

BỘ GIÁO DỤC ĐÀO TẠO
TRƯỜNG ĐẠI HỌC SƯ PHẠM VÀ KỸ THUẬT TP. HCM
KHOA CÔNG NGHỆ THÔNG TIN



HCMUTE

**BÁO CÁO CUỐI KỲ
Môn học: Kho dữ liệu**

Đề tài:

**THIẾT KẾ KHO DỮ LIỆU TỶ LỆ TỐT NGHIỆP
CỦA CÁC TRƯỜNG ĐẠI HỌC Ở MỸ**

Mã lớp học phần: DAWH430784_23_2_03CLC

Giảng viên hướng dẫn: Ths. Nguyễn Văn Thành

Sinh viên thực hiện: Nguyễn Phú Thành 21110299

Cao Thị Ngọc Phụng 21110276

Nguyễn Văn Hào 21110175

Tào Việt Đức 21110169

TP Hồ Chí Minh, tháng 5 năm 2024

DANH SÁCH SINH VIÊN THAM GIA

Mã học phần: DAWH430784_23_2_03CLC

Nhóm: 01

Tên đề tài: Thiết kế kho dữ liệu tỷ lệ tốt nghiệp của các trường Đại học ở Mỹ

STT	HỌ VÀ TÊN THÀNH VIÊN	MÃ SỐ SINH VIÊN	TỶ LỆ THAM GIA
1	Nguyễn Phú Thành	21110299	100%
2	Cao Thị Ngọc Phụng	21110276	100%
3	Nguyễn Văn Hào	21110175	100%
4	Tào Việt Đức	21110169	100%

Ghi chú:

Tỷ lệ %: Mức độ phần trăm hoàn thành của từng sinh viên tham gia.

Trưởng nhóm: Nguyễn Phú Thành

Điểm số:

Nhận xét của giáo viên

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp. Hồ Chí Minh - Tháng 5 năm 2024

PHỤ LỤC

1. BẢNG PHÂN CÔNG NHIỆM VỤ

Nhiệm vụ	Phú Thành	Văn Hào	Ngọc Phụng	Việt Đức
Tìm kiếm dữ liệu	X			
Xử lý dữ liệu	Xử lý về Carnegie	Xử lý về Institution Grads	Xử lý State Sector	Xử lý về Institution Detail
Xác định business process		Institution Detail		Institution Graduation
Xác định bảng dim	Dim Carnegie	Dim Location	Dim Cohort	Dim State
Xác định bảng fact		FactDetail		FactGraduation
Đẩy dữ liệu từ CSV – SQL Server	X	X	X	X
Tạo nguồn kết nối dữ liệu	X	X	X	X
Staging và load và các dim, fact	X	X	X	X
Nhập dữ liệu vào SSAS, tạo Data Cube	X	X	X	X
Phân tích SSAS	X		X	
Đặt 10 câu hỏi	X	X	X	X

Trả lời bằng SSAS	X		X	
Trả lời bằng Pivot Table		X		X
Report Power BI	X			

MỤC LỤC

PHỤ LỤC.....	1
1. BẢNG PHÂN CÔNG NHIỆM VỤ	1
LỜI CẢM ƠN.....	1
PHẦN MỞ ĐẦU	2
1. Lời mở đầu	2
2. Mục đích, nhiệm vụ của đề tài	2
3. Phạm vi thực hiện đề tài	2
PHẦN NỘI DUNG.....	3
CHƯƠNG 1: TỔNG QUAN VỀ DỮ LIỆU	3
1.1 Nguồn dữ liệu.....	3
1.2 Mô tả chi tiết	3
1.2.1 Mô tả dữ liệu	3
1.2.2 Thông số chi tiết	6
1.3 Công cụ sử dụng.....	9
CHƯƠNG 2: THIẾT KẾ XÂY DỰNG CƠ SỞ DỮ LIỆU TÁC NGHIỆP	10
2.1 Thiết kế nghiệp vụ.....	10
2.2 Thiết kế các bảng Dimension.....	10
2.2.1 DimState	10
2.2.2 DimCohort.....	11
2.2.3 DimLocation	11
2.2.4 DimCarnegie	11
2.3 Thiết kế bảng Fact.....	12
2.3.1 Fact Institution Details.....	12
2.3.2 Fact Institution Graduation.....	13
2.4 Thiết kế và tiền xử lý dữ liệu	14
2.4.1 Lược đồ hình bông tuyết.....	14
2.4.2 Tiền xử lý dữ liệu.....	14
CHƯƠNG 3: TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS)	21
3.1 Tạo nguồn dữ liệu từ CSV	21

3.2 Tạo College_Staging và project SSIS	25
3.3 Tạo kết nối nguồn và đích dữ liệu.....	28
3.4 Tiến hành Staging	29
3.5 Tiến hành Load	42
CHƯƠNG 4: PHÂN TÍCH DỮ LIỆU (SSAS)	61
4.1 Câu hỏi truy vấn.....	61
4.2 Xây dựng mô hình.....	62
4.3 Tạo Data Source View và Cube	66
4.4 Xây dựng Dim và Hierarchy	75
4.4.1 DimCohort.....	75
4.4.2 Dim (Fact) Detail	80
4.5 Trả lời câu hỏi Truy vấn.....	85
4.5.1 Câu 1: Tỉ lệ sinh viên theo học chương trình full time tại các khu vực	85
4.5.2 Câu 2: Tỉ lệ sinh viên nhận được trợ cấp Pell tại mỗi khu vực.....	86
4.5.3 Câu 2: Tỉ lệ sinh viên nhận được trợ cấp Pell tại mỗi khu vực.....	87
4.5.3 Câu 3:Tỉ lệ tốt nghiệp đúng hạn theo thời gian học tại các khu vực	88
4.5.4 Câu 4:Tỉ lệ sinh viên được giữ lại sau năm học thứ nhất	89
4.5.5 Câu 5: Tổng số sinh viên và chi phí ước tính cho mỗi giải thưởng học thuật và số lượng nhân viên trường đó	90
4.5.6 Câu 6: Số lượng sinh viên tốt nghiệp theo sắc tộc và giới tính.....	91
4.5.7 Câu 7: Số lượng sinh viên tốt nghiệp đúng hạn của từng khu vực.....	92
4.5.8 Câu 8: Số lượng sinh viên lấy bằng sau 150% thời gian học tiêu chuẩn theo giới tính và sắc tộc của từng năm.....	93
4.5.9 Câu 9: Số lượng sinh viên tốt nghiệp theo từng năm, phân theo loại trường học và cohort	94
4.5.10 Câu 10: Số lượng sinh viên tốt nghiệp theo sắc tộc và giới tính, phân theo quy mô chương trình học.....	95
CHƯƠNG 5: PHÂN TÍCH BẰNG POWER BI	96
5.1 Report về vị trí địa lý của các cơ sở Đại học	96
5.2 Report về tỷ lệ sinh viên theo học chương trình Full-time ở Trường Đại học	97

5.3 Report về tỷ lệ sinh viên nhận trợ cấp Pell ở Trường Đại học.....	98
PHẦN KẾT LUẬN	99
1. Ưu điểm	99
2. Nhược điểm.....	99
3. Hướng phát triển.....	99
TÀI LIỆU THAM KHẢO	1

LỜI CẢM ƠN

Lời đầu tiên, nhóm em xin gửi lời cảm ơn chân thành nhất đến *Thầy Nguyễn Văn Thành* – giảng viên bộ môn *Data WareHouse* của chúng em. Trong quá trình học tập và tìm hiểu bộ môn, chúng em đã nhận được sự quan tâm giúp đỡ, hướng dẫn rất tận tình và tâm huyết từ Thầy. Thầy đã giúp chúng em tích lũy thêm nhiều kiến thức để có cái nhìn sâu sắc và hoàn thiện hơn trong lĩnh vực Công nghệ thông tin. Để từ đó, ứng dụng những kiến thức mà Thầy truyền tải, nhóm em xin trình bày lại những gì mà mình đã học hỏi được thông qua việc thực hiện đề tài “*Thiết kế Kho dữ liệu tỷ lệ tốt nghiệp của các Trường Đại học ở Mỹ*”.

Kiến thức là vô hạn và sự tiếp nhận kiến thức của bản thân mỗi người luôn tồn tại những hạn chế nhất định. Do đó, trong phạm vi khả năng của bản thân, nhóm em đã rất cố gắng để hoàn thành đề tài một cách tốt nhất. Tuy nhiên, chắc chắn không tránh khỏi những thiếu sót, nhóm chúng em rất mong nhận được sự cảm thông và những ý kiến đóng góp đến từ Thầy để đề tài của nhóm em được hoàn thiện hơn.

Một lần nữa, nhóm em xin chân thành cảm ơn Thầy đã tận tình hướng dẫn, chỉ bảo các thành viên nhóm em trong suốt quá trình học tập và thực hiện đồ án này.

Kính chúc Thầy sức khỏe, hạnh phúc thành công trên con đường sự nghiệp giảng dạy.

Trân trọng
Đại diện nhóm
Nguyễn Phú Thành

PHẦN MỞ ĐẦU

1. Lời mở đầu

Nổi bật với sự đa dạng và cơ hội, nước Mỹ chính là điểm đến lý tưởng cho du học sinh quốc tế. Nơi đây sở hữu nền văn hóa, xã hội và kinh tế phong phú, tạo điều kiện cho du học sinh phát triển toàn diện. Hệ thống giáo dục Mỹ được đánh giá cao với chất lượng đào tạo xuất sắc và cơ hội nghề nghiệp rộng mở. Nổi bật trong số đó là hệ thống các trường đại học danh tiếng, nơi du học sinh được tiếp cận với chương trình giảng dạy tiên tiến, tài liệu giáo dục cập nhật cùng phương pháp giảng dạy đa dạng và hiệu quả. Nhờ sự tận tâm của đội ngũ giảng viên, du học sinh có cơ hội phát huy tối đa tiềm năng và gặt hái thành công trong tương lai. Với mong muốn hỗ trợ du học sinh lựa chọn ngôi trường phù hợp nhất, bản báo cáo thống kê về các trường đại học tại Mỹ đã ra đời. Báo cáo cung cấp thông tin chi tiết về chất lượng đào tạo, chi phí học tập và nhiều tiêu chí đánh giá quan trọng khác, giúp du học sinh đưa ra quyết định sáng suốt cho hành trình du học của mình.

2. Mục đích, nhiệm vụ của đề tài

Xây dựng và phát triển ứng dụng nhằm phục vụ việc phân tích, khai thác, nhằm nắm rõ xu hướng của các trường đại học và độ hiệu quả đến đến các nhóm đối tượng khác nhau.

Hướng tới đối tượng sử dụng là: các học sinh, sinh viên có nhu cầu qua du học tại các trường đại học ở Mỹ, những người quan tâm đến vấn đề giáo dục ở Mỹ

3. Phạm vi thực hiện đề tài

Đề tài này thực hiện chủ yếu các lý thuyết cơ bản về xây dựng kho dữ liệu. Áp dụng các kiến thức về phân tích dữ liệu và xây dựng các nghiệp vụ cần thiết, mô hình hóa nghiệp vụ, tích hợp dữ liệu vào kho (SSIS), tiến hành phân tích SSAS và trả lời các câu hỏi bằng SSAS, Pivot Excel và Power BI.

PHẦN NỘI DUNG

CHƯƠNG 1: TỔNG QUAN VỀ DỮ LIỆU

1.1 Nguồn dữ liệu

Dữ liệu được sử dụng để thiết kế và đầy vào kho được lấy từ trang DataWorld:

College Completion

Dữ liệu gốc là từ College Completion microsite được cung cấp bởi The Chronicle of Higher Education với sự hỗ trợ bởi Bill & Melinda Gates Foundation..

1.2 Mô tả chi tiết

1.2.1 Mô tả dữ liệu

Các Trường Cao Đẳng và Đại Học tại Mỹ

College Completion (Hoàn thành Cao đẳng/Đại học) nghiên cứu dữ liệu và xu hướng tại 3.800 cơ sở cấp bằng ở Mỹ (không bao gồm vùng lãnh thổ) đáp ứng các tiêu chí sau:

- + Theo dõi nhóm sinh viên theo học lần đầu, toàn thời gian và hướng tới bằng cử nhân bậc đại học.
- + Tổng số sinh viên bậc đại học đạt tối thiểu 100 người vào năm 2013. Cấp bằng cử nhân trong khoảng thời gian từ 2011 đến 2013.
- + Báo cáo cũng bao gồm các trường cao đẳng và đại học đáp ứng cùng tiêu chí vào năm 2010.

Tỷ lệ Tốt nghiệp

+ NCES/IPEDS: Dữ liệu tốt nghiệp từ Hệ thống Giáo dục Sau Trung học Tích hợp của Trung tâm Thông kê Giáo dục Quốc gia (NCES) chỉ giới hạn trong việc theo dõi tỷ lệ hoàn thành chương trình của nhóm sinh viên theo học lần đầu, toàn thời gian và hướng tới bằng cử nhân bậc đại học.

+ Nhóm sinh viên được nghiên cứu thường bắt đầu học đại học 6 năm trước tại các trường bốn năm và 3 năm trước tại các trường hai năm. Các trường báo cáo số lượng sinh viên hoàn thành chương trình trong khoảng thời gian 100% và 150% so với thời gian chuẩn. Đối với sinh viên theo học bằng cử nhân hoặc tương đương, điều này tương ứng với việc tốt nghiệp trong vòng bốn năm và sáu năm.

+ Các cơ sở báo cáo dữ liệu này dựa trên thời gian đào tạo của chương trình, có thể khác nhau đối với sinh viên theo học các loại bằng hoặc chứng chỉ khác.

+ Điều quan trọng cần lưu ý là dữ liệu tốt nghiệp không bao gồm thông tin về sinh viên bỏ học và quay lại học hoặc hoàn thành bằng cấp ở nơi khác.

+ Hệ thống Trách nhiệm Tự nguyện (VSA): Thông kê về tốt nghiệp và tiếp tục theo học đến từ tỷ lệ Thành công và Tiến bộ của Sinh viên thuộc Hệ thống Trách nhiệm Tự nguyện. Các con số dựa trên dữ liệu nhóm sinh viên từ Trung tâm Thanh toán Sinh viên Quốc gia cho sinh viên theo học lần đầu, toàn thời gian và sinh viên chuyển tiếp toàn thời gian. Khoảng một nửa số cơ sở công lập bốn năm được khảo sát đã báo cáo dữ liệu này. Vì được liên kết với từng sinh viên thay vì một cơ sở, các con số này có thể cho thấy kết quả cho sinh viên nhập học và rời khỏi một trường cao đẳng/đại học cụ thể theo cách mà tỷ lệ tốt nghiệp truyền thống không thể.

+ Mặc dù có những điểm tương đồng giữa các biện pháp hoàn thành lần đầu, toàn thời gian đến từ IPEDS và các biện pháp tồn tại như một phần của VSA, tỷ lệ này sẽ không bao giờ hoàn toàn khớp nhau do các lý do bao gồm sự khác biệt về khung thời gian được kiểm tra và nhóm sinh viên chính xác - các cơ sở không bị phạt đối với sinh viên không thể xác định thành công từ hồ sơ của Trung tâm Thanh toán Sinh viên Quốc gia.

Chủng tộc và Dân tộc

+ Cho đến năm 2009, NCES phân loại học sinh theo bảy cách: Da trắng, không phải gốc Tây Ban Nha; Da đen, không phải gốc Tây Ban Nha; Người Mỹ bản địa/Người Alaska bản địa; Người châu Á/Người Thái Bình Dương; Chủng tộc hoặc dân tộc không rõ; và Không cư trú. Bên cạnh việc tạo ra sự tách biệt mạnh mẽ hơn giữa các danh mục chủng tộc và dân tộc, hai danh mục chủng tộc mới đã được tạo ra: Người Hawaii bản địa hoặc Người Thái Bình Dương khác (trước đây được kết hợp với sinh viên châu Á) và sinh viên thuộc hai hoặc nhiều chủng tộc.

+ Phân loại chủng tộc mới không được tất cả các cơ sở áp dụng - họ được lựa chọn sử dụng các danh mục cũ, danh mục mới hoặc kết hợp cả hai.

+ Do đó, dữ liệu tốt nghiệp theo chủng tộc kể từ năm 2008 có thể bị ảnh hưởng bởi những khác biệt trong phân loại này.

+ Để khớp với các năm trước, Người Hawaii bản địa hoặc Người Thái Bình Dương khác đã được kết hợp với người châu Á. Sinh viên được báo cáo thuộc hai hoặc nhiều chủng tộc, không cư trú hoặc không rõ được bao gồm trong tổng số nhưng không được hiển thị riêng.

+ Tại các trường hai năm, số lượng sinh viên tốt nghiệp trong vòng 100% thời gian chuẩn không có sẵn trước năm 2009 và không bao giờ được phân theo giới tính

Hiệu quả Hoạt động

+ Số giải thưởng/100 sinh viên đại học toàn thời gian" bao gồm tất cả các bằng đại học, cao đẳng và chứng chỉ dưới 4 năm do cơ sở báo cáo cho NCES.

+ Số sinh viên đại học tương đương toàn thời gian được ước tính dựa trên số tín chỉ học tại cơ sở trong một năm học.

+ Doanh thu giáo dục bao gồm chi phí trực tiếp và gián tiếp cho giảng dạy, dịch vụ sinh viên, hỗ trợ học tập, hỗ trợ cho cơ sở, hoạt động và bảo trì.

+ Dữ liệu chi tiêu được báo cáo cho tất cả sinh viên, bao gồm cả sinh viên cao học

Table 1. New York State Awards per 100 FTE, all sectors, AY 2008-2009

2008-2009	Certificates	Associates	Bachelor	Undergrad FTE	Awards per 100 FTE
New York	4,616	59,048	122,186	875,045	21.2

$$\text{Awards per 100 FTE} = (4,616 + 59,048 + 122,186) / 875,045 = .212 * 100 = \mathbf{21.2}$$

1.2.2 Thông số chi tiết

Cc_institution_detail.csv

unitid	chromname	city	state	level	control	basic	hbcu	flagship	long_x								
100654	Alabama A&M University	Normal	Alabama	4-year	Public	Masters Colleges and Universities--larger programs	X	NULL	-86.56850								
100663	University of Alabama at Birmingham	Birmingham	Alabama	4-year	Public	Research Universities--very high research activity	NULL	NULL	-86.8091								
100670	Auburn University	Montgomery	Alabama	4-year	Private not-for-profit	Master's Colleges and Universities--medium programs	NULL	NULL	-86.1741								
100706	University of Alabama at Huntsville	Huntsville	Alabama	4-year	Public	Research Universities--very high research activity	NULL	NULL	-86.6268								
100724	Alabama State University	Montgomery	Alabama	4-year	Public	Master's Colleges and Universities--larger programs	X	NULL	-86.2956								
100751	University of Alabama at Tuscaloosa	Tuscaloosa	Alabama	4-year	Public	Research Universities--high research activity	NULL	X	-87.54571								
100760	Central Alabama Community College	Alexander City	Alabama	2-year	Public	Associates--Public Rural-serving Medium	NULL	NULL	-85.9461								
100830	Auburn University at Montgomery	Montgomery	Alabama	4-year	Public	Masters Colleges and Universities--larger programs	NULL	NULL	-86.17731								
100858	Auburn University	Auburn University	Alabama	4-year	Public	Research Universities--high research activity	NULL	NULL	-85.49241								
100937	Birmingham-Southern College	Birmingham	Alabama	4-year	Private not-for-profit	Baccalaureate Colleges--Arts & Sciences	NULL	NULL	-86.8536								
101001	Georgia Gwinnett College	Peachtree City	Alabama	2-year	Public	Associates--Public Suburban-serving Medium	NULL	NULL	-86.0234								
101073	Concordia College (Ala.)	Seima	Alabama	4-year	Private not-for-profit	Baccalaureate Colleges--Diverse Fields	X	NULL	-87.0235								
101116	South University at Montgomery	Montgomery	Alabama	4-year	Private for-profit	Baccalaureate/Associates Colleges	NULL	NULL	-86.21641								
101143	Enterprise State Community College	Enterprise	Alabama	2-year	Public	Associates--Public Rural-serving Medium	NULL	NULL	-85.83691								
101161	Faulkner State Community College	Bay Minette	Alabama	2-year	Public	Associates--Public Suburban-serving Multicampus	NULL	NULL	-87.77971								
101189	Faulkner University	Gadsden	Alabama	4-year	Private not-for-profit	Baccalaureate Colleges--Diverse Fields	NULL	NULL	-86.216								
101226	Golden State Community College	Dobson	Alabama	2-year	Public	Associates--Public Suburban-serving Large	X	NULL	-85.99121								
101286	Georgia Wallace Community College at Dothan	Hanceville	Alabama	2-year	Public	Associates--Public Rural-serving Medium	NULL	NULL	-86.7817								
101301	Wallace Community College at Selma	Selma	Alabama	2-year	Public	Associates--Public Rural-serving Medium	NULL	NULL	-87.013								
101365	Herzing University at Birmingham	Birmingham	Alabama	4-year	Private for-profit	Baccalaureate Colleges--Diverse Fields	NULL	NULL	-86.83221								
101435	Huntingdon College	Montgomery	Alabama	4-year	Private not-for-profit	Baccalaureate Colleges--Diverse Fields	NULL	NULL	-86.2853								
101462	J.F. Drake State Technical College	Huntsville	Alabama	2-year	Public	Associates--Public Rural-serving Small	X	NULL	-86.5738								
101482	Jefferson State Community College	Jacksonville	Alabama	4-year	Public	Masters Colleges and Universities--large programs	NULL	NULL	-85.7666								
101505	Jefferson State Community College	Birmingham	Alabama	2-year	Public	Associates--Public Urban-serving Multicampus	NULL	NULL	-86.70731								
101514	Cahoon Community College	Tanner	Alabama	2-year	Public	Associates--Public Rural-serving Large	NULL	NULL	-86.9491								
101541	Judson College (Ala.)	Marion	Alabama	4-year	Private not-for-profit	Baccalaureate Colleges--Arts & Sciences	NULL	NULL	-87.3161								
101569	Lawson State Community College	Birmingham	Alabama	2-year	Public	Associates--Public Urban-serving Multicampus	X	NULL	-86.8902								
101587	University of West Alabama	Livinston	Alabama	4-year	Public	Masters Colleges and Universities--larger programs	NULL	NULL	-88.1860								
101602	Mountain View Institute	Andalusia	Alabama	2-year	Public	Associates--Public Rural-serving Small	NULL	NULL	-86.45220								
101675	Miles College	Marion	Alabama	4-year	Private	Associates--Public Urban-serving Small	X	NULL	-87.3179								
101693	University of Mobile	Fairfield	Alabama	4-year	Private not-for-profit	Baccalaureate Colleges--Diverse Fields	NULL	NULL	-86.98661								
101709	University of Montevallo	Mobile	Alabama	4-year	Public	Baccalaureate Colleges--Diverse Fields	NULL	NULL	-88.1289								
101736	Northwest-Shoals Community College	Montevallo	Alabama	4-year	Public	Masters Colleges and Universities--medium programs	NULL	NULL	-86.8650								
101879	University of North Alabama	Muscle Shoals	Alabama	2-year	Public	Associates--Public Rural-serving Medium	NULL	NULL	-87.6777								
101897	Northeast Alabama Community College	Florence	Alabama	4-year	Public	Masters Colleges and Universities--larger programs	NULL	NULL	-87.6809								
101912	Oakwood University	Rainsville	Alabama	2-year	Public	Associates--Public Rural-serving Medium	NULL	NULL	-85.91161								
34.78337	www.aamu.edu/	4051	14.2	18.8	21.5	105331	75743	66436									
33.50223	www.uab.edu	11502	20.9	18.8	21.5	136546	75743	66436									
32.36261	www.amridgeuniversity.edu	322	29.9	17.8	22.5	58414	92268	101725									
34.72282	www.uah.edu	5696	20.9	18.8	21.5	64418	75743	66436									
32.36432	www.alasu.edu/email/index.aspx	5356	11.6	18.8	21.5	132407	75743	66436									
33.2144	www.ua.edu/	29440	18.3	18.8	21.5	75350	75743	66436									
32.92443	www.cacc.edu	1906	15.9	15.9	16.5	57572	42194	37780									
32.36994	www.aum.edu	4322	15.4	18.8	21.5	58541	75743	66436									
32.6002	www.auburn.edu	19799	21.5	18.8	21.5	71999	75743	66436									
33.51545	www.bsc.edu/	1188	23.5	17.8	22.5	113677	92268	101725									
32.42391	www.cv.edu	1837	13.6	15.9	16.5	60303	42194	37780									
32.42443	www.ccal.edu/	600	12.3	17.8	22.5	105461	92268	101725									
32.34268	southuniversity.edu	599	15	29.1	24.6	76191	17406	38763									
31.2975	www.escc.edu	2333	17.4	15.9	16.5	45911	42194	37780									
30.85205	www.faulknerstate.edu	4362	11.8	15.9	16.5	55426	42194	37780									
32.38418	www.faulkner.edu	2617	21.5	17.8	22.5	52010	92268	101725									
33.99187	www.gadsdenstate.edu	5797	14.5	15.9	16.5	54629	42194	37780									
31.31609	www.wallace.edu	4686	17.1	15.9	16.5	30885	42194	37780									
34.07341	www.wallacestate.edu	5281	18.8	15.9	16.5	42870	42194	37780									
32.44561	www.wccs.edu	1745	23.8	15.9	16.5	43923	42194	37780									
33.46811	www.herzing.edu/birmingham	320	50	29.1	24.6	24343	17406	38763									
32.35094	www.huntingdon.edu	1110	22.2	17.8	22.5	76919	92268	101725									
34.77212	www.drakestate.edu	1383	19.3	15.9	16.5	86504	42194	37780									
33.82213	www.jsu.edu/	7588	18.1	18.8	21.5	51692	75743	66436									
31.102	www.jdcc.edu	1091	20.1	15.9	16.5	40726	42194	37780									
33.65244	www.jeffstateonline.com	8542	15.8	15.9	16.5	32495	42194	37780									
34.65428	www.calhoun.edu	11186	12.8	15.9	16.5	32165	42194	37780									
32.63053	www.judson.edu	347	14.5	17.8	22.5	112204	92268	101725									
33.45167	www.lawsonstate.edu	3028	13.5	15.9	16.5	47256	42194	37780									
32.59244	www.uwa.edu	1999	15.3	18.8	21.5	40841	75743	66436									
31.32295	www.lbwc.edu	1570	20.9	15.9	16.5	35010	42194	37780									
32.62236	www.marietta.edu	418	26.1	15.9	16.5	123613	42194	37780									
33.48131	www.miles.edu	1666	12	17.8	22.5	115329	92268	101725									
30.79325	www.umobile.edu	1481	21.7	17.8	22.5	51098	92268	101725									
33.10625	www.montevallo.edu	2618	15.9	18.8	21.5	88163	75743	66436									
34.73957	www.mwsscc.edu	3854	16.3	15.9	16.5	33524	42194	37780									
34.80658	www.una.edu	5881	17.4	18.8	21.5	48127	75743	66436									
34.54547	www.nacc.edu	2834	23.7	15.9	16.5	18843	42194	37780									
34.75663	www.oakwood.edu	1861	16.9	17.8	22.5	96114	92268	101725									
exp_award_percentile	c_ipg	re_value	re_percentile	med_std_cat_value	med_std_percentile	aid_value	aid_percentile	median_value	median_percentile	med_100_value	med_100_percentile	med_150_value	med_150_percentile	exp_award_value	exp_award_state_value	exp_award_nati_value	exp_award_ijra_value
97	93.8	3096	33	0	7142	72	NUL	93	30	15	29.1	14	73.2	98.63.1	17		
97	72.7	10032	67	84	6088	50	24136	93	29.4	67	53.5	66	35.1	93.87.7	7		
90	62.7	294	12	NUL	2540	1	NUL	1	0	66.7	67.0	68.4	91.37.5	2			
81	59.4	403	43	89	6647	61	131502	91	16.5	24	48.4	54	38.9	72.7	1		
91	50.3	5053	41	830	7256	74	13202	84	8.8	11	25.2	9	82.7	100.62.2	15		
75	50.2	27148	96	1171	10390	94	193469	90	42.7	86	66.7	85	21.1	100.78.7	87		
53	50.9	11009	31	1100	5057	83	17239	23	7.7	39	53.3	53	93.4	93.57.7	5		
51	69.4	3571	30	970	4127	10	10736	79	9.9	15	27.3	12	40.1	56.63.2	17		
71	50.5	1311	47	177	20395	88	22092	92	70	53.2	67.0	87	16.9	10.89.5	92		
75	50.8	1311	47	177	20395	88	22092	70	53.2	67	53.9	64	19.4	19.95	25		
85	57.3	1431	20	NUL	4316	62	36	5	7.2	37	13.4	23	61.2	89.51.6	2.3		
72	30.5	515	20	NUL	4746	19	NUL	0	0	22.2	41	61	80.51.6	2.3			
81	51.5	577	49	NUL	4920	70	NUL	5	10	5.3	13.9	2	83.6	97.41.4	2		
65	56.5	2101	33	NUL	4936	80	428	37	8.3	44	12.9	21	51.2	74.54.1	33		
80	58.6	3703	55	NUL	4817	76	NUL	62	10	5.5	24.9	10	48.9	68.52.1	25		
73	58.6	2658	78	NUL	6667	15	15248	18	10.6	5	23.9	6	85.3	85.36.6	6		
79	54.9	4650	63	NUL	4849	77	649	46	7.4	38	14.4	27	60	88.55.9	42		
30	55.9	4486	63	NUL	4941	83	1131	23	5	28	55.5	42	55.5	81.56.7	42		
61	61.2	1574	23	NUL	4952	79	NUL	11.9	6.5	35.5	45.7	47	85.2	70.83.7	77		
13	50	338	21	NUL	3781	19	NUL	0	0	24.8	66	76.9	98.53.1	29			
49	61.8	1078	59	28	11447	34	41465	30	30.4	30	43.2	37	40.1	87.75.5	38		
34	45.9	921	9	NUL	3385	27	NUL	15.4	7.2	15.4	30	58.7	86.42.3	5			
34	76.2	6856	57	51	31036	57	6289	18	10.1	16	30.8	18	46.1	74.71.1	40		
34	56.5	935	103	NUL	3383	64	3141	31	13.6	68	29.8	80	48.8	77.55.4	23		
34	33.8	6041	74	NUL	3089	52	196	20	1.8	3	6.5	2	35.8	35.52.4	26		
34	38	7911	87	NUL	4589	70	NUL	6.2	32	13.3	22	40.3	46.57.5	49			
75	57.5	343	11	123	39910	78	4890	38	42	40.5	33	73.9	74.44	14			
67	59	2695	42	NUL	4127	57	NUL	18.4	81	18.5	43	68.2	95.52.5	27			
15	84.8	1744	19														

Cc Institution Grad.csv

unitid	year	gender	race	cohort	grad_cohort	grad_100	grad_150	grad_100_rate	grad_150_rate
100760	2011	B	X	2y all	446	73	105	16.4	23.5
100760	2011	M	X	2y all	185	NULL	40	NULL	21.6
100760	2011	F	X	2y all	261	NULL	65	NULL	24.9
100760	2011	B	W	2y all	348	NULL	86	NULL	24.7
100760	2011	M	W	2y all	162	NULL	35	NULL	21.6
100760	2011	F	W	2y all	186	NULL	51	NULL	27.4
100760	2011	B	B	2y all	89	NULL	18	NULL	20.2
100760	2011	M	B	2y all	21	NULL	5	NULL	23.8
100760	2011	F	B	2y all	68	NULL	13	NULL	19.1
100760	2011	B	H	2y all	0	NULL	0	NULL	NULL
100760	2011	M	H	2y all	0	NULL	0	NULL	NULL
100760	2011	F	H	2y all	0	NULL	0	NULL	NULL
100760	2011	B	Ai	2y all	0	NULL	0	NULL	NULL
100760	2011	M	Ai	2y all	0	NULL	0	NULL	NULL
100760	2011	F	Ai	2y all	0	NULL	0	NULL	NULL
100760	2011	B	A	2y all	0	NULL	0	NULL	NULL
100760	2011	M	A	2y all	0	NULL	0	NULL	NULL
100760	2011	F	A	2y all	0	NULL	0	NULL	NULL
100760	2012	B	X	2y all	594	40	87	6.7	14.6
100760	2012	M	X	2y all	234	NULL	32	NULL	13.7
100760	2012	F	X	2y all	360	NULL	55	NULL	15.3
100760	2012	B	W	2y all	410	NULL	69	NULL	16.8
100760	2012	M	W	2y all	181	NULL	27	NULL	14.9
100760	2012	F	W	2y all	229	NULL	42	NULL	18.3
100760	2012	B	B	2y all	174	NULL	18	NULL	10.3
100760	2012	M	B	2y all	47	NULL	5	NULL	10.6
100760	2012	F	B	2y all	127	NULL	13	NULL	10.2
100760	2012	B	H	2y all	3	NULL	0	NULL	0.0
100760	2012	M	H	2y all	1	NULL	0	NULL	0.0
100760	2012	F	H	2y all	2	NULL	0	NULL	0.0
100760	2012	B	Ai	2y all	1	NULL	0	NULL	0.0
100760	2012	M	Ai	2y all	1	NULL	0	NULL	0.0
100760	2012	F	Ai	2y all	0	NULL	0	NULL	NULL

Cc_state_sector_details.csv

stateid	state	state_abbr	state_post	level	control	schools_count	counted_pct	awards_per_state_value	awards_per_natl_value	exp_award_state_value	exp_award_natl_value
0	United States	US	U.S.	4-year	Public	632	NULL	NULL	21.5	NULL	664
0	United States	US	U.S.	4-year	Private not-for-profit	1180	NULL	NULL	22.5	NULL	1011
0	United States	US	U.S.	4-year	Private for-profit	527	NULL	NULL	24.6	NULL	381
0	United States	US	U.S.	2-year	Public	926	NULL	NULL	16.5	NULL	371
0	United States	US	U.S.	2-year	Private not-for-profit	68	NULL	NULL	25.9	NULL	349
0	United States	US	U.S.	2-year	Private for-profit	465	NULL	NULL	32.8	NULL	241
1	Alabama	AL	Ala.	4-year	Public	13	61	18.8	21.5	75743	664
1	Alabama	AL	Ala.	4-year	Private not-for-profit	16	62	17.8	22.5	92268	1011
1	Alabama	AL	Ala.	4-year	Private for-profit	9	37.8	29.1	24.6	17406	381
1	Alabama	AL	Ala.	2-year	Public	25	49	15.9	16.5	42194	371
1	Alabama	AL	Ala.	2-year	Private not-for-profit	1	NULL	17.1	25.9	16852	349
1	Alabama	AL	Ala.	2-year	Private for-profit	5	79.6	3.2	32.8	24274	241
2	Alaska	AK	Alaska	4-year	Public	3	51.7	16.3	21.5	103823	664
2	Alaska	AK	Alaska	4-year	Private not-for-profit	1	21.8	22.9	22.5	97275	1011
2	Alaska	AK	Alaska	4-year	Private for-profit	1	23.2	34.2	24.6	30416	381
2	Alaska	AK	Alaska	2-year	Public	2	15.6	9.1	16.5	160134	371
2	Alaska	AK	Alaska	2-year	Private not-for-profit	0	NULL	NULL	25.9	NULL	349
2	Alaska	AK	Alaska	2-year	Private for-profit	0	NULL	NULL	32.8	NULL	241
4	Arizona	AZ	Ariz.	4-year	Public	4	63.9	22	21.5	75690	664
4	Arizona	AZ	Ariz.	4-year	Private not-for-profit	4	62.3	21.7	22.5	35493	1011
4	Arizona	AZ	Ariz.	4-year	Private for-profit	23	48	32.8	24.6	25661	381
4	Arizona	AZ	Ariz.	2-year	Public	20	24.5	20.4	16.5	25969	371
4	Arizona	AZ	Ariz.	2-year	Private not-for-profit	0	NULL	NULL	25.9	NULL	349
4	Arizona	AZ	Ariz.	2-year	Private for-profit	12	44.9	36.8	32.8	21312	241
5	Arkansas	AR	Ark.	4-year	Public	10	67.3	19.6	21.5	47347	664
5	Arkansas	AR	Ark.	4-year	Private not-for-profit	12	78.3	19.4	22.5	82509	1011
5	Arkansas	AR	Ark.	4-year	Private for-profit	2	65.1	20.8	24.6	55570	381
5	Arkansas	AR	Ark.	2-year	Public	22	43.6	23.6	16.5	25729	371
5	Arkansas	AR	Ark.	2-year	Private not-for-profit	1	NULL	11.6	25.9	15410	349
5	Arkansas	AR	Ark.	2-year	Private for-profit	1	100	51.4	32.8	38603	241
6	California	CA	Calif.	4-year	Public	32	58.1	25	21.5	79310	664
6	California	CA	Calif.	4-year	Private not-for-profit	74	68.5	22.4	22.5	125902	1011

awards_per_state_value	awards_per_natl_value	exp_award_state_value	exp_award_natl_value	state_appr_value	state_appr_rank	grad_rate_rank	awards_per_rank
NULL	21.5	NULL	66436	NULL	NULL	23	NULL
NULL	22.5	NULL	101725	NULL	NULL	18	NULL
NULL	24.6	NULL	38763	NULL	NULL	8	NULL
NULL	16.5	NULL	37780	NULL	NULL	25	NULL
NULL	25.9	NULL	34510	NULL	NULL	12	NULL
NULL	32.8	NULL	24795	NULL	NULL	19	NULL
18.8	21.5	75743	66436	290.68	11	38	42
17.8	22.5	92268	101725	290.68	11	47	49
29.1	24.6	17406	38763	290.68	11	26	9
15.9	16.5	42194	37780	290.68	11	35	39
17.1	25.9	16852	34510	290.68	11	8	19
3.2	32.8	24274	24795	290.68	11	38	40
16.3	21.5	103823	66436	496.78	2	51	51
22.9	22.5	97275	101725	496.78	2	49	15
34.2	24.6	30416	38763	496.78	2	29	1
9.1	16.5	160134	37780	496.78	2	8	50
NULL	25.9	NULL	34510	496.78	2	NULL	NULL
NULL	32.8	NULL	24795	496.78	2	NULL	NULL
22	21.5	75690	66436	126.81	47	20	11
21.7	22.5	35493	101725	126.81	47	34	30
32.8	24.6	25661	38763	126.81	47	34	3
20.4	16.5	25969	37780	126.81	47	40	14
NULL	25.9	NULL	34510	126.81	47	NULL	NULL
36.8	32.8	21312	24795	126.81	47	12	19
19.6	21.5	47347	66436	306.32	9	50	34
19.4	22.5	82509	101725	306.32	9	38	42
20.8	24.6	55570	38763	306.32	9	38	33
23.6	16.5	25729	37780	306.32	9	20	3
11.6	25.9	15410	34510	306.32	9	7	26
51.4	32.8	38603	24795	306.32	9	32	4
25	21.5	79310	66436	230.7	26	9	1
22.4	22.5	125902	101725	230.7	26	4	21

Cc_state_sector_grads.csv

stateid	state	state_abbr	control	level	year	gender	race	cohort	grad_cohort	grad_100	grad_150	grad_100_rate	grad_150_rate	grad_cohort_ct
1	Alabama	AL	Private for-profit	4-year	2011	B	A	4y bach	0	0	0	NULL	NULL	9
1	Alabama	AL	Private for-profit	4-year	2011	B	Al	4y bach	1	0	0	0	0	9
1	Alabama	AL	Private for-profit	4-year	2011	B	B	4y bach	51	2	3 3.9	5.9	5.9	9
1	Alabama	AL	Private for-profit	4-year	2011	B	H	4y bach	1	0	0	0	0	9
1	Alabama	AL	Private for-profit	4-year	2011	B	W	4y bach	66	15	18 22.7	27.3	27.3	9
1	Alabama	AL	Private for-profit	4-year	2011	B	X	4y bach	209	39	49 18.7	23.4	23.4	9
1	Alabama	AL	Private for-profit	4-year	2011	F	A	4y bach	0	0	0	NULL	NULL	9
1	Alabama	AL	Private for-profit	4-year	2011	F	Al	4y bach	0	0	0	NULL	NULL	9
1	Alabama	AL	Private for-profit	4-year	2011	F	B	4y bach	26	0	0	0	0	9
1	Alabama	AL	Private for-profit	4-year	2011	F	H	4y bach	0	0	0	NULL	NULL	9
1	Alabama	AL	Private for-profit	4-year	2011	F	W	4y bach	21	6	6 28.6	28.6	28.6	9
1	Alabama	AL	Private for-profit	4-year	2011	F	X	4y bach	66	9	10 13.6	15.2	15.2	9
1	Alabama	AL	Private for-profit	4-year	2011	M	A	4y bach	0	0	0	NULL	NULL	9
1	Alabama	AL	Private for-profit	4-year	2011	M	Al	4y bach	1	0	0	0	0	9
1	Alabama	AL	Private for-profit	4-year	2011	M	B	4y bach	25	2	3 8	12	12	9
1	Alabama	AL	Private for-profit	4-year	2011	M	H	4y bach	1	0	0	0	0	9
1	Alabama	AL	Private for-profit	4-year	2011	M	W	4y bach	45	9	12 20	26.7	26.7	9
1	Alabama	AL	Private for-profit	4-year	2011	M	X	4y bach	143	30	39 21	27.3	27.3	9
1	Alabama	AL	Private for-profit	4-year	2012	B	A	4y bach	1	0	0	0	0	9
1	Alabama	AL	Private for-profit	4-year	2012	B	Al	4y bach	1	0	0	0	0	9
1	Alabama	AL	Private for-profit	4-year	2012	B	B	4y bach	75	15	19 20	25.3	25.3	9
1	Alabama	AL	Private for-profit	4-year	2012	B	H	4y bach	4	1	2 25	50	50	9
1	Alabama	AL	Private for-profit	4-year	2012	B	W	4y bach	89	30	34 33.7	38.2	38.2	9
1	Alabama	AL	Private for-profit	4-year	2012	B	X	4y bach	235	55	69 23.4	29.4	29.4	9
1	Alabama	AL	Private for-profit	4-year	2012	F	A	4y bach	0	0	0	NULL	NULL	9
1	Alabama	AL	Private for-profit	4-year	2012	F	Al	4y bach	0	0	0	NULL	NULL	9
1	Alabama	AL	Private for-profit	4-year	2012	F	B	4y bach	39	9	10 23.1	25.6	25.6	9
1	Alabama	AL	Private for-profit	4-year	2012	F	H	4y bach	3	0	1 0	33.3	33.3	9
1	Alabama	AL	Private for-profit	4-year	2012	F	W	4y bach	31	9	9 29	29	29	9
1	Alabama	AL	Private for-profit	4-year	2012	F	X	4y bach	88	19	22 21.6	25	25	9
1	Alabama	AL	Private for-profit	4-year	2012	M	A	4y bach	1	0	0	0	0	9
1	Alabama	AL	Private for-profit	4-year	2012	M	Al	4y bach	1	0	0	0	0	9

1.3 Công cụ sử dụng

Visual Studio 2022 tích hợp

- + Tích hợp dữ liệu SSIS
- + SQL Server Data Tools (SSDT)

SQL Server 2019

Ngôn ngữ lập trình: SQL, Python

Môi trường: Visual Studio Code

CHƯƠNG 2: THIẾT KẾ XÂY DỰNG CƠ SỞ DỮ LIỆU TÁC NGHIỆP

2.1 Thiết kế nghiệp vụ

Institution Details Report: Phân tích chi tiết về các cơ sở đại học. Người phân tích muốn theo dõi những thông số liên quan tới vị trí địa lý, loại trường dựa theo hệ thống phân loại Carnegie của Mỹ và các thông số đo lường liên quan tới Sinh viên của cơ sở InstitutionGraduation: Phân tích về tỉ lệ tốt nghiệp của các trường đại học

Instructions!					
Business Process Name	Fact Table	Fact Grain Type	Granularity	Facts	
DimLocation	DimCohort	DimCarnegie	DimState		
InstitutionDetailsReport	FactDetail	Periodic snapshot	one row per year	Student count, median SAT, pell value, ...	X X X
InstitutionGraduation_Cohort	FactGraduation	Periodic snapshot	one row per year	cohort size, grad_100, grad_150, ...	X X

2.2 Thiết kế các bảng Dimension

2.2.1 DimState

Thuộc tính	Mô tả
keyState	
name	tên tiểu bang và lãnh thổ
state abbr	tên viết tắt của tiểu bang và lãnh thổ
state code	mã số của tiểu bang và lãnh thổ
state post	mã của tiểu bang và lãnh thổ

2.2.2 DimCohort

Thuộc tính	Mô tả
keyCohort	
race	Chủng tộc
cohort	Nhóm sinh viên tốt nghiệp cùng một khóa hoặc một nhóm
gender	Giới tính
year	Năm thực hiện thống kê

2.2.3 DimLocation

Thuộc tính	Mô tả
keyLocation	
keyState	
City	Tên thành phố của một Bang

2.2.4 DimCarnegie

Thuộc tính	Mô tả
keyCarnegie	
institutionType	Loại trường đại học dựa trên phân loại của hệ thống Carnegie
programSize	Quy mô chương trình của Trường đại học

2.3 Thiết kế bảng Fact

2.3.1 Fact Institution Details

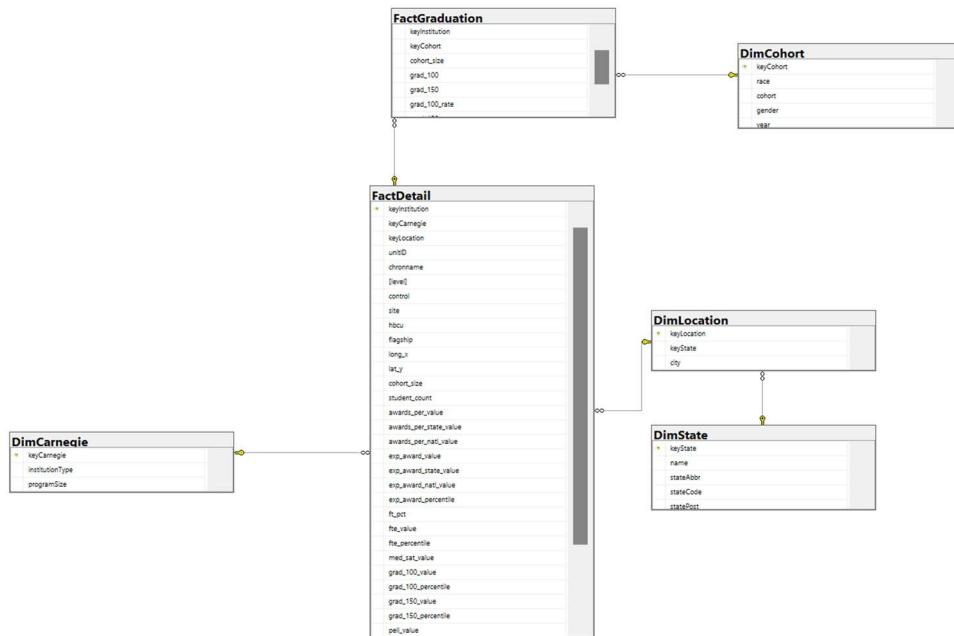
Thuộc tính	Mô tả
keyInstitution	
keyCarnegie	
keyLocation	
UnitID	Id trường đại học
chronname	tên trường đại học
Level	Trường đại học 4 năm hay 2 năm (Đại học hay cao đẳng)
Control	Kiểu trường (Công lập, tư nhân, tư nhân nhưng phi lợi nhuận)
site	Trang web trường
measure...	Các giá trị do lường (do có nhiều giá trị nên chỉ để là measure nói chung)

2.3.1 Fact Institution Gradution

Thuộc tính	Mô tả
keyInstitution	
keyCohort	
cohort_size	Số lượng sinh viên tham gia khảo sát
grad_100	Số sinh viên tốt nghiệp trong 100% thời gian bình thường hoặc mong đợi
grad_150	Số sinh viên tốt nghiệp trong 150% thời gian bình thường hoặc mong đợi
grad_100_rate	Tỷ lệ phần trăm sinh viên tốt nghiệp trong 150% thời gian bình thường hoặc mong đợi
grad_150_rate	Tỷ lệ phần trăm sinh viên tốt nghiệp trong 150% thời gian bình thường hoặc mong đợi.

2.4 Thiết kế và tiền xử lý dữ liệu

2.4.1 Lược đồ hình bông tuyết



2.4.2 Tiền xử lý dữ liệu

Ý tưởng: Với các dữ liệu ta cần phân tích riêng như “Mỗi bang có tỉ lệ tốt nghiệp như thế nào?” hay “Với mỗi nhóm đối tượng có sự chênh lệch với nhau không.” Ta sẽ tạo ra các dimension tương ứng. Từ dữ liệu gốc InstitutionDetails ban đầu có 62 cột, ta lọc những thuộc tính cần thiết ta được một dữ liệu mới có cùng độ dài nhưng.

Script tiền xử lý, làm sạch dữ liệu:

```
import pandas as pd
import numpy as np

cc_institutions_details = pd.read_csv('cc_institutions_details.csv')
cc_institutions_details["hbcu"] = cc_institutions_details["hbcu"].apply(lambda x: 1 if x == "X" or x == 1 else 0)
cc_institutions_details["flagship"] = cc_institutions_details["flagship"].apply(lambda x: 1 if x == "X" or x == 1 else 0)
cc_institutions_details.drop(columns=cc_institutions_details.iloc[:, 39:57].columns.tolist(), inplace=True)
columns_to_drop = ["aid_value", "aid_percentile", "med_sat_percentile", "endow_value", "endow_percentile", "nicknames", "counted_pct", "ft_fac_value", "ft_fac_percentile"]
cc_institutions_details.drop(columns=columns_to_drop, inplace=True)
cc_institutions_details.drop(columns=["state_sector_ct", "carnegie_ct"], inplace=True)
cc_institutions_details.to_csv('pre_cc_institutions_details.csv', index=False)
```

unitid	chronname	city	state	level	control	basic	hbcu	flagship	long_x	lat_y	site
100654	Alabama A&T Normal	Alabama	4-year	Public	Masters College	1	0	-86.568502	34.783368	www.aamu.edu	
100663	University of Birmingham	Alabama	4-year	Public	Research University	0	0	-86.80917	33.50223	www.uab.edu	
100690	Amridge University	Montgomery	Alabama	4-year	Private not-for-profit	Baccalaureate	0	0	-86.17401	32.362609	www.amridge.edu
100706	University of Huntsville	Alabama	4-year	Public	Research University	0	0	-86.63842	34.722818	www.uah.edu	
100724	Alabama State University	Montgomery	Alabama	4-year	Public	Masters College	1	0	-86.295677	32.364317	www.alasu.edu
100751	University of Tuscaloosa	Tuscaloosa	Alabama	4-year	Public	Research University	0	1	-87.545766	33.2144	www.ua.edu
100760	Central Alabama Community College	Alexander City	Alabama	2-year	Public	Associates	0	0	-85.94653	32.924429	www.cacc.edu
100830	Auburn University	Montgomery	Alabama	4-year	Public	Masters College	0	0	-86.177351	32.369939	www.aum.edu
100858	Auburn University	Auburn University	Alabama	4-year	Public	Research University	0	0	-85.492409	32.600201	www.auburn.edu
100937	Birmingham City University	Birmingham	Alabama	4-year	Private not-for-profit	Baccalaureate	0	0	-86.853636	33.515453	www.bsc.edu
101028	Chattahoochee Technical College	Phenix City	Alabama	2-year	Public	Associates	0	0	-85.031485	32.42391	www.cv.edu
101073	Concordia College	Selma	Alabama	4-year	Private not-for-profit	Baccalaureate	1	0	-87.023531	32.42443	www.ccac.edu
101116	South University	Montgomery	Alabama	4-year	Private for-profit	Baccalaureate	0	0	-86.216488	32.342684	southuniversity.edu
101143	Enterprise State Community College	Enterprise	Alabama	2-year	Public	Associates	0	0	-85.836956	31.297496	www.escc.edu

student_cohort	awards_per_pell	awards_per_fte	awards_per_fte_percent	exp_award	exp_award	exp_award	exp_award	ft_pct	fte_value	fte_percent	med_sat_value	grad_100_percent	grad_100_percent
4051	14.2	18.8	21.5	105331	75743	66436	90	93.8	3906	33	823	10	15
11502	20.9	18.8	21.5	136546	75743	66436	97	72.7	10032	67	1146	29.4	67
322	29.9	17.8	22.5	58414	92268	101725	30	62.7	294	12		0	0
5696	20.9	18.8	21.5	64418	75743	66436	61	74.4	5000	40	1180	16.5	34
5356	11.6	18.8	21.5	132407	75743	66436	96	91	5035	41	830	8.8	11
29440	18.3	18.8	21.5	75350	75743	66436	75	90.2	27148	96	1171	42.7	86
1906	15.9	15.9	16.5	57572	42194	37780	83	50	1489	21		7.7	39
4322	15.4	18.8	21.5	58541	75743	66436	51	69.4	3571	30	970	9.9	15
19799	21.5	18.8	21.5	71999	75743	66436	71	91	19635	87	1215	37.6	80
1188	23.5	17.8	22.5	113677	92268	101725	75	98.2	1331	47	1177	53.2	67
1837	13.6	15.9	16.5	60303	42194	37780	85	57.3	1431	20		7.2	37
600	12.3	17.8	22.5	105461	92268	101725	72	90.5	515	20		9.8	4
599	15	29.1	24.6	76191	17406	38763	81	51.9	577	49		5.3	10
2333	17.4	15.9	16.5	45911	42194	37780	65	56.5	2101	33		8.3	44

grad_150_value	grad_150_percent	pell_value	pell_percent	retain_value	retain_percent	state_sect	carnegie_ct	cohort_size
29.1	14	71.2	98	63.1	17	13	386	882
53.5	66	35.1	39	80.2	70	13	106	1376
66.7	72	68.4	91	37.5	2	16	252	3
48.4	54	32.8	32	81	72	13	106	759
25.2	9	82.7	100	62.2	15	13	386	1351
66.7	85	21.1	7	87	87	13	96	4438
9.1	8	65.1	93	42.7	5	25	289	594
27.1	12	40.1	56	63.2	17	13	386	536
67.9	87	16.9	3	89.5	92	13	96	4165
61.9	64	21.4	16	80.4	65	16	252	449
13.4	23	61.2	89	51.6	23	25	289	276
13.9	2	83.6	97	41	2	16	343	122
5.3	1	71.3	59	19.4	1	9	124	19
12.9	21	51.2	74	54.1	33	25	289	520

Trong tập dữ liệu, ta thấy có cột basic. Đó là một phân loại do một tổ chức “carnegie classification of institutions of higher education” (tổ chức phân loại các trường đại học) phân loại các trường đại học ở Mỹ. Lúc này ta cần có một dimension về phân loại các trường đại học để thuận lợi cho việc phân tích

```

import pandas as pd

# Đường dẫn đến tệp CSV
file_path =
"E:/TAI_LIEU_DAI_HOC/SEMESTER_6/DataWarehousing/Final_Project/data/cc_institution_details.csv"

# Đọc dữ liệu từ tệp CSV và tạo DataFrame
basic_set_df = pd.read_csv(file_path)

# Khởi tạo một từ điển rỗng
lst1 = {}

# Duyệt qua từng hàng của DataFrame
for index, row in basic_set_df.iterrows():
    # Lấy giá trị của cột "tên_cột" trong hàng hiện tại
    val = row["tên_cột"]
    # Tách giá trị dựa trên dấu --
    tmp = val.split("--")
    # Kiểm tra xem phần đầu của tmp đã là một khóa trong từ điển lst1 chưa
    if tmp[0] in lst1.keys():
        # Nếu đã tồn tại, kiểm tra xem có phần dữ liệu con không
        if len(tmp) > 1:
            subCarnegie = "--".join(tmp[1:])
            lst1[tmp[0]].add(subCarnegie)
        else:
            lst1[tmp[0]] = set()
    else:
        # Nếu không tồn tại, tạo một tập hợp mới và thêm phần dữ liệu con nếu có
        if len(tmp) > 1:
            lst1[tmp[0]] = set()
            subCarnegie = "--".join(tmp[1:])
            lst1[tmp[0]].add(subCarnegie)
        else:
            lst1[tmp[0]] = set()

# Tạo một từ điển mới để lưu trữ dữ liệu cho DataFrame
dic_csv = {"institutionType": [], "programSize": []}

# Duyệt qua các mục trong từ điển lst1
for key, val in lst1.items():
    # Nếu tập hợp có độ dài lớn hơn 0, thêm từng phần tử vào danh sách "institutionType" và "programSize"
    if len(val) > 0:
        for v in val:
            dic_csv["institutionType"].append(key)
            dic_csv["programSize"].append(v)
    else:
        # Nếu không có phần tử, chỉ thêm key vào danh sách "institutionType" và thêm pd.NA vào "programSize"

```

```

dic_csv["institutionType"].append(key)
dic_csv["programSize"].append(pd.NA)

# Tạo DataFrame từ từ điển dic_csv
Df_DimCarnegie = pd.DataFrame(dic_csv)

# Truy cập các thuộc tính của DataFrame
print("Head of DataFrame:")
print(Df_DimCarnegie.head()) # In ra 5 hàng đầu tiên

```

Sau khi xử lý, ta có thể tách ra được loại chương trình của Trường Đại học được phân loại theo hệ thống phân loại Carnegie và quy mô chương trình Đại học theo hệ thống phân loại đó

institutionType	programSize	basic
Masters Colleges and Universities	smaller programs	Masters Colleges and Universities--smaller programs
Masters Colleges and Universities	larger programs	Masters Colleges and Universities--larger programs
Masters Colleges and Universities	medium programs	Masters Colleges and Universities--medium programs
Research Universities	very high research activity	Research Universities--very high research activity
Research Universities	high research activity	Research Universities--high research activity
Baccalaureate Colleges	Diverse Fields	Baccalaureate Colleges--Diverse Fields
Baccalaureate Colleges	Arts & Sciences	Baccalaureate Colleges--Arts & Sciences
Associates	Private For-profit 4-year Primarily Associates	Associates--Private For-profit 4-year Primarily Associates
Associates	Public Rural-serving Small	Associates--Public Rural-serving Small
Associates	Public Urban-serving Single Campus	Associates--Public Urban-serving Single Campus
Associates	Public Urban-serving Multicampus	Associates--Public Urban-serving Multicampus
Associates	Public Rural-serving Large	Associates--Public Rural-serving Large
Associates	Public Rural-serving Medium	Associates--Public Rural-serving Medium
Associates	Public Suburban-serving Multicampus	Associates--Public Suburban-serving Multicampus
Associates	Private For-profit	Associates--Private For-profit
Associates	Public 2-year colleges under 4-year universities	Associates--Public 2-year colleges under 4-year universities
Associates	Public Special Use	Associates--Public Special Use
Associates	Private Not-for-profit 4-year Primarily Associates	Associates--Private Not-for-profit 4-year Primarily Associates
Associates	Private Not-for-profit	Associates--Private Not-for-profit
Associates	Public 4-year Primarily Associates	Associates--Public 4-year Primarily Associates
Associates	Public Suburban-serving Single Campus	Associates--Public Suburban-serving Single Campus
Baccalaureate/Associates Colleges		Baccalaureate/Associates Colleges
Theological seminaries- Bible colleges- and other faith-related institutions		Theological seminaries- Bible colleges- and other faith-related institutions
Not applicable- not in Carnegie universe		Not applicable- not in Carnegie universe
Schools of art- music- and design		Schools of art- music- and design
Other technology-related schools		Other technology-related schools
Tribal Colleges		Tribal Colleges
Doctoral/Research Universities		Doctoral/Research Universities
Other health professions schools		Other health professions schools
Schools of business and management		Schools of business and management
Schools of engineering		Schools of engineering
Other special-focus institutions		Other special-focus institutions

Dữ liệu gốc của cc_institution_grads.csv có 1302102 dòng. Nhưng hầu hết trong số đó dữ liệu đều bị trống và bằng 0. Ta tiến hành bỏ các dữ liệu đó để tiết kiệm thời gian khi ETL vào kho dữ liệu. Và các cột thuộc tính là các từ viết tắt ta cũng đưa thông tin cụ thể vào.

Script clean institution graduation.

```

import numpy as np
import pandas as pd
cc_institutions_grads = pd.read_csv('cc_institution_grads.csv')

cc_institutions_grads["grad_100_rate"] = np.round(cc_institutions_grads["grad_100"] * 100 /
cc_institutions_grads["grad_cohort"], 1)
cc_institutions_grads["grad_150_rate"] = np.round(cc_institutions_grads["grad_150"] * 100 /
cc_institutions_grads["grad_cohort"], 1)

clean_data = cc_institutions_grads[(~cc_institutions_grads["grad_100_rate"].isnull()) &
(~cc_institutions_grads["grad_150"].isnull())]
clean_data = clean_data[clean_data["grad_cohort"] > 0]
clean_data = clean_data.reset_index(drop=True)
gender_dic = {
    "M": "Male",
    "F": "Female",
    "B": "Both gender"
}

race_dic = {
    'X': "all students",
    'A1': "American Indian",
    'A': "Asian",
    'B': "Black",
    'H': "Hispanic",
    'W': "White"
}
cohort_dic = {
    '4y bach': "Bachelor's/equivalent-seeking cohort at 4-year institutions",
    '4y other': "Students seeking another type of degree or certificate at a 4-year institution",
    '2y all': "Degree-seeking students at 2-year institutions"
}

clean_data["gender"] = clean_data["gender"].transform(lambda x: gender_dic[x])
clean_data["race"] = clean_data["race"].transform(lambda x: race_dic[x])
clean_data["cohort"] = clean_data["cohort"].transform(lambda x: cohort_dic[x])

clean_data.to_csv('pre_institutions_grads.csv', index=False)

```

Dữ liệu sau khi xử lý còn 332061 dòng

A	B	C	D	E	F	G	H	I	J
unitid	year	gender	race	cohort	grad_cohort	grad_100	grad_150	grad_100_rate	grad_150_rate
2	100760	2011	Both gender all students	Degree-seeking students at 2-year institutions	446	73	105	16.4	23.5
3	100760	2012	Both gender all students	Degree-seeking students at 2-year institutions	594	40	87	6.7	14.6
4	100760	2013	Both gender all students	Degree-seeking students at 2-year institutions	594	46	54	7.7	9.1
5	101028	2011	Both gender all students	Degree-seeking students at 2-year institutions	261	25	42	9.6	16.1
6	101028	2012	Both gender all students	Degree-seeking students at 2-year institutions	281	41	41	14.6	14.6
7	101028	2013	Both gender all students	Degree-seeking students at 2-year institutions	276	20	37	7.2	13.4
8	101143	2011	Both gender all students	Degree-seeking students at 2-year institutions	461	126	174	27.3	37.7
9	101143	2012	Both gender all students	Degree-seeking students at 2-year institutions	553	103	122	18.6	22.1
10	101143	2013	Both gender all students	Degree-seeking students at 2-year institutions	520	43	67	8.3	12.9
11	101161	2011	Both gender all students	Degree-seeking students at 2-year institutions	868	118	129	13.6	14.9
12	101161	2012	Both gender all students	Degree-seeking students at 2-year institutions	1070	125	148	11.7	13.8
13	101161	2013	Both gender all students	Degree-seeking students at 2-year institutions	1043	99	104	9.5	10
14	101240	2011	Both gender all students	Degree-seeking students at 2-year institutions	1307	156	228	11.9	17.4
15	101240	2012	Both gender all students	Degree-seeking students at 2-year institutions	1573	164	244	10.4	15.5
16	101240	2013	Both gender all students	Degree-seeking students at 2-year institutions	1523	112	220	7.4	14.4
17	101286	2011	Both gender all students	Degree-seeking students at 2-year institutions	801	151	151	18.9	18.9
18	101286	2012	Both gender all students	Degree-seeking students at 2-year institutions	973	35	130	3.6	13.4
19	101286	2013	Both gender all students	Degree-seeking students at 2-year institutions	953	22	142	2.3	14.9
20	101295	2011	Both gender all students	Degree-seeking students at 2-year institutions	1026	212	212	20.7	20.7
21	101295	2012	Both gender all students	Degree-seeking students at 2-year institutions	1108	112	254	10.1	22.9
22	101295	2013	Both gender all students	Degree-seeking students at 2-year institutions	1086	129	212	11.9	19.5
23	101301	2011	Both gender all students	Degree-seeking students at 2-year institutions	394	59	107	15	27.2
24	101301	2012	Both gender all students	Degree-seeking students at 2-year institutions	462	54	111	11.7	24
25	101301	2013	Both gender all students	Degree-seeking students at 2-year institutions	415	44	103	10.6	24.8
26	101462	2011	Both gender all students	Degree-seeking students at 2-year institutions	180	37	37	20.6	20.6
27	101462	2012	Both gender all students	Degree-seeking students at 2-year institutions	237	27	29	11.4	12.2

```

pandas as pd
numpy as np

# Đọc dữ liệu từ tệp CSV
= pd.read_csv('cc_state_sector_details.csv', encoding='latin1')

# Loại bỏ các hàng trùng lặp và tạo cột mới "state_code"
= data.drop_duplicates(subset=["stateid", "state", "state_abbr", "state_post"])
["stateid", "state", "state_abbr", "state_post"]]
["state_code"] = data["stateid"].apply(lambda x: f'{x:02}')

# Loại bỏ cột "stateid" không cần thiết
.drop(columns=["stateid"], inplace=True)

print(data)

```

Code trên thực hiện các bước để xử lý dữ liệu. nó loại bỏ bất kỳ hàng trùng lặp nào và tạo một cột mới gọi là 'state_code', được tạo ra từ mã trạng thái. Cuối cùng, nó loại bỏ cột 'stateid' không cần thiết và in ra kết quả cuối cùng của DataFrame được cập nhật.

Đây là dữ liệu trước khi được xử lý

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	stateid	state	state_abbr	state_post	level	control	schools_co	counted_pc	awards_per	awards_per	exp_award	exp_award	state_appr	state_appr	grad_rate	r_awards_per	rank
2	0	United Stat US	U.S.	4-year	Public	632	NULL	NULL	21.5	NULL	66436	NULL	NULL	23	NULL		
3	0	United Stat US	U.S.	4-year	Private not-for	1180	NULL	NULL	22.5	NULL	101725	NULL	NULL	18	NULL		
4	0	United Stat US	U.S.	4-year	Private for-I	527	NULL	NULL	24.6	NULL	38763	NULL	NULL	8	NULL		
5	0	United Stat US	U.S.	2-year	Public	926	NULL	NULL	16.5	NULL	37780	NULL	NULL	25	NULL		
6	0	United Stat US	U.S.	2-year	Private not-for	68	NULL	NULL	25.9	NULL	34510	NULL	NULL	12	NULL		
7	0	United Stat US	U.S.	2-year	Private for-I	465	NULL	NULL	32.8	NULL	24795	NULL	NULL	19	NULL		
8	1	Alabama	AL	Ala.	4-year	Public	13	61	18.8	21.5	75743	66436	290.68	11	38	42	
9	1	Alabama	AL	Ala.	4-year	Private not-for	16	62	17.8	22.5	92268	101725	290.68	11	47	49	
10	1	Alabama	AL	Ala.	4-year	Private for-I	9	37.8	29.1	24.6	17406	38763	290.68	11	26	9	
11	1	Alabama	AL	Ala.	2-year	Public	25	49	15.9	16.5	42194	37780	290.68	11	35	39	
12	1	Alabama	AL	Ala.	2-year	Private not-for	1	NULL	17.1	25.9	16852	34510	290.68	11	8	19	
13	1	Alabama	AL	Ala.	2-year	Private for-I	5	79.6	3.2	32.8	24274	24795	290.68	11	38	40	
14	2	Alaska	AK	Alaska	4-year	Public	3	51.7	16.3	21.5	103823	66436	496.78	2	51	51	
15	2	Alaska	AK	Alaska	4-year	Private not-for	1	21.8	22.9	22.5	97275	101725	496.78	2	49	15	
16	2	Alaska	AK	Alaska	4-year	Private for-I	1	23.2	34.2	24.6	30416	38763	496.78	2	29	1	
17	2	Alaska	AK	Alaska	2-year	Public	2	15.6	9.1	16.5	160134	37780	496.78	2	8	50	
18	2	Alaska	AK	Alaska	2-year	Private not-for	0	NULL	NULL	25.9	NULL	34510	496.78	2	NULL	NULL	
19	2	Alaska	AK	Alaska	2-year	Private for-I	0	NULL	NULL	32.8	NULL	24795	496.78	2	NULL	NULL	

Đây là kết quả dữ liệu sau khi được xử lý

A	B	C	D
state	state_abbr	state_post	state_code
United Stat US		U.S.	0
Alabama	AL	Ala.	1
Alaska	AK	Alaska	2
Arizona	AZ	Ariz.	4
Arkansas	AR	Ark.	5
California	CA	Calif.	6
Colorado	CO	Colo.	8
Connecticut	CT	Conn.	9
Delaware	DE	Del.	10
District of C	DC	D.C.	11
Florida	FL	Fla.	12
Georgia	GA	Ga.	13
Hawaii	HI	Hawaii	15
Idaho	ID	Idaho	16
Illinois	IL	Ill.	17
Indiana	IN	Ind.	18
Iowa	IA	Iowa	19
Kansas	KS	Kan.	20
Kentucky	KY	Ky.	21
Louisiana	LA	La.	22
Maine	ME	Me.	23

CHƯƠNG 3: TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS)

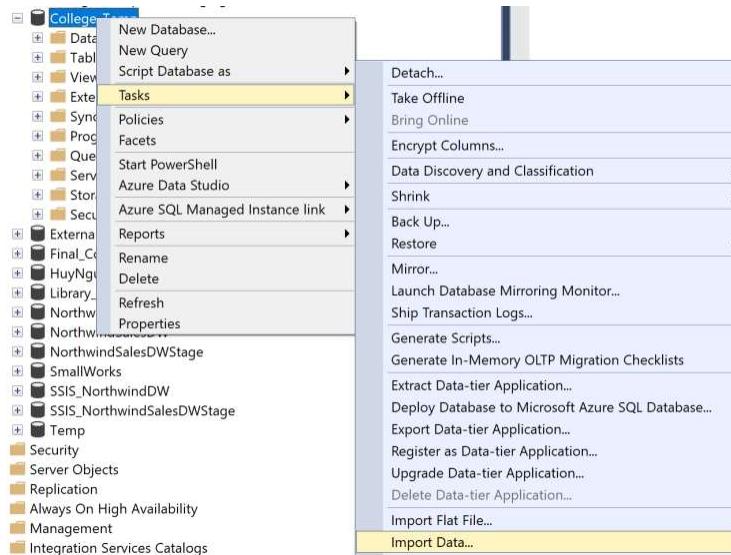
3.1 Tạo nguồn dữ liệu từ CSV

Thêm các dữ liệu đã xử lý từ file .csv vào một database trước khi staging

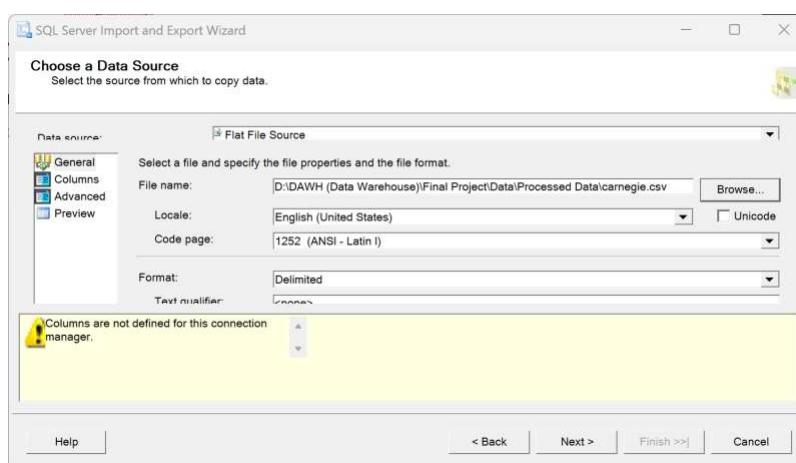
```
CREATE DATABASE College_Temp;
```



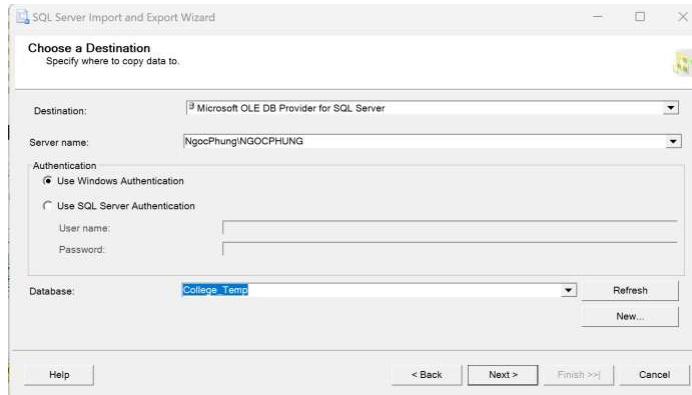
College_Temp → Tasks → Import Data



Chọn Data Source là Flat File Source à Mục File name

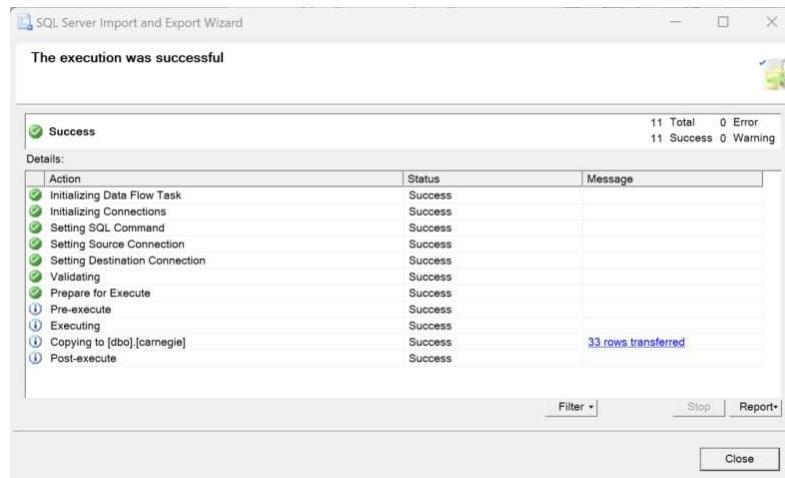


Chọn Data Destination là SQL Server

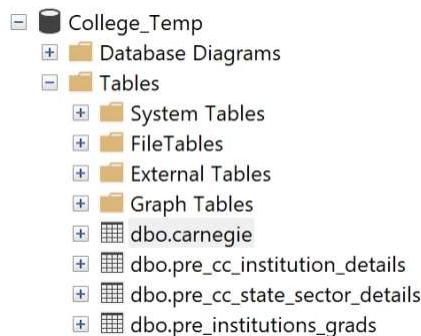


Chọn Edit Mappings để kiểm tra xem kiểu dữ liệu

Khi dữ liệu được thêm thành công hết vào không cáo bất kì lỗi gì



Cứ thực hiện các bước trên với 3 file còn lại



Kết quả bảng carnegie

institutionType	programSize	basic
1 Masters Colleges and Universities	smaller programs	Masters Colleges and Universities--smaller programs
2 Masters Colleges and Universities	larger programs	Masters Colleges and Universities--larger programs
3 Masters Colleges and Universities	medium programs	Masters Colleges and Universities--medium programs
4 Research Universities	very high research activity	Research Universities--very high research activity
5 Research Universities	high research activity	Research Universities--high research activity
6 Baccalaureate Colleges	Diverse Fields	Baccalaureate Colleges--Diverse Fields
7 Baccalaureate Colleges	Arts & Sciences	Baccalaureate Colleges--Arts & Sciences
8 Associates	Private For-profit 4-year Primarily Associates	Associates--Private For-profit 4-year Primarily Assoc...
9 Associates	Public Rural-serving Small	Associates--Public Rural-serving Small
10 Associates	Public Urban-serving Single Campus	Associates--Public Urban-serving Single Campus
11 Associates	Public Urban-serving Multicampus	Associates--Public Urban-serving Multicampus
12 Associates	Public Rural-serving Large	Associates--Public Rural-serving Large
13 Associates	Public Rural-serving Medium	Associates--Public Rural-serving Medium
14 Associates	Public Suburban-serving Multicampus	Associates--Public Suburban-serving Multicampus
15 Associates	Private For-profit	Associates--Private For-profit
16 Associates	Public 2-year colleges under 4-year universities	Associates--Public 2-year colleges under 4-year uni...
17 Associates	Public Special Use	Associates--Public Special Use
18 Associates	Private Not-for-profit 4-year Primarily Associates	Associates--Private Not-for-profit 4-year Primarily As...
19 Associates	Private Not-for-profit	Associates--Private Not-for-profit
20 Associates	Public 4-year Primarily Associates	Associates--Public 4-year Primarily Associates

Kết quả bảng pre_cc_institution_details

	unitid	chronname	city	state	level	control	basic	hbcu	flagship	long_x	lat_y
1	100654	Alabama A&M University	Normal	Alabama	4-year	Public	Masters Colleges and Universities--larger programs	1	0	-86.29568	32.36432
2	100663	University of Alabama at Birmingham	Birmingham	Alabama	4-year	Public	Research Universities--very high research activity	0	0	-86.80917	33.50223
3	100690	Amridge University	Montgomery	Alabama	4-year	Private not-for-profit	Baccalaureate Colleges--Arts & Sciences	0	0	-86.17401	32.36261
4	100706	University of Alabama at Huntsville	Huntsville	Alabama	4-year	Public	Research Universities--very high research activity	0	0	-86.63842	34.72282
5	100724	Alabama State University	Montgomery	Alabama	4-year	Public	Masters Colleges and Universities--larger programs	1	0	-86.29568	32.36432
6	100751	University of Alabama at Tuscaloosa	Tuscaloosa	Alabama	4-year	Public	Research Universities--high research activity	0	1	-87.54577	33.2144
7	100760	Central Alabama Community College	Alexander City	Alabama	2-year	Public	Associates--Public Rural-serving Medium	0	0	-85.94653	32.92443
8	100830	Auburn University at Montgomery	Montgomery	Alabama	4-year	Public	Masters Colleges and Universities--larger programs	0	0	-86.17735	32.36994
9	100858	Auburn University	Auburn University	Alabama	4-year	Public	Research Universities--high research activity	0	0	-85.49241	32.6000
10	100937	Birmingham-Southern College	Birmingham	Alabama	4-year	Private not-for-profit	Baccalaureate Colleges--Arts & Sciences	0	0	-86.85364	33.51545
11	101028	Chattahoochee Valley Community College	Phenix City	Alabama	2-year	Public	Associates--Public Rural-serving Medium	0	0	-85.03149	32.42391
12	101073	Concordia College (Ala.)	Selma	Alabama	4-year	Private not-for-profit	Baccalaureate Colleges--Diverse Fields	1	0	-87.02553	32.42443
13	101118	South University at Montgomery	Montgomery	Alabama	4-year	Private for-profit	Baccalaureate/Associates Colleges	0	0	-86.21649	32.34269
14	101143	Enterprise State Community College	Enterprise	Alabama	2-year	Public	Associates--Public Rural-serving Medium	0	0	-85.83695	31.2975
15	101161	Faulkner State Community College	Bay Minette	Alabama	2-year	Public	Associates--Public Suburban-serving Multicampus	0	0	-87.77975	30.85205
16	101189	Faulkner University	Montgomery	Alabama	4-year	Private not-for-profit	Baccalaureate Colleges--Diverse Fields	0	0	-86.21641	32.38418
17	101240	Gadsden State Community College	Gadsden	Alabama	2-year	Public	Associates--Public Rural-serving Large	1	0	-85.9913	33.99187
18	101286	George C. Wallace Community College at Dothan	Dothan	Alabama	2-year	Public	Associates--Public Rural-serving Medium	0	0	-85.46355	31.31609
19	101295	Wallace State Community College at Hanceville	Hanceville	Alabama	2-year	Public	Associates--Public Rural-serving Medium	0	0	-86.78175	34.07341

Kết quả bảng pre_cc_state_sector_details

	state	state_abbr	state_post	state_code
1	United States	US	U.S.	0
2	Alabama	AL	Ala.	1
3	Alaska	AK	Alaska	2
4	Arizona	AZ	Ariz.	4
5	Arkansas	AR	Ark.	5
6	California	CA	Calif.	6
7	Colorado	CO	Colo.	8
8	Connecticut	CT	Conn.	9
9	Delaware	DE	Del.	10
10	District of Columbia	DC	D.C.	11
11	Florida	FL	Fla.	12
12	Georgia	GA	Ga.	13
13	Hawaii	HI	Hawaii	15
14	Idaho	ID	Idaho	16
15	Illinois	IL	Ill.	17
16	Indiana	IN	Ind.	18
17	Iowa	IA	Iowa	19
18	Kansas	KS	Kan.	20
19	Kentucky	KY	Ky.	21
20	Louisiana	LA	La.	22

Kết quả bảng pre_institutions_grads

	unitid	year	gender	race	cohort	grad_cohort	grad_100	grad_150	grad_100_rate	grad_150_rate
1	206349	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	5	0	0	0	0
2	206589	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	5	2	2	2	1.3
3	206695	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	24	12	15	12	10
4	206941	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	17	1	6	1	4
5	207388	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	43	10	25	10	16.7
6	208725	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	2	1	1	1	0.7
7	208822	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	12	3	4	3	2.7
8	209065	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	24	11	14	11	9.3
9	209506	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	13	5	6	5	4
10	209612	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	41	18	26	18	17.3
11	209807	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	88	12	29	12	19.3
12	210571	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	9	7	8	7	5.3
13	211158	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	13	6	8	6	5.3
14	211361	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	3	0	1	0	0.7
15	211644	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	3	1	2	1	1.3
16	212054	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	144	16	97	16	64.7
17	212197	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	5	2	3	2	2
18	212984	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	5	2	4	2	2.7
19	213385	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	11	6	6	6	4
20	213668	2002	Both gender	Asian	Bachelor's/equivalent-seeking cohort at 4-year i...	4	3	4	3	2.7

3.2 Tạo College_Staging và project SSIS

```
/*
-- Drop existing College_Staging database (optional, ensures clean slate)
IF EXISTS (SELECT name FROM sys.databases WHERE name = 'College_Completion_Staging')
    DROP DATABASE College_Completion_Staging;
GO

-- Create College_Staging database
CREATE DATABASE College_Completion_Staging;
GO
*/
-- Use College_Staging database
USE College_Completion_Staging;
GO

-- Create Institution table
CREATE TABLE Institution (
    unitid int,
    city varchar(50),
    state varchar(50),
    basic varchar(255),
    chronname varchar(max),
    level varchar(50),
    control varchar(50),
    hbcu bit,
    flagship smallint,
    long_x real,
    lat_y real,
    site varchar(max),
    student_count int,
    awards_per_value real,
    awards_per_state_value real,
    awards_per_natl_value real,
    exp_award_value int,
```

```

exp_award_state_value int,
exp_award_natl_value int,
exp_award_percentile int,
ft_pct real,
fte_value int,
fte_percentile smallint,
med_sat_value real,
grad_100_value real,
grad_100_percentile real,
grad_150_value real,
grad_150_percentile real,
pell_value real,
pell_percentile real,
retain_value real,
retain_percentile real,
cohort_size real
);

-- Create InstitutionGrad table
CREATE TABLE InstitutionGraduation (
    unitid int,
    year int,
    gender varchar(50),
    race varchar(50),
    cohort varchar(255),
    grad_cohort int,
    grad_100 float,
    grad_150 float,
    grad_100_rate float,
    grad_150_rate float
);

-- Create State table
CREATE TABLE State (
    state varchar(255),
    state_abbr varchar(50),
    state_post varchar(50),
    state_code varchar(50)
);

-- Create Location table
CREATE TABLE Location (
    state varchar(255),
    city varchar(50)
);

-- Create Carnegie table with auto-incrementing primary key
CREATE TABLE Carnegie (
    id int IDENTITY(1, 1) PRIMARY KEY,
    institutionType varchar(255),
    programSize varchar(255),
    basic varchar(255)
);

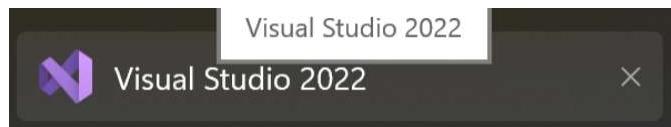
-- Create Cohort table

```

```
CREATE TABLE Cohort (
    cohort varchar(255),
    race varchar(50),
    gender varchar(50),
    year int
);
```

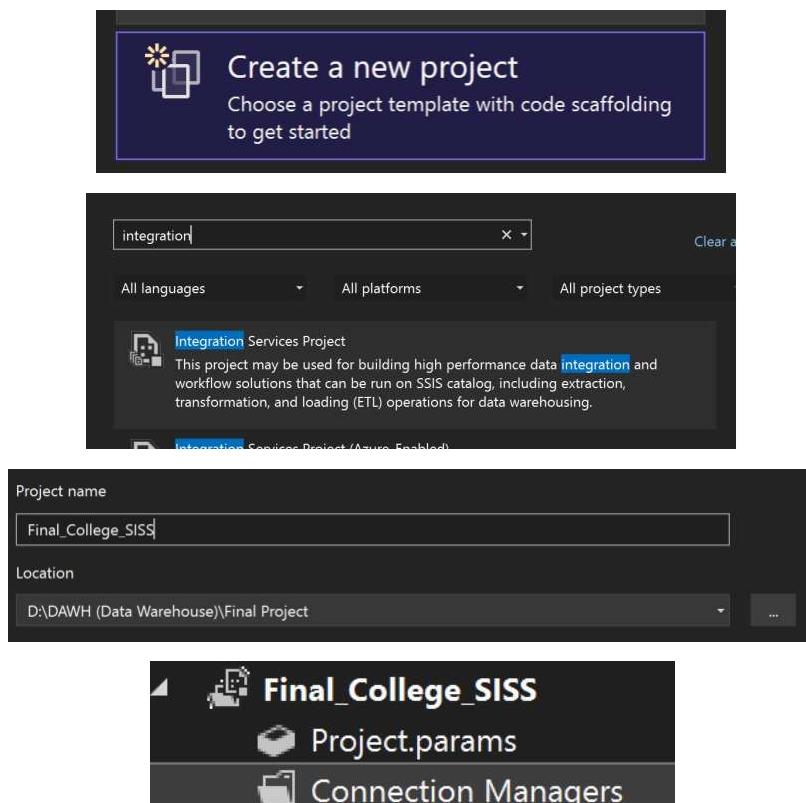
College_Completion_Staging

Mở Visual Studio 2022



College_Completion_Staging

Tạo project với tên “Final_College_SSIS”



Create a new project
Choose a project template with code scaffolding to get started

integration

All languages All platforms All project types

Integration Services Project
This project may be used for building high performance data integration and workflow solutions that can be run on SSIS catalog, including extraction, transformation, and loading (ETL) operations for data warehousing.

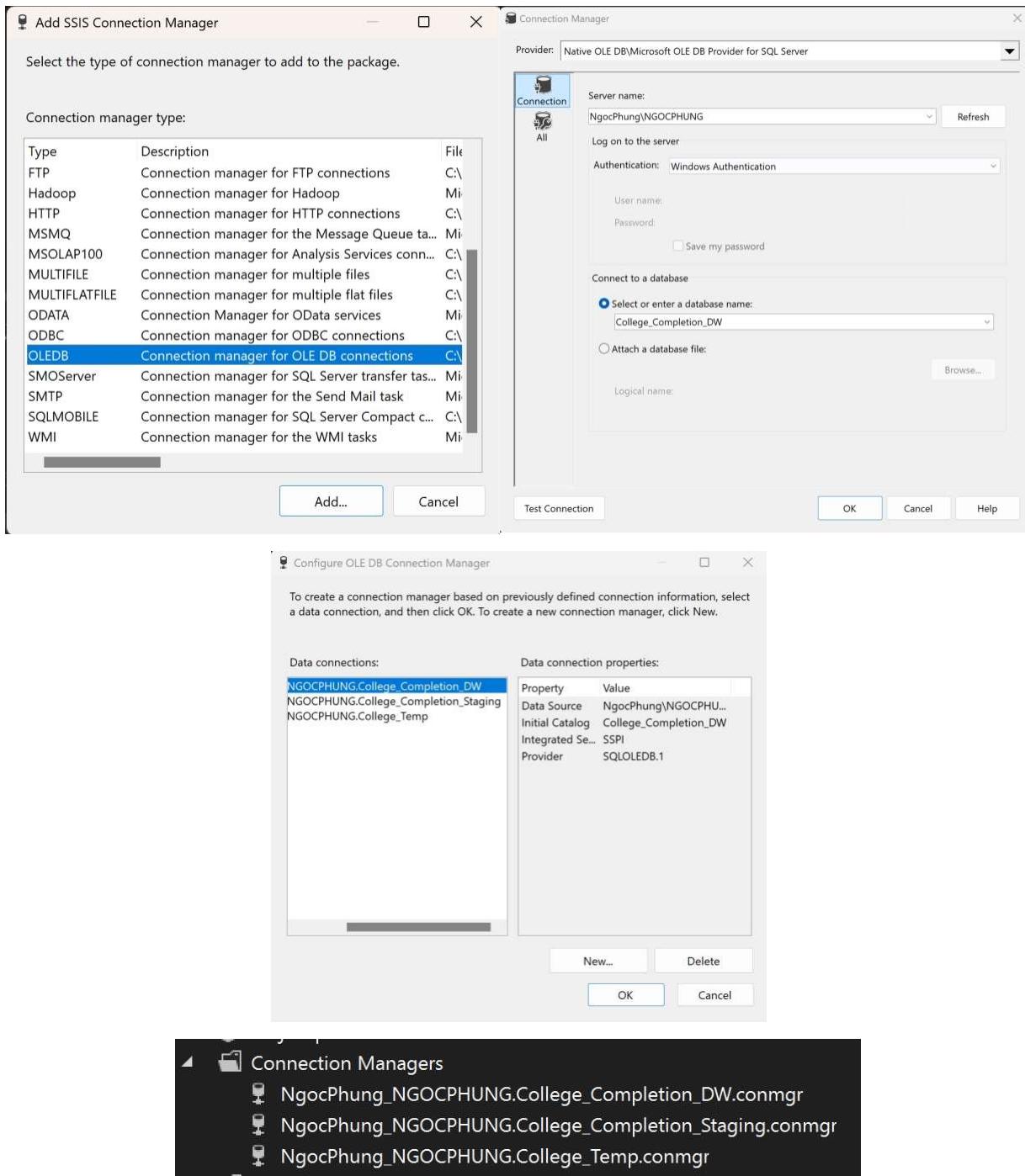
Project name: Final_College_SSIS

Location: D:\DAWH (Data Warehouse)\Final Project

Final_College_SSIS

- Project.params
- Connection Managers

3.3 Tạo kết nối nguồn và đích dữ liệu



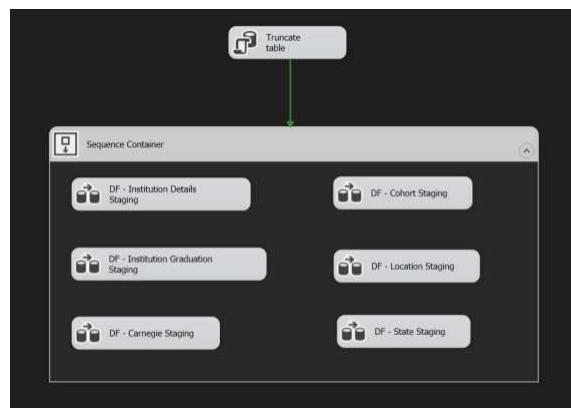
3.4 Tiến hành Staging

Cấu hình package “College_Completioon_Staging”

Tạo package tên “College_Completioon_Staging”



Tạo các data flow

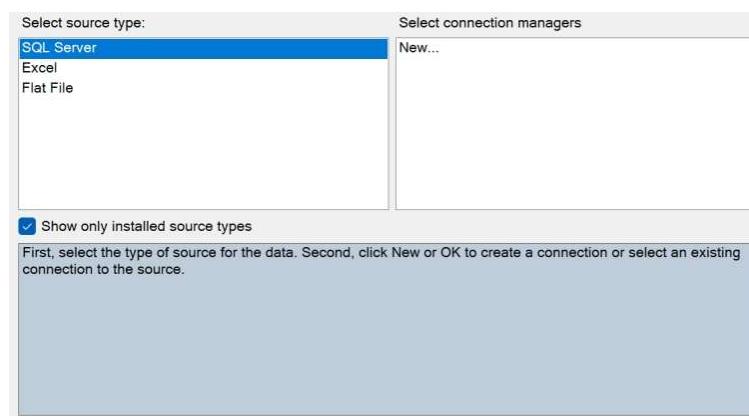


Cấu hình Data Flow Institution Details Staging

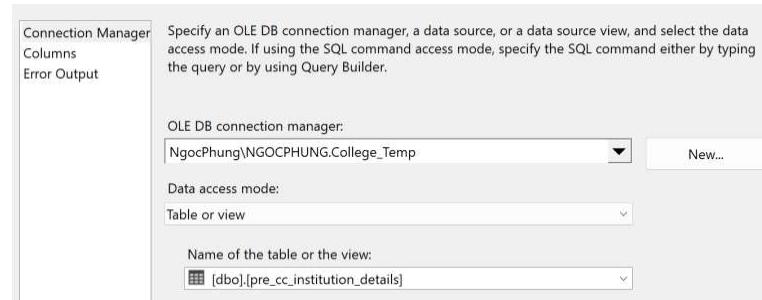


Cấu hình Source Assistant tên “InstitutionDetails - SQL”

Chọn loại source là SQL



Chọn database là College_Temp bảng pre_cc_institution_details



Cấu hình Destination tên “InstitutionDetails”

Connection Manager
Mappings
Error Output

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder. For fast-load data access, set the table update options.

OLE DB connection manager:
NgocPhung_NGOCPHUNG.College_Completion_Staging

Data access mode:
Table or view - fast load

Name of the table or the view:
[dbo].[Institution]

Available Input Columns

Name
unitid
chronname
city
state
level
control
basic
hbcu
flagship
long_x
lat_y
site
student_count
awards_per_value
awards_per_state_value
awards_per_natl_value
exp_award_value
exp_award_state_value
exp_award_natl_value
exp_award_percentile
ft_pct

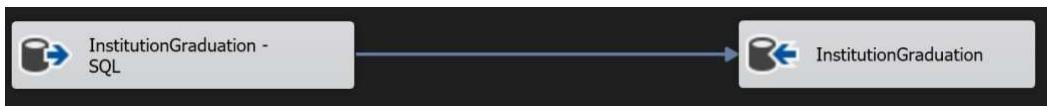
Available Destination Columns

Name
unitid
city
state
basic
chronname
level
control
hbcu
flagship
long_x
lat_y
site
student_count
awards_per_value
awards_per_state_value
awards_per_natl_value
exp_award_value
exp_award_state_value
exp_award_natl_value
exp_award_percentile
ft_pct

Input Column Destination Column

Input Column	Destination Column
unitid	unitid
city	city
state	state
basic	basic

Cấu hình Data Flow Institution Graduation Staging



Cấu hình Source Assistant tên “InstitutionGraduation - SQL”

Connection Manager
Columns
Error Output

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder.

OLE DB connection manager:
NgocPhung_NGOCPHUNG.College_Temp

Data access mode:
Table or view

Name of the table or the view:
[dbo].[pre_institutions_grads]

Available External Columns

External Column	Output Column
unitid	unitid
year	year
gender	gender
race	race
cohort	cohort
grad_cohort	grad_cohort
grad_100	grad_100
grad_150	grad_150

Cấu hình Destination Assistant tên “InstitutionGraduation”

Connection Manager
Mappings
Error Output

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder. For fast-load data access, set the table update options.

OLE DB connection manager:
NgocPhung_NGOCPHUNG.College_Completion_Staging

Data access mode:
Table or view - fast load

Name of the table or the view:
[dbo].[InstitutionGraduation]

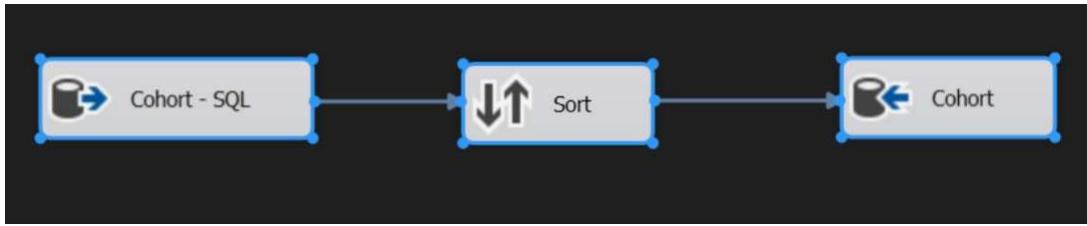
Connection Manager
Mappings
Error Output

Available Input Columns
Name
unitid
year
gender
race
cohort
grad_cohort
grad_100
grad_150
grad_100_rate
grad_150_rate

Available Destination Columns
Name
unitid
year
gender
race
cohort
grad_cohort
grad_100
grad_150
grad_100_rate
grad_150_rate

Input Column	Destination Column
unitid	unitid
year	year
gender	gender
race	race
cohort	cohort
grad_cohort	grad_cohort
grad_100	grad_100
grad_150	grad_150
grad_100_rate	grad_100_rate
grad_150_rate	grad_150_rate

Cấu hình Data Flow Cohort Staging



Cấu hình Source Assistant tên “Cohort - SQL”

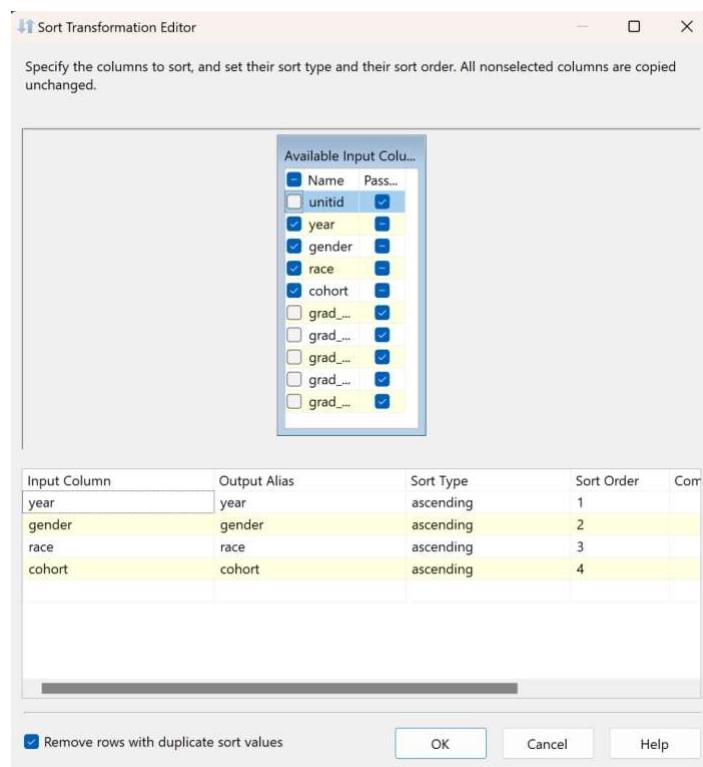
The Source Assistant configuration for the 'Cohort - SQL' component includes:

- Connection Manager:** NgocPhung\NGOCPHUNG.College_Temp
- Columns:** Available External Columns:
 - Name
 - unitid
 - year
 - gender
 - race
 - cohort
 - grad_cohort
 - grad_100
 - grad_150
 - grad_100_rate
 - grad_150_rate
- Error Output:** None

The 'Available External Columns' list shows all columns selected with checkboxes. The 'External Column' and 'Output Column' mapping table shows the following mappings:

External Column	Output Column
unitid	unitid
year	year
gender	gender
race	race
cohort	cohort
grad_cohort	grad_cohort
grad_100	grad_100
grad_150	grad_150
grad_100_rate	grad_100_rate
grad_150_rate	grad_150_rate

Cấu hình “Sort”



Cấu hình Destination Assistant tên “Cohort”

The screenshot shows the 'Destination Assistant' configuration window for the 'Cohort' destination. It includes sections for 'Connection Manager', 'Mappings', and 'Error Output'. The 'Mappings' section displays a mapping grid where input columns from the source are mapped to destination columns. The 'Available Input C...' list contains columns: cohort, gender, grad_100, grad_150, grad_150_rate, grad_cohort, race, unitid, and year. The 'Available Desti...' list contains columns: cohort, race, gender, and year. The mapping grid shows cohort to cohort, race to race, gender to gender, and year to year.

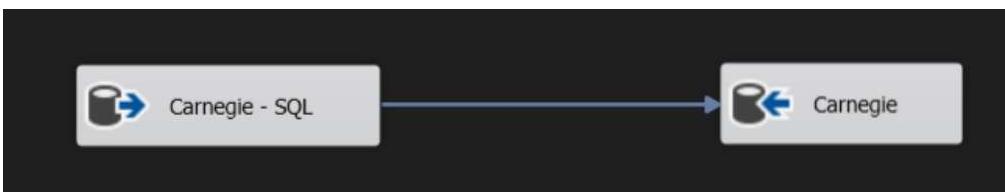
Input Column	Destination Column
cohort	cohort
race	race
gender	gender
year	year

Connection Manager: NgocPhung_NGOCPHUNG.College_Staging

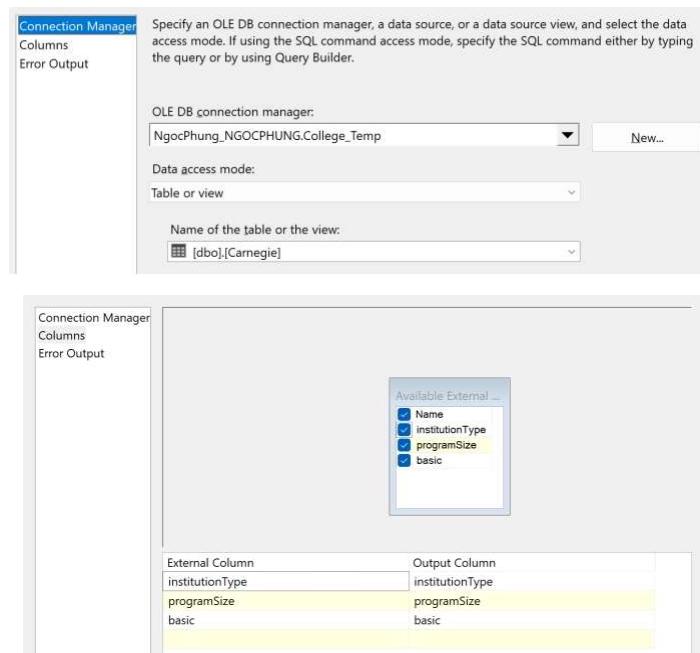
Data access mode: Table or view - fast load

Name of the table or the view: [dbo].[Cohort]

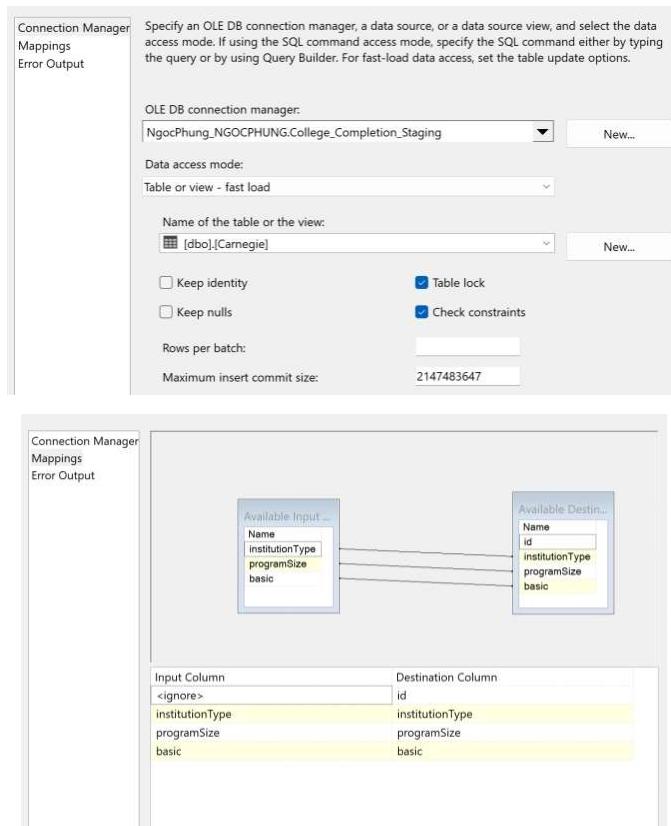
Cấu hình Data Flow Carnegie Staging



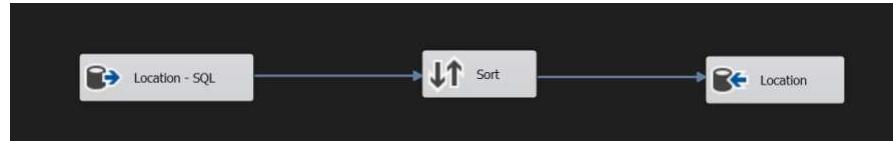
Cấu hình Source Assistant tên “Carnegie - SQL”



Cấu hình Destination Assistant tên “Carnegie”



Cấu hình Data Flow Location Staging



Cấu hình Source Assistant tên “Location - SQL”

Connection Manager
Columns
Error Output

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder.

OLE DB connection manager:
NgocPhung.NGOCPHUNG.College_Temp

Data access mode:
Table or view

Name of the table or the view:
[dbo].[pre_cc_institution_details]

Available External Columns

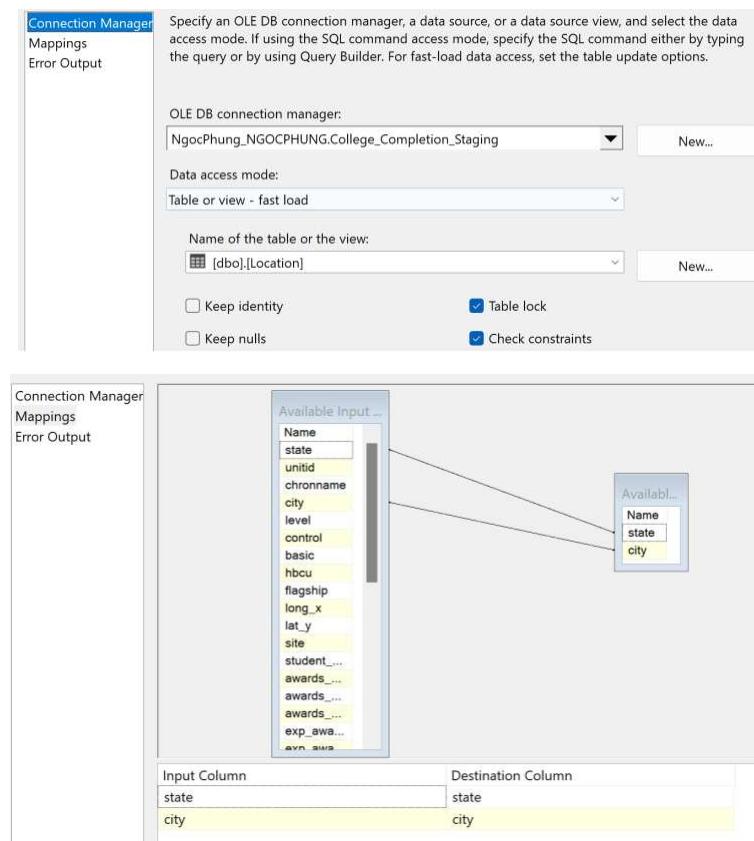
External Column	Output Column
state	state
unitid	unitid
chronname	chronname
city	city
level	level
control	control
basic	basic

Cấu hình Sort

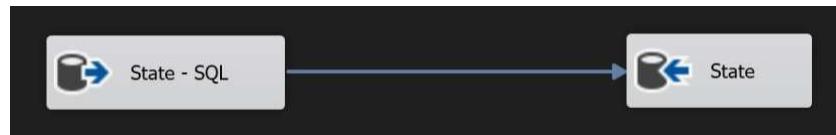
Available Input Columns

Input Column	Output Alias	Sort Type	Sort Order	Com
state	state	ascending	1	
city	city	ascending	2	

Cấu hình Destination Assistant tên “Location”



Cấu hình Data Flow State Staging



Cấu hình Source Assistant tên “State - SQL”

The screenshot shows the 'Source Assistant' configuration window for an OLE DB connection manager named 'NgocPhung_NGOCPHUNG.College_Temp'. The 'Data access mode' is set to 'Table or view', and the 'Name of the table or the view' is '[dbo].[pre_cc_state_sector_details]'. The 'Columns' tab is selected, displaying a list of available external columns: Name, state, state_abbr, state_post, and state_code. These columns are mapped to output columns: state, state_abbr, state_post, and state_code respectively.

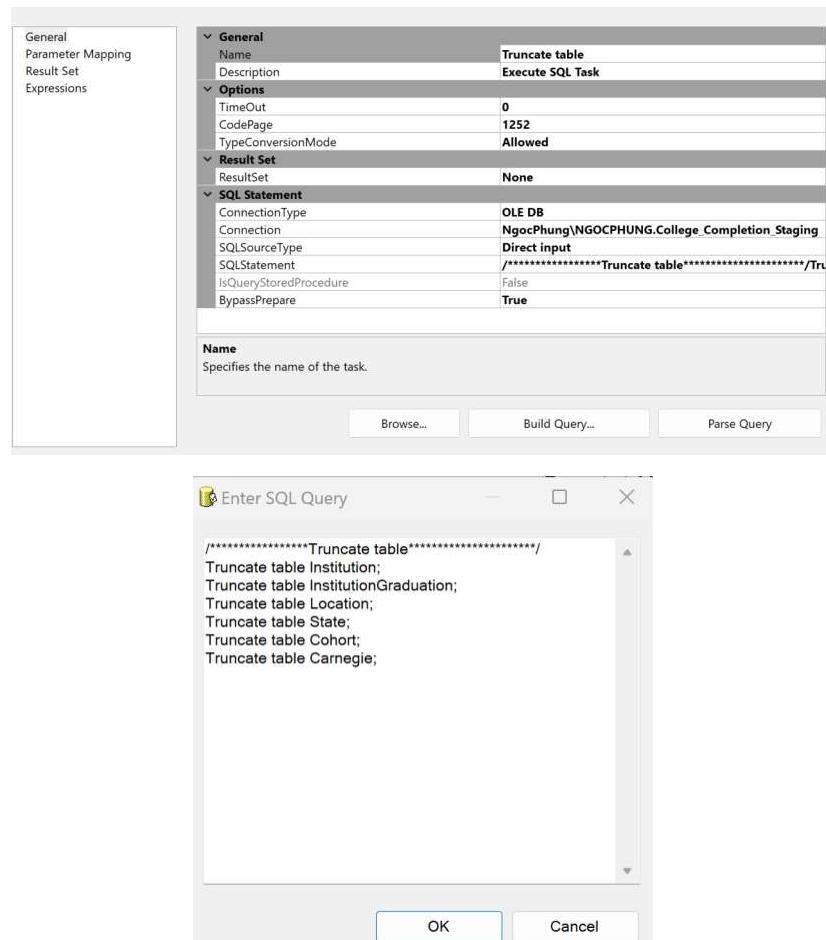
External Column	Output Column
state	state
state_abbr	state_abbr
state_post	state_post
state_code	state_code

Cấu hình Destination Assistant tên “State”

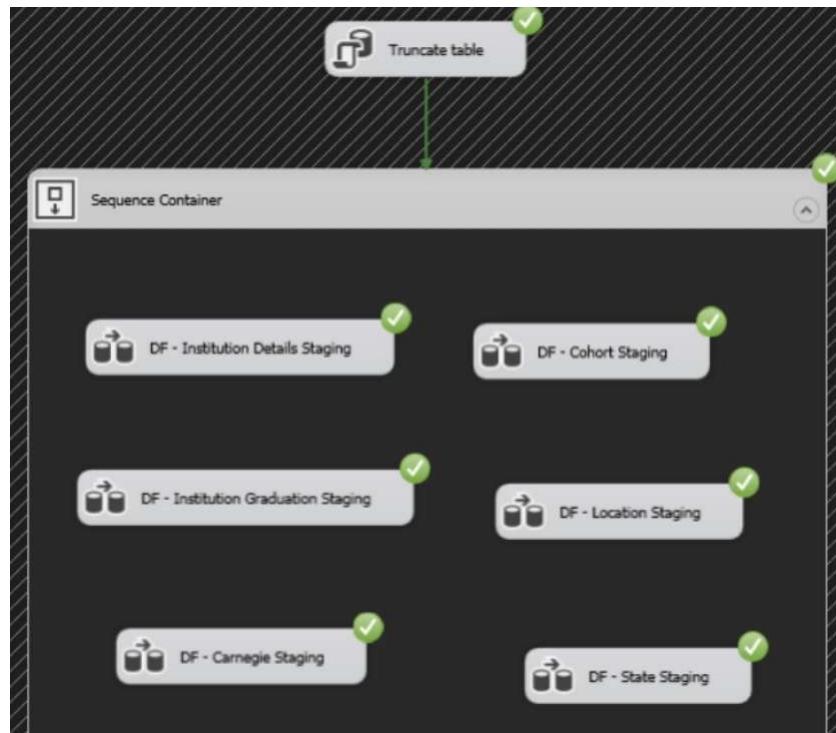
The screenshot shows the 'Destination' configuration window in the SSIS Designer. The left sidebar lists 'Connection Manager', 'Mappings', and 'Error Output'. The main area has a descriptive header: 'Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder. For fast-load data access, set the table update options.' Below this, the 'OLE DB connection manager:' dropdown is set to 'NgocPhung_NGOCPHUNG.College_Completion_Staging'. The 'Data access mode:' dropdown is set to 'Table or view - fast load'. The 'Name of the table or the view:' dropdown is set to '[dbo].[State]'. Underneath these settings are several checkboxes: 'Keep identity' (unchecked), 'Table lock' (checked), 'Keep nulls' (unchecked), and 'Check constraints' (checked). The 'Rows per batch:' input field is empty. The 'Maximum insert commit size:' input field contains the value '2147483647'. At the bottom of the main panel, there are two 'Available In...' and 'Available De...' panes showing column lists for mapping. The 'Mappings' tab is selected, displaying a table with 'Input Column' and 'Destination Column' columns. The table data is as follows:

Input Column	Destination Column
state	state
state_abbr	state_abbr
state_post	state_post
state_code	state_code

Cấu hình Truncate



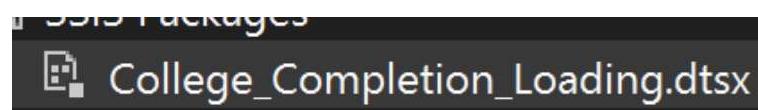
Kết quả của quá trình staging



3.5 Tiến hành Load

Cấu hình package “College_Completion>Loading”

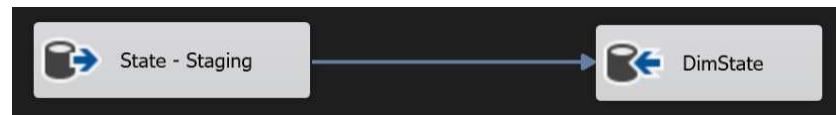
Tạo package tên College_Completioon>Loading



Gồm có các data flow và các truncate



Cấu hình Data Flow DimState



Cấu hình Source Assistant tên “State – Staging”

The screenshot shows the 'Source Assistant' window for the 'State - Staging' source. On the left, there are tabs for 'Connection Manager', 'Columns', and 'Error Output'. The main area displays configuration options for an OLE DB connection manager:

- 'OLE DB connection manager:' dropdown set to 'NgocPhung_NGOCPHUNG.College_Completion_Staging' with a 'New...' button.
- 'Data access mode:' dropdown set to 'Table or view'.
- 'Name of the table or the view:' dropdown set to '[dbo].[State]'.

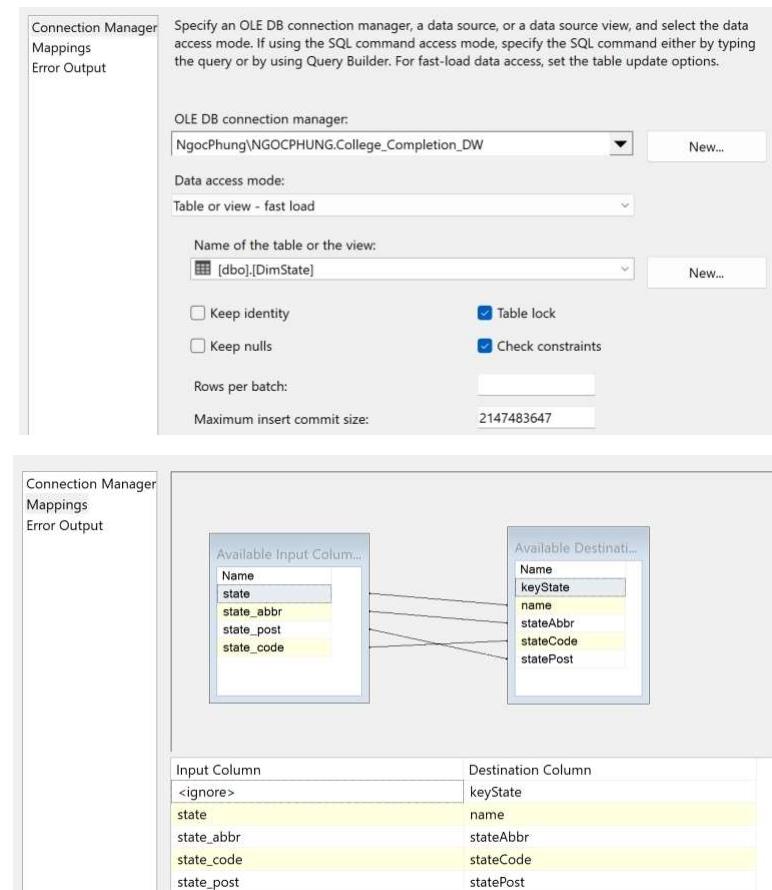
On the right, under 'Available External...', a list of columns is shown:

- Name
- state
- state_abbr
- state_post
- state_code

Below this, a mapping grid shows the 'External Column' and 'Output Column' for each row:

External Column	Output Column
state	state
state_abbr	state_abbr
state_post	state_post
state_code	state_code

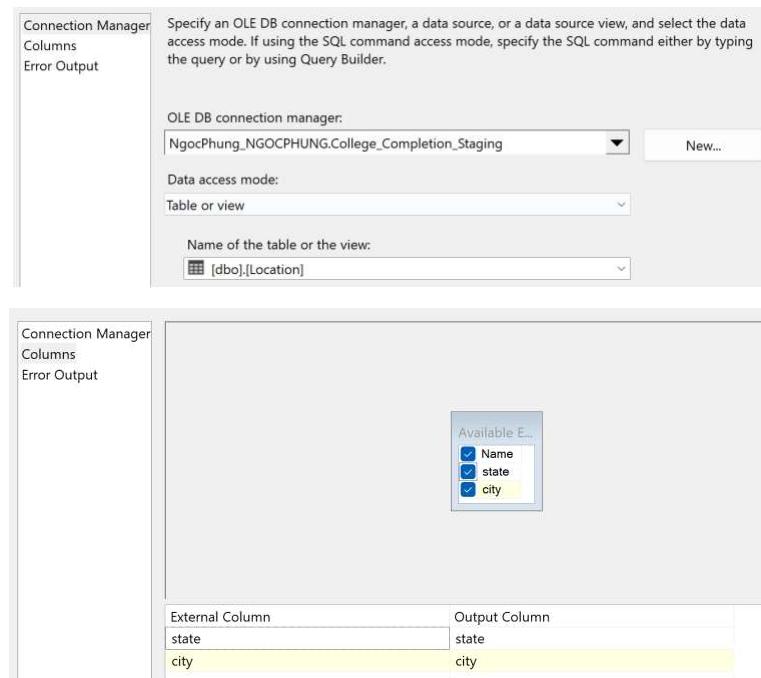
Cấu hình Destination Assisstant tên “DimState”



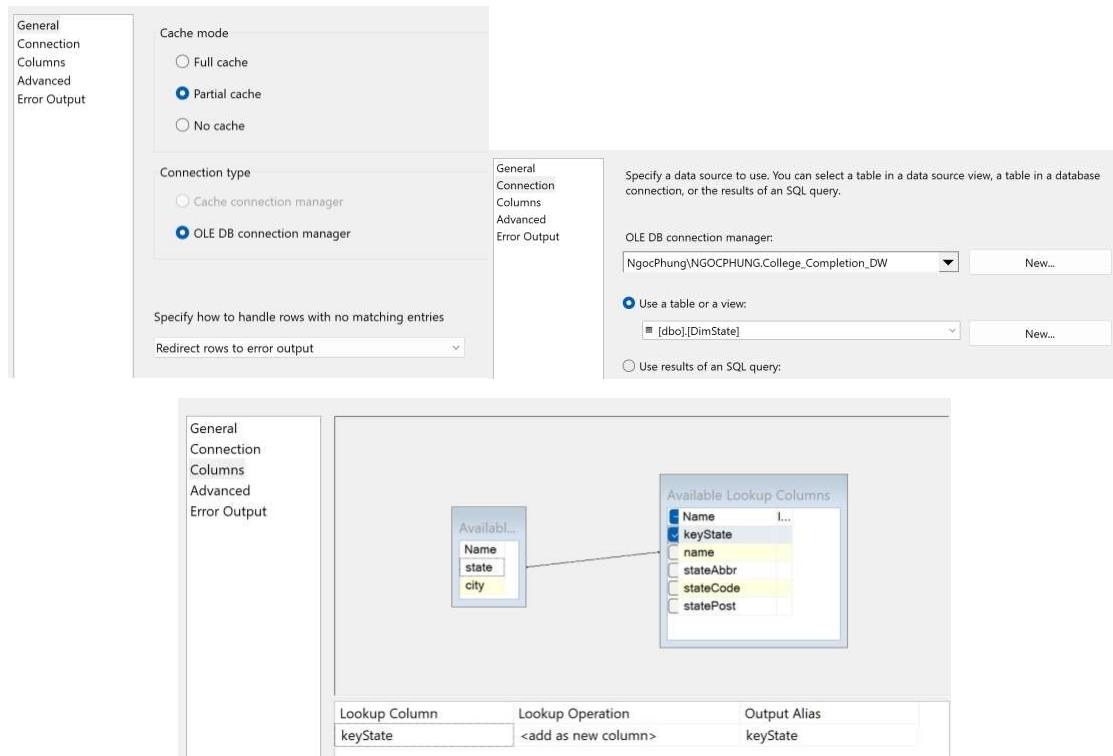
Cấu hình Data Flow DimLocation



Cấu hình Source Assisstant tên “Location - Staging”



Cấu hình Lookup



Cấu hình Sort

The screenshot shows the configuration for a Sort transformation. On the left, there is a list of available input columns: Name, Pass..., state, city, and keyState. The 'state' and 'city' columns are selected and highlighted in yellow. In the main pane, there is a table with five columns: Input Column, Output Alias, Sort Type, Sort Order, and Collation. The rows are:

Input Column	Output Alias	Sort Type	Sort Order	Collation
state	state	ascending	1	
city	city	ascending	2	

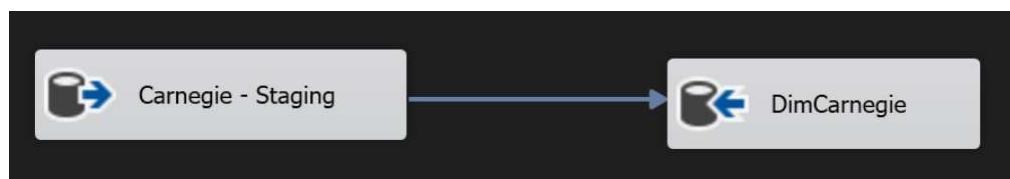
Cấu hình Destination Assisstant tên “DimLocation”

The screenshot shows the configuration for a destination table named [dbo].[DimLocation]. The 'OLE DB connection manager' is set to 'NgocPhung.NGOCPHUNG.College_Completion_DW'. The 'Data access mode' is 'Table or view - fast load'. The 'Name of the table or the view' is '[dbo].[DimLocation]'. Under 'Table lock' options, 'Keep identity' is unchecked and 'Table lock' is checked. Under 'Check constraints', 'Keep nulls' is unchecked and 'Check constraints' is checked. The 'Rows per batch:' field is empty, and the 'Maximum insert commit size:' field contains the value '2147483647'.

Below this, the 'Mappings' tab is selected in the left sidebar. It shows the mapping between the input columns 'Name', 'state', and 'city' and the destination columns 'keyLocation', 'keyState', and 'city'. The 'keyState' and 'city' mappings are highlighted in yellow.

Input Column	Destination Column
<ignore>	keyLocation
keyState	keyState
city	city

Cấu hình Data Flow DimCarnegie



Cấu hình Destination Assisstant tên “DimLocation”

Connection Manager
Columns
Error Output

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder.

OLE DB connection manager: NgocPhung_NGOCPHUNG.College_Completion_Staging New...

Data access mode: Table or view

Name of the table or the view: [dbo].[Carnegie]

Connection Manager
Columns
Error Output

External Column	Output Column
id	id
institutionType	institutionType
programSize	programSize
basic	basic

Cấu hình Destination Assisstant tên “DimCarnegie”

Connection Manager
Mappings
Error Output

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder. For fast-load data access, set the table update options.

OLE DB connection manager: NgocPhung_NGOCPHUNG.College_Completion_DW New...

Data access mode: Table or view - fast load

Name of the table or the view: [dbo].[DimCarnegie] New...

Keep identity Table lock

Keep nulls Check constraints

Rows per batch:

Maximum insert commit size: 2147483647

Connection Manager
Mappings
Error Output

Input Column	Destination Column
<ignore>	keyCarnegie
institutionType	institutionType
programSize	programSize

Cấu hình Data Flow DimCohort



Cấu hình Source Assisstant tên “Cohort - Staging”

The screenshot shows the 'Source' configuration dialog in SSIS. On the left, there are tabs for 'Connection Manager', 'Columns', and 'Error Output'. The 'Connection Manager' tab is selected, showing a dropdown for 'OLE DB connection manager' set to 'NgocPhung.NGOCPHUNG.College_Completion_Staging' and a 'Data access mode' dropdown set to 'Table or view'. Below these are fields for 'Name of the table or the view' containing '[dbo].[Cohort]' and a preview icon.

The main area displays a mapping grid:

External Column	Output Column
cohort	cohort
race	race
gender	gender
year	year

A 'Available External...' button is visible above the grid, and a preview window on the right shows the column names: Name, cohort, race, gender, and year.

Cấu hình Destination Assisstant tên “DimCohort”

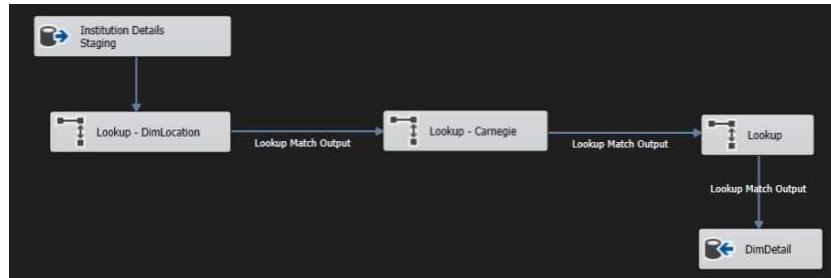
The screenshot shows the 'Destination' configuration dialog in SSIS. On the left, there are tabs for 'Connection Manager', 'Mappings', and 'Error Output'. The 'Connection Manager' tab is selected, showing a dropdown for 'OLE DB connection manager' set to 'NgocPhung.NGOCPHUNG.College_Completion_DW' and a 'Data access mode' dropdown set to 'Table or view - fast load'. Below these are fields for 'Name of the table or the view' containing '[dbo].[DimCohort]', checkboxes for 'Keep identity', 'Table lock', 'Keep nulls', and 'Check constraints', and input fields for 'Rows per batch' and 'Maximum insert commit size'.

The main area displays a mapping grid:

Input Column	Destination Column
<ignore>	keyCohort
race	race
cohort	cohort
gender	gender
year	year

Two 'Available...' buttons are visible above the grid, and a preview window on the right shows the destination columns: keyCohort, race, cohort, gender, and year.

Cấu hình Data Flow FactDetail



Cấu hình Source Assisstant tên “Institution Details - Staging”

The screenshot shows the 'Source Assistant' configuration window for the 'Institution Details - Staging' source. The 'Connection Manager' tab is selected, showing the connection manager 'NgocPhung_NGOCPHUNG.College_Completion_Staging' and the data access mode 'Table or view' set to '[dbo].[Institution]'. The 'Columns' tab is also visible, displaying a list of columns from the external table: Name, unitid, city, state, basic, chronname, level, control, hbcu, flagship, long_x, lat_y, and site. These columns are mapped to the corresponding columns in the output table.

External Column	Output Column
unitid	unitid
city	city
state	state
basic	basic
chronname	chronname
level	level
control	control
hbcu	hbcu
flagship	flagship
long_x	long_x
lat_y	lat_y
site	site

Cấu hình Lookup – DimLocation

The screenshot shows the 'General' tab of a configuration dialog for a Lookup transformation. On the left, a sidebar lists tabs: General, Connection, Columns, Advanced, and Error Output. The 'General' tab is selected.

Cache mode: The 'Full cache' radio button is selected.

Connection type: The 'OLE DB connection manager' radio button is selected.

Specify how to handle rows with no matching entries: The dropdown menu is set to 'Redirect rows to error output'.

The screenshot shows the 'Connection' tab of the configuration dialog. The sidebar shows tabs: General, Connection, Columns, Advanced, and Error Output. The 'Connection' tab is selected.

OLE DB connection manager: A dropdown menu is open, showing 'NgocPhung_NGOCPHUNG.College_Completion_DW' as the selected item. There is also a 'New...' button.

Use a table or a view: This option is not selected.

Use results of an SQL query: This option is selected. A SQL query is displayed in the text area:

```
SELECT DimState.name, DimLocation.city,
DimLocation.keyLocationFROM DimLocation INNER JOIN
DimState ON DimLocation.keyState =
DimState.keyState
```

Next to the query, there are three buttons: 'Build Query...', 'Browse...', and 'Parse Query'.

The screenshot shows the 'Columns' tab of the configuration dialog. The sidebar shows tabs: General, Connection, Columns, Advanced, and Error Output. The 'Columns' tab is selected.

Available columns: A list of columns from the source table is shown, including 'unitid', 'city', 'state', 'basic', 'chr...', 'level', 'con...', 'hbcu', 'flag...', 'lon...', 'lat...', and '...'. The 'city' column is highlighted.

Available Lookups: A list of columns from the target table is shown, including 'name', 'basic', 'chr...', 'level', 'con...', 'hbcu', 'flag...', 'lon...', 'lat...', and '...'. The 'name' column is highlighted.

Lookup Column: 'keyLocation' is selected.

Lookup Operation: '<add as new column>' is selected.

Output Alias: 'keyLocation' is selected.

Cấu hình Lookup – Carnegie

The screenshot shows the configuration interface for a Lookup transformation in SSIS. The left sidebar lists tabs: General, Connection, Columns, Advanced, and Error Output. The main area is divided into several sections:

- Cache mode:** Partial cache (selected).
- Connection type:** OLE DB connection manager (selected).
- Specify how to handle rows with no matching entries:** Redirect rows to no match output.
- General tab (Connection):** Shows the connection manager dropdown set to "NgocPhung_NGOCPHUNG.College_Completion_Staging".
- Connection tab (OLE DB connection manager):** Shows the table selection dropdown set to "[dbo].[Carnegie]".
- Columns tab:** Displays two columns: "programSize" and "institutionType". For "programSize", the "Lookup Column" is "programSize", "Lookup Operation" is "<add as new column>", and "Output Alias" is "programSize". For "institutionType", the "Lookup Column" is "institutionType", "Lookup Operation" is "<add as new column>", and "Output Alias" is "institutionType".

Cấu hình Lookup

The screenshot shows the configuration interface for a Lookup transformation in SSIS. The left sidebar lists tabs: General (selected), Connection, Columns, Advanced, and Error Output.

Cache mode: Full cache (selected).

Connection type: OLE DB connection manager (selected).

Specify how to handle rows with no matching entries: Redirect rows to no match output (selected).

Connection Tab:

- Specify a data source to use. You can select a table in a data source view, a table in a database connection, or the results of an SQL query.
- OLE DB connection manager: NgocPhung_NGOCPHUNG.College_Completion_DW (selected).
- Use a table or a view: [dbo].[DimCarnegie] (selected).

Columns Tab:

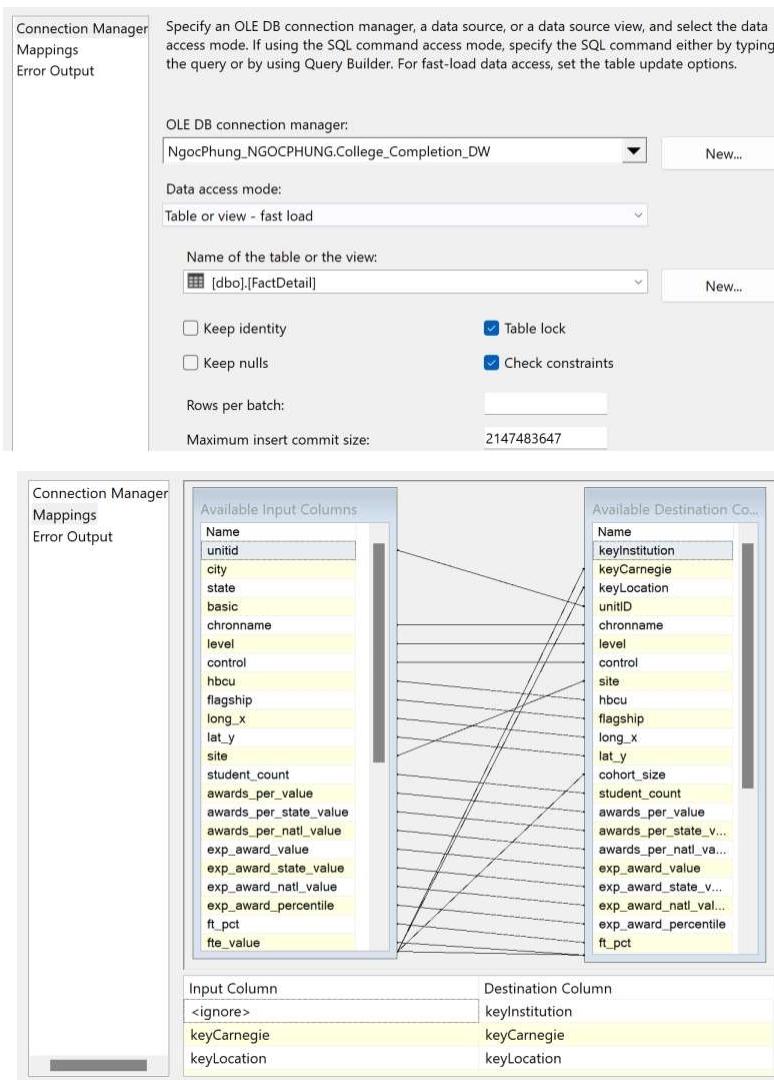
- Available Input Columns: awards_per_value, awards_per_state_value, awards_per_nati_value, exp_award_value, exp_award_state_value, exp_award_nati_value, exp_award_percentile, ft_pct, fts_value, fts_percentile, med_pct_value, grad_100_value, grad_100_percentile, grad_150_value, grad_150_percentile, pell_value, pell_percentile, retain_value, retain_percentile, cohort_size, keyLocation, programSize, institutionType.
- Available Lookup Columns: Name, keyCarnegie, institutionType, programSize.

Lookup Column: keyCarnegie

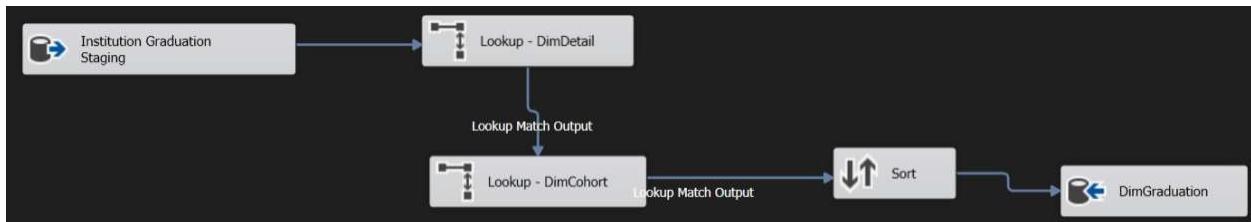
Lookup Operation: <add as new column>

Output Alias: keyCarnegie

Cáu hình Destination Assisstant tên “DimDetail”



Cáu hình Data Flow FactGraduation



Cấu hình Destination Assisstant tên “DimDetail”

The screenshot shows the 'Destination' configuration dialog box in the SSIS Designer. The left pane lists 'Connection Manager', 'Columns', and 'Error Output'. The main area has three tabs: 'General', 'Columns', and 'Error Output'. The 'General' tab is selected, showing the following settings:

- OLE DB connection manager:** NgocPhung_NGOCPHUNG.College_Completion_Staging
- Data access mode:** Table or view
- Name of the table or the view:** [dbo].[InstitutionGraduation]

The 'Columns' tab displays a list of available external columns from the source table:

Available External ...
Name
unitid
year
gender
race
cohort
grad_cohort
grad_100
grad_150
grad_100_rate

The 'Columns' tab also shows the mapping between external columns and output columns:

External Column	Output Column
unitid	unitid
year	year
gender	gender
race	race
cohort	cohort
grad_cohort	grad_cohort
grad_100	grad_100
grad_150	grad_150
grad_100_rate	grad_100_rate
grad_150_rate	grad_150_rate

Cấu hình Lookup – DimDetail

General

- Connection
- Columns
- Advanced
- Error Output

Cache mode

- Full cache
- Partial cache
- No cache

Connection type

- Cache connection manager
- OLE DB connection manager

Specify how to handle rows with no matching entries

Redirect rows to no match output

Connection

Specify a data source to use. You can select a table in a data source view, a table in a database connection, or the results of an SQL query.

OLE DB connection manager:

NgocPhung_NGOCPHUNG.College_Completion_DW

New...

Use a table or a view:

[dbo].[FactDetail]

New...

Use results of an SQL query:

General

Available Input ...

Name
unitid
year
gender
race
cohort
grad_cohort
grad_100
grad_150
grad_100_rate
grad_150_rate

Available Lookup Colu...

Name
keyInstitution
keyCarnegie
keyLocation
unitID
chromename
level
control
site
hbcu
flagship
long_x
lat_y

Lookup Column Lookup Operation Output Alias

keyInstitution <add as new column> keyInstitution

Cấu hình Lookup – DimCohort

The screenshot shows three configuration panels for a SQL Server Integration Services (SSIS) lookup transformation.

Top Panel: General tab settings.

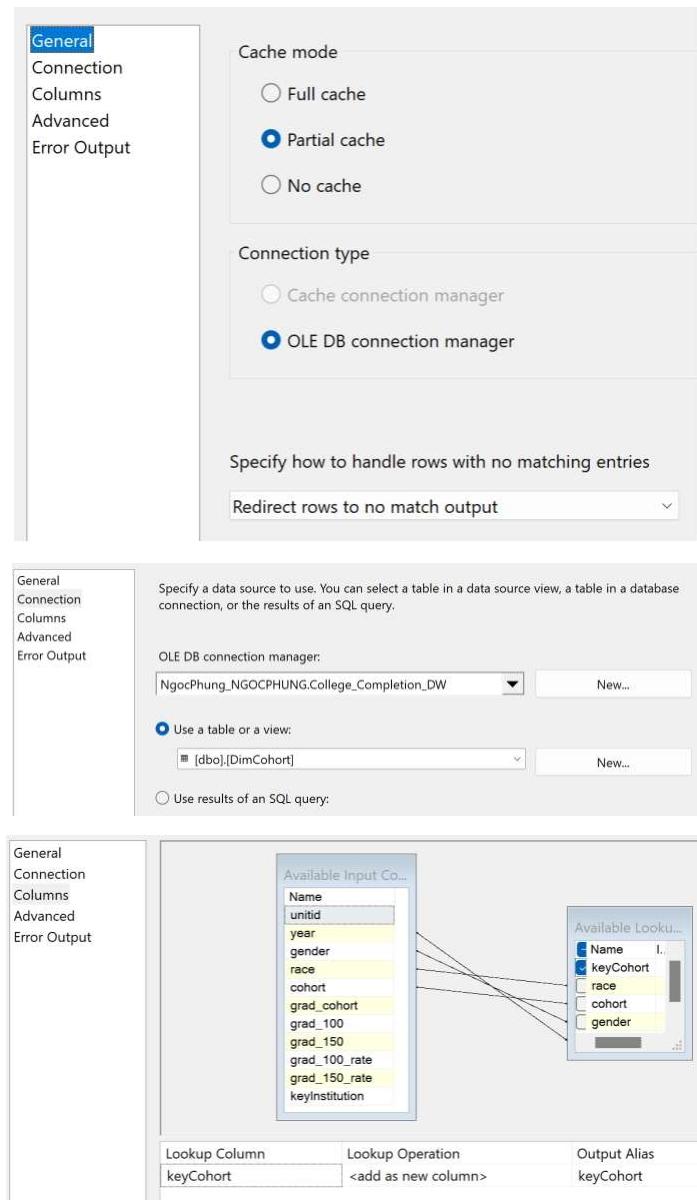
- Cache mode:** Partial cache (selected).
- Connection type:** OLE DB connection manager (selected).
- Specify how to handle rows with no matching entries:** Redirect rows to no match output (selected).

Middle Panel: Connection tab settings.

- Data source:** Specified as a table in a database connection.
- OLE DB connection manager:** NgocPhung_NGOCPHUNG.College_Completion_DW (selected).
- Use a table or a view:** [dbo].[DimCohort] (selected).
- Use results of an SQL query:** (unchecked).

Bottom Panel: Columns tab settings.

- Available Input Columns:** A list of columns including unitid, year, gender, race, cohort, grad_cohort, grad_100, grad_150, grad_100_rate, grad_150_rate, and keyInstitution.
- Available Lookups:** A list of lookups including keyCohort, race, cohort, and gender.
- Lookup Column:** keyCohort.
- Lookup Operation:** <add as new column>
- Output Alias:** keyCohort.



Cấu hình Sort

The screenshot shows the configuration for a Sort transformation. On the left, there is a list of available input columns with checkboxes next to them. Below this is a table mapping input columns to output aliases and sort types.

Input Column	Output Alias	Sort Type	Sort Order
keyInstitution	Sort.keyInstitution	ascending	1
keyCohort	keyCohort	ascending	2

Cấu hình Destination Assistant tên “DimGraduation”

The screenshot shows the configuration for a destination table named [dbo].[FactGraduation]. It includes settings for OLE DB connection manager, data access mode (Table or view - fast load), and various insert options like Keep identity, Keep nulls, Table lock, and Check constraints.

OLE DB connection manager: NgocPhung_NGOCPHUNG.College_Completion_DW

Data access mode: Table or view - fast load

Name of the table or the view: [dbo].[FactGraduation]

Insert Options:

- Keep identity
- Table lock
- Keep nulls
- Check constraints

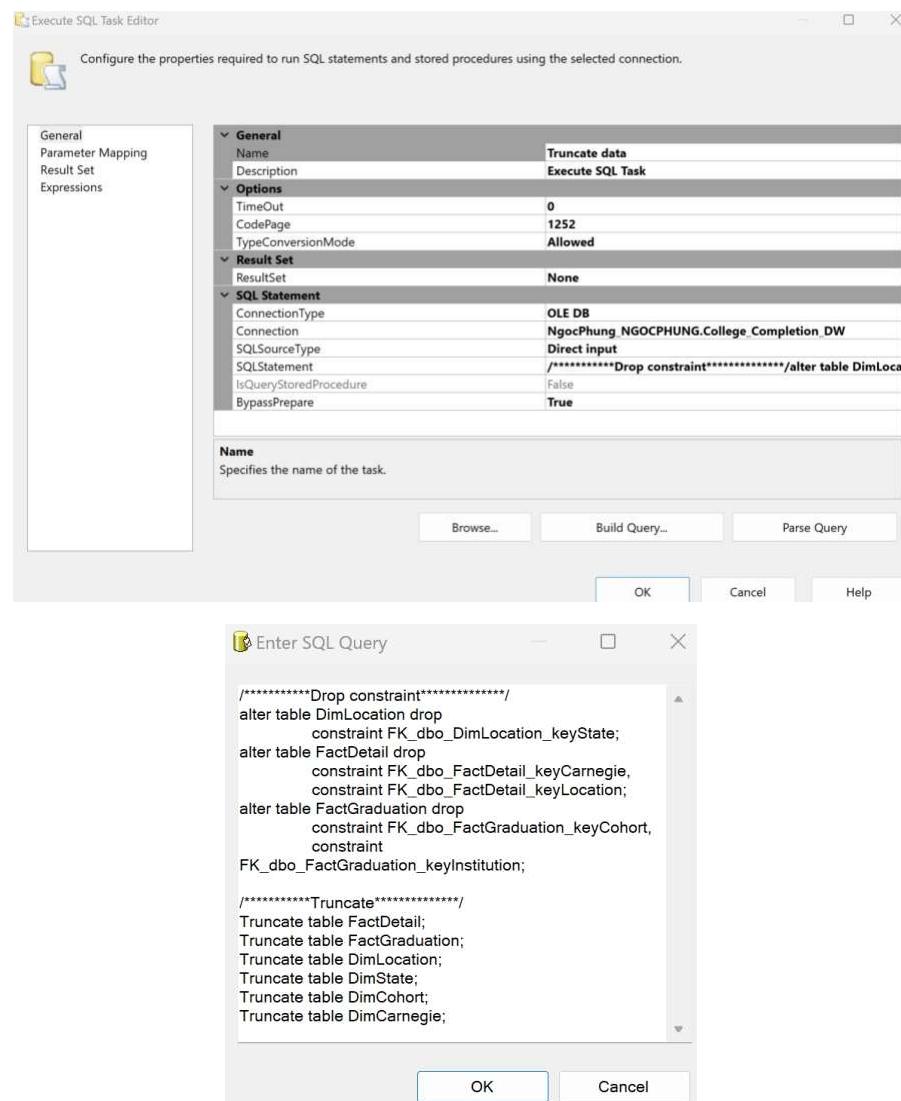
Rows per batch: 1000000

Maximum insert commit size: 2147483647

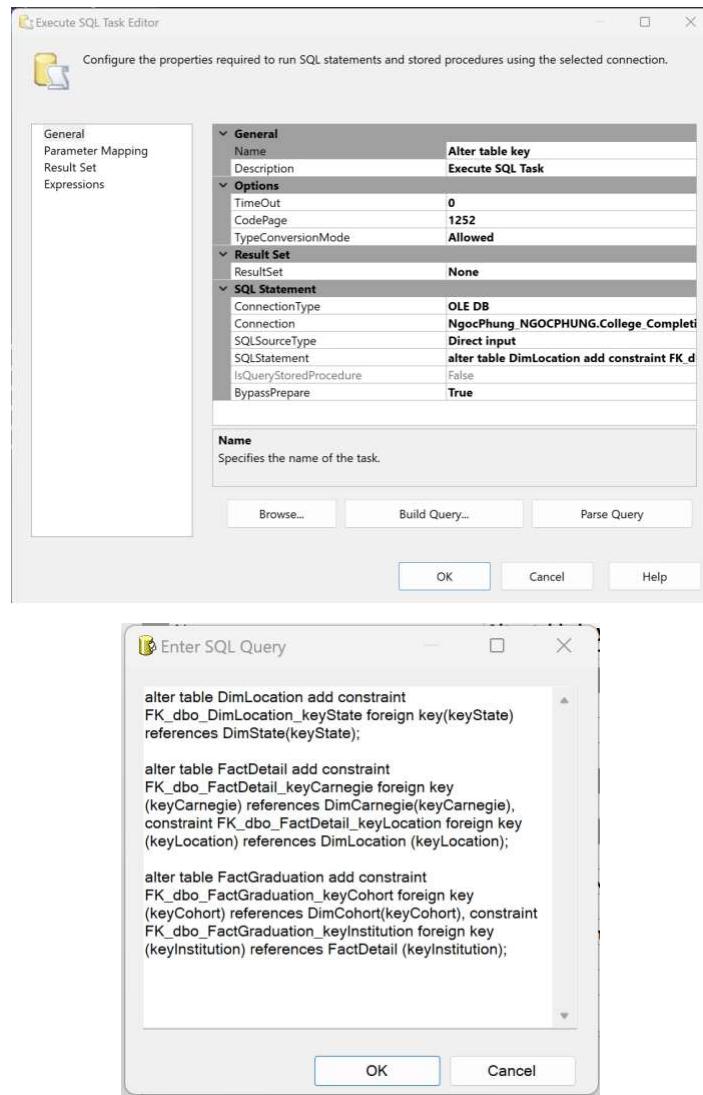
This screenshot shows the mapping between input columns from the previous Sort transformation and destination columns in the [dbo].[FactGraduation] table. The 'Available Input Columns' list on the left maps to the 'Available Destination Columns' list on the right via lines connecting them.

Input Column	Destination Column
Sort.keyInstitution	keyInstitution
keyCohort	keyCohort
grad_cohort	cohort_size
grad_100	grad_100
grad_150	grad_150
grad_100_rate	grad_100_rate
grad_150_rate	grad_150_rate

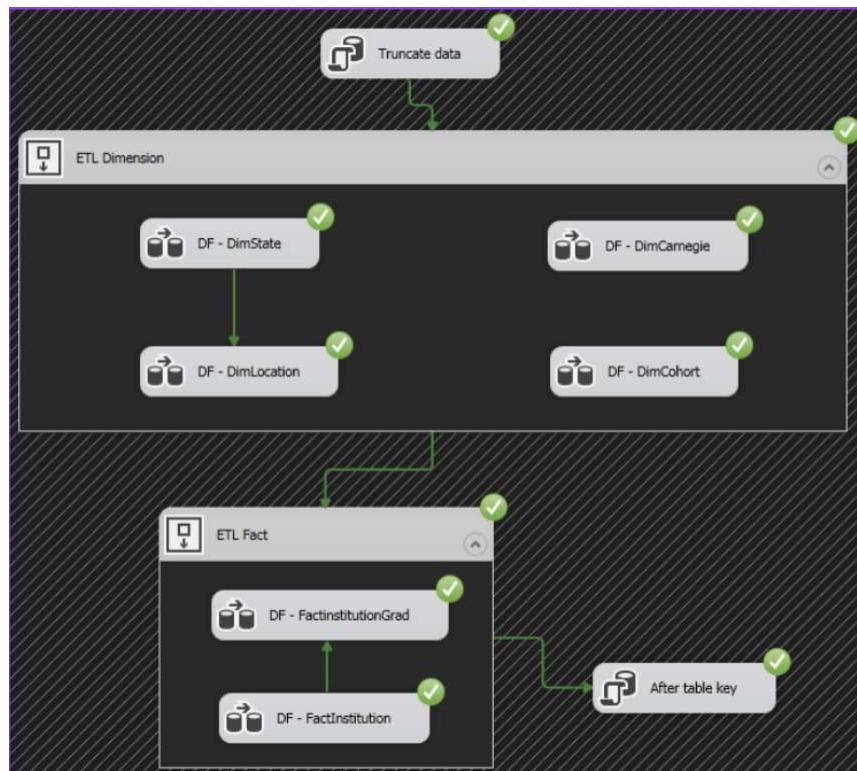
Cấu hình Truncate “Truncate data”



Câu hình Truncate “Alter table key”



Kết quả loading



CHƯƠNG 4: PHÂN TÍCH DỮ LIỆU (SSAS)

4.1 Câu hỏi truy vấn

Từ dữ liệu đã được xử lý và tích hợp, người phân tích dữ liệu đặt ra các câu hỏi truy vấn để tiến hành phân tích

Câu 1: Tỉ lệ sinh viên theo học toàn thời gian tại các khu vực

Câu 2: Tỉ lệ sinh viên được nhận trợ cấp Pell tại mỗi khu vực

Câu 3: Tỉ lệ tốt nghiệp đúng hạn theo thời gian học tại các khu vực

Câu 4: Tỉ lệ sinh viên được giữ lại sau năm học thứ nhất của mỗi khu vực

Câu 5: Tổng số sinh viên và chi phí ước tính để hoàn tất khoá học của mỗi sinh viên

Câu 6: Số lượng sinh viên tốt nghiệp để tốt nghiệp theo sắc tộc và giới tính

Câu 7: Số lượng sinh viên tốt nghiệp đúng hạn theo từng khu vực

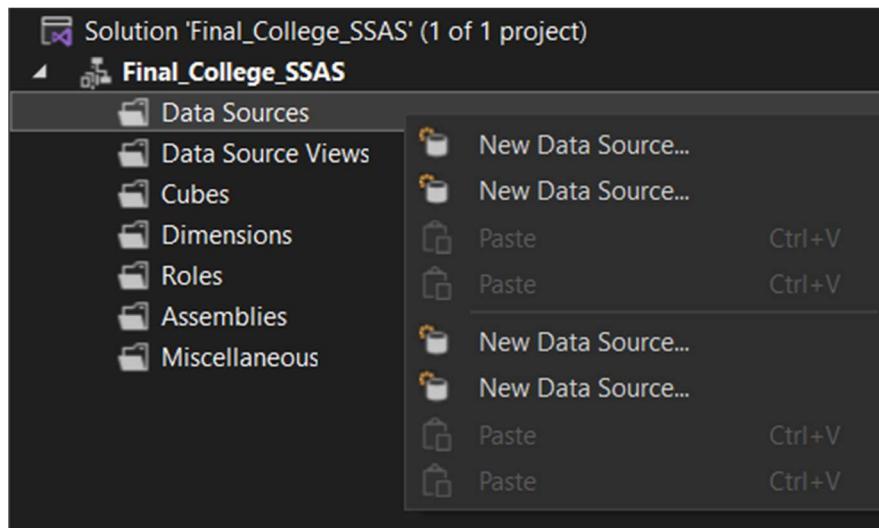
Câu 8: Số lượng sinh viên lấy bằng sau 150% thời gian học tiêu chuẩn theo giới tính và sắc tộc của từng năm

Câu 9: Số lượng sinh viên tốt nghiệp theo từng năm, phân theo loại trường học và cohort

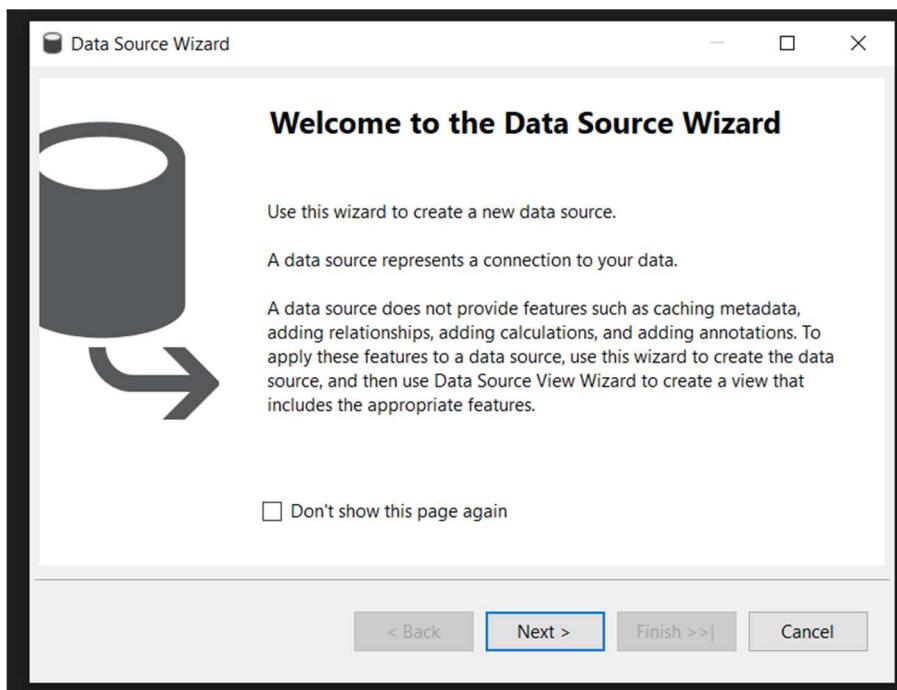
Câu 10: Số lượng sinh viên tốt nghiệp theo sắc tộc và giới tính, phân theo quy mô chương trình học

4.2 Xây dựng mô hình

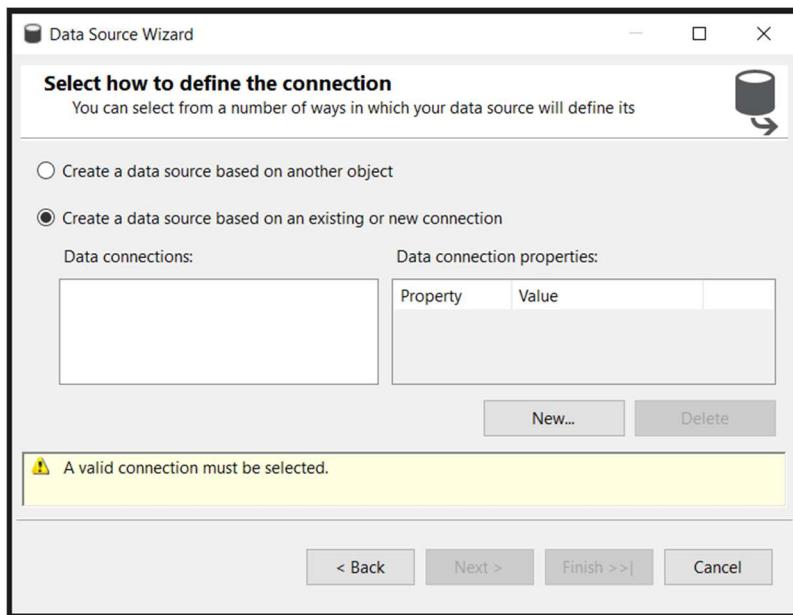
Chọn vào Data Sources, sau đó chọn New Data Source...



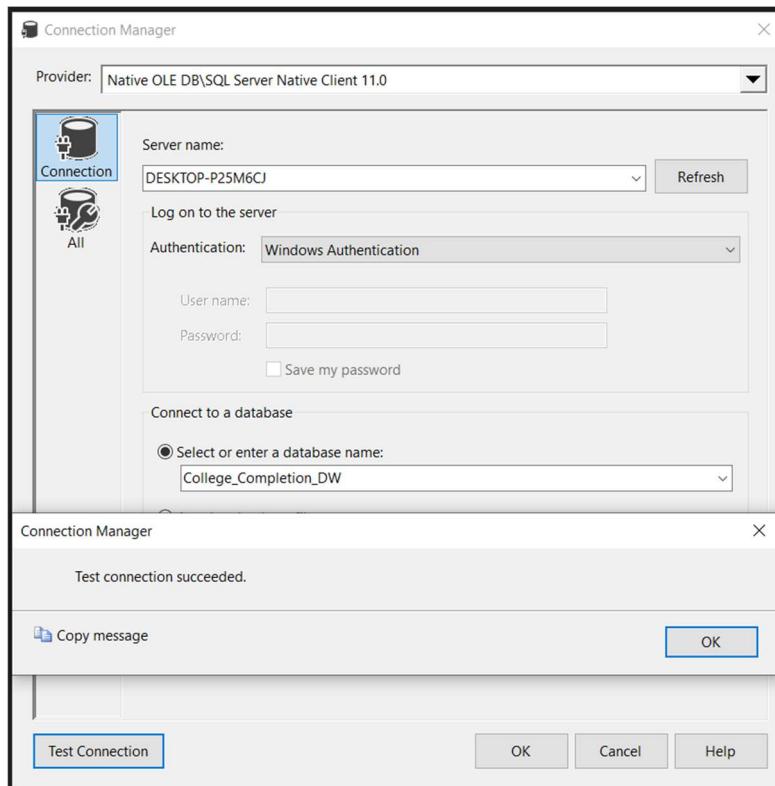
Cửa sổ Data Source Wizard hiện ra, chọn Next



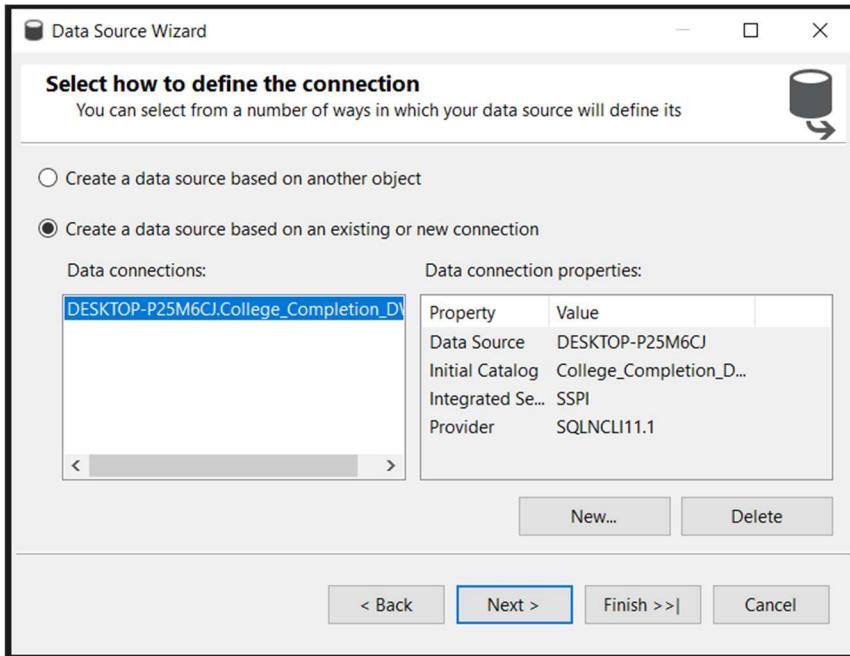
Chọn Create a data source based on an existing or new connection, sau đó chọn New



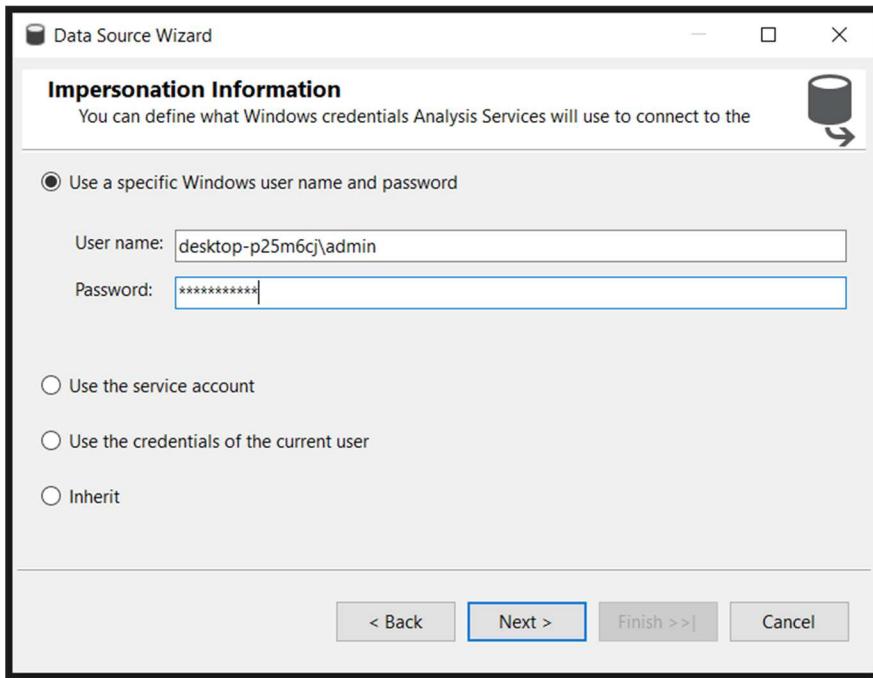
Cửa sổ Connection Manager hiện ra, điền vào tên server tương ứng và chọn tên của database, sau đó chọn Test Connection để kiểm tra kết nối, sau đó chọn OK



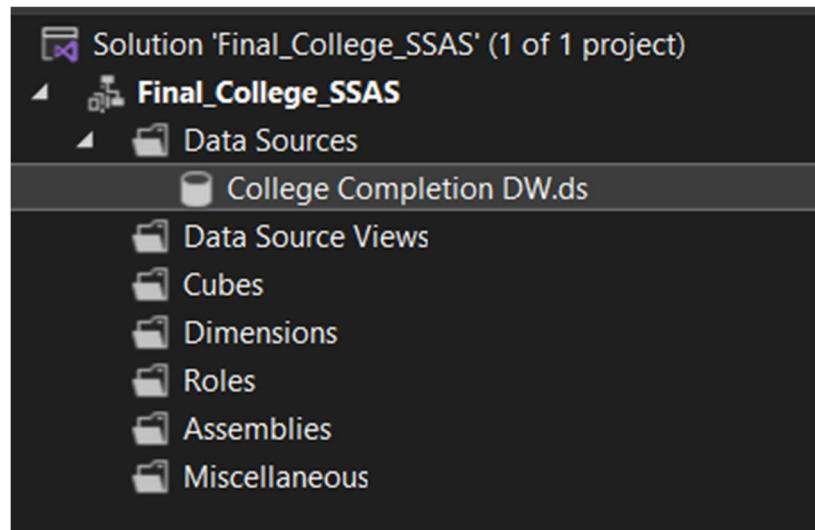
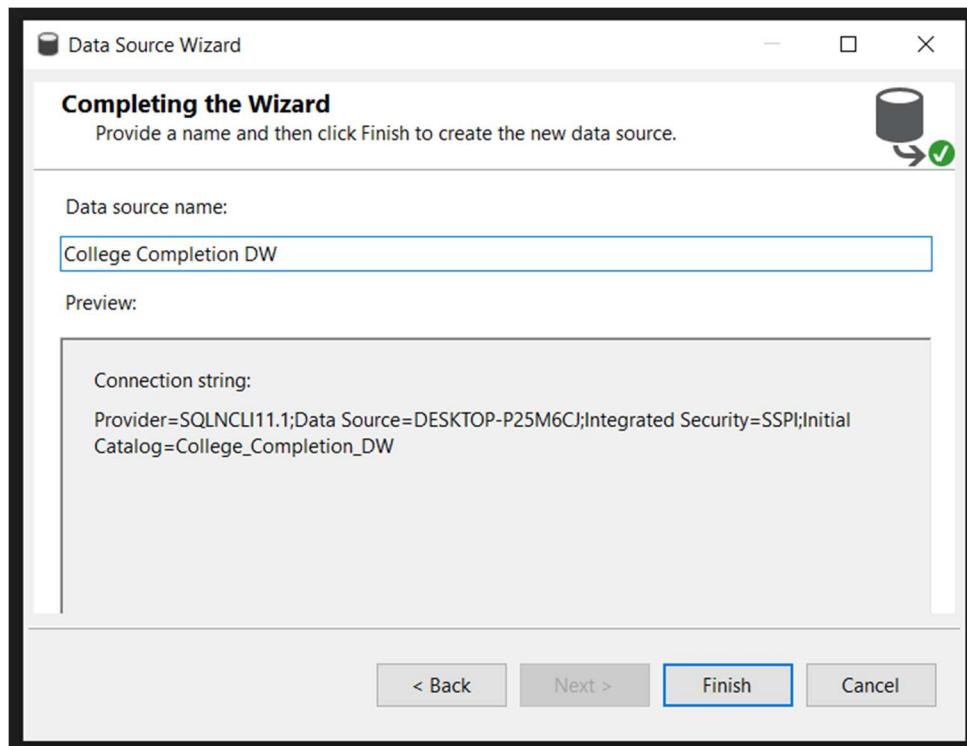
Màn hình sẽ hiển thị như bên dưới, chọn Next



Tiếp theo, điền thông tin User name và password của người dùng máy Window, tiếp đến chọn Next

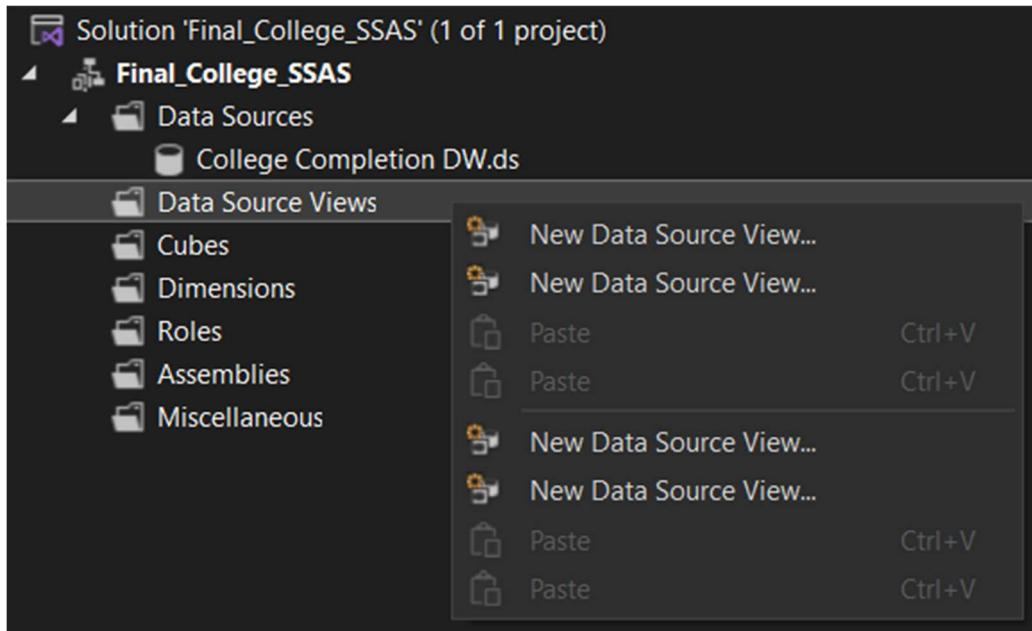


Kế tiếp, đặt tên cho Data Source, cuối cùng nhấn Finish để hoàn thành.

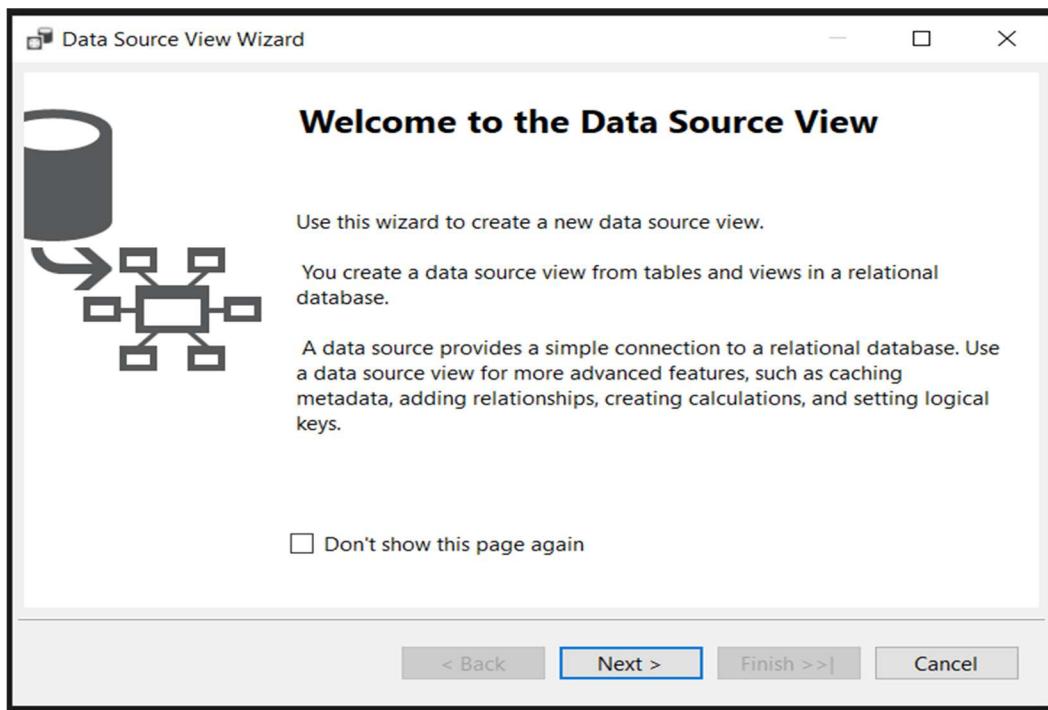


4.3 Tạo Data Source View và Cube

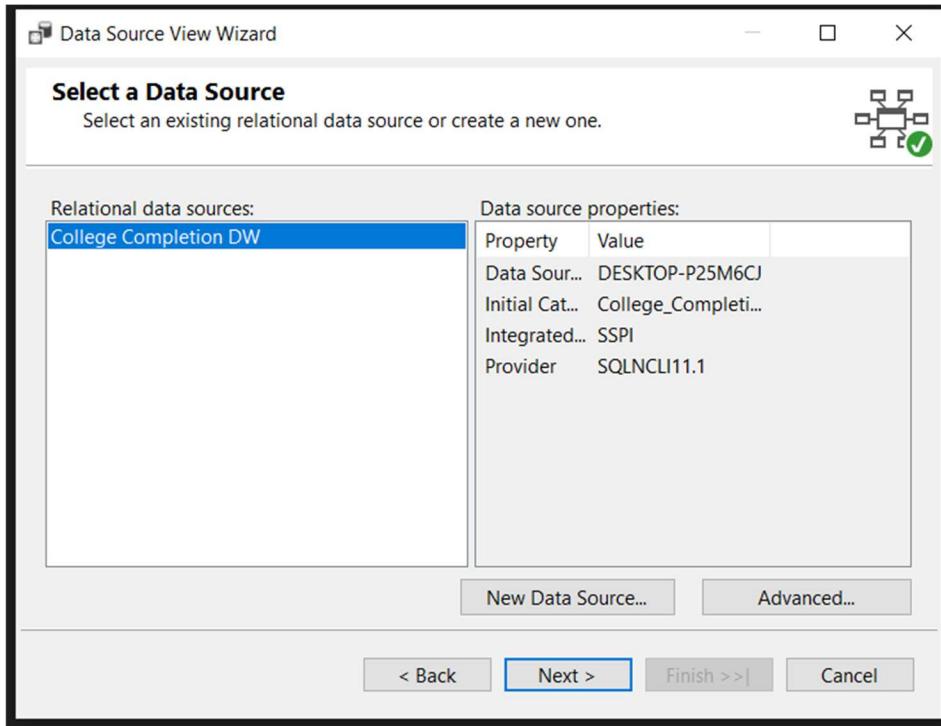
Chọn Data Source View và chọn New Data Source View...



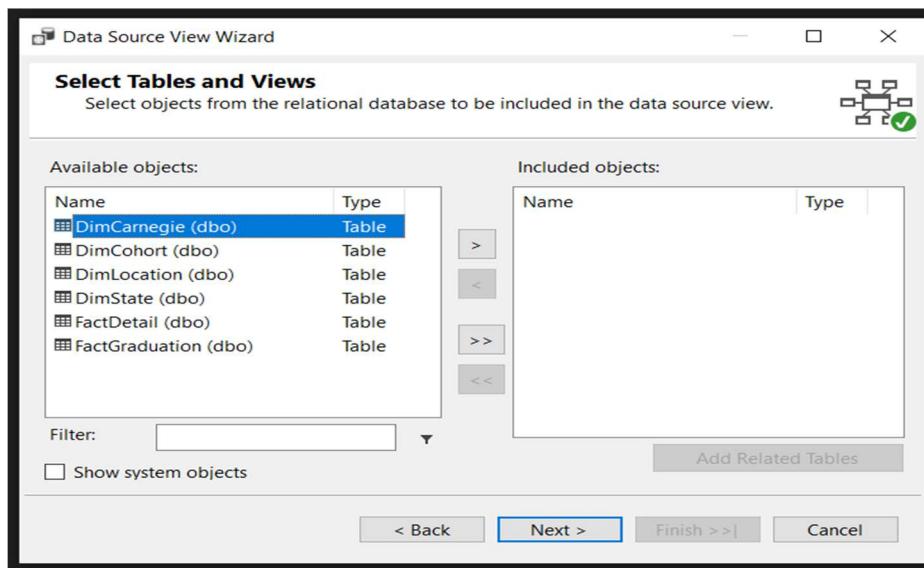
Cửa sổ Data Source View Wizard hiển thị, chọn Next



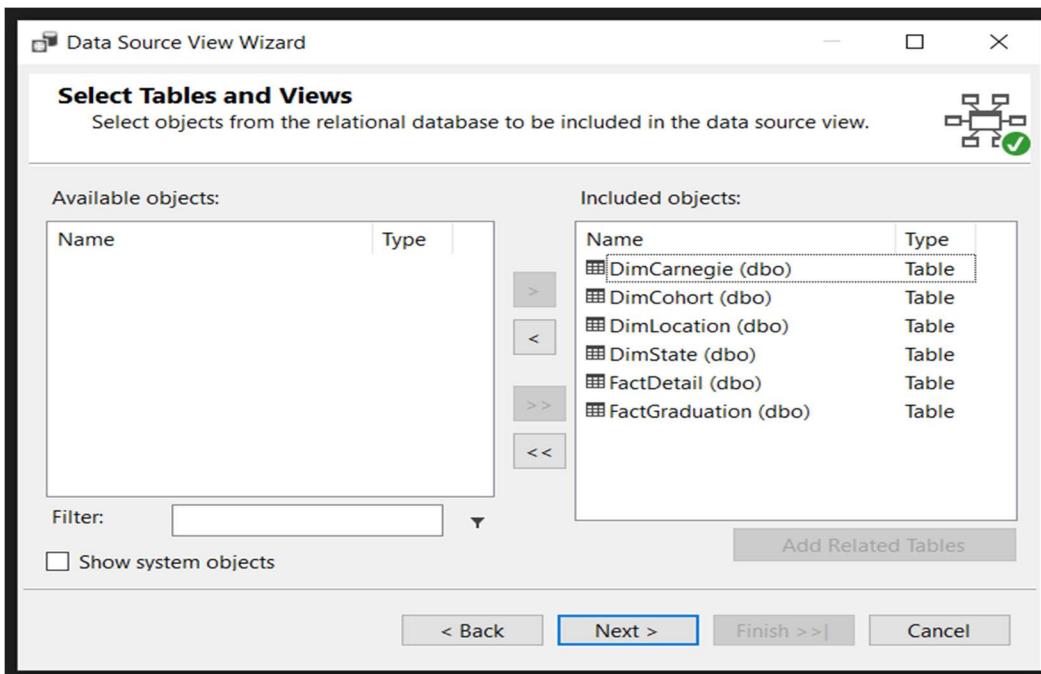
Tiếp theo, chọn Data Source đã tạo và chọn Next



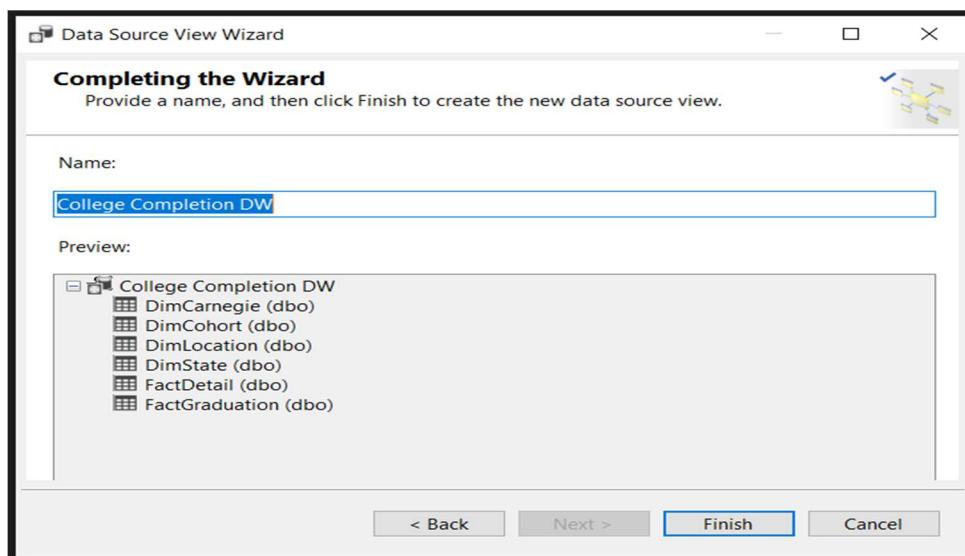
Kế đến, ta sẽ chọn các bảng và views

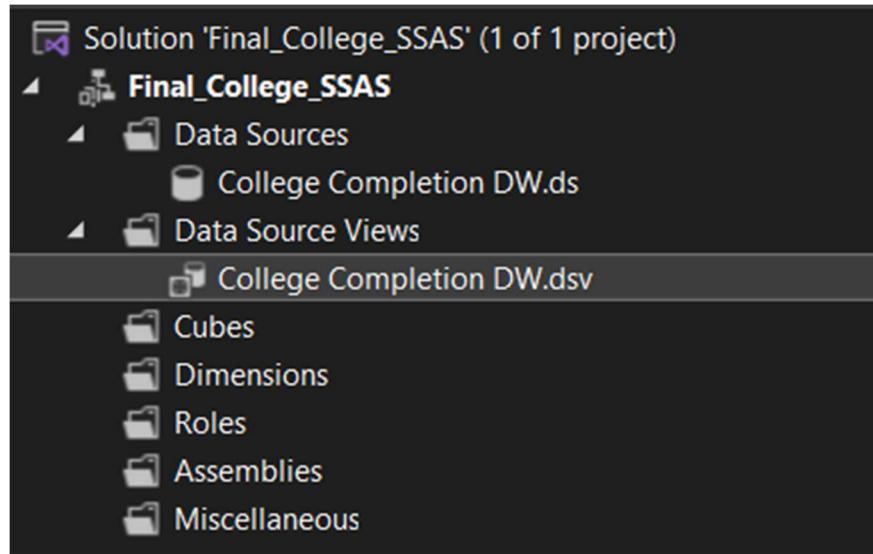


Chọn tất cả các bảng hiện có trong Available objects sang Included objects và chọn Next



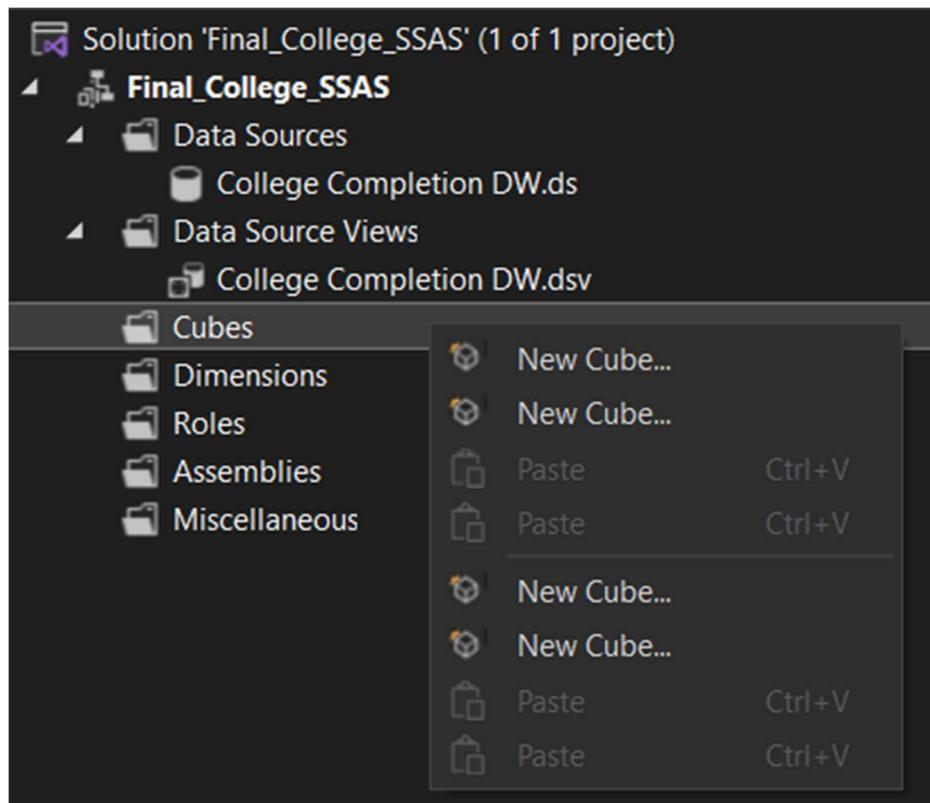
Chỉ định tên của Data Source View và nhấn Finish để hoàn thành.



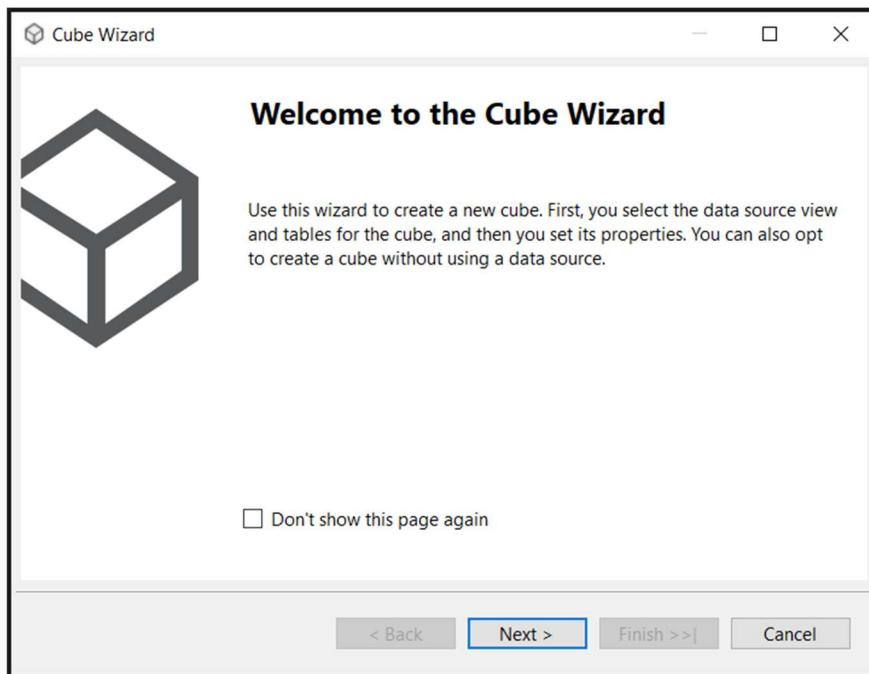


Tạo 2 Cube cho 2 fact: Institution Detail và Institution Graduation cùng các dimension liên quan

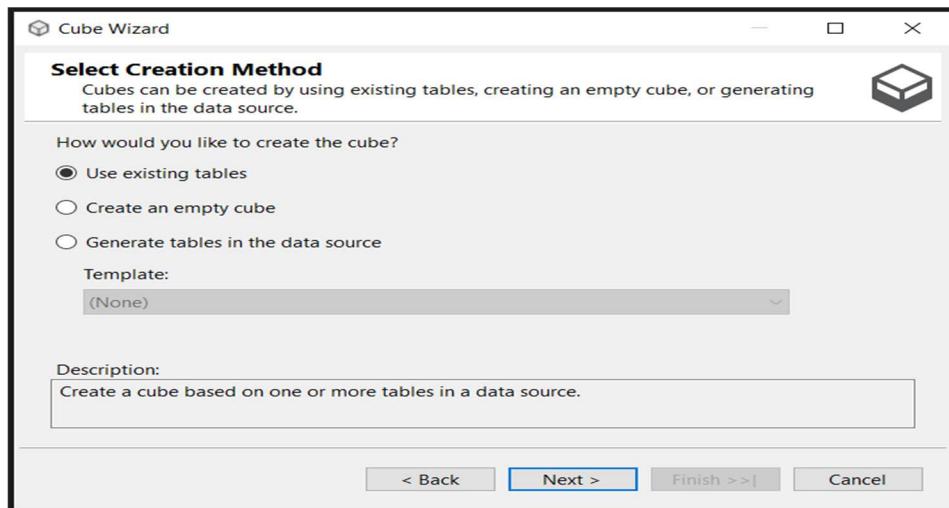
Chọn Cube và chọn New Cube...



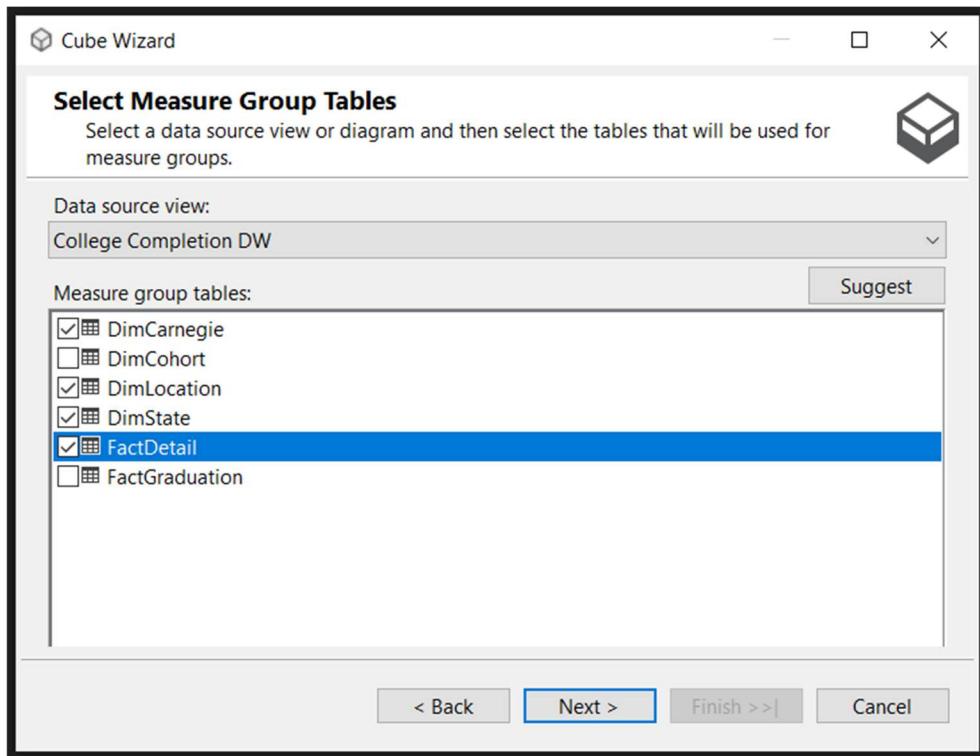
Cửa sổ Cube Wizard xuất hiện, chọn Next



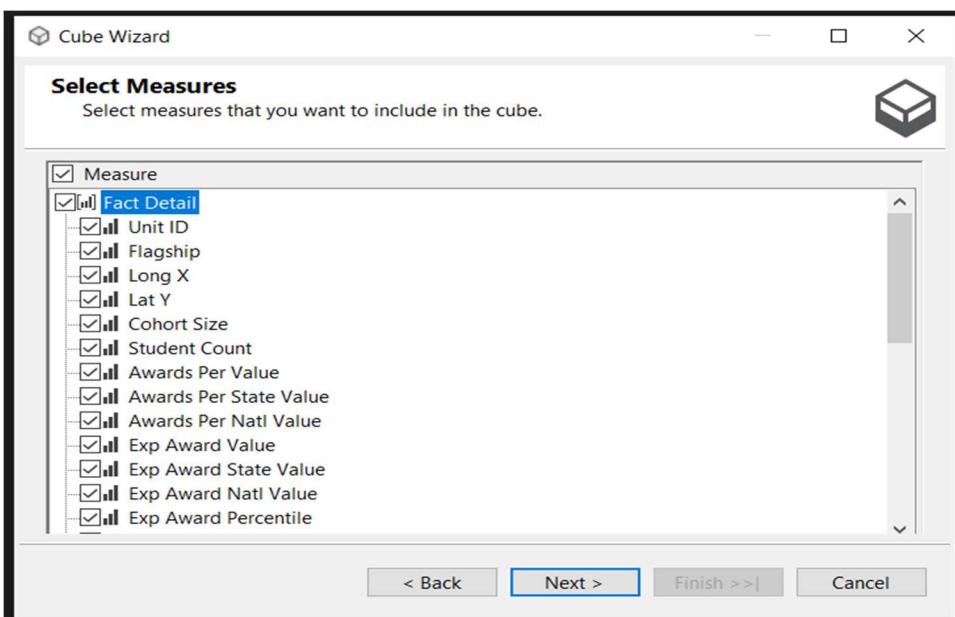
Chọn Use existing tables để chỉ định rằng Cube được tạo từ các bảng đã có, sau đó chọn Next



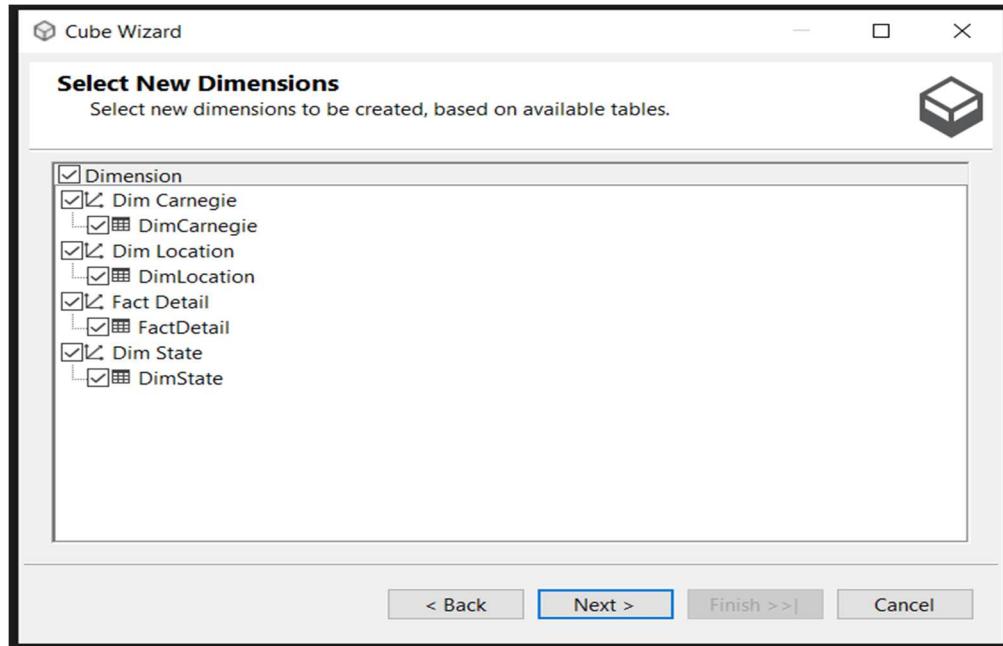
Chọn Measure từ các bảng đưa vào trong Cube, do đang thực hiện cho Fact Institution Detail nên ta chọn các measure ở các bảng: DimCarnegie, DimLocation, DimState và cuối cùng là FactDetail. Tiếp đến chọn Next



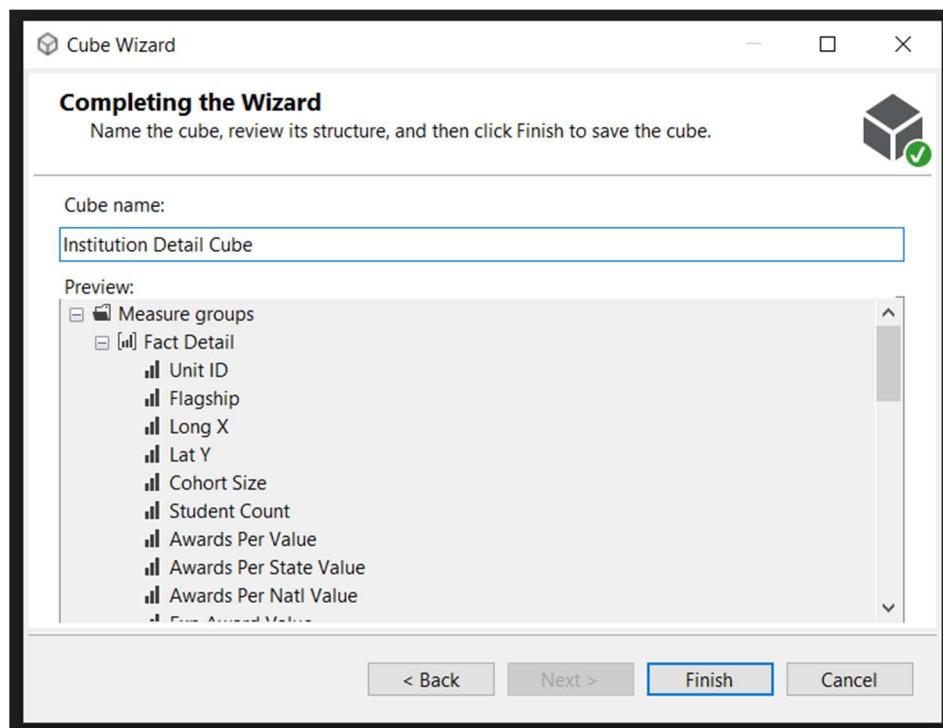
Màn hình hiển thị chọn các measure đưa vào trong Cube như bên dưới, chọn Next

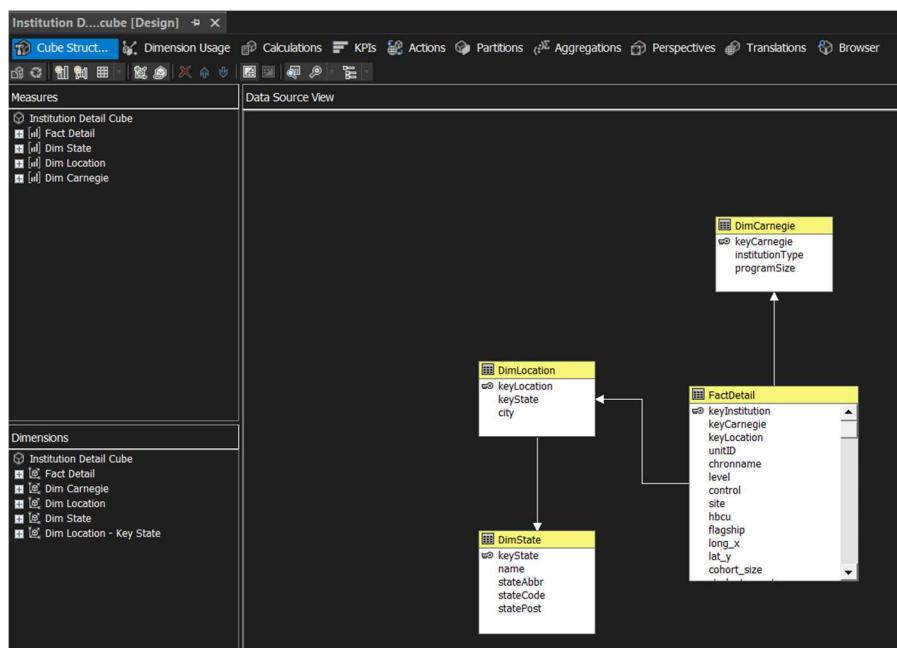
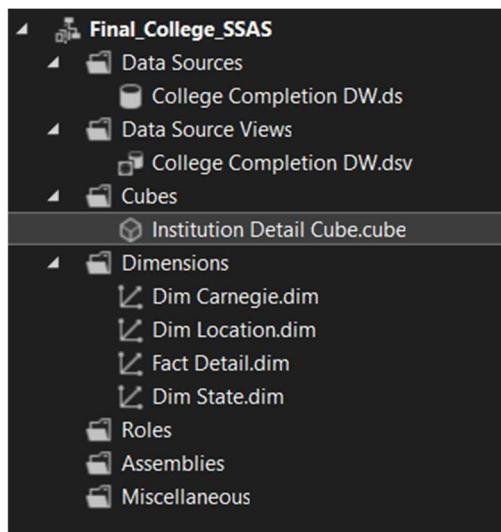


Tiếp đến, chỉ định các dimension như bên dưới và chọn Next

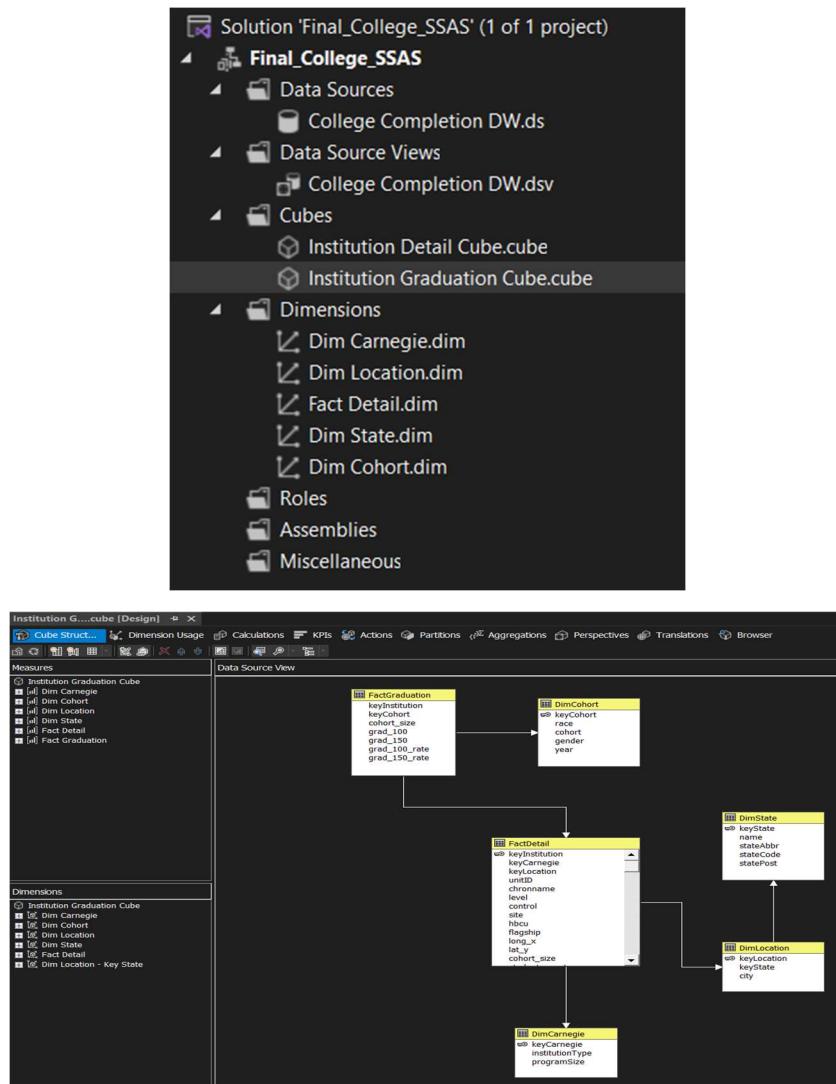


Đặt tên cho Cube và nhấn Finish để hoàn thành





Tương tự, tạo Cube cho Fact Institution Graduation



4.4 Xây dựng Dim và Hierarchy

4.4.1 DimCohort

Nhấn đúp vào DimCohort, kéo thả các thuộc tính của DimCohort trong Data Source

View vào Attributes

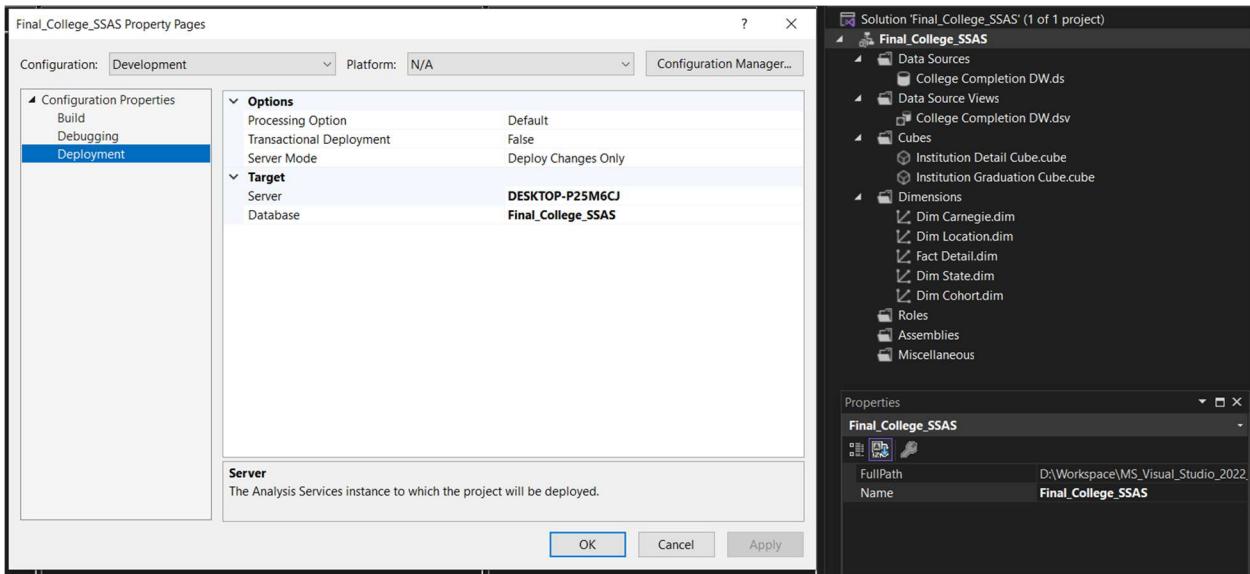
The screenshot shows the 'Dim Cohort.dim [Design]' window. The 'Attributes' pane on the left lists attributes: Dim Cohort (Cohort, Gender, Key Cohort, Race, Year). The 'Hierarchies' pane in the center has a placeholder message: 'To create a new hierarchy, drag an attribute here.' The 'Data Source View' pane on the right shows a list of attributes: DimCohort, keyCohort, race, cohort, gender, year.

Tiếp theo, kéo các thuộc tính ở Attributes vào Hierarchies và đổi tên thành Cohort

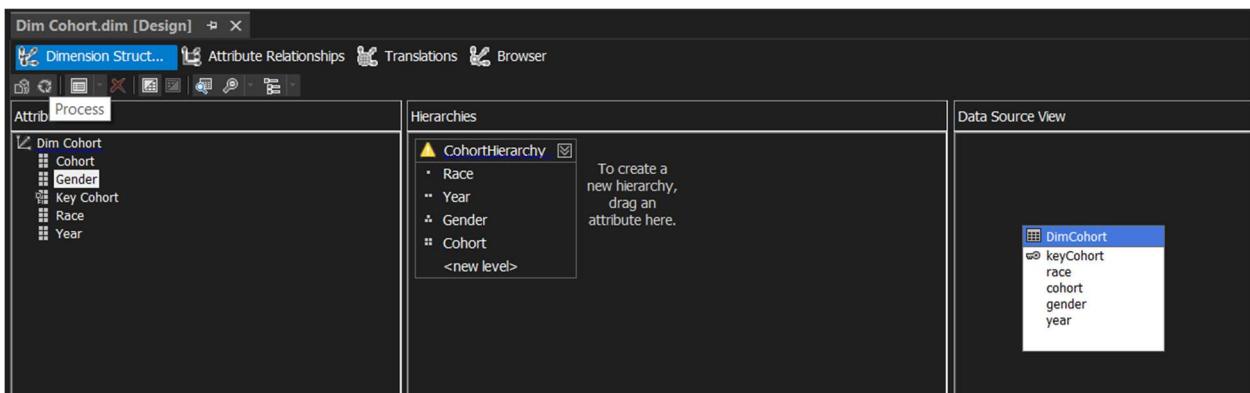
Hierarchy

The screenshot shows the 'Dim Cohort.dim [Design]' window. The 'Attributes' pane on the left lists attributes: Dim Cohort (Cohort, Gender, Key Cohort, Race, Year). The 'Hierarchies' pane in the center displays a new hierarchy named 'CohortHierarchy': Race, Year, Gender, Cohort, <new level>. The 'Data Source View' pane on the right shows a list of attributes: DimCohort, keyCohort, race, cohort, gender, year.

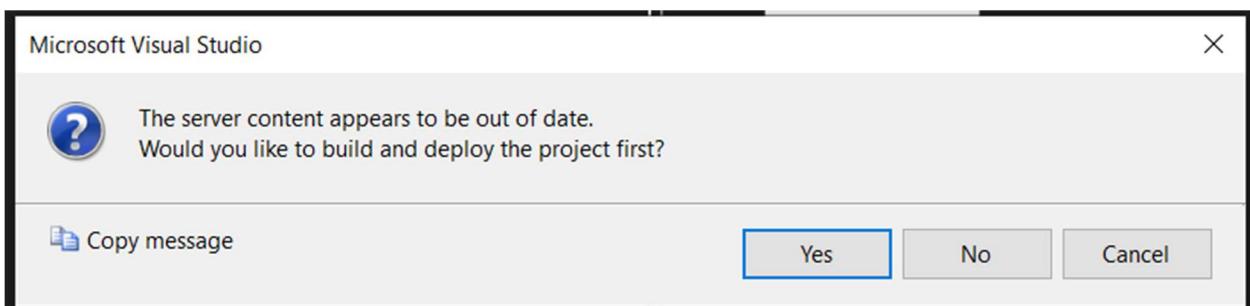
Nhấn đúp vào tên project, chọn properties, đổi tên Server trong phần Deployment từ local thành tên của server trên máy



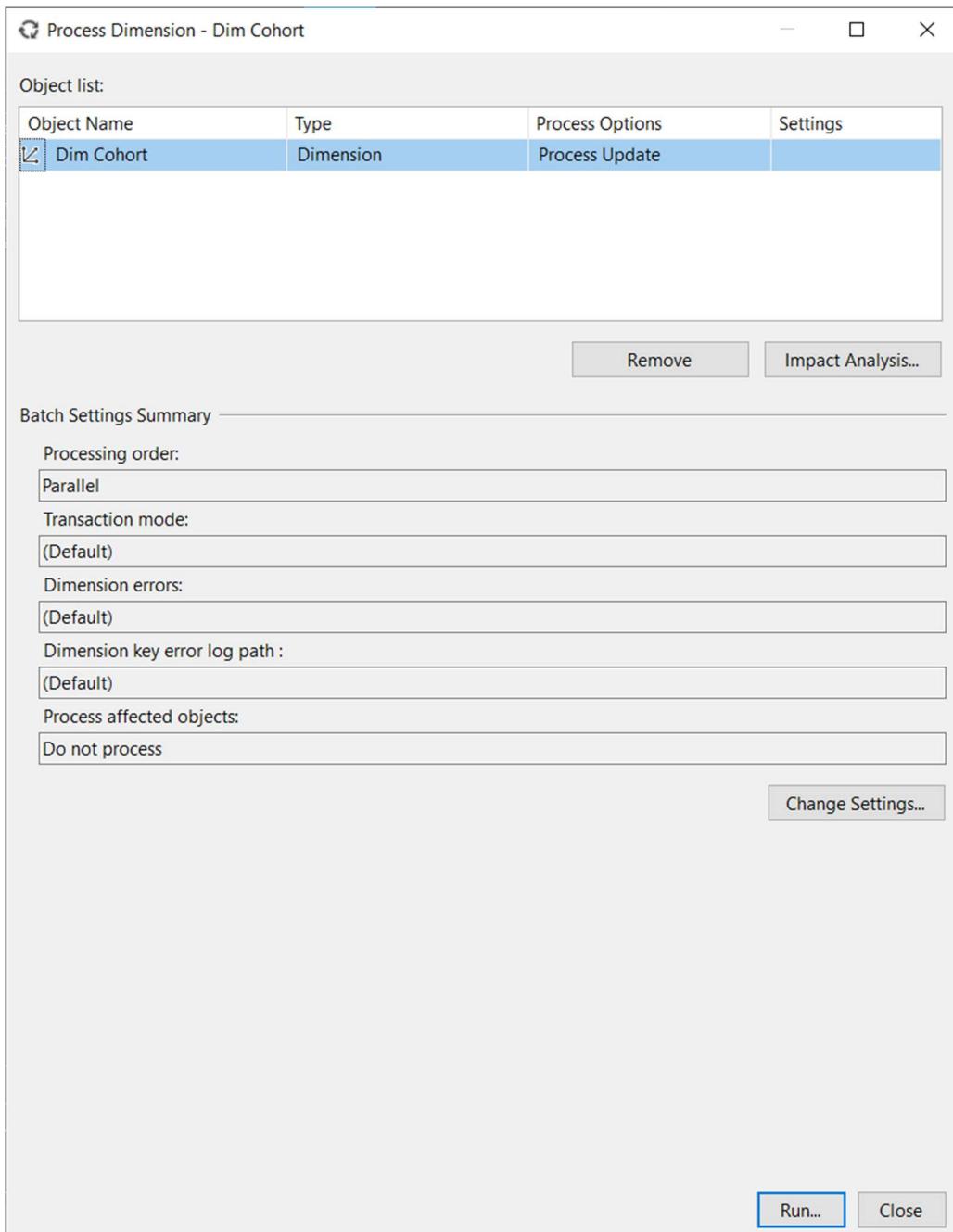
Tiếp đến chọn process DimCohort



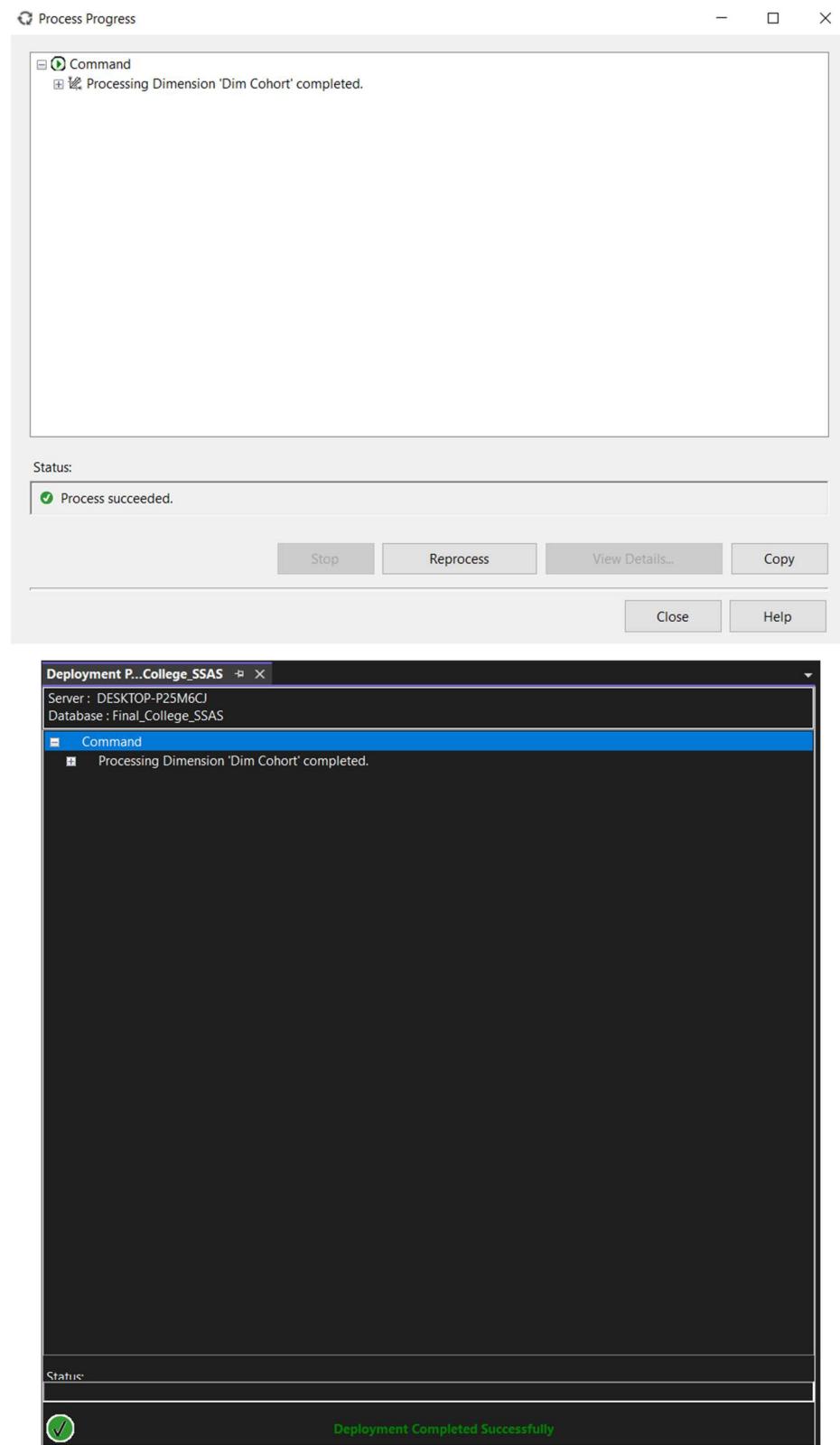
Màn hình sẽ hiển thị như bên dưới, chọn Yes

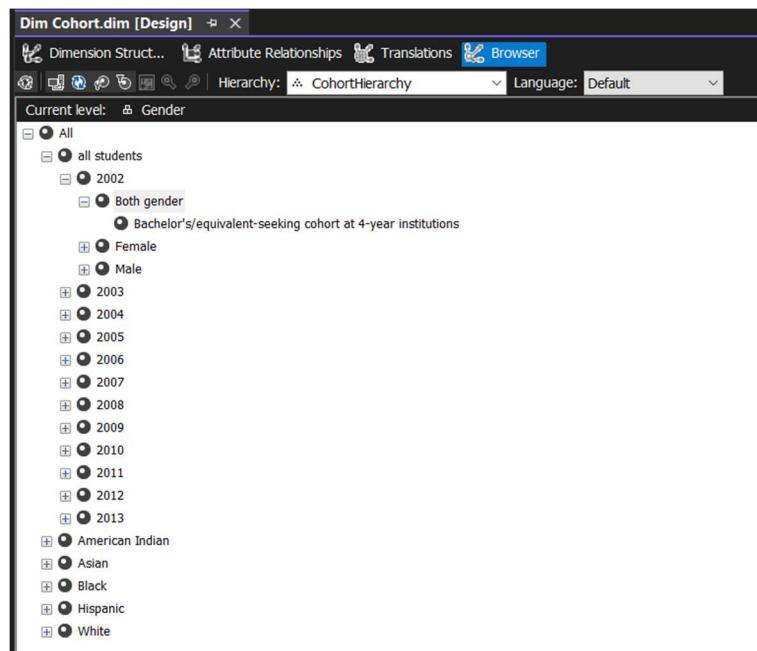


Một cửa sổ như sau sẽ xuất hiện, chọn Run



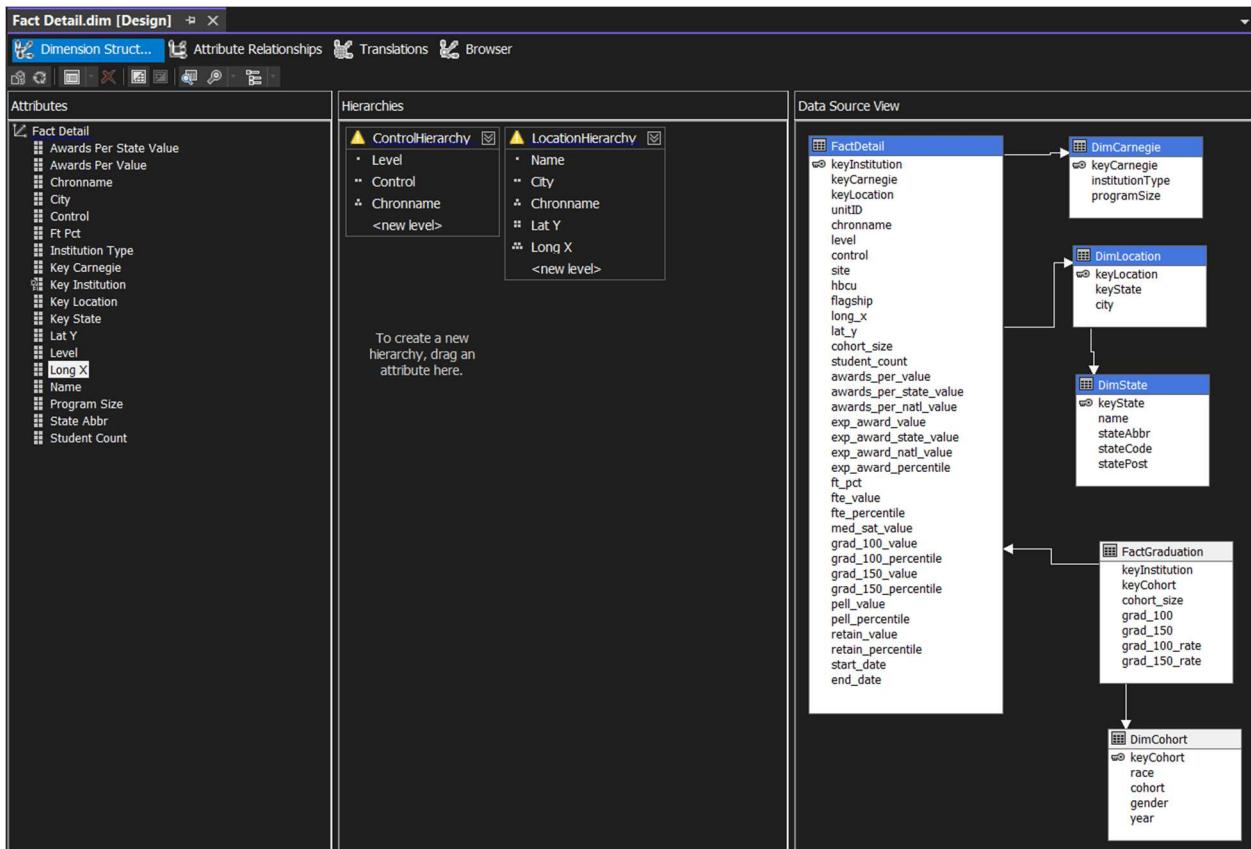
Nếu quá trình chạy thành công, kết quả sẽ hiển thị như sau





4.4.2 Dim (Fact) Detail

Cách tạo tương tự như ở DimCohort, kéo thả các thuộc tính trong Data Source View vào trong Attributes, sau đó đưa từ Attributes vào Hierarchy tạo thành LocationHierarchy và ControlHierarchy, sau đó process cho Dim (Fact) Detail



Màn hình sẽ hiển thị như bên dưới, chọn Run

Process Dimension - Fact Detail

Object list:

Object Name	Type	Process Options	Settings
Fact Detail	Dimension	Process Full	

Remove Impact Analysis...

Batch Settings Summary

Processing order:
Parallel

Transaction mode:
(Default)

Dimension errors:
(Default)

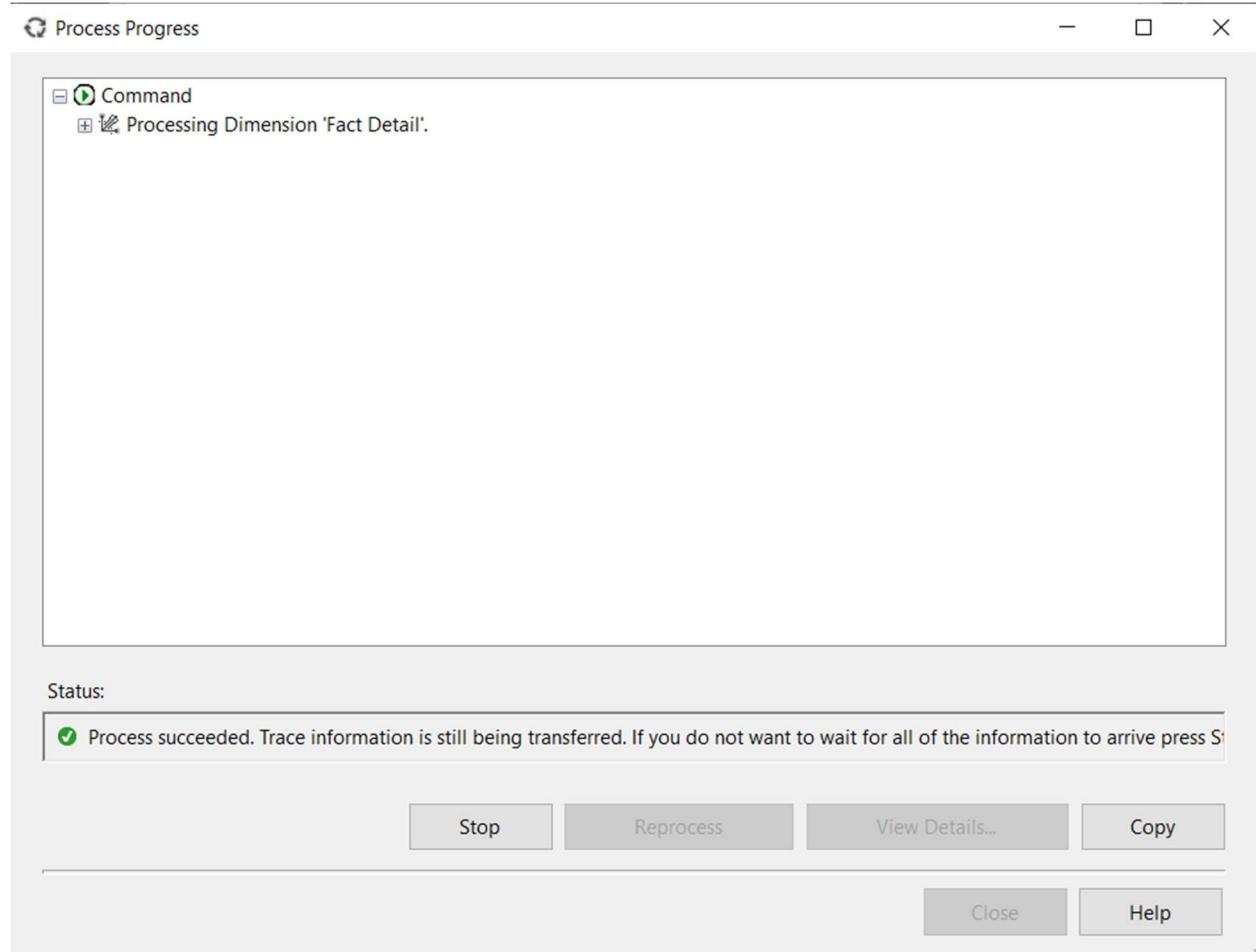
Dimension key error log path :
(Default)

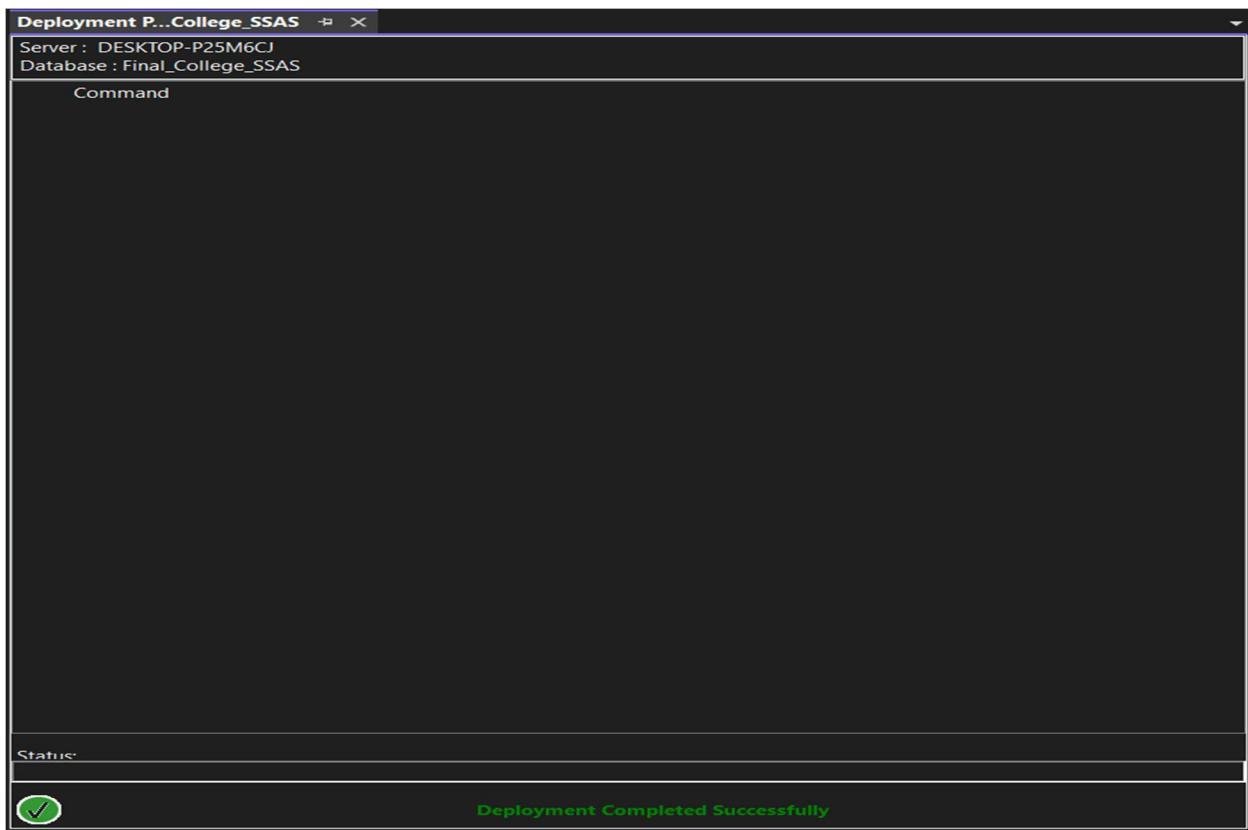
Process affected objects:
Do not process

Change Settings...

Run... Close

Quá trình chạy thành công, ta sẽ nhận được các kết quả sau





Đối với ControlHierarchy

The screenshot shows the "Fact Detail.dim [Design]" window. The toolbar includes icons for Dimension Structure, Attribute Relationships, Translations, and Browser. The "Hierarchy" dropdown is set to "ControlHierarchy". The "Language" dropdown is set to "Default". The "Current level" dropdown is set to "(All)". The main pane displays a hierarchical tree under the "All" node, which branches into "2-year", "Private for-profit", "Private not-for-profit", "Public", and "4-year".

Đối với LocationHierarchy

Fact Detail.dim [Design] ✖

Dimension Struct... Attribute Relationships Translations Browser

Hierarchy: LocationHierarchy Language: Default

Current level: (All)

- [-] All
 - [+] Alabama
 - [+] Alaska
 - [+] Anchorage
 - [+] Alaska Pacific University
 - [+] 61.191235
 - [+] -149.80424
 - [+] Charter College
 - [+] University of Alaska at Anchorage
 - [+] Barrow
 - [+] Fairbanks
 - [+] Juneau
 - [+] Valdez
 - [+] Arizona
 - [+] Arkansas
 - [+] California
 - [+] Colorado
 - [+] Connecticut
 - [+] Delaware
 - [+] District of Columbia
 - [+] Florida
 - [+] Georgia
 - [+] Hawaii
 - [+] Idaho
 - [+] Illinois
 - [+] Indiana
 - [+] Iowa
 - [+] Kansas
 - [+] Kentucky
 - [+] Louisiana
 - [+] Maine
 - [+] Maryland
 - [+] Massachusetts
 - [+] Michigan

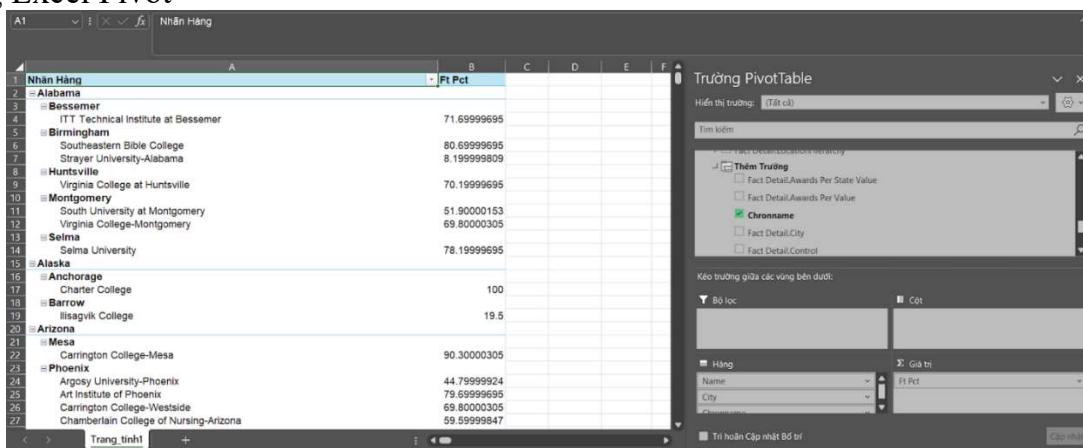
4.5 Trả lời câu hỏi Truy vấn

4.5.1 Câu 1: Tỉ lệ sinh viên theo học chương trình full time tại các khu vực

Dùng SSAS để phân tích

Name	City	Chronname	Ft Pct
Alabama	Bessemer	ITT Technical Institute at Bessemer	71.7
Alabama	Birmingham	Southeastern Bible College	80.7
Alabama	Birmingham	Strayer University-Alabama	8.2
Alabama	Huntsville	Virginia College at Huntsville	70.2
Alabama	Montgomery	South University at Montgomery	51.9
Alabama	Montgomery	Virginia College-Montgomery	69.8
Alabama	Selma	Selma University	78.2
Alaska	Anchorage	Charter College	100
Alaska	Barrow	Ilisagvik College	19.5
Arizona	Mesa	Carrington College-Mesa	90.3
Arizona	Phoenix	Argosy University-Phoenix	44.8
Arizona	Phoenix	Art Institute of Phoenix	79.7
Arizona	Phoenix	Carrington College-Westside	69.8
Arizona	Phoenix	Chamberlain College of Nursing-Arizona	59.6
Arizona	Phoenix	International Institute of the Americas at Gl...	100
Arizona	Phoenix	University of Phoenix Online	100
Arizona	Scottsdale	Le Cordon Bleu College of Culinary Arts - S...	100
Arizona	Sells	Tohono O'odham Community College	30
Arizona	Tempe	ITT Technical Institute at Tempe	78.7
Arizona	Tsaila	Dine College	62.2
Arizona	Tucson	Art Center Design College (Ariz.)	58.3
Arizona	Tucson	Art Institute of Tucson	71.6
Arizona	Tucson	Carrington College-Tucson	99.3
Arizona	Tucson	ITT Technical Institute at Tucson	74.6
Arkansas	Fort Smith	University of Arkansas at Fort Smith	71.4
Arkansas	Little Rock	Arkansas Baptist College	84.4
Arkansas	Little Rock	ITT Technical Institute at Little Rock	81.7
Arkansas	Little Rock	University of Arkansas at Little Rock	55.7
Arkansas	Little Rock	University of Phoenix at Little Rock	100
Arkansas	Springdale	Ecclesia College	80.9
California	Alhambra	Platt College (Alhambra, Calif.)	100
California	Anaheim	Bethesda Christian University	78.8
California	Anaheim	Everest College-Anaheim	88.9

Dùng Excel Pivot

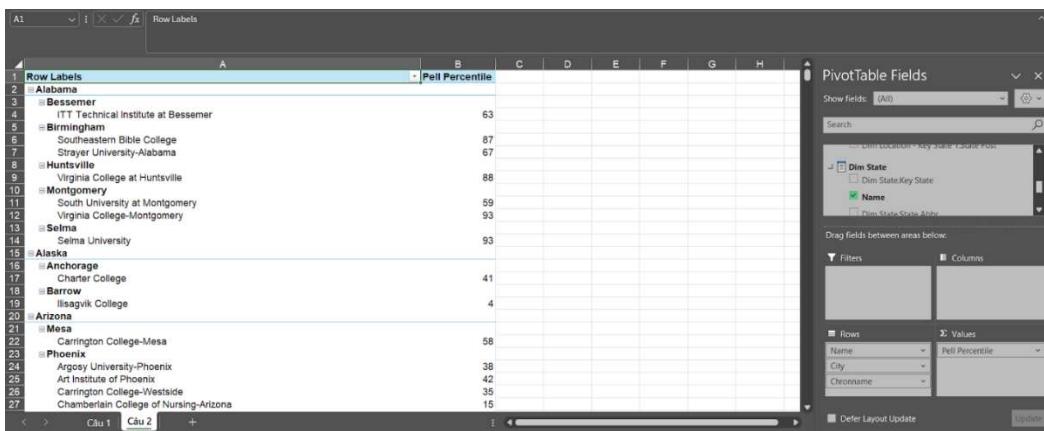


4.5.2 Câu 2: Tỉ lệ sinh viên nhận được trợ cấp Pell tại mỗi khu vực

Dùng SSAS để phân tích

Name	City	Chronname	Pell Percentile
Alabama	Huntsville	Virginia College at Huntsville	88
Alabama	Montgomery	South University at Montgomery	59
Alabama	Montgomery	Virginia College-Montgomery	93
Alabama	Selma	Selma University	93
Alaska	Anchorage	Charter College	41
Alaska	Barrow	Ilisagvik College	4
Arizona	Mesa	Carrington College-Mesa	58
Arizona	Phoenix	Argosy University-Phoenix	38
Arizona	Phoenix	Art Institute of Phoenix	42
Arizona	Phoenix	Carrington College-Westside	35
Arizona	Phoenix	Chamberlain College of Nursing-Arizona	15
Arizona	Phoenix	International Institute of the Americas at Glendale	42
Arizona	Phoenix	University of Phoenix Online	30
Arizona	Scottsdale	Le Cordon Bleu College of Culinary Arts - Scottsdale	29
Arizona	Sells	Tohono O'odham Community College	43
Arizona	Tempe	ITT Technical Institute at Tempe	47
Arizona	Tsaile	Dine College	100
Arizona	Tucson	Art Center Design College (Ariz.)	17
Arizona	Tucson	Art Institute of Tucson	54
Arizona	Tucson	Carrington College-Tucson	39
Arizona	Tucson	ITT Technical Institute at Tucson	65
Arkansas	Fort Smith	University of Arkansas at Fort Smith	84
Arkansas	Little Rock	Arkansas Baptist College	92
Arkansas	Little Rock	ITT Technical Institute at Little Rock	68
Arkansas	Little Rock	University of Arkansas at Little Rock	57
Arkansas	Little Rock	University of Phoenix at Little Rock	74
Arkansas	Springdale	Ecclesia College	76
California	Alhambra	Platt College (Alhambra, Calif.)	28
California	Anaheim	Bethesda Christian University	93
California	Anaheim	Everest College-Anaheim	47
California	Anaheim	West Coast University-Orange County	6
California	Azusa	Azusa Pacific University	30

Dùng Excel Pivot



4.5.3 Câu 2: Tỉ lệ sinh viên nhận được trợ cấp Pell tại mỗi khu vực

Dùng SSAS để phân tích

Name	City	Chronname	Pell Percentile
Alabama	Huntsville	Virginia College at Huntsville	88
Alabama	Montgomery	South University at Montgomery	59
Alabama	Montgomery	Virginia College-Montgomery	93
Alabama	Selma	Selma University	93
Alaska	Anchorage	Charter College	41
Alaska	Barrow	Ilisagvik College	4
Arizona	Mesa	Carrington College-Mesa	58
Arizona	Phoenix	Argosy University-Phoenix	38
Arizona	Phoenix	Art Institute of Phoenix	42
Arizona	Phoenix	Carrington College-Westside	35
Arizona	Phoenix	Chamberlain College of Nursing-Arizona	15
Arizona	Phoenix	International Institute of the Americas at Glendale	42
Arizona	Phoenix	University of Phoenix Online	30
Arizona	Scottsdale	Le Cordon Bleu College of Culinary Arts - Scottsdale	29
Arizona	Sells	Tohono O'odham Community College	43
Arizona	Tempe	ITT Technical Institute at Tempe	47
Arizona	Tsailie	Dine College	100
Arizona	Tucson	Art Center Design College (Ariz.)	17
Arizona	Tucson	Art Institute of Tucson	54
Arizona	Tucson	Carrington College-Tucson	39
Arizona	Tucson	ITT Technical Institute at Tucson	65
Arkansas	Fort Smith	University of Arkansas at Fort Smith	84
Arkansas	Little Rock	Arkansas Baptist College	92
Arkansas	Little Rock	ITT Technical Institute at Little Rock	68
Arkansas	Little Rock	University of Arkansas at Little Rock	57
Arkansas	Little Rock	University of Phoenix at Little Rock	74
Arkansas	Springdale	Ecclesia College	76
California	Alhambra	Platt College (Alhambra, Calif.)	28
California	Anaheim	Bethesda Christian University	93
California	Anaheim	Everest College-Anaheim	47
California	Anaheim	West Coast University-Orange County	6
California	Azusa	Azusa Pacific University	30

Dùng Excel Pivot

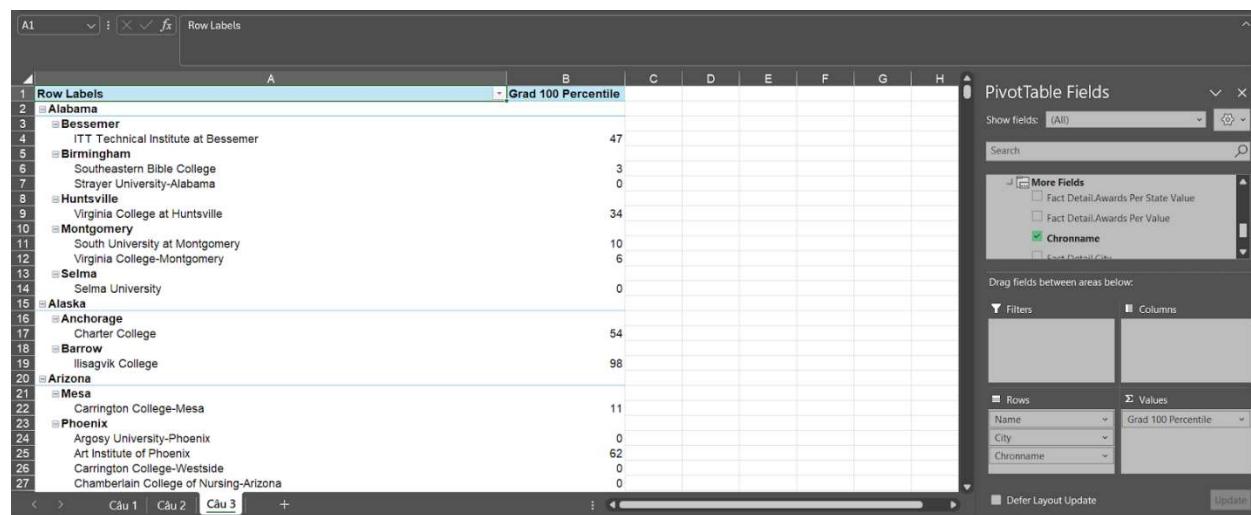
Row Labels	Pell Percentile
Alabama	63
Bessemer	
IT Technical Institute at Bessemer	
Birmingham	87
Southeastern Bible College	
Strayer University-Alabama	67
Huntsville	88
Virginia College at Huntsville	
Montgomery	59
South University at Montgomery	
Virginia College-Montgomery	93
Selma	
Selma University	93
Alaska	
Anchorage	41
Charter College	
Barrow	4
Ilisagvik College	
Arizona	
Mesa	
Carrington College-Mesa	58
Phoenix	
Argosy University-Phoenix	38
Art Institute of Phoenix	42
Carrington College-Westside	35
Chamberlain College of Nursing-Arizona	15

4.5.3 Câu 3:Tỉ lệ tốt nghiệp đúng hạn theo thời gian học tại các khu vực

Dùng SSAS để phân tích

Name	City	Chronname	Grad 100 Percentile
Alabama	Bessemer	ITT Technical Institute at Bessemer	47
Alabama	Birmingham	Southeastern Bible College	3
Alabama	Birmingham	Strayer University-Alabama	0
Alabama	Huntsville	Virginia College at Huntsville	34
Alabama	Montgomery	South University at Montgomery	10
Alabama	Montgomery	Virginia College-Montgomery	6
Alabama	Selma	Selma University	0
Alaska	Anchorage	Charter College	54
Alaska	Barrow	Ilisagvik College	98
Arizona	Mesa	Carrington College-Mesa	11
Arizona	Phoenix	Argosy University-Phoenix	0
Arizona	Phoenix	Art Institute of Phoenix	62
Arizona	Phoenix	Carrington College-Westside	0
Arizona	Phoenix	Chamberlain College of Nursing-Arizona	0
Arizona	Phoenix	International Institute of the Americas a...	82
Arizona	Phoenix	University of Phoenix Online	4
Arizona	Scottsdale	Le Cordon Bleu College of Culinary Arts ...	94
Arizona	Sells	Tohono O'odham Community College	0
Arizona	Tempe	ITT Technical Institute at Tempe	0
Arizona	Tsai	Dine College	0
Arizona	Tucson	Art Center Design College (Ariz.)	23
Arizona	Tucson	Art Institute of Tucson	0
Arizona	Tucson	Carrington College-Tucson	2
Arizona	Tucson	ITT Technical Institute at Tucson	60
Arkansas	Fort Smith	University of Arkansas at Fort Smith	13
Arkansas	Little Rock	Arkansas Baptist College	2
Arkansas	Little Rock	ITT Technical Institute at Little Rock	19
Arkansas	Little Rock	University of Arkansas at Little Rock	6
Arkansas	Little Rock	University of Phoenix at Little Rock	0
Arkansas	Springdale	Ecclesia College	10
California	Alhambra	Platt College (Alhambra, Calif.)	0
California	Anaheim	Bethesda Christian University	82
California	Anaheim	Everest College-Anaheim	74

Dùng Excel Pivot



4.5.4 Câu 4:Tỉ lệ sinh viên được giữ lại sau năm học thứ nhất

Dùng SSAS để phân tích

Name	City	Chronname	Retain Value
Alabama	Bessemer	ITT Technical Institute at Bessemer	0
Alabama	Birmingham	Southeastern Bible College	64.3
Alabama	Birmingham	Strayer University-Alabama	100
Alabama	Huntsville	Virginia College at Huntsville	40.9
Alabama	Montgomery	South University at Montgomery	19.4
Alabama	Montgomery	Virginia College-Montgomery	62.1
Alabama	Selma	Selma University	42
Alaska	Anchorage	Charter College	43.3
Alaska	Barrow	Ilisagvik College	40
Arizona	Mesa	Carrington College-Mesa	71.8
Arizona	Phoenix	Argosy University-Phoenix	42.9
Arizona	Phoenix	Art Institute of Phoenix	50.5
Arizona	Phoenix	Carrington College-Westside	61.5
Arizona	Phoenix	Chamberlain College of Nursing-Arizona	83.3
Arizona	Phoenix	International Institute of the Americas at Glendale	44.7
Arizona	Phoenix	University of Phoenix Online	34.7
Arizona	Scottsdale	Le Cordon Bleu College of Culinary Arts - Scottsdale	0
Arizona	Sells	Tohono O'odham Community College	42.3
Arizona	Tempe	ITT Technical Institute at Tempe	0
Arizona	Tsaike	Dine College	80
Arizona	Tucson	Art Center Design College (Ariz.)	84.6
Arizona	Tucson	Art Institute of Tucson	46.3
Arizona	Tucson	Carrington College-Tucson	72.7
Arizona	Tucson	ITT Technical Institute at Tucson	0
Arkansas	Fort Smith	University of Arkansas at Fort Smith	62.2
Arkansas	Little Rock	Arkansas Baptist College	39.5
Arkansas	Little Rock	ITT Technical Institute at Little Rock	0
Arkansas	Little Rock	University of Arkansas at Little Rock	69.7
Arkansas	Little Rock	University of Phoenix at Little Rock	29.3
Arkansas	Springdale	Ecclesia College	47.2
California	Alhambra	Platt College (Alhambra, Calif.)	88.9
California	Anaheim	Bethesda Christian University	92.9
California	Anaheim	Everest College-Anaheim	69.5

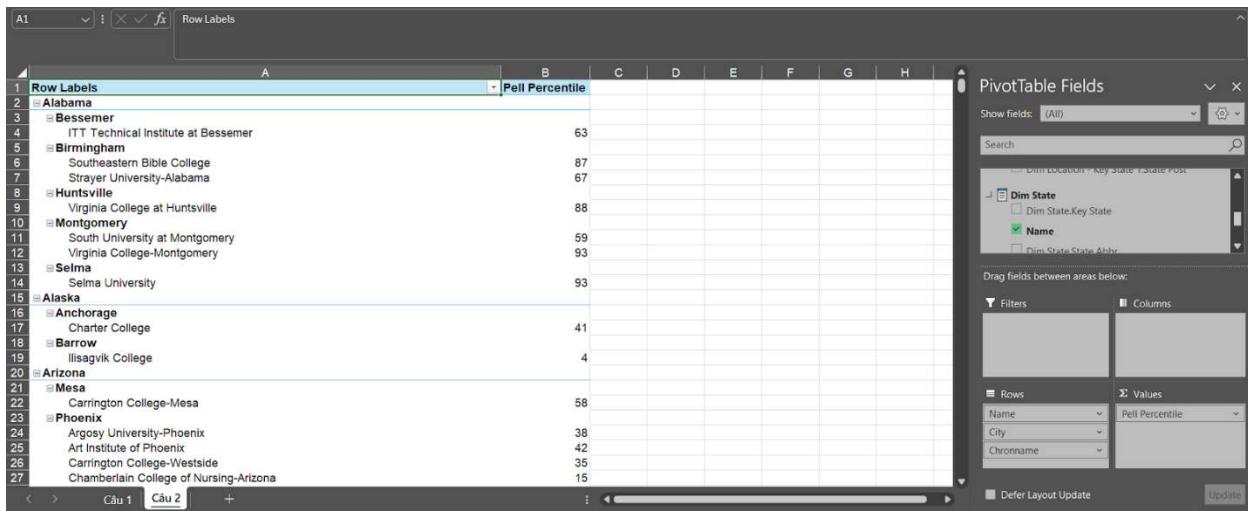
Dùng Excel Pivot

Row Labels	Retain Value
Alabama	
Bessemer	0
Birmingham	64.30000305
Huntsville	100
Montgomery	40.90000153
Selma	42
Alaska	
Anchorage	43.29999924
Barrow	40
Arizona	
Mesa	71.80000305
Phoenix	42.90000153
Tucson	50.5
Tempe	61.5
Scottsdale	83.30000305

4.5.5 Câu 5: Tổng số sinh viên và chi phí ước tính cho mỗi giải thưởng học thuật và số lượng nhân viên trường đó

Dùng SSAS để phân tích

Dùng Excel Pivot



4.5.6 Câu 6: Số lượng sinh viên tốt nghiệp theo sắc tộc và giới tính

Dùng SSAS để phân tích

Race	Gender	Grad 100	Grad 150
all students	Both gender	411470	687766
all students	Female	235533	366316
all students	Male	145199	261046
American Indian	Both gender	1613	2898
American Indian	Female	992	1705
American Indian	Male	621	1193
Asian	Both gender	20262	31558
Asian	Female	12776	19049
Asian	Male	7486	12509
Black	Both gender	37725	82484
Black	Female	27014	54478
Black	Male	10711	28006
Hispanic	Both gender	23071	37921
Hispanic	Female	14758	23118
Hispanic	Male	8313	14803
White	Both gender	261432	418914
White	Female	159147	238465
White	Male	102285	180449

Dùng Excel Pivot

Row Labels	Grad 100	Grad 150
Both gender		
all students	411470	687766
American Indian	1613	2898
Asian	20262	31558
Black	37725	82484
Hispanic	23071	37921
White	261432	418914
Female		
all students	235533	366316
American Indian	992	1705
Asian	12776	19049
Black	27014	54478
Hispanic	14758	23118
White	159147	238465
Male		
all students	145199	261046
American Indian	621	1193
Asian	7486	12509
Black	10711	28006
Hispanic	8313	14803
White	102285	180449
Grand Total	1480408	2462678

4.5.7 Câu 7: Số lượng sinh viên tốt nghiệp đúng hạn của từng khu vực

Dùng SSAS để phân tích

Name	Cohort	Grad 100
Alabama	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
Alabama	Degree-seeking students at 2-year institutions	30731
Alabama	Students seeking another type of degree or certificate at a 4-year institution	7
Alaska	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
Alaska	Degree-seeking students at 2-year institutions	30731
Alaska	Students seeking another type of degree or certificate at a 4-year institution	7
Arizona	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
Arizona	Degree-seeking students at 2-year institutions	30731
Arizona	Students seeking another type of degree or certificate at a 4-year institution	7
Arkansas	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
Arkansas	Degree-seeking students at 2-year institutions	30731
Arkansas	Students seeking another type of degree or certificate at a 4-year institution	7
California	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
California	Degree-seeking students at 2-year institutions	30731
California	Students seeking another type of degree or certificate at a 4-year institution	7
Colorado	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
Colorado	Degree-seeking students at 2-year institutions	30731
Colorado	Students seeking another type of degree or certificate at a 4-year institution	7
Connecticut	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
Connecticut	Degree-seeking students at 2-year institutions	30731
Connecticut	Students seeking another type of degree or certificate at a 4-year institution	7
Delaware	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670

Dùng Excel Pivot

■ Alabama	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
	Degree-seeking students at 2-year institutions	30731
	Students seeking another type of degree or certificate at a 4-year institution	7
■ Alaska	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
	Degree-seeking students at 2-year institutions	30731
	Students seeking another type of degree or certificate at a 4-year institution	7
■ Arizona	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
	Degree-seeking students at 2-year institutions	30731
	Students seeking another type of degree or certificate at a 4-year institution	7
■ Arkansas	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
	Degree-seeking students at 2-year institutions	30731
	Students seeking another type of degree or certificate at a 4-year institution	7
■ California	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
	Degree-seeking students at 2-year institutions	30731
	Students seeking another type of degree or certificate at a 4-year institution	7
■ Colorado	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
	Degree-seeking students at 2-year institutions	30731
	Students seeking another type of degree or certificate at a 4-year institution	7
■ Connecticut	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
	Degree-seeking students at 2-year institutions	30731
	Students seeking another type of degree or certificate at a 4-year institution	7
■ Delaware	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
	Degree-seeking students at 2-year institutions	30731
	Students seeking another type of degree or certificate at a 4-year institution	7
■ District of Columbia	Bachelor's/equivalent-seeking cohort at 4-year institutions	1449670
	Degree-seeking students at 2-year institutions	30731

4.5.8 Câu 8: Số lượng sinh viên lấy bằng sau 150% thời gian học tiêu chuẩn theo giới tính và sắc tộc của từng năm.

Dùng SSAS để phân tích

Year	Gender	Race	Grad 150	Grad 150 Rate
2002	Both gender	Black	5468	3645.8
2002	Both gender	Hispanic	1840	1227.3
2002	Both gender	White	26794	17863.2
2002	Female	all students	22478	14985.8
2002	Female	American Indian	85	57.2
2002	Female	Asian	971	648
2002	Female	Black	3704	2470
2002	Female	Hispanic	1080	720.8
2002	Female	White	15353	10235.2
2002	Male	all students	15799	10532.5
2002	Male	American Indian	76	51.5
2002	Male	Asian	650	434.5
2002	Male	Black	1764	1176.4
2002	Male	Hispanic	760	507.3
2002	Male	White	11441	7627.3
2003	Both gender	all students	41128	27419.1
2003	Both gender	American Indian	185	124
2003	Both gender	Asian	1844	1229.5
2003	Both gender	Black	5808	3873
2003	Both gender	Hispanic	1869	1246.5
2003	Both gender	White	28678	19118.3
2003	Female	all students	23955	15970.2
2003	Female	American Indian	115	77.2
2003	Female	Asian	1112	741.8

Dùng Excel Pivot

Row Labels	Grad 150 Rate	Grad 150
2002		
Both gender		
all students	25518.4	38277
American Indian	108.1	161
Asian	1081.1	1621
Black	3645.8	5468
Hispanic	1227.3	1840
White	17863.2	26794
Female		
all students	14985.8	22478
American Indian	57.2	85
Asian	648	971
Black	2470	3704
Hispanic	720.8	1080
White	10235.2	15353
Male		
all students	10532.5	15799
American Indian	51.5	76
Asian	434.5	650
Black	1176.4	1764
Hispanic	507.3	760
White	7627.3	11441
2003		
Both gender		
all students	27419.1	41128
American Indian	124	185
Asian	1229.5	1844
Black	3873	5808
Hispanic	1246.5	1869

4.5.9 Câu 9: Số lượng sinh viên tốt nghiệp theo từng năm, phân theo loại trường học và cohort

Dùng SSAS để phân tích

Institution Type	Control	Level	Year	Grad 100
Baccalaureate/Associates Colleges	Private for-profit	4-year	2002	636
Baccalaureate/Associates Colleges	Private for-profit	4-year	2003	180
Baccalaureate/Associates Colleges	Private for-profit	4-year	2004	350
Baccalaureate/Associates Colleges	Private for-profit	4-year	2005	544
Baccalaureate/Associates Colleges	Private for-profit	4-year	2006	400
Baccalaureate/Associates Colleges	Private for-profit	4-year	2007	564
Baccalaureate/Associates Colleges	Private for-profit	4-year	2008	372
Baccalaureate/Associates Colleges	Private for-profit	4-year	2009	2402
Baccalaureate/Associates Colleges	Private for-profit	4-year	2010	1280
Baccalaureate/Associates Colleges	Private for-profit	4-year	2011	1424
Baccalaureate/Associates Colleges	Private for-profit	4-year	2012	1186
Baccalaureate/Associates Colleges	Private for-profit	4-year	2013	1384
Baccalaureate/Associates Colleges	Private not-for-profit	4-year	2002	350
Baccalaureate/Associates Colleges	Private not-for-profit	4-year	2003	520
Baccalaureate/Associates Colleges	Private not-for-profit	4-year	2004	490
Baccalaureate/Associates Colleges	Private not-for-profit	4-year	2005	682
Baccalaureate/Associates Colleges	Private not-for-profit	4-year	2006	934
Baccalaureate/Associates Colleges	Private not-for-profit	4-year	2007	1650
Baccalaureate/Associates Colleges	Private not-for-profit	4-year	2008	1418
Baccalaureate/Associates Colleges	Private not-for-profit	4-year	2009	1300
Baccalaureate/Associates Colleges	Private not-for-profit	4-year	2010	3938
Baccalaureate/Associates Colleges	Private not-for-profit	4-year	2011	3086
Baccalaureate/Associates Colleges	Private not-for-profit	4-year	2012	2830
Baccalaureate/Associates Colleges	Private not-for-profit	4-year	2013	3364

Dùng Excel Pivot

Row Labels	Grad 100
■ Baccalaureate/Associates Colleges	
■ 4-year	
■ Private for-profit	
■ Bachelor's/equivalent-seeking cohort at 4-year institutions	
2002	636
2003	180
2004	350
2005	544
2006	400
2007	564
2008	372
2009	2402
2010	1280
2011	1424
2012	1186
2013	1384
■ Private not-for-profit	
■ Bachelor's/equivalent-seeking cohort at 4-year institutions	
2002	350
2003	520
2004	490
2005	682
2006	934
2007	1650
2008	1418
2009	1300
2010	3938
2011	3086
2012	2830

4.5.10 Câu 10: Số lượng sinh viên tốt nghiệp theo sắc tộc và giới tính, phân theo quy mô chương trình học

Dùng SSAS để phân tích

Year	Gender	Race	Grad 150
2002	Both gender	all students	38277
	Both gender	American Indian	161
	Both gender	Asian	1621
	Both gender	Black	5468
	Both gender	Hispanic	1840
	Both gender	White	26794
	Female	all students	22478
	Female	American Indian	85
	Female	Asian	971
	Female	Black	3704
	Female	Hispanic	1080
	Female	White	15353
	Male	all students	15799
	Male	American Indian	76
	Male	Asian	650
	Male	Black	1764
	Male	Hispanic	760
	Male	White	11441
2003	Both gender	all students	41128
	Both gender	American Indian	185
	Both gender	Asian	1844
	Both gender	Black	5808
	Both gender	Hispanic	1869
2003	Both gender	White	28678

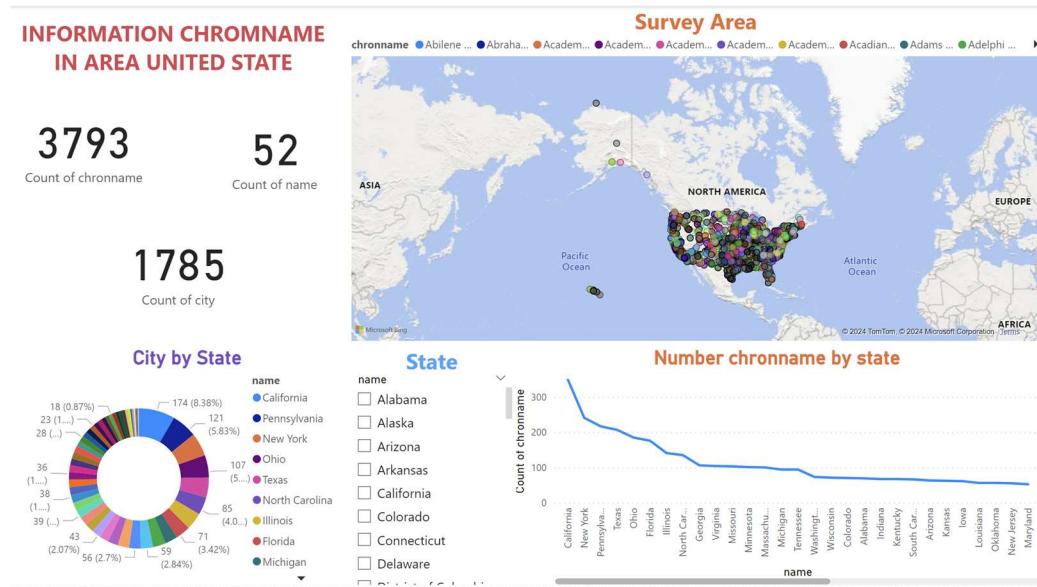
Dùng Excel Pivot

Grad 150	Column Labels								
Row Labels		all students	American Indian	Asian	Black	Hispanic	White	Grand Total	
2002	Both gender	38277		161	1621	5468	1840	26794	74161
	Female	22478		85	971	3704	1080	15353	43671
	Male	15799		76	650	1764	760	11441	30490
2003	Both gender	41128		185	1844	5808	1869	28678	79512
	Female	23955		115	1112	3892	1139	16279	46492
	Male	17173		70	732	1916	730	12399	33020
2004	Both gender	44171		170	2127	6370	2215	30302	85355
	Female	25947		95	1284	4299	1353	17348	50326
	Male	18224		75	843	2071	862	12954	35029
2005	Both gender	46257		216	2155	6299	2451	31679	89057
	Female	27171		125	1355	4206	1469	18125	52451
	Male	19086		91	800	2093	982	13554	36606
2006	Both gender	49606		201	2414	6510	2770	33680	95181
	Female	29053		114	1481	4275	1686	19370	55979
	Male	20553		87	933	2235	1084	14310	39202
2007	Both gender	52216		224	2548	6896	3141	34734	99759
	Female	30389		131	1577	4561	1855	19672	58185
	Male	21827		93	971	2335	1286	15062	41574
2008	Both gender	53087		266	2717	6922	3092	35991	102075
	Female	31115		156	1674	4581	1918	20464	59908
	Male	21972		110	1043	2341	1174	15527	42167
2009	Both gender	57914		292	2743	7465	3637	38452	110503
	Female	33526		160	1621	4950	2227	22027	64511
	Male	23756		132	1122	2515	1410	16425	45360
2010		22426		205	2010	7045	2210	22005	102074

CHƯƠNG 5: PHÂN TÍCH BẰNG POWER BI

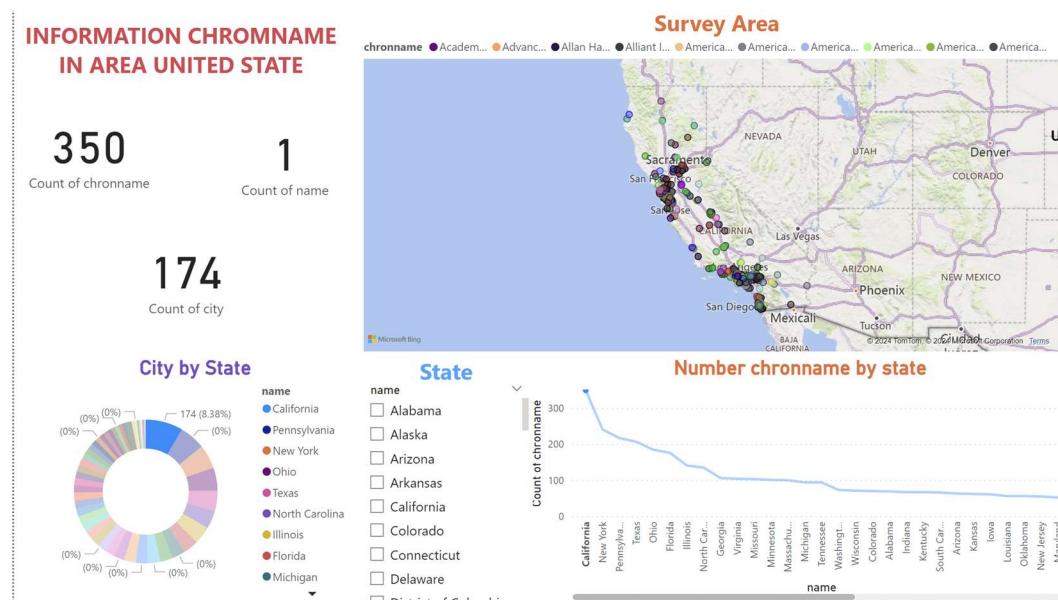
5.1 Report về vị trí địa lý của các cơ sở Đại học

Phân tích chi tiết về vị trí địa lý của các cơ sở Đại học bằng MapChart. Cho biết Bang đó có bao nhiêu thành phố, có bao nhiêu Trường Đại học. Hiện các vị trí trường đó trên bản đồ



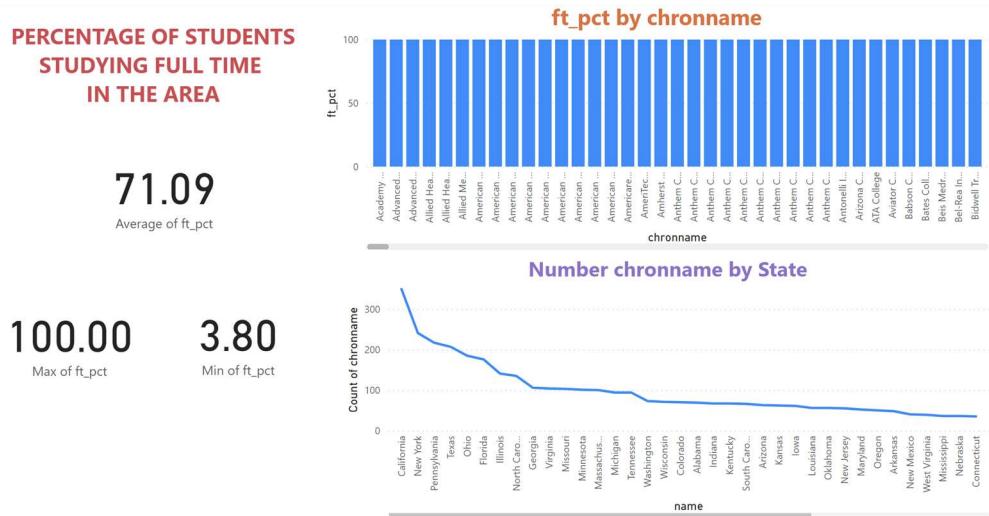
Khi ta nhấn vào một Bang bất kỳ, thì sẽ hiện ra các trường và các thành phố của Bang đó.

Ví dụ nhấn vào bang California

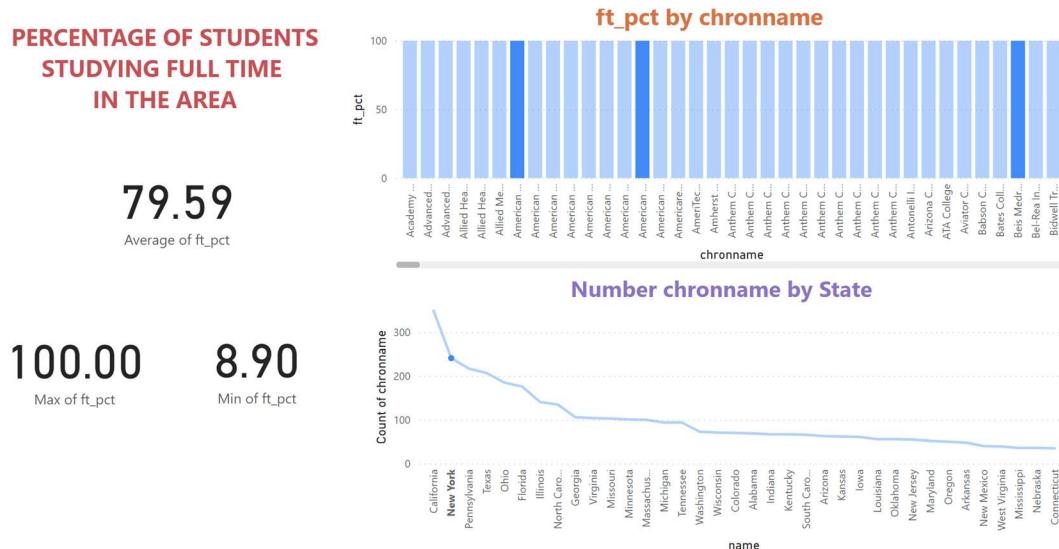


5.2 Report về tỷ lệ sinh viên theo học chương trình Full-time ở Trường Đại học

Phân tích chi tiết về tỷ lệ sinh viên theo học chương trình Full-time. Ta thấy trung bình ở nước Mỹ, tỷ lệ này khá cao chiếm 71.09 %....

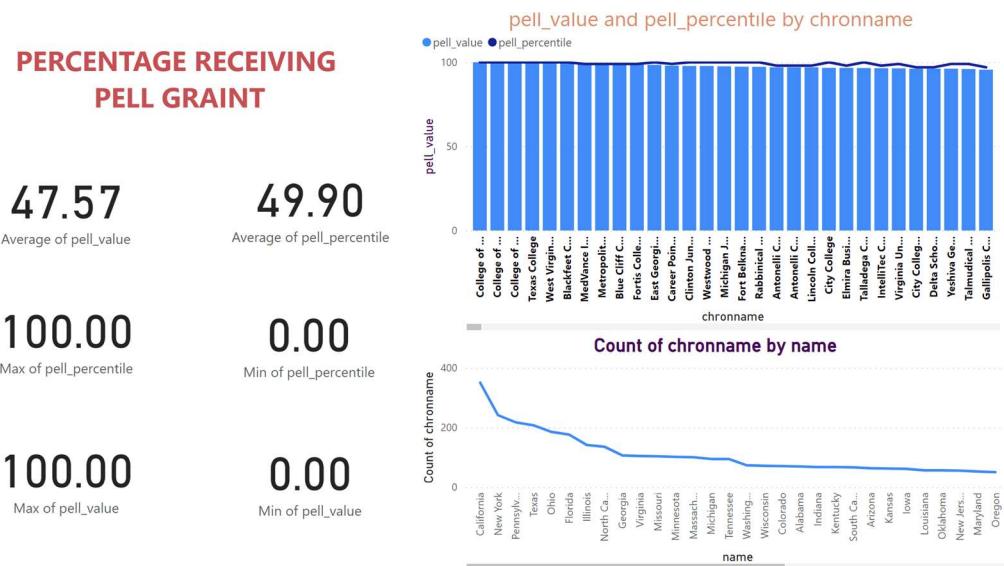


Khi ta nhấp vào một Bang bất kỳ, thì sẽ hiện ra các trường và các thành phố của Bang đó. Ví dụ nhấp vào bang New York, ta thấy tỷ lệ sinh viên theo học chương trình full tile ở đây là 79.59%, thấp nhất là 8.9%



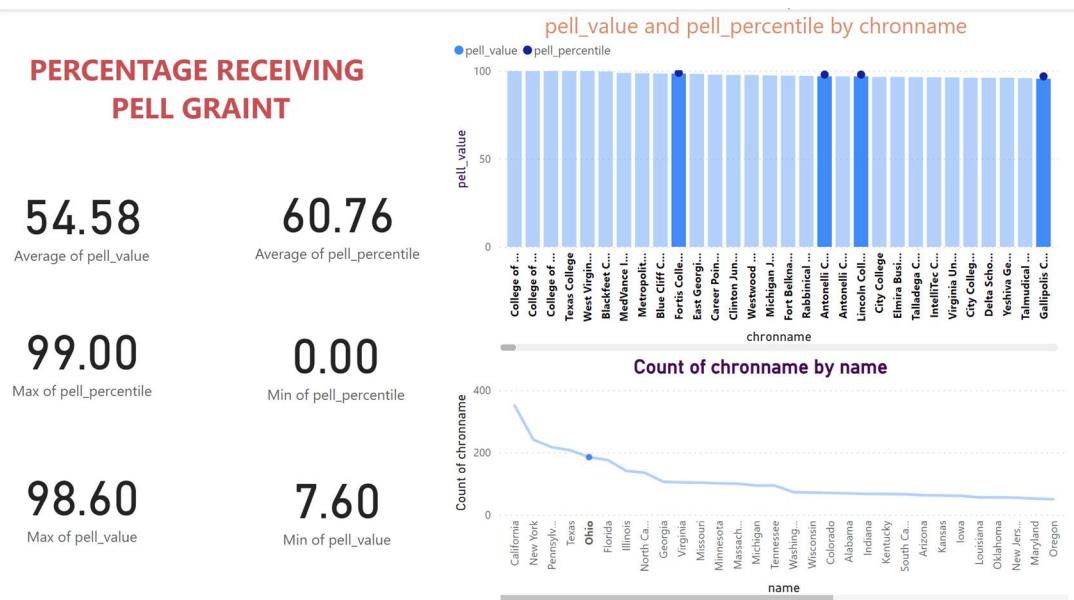
5.3 Report về tỷ lệ sinh viên nhận trợ cấp Pell ở Trường Đại học

Phân tích chi tiết về tỷ lệ sinh viên nhận trợ cấp Pell của Mỹ. Ta thấy trung bình ở nước Mỹ, tỷ lệ trung bình của toàn nước Mỹ là 47.57%



Khi ta nhán vào một Bang bất kỳ, thì sẽ hiện ra các trường và các thành phố của Bang đó.

Ví dụ nhán vào bang Ohio, thì giá trị trung bình của tỷ lệ sinh viên nhận trợ cấp Pell là 54.58%



PHẦN KẾT LUẬN

1. Ưu điểm

Nắm rõ các khái niệm cơ bản về kho dữ liệu và OLAP, các tính chất của một kho dữ liệu cần có. Từ đó vận dụng để xây dựng một kho dữ liệu hoàn chỉnh dùng để khai thác dữ liệu. Xây dựng mô hình kho dữ liệu và trang bị kiến thức về các công cụ SSIS, SSAS, Report.

2. Nhược điểm

Do thời gian tìm hiểu giới hạn, nhóm chưa tìm hiểu sâu và khai thác tối ưu. Những câu hỏi truy vấn còn đơn giản và chưa được đa dạng

3. Hướng phát triển

Từ những nghiên cứu ban đầu về Kho dữ liệu, nhóm sẽ tích hợp thêm các công cụ, xác định những nghiệp vụ phức tạp hơn để nghiên c

TÀI LIỆU THAM KHẢO

ThS. Nguyễn Văn Thành “Bài giảng bộ môn Kho Dữ liệu”

Jonathan Ortiz, “*Data College Completion*”, Data Worls, 2017.

[College Completion - dataset by databaseats | data.world](#)