUMMARY\_GROUP.YEAR END AS N'NĂM',

COUNT(SUMMARY\_GROUP.Id) AS N'TỔNG SINH VIÊN'

FROM SUMMARY\_GROUP

GROUP BY ROLLUP(SUMMARY\_GROUP.CId, SUMMARY\_GROUP.YEAR)



DANH SÁCH TỔNG TIỀN CỦA TẤT CẢ ĐOÀN QUA TỪNG NĂM(CUBE)

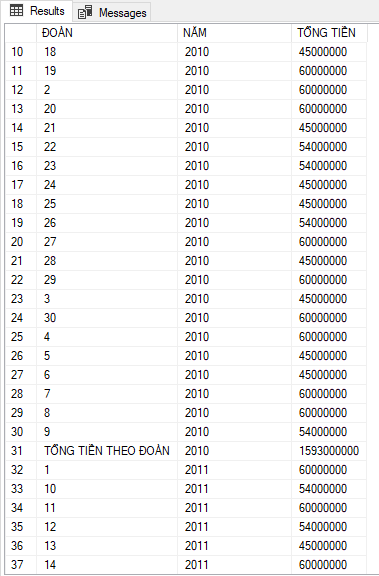
SELECT CASE WHEN GROUPING(SUMMARY\_GROUP1.CId) = 1 THEN N'TỔNG TIỀN THEO ĐOÀN' ELSE SUMMARY\_GROUP1.CId END AS N'ĐOÀN',

CASE WHEN GROUPING(SUMMARY\_GROUP1.YEAR) = 1 THEN N'TỔNG TIỀN THEO NĂM' ELSE SUMMARY\_GROUP1.YEAR END AS N'NĂM',

SUM(SUMMARY\_GROUP1.TIEN) AS N'TỔNG TIỀN'

FROM SUMMARY\_GROUP1

GROUP BY CUBE(SUMMARY\_GROUP1.CId, SUMMARY\_GROUP1.YEAR)



DANH SÁCH TỔNG SINH VIÊN CỦA CÁC ĐOÀN QUA TỪNG NĂM(CUBE)

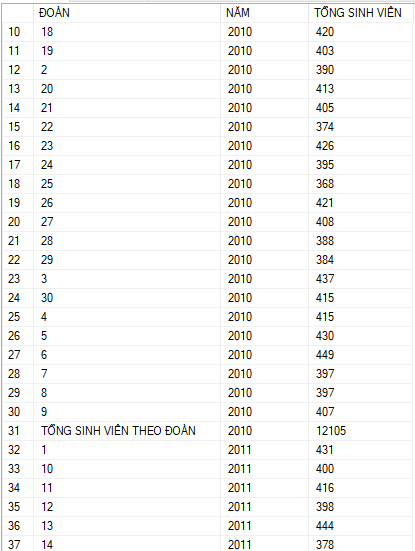
SELECT CASE WHEN GROUPING(SUMMARY\_GROUP.CID) = 1 THEN N'TỔNG SINH VIÊN THEO ĐOÀN' ELSE SUMMARY\_GROUP.CId END AS N'ĐOÀN',

CASE WHEN GROUPING(SUMMARY\_GROUP.YEAR) = 1 THEN N'TỔNG SINH VIÊN THEO NĂM' ELSE SUMMARY\_GROUP.YEAR END AS N'NĂM',

COUNT(SUMMARY\_GROUP.Id) AS N'TỔNG SINH VIÊN'

FROM SUMMARY\_GROUP

GROUP BY CUBE(SUMMARY\_GROUP.CId, SUMMARY\_GROUP.YEAR)



TỔNG TIỀN (GROUPING SETS)

ALTER VIEW T3 AS

SELECT CASE WHEN GROUPING(SUMMARY\_GROUP1.CId) = 1 THEN N'TỔNG TIỀN THEO ĐOÀN' ELSE SUMMARY\_GROUP1.CId END AS N'ĐOÀN',

CASE WHEN GROUPING(SUMMARY\_GROUP1.YEAR) = 1 THEN N'TỔNG TIỀN THEO NĂM' ELSE SUMMARY\_GROUP1.YEAR END AS N'NĂM',

SUM(SUMMARY\_GROUP1.TIEN) AS N'TỔNG TIỀN'

FROM SUMMARY\_GROUP1

GROUP BY GROUPING SETS (

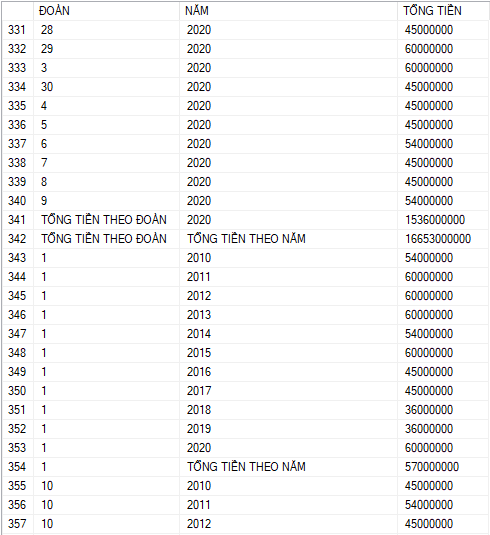
(SUMMARY\_GROUP1.CId, SUMMARY\_GROUP1.YEAR) ,

(SUMMARY\_GROUP1.YEAR, SUMMARY\_GROUP1.CId),

(SUMMARY\_GROUP1.CId),

(SUMMARY\_GROUP1.YEAR),

())



TỔNG SINH VIÊN

ALTER VIEW T4 AS

SELECT CASE WHEN GROUPING(SUMMARY\_GROUP.CId) = 1 THEN N'TỔNG SINH VIÊN THEO ĐOÀN' ELSE SUMMARY\_GROUP.CId END AS N'ĐOÀN',

CASE WHEN GROUPING(SUMMARY\_GROUP.YEAR) = 1 THEN N'TỔNG SINH VIÊN THEO NĂM' ELSE SUMMARY\_GROUP.YEAR END AS N'NĂM',

COUNT(SUMMARY\_GROUP.Id) AS N'TỔNG SINH VIÊN'

FROM SUMMARY\_GROUP

GROUP BY GROUPING SETS (

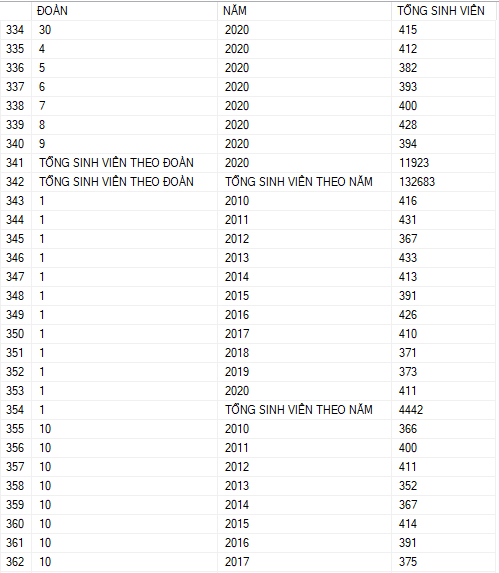
(SUMMARY\_GROUP.CId, SUMMARY\_GROUP.YEAR),

(SUMMARY\_GROUP.YEAR, SUMMARY\_GROUP.CId),

(SUMMARY\_GROUP.CId),

(SUMMARY\_GROUP.YEAR),

())

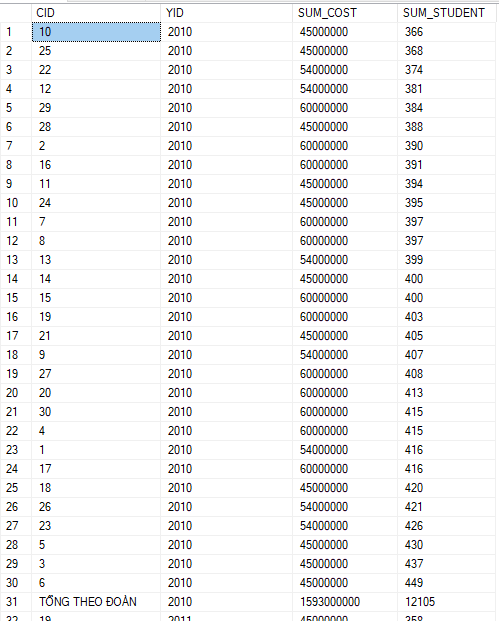


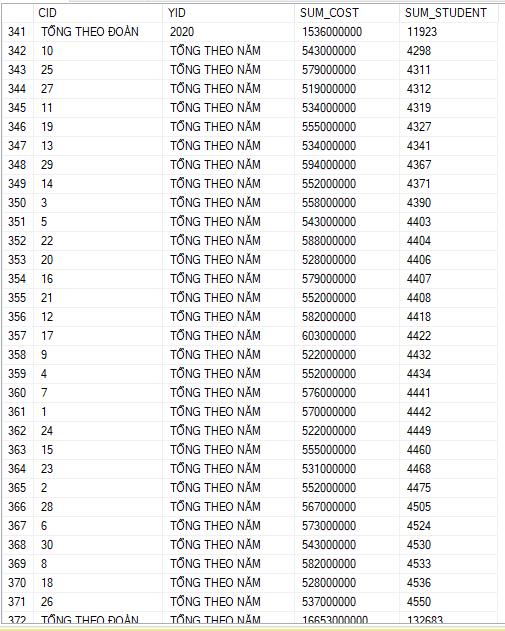
TẠO VIEW T5 ĐỂ CÓ ĐƯỢC DANH SÁCH TỔNG SINH VIÊN VÀ TỔNG TIỀN CỦA TẤT CÁC ĐOÀN QUA CÁC NĂM

ALTER VIEW T5 AS

SELECT DISTINCT T3.CID,T3.YID,T3.SUM\_COST,T4.SUM\_STUDENT

FROM T3 INNER JOIN T4 ON T3.CID=T4.CID AND T3.YID=T4.YID





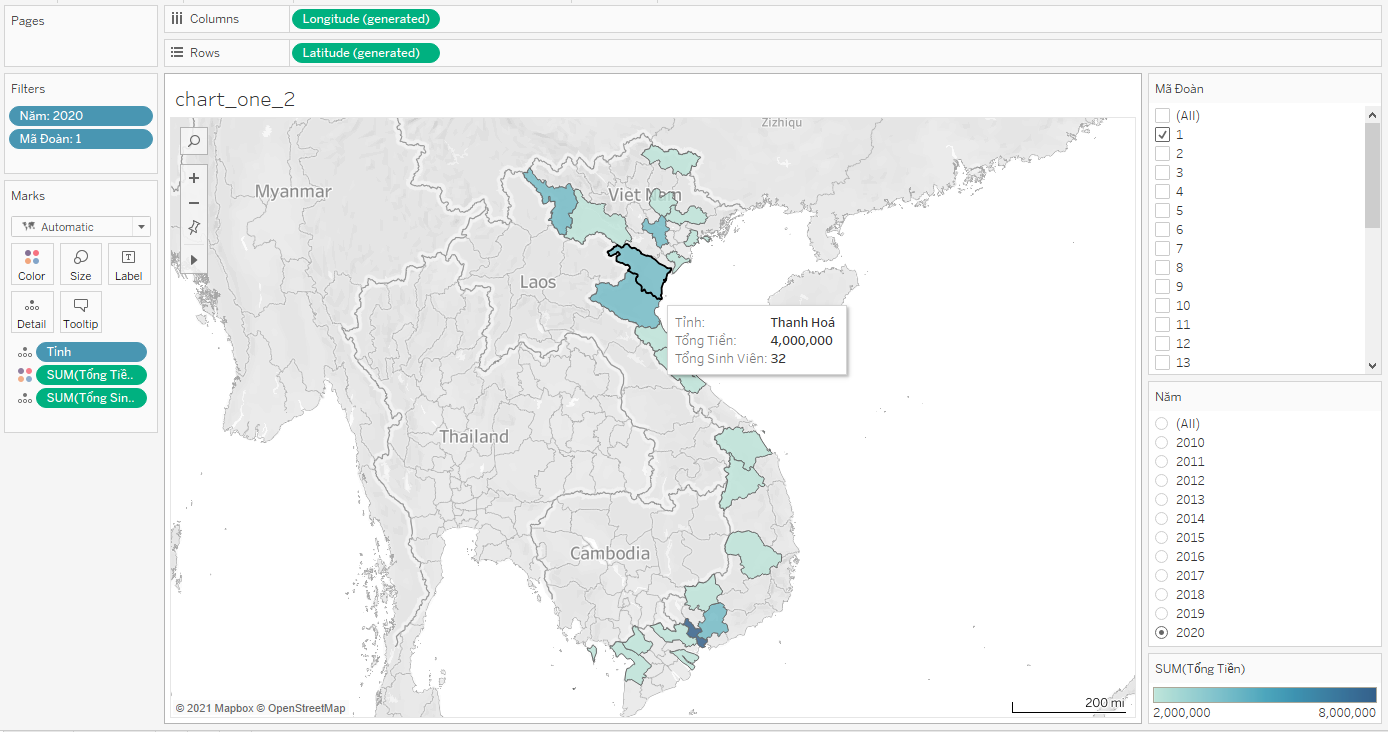
* + 1. **Sử dụng Tool Tableau để đưa dữ liệu lên các chart**

Sử dụng phần mềm Tableau để vẽ các chart: gồm 3 chart

Qua biểu đồ sẽ thể hiện: Tổng số tiền có ảnh hưởng đến tổng số sinh viên hay không và ngược lại, bên cạnh đó ta thấy được tổng số tiền và tổng số sinh viên của mỗi đoàn ở mỗi tỉnh qua các năm.

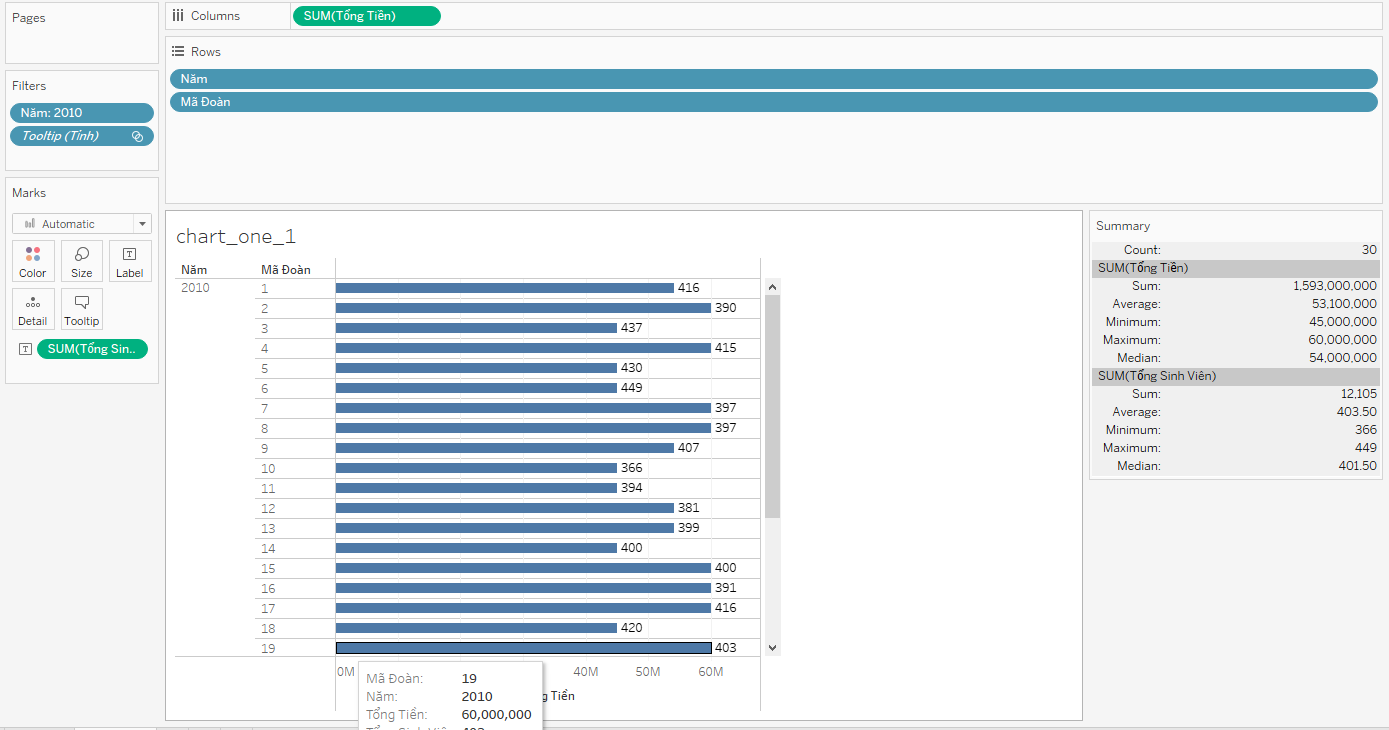
+ Chart 1: Sơ lược về tổng sinh viên và số tiền của các đoàn ở mỗi tỉnh qua các năm, giúp người sử dụng có cái tổng quan hơn về dữ liệu.

Ví dụ: Năm 2010 thì đoàn số 1 tuyển được 32 sinh viên ở tỉnh Thanh Hóa với chi phí là 4000000.

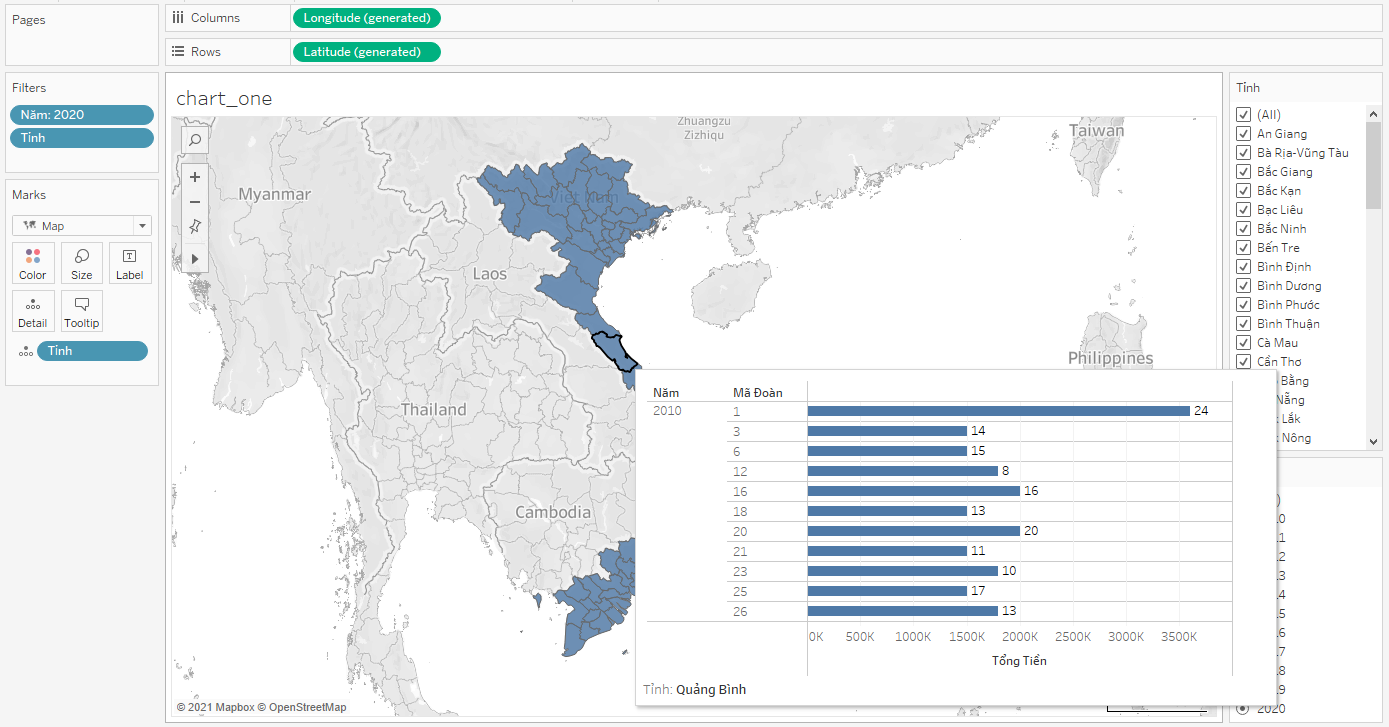


+ Chart 2:Bằng mắt thường có thể thấy được giữa tổng tiền và tổng sinh viên của mỗi đoàn không có sự liên quan với nhau. Số tiền càng lớn không đồng nghĩa là tổng số sinh viên được tuyển vào sẽ cao và ngược lại.

Vi dụ: Trong biểu đồ ta thấy năm 2010 thì tổng số tiền của đoàn 1 là 54000000 nhỏ hơn đoàn 2 là 60000000 những đoàn 1 lại có tổng sinh viên cao hơn là 416 sinh viên so với 390 sinh viên.



+ Chart 3: Tổng sinh viên và tổng tiền của mỗi đoàn qua mỗi năm trên bản đồ Việt Nam, việc ghép 2 bản đồ lại với nhau giúp người sử dụng có thể thấy được qua mỗi năm thì tổng số sinh viên và tổng tiền được biểu diễn ra sao.



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

## Kết luận

### Những kết quả đạt được

ĐỒ ÁN HỌC PHẦN đã giải quyết được vấn đề:

- Biết được chất lượng của sinh viên ở khu vực nào tốt hơn khu vực nào.

- Biết được khu vực nào thường học những ngành nào

- Biết được xu thế chọn nghề nghiệp của sinh viên hiện nay

- Biết được chất lượng tuyển sinh của các đoàn tuyển sinh được phân công

- Từ những ý trên, các trường Đại Học có thể đưa ra phương án tuyển sinh hợp lý hơn trong các kỳ tuyển sinh

### Hạn chế

- Dữ liệu được sử dụng trong bảng Student chính là dữ liệu được tạo ngẫu nhiên, tuy nhiên dữ liệu các bảng khác được tham khảo 90-95% dữ liệu thực tế của trường ĐH Sư phạm Kỹ thuật TP. HCM nên kết quả phân tích không bị lệch xa so với thực tế.

## Hướng phát triển

### Hướng khắc phục các hạn chế

- Khi đưa dự án vào thực tế thì vấn đề về dữ liệu sẽ được giải quyết vì lúc đó có thể sử dụng dữ liệu trực tiếp tại trường sử dụng dự án.

### Hướng mở rộng ĐỒ ÁN HỌC PHẦN

+ Thay vì sử dụng bộ nhớ ổ cứng một máy tính, ta có thể lưu phân tán dữ liệu khi dữ liệu quá nhiều bằng Framework Hadoop Distributed File System để tăng khả năng chị lỗi cũng như bảo mật dữ liệu.

+ Sử dụng Framework Spark và Framework HiveQL để tạo các Data Pipeline giúp dữ liệu trong DWH được cập nhật một cách tuần tự và nhanh chóng.

# 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).

# CÁC PHỤ LỤC

(nếu có)

# 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 phân hệ cụ thể cho từng thành viên thuộc nhóm  + Xây dựng và thực nghiệm Hệ thống Kho dữ liệu theo nhóm và phân hệ 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 |