

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
ĐẠI HỌC CÔNG NGHỆ THÔNG TIN



BÁO CÁO

Kho dữ liệu và OLAP

**Đề tài: Phân tích kho dữ liệu
Các Cuộc Vận Động Biểu Tình
1990 - 2020**

Lớp: IS217.M22.HTCL

Giảng viên: Đỗ Thị Minh Phụng

Sinh Viên : 19521713 - Trần Nhật Khuê
19522305 - Lê Ngọc Minh Thư

CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ KHO DỮ LIỆU	6
1. PHÁT BIỂU VỀ DỮ LIỆU	6
1.1 Lý do chọn đề tài.....	6
1.2 Mô tả dữ liệu	6
1.3 Nguồn kho dữ liệu.....	7
2. XÂY DỰNG KHO DỮ LIỆU	19
2.1 Sơ đồ hình sao minh họa.....	19
2.2 DIM_PARTICIPANT	19
2.3 DIM_LOCATION	20
2.4 DIM_RESPONSE.....	20
2.5 DIM_IDENTIFICATION	23
2.6 DIM_DATE	23
2.7 DIM_DEMAND	24
2.8 DIM_VIOLENCE.....	26
2.9 FACT_MASS_PROTEST	26
3. NỘI DUNG 15 CÂU TRUY VẤN	28
CHƯƠNG 2: TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS).....	30
1. CHUẨN BỊ CÔNG CỤ VÀ DATA WAREHOUSE.....	30
2. QUÁ TRÌNH SSIS	32
2.1 Quá trình nhập dữ liệu vào SQL Server.....	32
2.2 Quá trình SSIS trong Visual Studio 2019	35
2.2.1 Tạo SSIS Project	35
2.2.2 Làm sạch dữ liệu và tạo các bảng DIM.....	37
2.2.3 Tạo bảng Fact.....	107
2.2.4 Tạo các ràng buộc khóa ngoại và thực thi toàn bộ quá trình SSIS.....	114
CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU TRONG KHO (SSAS)	117
1. QUÁ TRÌNH SSAS TRONG VISUAL STUDIO 2019	117
1.1 Tạo project tại Visual Studio 2019 (Define Data Source)	117
1.2 Xác định dữ liệu nguồn	117
1.3 Xác định khung dữ liệu nguồn (Define Data Source View)	120
2. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ SSAS TRÊN CÁC KHỐI CUBE.....	133
2.1 Cho biết tổng số cuộc biểu tình ở Russia theo từng năm.....	133
2.2 Cho biết thông tin cuộc biểu tình, nước xảy ra biểu tình có tổng số lượng người ước tính tham gia biểu tình cao nhất trong quý 3/1990.....	135
2.3 Cho biết tên quốc gia có tổng số lượng người tham gia > 500000 tại Châu Á	138

2.4 Theo từng tháng, quý, năm liệt kê số lần diễn ra biểu tình tại các nước ở Châu Á	140
2.5 Cho biết thông tin nhóm người biểu tình và tổng số cuộc biểu tình của cuộc biểu tình đó lớn hơn 15 được đặt theo từng quốc gia.....	141
2.6 Với từng châu lục liệt kê số ngày biểu tình theo từng năm, trừ Châu Âu.....	144
2.7 Đưa ra top 5 quốc gia có số cuộc biểu tình diễn ra nhiều nhất trong năm 2017.....	145
2.8 Thống kê số cuộc biểu tình, số người tham gia và số ngày biểu tình của mục đích biểu tình thứ nhất, sắp xếp số ngày biểu tình theo thứ tự giảm dần.	147
2.9 Cho biết quốc gia có số ngày biểu tình cao nhất và quốc gia có số lần biểu tình thấp nhất.	149
2.10 Liệt kê những quốc gia có số cuộc biểu tình từ 3 đến 10.....	151
2.11 Xuất ra top 2 những thành phố thuộc Châu Phi có số lần biểu tình thấp nhất trong năm 2018.	152
2.12 Với mỗi châu lục, đưa ra 3 quốc gia có tổng số ngày biểu tình cao nhất.....	153
2.13 Liệt kê những cuộc biểu tình ở các quốc gia tại Bắc Mĩ, xảy ra từ năm 1998 đến 2002 với số người tham gia < 1000	155
2.14 Liệt kê những quốc gia có các cuộc biểu tình diễn ra với hình thức ôn hòa vào năm 2019, cũng như số ngày diễn ra cuộc biểu tình đó.....	157
2.15 Cho biết số ngày diễn ra những cuộc biểu tình có sử dụng bạo lực vũ trang cũng như phản ứng đầu tiên của chính phủ đối với cuộc biểu tình đó tại Mexico, sắp xếp theo số ngày biểu tình tăng dần.	158
3. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ PIVOT EXCEL	160
3.1 Cho biết tổng số cuộc biểu tình ở Russia theo từng năm.....	160
3.2 Cho biết thông tin cuộc biểu tình, nước xảy ra biểu tình có tổng số lượng người ước tính tham gia biểu tình cao nhất trong quý 3/1990.....	162
3.3 Cho biết tên quốc gia có tổng số lượng người tham gia >500000 tại Châu Á.....	163
3.4 Theo từng tháng, quý, năm liệt kê số lần diễn ra biểu tình tại các nước ở Châu Á.	166
3.5 Cho biết thông tin nhóm người biểu tình và tổng số cuộc biểu tình của cuộc biểu tình đó lớn hơn 15 được đặt theo từng quốc gia.....	168
3.6 Với từng châu lục liệt kê số ngày biểu tình theo từng năm, trừ Châu Âu.....	169
3.7 Đưa ra top 5 quốc gia có số cuộc biểu tình diễn ra nhiều nhất trong năm 2017.....	171
3.8 Thống kê số cuộc biểu tình, số người tham gia và số ngày biểu tình của mục đích biểu tình thứ nhất, sắp xếp số ngày biểu tình theo thứ tự giảm dần.	172
3.9 Cho biết quốc gia có số ngày biểu tình cao nhất và quốc gia có số lần biểu tình thấp nhất.	173
3.10 Liệt kê những quốc gia có số cuộc biểu tình từ 3 đến 10.....	174
3.11 Xuất ra top 2 những thành phố thuộc Châu Phi có số lần biểu tình thấp nhất trong năm 2018.	175
3.12 Với mỗi châu lục, đưa ra 3 quốc gia có tổng số ngày biểu tình cao nhất.....	176
3.13 Liệt kê những cuộc biểu tình ở các quốc gia tại Bắc Mĩ, xảy ra từ năm 1998 đến 2002 với số người tham gia < 1000	177
3.14 Liệt kê những quốc gia có các cuộc biểu tình diễn ra với hình thức ôn hòa vào năm 2019, cũng như số ngày diễn ra cuộc biểu tình đó.....	179

3.15 Cho biết số ngày diễn ra những cuộc biểu tình có sử dụng bạo lực vũ trang cũng như phản ứng đầu tiên của chính phủ đối với cuộc biểu tình đó tại Mexico, sắp xếp theo số ngày biểu tình tăng dần.	181
4. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG NGÔN NGỮ MDX.....	182
4.1 Cho biết tổng số cuộc biểu tình ở Russia theo từng năm.....	182
4.2 Cho biết thông tin cuộc biểu tình, nước xảy ra biểu tình có tổng số lượng người ước tính tham gia biểu tình cao nhất trong quý 3/1990.....	182
4.3 Cho biết tên những quốc gia mà có tổng số lượng người tham gia >500000 tại Châu Á.....	183
4.4 Theo từng tháng, quý, năm liệt kê số lần diễn ra biểu tình tại các nước ở Châu Á	184
4.5 Cho biết thông tin nhóm người biểu tình và tổng số cuộc biểu tình của cuộc biểu tình đó lớn hơn 15 được đặt theo từng quốc gia.....	185
4.6 Với từng châu lục liệt kê số ngày biểu tình theo từng năm, trừ Châu Âu.....	185
4.7 Đưa ra top 5 quốc gia có số cuộc biểu tình diễn ra nhiều nhất trong năm 2017	186
4.8 Thống kê số cuộc biểu tình, số người tham gia và số ngày biểu tình của mục đích biểu tình thứ nhất, sắp xếp số ngày biểu tình theo thứ tự giảm dần.	187
4.9 Cho biết quốc gia có số ngày biểu tình cao nhất và quốc gia có số lần biểu tình thấp nhất.	187
4.10 Liệt kê những quốc gia có số cuộc biểu tình từ 3 đến 10.....	188
4.11 Xuất ra top 2 những thành phố thuộc Châu Phi có số lần biểu tình thấp nhất trong năm 2018.189	189
4.12 Với mỗi châu lục, đưa ra 3 quốc gia có tổng số ngày biểu tình cao nhất.....	189
4.13 Liệt kê những cuộc biểu tình ở các quốc gia tại Bắc Mĩ, xảy ra từ năm 1998 đến 2002 với số người tham gia < 1000	190
4.14 Liệt kê những quốc gia có các cuộc biểu tình diễn ra với hình thức ôn hòa vào năm 2019, cũng như số ngày diễn ra cuộc biểu tình đó.....	191
4.15 Cho biết số ngày diễn ra những cuộc biểu tình có sử dụng bạo lực vũ trang cũng như phản ứng đầu tiên của chính phủ đối với cuộc biểu tình đó tại Mexico, sắp xếp theo số ngày biểu tình tăng dần.	192
CHƯƠNG 4: QUY TRÌNH LẬP BÁO CÁO (REPORT)	193
1. QUÁ TRÌNH TẠO BÁO CÁO BẰNG CÔNG CỤ VISUAL STUDIO 2019 (SSRS)	193
1.2 Khởi tạo project SSRS	193
1.3 Thực hiện tạo report trên Visual Studio	198
1.3.1 Báo cáo thống kê số cuộc biểu tình, số ngày biểu tình, số người tham gia biểu tình tại Nga từ 1992 đến 1999.....	198
1.3.2 Báo cáo thống kê số người tham gia biểu tình tại Nam Mĩ và Bắc Mĩ trong năm 2019.....	205
1.3.3 Báo cáo Thống kê số người tham gia biểu tình ở MENA trong năm 2019 - 2020 (theo quý năm)	209
2. QUÁ TRÌNH TẠO BÁO CÁO BẰNG CÔNG CỤ POWER BI.....	212
2.1 Nhập dữ liệu nguồn (Data Source).....	212
2.2 Các bước thực hiện	213

2.2.1	Thống kê số cuộc biểu tình, số ngày biểu tình trong 5 năm (2016 – 2020) tại các châu lục	213
2.2.2	Thống kê số người tham gia các cuộc biểu tình ôn hòa tại MENA	216
2.2.3	Thống kê và so sánh số cuộc biểu tình tại Châu Âu, Châu Phi và Châu Á giữa các năm ...	222
2.2.4	Thống kê biểu tình tại Châu Á	229
CHƯƠNG 5:	QUY TRÌNH KHAI THÁC DỮ LIỆU	244
1.	TỔNG QUAN DỮ LIỆU	244
1.1	Giới thiệu dữ liệu	244
1.2	Mô tả các thuộc tính.....	245
2.	QUÁ TRÌNH THỰC HIỆN SSIS, SSAS.....	247
3.	QUÁ TRÌNH THỰC HIỆN KHAI THÁC DỮ LIỆU BẰNG THUẬT TOÁN CÂY QUYẾT ĐỊNH	253
3.1	Tạo và deploy cấu trúc thuật toán cây quyết định.....	253
3.2	Phân tích và đưa ra tập luật	261
4.	QUÁ TRÌNH THỰC HIỆN KHAI THÁC DỮ LIỆU BẰNG THUẬT TOÁN CLUSTERING VÀ NAÏVE BAYES	282
4.1	Tạo và deploy cấu trúc thuật toán Clustering và Naïve Bayes.....	282
4.2	Phân tích và đưa ra tập luật	286
5.	SO SÁNH GIỮA CÁC THUẬT TOÁN BẰNG ĐỘ THỊ LIFT	291

CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ KHO DỮ LIỆU

1. PHÁT BIỂU VỀ DỮ LIỆU

1.1 Lý do chọn đề tài

- Sau khi tìm hiểu về kho dữ liệu: về khái niệm cũng như những thông tin xoay quanh kho dữ liệu thì nhóm nhận thấy rằng chủ đề về kho dữ liệu vô cùng đa dạng, phong phú. Đường như tất cả mọi thứ xoay quanh xã hội, con người, tự nhiên đều có thể làm một chủ đề trong kho dữ liệu. Cũng chính vì vậy, mà việc chọn lựa đề tài cho môn học cũng là một việc khá khó vì phải chọn ra được một đề tài ưng ý và phù hợp nhất với nhóm.
- Từ những suy nghĩ trên nhóm đã đưa ra tiêu chí để chọn đề tài là đề tài phải lạ, có nhiều cột dữ liệu phong phú cũng như là nhiều dòng để đáp ứng được nội dung đồ án. Và sau bao ngày họp nhóm thì cuối cùng nhóm đã quyết định chọn đề tài về biểu tình. Theo nhóm thấy thì đây là một đề tài khá mới cũng như đáp ứng khá đầy đủ yêu cầu đối với đề tài của nhóm.

1.2 Mô tả dữ liệu

- Dữ liệu về những hoạt động biểu tình được lấy từ dự án ‘*Mass Mobilization*’ (dự án Biểu Tình Quần Chúng). Dự án này cung cấp dữ liệu dựa trên những cuộc biểu tình, vận động quần chúng đối với chính phủ nhà nước
- Cuộc biểu tình là sự biểu hiện, thể hiện các suy nghĩ, hành động, bất đồng quan điểm đối với các quy định, liên quan đến quyền lợi, của các tổ chức hay cá nhân, hoặc tập hợp các nhóm người ủng hộ cho mục đích chính trị, hoặc nguyên nhân khác. Nó thường bao gồm việc đi bộ

để diễu hành hàng loạt và bắt đầu với một cuộc gặp gỡ tại một địa điểm được chỉ định sẵn hoặc không chỉ định sẵn.

- Dự án ‘Mass Mobilization’ là một hoạt động nỗ lực để có thể thâu hiểu hơn về những hành động của dân chúng với chính quyền của họ. Từ đó, ta biết được những nguyện vọng, lợi ích mà họ muốn có được khi thực hiện diễu hành/ biểu tình phản đối chính phủ và cũng như cách mà chính phủ họ giải quyết những vấn đề này.

1.3 Nguồn kho dữ liệu

- Kho dữ liệu tên: Mass Mobilization Protest (1990-2020), gồm hơn 17 000 dòng và 31 thuộc tính.
- Kho dữ liệu này được tạo từ dự án ‘Mass Mobilization’ của hai tác giả David H. Clark (đến từ Đại học Binghamton) và Patrick M. Regan (Đại học Notre Dame). Kho dữ liệu này thu nhập trong phạm vi khắp toàn cầu, bao gồm 162 đất nước, giai đoạn từ 1990-2020 và được cập nhập gần nhất vào ngày 09/01/2021.
- Các nguồn chính được các tác giả sử dụng để thu nhập dữ liệu:
 - New York Times.
 - Washington Post.
 - Christian Science Monitor.
 - Times of London.
 - Jerusalem Post.
 - Link nguồn kho dữ liệu:

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/D>

[VN/HTTWYL](#)



1.4 Thuộc tính của kho dữ liệu

STT	Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa
1	id	Numberial	Mã khóa chính trong kho dữ liệu
2	country	Text	Đất nước – nơi diễn ra cuộc biểu tình
3	ccode	Numberial	Mã đất nước (country code) – nơi diễn ra cuộc biểu tình <ul style="list-style-type: none"> ▪ South America – country codes 100-165 ▪ Central America – country codes 40-95

			<ul style="list-style-type: none"> ▪ North America – country codes 20-70 ▪ Europe – country codes 200-390 ▪ Asia – country codes 700-910 ▪ MENA (Middle East and North Africa) – country codes 600-698 ▪ Africa – country codes 402-590
4	year	Numberial	Năm diễn ra
5	region	Text	Châu lục – nơi diễn ra cuộc biểu tình
6	protest	Numberial	<p>Mã xác định có xảy ra cuộc biểu tình hay không trong 1 khoảng thời gian nhất định.</p> <p>0. Đối với những cuộc biểu tình thể hiện sự phản đối với chính phủ hay chính sách của một nước khác (*)</p>

			<p>1. Đối với những cuộc biểu tình chống đối chính phủ hay chính sách tại đất nước họ biểu tình.</p>
7	protestnumber	Numberial	Số đợt/ lần diễu hành/ biểu tình trong cuộc biểu tình đó
8	startday	Numberial	Ngày bắt đầu biểu tình
9	startmonth	Numberial	Tháng bắt đầu biểu tình
10	startyear	Numberial	Năm bắt đầu biểu tình
11	endday	Numberial	Ngày kết thúc biểu tình
12	endmonth	Numberial	Tháng kết thúc biểu tình
13	endyear	Numberial	Năm kết thúc biểu tình
14	protesterviolence	Numberial	<p>Mã xác định xem cuộc biểu tình này có sử dụng bạo lực (đập phá, phá hoại tài sản, bắn cảnh sát/ quân đội,...) hay không</p> <p>1 Có sử dụng bạo lực</p> <p>0 Không sử dụng bạo lực</p>
15	location	Text	Địa điểm diễn ra biểu tình (một số địa điểm sẽ không được ghi chi tiết, hoặc chỉ có thể cung

			cấp tên thành phố, khu vực hoặc là tên tỉnh)
16	participant_category	Text	<p>Ước tính khoảng người tham gia biểu tình (nhỏ nhất – lớn nhất).</p> <p>Gồm những nhóm như:</p> <ul style="list-style-type: none"> <input type="radio"/> 50-99 <input type="radio"/> 100-999 <input type="radio"/> 1000-1999 <input type="radio"/> 2000-4999 <input type="radio"/> 5000-10000 <input type="radio"/> >10000
17	participant	Text	Ước tính số người tham gia
18	protesteridentity	Text	<p>Miêu tả về nhóm người tham gia biểu tình</p> <p>(người da trắng, người da đen, người thuộc cộng đồng LGBT, học sinh, ca sĩ, nhạc sĩ, công nhân,)</p>
19	protesterdemand1	Text	Vấn đề/ nguyên nhân thúc đẩy người dân biểu tình.

20	protesterdemand2	Text	Bất cứ sự kiện biểu tình nào cũng có thể có nhiều nguyên nhân khiến họ tổ chức biểu tình (trong kho dữ liệu quy định mỗi cuộc biểu tình sẽ có nhiều nhất là 4 nguyên nhân)
21	protesterdemand3	Text	Có 7 loại/ nguyên nhân chính thúc đẩy người dân biểu tình. (**):
22	protesterdemand4	Text	<ol style="list-style-type: none"> 1. <i>Labor or wage dispute:</i> Vấn đề về nhân công/ lao động, lương 2. <i>Land tenure or farm issues:</i> Quyền sở hữu đất đai và vấn đề về trang trại, nông trại 3. <i>Police brutality or arbitrary actions:</i> Hành động bạo lực và tùy tiện của cảnh sát 4. <i>Political behavior/processes:</i> Vấn đề về hành vi chính trị, quy trình chính sách

			<p>5. <i>Price increases or tax policy:</i> Chính sách tăng giá/thuế</p> <p>6. <i>Removal of corrupt or reviled political person:</i> Người nhà nước/người trong chính phủ liên quan đến tham nhũng / phê truất</p> <p>7. <i>Social restrictions:</i> Vấn đề xã hội (ràng buộc xã hội), như tranh cãi về quyền được sử dụng khăn trùm đầu của người Hồi, ...</p>
23	stateresponse1	Text	Có 7 loại hành động (hoặc không hành động) mà chính phủ thực hiện để đáp lại các cuộc biểu tình:
24	stateresponse2	Text	<p>1. <i>Accommodation of demands, indicated by agreeing, negotiating:</i> đáp ứng các yêu cầu, có thể là thương lượng, đồng ý tất cả yêu cầu. Gặp người lãnh</p>
25	stateresponse3	Text	
26	stateresponse4	Text	

27	stateresponse5	Text	<p>đạo cuộc biểu tình để thương lượng, đưa ra yêu cầu giữa hai bên.</p>
28	stateresponse6	Text	<p>2. <i>Arrest</i>: Bắt giữ</p> <p>3. <i>Beatings</i>: đánh đập</p>
29	stateresponse7	Text	<p>4. <i>Crowd dispersal mechanisms</i>: giải tán đám đông bằng một vài biện pháp như: dùng hơi cay, đưa ra cảnh báo, sử dụng quân đội để giải tán đám đông.</p> <p>5. <i>Ignore</i>: không quan tâm. Chia làm 2 trường hợp: Một là báo chí, hãng tin tức đã bỏ qua phản hồi từ chính phủ, hai là chính phủ không phản hồi lại các cuộc biểu tình.</p> <p>6. <i>Killings</i>: Giết người biểu tình</p> <p>7. <i>Shootings</i>: Bắn vào người biểu tình</p>

30	source	Text	Các nguồn cung cấp thông tin như các bài báo, tin tức truyền thông về cuộc biểu tình cụ thể.
31	note	Text	Ghi chú cũng như bình luận về các dữ kiện của các cuộc biểu tình và được trích dẫn từ các nguồn chính thức.

- **Phản giải nghĩa bổ sung (dựa trên tài liệu cung cấp kèm dataset):**

(*), Mục đích của dự án Mass Mobilization:

- Mục tiêu của dự án này là để tìm kiếm những sự kiện biểu tình mà nhắm vào chính phủ nhà nước, và những nhóm này từ 50 người trở lên.
- Chính vì vậy, các hành động biểu tình của người dân thường hướng đến chính phủ hoặc **chính sách của nhà nước** đó.
- Do vậy, kho dữ liệu này chỉ tập trung những cuộc biểu tình trong nội bộ đất nước nên sẽ không chú trọng những cuộc biểu tình tranh đấu liên cộng đồng với nhau. Ví dụ như giữa nhóm người Hồi giáo và người Kito Giáo trong một cộng đồng có thể có nhóm xảy ra xích mích, diễu hành phản đối nhóm người còn lại. Mặc dù, sự việc đó có thể có sự tham gia của cảnh sát (người của chính phủ) nhưng cũng không được tác giả kho dữ liệu, và không được xếp loại vào một hành động ‘anti-state protest’ (biểu tình chống chính phủ)

- Những trường hợp như các đạo quân nổi loạn có trang bị vũ khí tiến hành tấn công cảnh sát, quân đội thì cũng sẽ không được tác giả đưa vào kho dữ liệu.

() 7 loại/ nguyên nhân chính thúc đẩy người dân biểu tình:**

1) Labor or wage dispute

- Những vấn đề liên quan đến chính sách của mỗi nước về điều kiện lao động (giờ lao động, lương tối thiểu, ...) đối với người dân và doanh nghiệp

2) Land tenure or farm issues

- Quyền sở hữu đất đai và vấn đề về trang trại, nông trại
- Nếu chính sách của nhà nước làm ảnh hưởng đến quyền sử dụng đất và chính sách đó mà tạo ra một hành động biểu tình phản đối hay yêu cầu sửa đổi chính sách đó thì nó sẽ được liệt kê vào trong kho dữ liệu.
- Ví dụ về một bang/ chính phủ sử dụng đất của những người nông dân cho một dự án xây đập thì có thể làm nổi dậy biểu tình đòi trả đất từ những người nông dân đó.

3) Police brutality or arbitrary actions - Hành động bạo lực và vô cớ từ cảnh sát

- Mục này gồm những hành động phản đối các hành vi ứng xử không đúng đắn bởi người của chính phủ đối với người dân. Hành vi đánh

đập, bỏ tù người dân với những lý do vô lý/ tùy tiện bởi cảnh sát hay người có thẩm quyền đối với cá nhân hay nhóm người nào đó thì sẽ được lưu trữ vài khi dữ liệu này.

4) *Political behavior/processes:*

- Liên quan đến hành vi bất đồng quan điểm/ phản đối về chính trị, quy trình chính sách hay kêu gọi tranh cử.

5) *Price increases or tax policy:*

- Vấn đề về chính sách tăng giá/ thuế như trợ cấp, tăng thuế, chi phí thực phẩm, điện nước, và nhu cầu thiết yếu khác ...

6) *Removal of corrupt or reviled political person:*

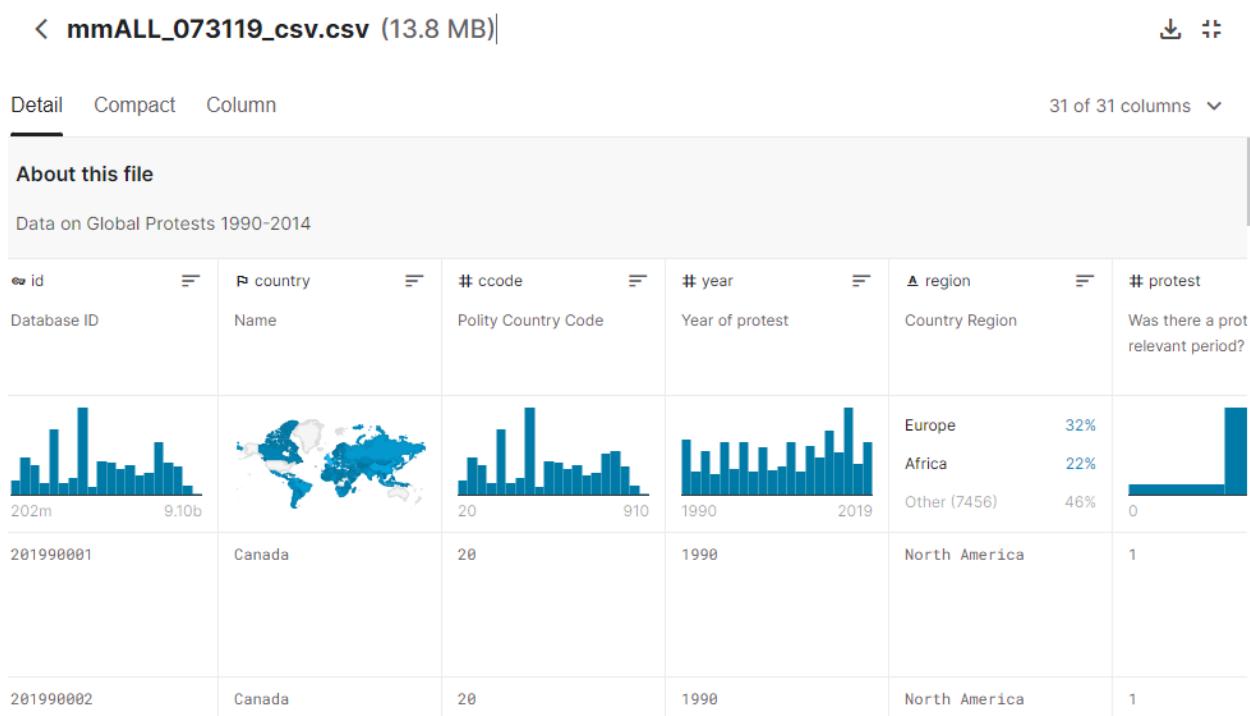
- Người nhà nước/người trong chính phủ liên quan đến tham nhũng / phế truất. Ở đây, tác giả dự án chủ yếu tìm kiếm vụ tham nhũng có hệ thống mà làm thúc đẩy người dân nổi dậy phản đối, yêu cầu phế truất một cá nhân hoặc một nhóm nhỏ trong chính phủ.

7) *Social restrictions:*

- Vấn đề xã hội (rằng buộc xã hội), như tranh cãi về quyền được sử dụng khăn trùm đầu của người Hồi, quyền bình đẳng cho cộng đồng LGBT, ...

The screenshot shows the Kaggle interface. On the left, there's a sidebar with navigation links like Home, Competitions, Datasets, Code, Discussions, Courses, and More. Below these are sections for Your Work and Recently Viewed datasets. The main area displays a dataset titled 'Mass Mobilization Protest Data' with a thumbnail image of a protest crowd. Below the thumbnail, it says 'Data on Global Protests 1990-2014' and 'Sabine Bot • updated a year ago (Version 1)'. There are tabs for Data, Code (1), Discussion, Activity, and Metadata. A prominent 'Download (14 MB)' button is visible, along with a 'New Notebook' button. Below the download button, there are 'Usability 7.1' and 'Tags' sections.

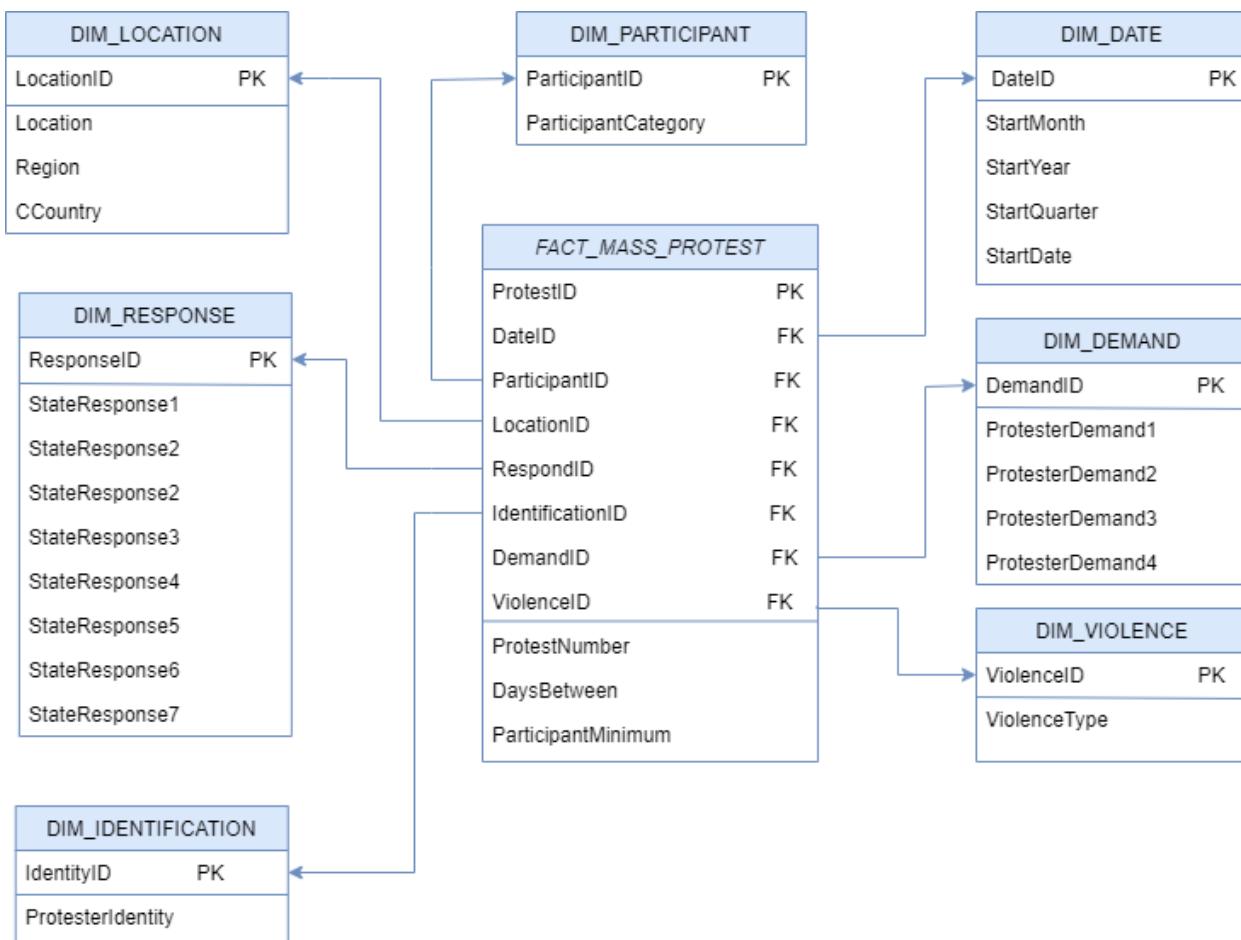
Hình 1. Hình ảnh kho dữ liệu ‘Mass Mobilization Protest’ trên trang Kaggle



Hình 2. Kho dữ liệu khi chưa xử lý

2. XÂY DỰNG KHO DỮ LIỆU

2.1 Sơ đồ hình sao minh họa



Hình 3. Sơ đồ hình sao minh họa

2.2 DIM_PARTICIPANT

Khóa chính	Tên thuộc tính	Kiểu dữ liệu	Mô tả
🔑	ParticipantID	int	Mã ước lượng khoảng người tham gia

	ParticipantCategory	Varchar	<p>Ước tính khoảng người tham gia biểu tình (nhỏ nhất – lớn nhất). Gồm những nhóm như:</p> <ul style="list-style-type: none"> <input type="radio"/> 50-99 <input type="radio"/> 100-999 <input type="radio"/> 1000-1999 <input type="radio"/> 2000-4999 <input type="radio"/> 5000-10000 <input type="radio"/> >10000
--	---------------------	---------	--

2.3 DIM_LOCATION

Khóa chính	Tên thuộc tính	Kiểu dữ liệu	Mô tả
	LocationID	Int	Mã địa điểm diễn ra biểu tình
	Location	Varchar	Địa điểm diễn ra biểu tình
	Region	Varchar	Châu lục diễn ra biểu tình
	Ccountry	varchar	Đất nước diễn ra biểu tình

2.4 DIM_RESPONSE

Khóa chính	Tên thuộc tính	Kiểu dữ liệu	Mô tả
	ResponseID	Int	Mã loại hành động của chính phủ
	StateResponse1	Varchar	

	StateResponse2	Varchar	Các loại hành động (hoặc không hành động) mà chính phủ thực hiện để đáp lại các cuộc biểu tình.
	StateResponse3	Varchar	
	StateResponse4	Varchar	
	StateResponse5	Varchar	
	StateResponse6	Varchar	
	StateResponse7	Varchar	<p>Gồm 7 loại:</p> <ul style="list-style-type: none"> ▪ <i>Accommodation of demands, indicated by agreeing, negotiating:</i> đáp ứng các yêu cầu, có thể là thương lượng, đồng ý tất cả yêu cầu. ▪ Gặp người lãnh đạo cuộc biểu tình để thương lượng, đưa ra yêu cầu giữa hai bên. ▪ <i>Arrest:</i> bắt giữ ▪ <i>Beatings:</i> đánh đập ▪ <i>Crowd dispersal mechanisms:</i> giải tán đám đông bằng một vài biện pháp như: dùng hơi cay, đưa ra cảnh báo, sử dụng quân đội để giải tán đám đông.

			<ul style="list-style-type: none">▪ <i>Ignore</i>: không quan tâm. Chia làm 2 trường hợp: Một là báo chí, hằng tin tức đã bỏ qua phản hồi từ chính phủ, hai là chính phủ không phản hồi lại các cuộc biểu tình.▪ <i>Killings</i>: Giết người biểu tình▪ <i>Shootings</i>: Bắn vào người biểu tình▪ Bất cứ sự kiện biểu tình nào cũng có thể có nhiều hành động từ chính phủ để đáp trả lại cuộc biểu tình (trong kho dữ liệu quy định mỗi cuộc biểu tình sẽ có nhiều nhất là 7 hành động đáp trả từ chính phủ)
--	--	--	---

2.5 DIM_IDENTIFICATION

Khóa chính	Tên thuộc tính	Kiểu dữ liệu	Mô tả
	IndentificationID	int	Mã miêu tả nhóm người tham gia biểu tình
	ProtesterIdentity	varchar	Miêu tả về nhóm người tham gia biểu tình (người da trắng, người da đen, người thuộc cộng đồng LGBT, học sinh, ca sĩ, nhạc sĩ, công nhân,)

2.6 DIM_DATE

Khóa chính	Tên thuộc tính	Kiểu dữ liệu	Mô tả
	DateID	int	Mã ngày sự kiện
	StartMonth	int	Tháng bắt đầu biểu tình
	StartYear	int	Năm bắt đầu biểu tình
	StartQuarter	int	Quý bắt đầu biểu tình
	StartDate	date	Thời gian bắt đầu biểu tình (ngày/tháng/năm)

- *Giải thích bổ sung những thuộc tính mới:*

Thuộc tính	Kiểu dữ liệu	Mô tả	Ghi chú
StartQuarter	int	Quý bắt đầu biểu tình	
StartDate	date	Thời gian bắt đầu biểu tình (ngày/ tháng/ năm)	tạo từ <i>StartDay</i> , <i>StartMonth</i> , <i>StartYear</i>

2.7 DIM_DEMAND

Khóa chính	Tên thuộc tính	Kiểu dữ liệu	Mô tả
	DemandID	int	Mã yêu cầu của người dân
	ProtesterDemand1	varchar	Vấn đề/ nguyên nhân thúc đẩy
	ProtesterDemand2	varchar	người dân biểu tình.
	ProtesterDemand3	varchar	Có 7 loại/ nguyên nhân chính thúc
	ProtesterDemand4	varchar	đẩy người dân biểu tình. 1) <i>Labor or wage dispute</i> : Vấn đề về nhân công/ lao động, lương 2) <i>Land tenure or farm issues</i> : Quyền sở hữu đất đai và vấn đề về trang trại, nông trại

			<p>3) <i>Police brutality or arbitrary actions:</i> Hành động bạo lực và tùy tiện của cảnh sát</p> <p>4) <i>Political behavior/processes:</i> Vấn đề về hành vi chính trị, quy trình chính sách</p> <p>5) <i>Price increases or tax policy:</i> Chính sách tăng giá/ thuế</p> <p>6) <i>Removal of corrupt or reviled political person:</i> Người nhà nước/người trong chính phủ liên quan đến tham nhũng / phê truất</p> <p>7) <i>Social restrictions:</i> Vấn đề xã hội (ràng buộc xã hội), như tranh cãi về quyền được sử dụng khăn trùm đầu của người Hồi,...</p> <p>8) Bất cứ sự kiện biểu tình nào cũng có thể có nhiều nguyên nhân khiến họ tổ chức biểu tình (trong kho dữ liệu quy định mỗi cuộc biểu tình sẽ</p>
--	--	--	---

			có nhiều nhất là 4 nguyên nhân)
--	--	--	---------------------------------

2.8 DIM_VIOLENCE

Khóa chính	Tên thuộc tính	Kiểu dữ liệu	Mô tả
	ViolenceID	int	Mã loại biểu tình
	ViolenceType	varchar	Loại biểu tình 0. Biểu tình ôn hòa 1. Biểu tình bạo lực (có sử dụng vũ khí, bạo lực vũ trang)

2.9 FACT_MASS_PROTEST

Khóa chính	Tên thuộc tính	Kiểu dữ liệu	Mô tả
	ProtestID	int	Mã cuộc biểu tình
	DateID	int	Mã ngày sự kiện
	ParticipantID	int	Mã miêu tả nhóm người tham gia
	LocationID	int	Mã địa điểm diễn ra
	RespondID	int	Mã hành động phản ứng của chính phủ đối với cuộc biểu tình

	DemandID	int	Mã yêu cầu của người dân
	ViolenceID	int	Mã loại biểu tình
	ProtestNumber	int	Số lần biểu tình
	ProtestMinimum	int	Ước tính khoảng người tham gia biểu tình (ít nhất)
	DaysBetween	int	Số ngày diễn ra biểu tình.

- *Giải thích bổ sung những thuộc tính mới:*

Thuộc tính	Kiểu dữ liệu	Mô tả	Ghi chú
ProtestMinimum	int	Ước tính khoảng người tham gia biểu tình (ít nhất)	<p>Lấy dữ liệu từ thuộc tính ParticipantCategory (lấy số ít nhất người tham gia biểu tình)</p> <p>Vd: ở ParticipantCategory</p> <p>Ta biết được cuộc biểu tình đó có khoảng người ước lượng tham gia từ 50-99 người.</p>

			Thì ProtestMinimum sẽ có dữ liệu là 50
DaysBetween	int	Số ngày diễn ra biểu tình	<p>Lấy thời gian kết thúc trừ đi thời gian bắt đầu để tính số ngày diễn ra</p> <p>DaysBetween = EndDate – StartDate</p> <p>-EndDate: gộp từ 3 cột EndDay, EndMonth, EndYear</p>

3. NỘI DUNG 15 CÂU TRUY VẤN

Câu 1: Cho biết tổng số cuộc biểu tình ở Russia theo từng năm.

Câu 2: Cho biết thông tin cuộc biểu tình, nước xảy ra biểu tình có tổng số lượng người ước tính tham gia biểu tình cao nhất trong quý 3/1990 .

Câu 3: Cho biết tên quốc gia có tổng số lượng người tham gia >500000 tại Châu Á.

Câu 4: Theo từng tháng, quý, năm liệt kê số lần diễn ra biểu tình tại các nước ở Châu Á.

Câu 5: Cho biết thông tin nhóm người biểu tình và tổng số cuộc biểu tình của cuộc biểu tình đó lớn hơn 15 được đặt theo từng quốc gia.

Câu 6: Với từng châu lục liệt kê số ngày biểu tình theo từng năm, trừ Châu Âu.

Câu 7: Đưa ra top 5 quốc gia có số cuộc biểu tình diễn ra nhiều nhất trong năm 2017.

Câu 8: Thống kê số cuộc biểu tình, số người tham gia và số ngày biểu tình của mục đích biểu tình thứ nhất, sắp xếp số ngày biểu tình theo thứ tự giảm dần.

Câu 9: Cho biết quốc gia có số ngày biểu tình cao nhất và quốc gia có số lần biểu tình thấp nhất.

Câu 10: Liệt kê những quốc gia có số cuộc biểu tình từ 3 đến 10.

Câu 11: Xuất ra top 2 những thành phố thuộc Châu Phi có số lần biểu tình thấp nhất trong năm 2018.

Câu 12: Với mỗi châu lục, đưa ra 3 quốc gia có tổng số ngày biểu tình cao nhất.

Câu 13: Liệt kê những cuộc biểu tình ở các quốc gia tại Bắc Mĩ, xảy ra từ năm 1998 đến 2002 với số người tham gia < 1000

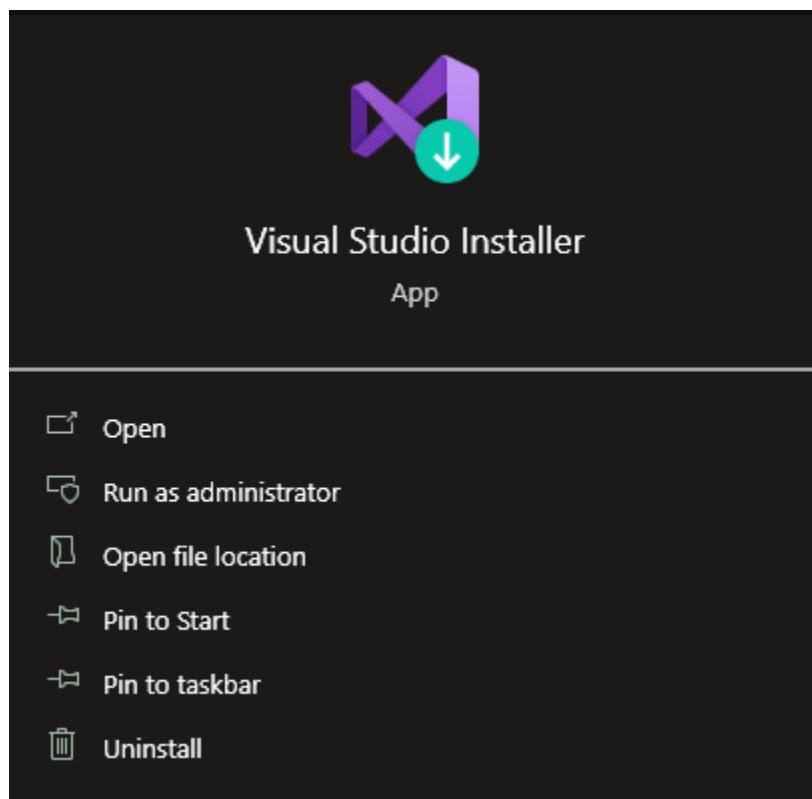
Câu 14: Liệt kê những quốc gia có các cuộc biểu tình diễn ra với hình thức ôn hòa vào năm 2019, cũng như số ngày diễn ra cuộc biểu tình đó.

Câu 15: Cho biết số ngày diễn ra những cuộc biểu tình có sử dụng bạo lực vũ trang cũng như phản ứng đầu tiên của chính phủ đối với cuộc biểu tình đó tại Mexico, sắp xếp theo số ngày biểu tình tăng dần.

CHƯƠNG 2: TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS)

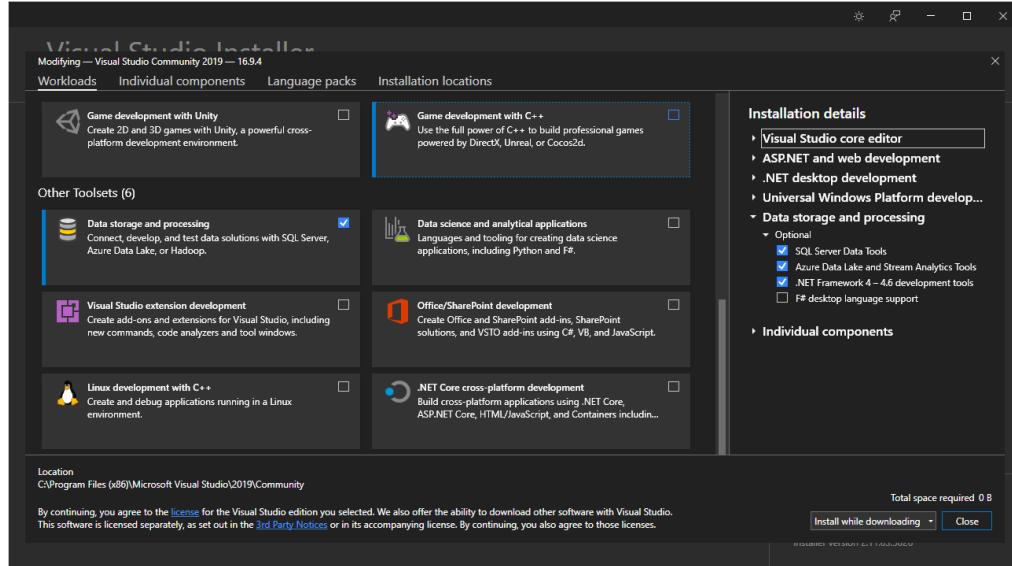
1. CHUẨN BỊ CÔNG CỤ VÀ DATA WAREHOUSE

- Tải và cài đặt công cụ **SSDT** (một công cụ phát triển hiện đại để xây dựng cơ sở dữ liệu quan hệ SQL Server)
- Mở **Visual Studio Installer**:



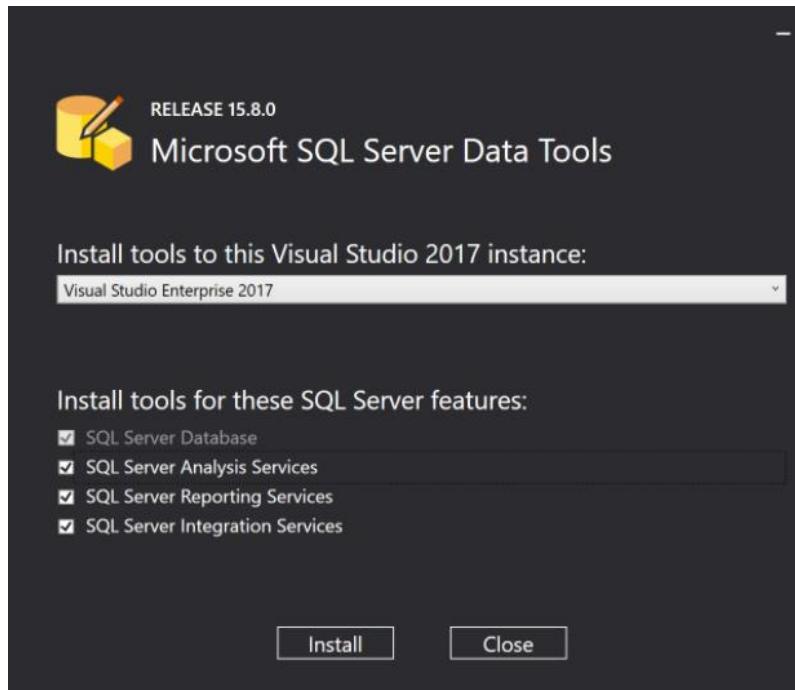
Hình 4. Quá trình cài SQL Server Data Tools

- Cài đặt công cụ SQL Server Data Tools: **Data Storage and Processing**
- Chọn thêm các lựa chọn kèm theo trong gói cài đặt (trong mục **Installation details**)
 - SQL Server Data Tools
 - Azure Data Lake and Stream Analytics Tools
 - .NET Framework 4 - 4.0 development tools



Hình 5. Cài đặt SQL Server Data Tools

- Cài đặt các extension: SQL Analysis Services, Integration Services, and Reporting Services tools



Hình 6. Cài đặt các Extension

2. QUÁ TRÌNH SSIS

2.1 Quá trình nhập dữ liệu vào SQL Server

- **Bước 1:** Tạo 2 database

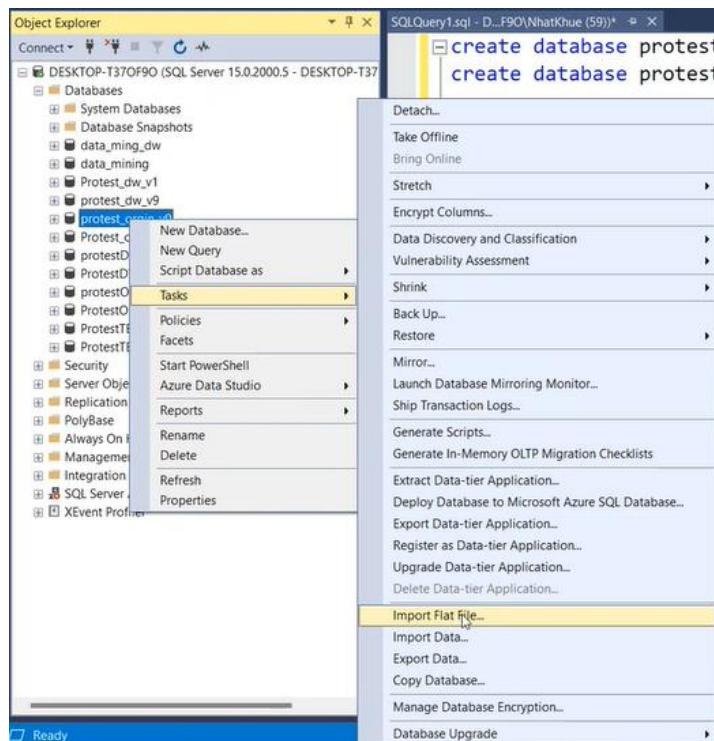
- ProtestOrigin: kho dữ liệu chứa dữ liệu gốc.
- ProtestDW: kho dữ liệu đích, chứa dữ liệu sau khi xử lý làm sạch.

```
--Database du lieu goc---
create database ProtestOrigin

---Database du lieu dich (Destination)
create database ProtestDW
```

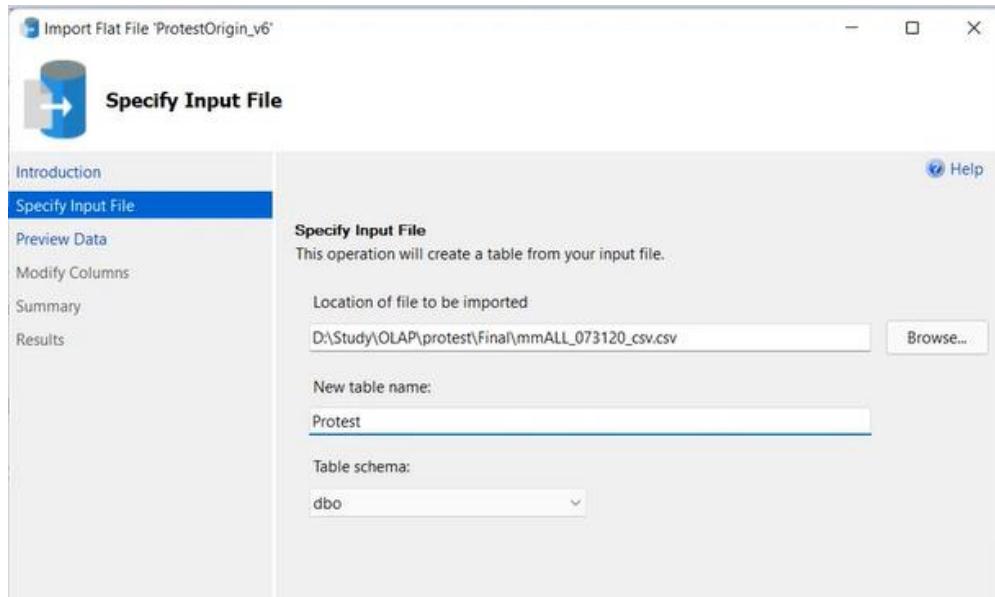
Hình 7. Tạo 2 cơ sở dữ liệu trong SQL Server

- **Bước 2:** Tại kho dữ liệu gốc (ProtestOrigin) > Task > Import Flat File...



Hình 8. Import data gốc vào kho dữ liệu gốc ProtestOrigin

- **Bước 3:** Tại mục *Specify Input File* > chọn *Browse* để chọn file dữ liệu mình cần import vào SQL Server > Đặt tên lại tên bảng tại *New table name* > *Next*



Hình 9. Quá trình dẫn link dữ liệu và đặt tên bảng

- **Bước 4:** Tại mục *Preview Data* > kiểm tra lại dữ liệu mình nhập vào

The screenshot shows the 'Specify Input File' dialog with the 'Preview Data' tab selected. The preview area displays the first 50 rows of the dataset. The columns are labeled: id, country, ccode, year, region, protest, and prot. The data shows various entries for Canada from 1990 to 1997.

id	country	ccode	year	region	protest	prot
201990001	Canada	20	1990	North Amer...	1	1
201990002	Canada	20	1990	North Amer...	1	2
201990003	Canada	20	1990	North Amer...	1	3
201990004	Canada	20	1990	North Amer...	1	4
201990005	Canada	20	1990	North Amer...	1	5
201990006	Canada	20	1990	North Amer...	1	6
201991001	Canada	20	1991	North Amer...	1	1
201991002	Canada	20	1991	North Amer...	1	2
201992001	Canada	20	1992	North Amer...	1	1
201993001	Canada	20	1993	North Amer...	1	1
201993002	Canada	20	1993	North Amer...	1	2
201994001	Canada	20	1994	North Amer...	1	1
201994002	Canada	20	1994	North Amer...	1	2
201995001	Canada	20	1995	North Amer...	1	1
201995002	Canada	20	1995	North Amer...	1	2
201996001	Canada	20	1996	North Amer...	1	1
201997001	Canada	20	1997	North Amer...	1	1
201997002	Canada	20	1997	North Amer...		

Hình 10. Kết quả preview data

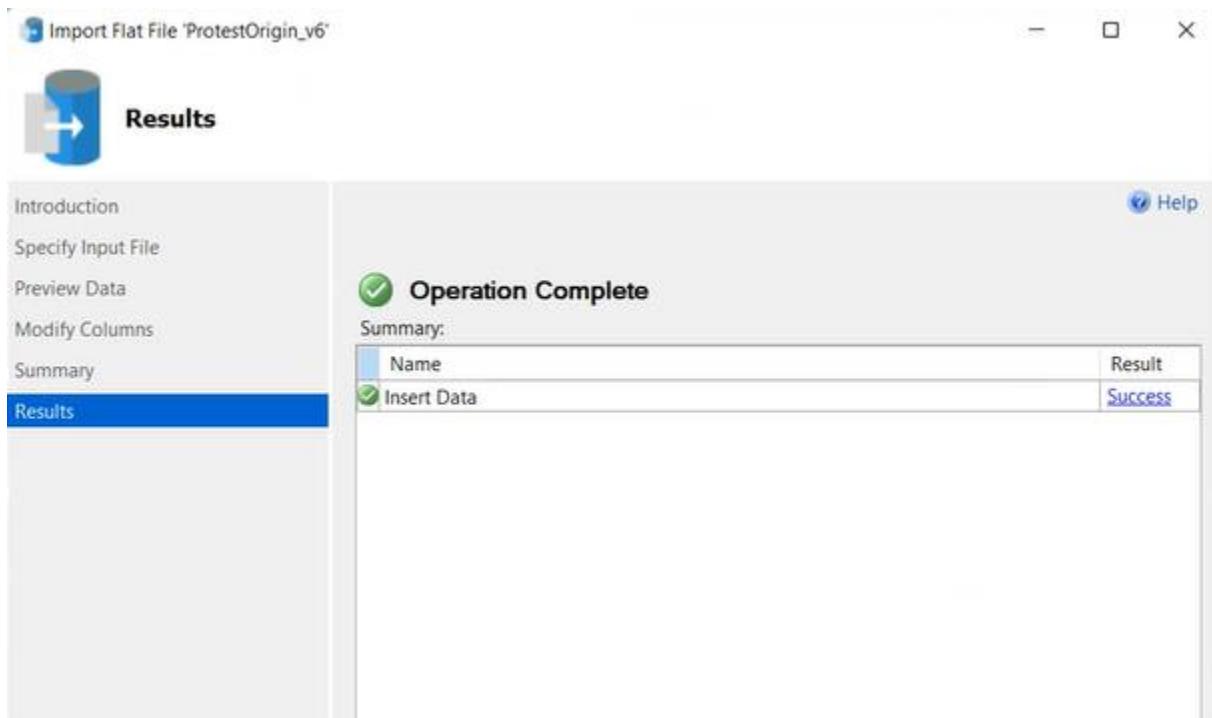
- **Bước 5:** Tại mục *Modify Columns* > thay đổi kiểu dữ liệu, điều chỉnh khóa chính, giá trị Nulls > *Next*

The screenshot shows the 'Modify Columns' dialog with the 'Modify Columns' tab selected. The table schema is displayed with columns: Column Name, Data Type, Primary Key, and Allow Nulls. The data type for most columns is bigint, except for year, region, protestnumber, startday, startmonth, startyear, stardate, endday, endmonth, endyear, enddate, DaysBetween, protesterviolence, location, participants_category, and participants, which are varchar.

Column Name	Data Type	Primary Key	Allow Nulls
id	bigint	<input checked="" type="checkbox"/>	<input type="checkbox"/>
country	varchar(50)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
ccode	int	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
year	int	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
region	varchar(50)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
protestnumber	int	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
startday	int	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
startmonth	int	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
startyear	int	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
stardate	date	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
endday	int	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
endmonth	int	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
endyear	int	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
enddate	date	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
DaysBetween	int	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
protesterviolence	varchar(50)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
location	varchar(MAX)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
participants_category	varchar(50)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
participants	varchar(MAX)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Hình 11. Điều chỉnh thuộc tính

- **Bước 6:** Tại mục *Results* > hiển thị thông báo kết quả import dữ liệu vào SQL Server

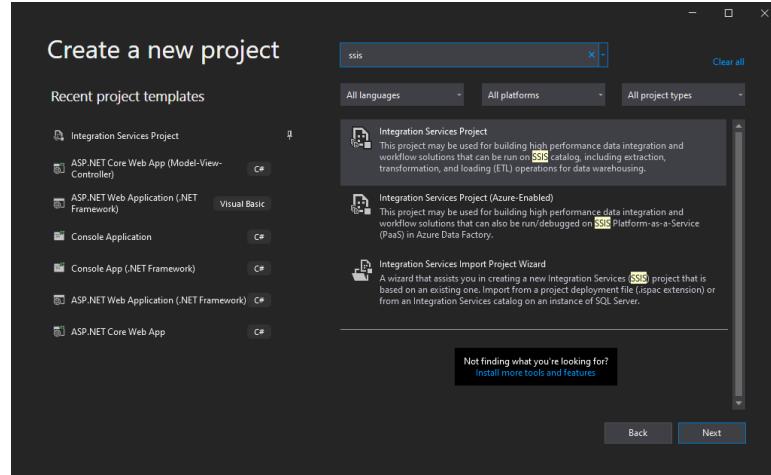


Hình 12. Kết quả import data

2.2 Quá trình SSIS trong Visual Studio 2019

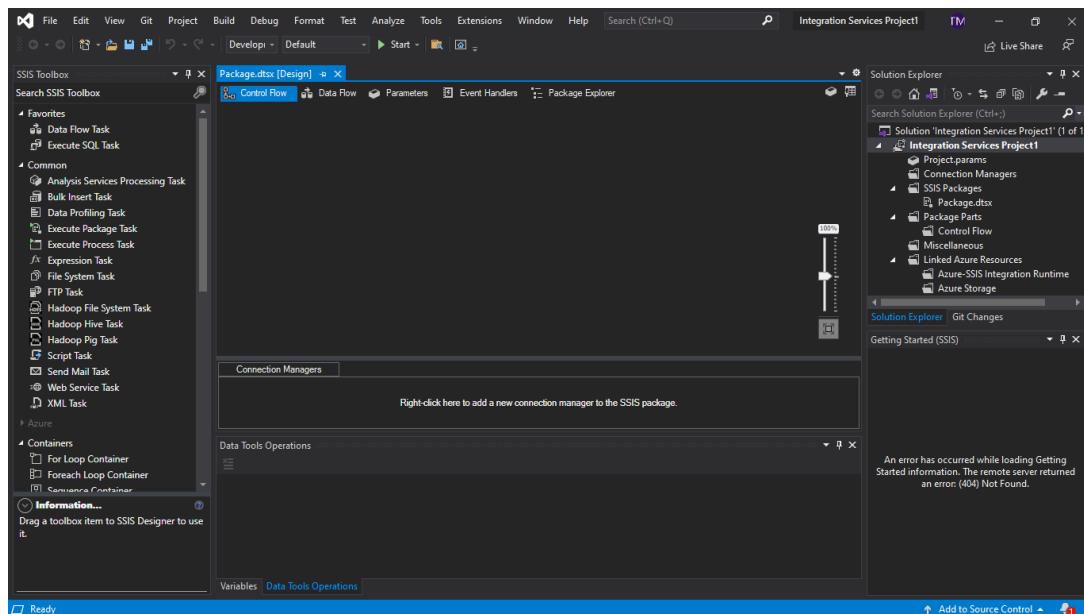
2.2.1 Tạo SSIS Project

- Mở Visual Studio 2019 > *New Project* > search ‘ssis’ > *Integration Services Project* > chọn đường dẫn và tên lưu project của mình



Hình 13. Quá trình tạo project SSIS

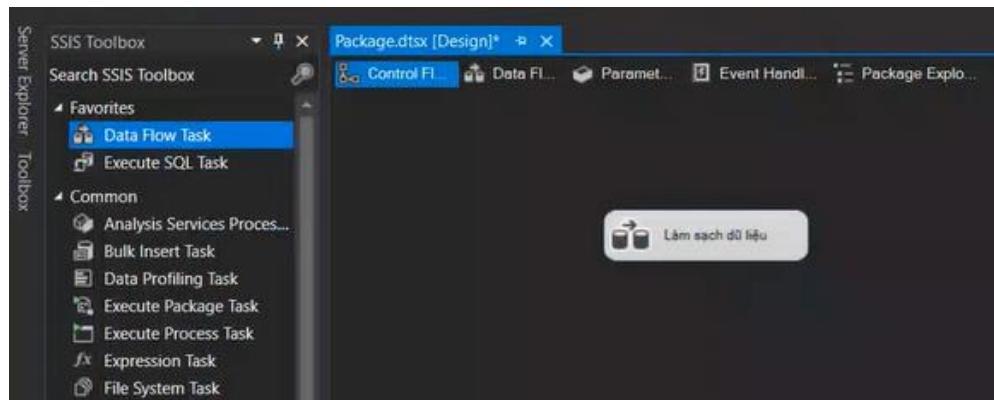
- Giao diện của công cụ BI quá trình SSIS
 - Các chức năng, công cụ chính cho quá trình SSIS nằm bên trái màn hình



Hình 14. Giao diện quá trình SSIS

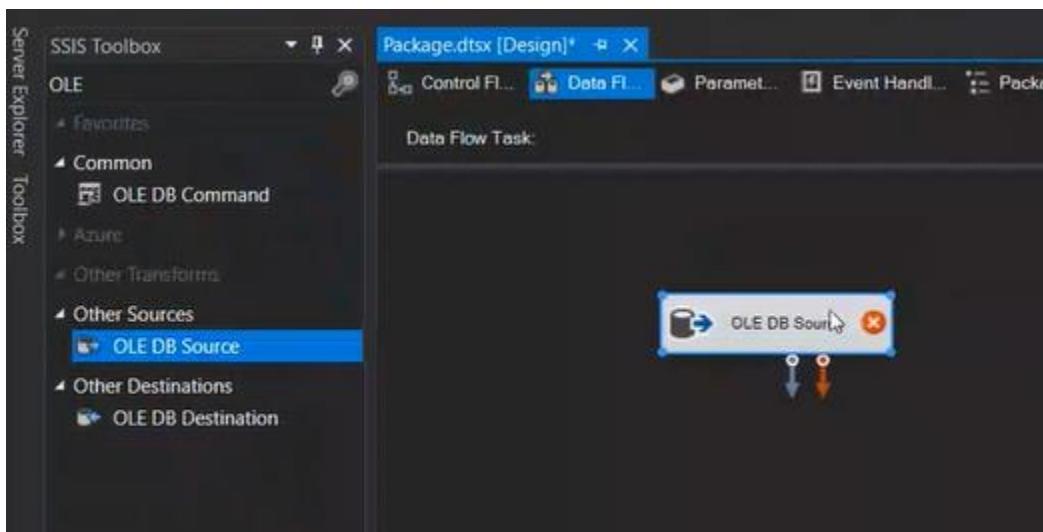
2.2.2 Làm sạch dữ liệu và tạo các bảng DIM

- **Bước 1:** Kéo thả chức năng Data Flow Task từ SSIS Toolbox vào Control Flow Task



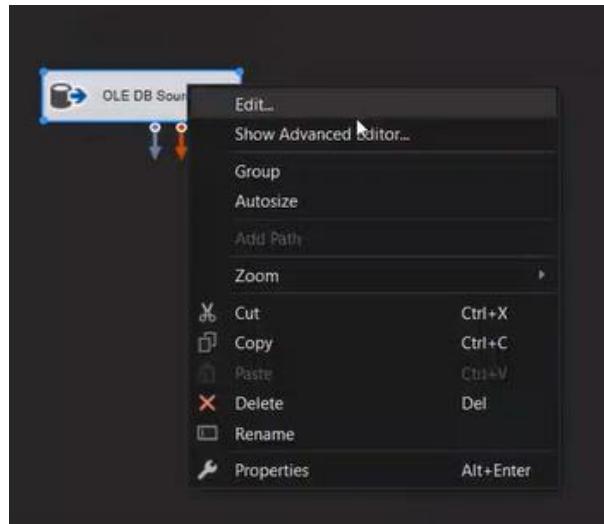
Hình 15. Thao tác sử dụng chức năng Data Flow Task

- **Bước 2:** Nhấn đúp vào Data Flow Task > Kéo thả **OLE DB Source** vào Data Flow Task



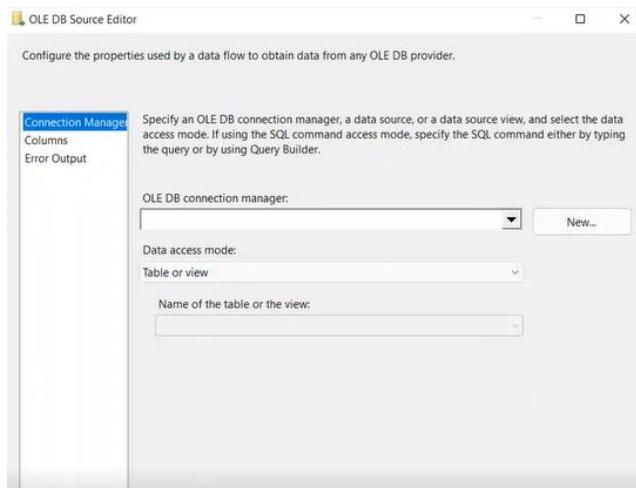
Hình 16. Thao tác sử dụng chức năng OLE DB Source

- **Bước 3:** tiến hành import dữ liệu đã nhập từ *SQL Server*
 - Chuột phải > *Edit*



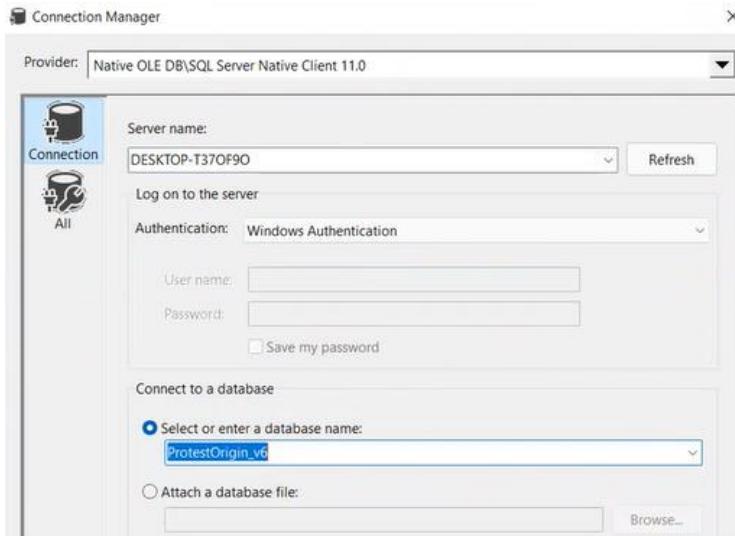
Hình 17. Thao tác tùy chỉnh OLE DB Source

- **Bước 4:** Tạo mới một kết nối tới SQL Server tại *OLE DB connection manager* > chọn *New*



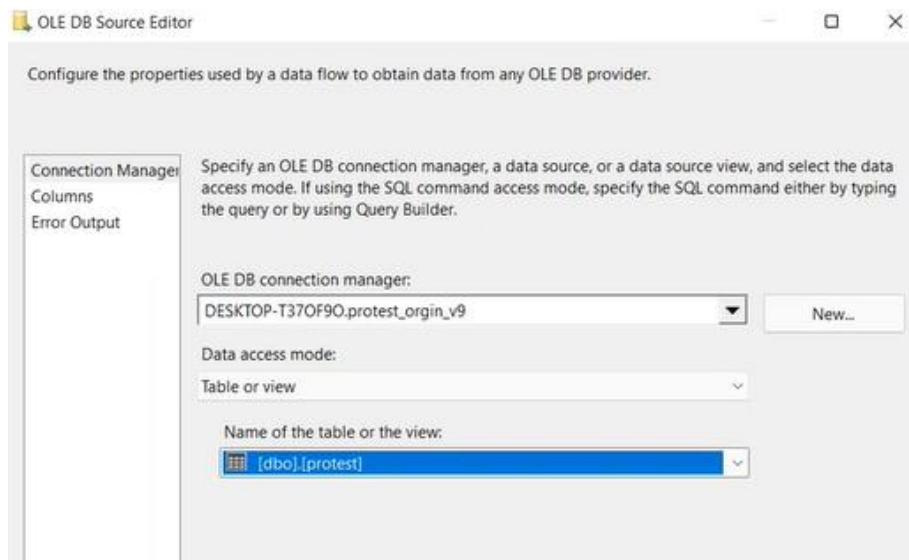
Hình 18. Thao tác chọn kết nối OLE DB

- Nhập *Server Name* của SQL Server và chọn Database gốc (ProtestOrigin) > *OK*



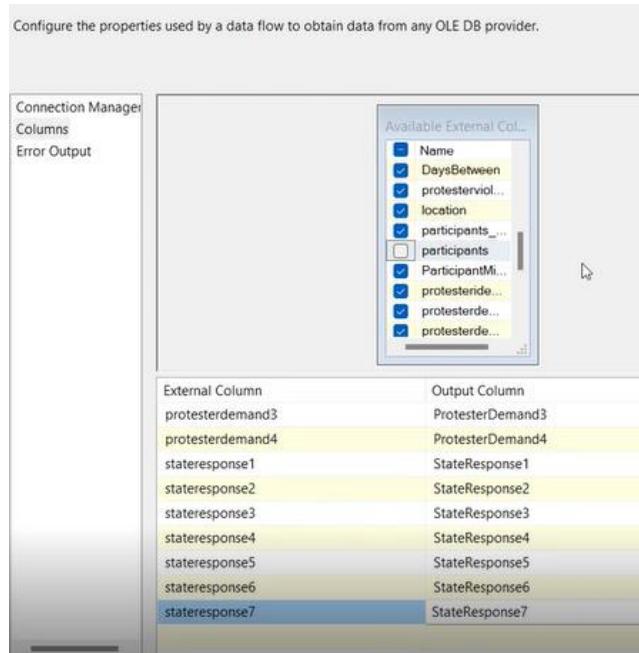
Hình 19. Thao tác nhập Server name và đặt tên database

- **Bước 5:** Quay lại tại trang **OLE DB Source Editor** > chọn bảng (table) mình đã import vào SQL Server. (*protest*)



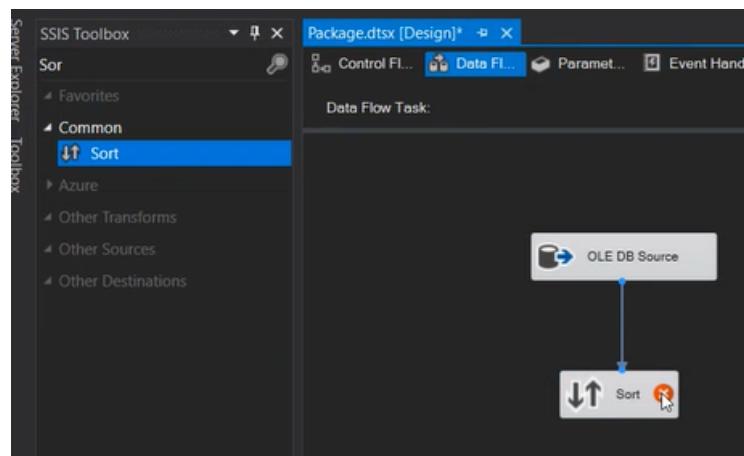
Hình 20. Thao tác chọn bảng mình cần nhập

- **Bước 6:** Tại mục **Columns** > chọn thuộc tính mình cần trong các quá trình sau > thực hiện đổi tên thuộc tính > **OK**



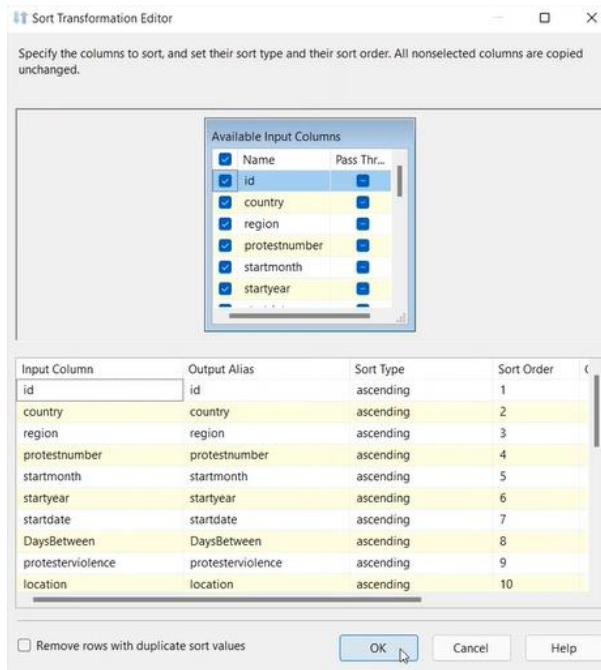
Hình 21. Thao tác chọn thuộc tính cần sử dụng

- Bước 7:** Chọn chức năng *Sort* trong SSIS Toolbox > Kéo mũi tên xanh từ *OLE DB Source* vào ô *Sort* > chuột phải chọn *Edit*



Hình 22. Thao tác với ô Sort

- Bước 8:** Tại màn hình *Sort Transformation Editor* > Chọn những thuộc tính mình muốn sắp xếp theo thứ tự > Kiểu sắp xếp tăng dần hoặc giảm dần

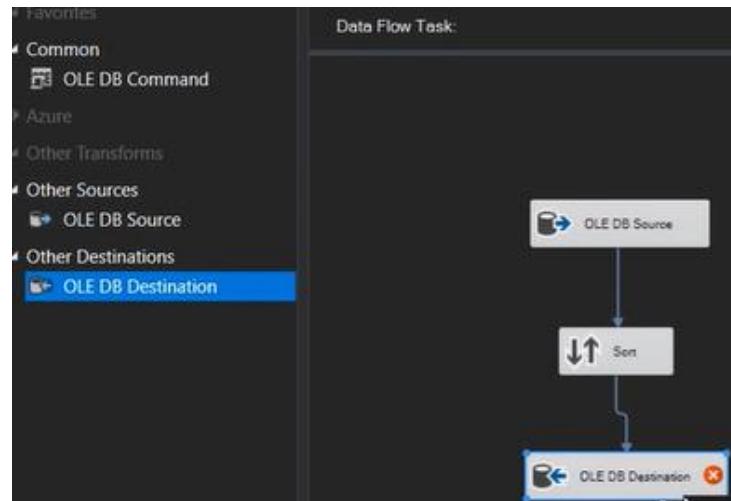
*Hình 23. Thực hiện sắp xếp dữ liệu*

- Lỗi không sắp xếp được thuộc tính **Location** và **ProtesterIdentity** do kí tự nhiều và dài.
⇒ Vào SQL Server, chạy lệnh thay đổi kiểu dữ liệu cho 2 thuộc tính trên

```
--Chinh kieu du lieu cua thuoc tinh---
ALTER TABLE dbo.Protest ALTER COLUMN protesteridentity nvarchar(4000);
ALTER TABLE dbo.Protest ALTER COLUMN location nvarchar(4000);
```

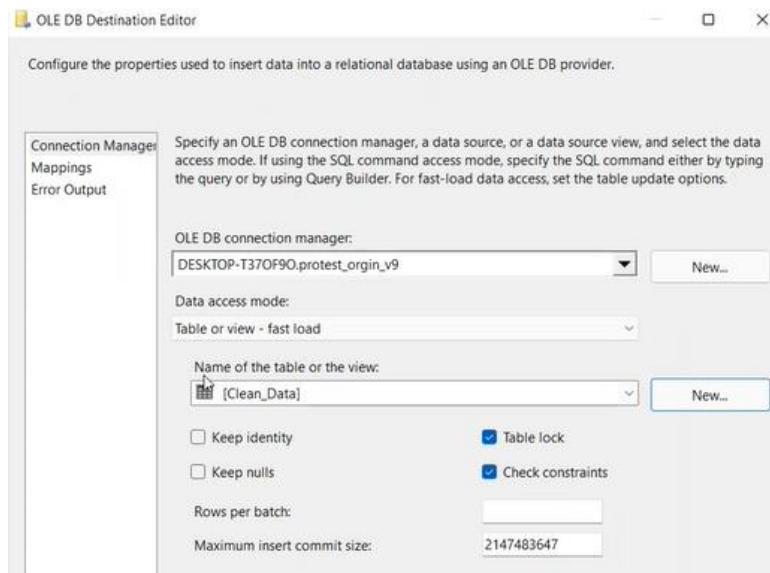
Hình 24. Câu lệnh điều chỉnh kiểu dữ liệu của thuộc tính

- ⇒ Thực hiện lại bước Sort
- **Bước 9:** Lưu trữ dữ liệu đã lọc và sắp xếp vào 1 destination. Kéo thả **OLE DB Destination** > kéo mũi tên từ Sort xuống Destination > **Edit**



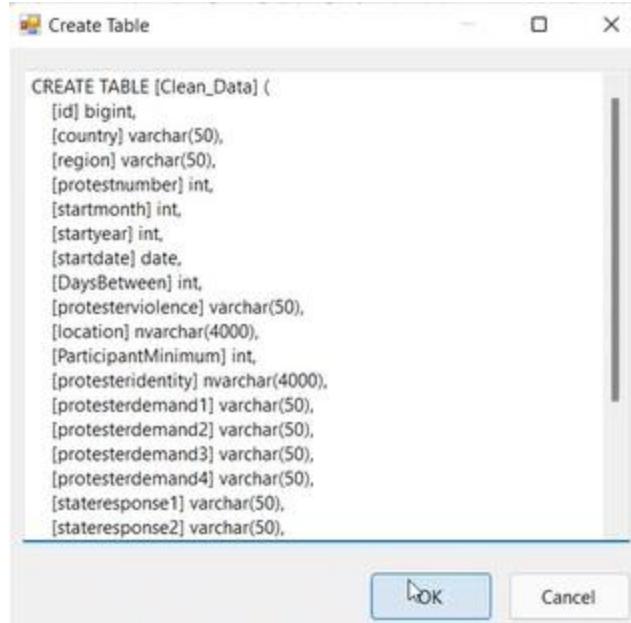
Hình 25. Thao tác với ô OLE DB Destination

- **Bước 10:** Tại **OLE DB Destination Editor** > Tạo mới bảng (table)
> Nhấn **New...** tại Name of table or view



Hình 26. Thao tác chọn tạo mới bảng dữ liệu

- **Bước 11:** Tại màn hình **Create Table** > đổi tên bảng ‘Clean_Data’ > chỉnh sửa thuộc tính, kiểu dữ liệu

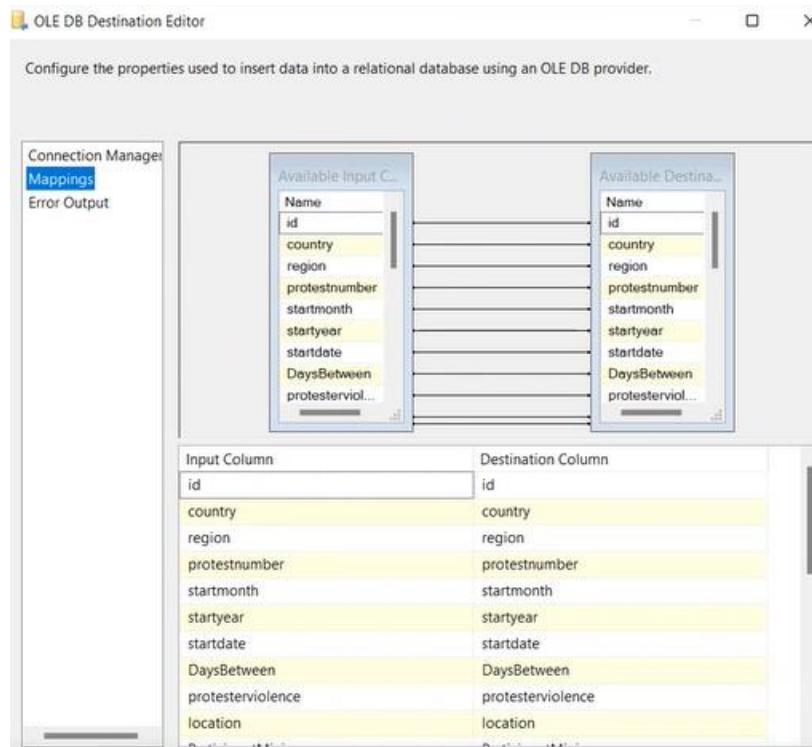


```
CREATE TABLE [Clean_Data] (
    [id] bigint,
    [country] varchar(50),
    [region] varchar(50),
    [protestnumber] int,
    [startmonth] int,
    [startyear] int,
    [startdate] date,
    [DaysBetween] int,
    [protesterviolence] varchar(50),
    [location] nvarchar(4000),
    [ParticipantMinimum] int,
    [protesteridentity] nvarchar(4000),
    [protesterdemand1] varchar(50),
    [protesterdemand2] varchar(50),
    [protesterdemand3] varchar(50),
    [protesterdemand4] varchar(50),
    [stateresponse1] varchar(50),
    [stateresponse2] varchar(50),
)
```

OK Cancel

Hình 27. Câu lệnh tạo bảng Clean Data

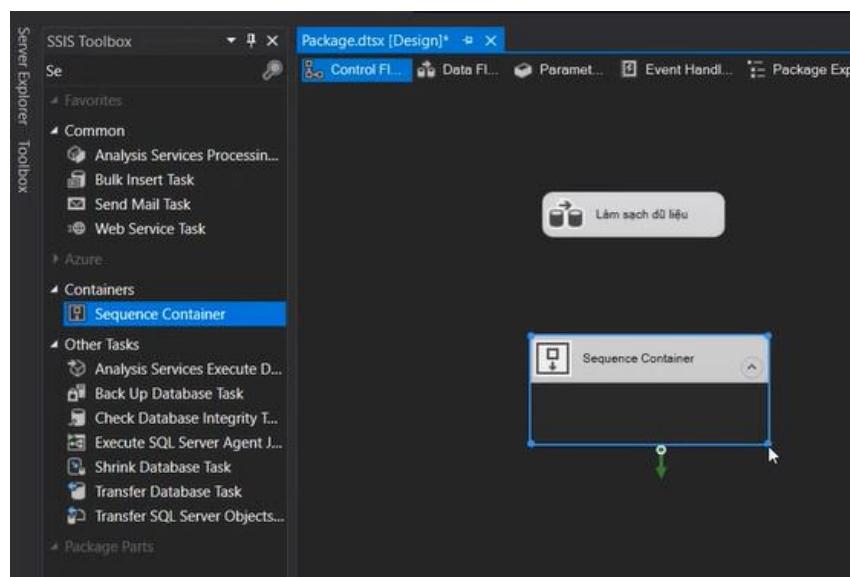
- **Bước 12:** Kiểm tra Mappings



Hình 28. Kiểm tra Mapping các thuộc tính

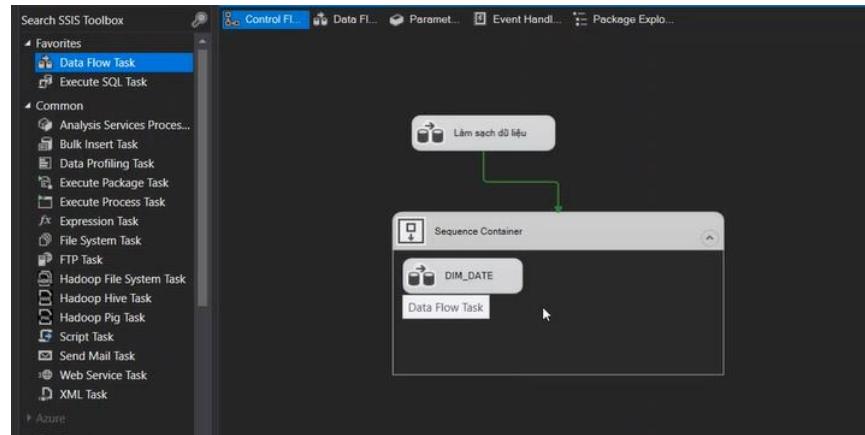
2.2.3 Tạo các bảng Dimension

- **Bước 1:** Quay trở lại *Control Flow Task* > Kéo thả *Sequence Container* vào *Control Flow Task*

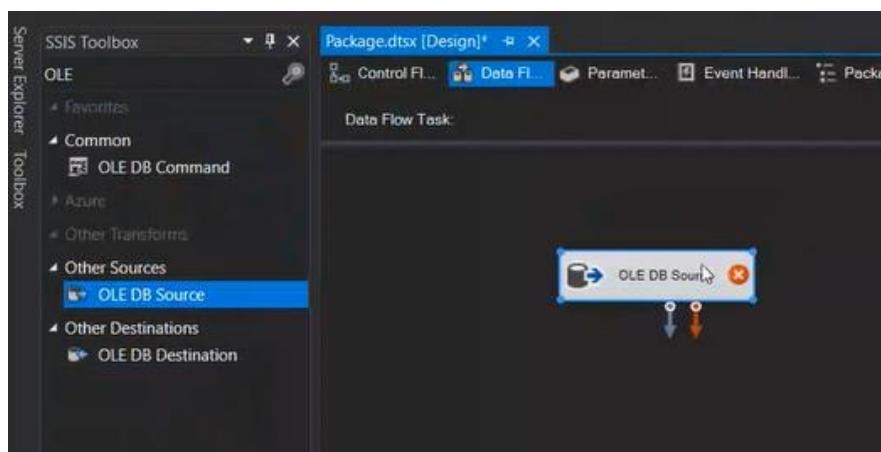


Hình 29. Thao tác với chức năng Sequence Container

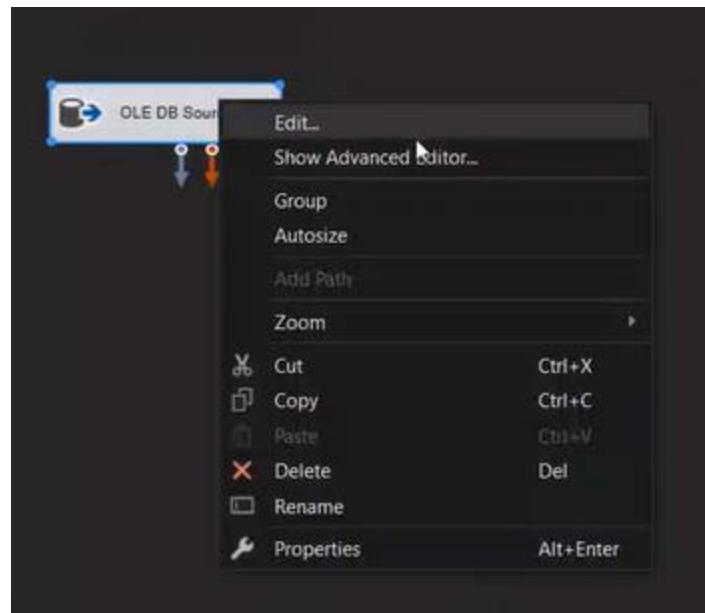
- **Bước 2:** Tạo bảng ***Dim_DATE*** > Kéo thả **Data Flow Task** vào **Sequence Container** > Đổi thành tên bảng

*Hình 30. Thao tác với ô Data Task Flow Dim_Date*

- **Bước 3:** Nhấn đúp vào **Data Flow Task** của ***DIM_DATE*** > Kéo thả **OLE DB Source** vào Data Flow Task

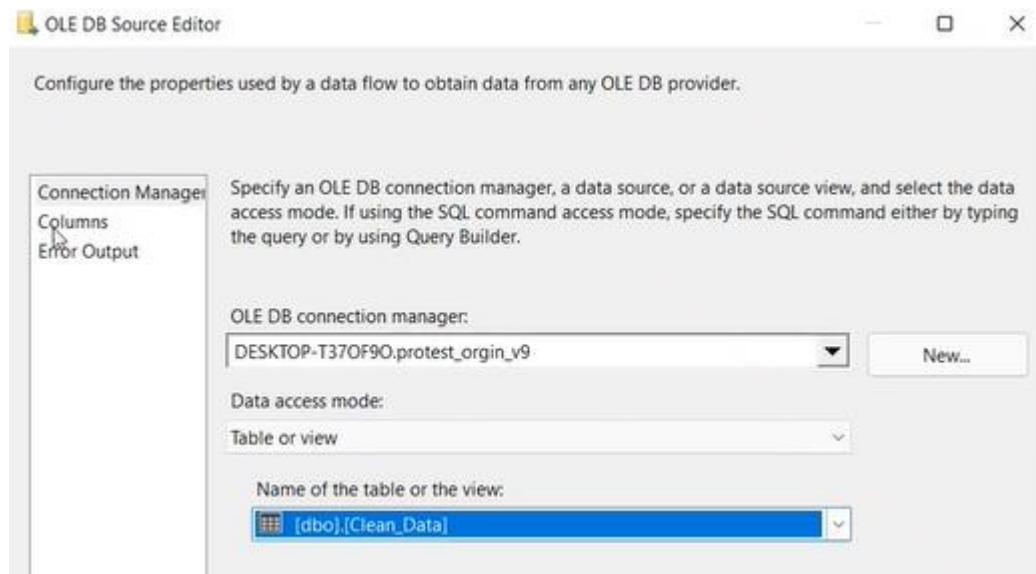
*Hình 31. Thao tác với ô OLE DB Source cho Dim_Date*

- **Bước 4:** tiến hành import dữ liệu đã nhập từ **SQL Server**
 - Chuột phải > **Edit**



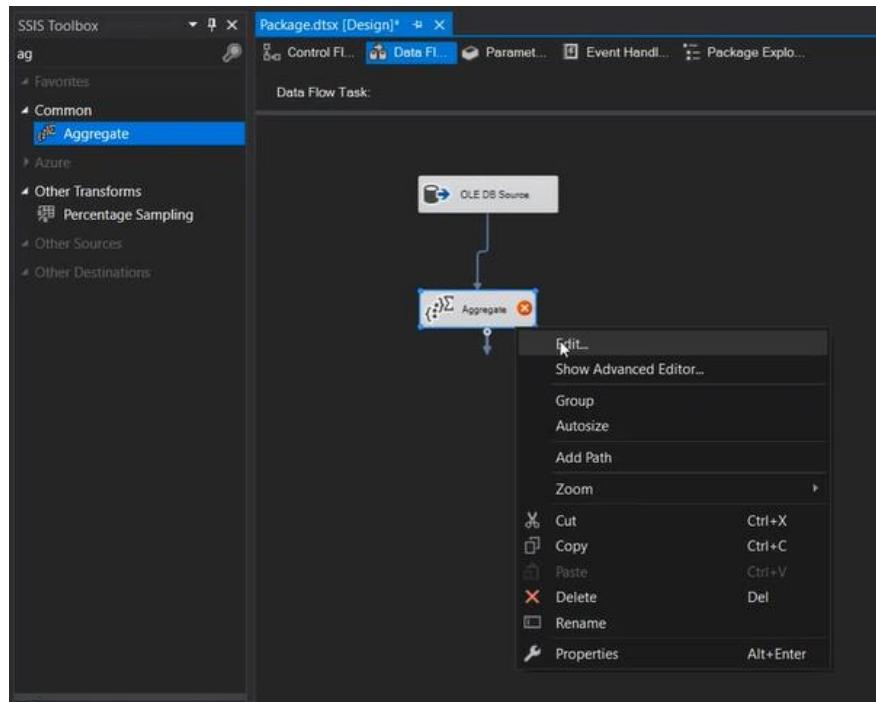
Hình 32. Thao tác chỉnh sửa ô Dim_Date

- **Bước 5:** Chọn bảng [*Clean_Data*] vừa tạo ở quá trình làm sạch dữ liệu > **OK**



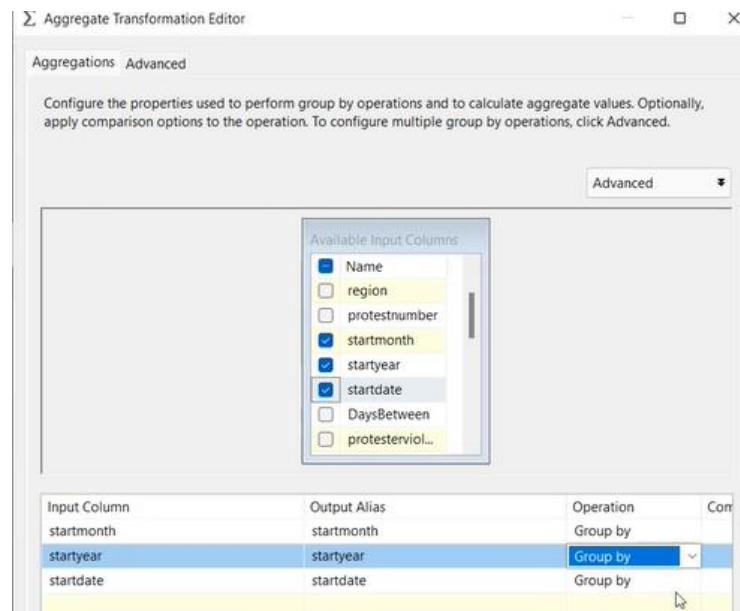
Hình 33. Thao tác chọn bảng ‘Clean Data’

- **Bước 6:** Kéo thả chức năng *Aggregate* > Chuột phải chọn *Edit*



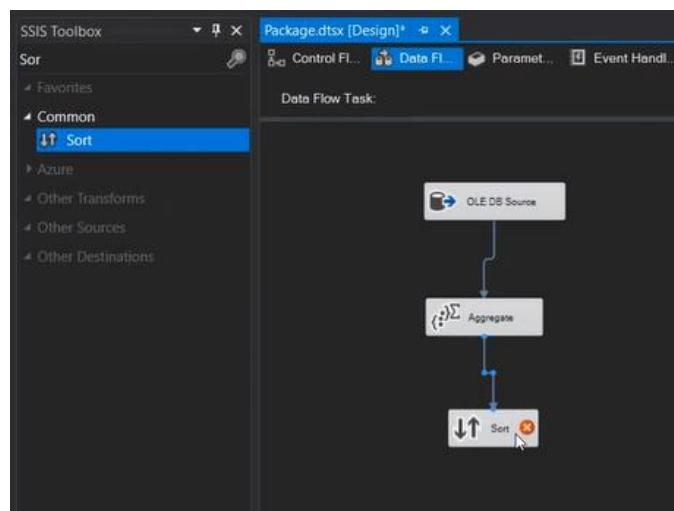
Hình 34. Thao tác với ô Aggregate của Dim_Date

- **Bước 7:** Tại màn hình *Aggregate Transformation Editor* > Tick chọn những thuộc tính có trong bảng DIM_DATE (dựa trên lược đồ hình sao) > Operation = ‘Group by’



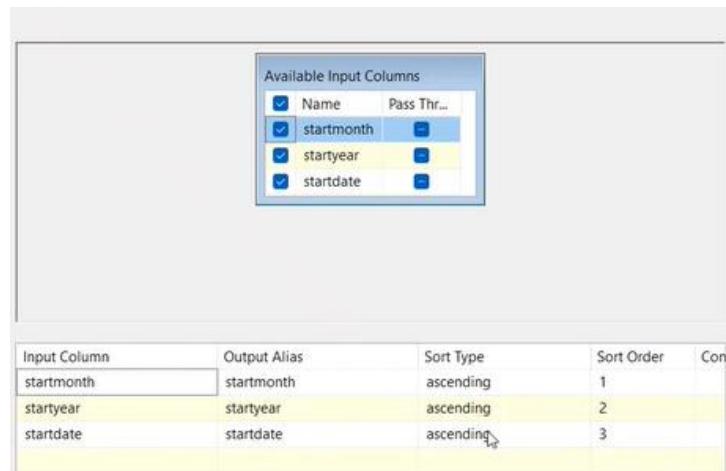
Hình 35. Thao tác chỉnh sửa bằng Aggregate

- **Bước 8:** Kéo thả ô Sort để sắp xếp dữ liệu thuộc tính > **Edit**



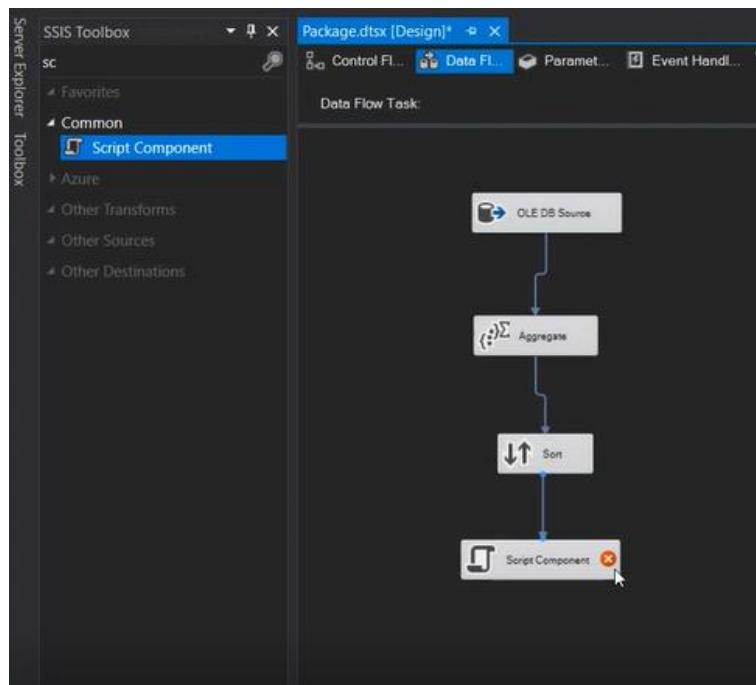
Hình 36. Thao tác với ô Sort

- **Bước 9:** Chọn những thuộc tính, kiểu sắp xếp



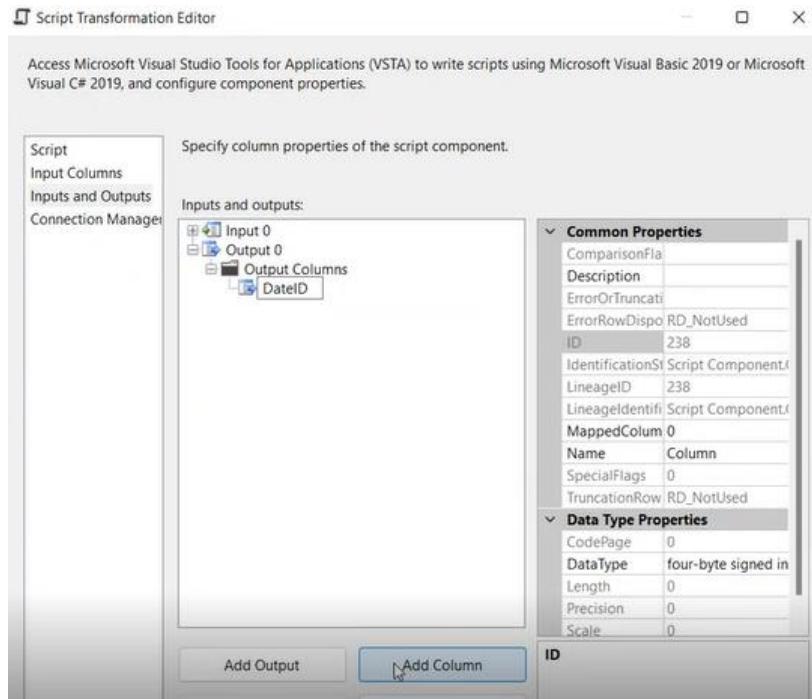
Hình 37. Thực hiện sắp xếp dữ liệu

- Bước 10:** Kéo thả chức năng *Script Component* để thực hiện tạo thêm một biến tăng tự động > Kéo mũi tên từ *Sort* vào *Script Component* > Chuột phải *Edit*



Hình 38. Thao tác với chức năng *Script Component*

- **Bước 11:** Tại *Script Transformation Editor* > chọn mục *Inputs and Outputs* > *Add Column* tại Output Column > Tạo thuộc tính mới tên DateID



Hình 39. Thực hiện thêm cột thuộc tính mới

- **Bước 12:** Tại mục *Script* > Chọn *Edit Script* > Hệ thống sẽ mở giao diện code của Visual Studio > Tạo biến đếm count = 0 > cho chạy tăng tự động.

```

1  Help: Introduction to the Script Component
8
9  Namespaces
15
16  </summary>
17  /// This is the class to which to add your code. Do not change the name, attributes, or parent
18  /// of this class.
19  </summary>
20  [Microsoft.SqlServer.Dts.Pipeline.SSIScriptComponentEntryPointAttribute]
- references
21  public class ScriptMain : UserComponent
22  {
23      int count = 1;
24
25      Help: Using Integration Services variables and parameters
26
27      Help: Using Integration Services Connection Managers
28
29      Help: Firing Integration Services Events
30
31      </summary>
32      /// This method is called once, before rows begin to be processed in the data flow.
33
34      /// You can remove this method if you don't need to do anything here.
35      </summary>
- references
36      public override void PreExecute()
37      {
38          base.PreExecute();
39          /*
40           * Add your code here
41           */
42      }
43
44      </summary>
45      /// This method is called after all the rows have passed through this component.
46
47      /// You can delete this method if you don't need to do anything here.
48      </summary>
- references

```

Hình 40.1 Thực hiện thêm ID tự động

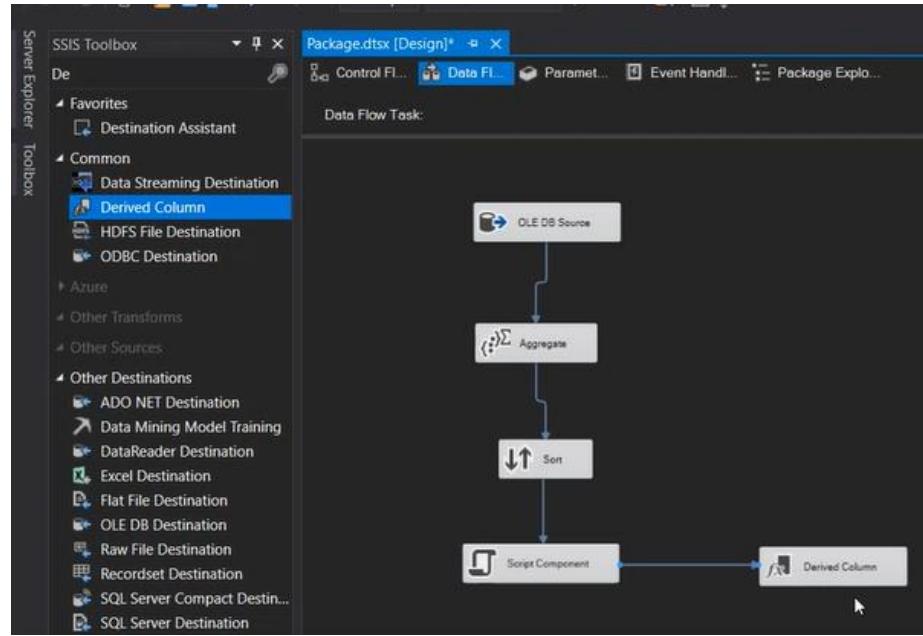
```

/// <param name="Row">The row that is currently passing through the component.
2 references
public override void Input0_ProcessInputRow(Input0Buffer Row)
{
    Row.DateID = count;
    count++;
}

```

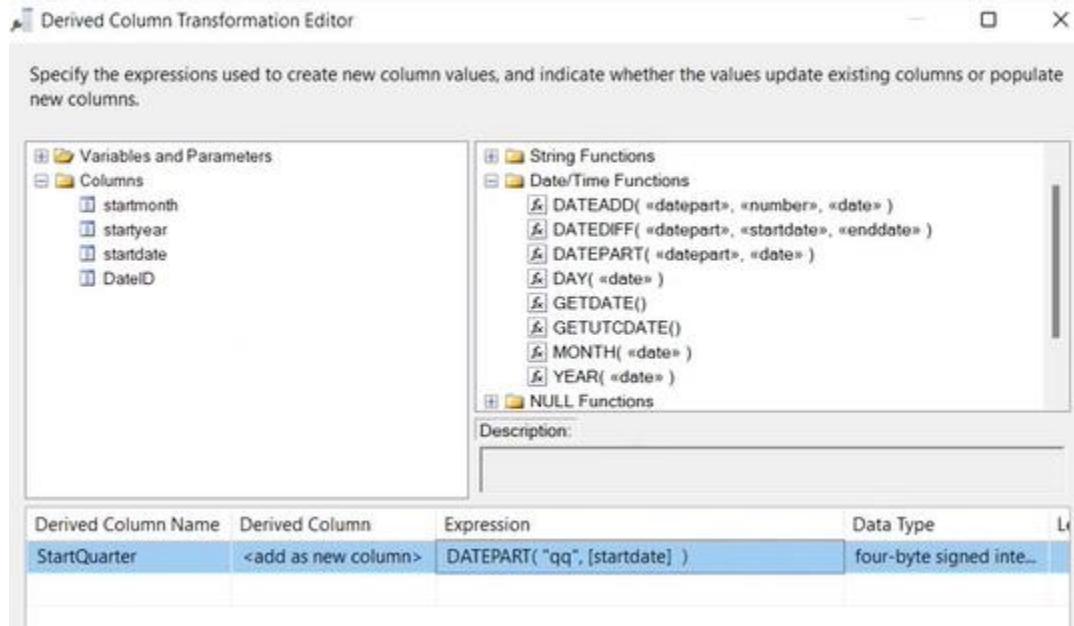
Hình 40.2 Thực hiện thêm ID tự động

- **Bước 13:** Kéo thả chức năng *Derived Column* để thực hiện thêm thuộc tính *StartQuarter* và *EndQuarter* > *Edit*



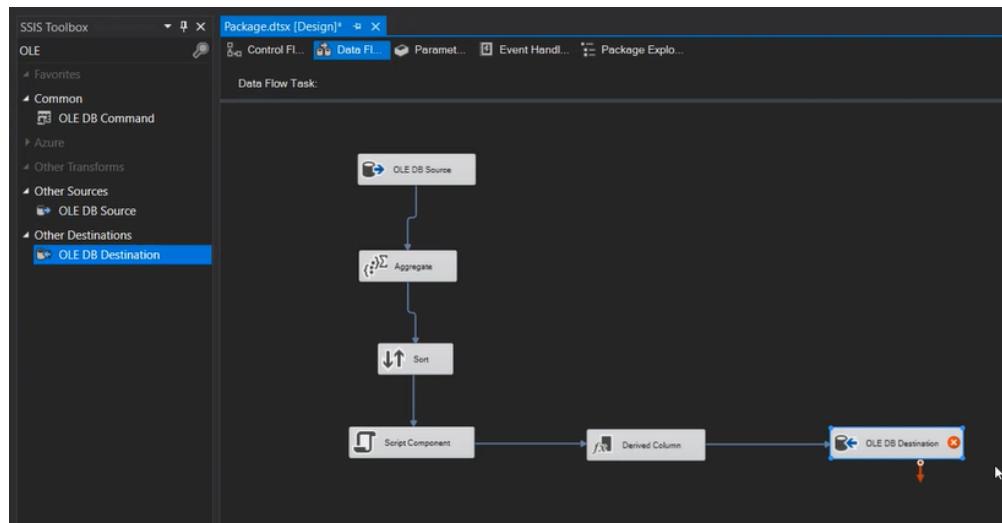
Hình 41. Thực hiện với chức năng Derived Column

- Bước 14:** Tạo 1 cột mới tên <StartQuarter> Expression sử dụng hàm **DATEPART("datepart", "date")** trong thư mục Date Time Functions



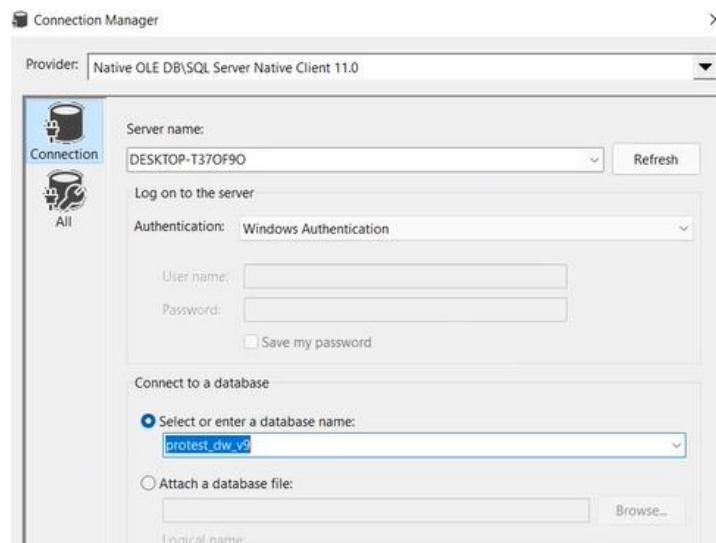
Hình 42. Thực hiện thêm thuộc tính quý

- **Bước 15:** Kéo thả chức năng ***OLE DB Destination*** để truyền dữ liệu bảng ***DIM_DATE*** vào kho dữ liệu đích > *Edit*



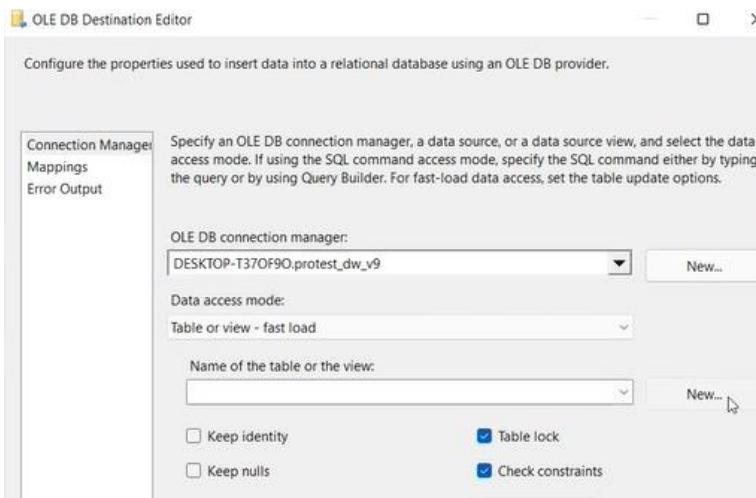
Hình 43. Thao tác với chức năng ***OLE DB Destination***

- **Bước 16:** Tại ***Connection Manager***, chọn ***server name*** của SQL Server và kho dữ liệu đích là ProtestDW



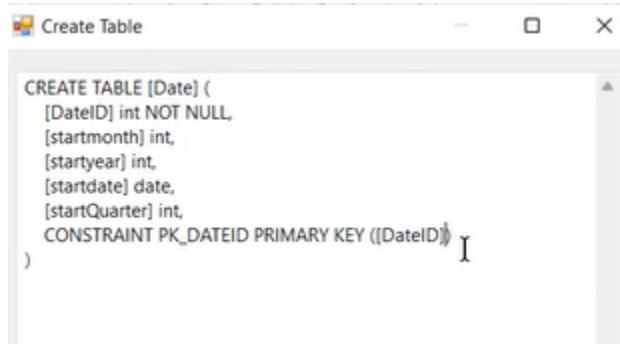
Hình 44. Thực hiện nhập ***server name*** và chọn ***database***

- **Bước 17:** Tại **OLE DB Destination Editor** > Tạo câu lệnh tạo mới bảng cho DIM_DATE > **New ...**



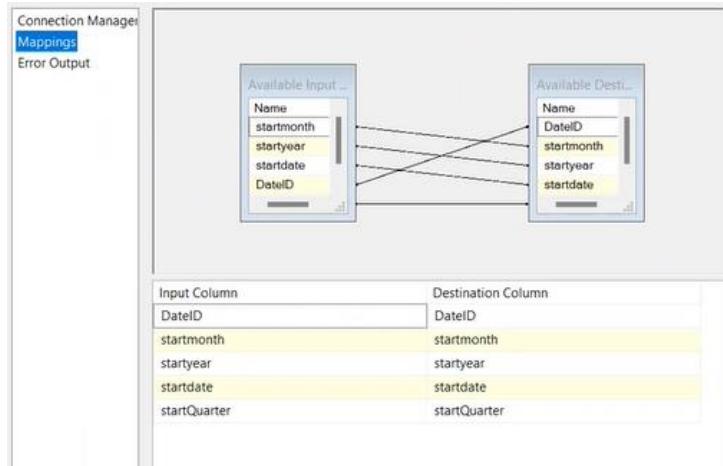
Hình 45. Thực hiện tạo bảng Dim_Date

- **Bước 18:** Tại màn hình câu lệnh, chỉnh sửa ràng buộc thuộc tính khóa chính, thứ tự trong bảng,...



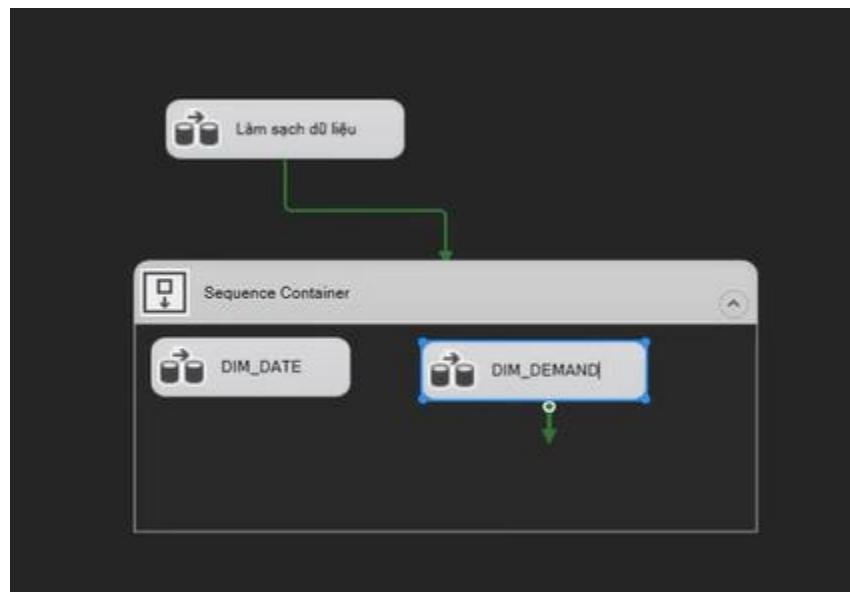
Hình 46. Câu lệnh tạo bảng Date

- **Bước 19:** Kiểm tra mappings > *OK*



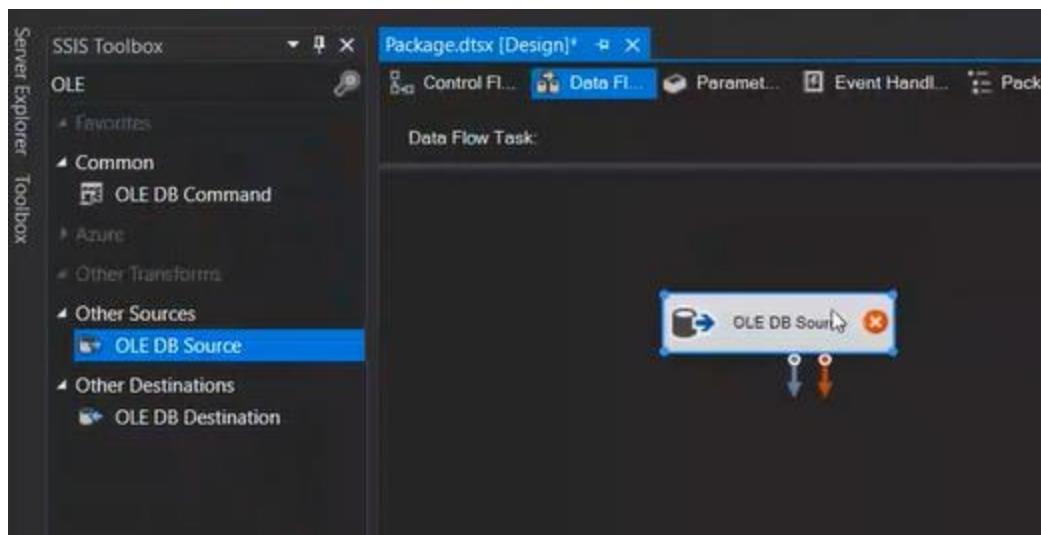
Hình 47. Kiểm tra mapping giữa các thuộc tính bảng Date

- **Bước 20:** Tạo bảng ***Dim_DEMAND*** > Kéo thả **Data Flow Task** vào **Sequence Container** > Đổi tên bảng



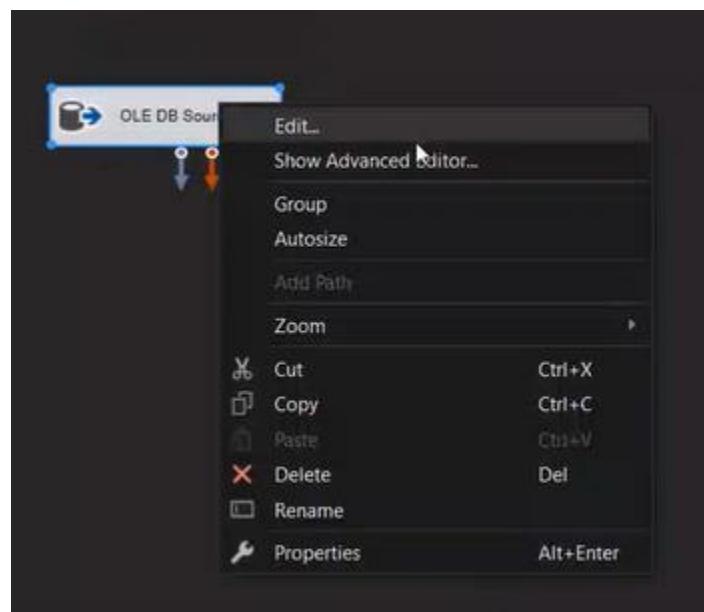
Hình 48. Thao tác với chức năng Data Flow Task

- **Bước 21:** Nhấn đúp vào **Data Flow Task** của ***DIM_DEMAND*** > Kéo thả **OLE DB Source** vào Data Flow Task



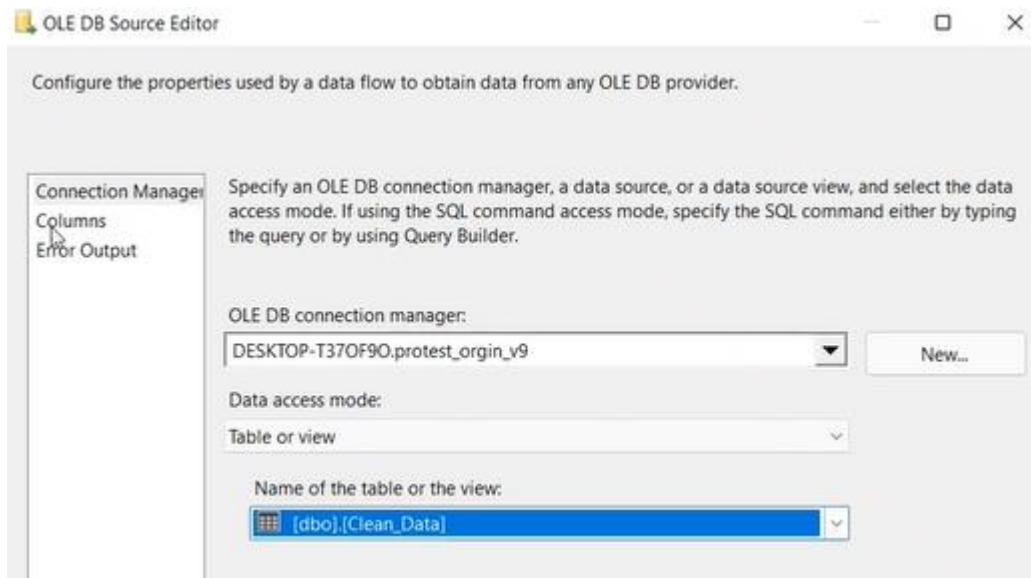
Hình 49. Thao tác với chức năng OLE DB Source

- **Bước 22:** tiến hành import dữ liệu đã nhập từ *SQL Server*
 - Chuột phải > *Edit*



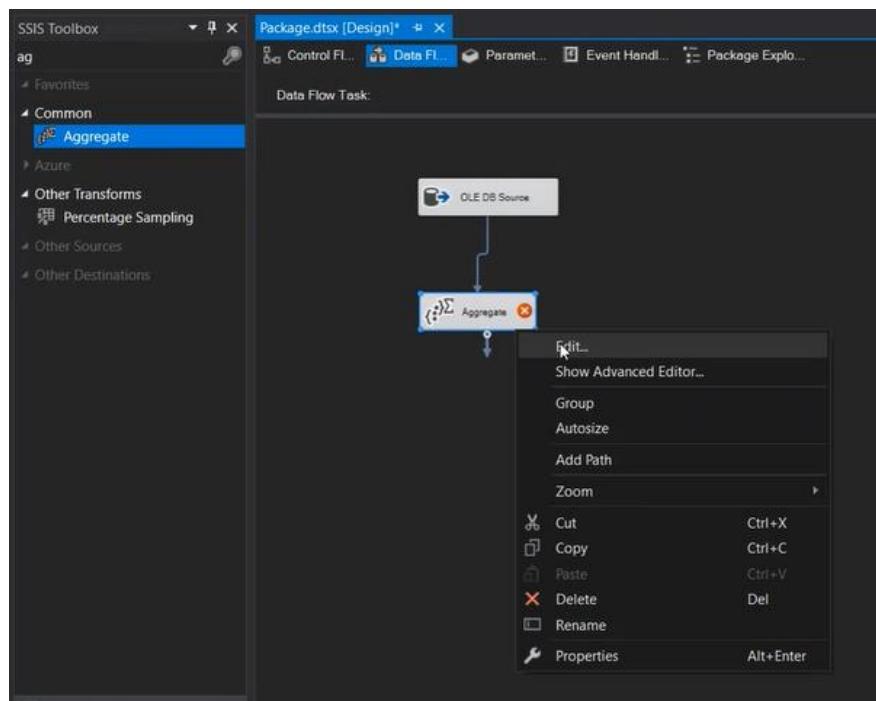
Hình 50. Thực hiện chỉnh sửa OLE DB Source

- **Bước 23:** Chọn bảng [*Clean_Data*] vừa tạo ở quá trình làm sạch dữ liệu > ***OK***



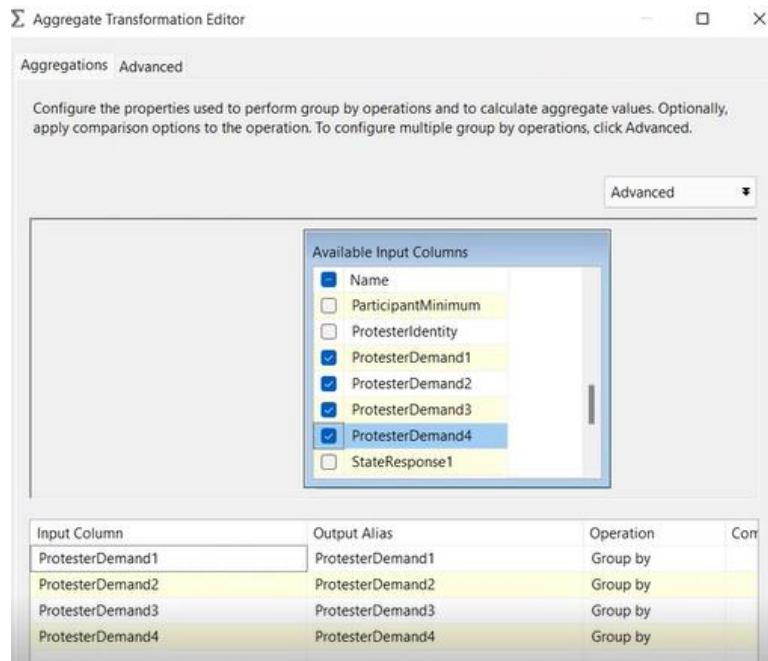
Hình 51. Thực hiện chọn bảng Clean_Data làm dữ liệu nguồn

- **Bước 24:** Kéo thả chức năng *Aggregate* > Chuột phải chọn *Edit*



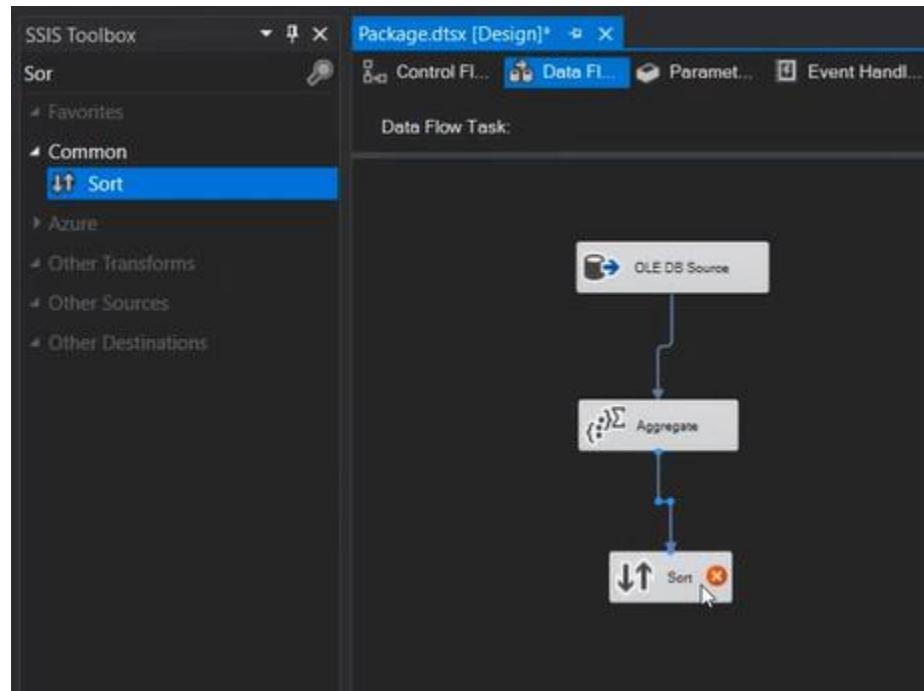
Hình 52. Thao tác với chức năng Aggregate

- **Bước 25:** Tại màn hình *Aggregate Transformation Editor* > Tick chọn những thuộc tính có trong bảng DIM_DEMAND (dựa trên lược đồ hình sao) > Operation = ‘Group by’



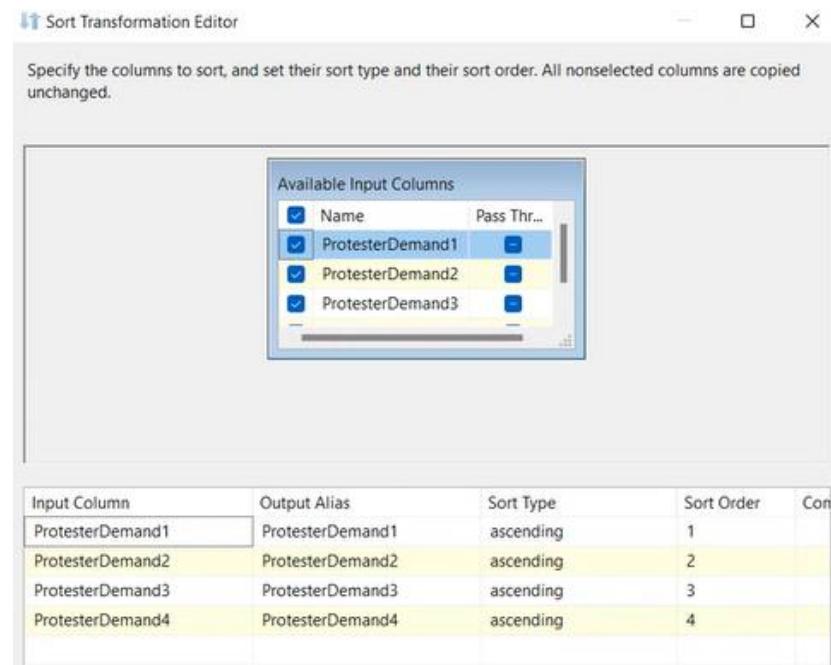
Hình 53. Thực hiện điều chỉnh thuộc tính bảng Aggregate

- **Bước 26:** Kéo thả chức năng Sort để sắp xếp dữ liệu thuộc tính > *Edit*



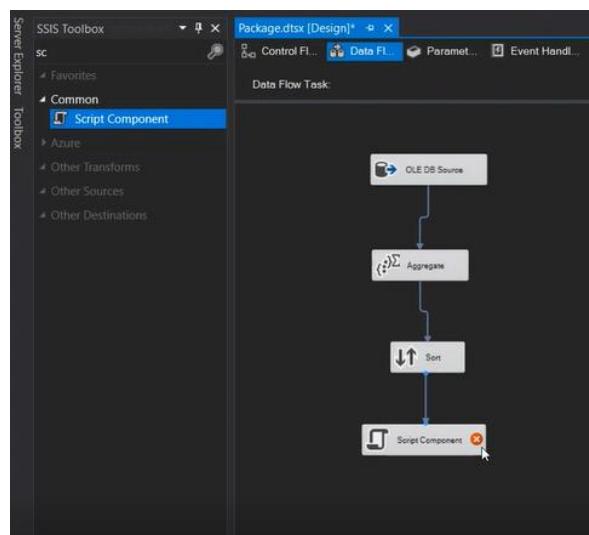
Hình 54. Thao tác với ô Sort

- **Bước 27:** Chọn những thuộc tính, kiểu sắp xếp



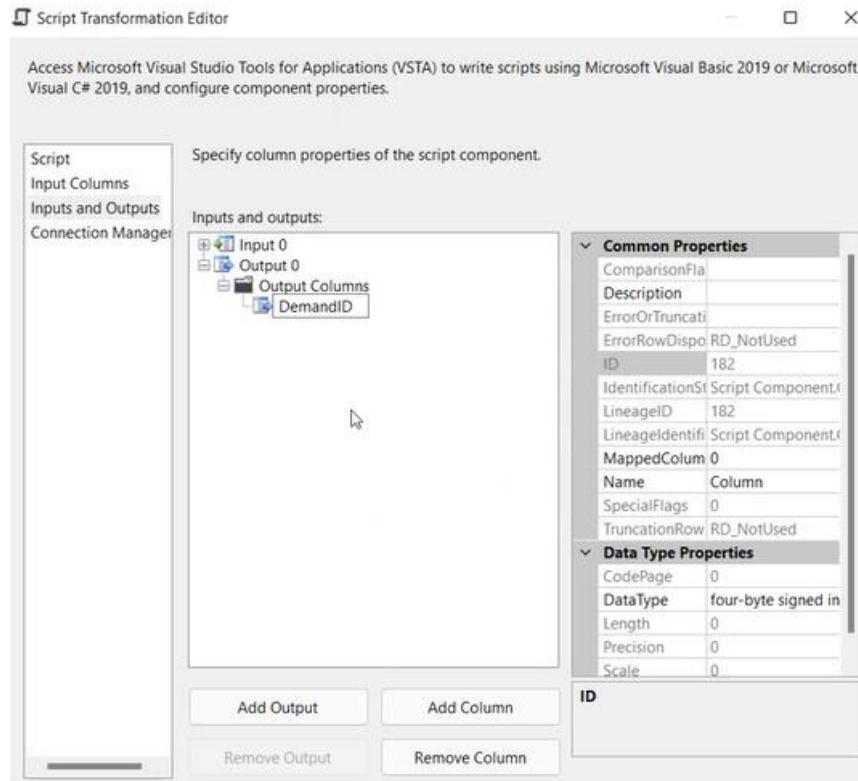
Hình 55. Thao tác sắp xếp dữ liệu

- **Bước 28:** Kéo thả chức năng *Script Component* để thực hiện tạo thêm một biến tăng tự động > Kéo mũi tên từ *Sort* vào *Script Component* > Chuột phải *Edit*



Hình 56. Thao tác với chức năng Script Component

- **Bước 29:** Tại *Script Transformation Editor* > chọn mục *Inputs and Outputs* > *Add Column* tại Output Column > Tạo thuộc tính mới tên DemandID



Hình 57. Thực hiện thêm thuộc tính Demand_ID

- **Bước 30:** Tại mục *Script* > Chọn *Edit Script* > Hệ thống sẽ mở giao diện code của Visual Studio > Tạo biến đếm count = 0 > cho chạy tăng tự động.

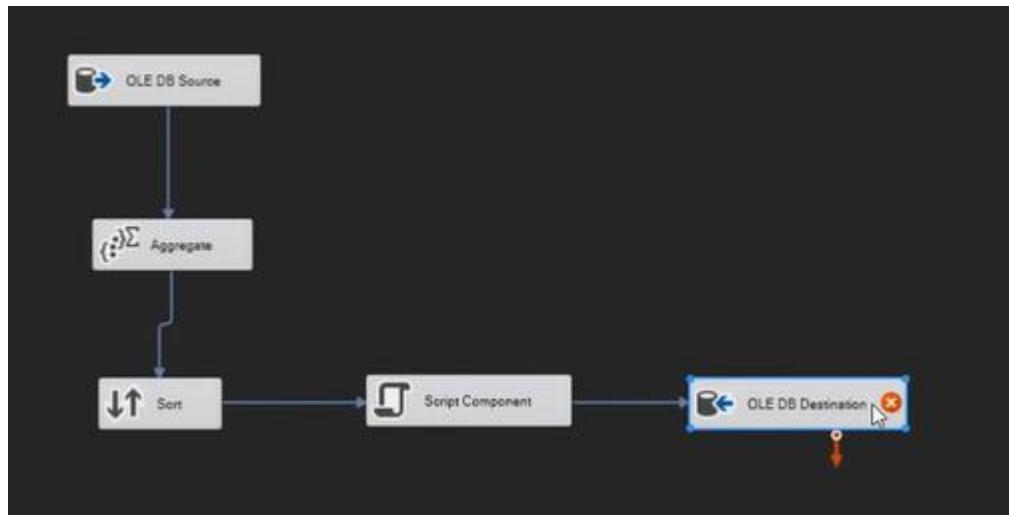
```
1  [Help: Introduction to the Script Component]
2
3  [Namespaces]
4
5  [/summary]
6  [This is the class to which to add your code. Do not change the name, attributes, or parent
7  [of this class.]
8  [/summary]
9  [Microsoft.SqlServer.Dts.Pipeline.SSIScriptComponentEntryPointAttribute]
10 [/references]
11 public class ScriptMain : UserComponent
12 {
13     [int count = 1;
14     [Help: Using Integration Services variables and parameters]
15
16     [Help: Using Integration Services Connection Managers]
17
18     [Help: Firing Integration Services Events]
19
20     [/summary]
21     [This method is called once, before rows begin to be processed in the data flow.
22     [/]
23     [You can remove this method if you don't need to do anything here.]
24     [/summary]
25     [/references]
26     public override void PreExecute()
27     {
28         [base.PreExecute();
29         [/*
30         [ * Add your code here
31         [ */
32     }
33
34     [/summary]
35     [This method is called after all the rows have passed through this component.
36     [/]
37     [You can delete this method if you don't need to do anything here.]
38     [/summary]
39 }
```

Hình 58.1 Thực hiện thêm thuộc tính ID tăng tự động

```
2 references
public override void Input0_ProcessInputRow(Input0Buffer Row)
{
    Row.DemandID = count;
    count++;
}
```

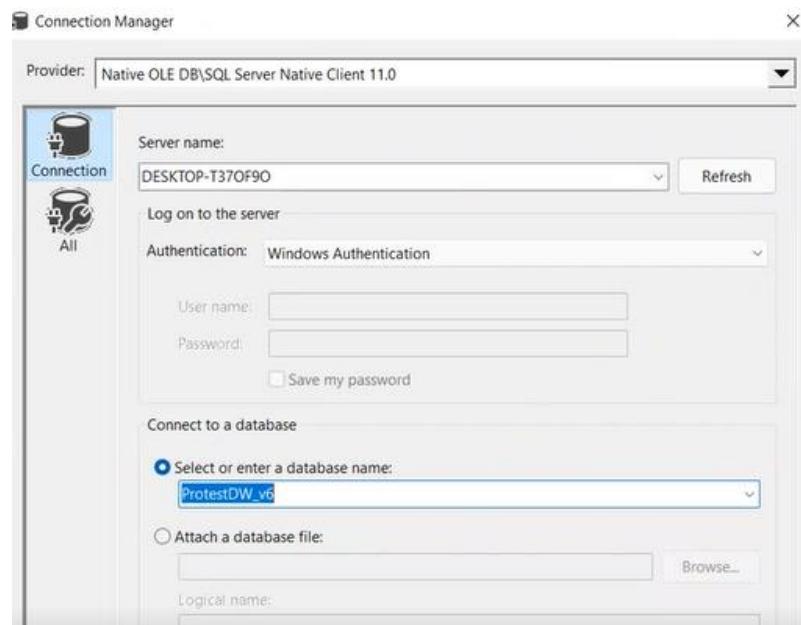
Hình 58.2 Thực hiện thêm thuộc tính ID tăng tự động

- **Bước 31:** Kéo thả chức năng ***OLE DB Destination*** để truyền dữ liệu bảng ***DIM_DEMAND*** vào kho dữ liệu đích > ***Edit***



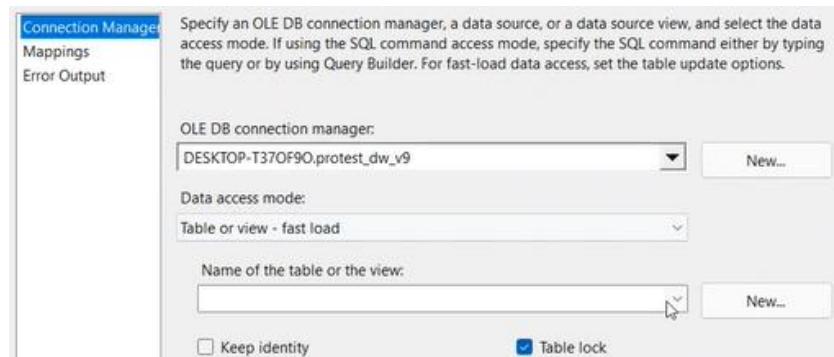
Hình 59. Thao tác với chức năng *OLE DB Destination*

- **Bước 32:** Tại ***Connection Manager***, chọn ***server name*** của SQL Server và kho dữ liệu đích là ProtestDW



Hình 60. Thực hiện nhập tên server và chọn database

- **Bước 33:** Tại *OLE DB Destination Editor* > Tạo câu lệnh tạo mới bảng cho DIM_DEMAND > *New ...*



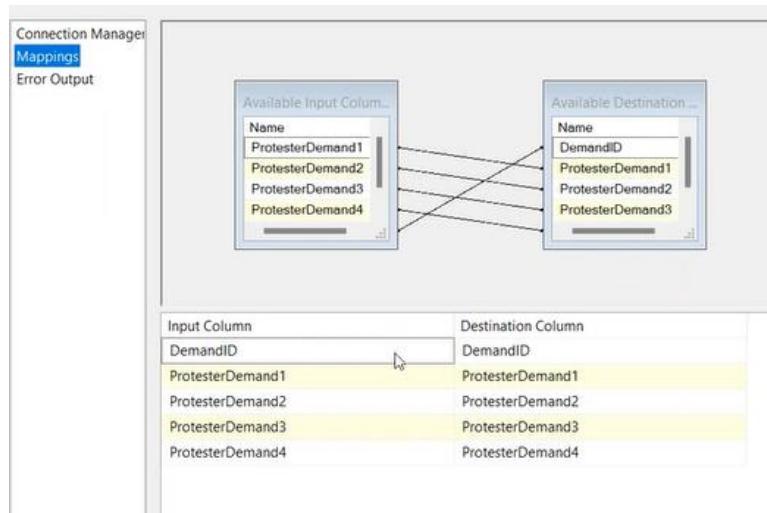
Hình 61. Thực hiện tạo bảng Demand ở kho dữ liệu đích

- **Bước 34:** Tại màn hình câu lệnh, chỉnh sửa ràng buộc thuộc tính khóa chính, thứ tự trong bảng,...



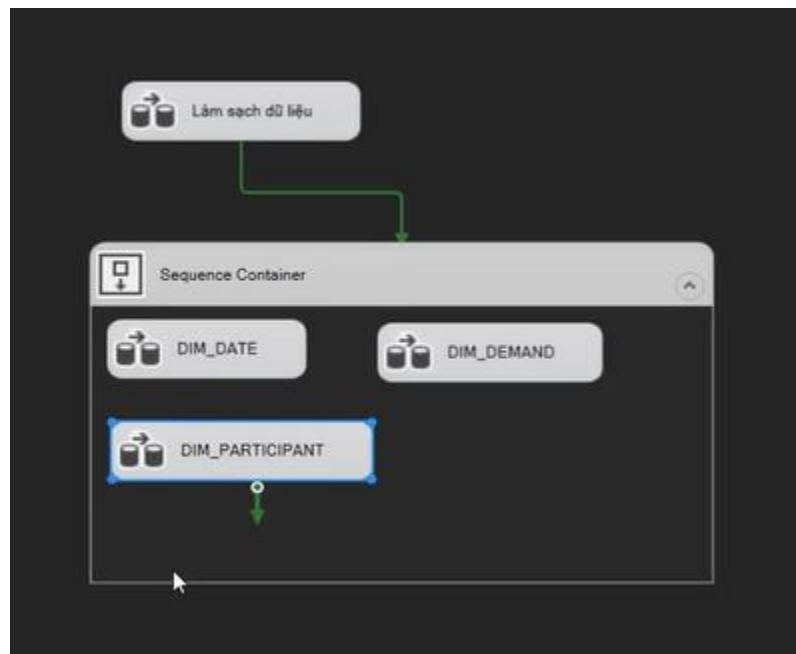
Hình 62. Câu lệnh tạo bảng Demand

- **Bước 35:** Kiểm tra mappings > ***OK***



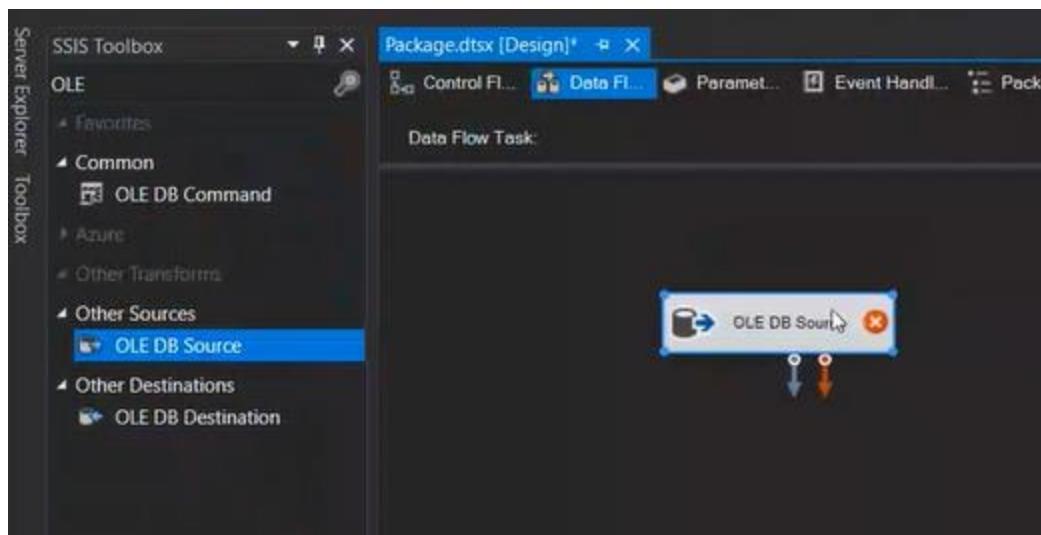
Hình 63. Kiểm tra mapping của các thuộc tính

- **Bước 36:** Tạo bảng ***Dim_PARTICIPANT*** > Kéo thả **Data Flow Task** vào ***Sequence Container*** > Đổi tên bảng



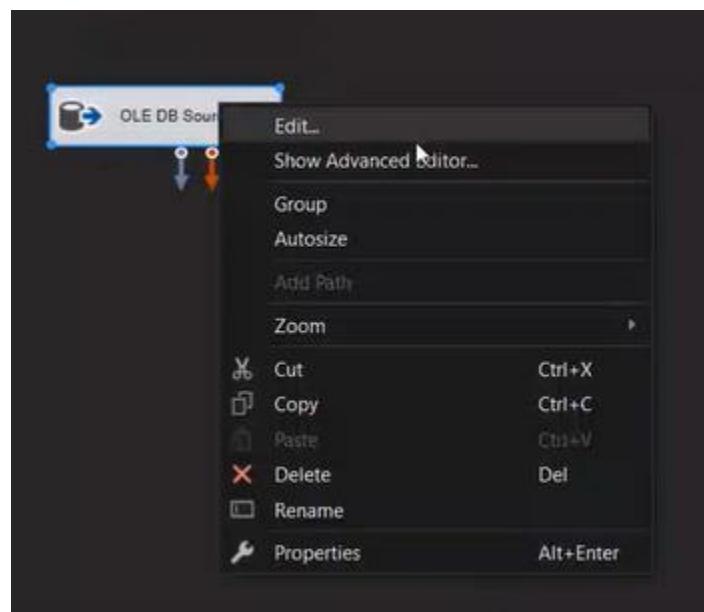
Hình 64. Thao tác với chức năng Data Flow Task

- **Bước 37:** Nhấn đúp vào **Data Flow Task** của ***DIM_PARTICIPANT*** > Kéo thả ***OLE DB Source*** vào Data Flow Task



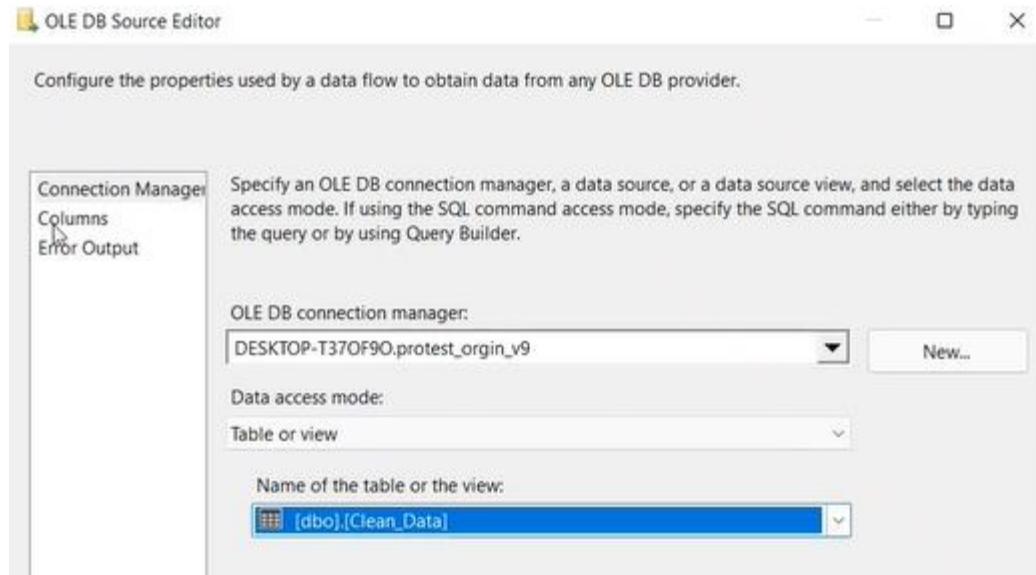
Hình 65. Thao tác với OLE DB Source cho bảng Dim_Participant

- **Bước 38:** tiến hành import dữ liệu đã nhập từ *SQL Server*
 - Chuột phải > *Edit*



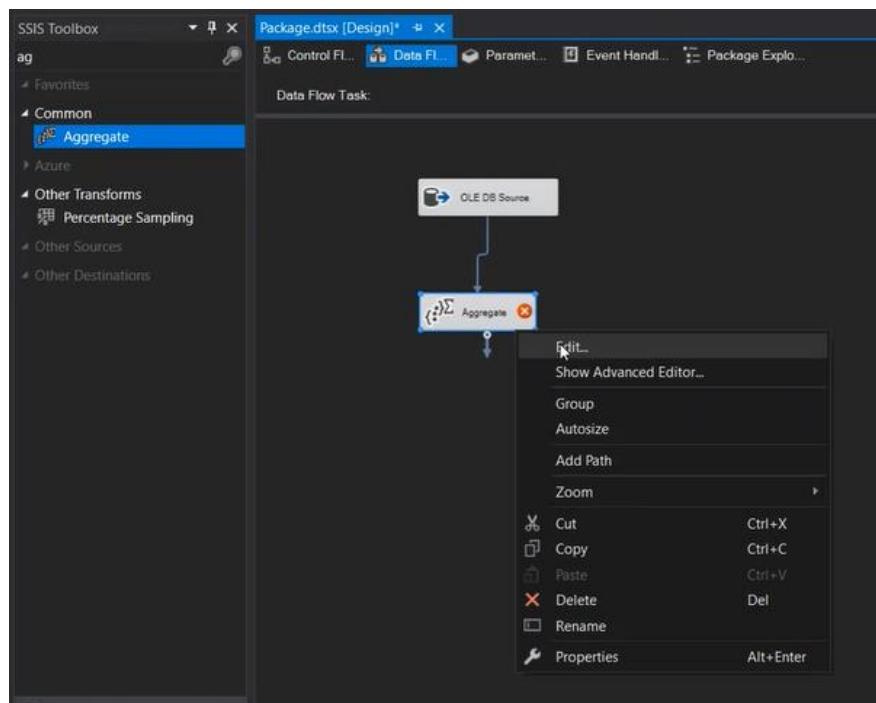
Hình 66. Thực hiện chỉnh sửa OLE DB Source của Dim_Participant

- **Bước 39:** Chọn bảng [*Clean_Data*] vừa tạo ở quá trình làm sạch dữ liệu > *OK*



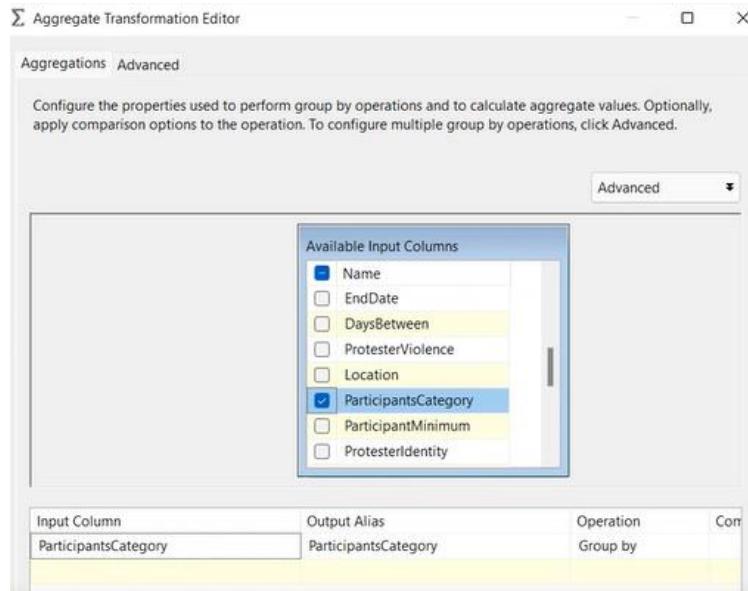
Hình 67. Thực hiện chọn bảng Clean_Data làm dữ liệu nguồn

- **Bước 40:** Kéo thả chức năng **Aggregate** > Chuột phải chọn **Edit**



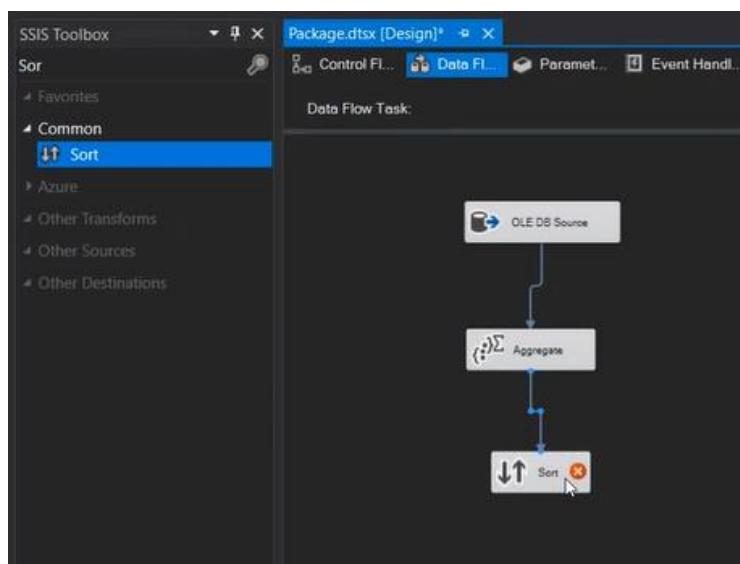
Hình 68. Thực hiện chỉnh sửa ô Aggregate

- **Bước 41:** Tại màn hình *Aggregate Transformation Editor* > Tick chọn những thuộc tính có trong bảng DIM_PARTICIPANT (dựa trên lược đồ hình sao) > Operation = ‘Group by’



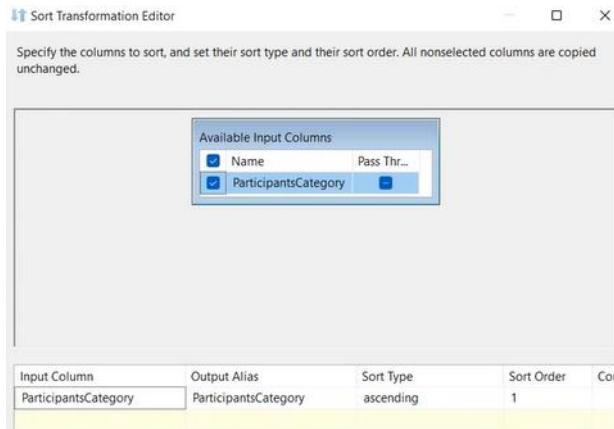
Hình 69. Thực hiện chỉnh sửa thuộc tính bằng Aggregate Editor

- **Bước 42:** Kéo thả chức năng Sort để sắp xếp dữ liệu > *Edit*



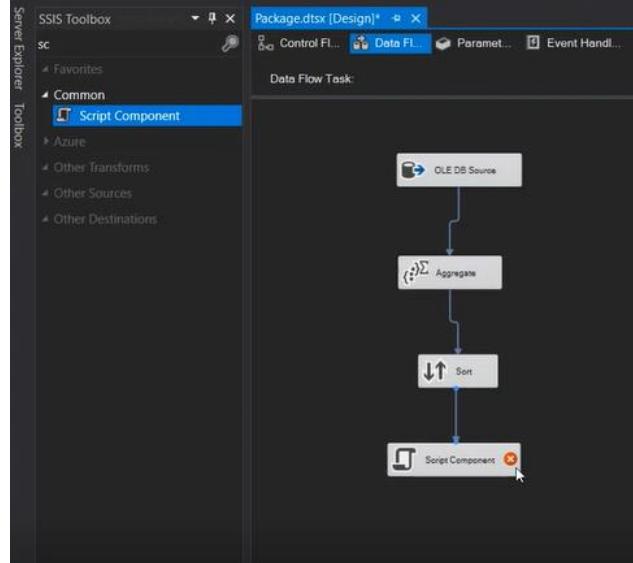
Hình 70. Thao tác với ô Sort

- **Bước 43:** Chọn những thuộc tính, kiểu sắp xếp



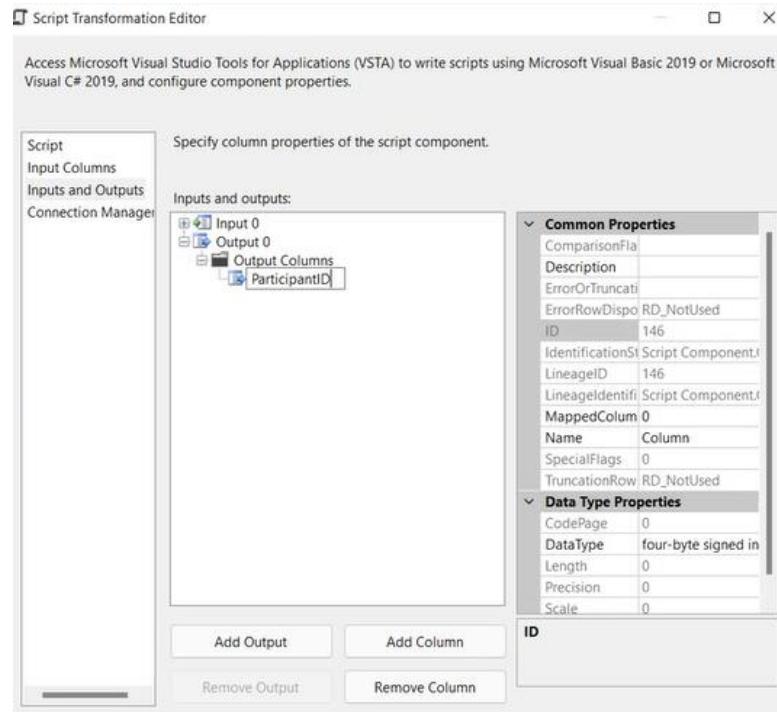
Hình 71. Sắp xếp dữ liệu thuộc tính ParticipantCategory

- **Bước 44:** Kéo thả chức năng **Script Component** để thực hiện tạo thêm một biến tăng tự động > Kéo mũi tên từ **Sort** vào **Script Component** > Chuột phải **Edit**



Hình 72. Thao tác với chức năng Script Component

- **Bước 45:** Tại *Script Transformation Editor* > chọn mục *Inputs and Outputs* > *Add Column* tại Output Column > Tạo thuộc tính mới tên ParticipantID



Hình 73. Thực hiện tạo thuộc tính ParticipantID

- **Bước 46:** Tại mục *Script* > Chọn *Edit Script* > Hệ thống sẽ mở giao diện code của Visual Studio > Tạo biến đếm count = 0 > cho chạy tăng tự động.

```

1  [Help: Introduction to the Script Component]
2
3  [Namespaces]
4
5  [/summary]
6  [/ This is the class to which to add your code. Do not change the name, attributes, or parent
7  [/ of this class.]
8  [/summary]
9  [Microsoft.SqlServer.Dts.Pipeline.SSIScriptComponentEntryPointAttribute]
10 [references]
11 [public class ScriptMain : UserComponent
12 {
13     [int count = 1;
14     [Help: Using Integration Services variables and parameters]
15     [Help: Using Integration Services Connection Managers]
16     [Help: Firing Integration Services Events]
17
18     [/summary]
19     [/ This method is called once, before rows begin to be processed in the data flow.
20     [/ You can remove this method if you don't need to do anything here.]
21     [/summary]
22     [public override void PreExecute()
23     {
24         [base.PreExecute();
25         [/*
26         [ * Add your code here
27         [ */
28     }
29
30     [/summary]
31     [/ This method is called after all the rows have passed through this component.
32     [/ You can delete this method if you don't need to do anything here.]
33     [/summary]
34 }

```

Hình 74.1 Thực hiện cho thuộc tính *ParticipantID* tăng tự động

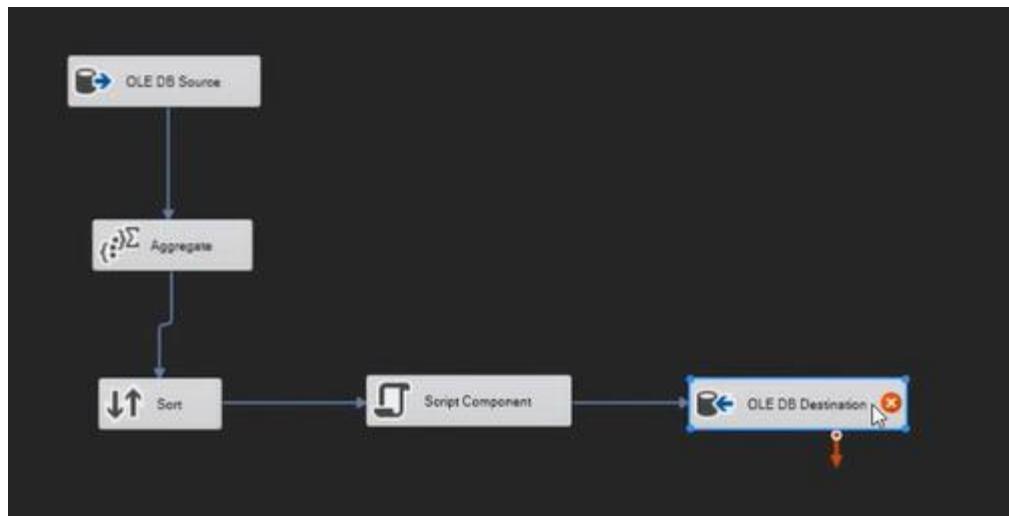
```

public override void Input0_ProcessInputRow(Input0Buffer Row)
{
    Row.ParticipantID = count;
    count++;
}

```

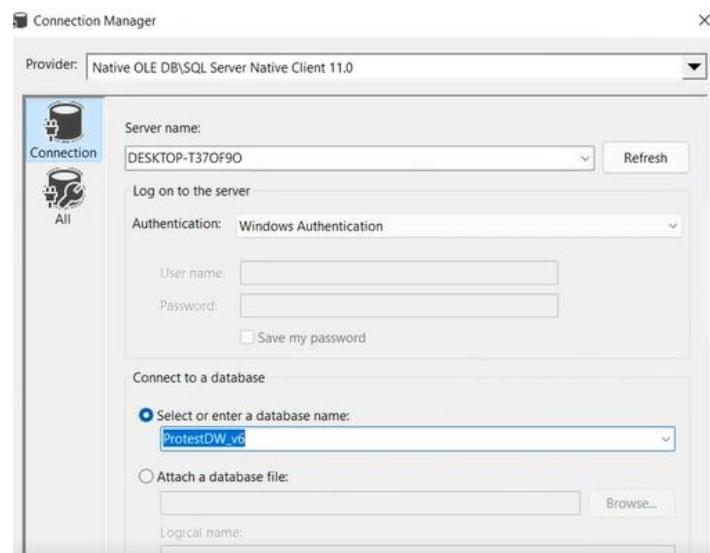
Hình 74.2 Thực hiện cho thuộc tính *ParticipantID* tăng tự động

- **Bước 47:** Kéo thả chức năng ***OLE DB Destination*** để truyền dữ liệu bảng ***DIM_PARTICIPANT*** vào kho dữ liệu đích > ***Edit***



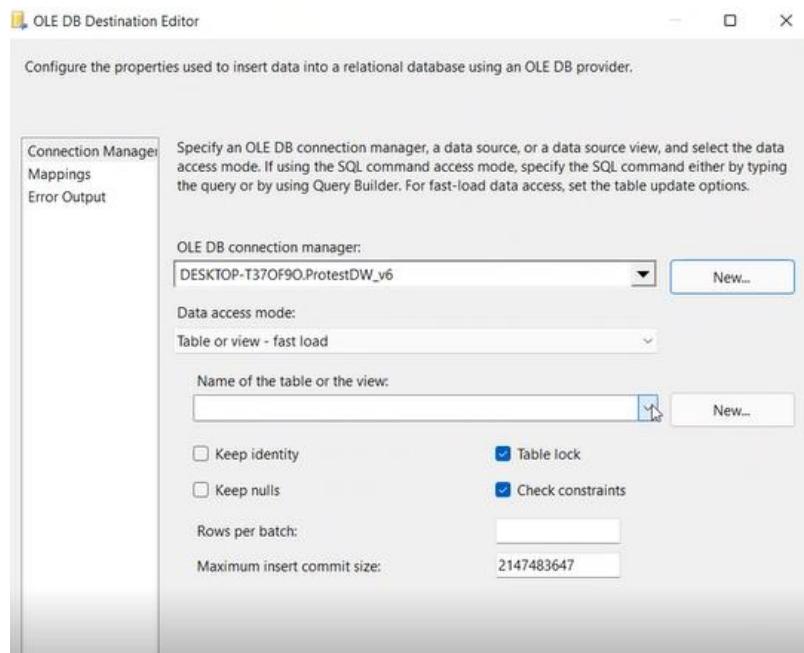
Hình 75. Thao tác với chức năng OLE DB Source cho Dim _ Participant

- Bước 48:** Tại **Connection Manager**, chọn **server name** của SQL Server và kho dữ liệu đích là ProtestDW



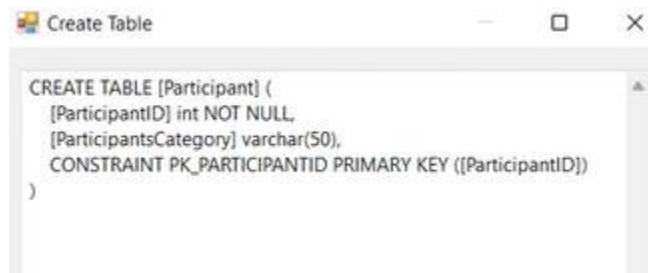
Hình 76. Thực hiện nhập server name và chọn data warehouse

- Bước 49:** Tại **OLE DB Destination Editor** > Tạo câu lệnh tạo mới bảng cho DIM_PARTICIPANT > **New ...**



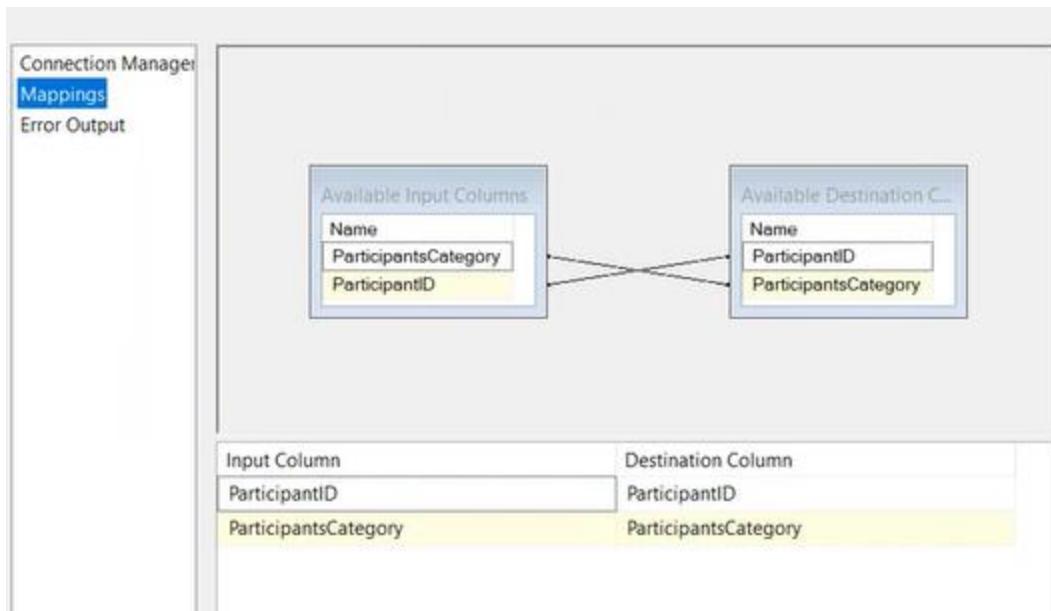
Hình 77. Thực hiện tạo bảng Participant tại data warehouse

- **Bước 50:** Tại màn hình câu lệnh, chỉnh sửa ràng buộc thuộc tính khóa chính, thứ tự trong bảng,...



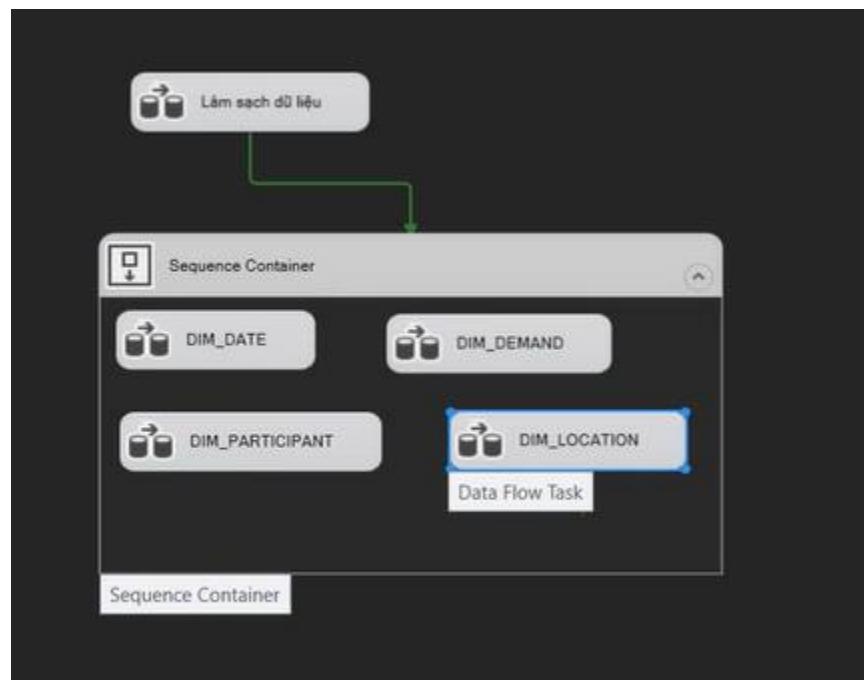
Hình 78. Câu lệnh tạo bảng Participant

- **Bước 51:** Kiểm tra mappings > **OK**



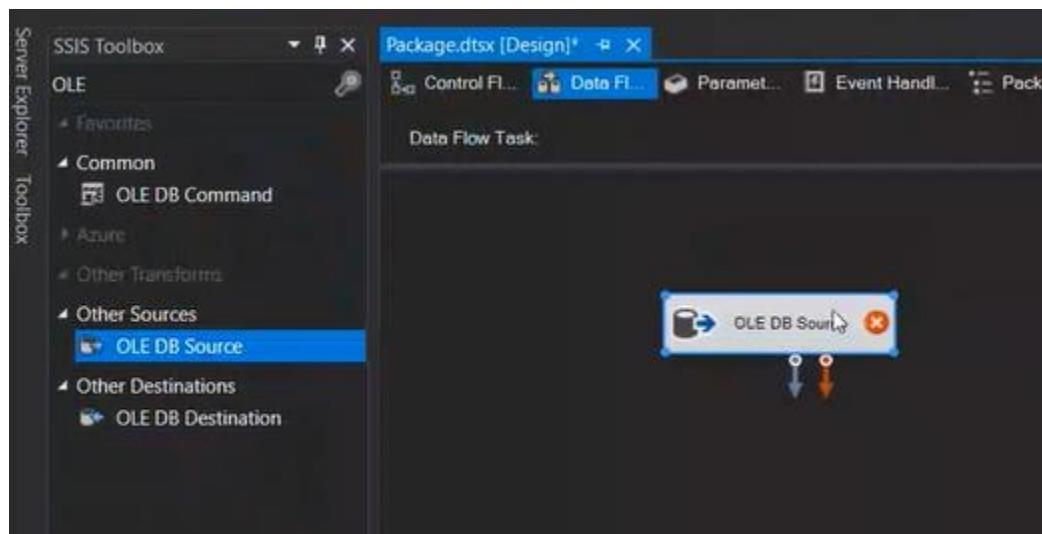
Hình 79. Kiểm tra Mapping của các thuộc tính

- Bước 52:** Tạo bảng **Dim_LOCATION** > Kéo thả **Data Flow Task** vào **Sequence Container** > Đổi thành tên bảng



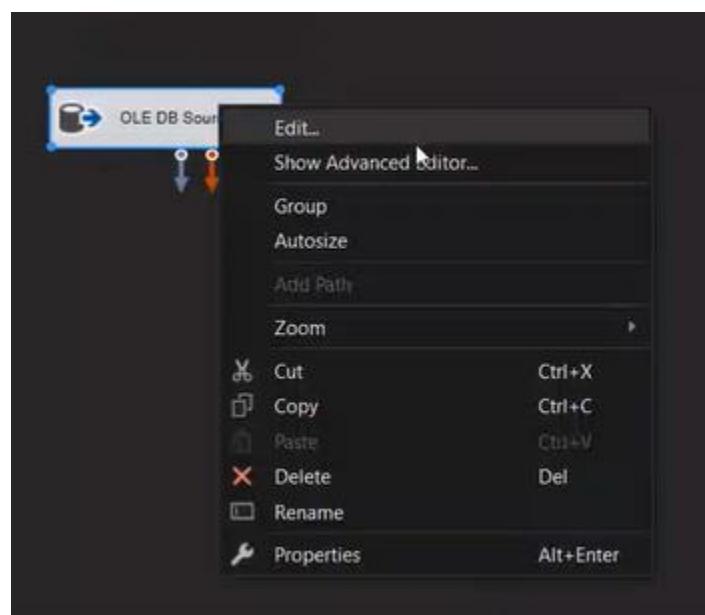
Hình 80. Thao tác với chức năng Data Flow Task

- **Bước 53:** Nhấn đúp vào *Data Flow Task* của *DIM_LOCATION* > Kéo thả *OLE DB Source* vào Data Flow Task



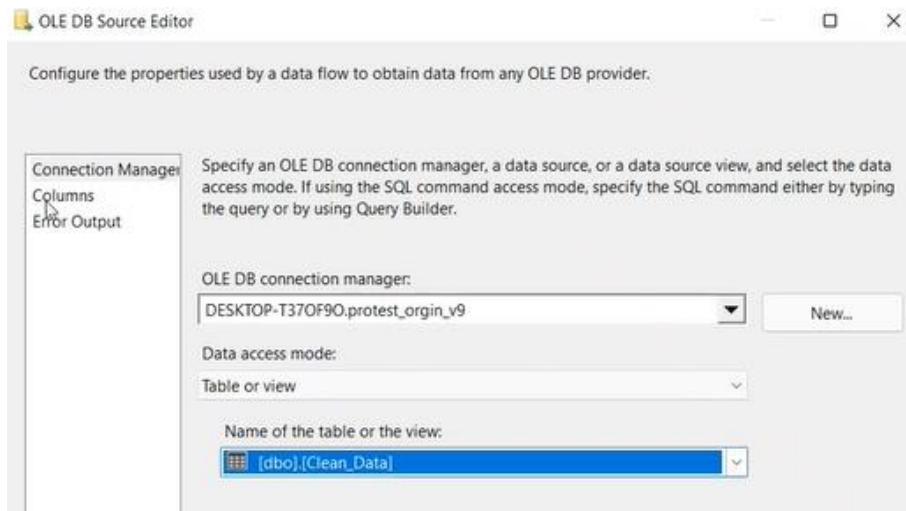
Hình 81. Thao tác với chức năng OLE DB Source bằng Location

- **Bước 54:** tiến hành import dữ liệu đã nhập từ *SQL Server*
 - Chuột phải > *Edit*



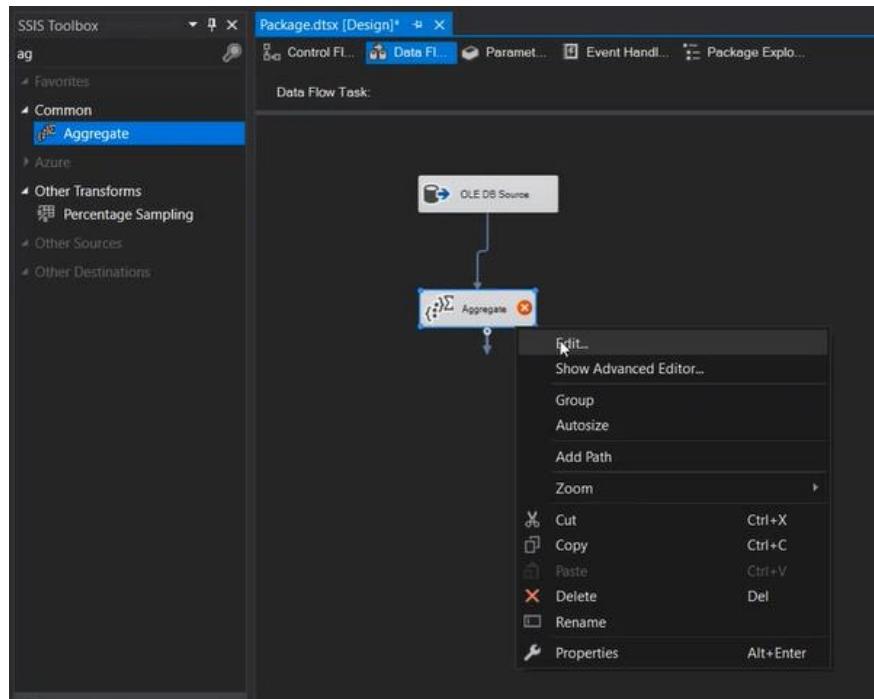
Hình 82. Thực hiện chỉnh sửa OLE DB Source

- **Bước 55:** Chọn bảng [*Clean_Data*] vừa tạo ở quá trình làm sạch dữ liệu > ***OK***



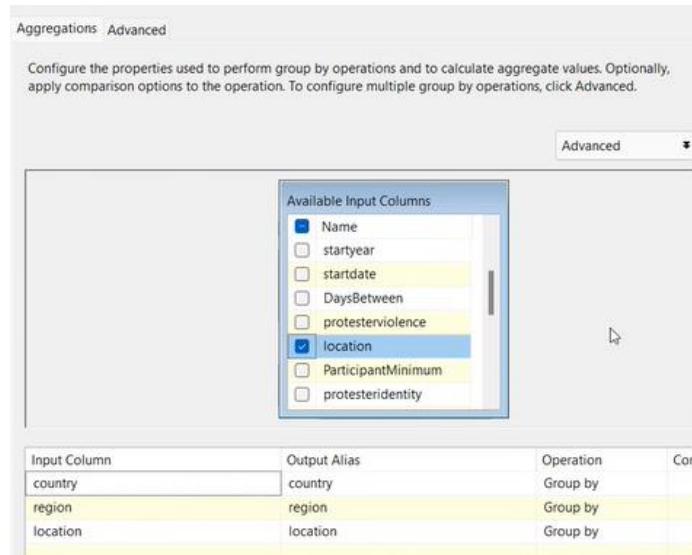
Hình 83. Thực hiện chọn bảng *Clean_Data* làm dữ liệu nguồn

- **Bước 56:** Kéo thả chức năng *Aggregate* > Chuột phải chọn **Edit**



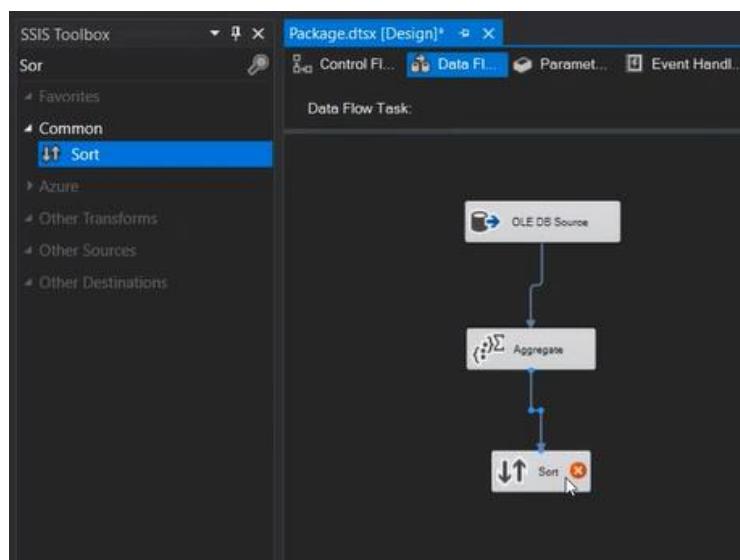
Hình 84. Thực hiện chỉnh sửa ô Aggregate

- **Bước 57:** Tại màn hình *Aggregate Transformation Editor* > Tick chọn những thuộc tính có trong bảng DIM_LOCATION (dựa trên lược đồ hình sao) > Operation = ‘Group by’



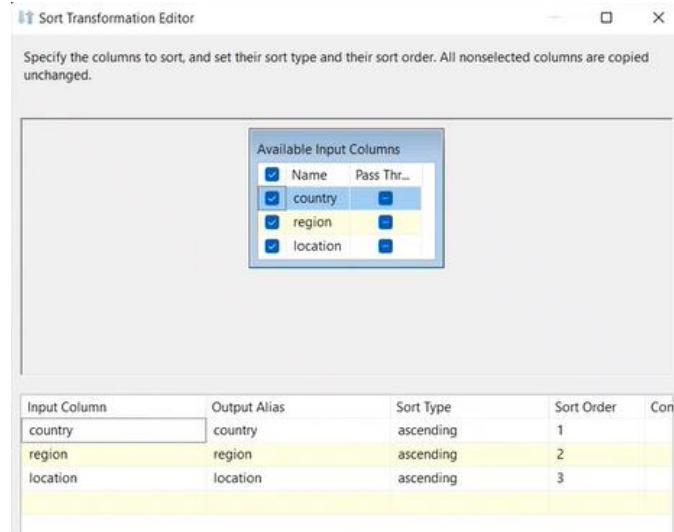
Hình 85. Thực hiện chỉnh sửa thuộc tính trong bảng Location

- **Bước 58:** Kéo thả chức năng Sort để sắp xếp dữ liệu thuộc tính > *Edit*



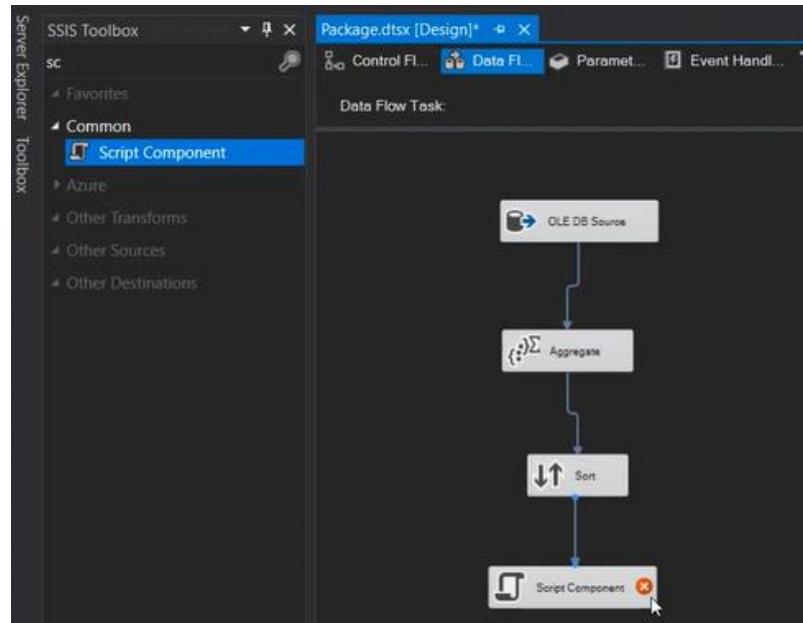
Hình 86. Thao tác với ô Sort

- **Bước 59:** Chọn những thuộc tính, kiểu sắp xếp



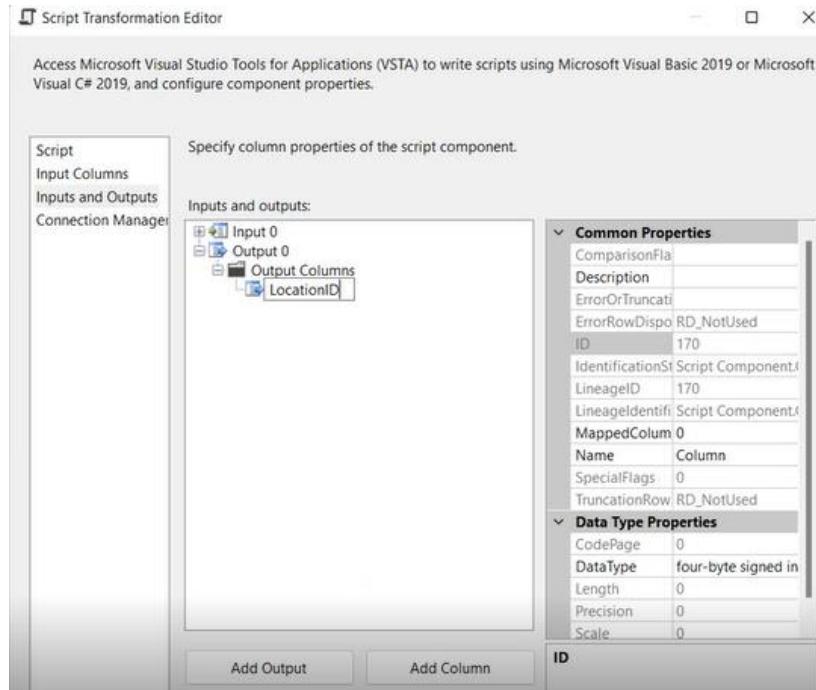
Hình 87. Thực hiện sắp xếp dữ liệu cho các thuộc tính Location

- **Bước 60:** Kéo thả chức năng *Script Component* để thực hiện tạo thêm một biến tăng tự động > Kéo mũi tên từ *Sort* vào *Script Component* > Chuột phải *Edit*



Hình 88. Thao tác với ô *Script Component*

- **Bước 61:** Tại *Script Transformation Editor* > chọn mục *Inputs and Outputs* > *Add Column* tại Output Column > Tạo thuộc tính mới tên LocationID



Hình 89. Thực hiện tạo thuộc tính LocationID

- **Bước 62:** Tại mục *Script* > Chọn *Edit Script* > Hệ thống sẽ mở giao diện code của Visual Studio > Tạo biến đếm count = 0 > cho chạy tăng tự động.

```

1  [Help: Introduction to the Script Component]
2
3  [Namespaces]
4
5  [/summary]
6  [/ This is the class to which to add your code. Do not change the name, attributes, or parent
7  [/ of this class.]
8  [/summary]
9  [Microsoft.SqlServer.Dts.Pipeline.SSIScriptComponentEntryPointAttribute]
10 [references]
11 [public class ScriptMain : UserComponent
12 {
13     [int count = 1;
14     [Help: Using Integration Services variables and parameters]
15     [Help: Using Integration Services Connection Managers]
16     [Help: Firing Integration Services Events]
17
18     [/summary]
19     [/ This method is called once, before rows begin to be processed in the data flow.
20     [/ You can remove this method if you don't need to do anything here.]
21     [/summary]
22     [public override void PreExecute()
23     {
24         [base.PreExecute();
25         [/*
26         [ * Add your code here
27         [ */
28     }
29
30     [/summary]
31     [/ This method is called after all the rows have passed through this component.
32     [/ You can delete this method if you don't need to do anything here.]
33     [/summary]
34 }

```

Hình 90.1 Thực hiện chỉnh cho thuộc tính LocationID tăng tự động

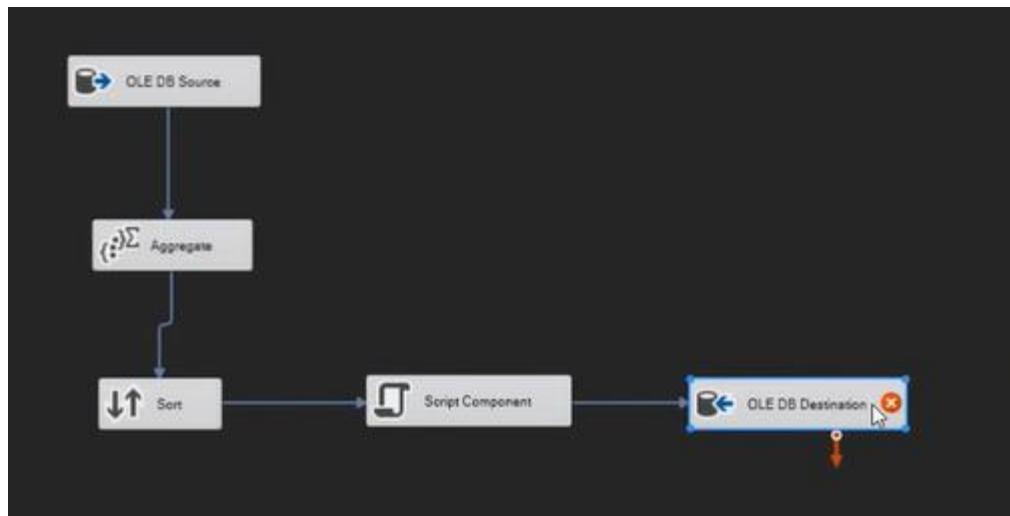
```

[references]
[public override void Input0_ProcessInputRow(Input0Buffer Row)
{
[    Row.LocationID = count;
[    count++;
}

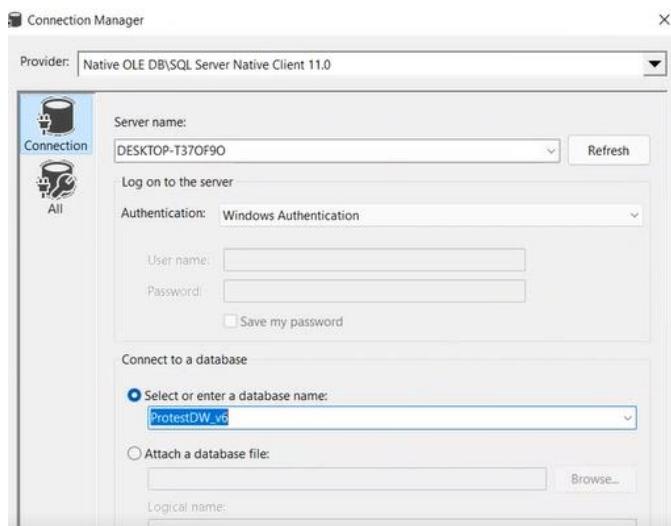
```

Hình 90.2 Thực hiện chỉnh cho thuộc tính LocationID tăng tự động

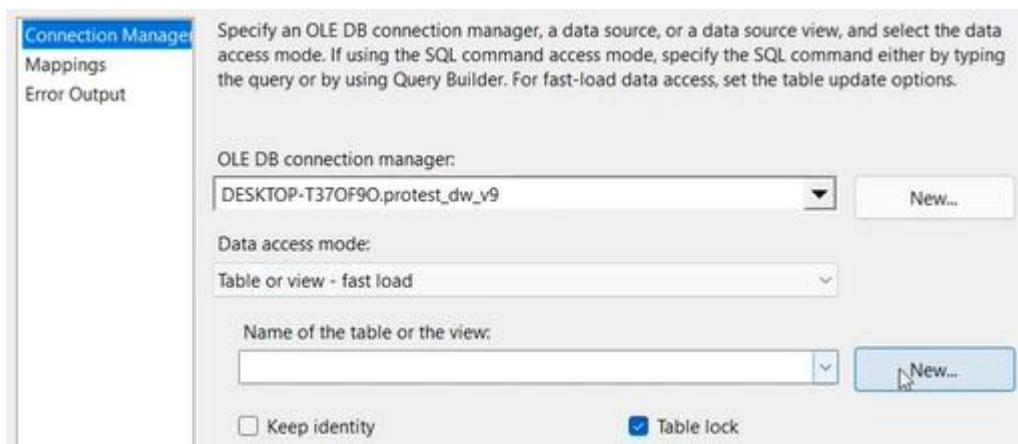
- Bước 63:** Kéo thả chức năng **OLE DB Destination** để truyền dữ liệu bảng **DIM_LOCATION** vào kho dữ liệu đích > **Edit**

*Hình 91. Thao tác với ô OLE DB Destination*

- **Bước 64:** Tại **Connection Manager**, chọn **server name** của SQL Server và kho dữ liệu đích là ProtestDW

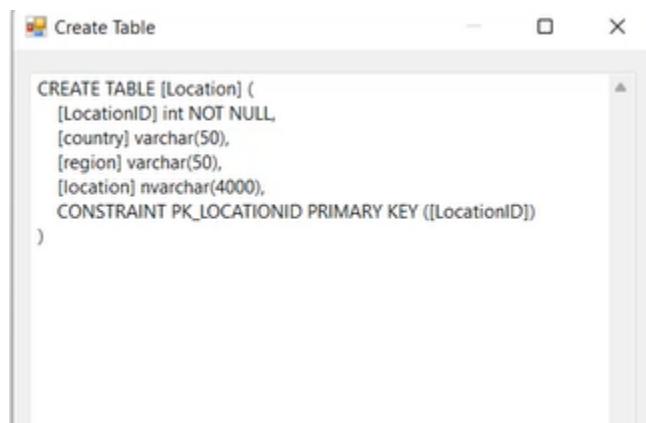
*Hình 92. Thực hiện nhập server name và data warehouse*

- **Bước 65:** Tại **OLE DB Destination Editor** > Tạo câu lệnh tạo mới bảng cho DIM_LOCATION > **New ...**



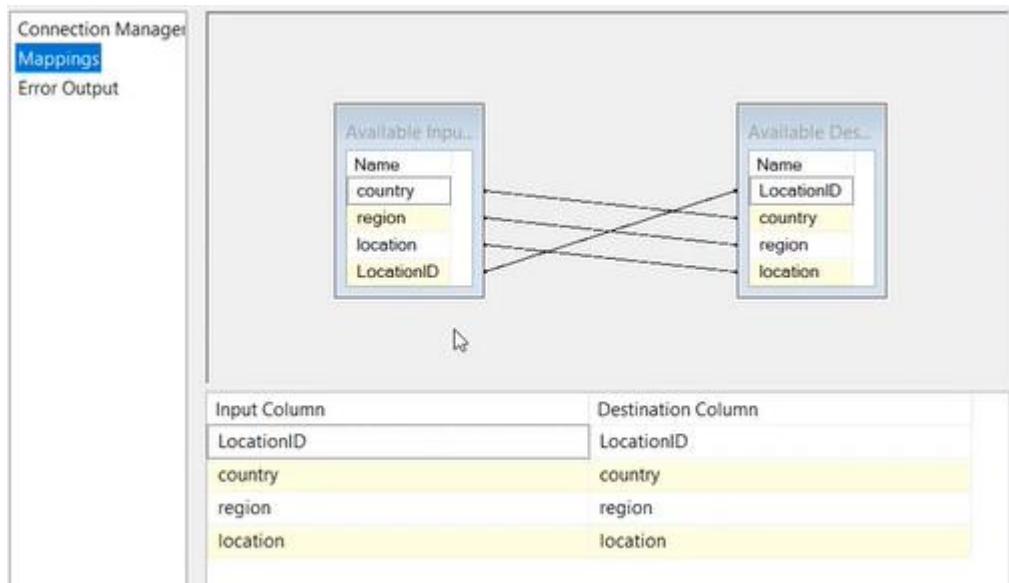
Hình 93. Thực hiện tạo bảng Location cho kho dữ liệu đích

- **Bước 66:** Tại màn hình câu lệnh, chỉnh sửa ràng buộc thuộc tính khóa chính, thứ tự trong bảng,...



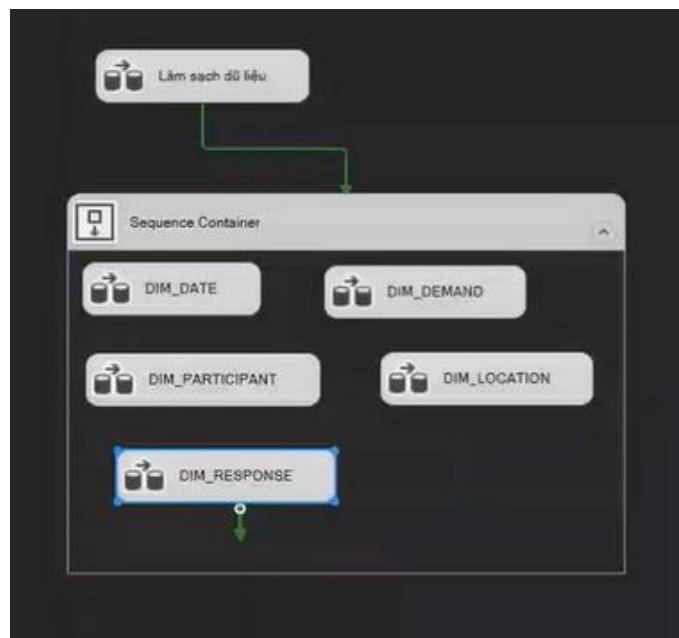
Hình 94. Thực hiện tạo bảng Location

- **Bước 67:** Kiểm tra mappings > ***OK***



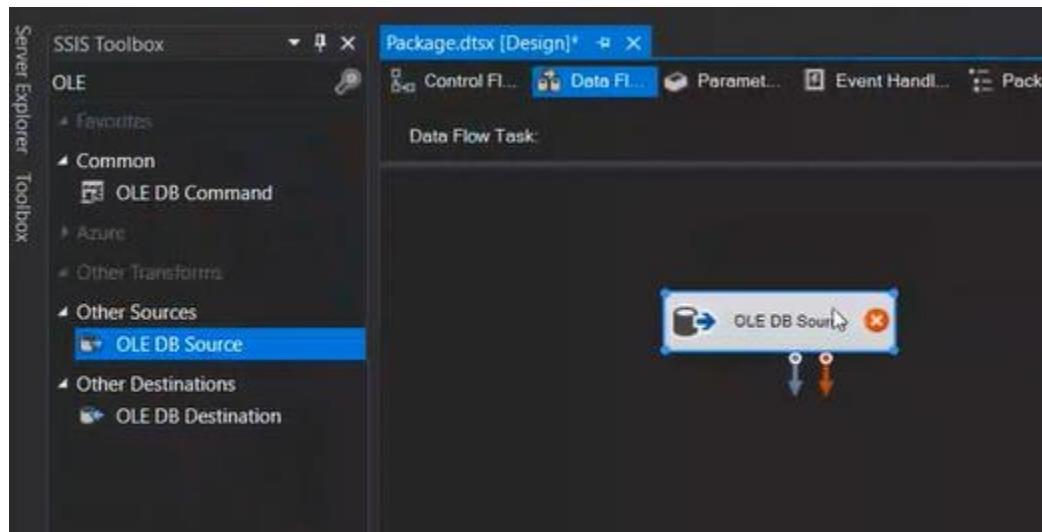
Hình 95. Kiểm tra Mapping các thuộc tính trong Location

- **Bước 68:** Tạo bảng **DIM_RESPONSE** > Kéo thả **Data Flow Task** vào **Sequence Container** > Đổi tên bảng



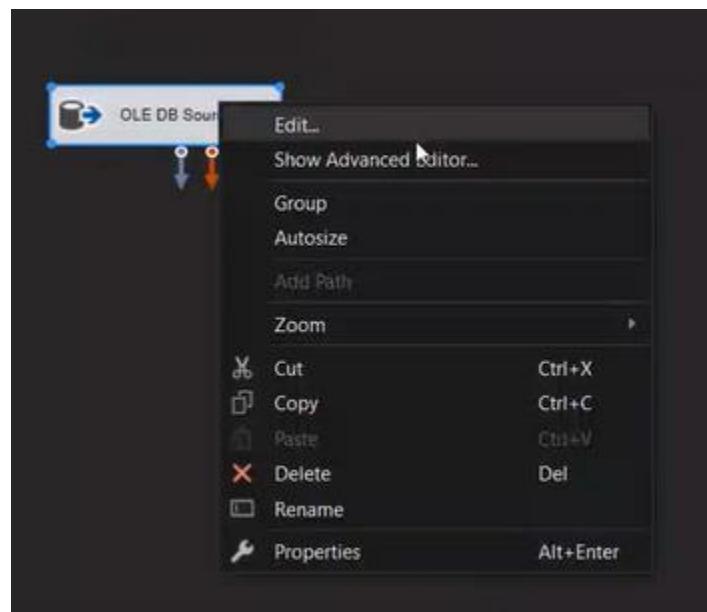
Hình 96. Thao tác với ô Data Flow Task

- **Bước 69:** Nhấn đúp vào *Data Flow Task* của *DIM_RESPONSE* > Kéo thả *OLE DB Source* vào Data Flow Task



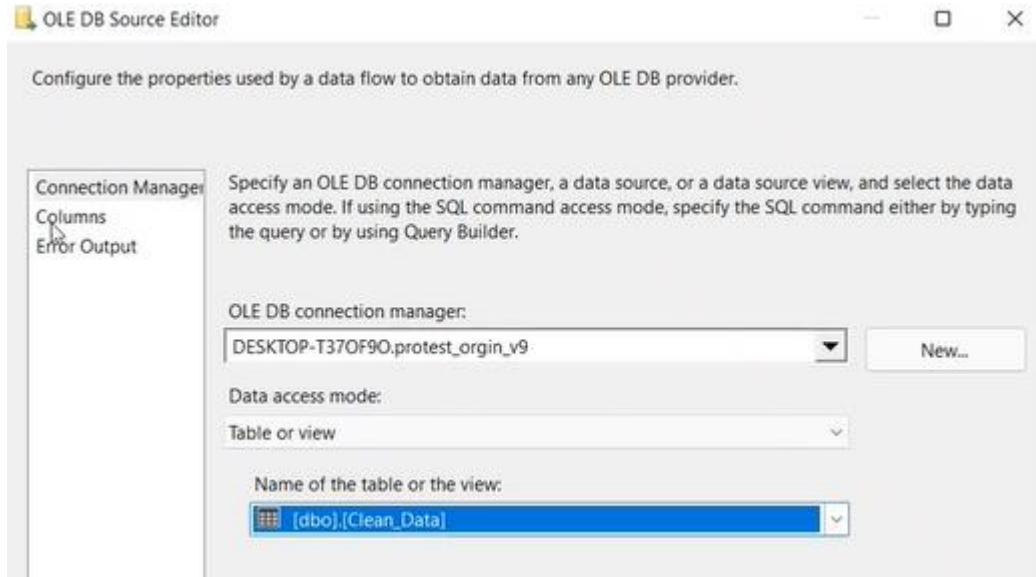
Hình 97. Thao tác với ô *OLE DB Source*

- **Bước 70:** tiến hành import dữ liệu đã nhập từ *SQL Server*
 - Chuột phải > *Edit*



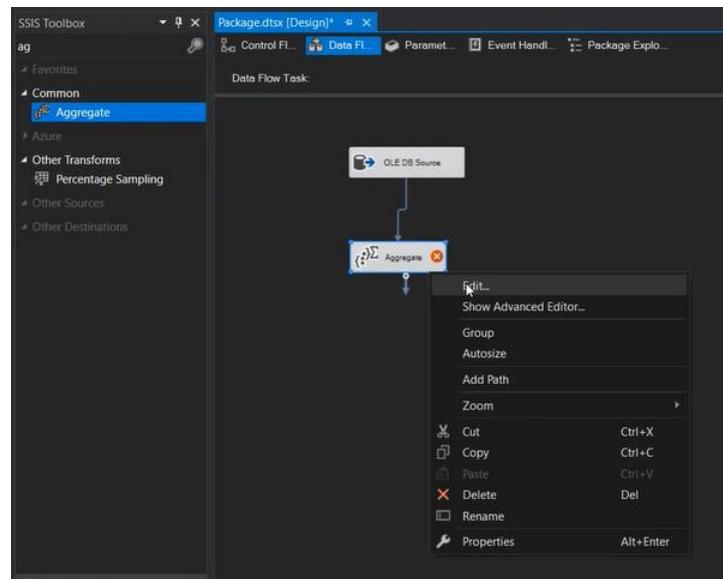
Hình 98. Thực hiện chỉnh sửa *OLE DB Source*

- **Bước 71:** Chọn bảng **[Clean_Data]** vừa tạo ở quá trình làm sạch dữ liệu > **OK**



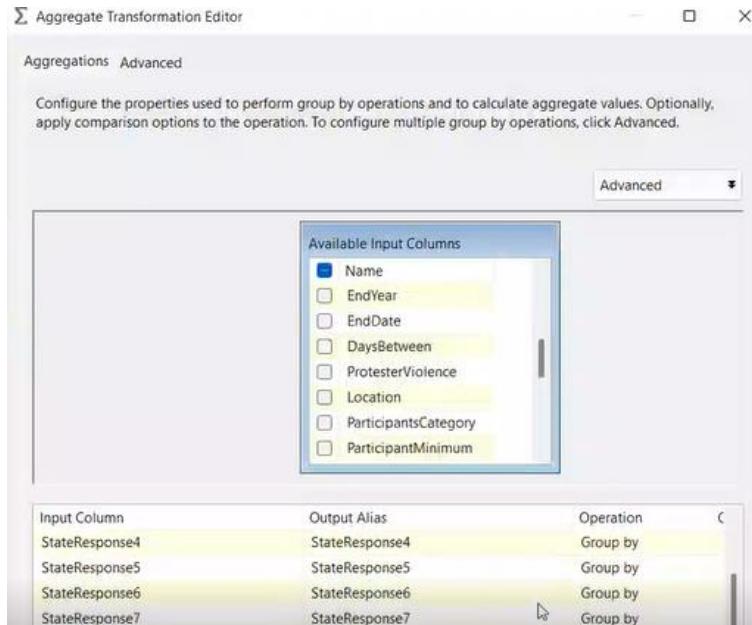
Hình 99. Thực hiện chọn bảng Clean_Data làm dữ liệu nguồn

- **Bước 72:** Kéo thả chức năng **Aggregate** > Chuột phải chọn **Edit**



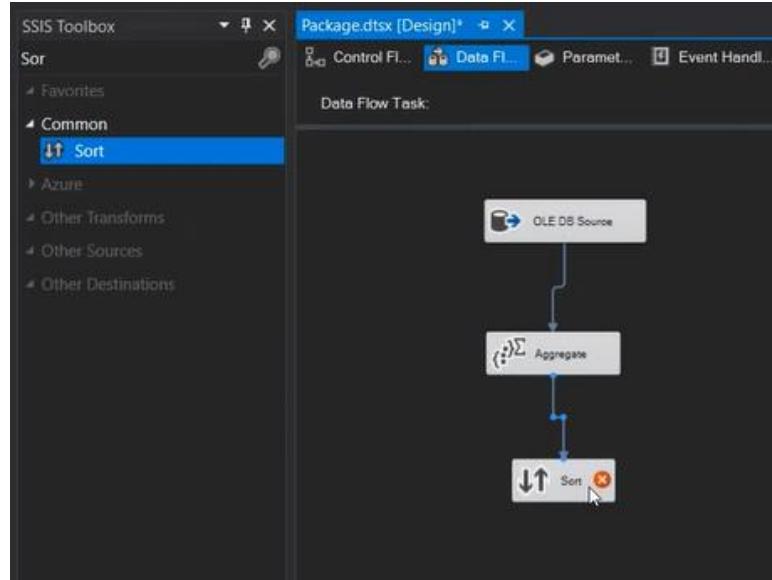
Hình 100. Thực hiện chỉnh sửa ô Aggregate

- **Bước 73:** Tại màn hình *Aggregate Transformation Editor* > Tick chọn những thuộc tính có trong bảng DIM_RESPONSE (dựa trên lược đồ hình sao) > Operation = ‘Group by’



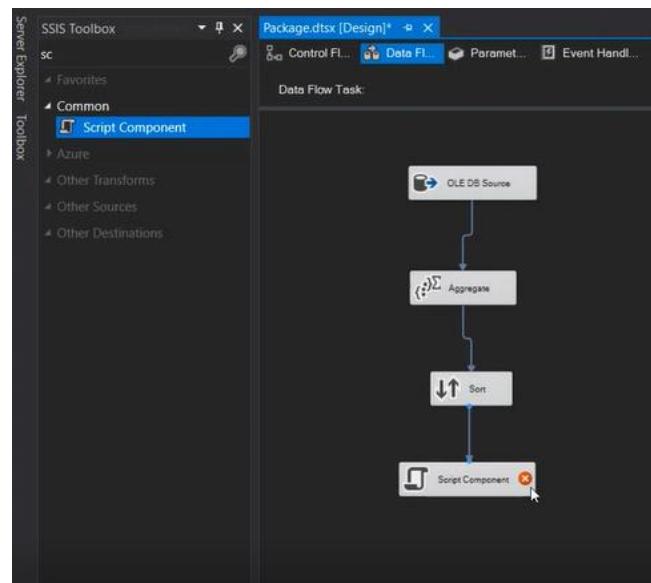
Hình 101. Thực hiện chỉnh sửa thuộc tính tại Aggregate Editor

- **Bước 74:** Kéo thả chức năng Sort để sắp xếp dữ liệu thuộc tính > *Edit*



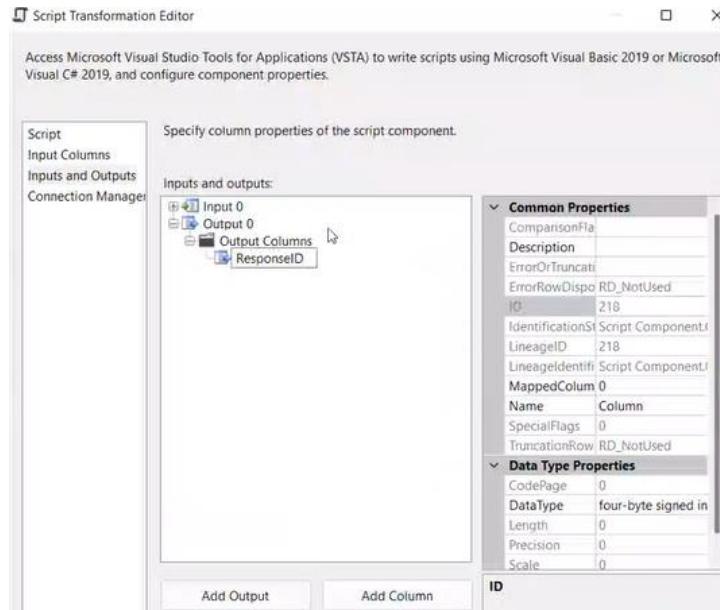
Hình 102. Thao tác với ô Sort

- **Bước 75:** Chọn những thuộc tính, kiểu sắp xếp
- **Bước 76:** Kéo thả chức năng *Script Component* để thực hiện tạo thêm một biến tăng tự động > Kéo mũi tên từ *Sort* vào *Script Component* > Chuột phải *Edit*



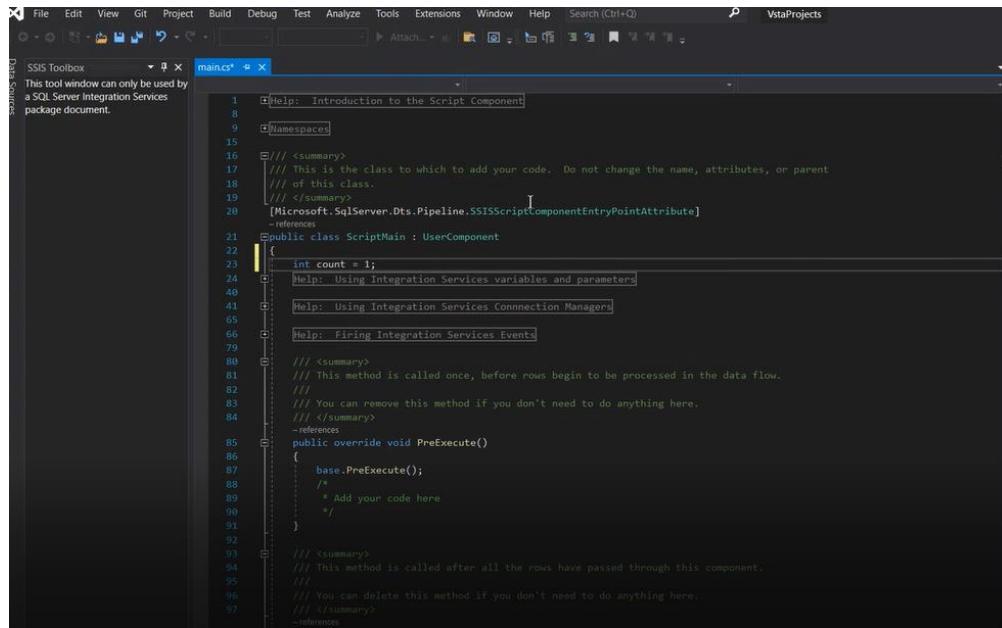
Hình 103. Thao tác với ô Script Component

- **Bước 77:** Tại *Script Transformation Editor* > chọn mục *Inputs and Outputs* > *Add Column* tại Output Column > Tạo thuộc tính mới tên ResponseID



Hình 104. Thực hiện tạo thuộc tính ResponseID

- **Bước 78:** Tại mục *Script* > Chọn *Edit Script* > Hệ thống sẽ mở giao diện code của Visual Studio > Tạo biến đếm count = 0 > cho chạy tăng tự động.



```

1  [Help: Introduction to the Script Component]
2
3  [Namespaces]
4
5  [/summary]
6  [This is the class to which to add your code. Do not change the name, attributes, or parent
7  [of this class.]
8  [/summary]
9  [Microsoft.SqlServer.Dts.Pipeline.SSIScriptComponentEntryPointAttribute]
10 [references]
11 [public class ScriptMain : UserComponent
12 {
13     int count = 1;
14     [Help: Using Integration Services variables and parameters]
15     [Help: Using Integration Services Connection Managers]
16     [Help: Firing Integration Services Events]
17
18     [summary]
19     [This method is called once, before rows begin to be processed in the data flow.
20     [You can remove this method if you don't need to do anything here.]
21     [/summary]
22     [/references]
23     public override void PreExecute()
24     {
25         base.PreExecute();
26         /*
27             * Add your code here
28         */
29     }
30
31     [summary]
32     [This method is called after all the rows have passed through this component.
33     [You can delete this method if you don't need to do anything here.]
34     [/summary]
35 }

```

Hình 105.1 Thực hiện chỉnh cho thuộc tính ResponseID tăng tự động

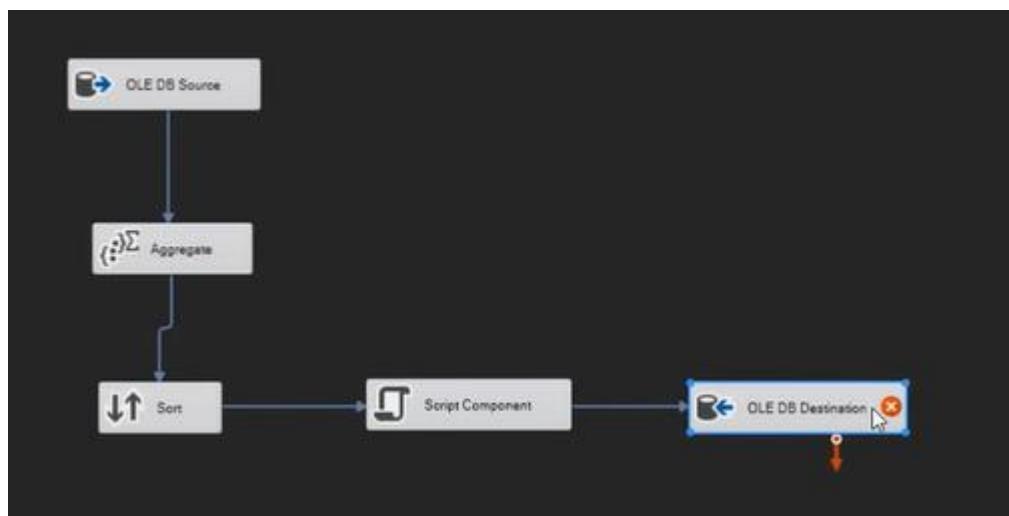
```

public override void Input0_ProcessInputRow(Input0Buffer Row)
{
    Row.ResponseID = count;
    count++;
}

```

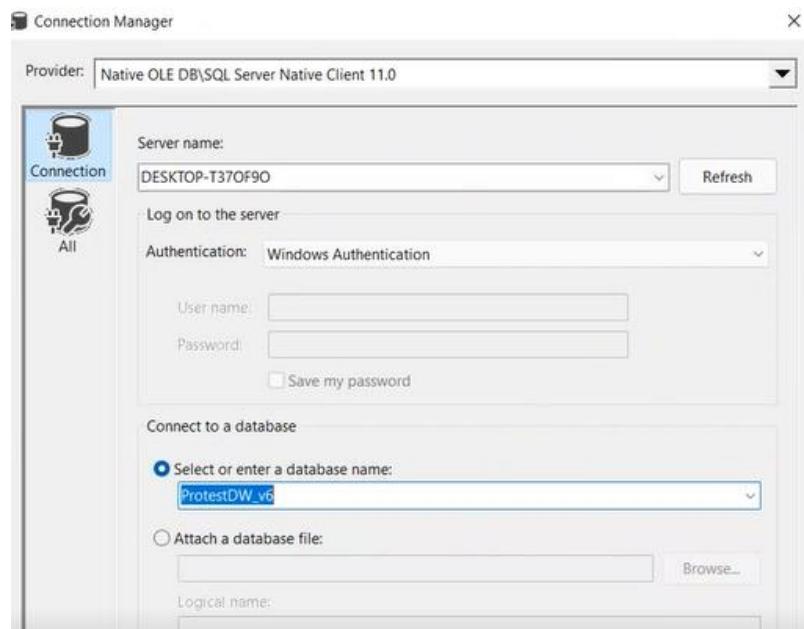
Hình 105.2 Thực hiện chỉnh cho thuộc tính ResponseID tăng tự động

- Bước 79:** Kéo thả chức năng **OLE DB Destination** để truyền dữ liệu bảng **DIM_RESPONSE** vào kho dữ liệu đích > **Edit**

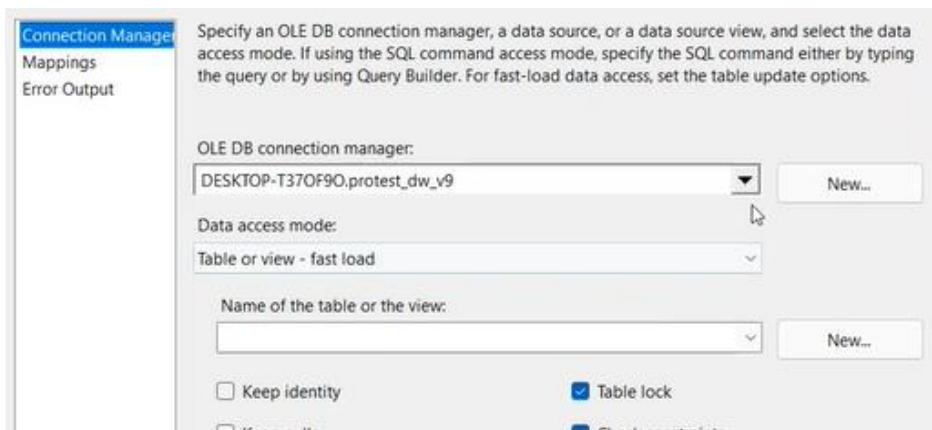


Hình 106. Thao tác với ô OLE DB Destination

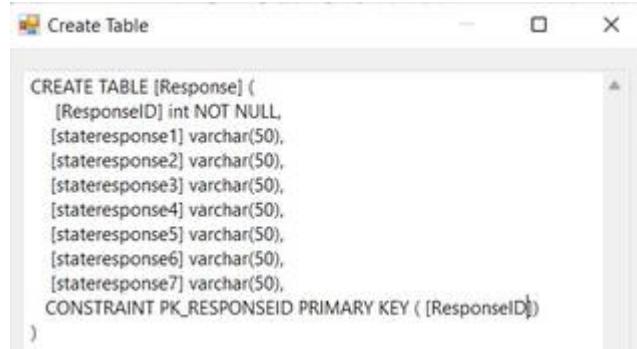
- **Bước 80:** Tại **Connection Manager**, chọn **server name** của SQL Server và kho dữ liệu đích là ProtestDW

*Hình 106. Thực hiện nhập server name và chọn data warehouse*

- **Bước 81:** Tại **OLE DB Destination Editor** > Tạo câu lệnh tạo mới bảng cho DIM_RESPONSE > **New ...**

*Hình 107. Thực hiện tạo bảng Response vào kho dữ liệu đích*

- **Bước 82:** Tại màn hình câu lệnh, chỉnh sửa ràng buộc thuộc tính khóa chính, thứ tự trong bảng,...



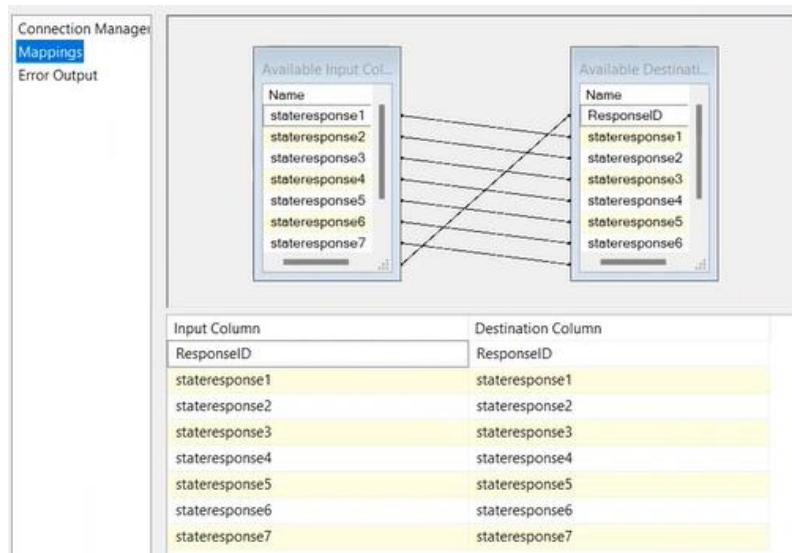
```

CREATE TABLE [Response] (
    [ResponseID] int NOT NULL,
    [stateresponse1] varchar(50),
    [stateresponse2] varchar(50),
    [stateresponse3] varchar(50),
    [stateresponse4] varchar(50),
    [stateresponse5] varchar(50),
    [stateresponse6] varchar(50),
    [stateresponse7] varchar(50),
    CONSTRAINT PK_RESPONSEID PRIMARY KEY ( [ResponseID])
)

```

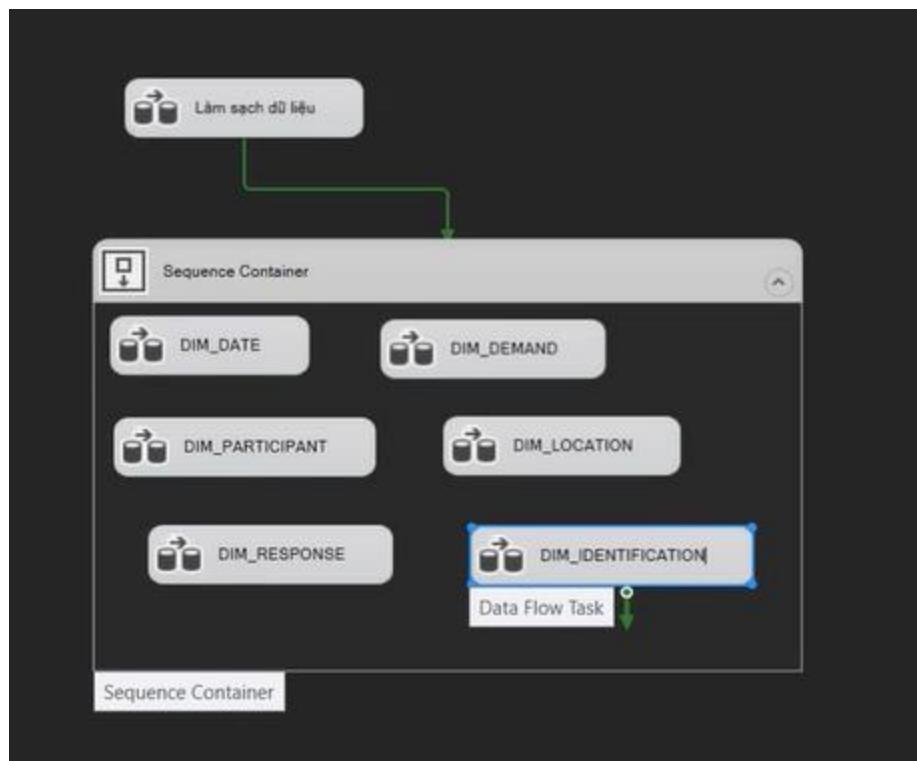
Hình 108. Câu lệnh tạo bảng Response

- **Bước 83:** Kiểm tra mappings > **OK**



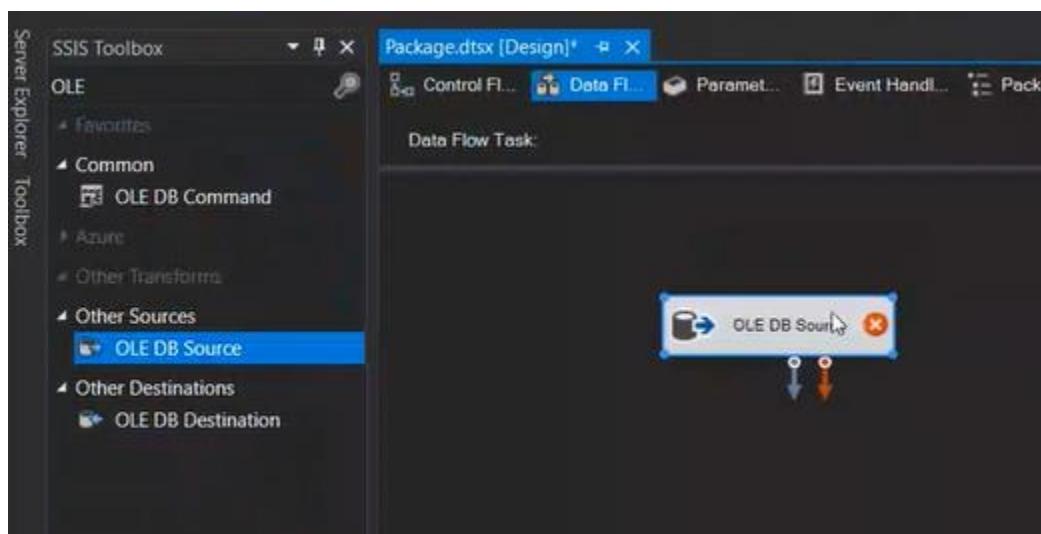
Hình 109. Kiểm tra Mapping các thuộc tính của bảng Response

- **Bước 84:** Tạo bảng **DIM_IDENTIFICATION** > Kéo thả **Data Flow Task** vào **Sequence Container** > Đổi thành tên bảng



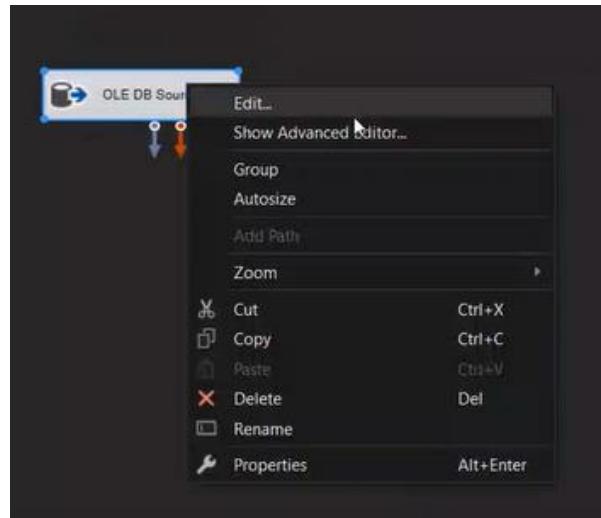
Hình 110. Thao tác với ô *Data Flow Task* của Identification

- **Bước 85:** Nhấn đúp vào *Data Flow Task* của **DIM_IDENTIFICATION**> Kéo thả **OLE DB Source** vào Data Flow Task



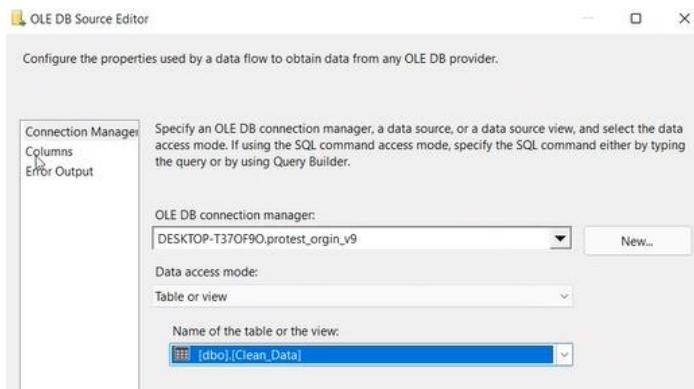
Hình 111. Thao tác với ô *OLE DB Source* của bảng Identification

- **Bước 86:** tiến hành import dữ liệu đã nhập từ *SQL Server*
 - Chuột phải > *Edit*



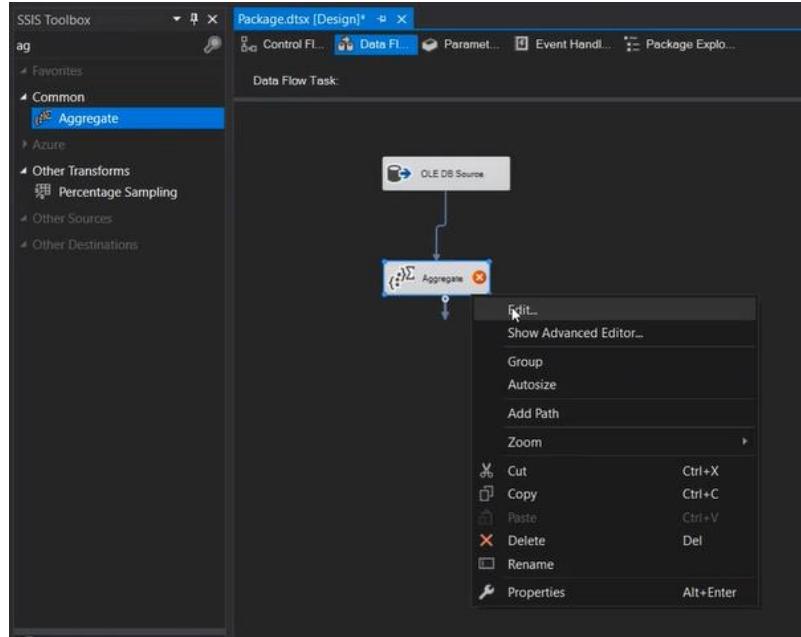
Hình 112. Thực hiện chỉnh sửa tại OLE DB Source

- **Bước 87:** Chọn bảng [*Clean_Data*] vừa tạo ở quá trình làm sạch dữ liệu > *OK*



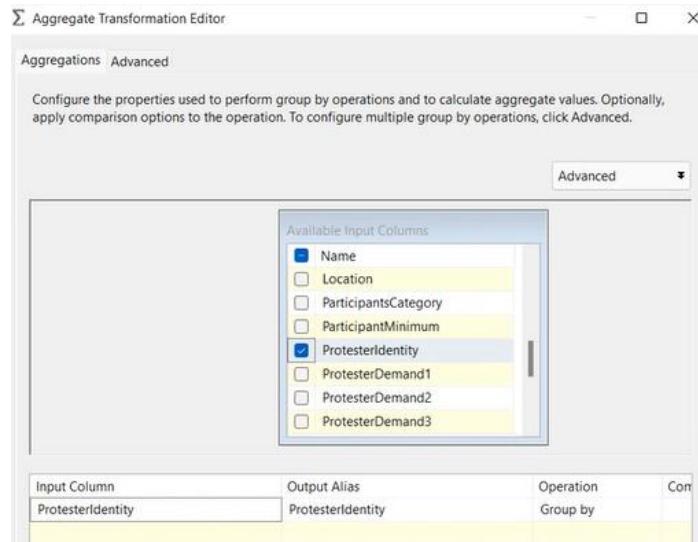
Hình 112. Chọn bảng *Clean_Data* làm dữ liệu nguồn

- **Bước 88:** Kéo thả chức năng *Aggregate* > Chuột phải chọn *Edit*



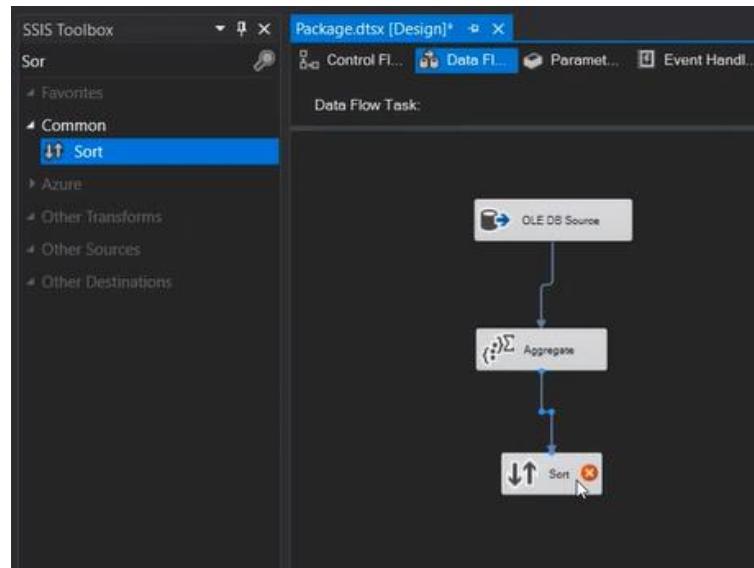
Hình 113. Thực hiện chỉnh sửa tại ô Aggregate

- Bước 89:** Tại màn hình *Aggregate Transformation Editor* > Tick chọn những thuộc tính có trong bảng DIM_IDENTIFICATION (dựa trên lược đồ hình sao) > Operation = ‘Group by’



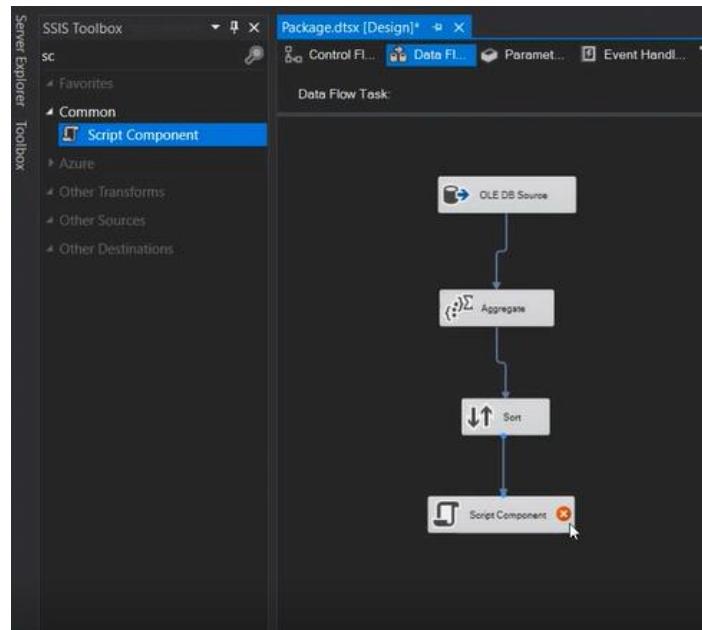
Hình 112. Thực hiện sắp xếp dữ liệu thuộc tính ProtesterIdentity

- **Bước 90:** Kéo thả chức năng Sort để sắp xếp dữ liệu thuộc tính > *Edit*



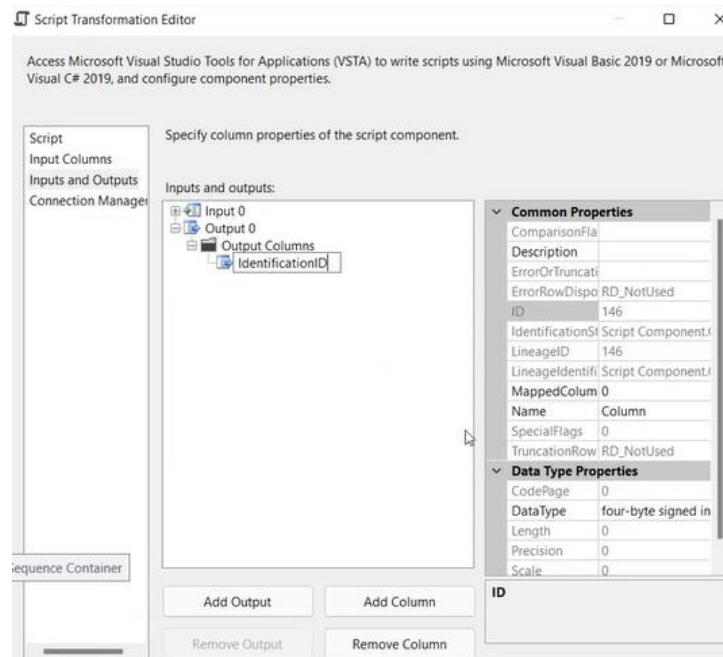
Hình 113. Thực hiện với ô Sort

- **Bước 91:** Chọn những thuộc tính, kiểu sắp xếp
- **Bước 92:** Kéo thả chức năng *Script Component* để thực hiện tạo thêm một biến tăng tự động > Kéo mũi tên từ *Sort* vào *Script Component* > Chuột phải *Edit*



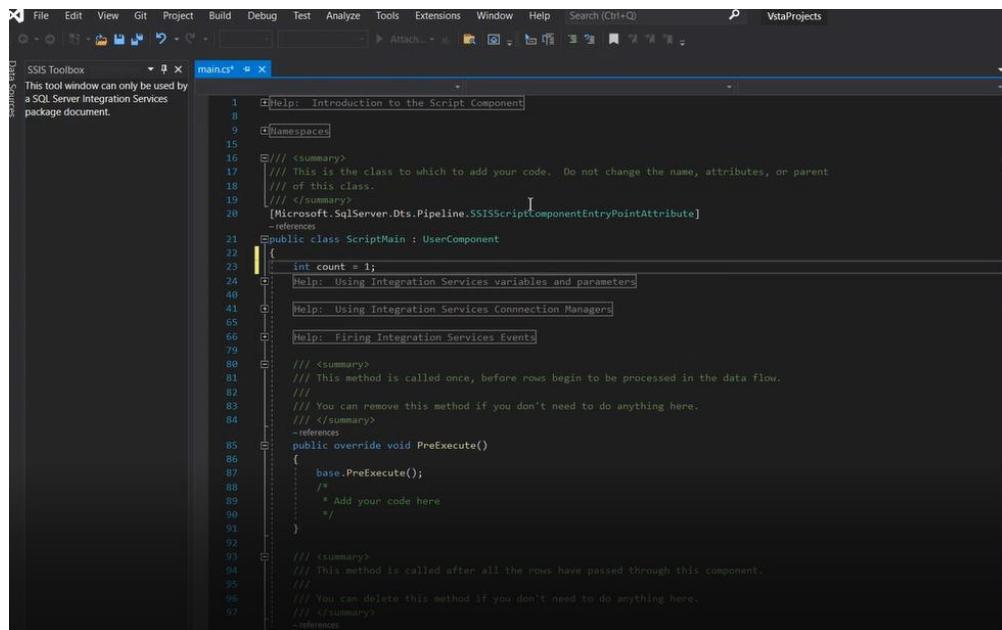
Hình 114. Thực hiện với ô Script Component

- **Bước 93:** Tại *Script Transformation Editor* > chọn mục *Inputs and Outputs* > *Add Column* tại Output Column > Tạo thuộc tính mới tên IdentificationID



Hình 115. Thực hiện tạo thuộc tính IdentificationID

- **Bước 94:** Tại mục *Script* > Chọn *Edit Script* > Hệ thống sẽ mở giao diện code của Visual Studio > Tạo biến đếm count = 0 > cho chạy tăng tự động.



```

1  Help: Introduction to the Script Component
8
9  Namespaces
35
36  /// <summary>
37  /// This is the class to which to add your code. Do not change the name, attributes, or parent
38  /// of this class.
39  /// </summary>
40  [Microsoft.SqlServer.Dts.Pipeline.SSIScriptComponentEntryPointAttribute]
-references
21  public class ScriptMain : UserComponent
22  {
23      int count = 1;
24  }
Help: Using Integration Services variables and parameters
40
41  Help: Using Integration Services Connection Managers
65
66  Help: Firing Integration Services Events
79
80  /// <summary>
81  /// This method is called once, before rows begin to be processed in the data flow.
82  ///
83  /// You can remove this method if you don't need to do anything here.
84  /// </summary>
-references
85  public override void PreExecute()
86  {
87      base.PreExecute();
88      /*
89          * Add your code here
90      */
91  }
92
93  /// <summary>
94  /// This method is called after all the rows have passed through this component.
95  ///
96  /// You can delete this method if you don't need to do anything here.
97  /// </summary>

```

Hình 116.1 Thực hiện chỉnh thuộc tính IdentificationID tăng tự động

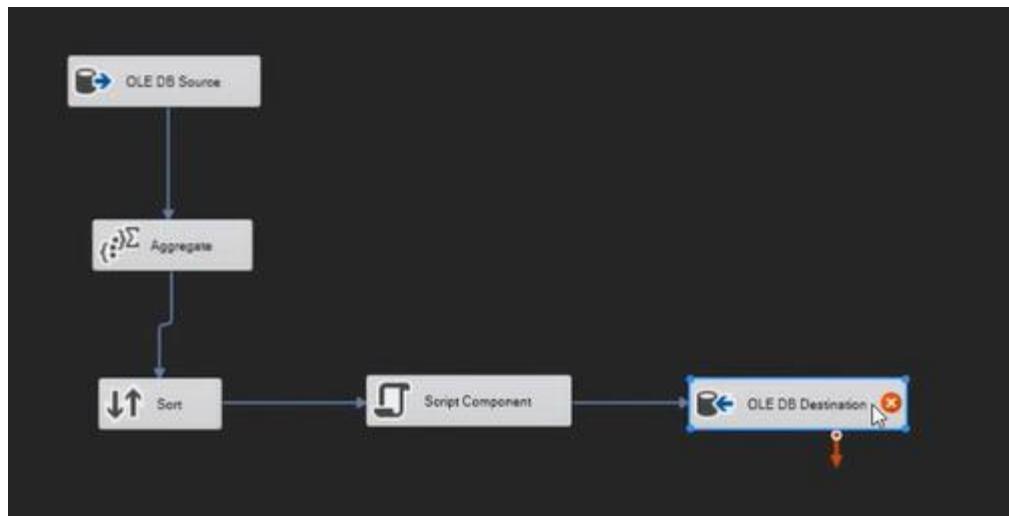
```

public override void Input0_ProcessInputRow(Input0Buffer Row)
{
    Row.IdentificationID = count;
    count++;
}

```

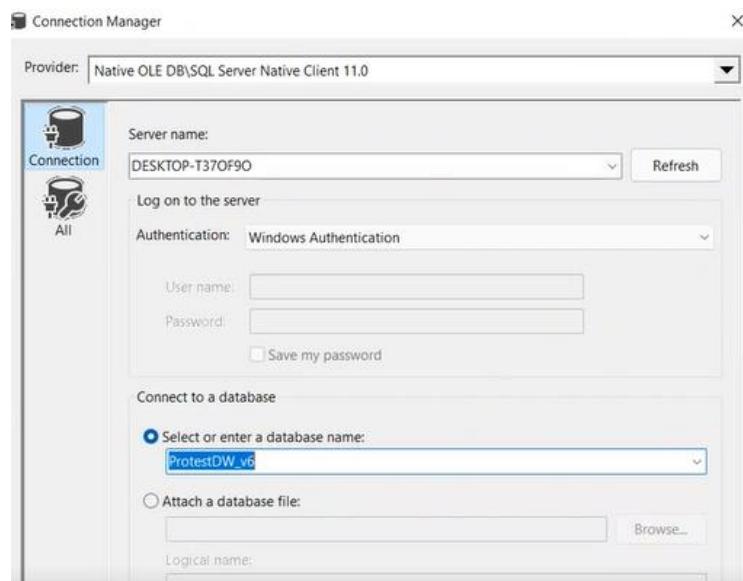
Hình 116.2 Thực hiện chỉnh thuộc tính IdentificationID tăng tự động

- **Bước 95:** Kéo thả chức năng *OLE DB Destination* để truyền dữ liệu bảng **DIM_IDENTIFICATION** vào kho dữ liệu đích > *Edit*



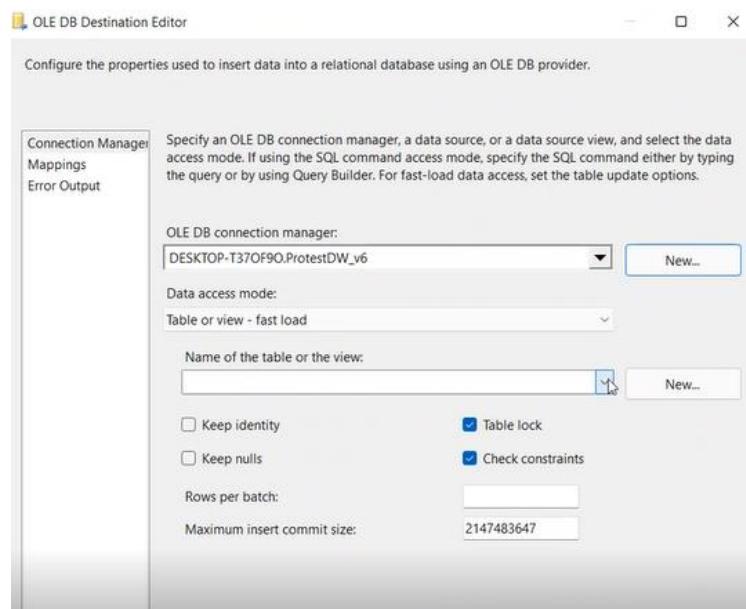
Hình 117. Thao tác tại ô OLE DB Destination

- **Bước 96:** Tại **Connection Manager**, chọn **server name** của SQL Server và kho dữ liệu đích là ProtestDW



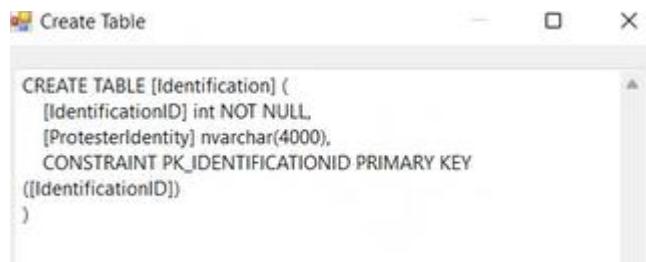
Hình 118. Thực hiện nhập server name và chọn data warehouse

- **Bước 97:** Tại **OLE DB Destination Editor** > Tạo câu lệnh tạo mới bảng cho DIM_IDENTIFICATION > **New ...**



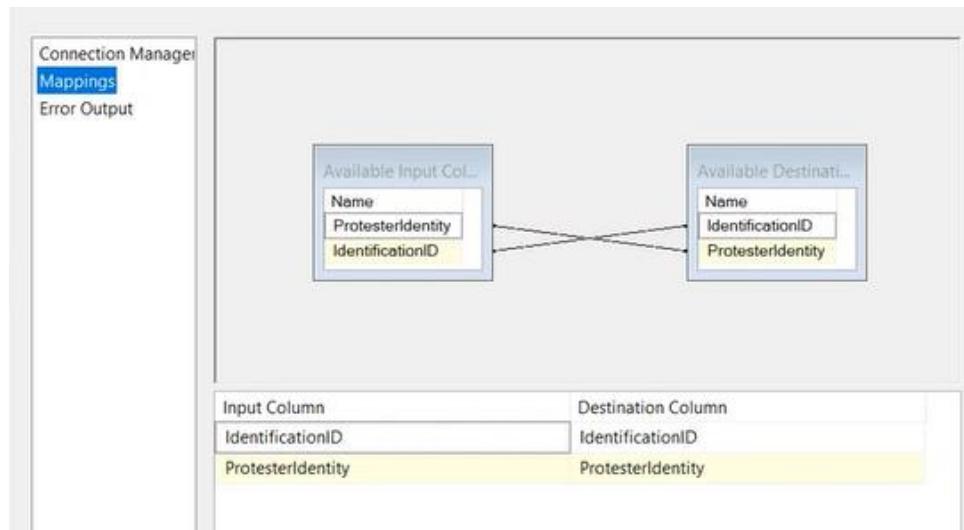
Hình 119. Thực hiện tạo bảng Identification cho dữ liệu đích

- **Bước 98:** Tại màn hình câu lệnh, chỉnh sửa ràng buộc thuộc tính khóa chính, thứ tự trong bảng,...



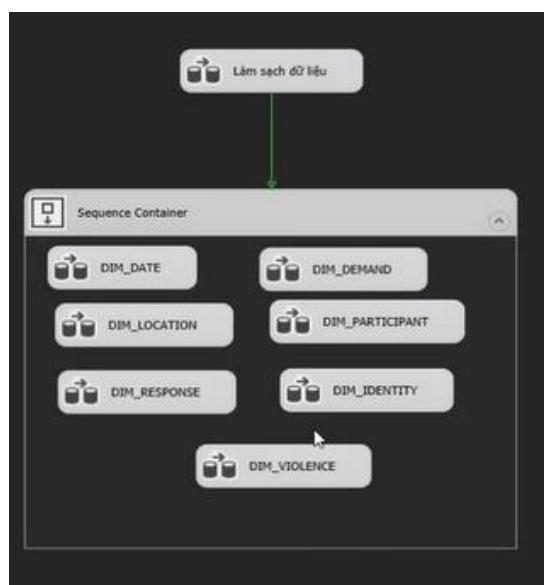
Hình 120. Câu lệnh tạo bảng Identification

- **Bước 99:** Kiểm tra mappings > ***OK***



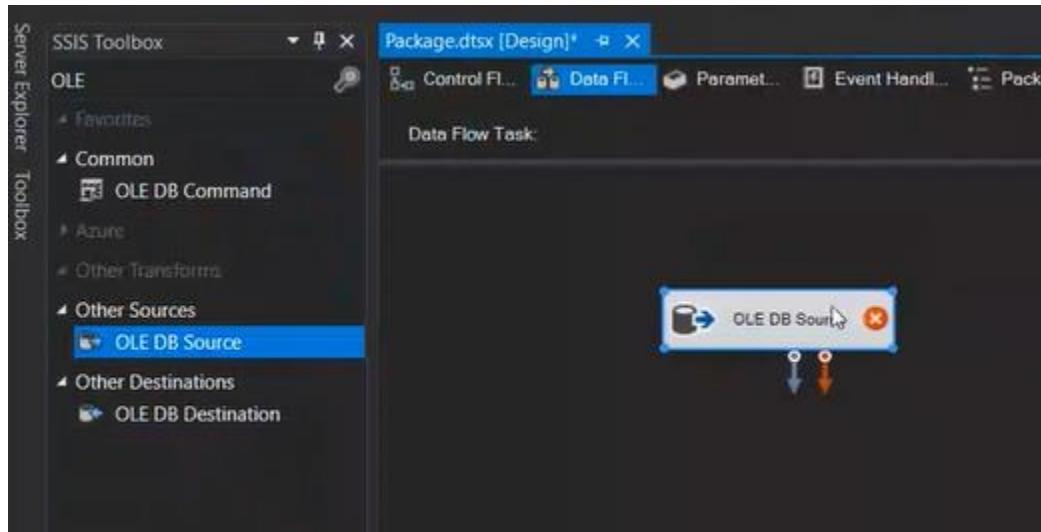
Hình 121. Kiểm tra Mapping các thuộc tính bảng Identification

- **Bước 100:** Tạo bảng **DIM_VIOLENCE** > Kéo thả **Data Flow Task** vào **Sequence Container** > Đổi tên bảng



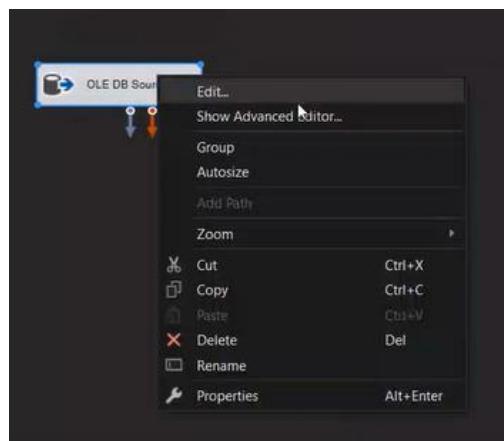
Hình 122. Thao tác với ô Data Flow Task (Dim_Violence)

- **Bước 101:** Nhấn đúp vào *Data Flow Task* của *DIM_VIOLENCE* > Kéo thả *OLE DB Source* vào Data Flow Task



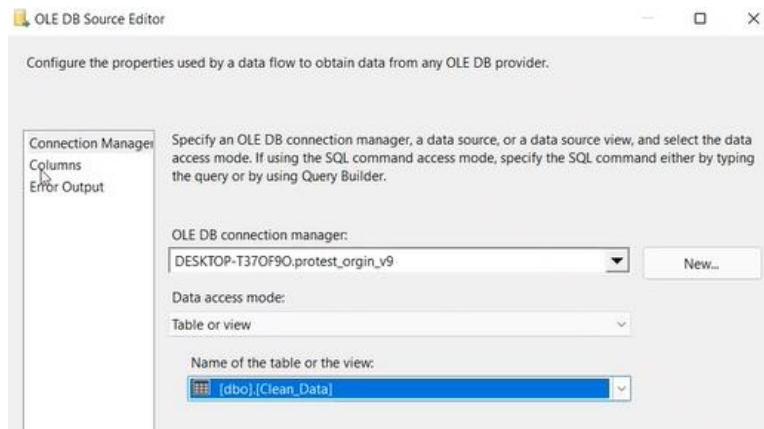
Hình 123. Thao tác với ô *OLE DB Source*

- **Bước 102:** tiến hành import dữ liệu đã nhập từ *SQL Server*
 - Chuột phải > *Edit*



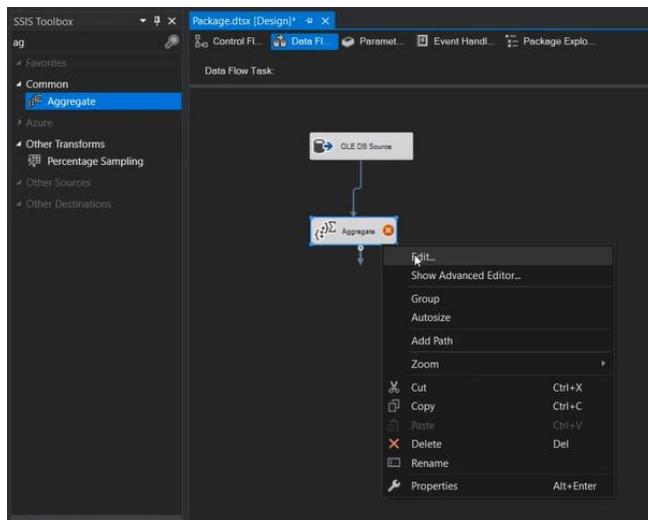
Hình 124. Thực hiện chỉnh sửa tại ô *OLE DB Sourcr*

- **Bước 103:** Chọn bảng [*Clean_Data*] vừa tạo ở quá trình làm sạch dữ liệu > *OK*



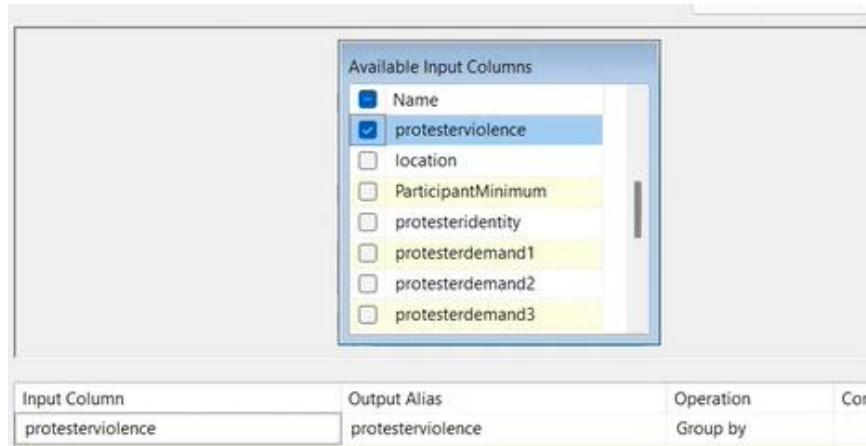
Hình 125. Thực hiện chọn bảng Clean_Data làm dữ liệu nguồn

- **Bước 104:** Kéo thả chức năng *Aggregate* > Chuột phải chọn *Edit*



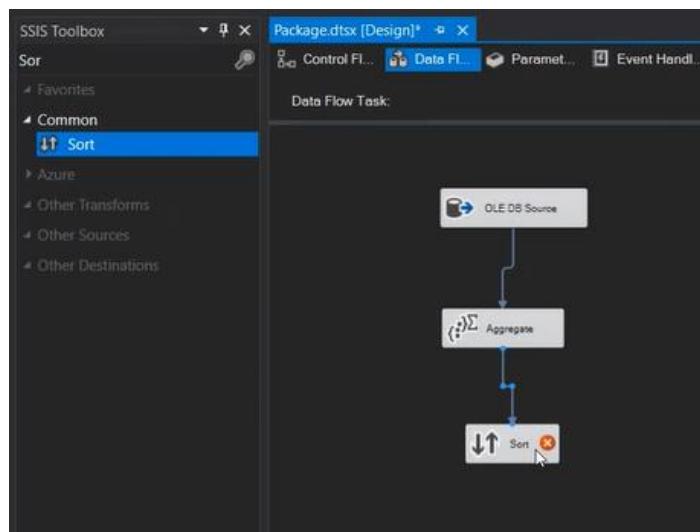
Hình 126. Thực hiện chỉnh sửa tại Aggregate

- **Bước 105:** Tại màn hình *Aggregate Transformation Editor* > Tick chọn những thuộc tính có trong bảng DIM_VIOLENCE (dựa trên lược đồ hình sao) > Operation = ‘Group by’



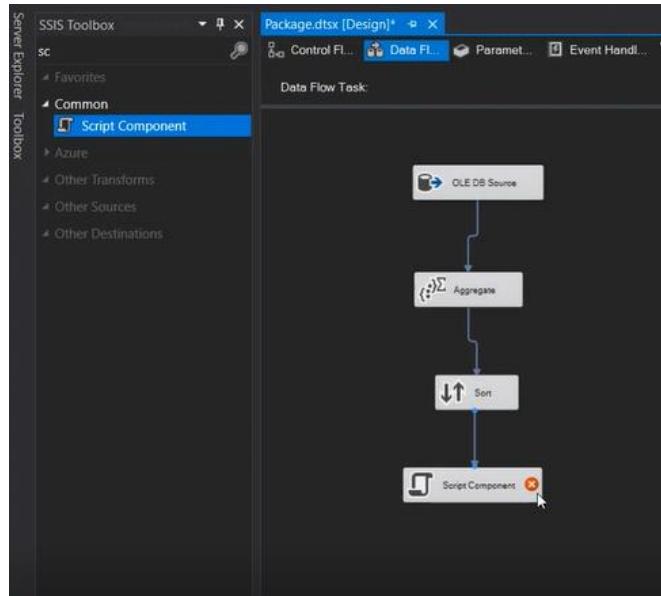
Hình 127. Thực hiện chỉnh sửa thuộc tính tại Aggregate Editor

- **Bước 106:** Kéo thả chức năng Sort để sắp xếp dữ liệu thuộc tính > *Edit*



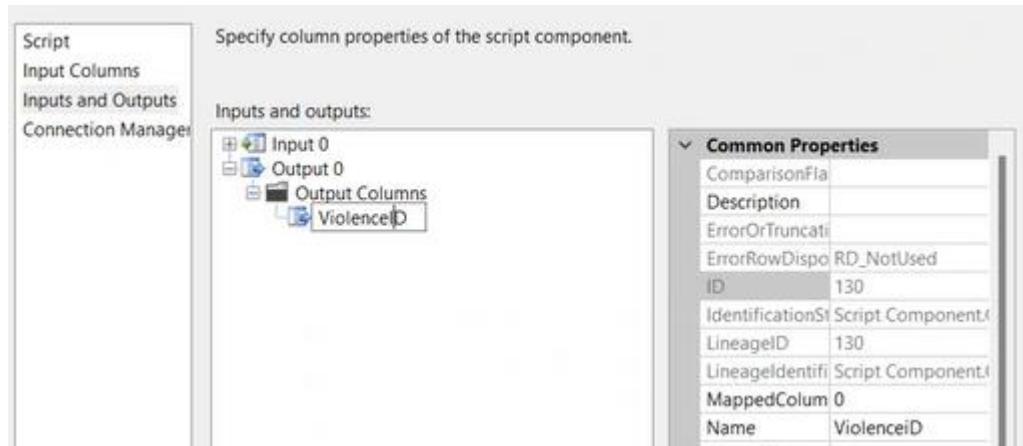
Hình 128. Thao tác tại ô Sort

- **Bước 107:** Chọn những thuộc tính, kiểu cần sắp xếp
- **Bước 108:** Kéo thả chức năng *Script Component* để thực hiện tạo thêm một biến tăng tự động > Kéo mũi tên từ *Sort* vào *Script Component* > Chuột phải *Edit*



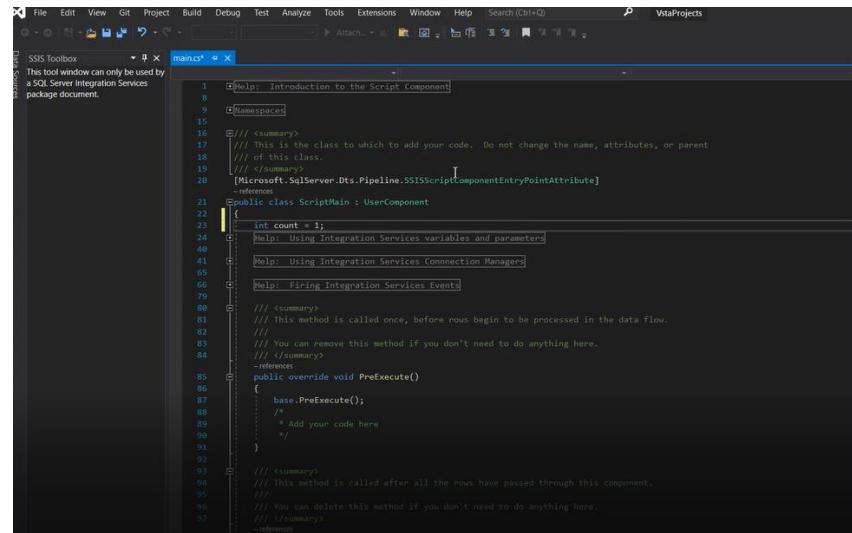
Hình 129. Thao tác với ô *Script Component*

- **Bước 109:** Tại *Script Transformation Editor* > chọn mục *Inputs and Outputs* > *Add Column* tại Output Column > Tạo thuộc tính mới tên ViolenceID



Hình 130. Thêm thuộc tính ViolenceID

- **Bước 110:** Tại mục *Script* > Chọn *Edit Script* > Hệ thống sẽ mở giao diện code của Visual Studio > Tạo biến đếm count = 0 > cho chạy tăng tự động.

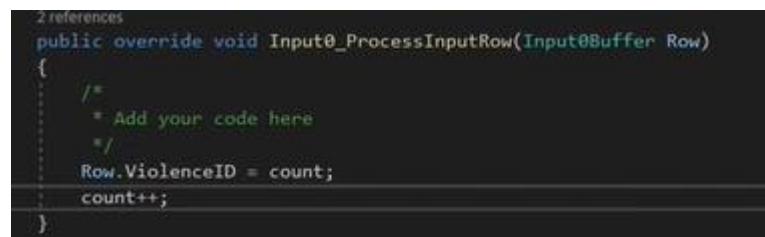


```

1  /// <summary>
2  /// This is the class to which to add your code. Do not change the name, attributes, or parent
3  /// of this class.
4  /// </summary>
5  [Microsoft.SqlServer.Dts.Pipeline.SSIScriptComponentEntryPointAttribute]
6  -references
7  public class ScriptMain : UserComponent
8  {
9      int count = 1;
10     Help: Using Integration Services variables and parameters
11     Help: Using Integration Services Connection Manager
12     Help: Firing Integration Services Events
13
14     /// <summary>
15     /// This method is called once, before rows begin to be processed in the data flow.
16     ///
17     /// You can remove this method if you don't need to do anything here.
18     /// </summary>
19     Help: Using Integration Services Events
20
21     public override void PreExecute()
22     {
23         base.PreExecute();
24         /*
25             * Add your code here
26         */
27         Row.ViolenceID = count;
28         count++;
29     }
30
31     /// <summary>
32     /// This method is called after all the rows have passed through this component.
33     ///
34     /// You can delete this method if you don't need to do anything here.
35     /// </summary>
36     -references
37
38 }

```

Hình 131.1 Thực hiện chỉnh thuộc tính ViolenceID tăng tự động



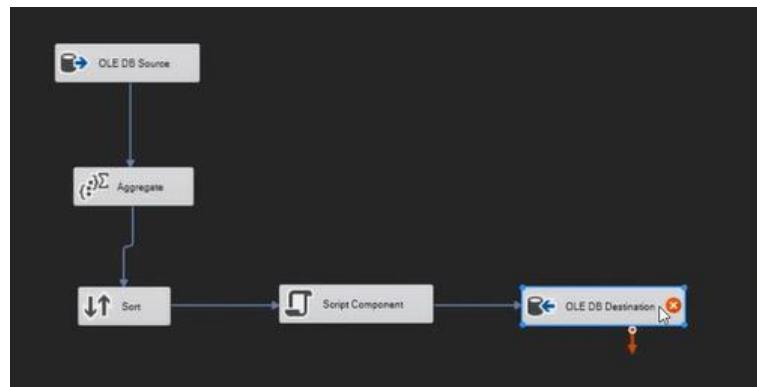
```

2 references
public override void Input0_ProcessInputRow(Input0Buffer Row)
{
    /*
     * Add your code here
     */
    Row.ViolenceID = count;
    count++;
}

```

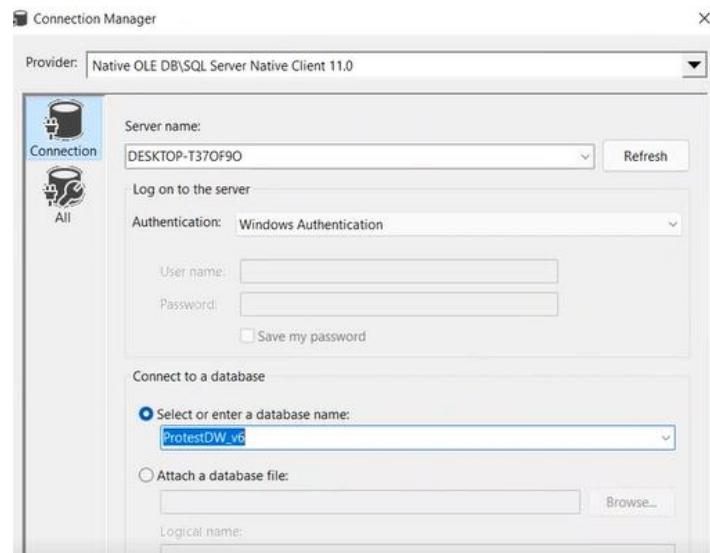
Hình 131.2 Thực hiện chỉnh thuộc tính ViolenceID tăng tự động

- Bước 111:** Kéo thả chức năng **OLE DB Destination** để truyền dữ liệu bảng **DIM_VIOLENCE** vào kho dữ liệu đích > **Edit**



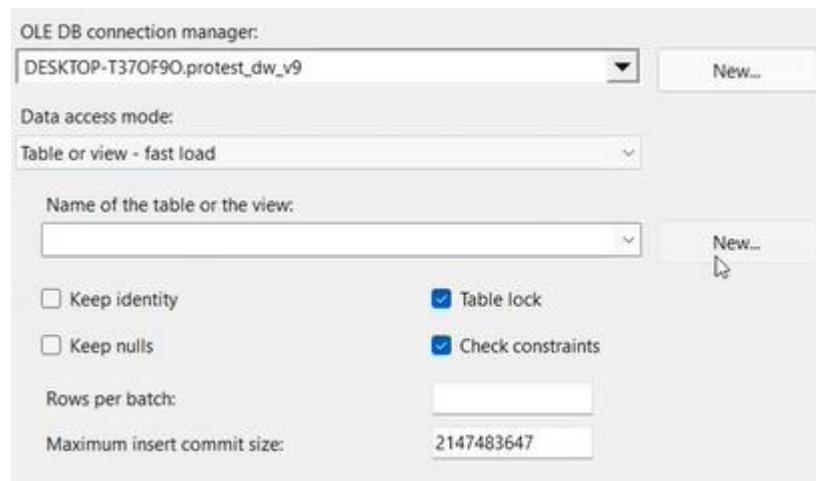
Hình 132. Thao tác tại ô OLE DB Destination

- **Bước 112:** Tại *Connection Manager*, chọn *server name* của SQL Server và kho dữ liệu đích là ProtestDW



Hình 133. Thực hiện nhập server name và chọn data warehouse

- **Bước 113:** Tại *OLE DB Destination Editor* > Tạo câu lệnh tạo mới bảng cho DIM_IDENTIFICATION > *New ...*



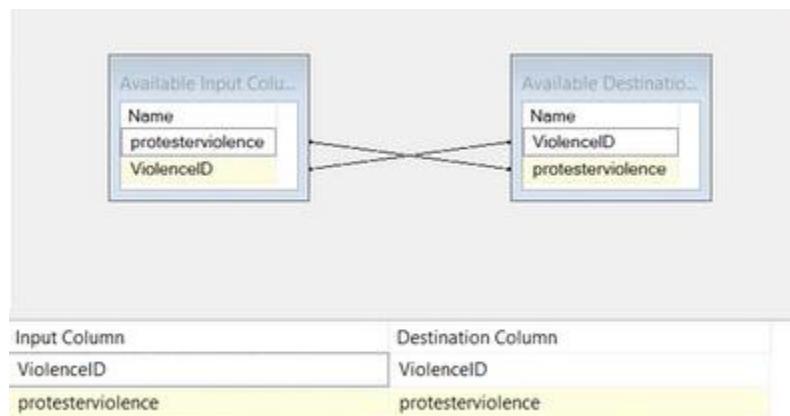
Hình 134. Thực hiện tạo bảng Violence tại kho dữ liệu đích

- **Bước 114:** Tại màn hình câu lệnh, chỉnh sửa ràng buộc thuộc tính khóa chính, thứ tự trong bảng,...

```
CREATE TABLE [Violence] (
    [ViolenceID] int NOT NULL,
    [protesterviolence] varchar(50),
    CONSTRAINT PK_VIOLENCEID PRIMARY KEY ([ViolenceID])
)
```

Hình 135. Câu lệnh tạo bảng Violence

- **Bước 115:** Kiểm tra mappings > *OK*



Hình 136. Thực hiện kiểm tra Mapping các thuộc tính bảng Violence

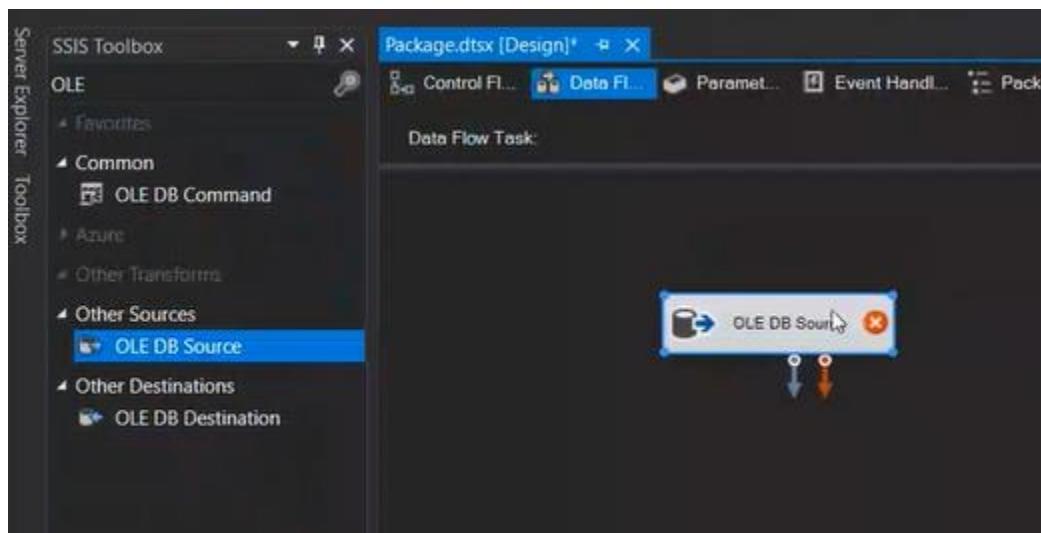
2.2.3 Tạo bảng Fact

- **Bước 1:** Tạo bảng **FACT** > Kéo thả **Data Flow Task** vào **Sequence Container** > Đổi thành tên bảng



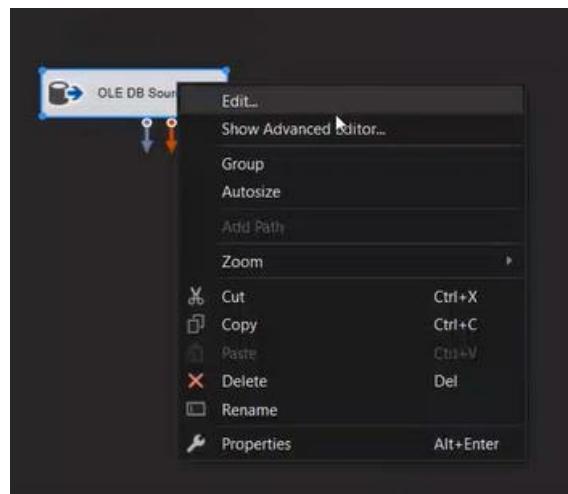
Hình 137. Thao tác tại ô Data Flow Task (Fact)

- **Bước 2:** Nhấn đúp vào **Data Flow Task** của **DIM_VIOLENCE** > Kéo thả **OLE DB Source** vào Data Flow Task



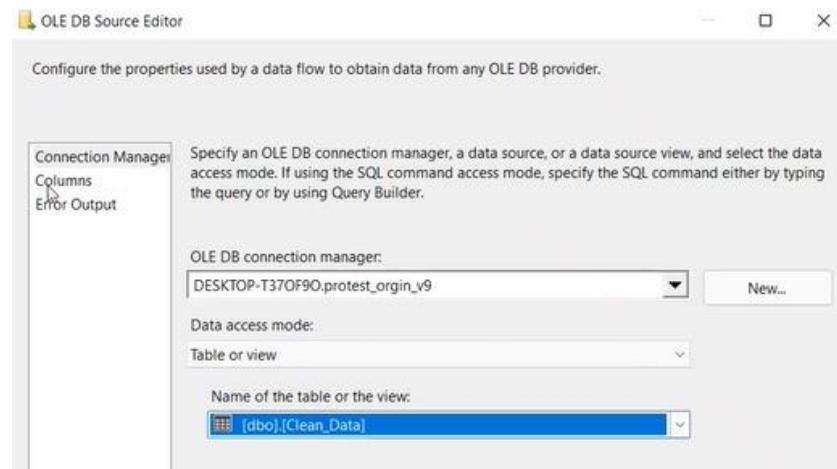
Hình 138. Thao tác tại ô OLE DB Source

- **Bước 3:** tiến hành import dữ liệu đã nhập từ **SQL Server**
 - Chuột phải > **Edit**



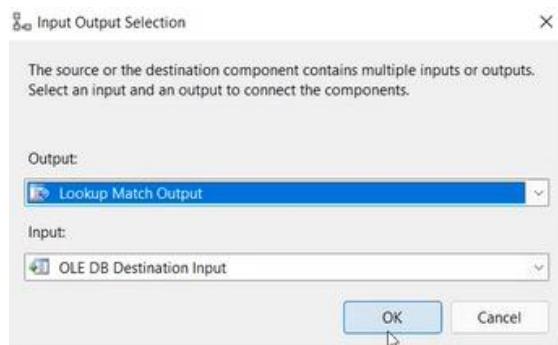
Hình 139. Thao tác chỉnh sửa tại ô OLE DB Source

- **Bước 4:** Chọn bảng [*Clean_Data*] vừa tạo ở quá trình làm sạch dữ liệu > **OK**



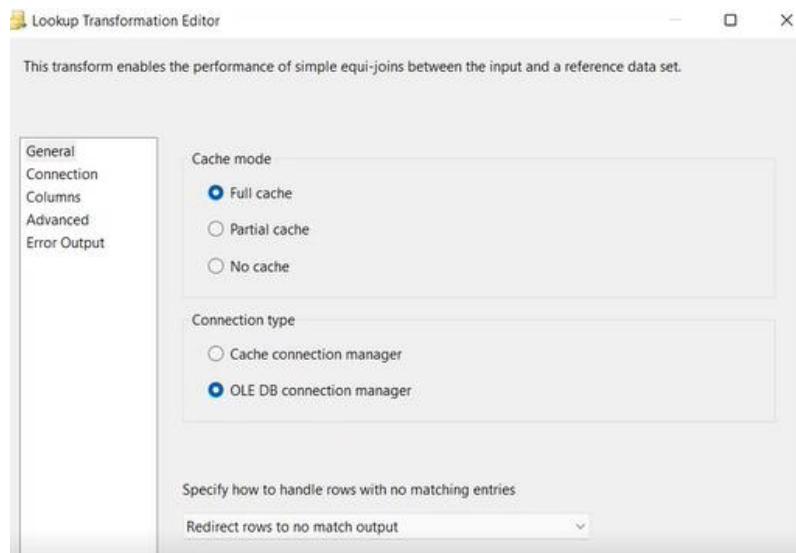
Hình 140. Thực hiện chọn bảng Clean_Data làm dữ liệu nguồn

- **Bước 5:** Kéo thả chức năng *Lookup* > đổi tên > kéo từ *OLE DB Source* sang ô LookUp > chọn Output: ‘*Lookup Match Output*’ > **Edit**



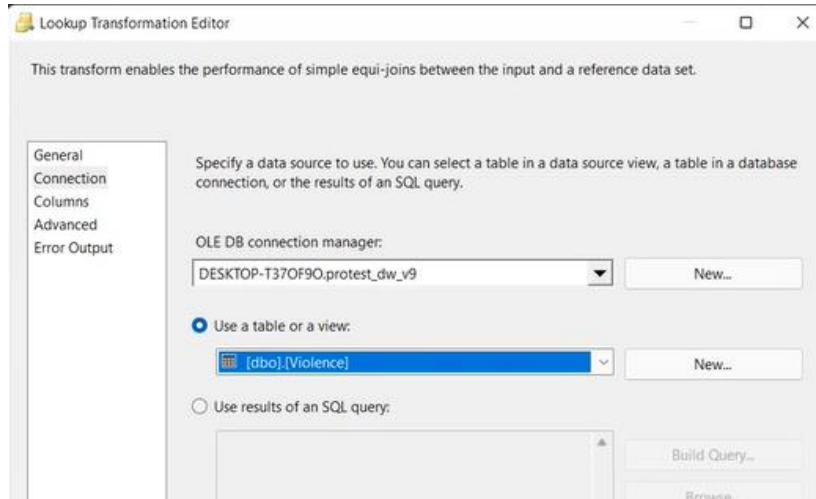
Hình 141. Thực hiện lựa chọn đầu ra cho ô LookUp

- **Bước 6:** Specify how to handle rows with no matching entries: chọn ‘*Redirect rows to match output*’



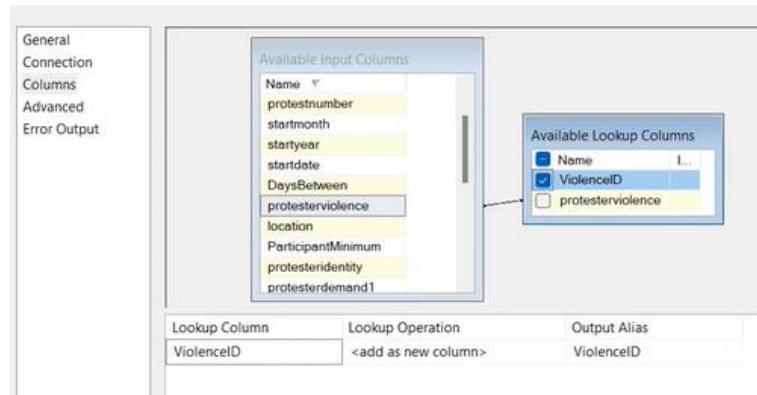
Hình 142. Lựa chọn kiểu kết nối trong Lookup Editor

- **Bước 7:** Tại phần Connection chọn bảng mình muốn sử dụng
lookup > ***Violence***



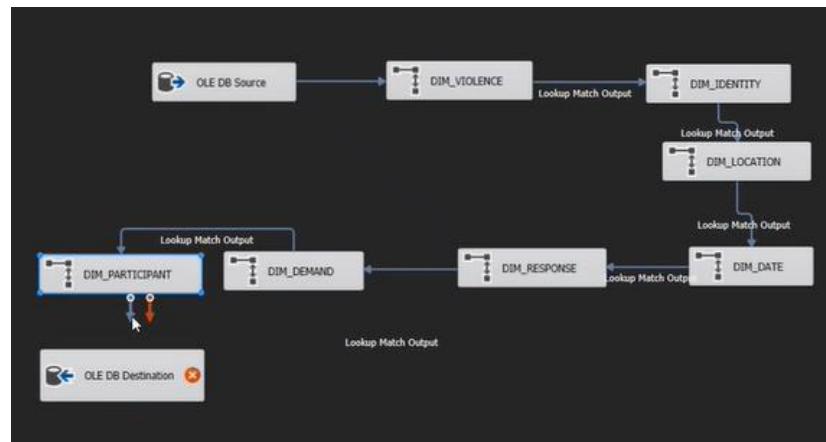
Hình 143. Thực hiện chọn bảng Violence trong Lookup Editor

- **Bước 8:** Chọn thuộc tính mình cần và kéo sang bảng phải để thực hiện lookup



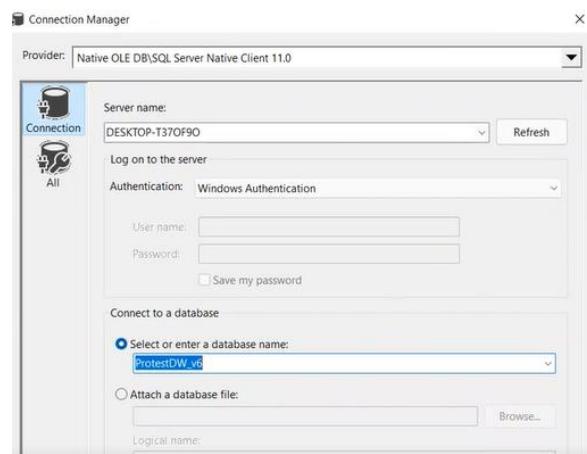
Hình 142. Thực hiện lookup dữ liệu mình cần

- Thực hiện Lookup cho các bảng còn lại: thực hiện tương tự như các bước từ 1 đến 8
- **Bước 9:** Kéo thả chức năng **OLE DB Destination** để truyền dữ liệu bảng **Fact** vào kho dữ liệu đích > **Edit**



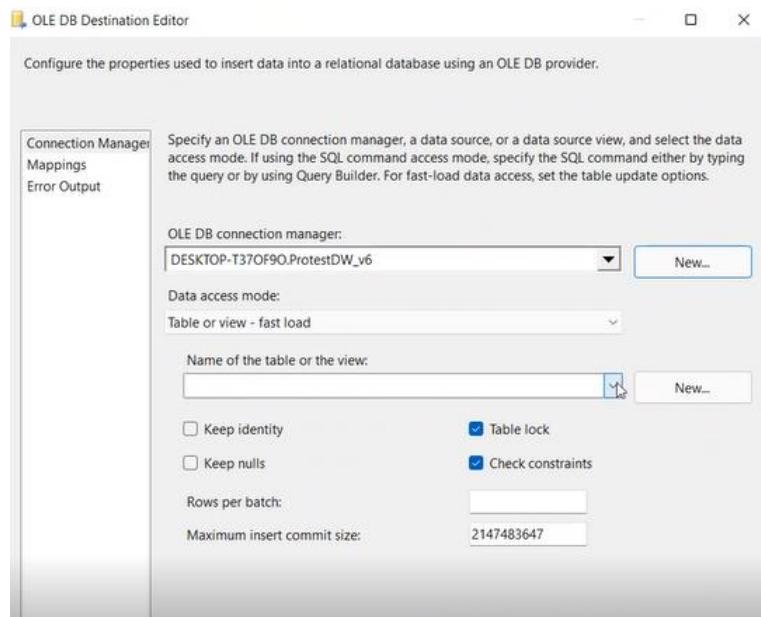
Hình 143. Toàn bộ quá trình của Data Task Flow (Fact)

- Bước 10:** Tại **Connection Manager**, chọn **server name** của SQL Server và kho dữ liệu đích là ProtestDW



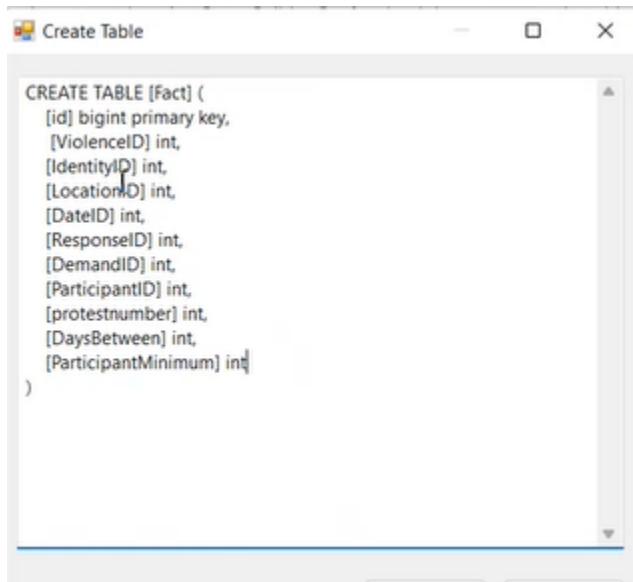
Hình 144. Thực hiện nhập Server name và chọn data warehouse

- Bước 11:** Tại **OLE DB Destination Editor** > Tạo câu lệnh tạo mới bảng cho Fact > **New ...**



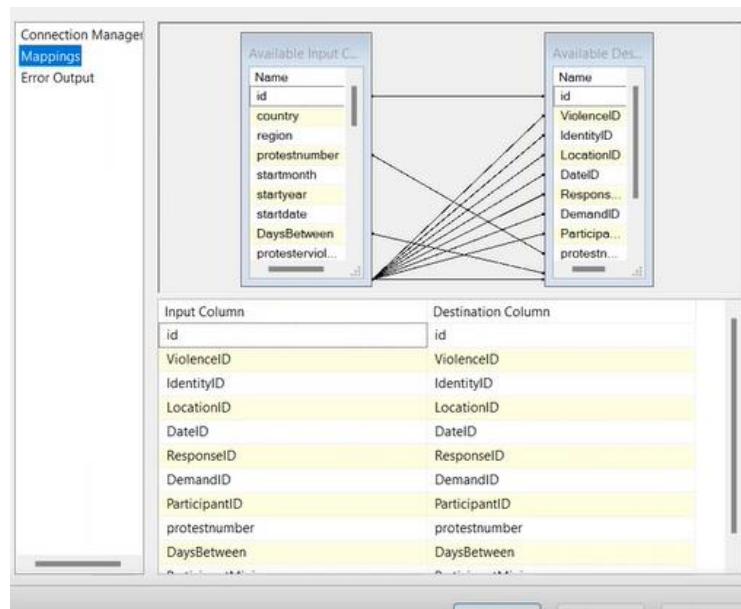
Hình 145. Tạo bảng Fact cho kho dữ liệu đích

- **Bước 12:** Tại màn hình câu lệnh, chỉnh sửa ràng buộc thuộc tính khóa chính, thứ tự trong bảng,...



Hình 146. Câu lệnh tạo bảng Fact

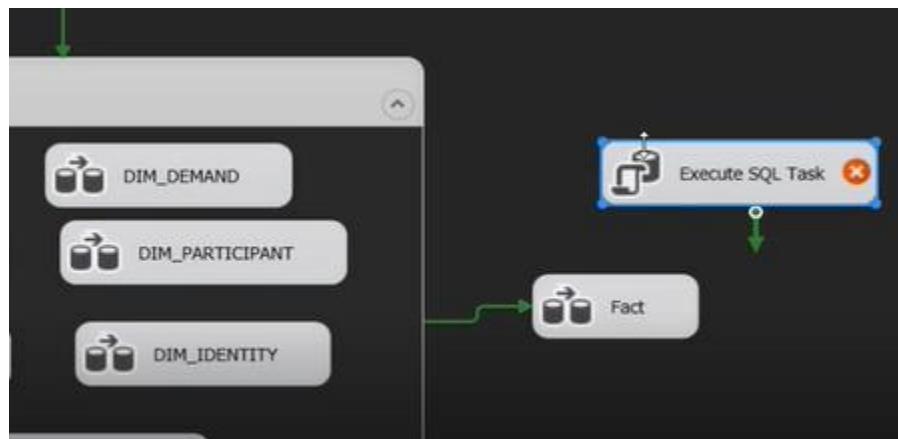
- **Bước 13:** Kiểm tra mappings > **OK**



Hình 146. Kiểm tra Mapping các thuộc tính bảng Fact

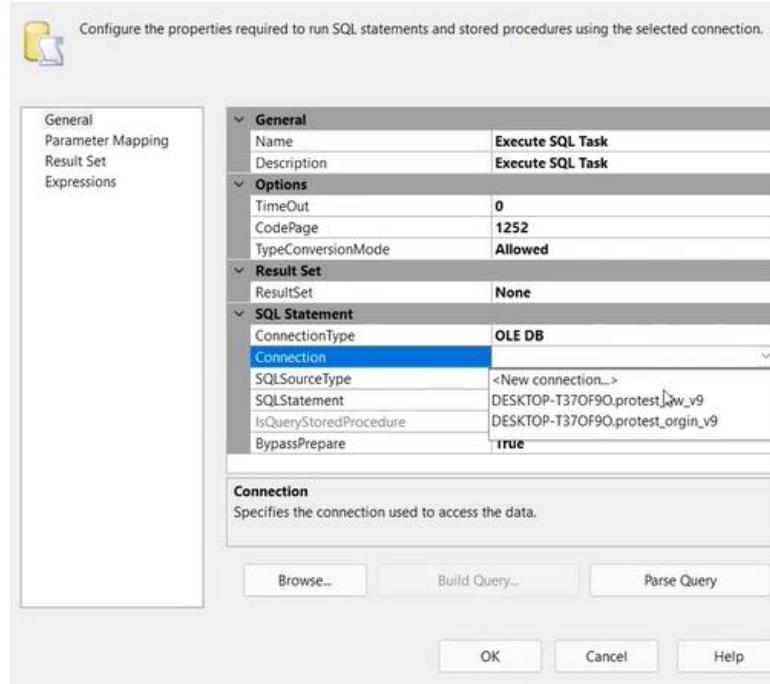
2.2.4 Tạo các ràng buộc khóa ngoại và thực thi toàn bộ quá trình SSIS

- Bước 1:** Kéo thả chức năng *Execute SQL Task* để thực hiện ràng buộc khóa ngoại > *Edit*



Hình 147. Thực hiện tạo ô chức năng Execute SQL Task

- Bước 2:** Chính Connection kết nối đến kho dữ liệu đích tại SQL Server



Hình 148. Chọn đường dẫn đến SQL Server

- **Bước 3:** Tại *SQL Statement* dán đoạn lệnh ràng buộc khóa ngoại vào

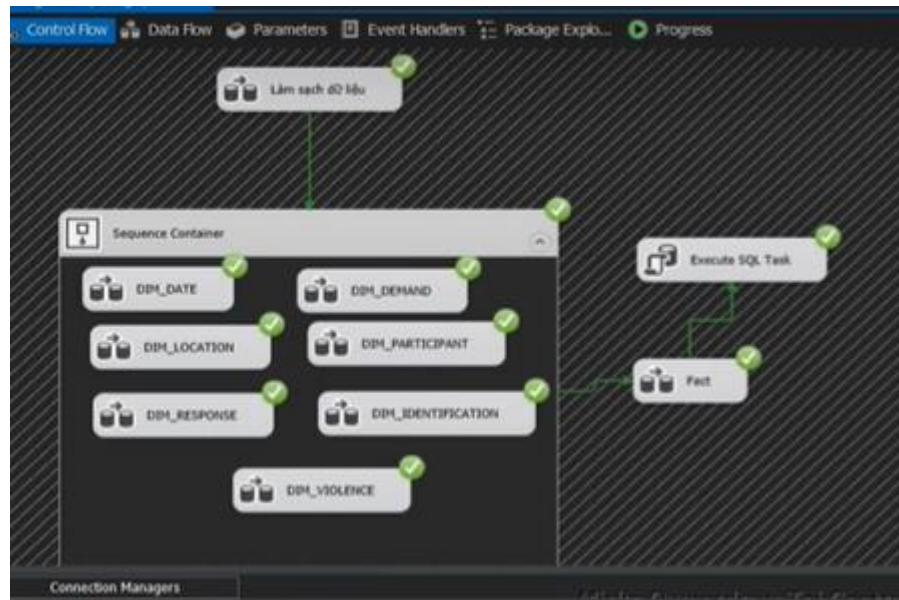
```

foreign key (ResponseID) references Response(ResponseID);
go
alter table Fact
add constraint fk_demandid
foreign key (DemandID) references Demand(DemandID);
go
alter table Fact
add constraint fk_locationid
foreign key (LocationID) references Location(LocationID);
go
alter table Fact
add constraint fk_identityid
foreign key (IdentityID) references Identity(IdentityID);
go
alter table Fact
add constraint fk_violenceid
foreign key (ViolenceID) references Violence(ViolenceID);
go

```

Hình 149. Câu lệnh ràng buộc khóa ngoại

- **Bước 4:** Tại folder SSIS Packages chọn file *Package.dtsx* > *Excecute Package*



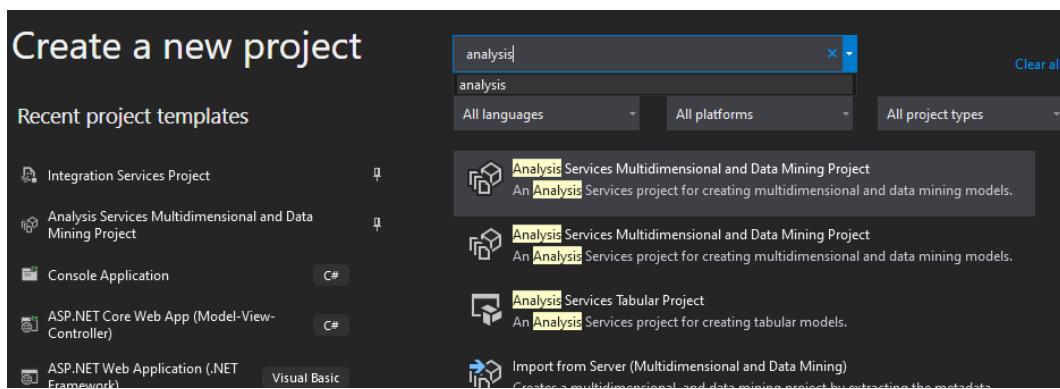
Hình 150. Kết quả chạy quá trình SSIS

CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU TRONG KHO (SSAS)

1. QUÁ TRÌNH SSAS TRONG VISUAL STUDIO 2019

1.1 Tạo project tại Visual Studio 2019 (Define Data Source)

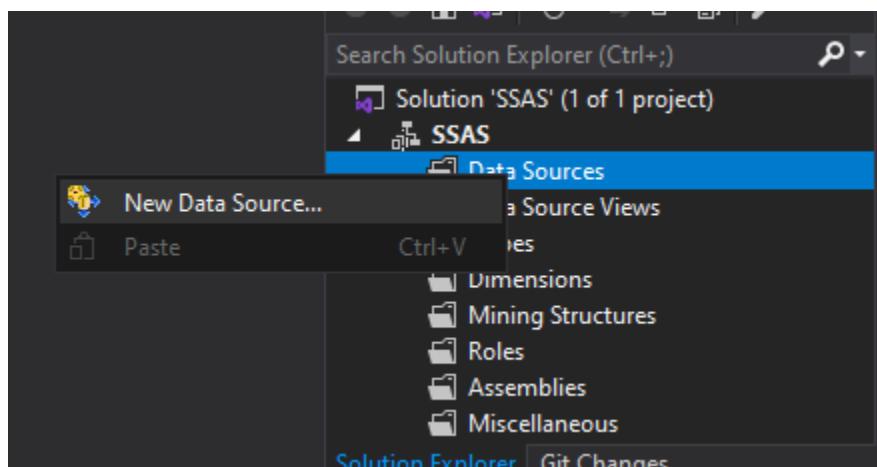
Mở Visual Studio 2019 > *Create a new project* > ‘*Analysis Service Multidimensional and Data Mining Project*’ > Đặt tên > lưu



Hình 151. Chọn Project SSAS trong Visual Studio 2019

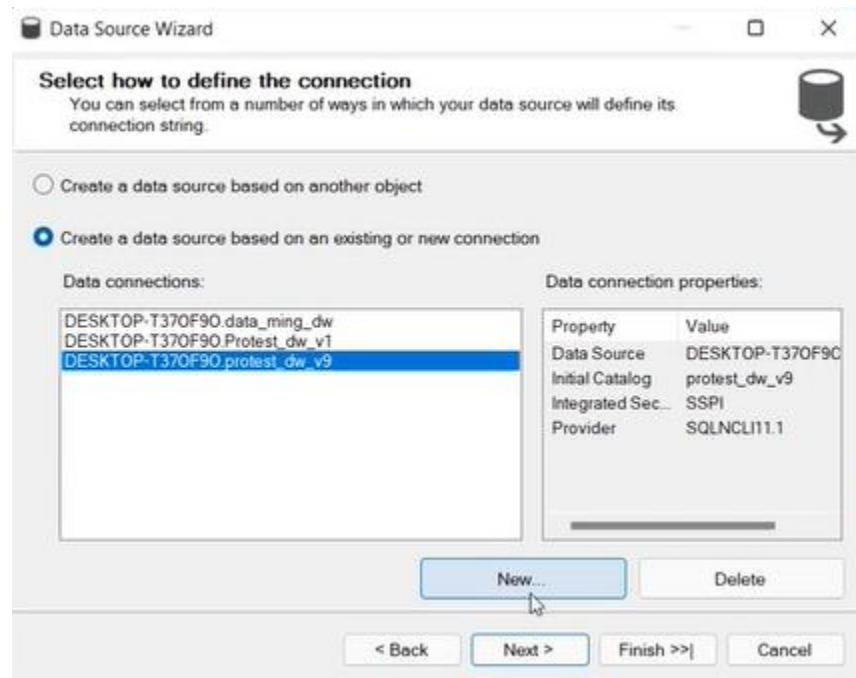
1.2 Xác định dữ liệu nguồn

Bước 1: Tại mục *Solution* của project > Chuột phải *New Data Source* để kết nối dữ liệu từ kho dữ liệu SQL Server

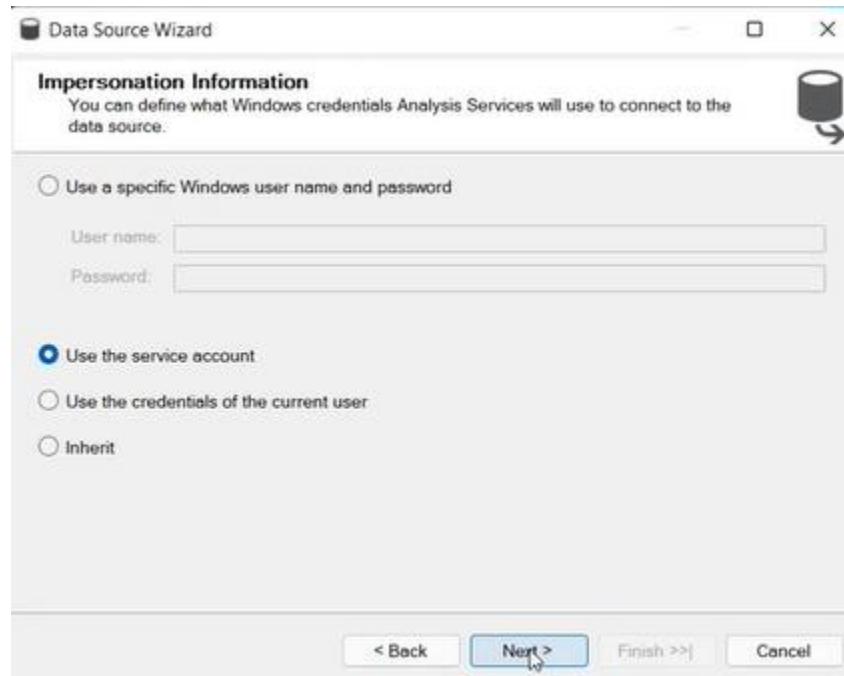


Hình 152. Thực hiện tạo mới Data Source

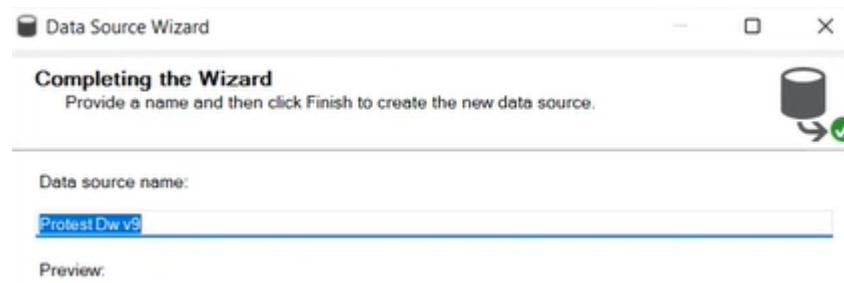
Bước 2: Sau khi màn hình Data Source Wizard mở > Next > Xác định dữ liệu nguồn > Chọn kết nối đã tạo sẵn ‘*Create a data source base on an existing or new connection*’ > Chọn Data Connection



Hình 153. Chọn kết nối dữ liệu đến kho dữ liệu trong SQL Server

Bước 3: Chọn ‘*Use the service account*’

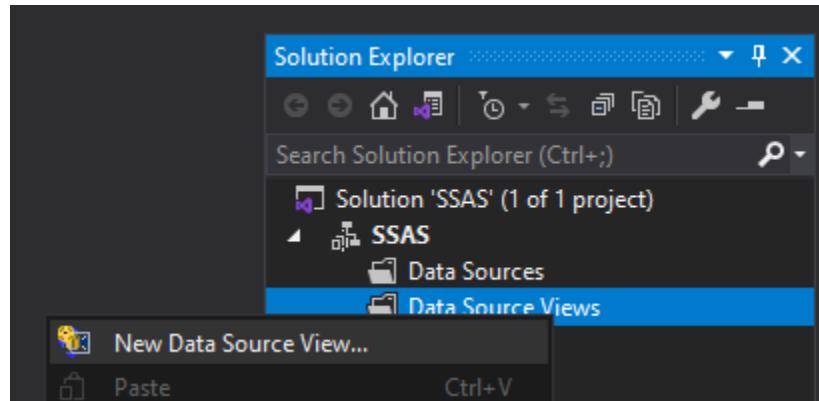
Hình 154. Chọn tài khoản kết nối Analysis Services

Bước 4: Đặt tên cho dữ liệu nguồn

Hình 155. Đặt tên nguồn dữ liệu

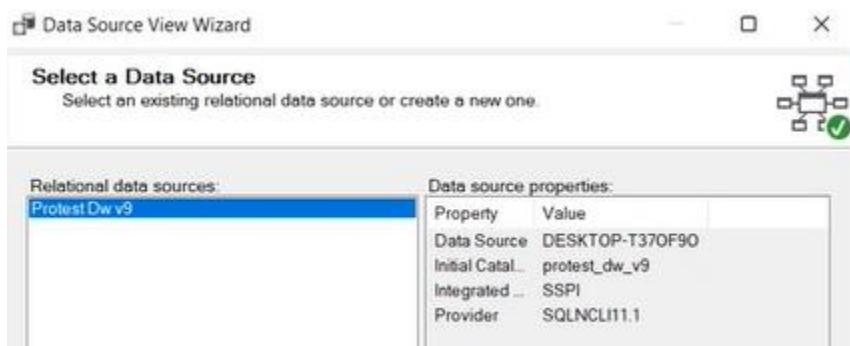
1.3 Xác định khung dữ liệu nguồn (Define Data Source View)

Bước 1: Chuột phải chọn *New Data Source View* > Next



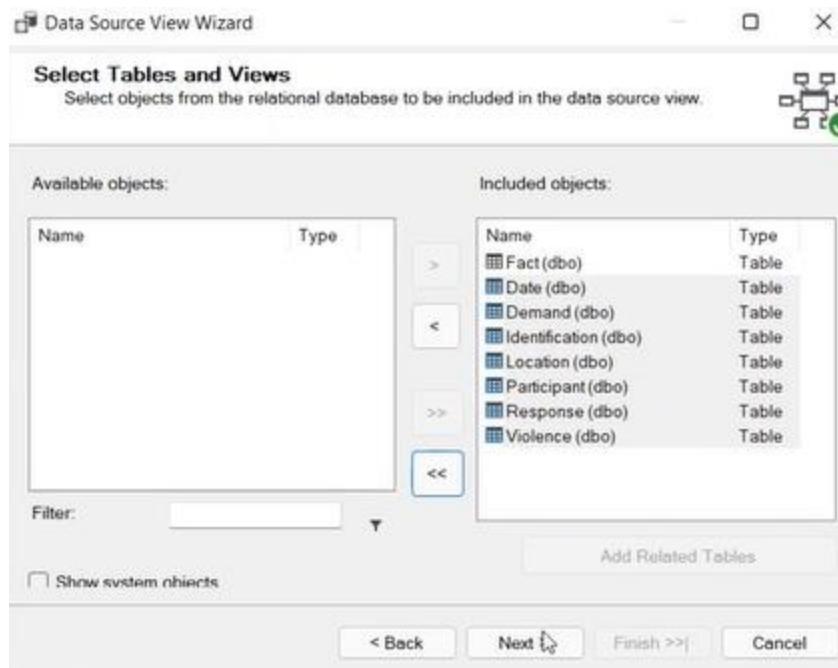
Hình 156. Thực hiện tạo Data Source View

Bước 2: Chọn dữ liệu nguồn vừa thêm vào ở mục 2.3.2



Hình 157. Thực hiện chọn dữ liệu nguồn cho khung dữ liệu

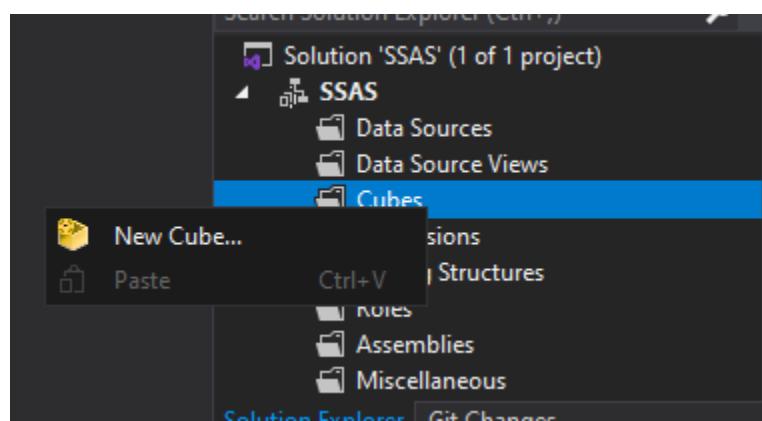
Bước 3: Chọn các bảng sẽ sử dụng vào khung '**Included objects**' > Next
> Finish kết thúc quá trình tạo khung dữ liệu



Hình 158. Thực hiện tạo Data Source View

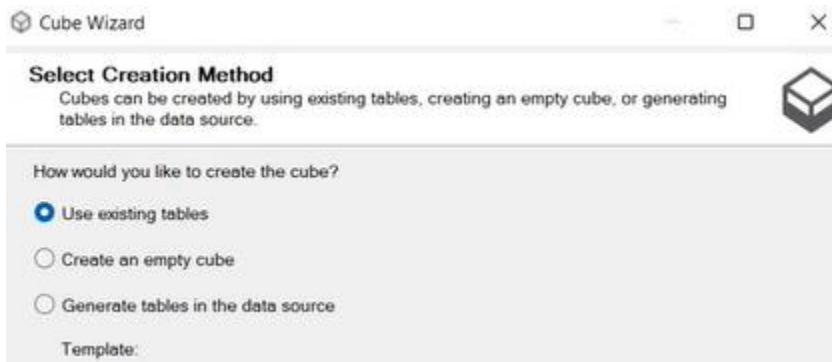
1.3.1 XÁC ĐỊNH KHỐI (DEFINE CUBE)

Bước 1: Chuột phải *New Cube*



Hình 159. Chọn tạo khối mới – New Cube

Bước 2: Chọn cách tạo cube ‘use existing tables’ – bảng có sẵn



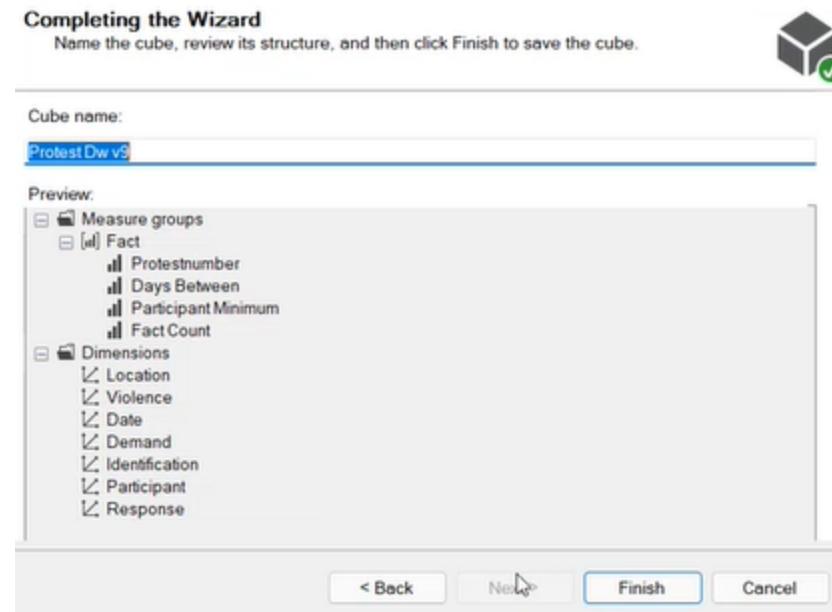
Hình 160. Chọn phương thức tạo khối

Bước 3: Chọn bảng Fact để tạo measure group



Hình 161. Thực hiện chọn bảng tạo Measure Group

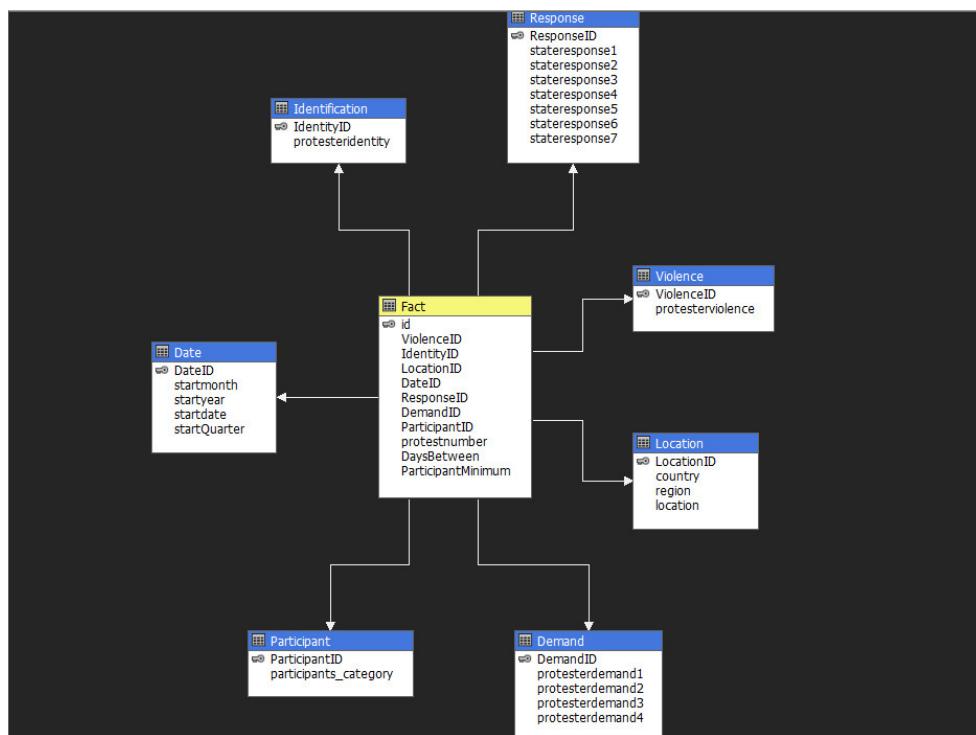
Bước 4: Kiểm tra lại và đặt tên cho cube > *Finish*



Hình 162. Kiểm tra lại các bảng Dimension và Measure Groups

Bước 5: Kiểm tra lại và đặt tên cho cube > **Finish**

Kết quả ‘Data Source View’:



Hình 163. Kết quả khung dữ liệu

1.3.2 TẠO HIERACHY

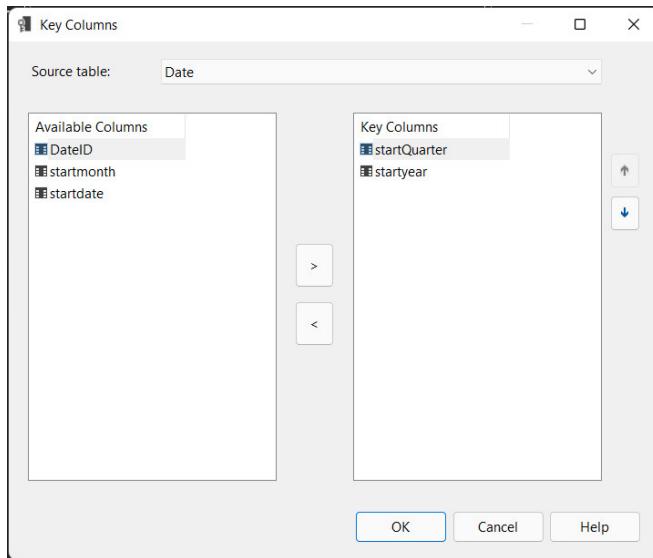
Sau khi tạo cube, Visual Studio sẽ tự tạo các dimension của các bảng. Tiếp đến tiến hành tạo Hierachies – phân cấp theo thời gian, địa điểm.

Bước 1: Tạo phân cấp theo thời gian, kéo những thuộc tính cần qua cửa sổ

Hierachies > Đặt tên cho các bảng Hierachies

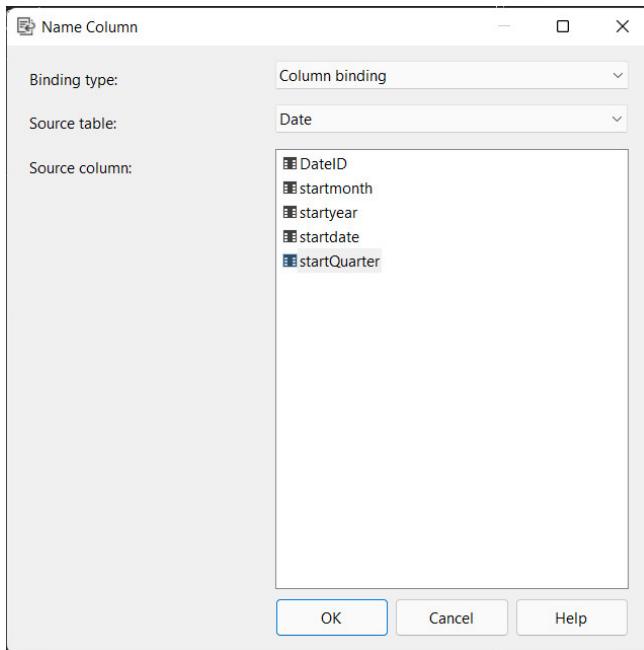
*Hình 164. Thực hiện phân cấp thuộc tính*

Bước 2: Vào Attributes, chỉnh khóa dòng thuộc tính *StartQuarter*. (Cấp nhỏ hơn: *StartYear*)



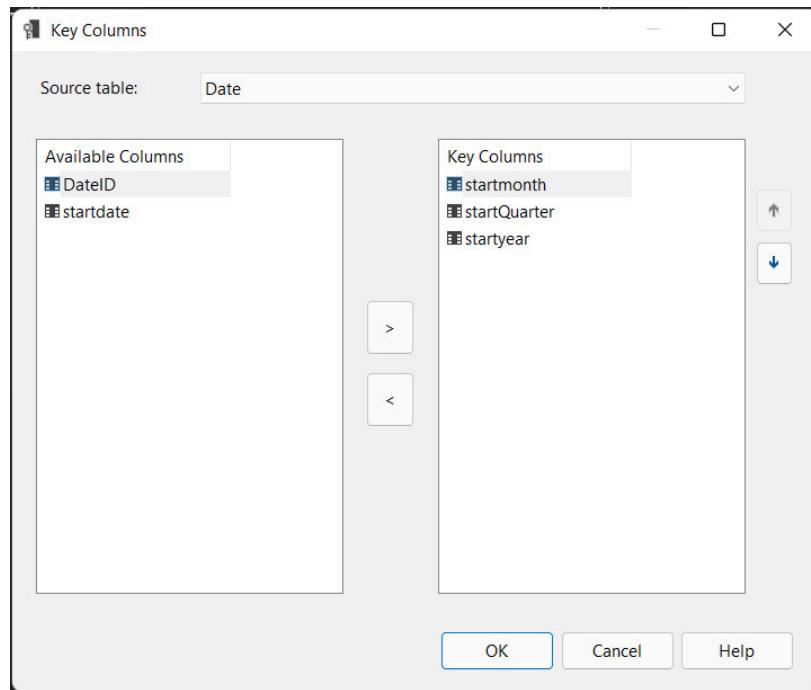
Hình 165. Thực hiện chọn thuộc tính phân cấp cho StartQuarter

Bước 3: Tạo tên dòng của thuộc tính **OrderQuarter**



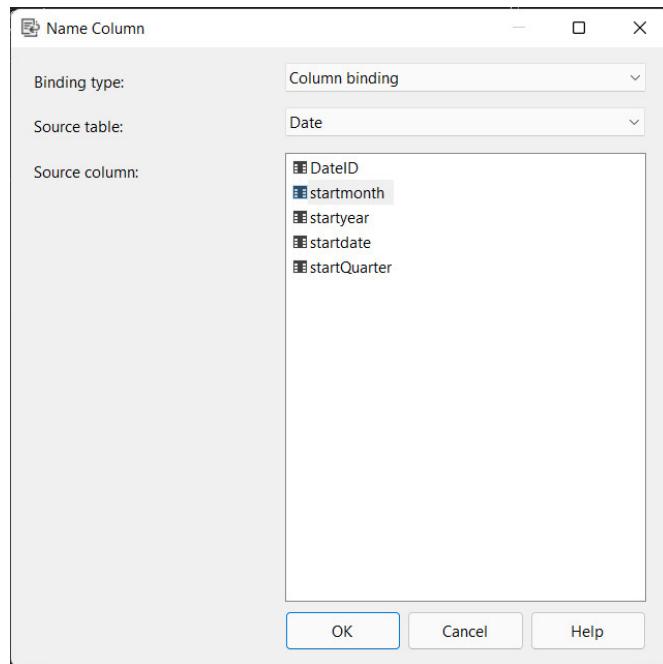
Hình 166. Tạo tên cho thuộc tính StartQuarter

Bước 4: Vào Attributes, chỉnh khóa dòng và tên dòng thuộc tính *StartMonth*. (Cấp nhỏ hơn: *StartYear*, *StartQuarter*)



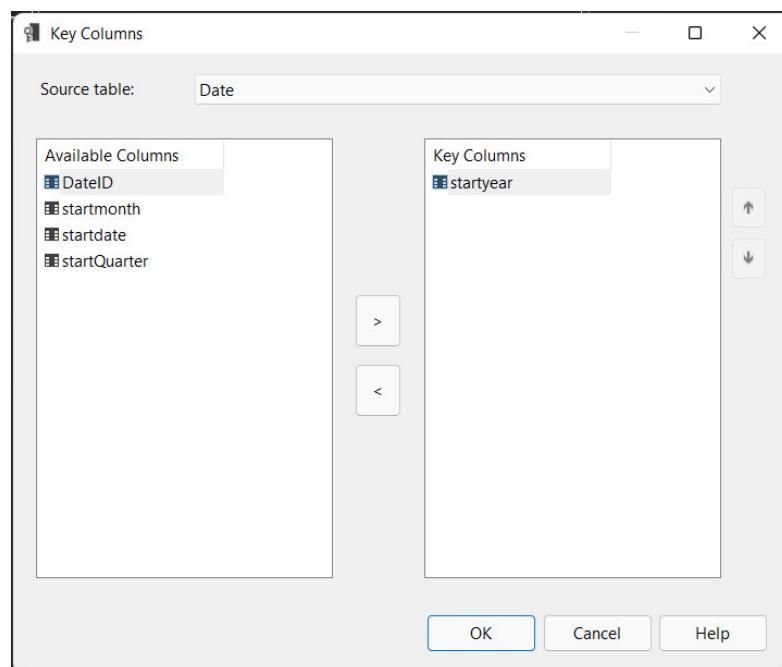
Hình 167. Thực hiện chọn thuộc tính phân cấp cho StartMonth

Bước 5: Tạo tên dòng của thuộc tính *StartMonth*.



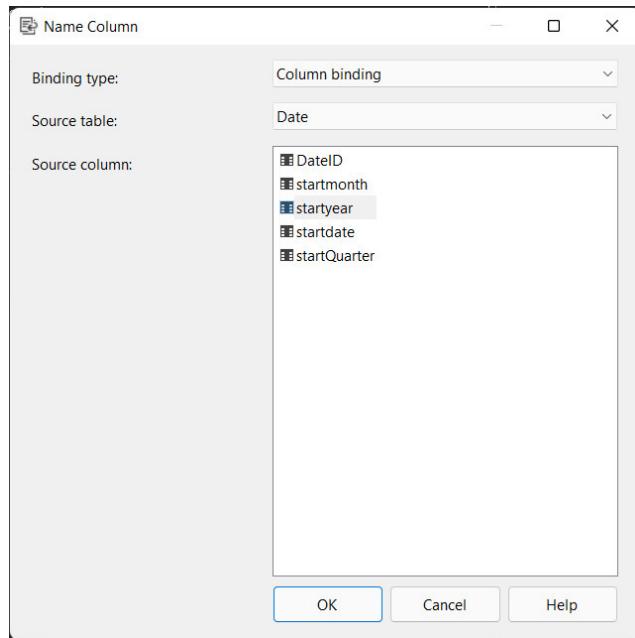
Hình 168. Tạo tên cho thuộc tính StartMonth

Bước 6: Vào Attributes, chỉnh khóa dòng và tên dòng thuộc tính *StartYear*.



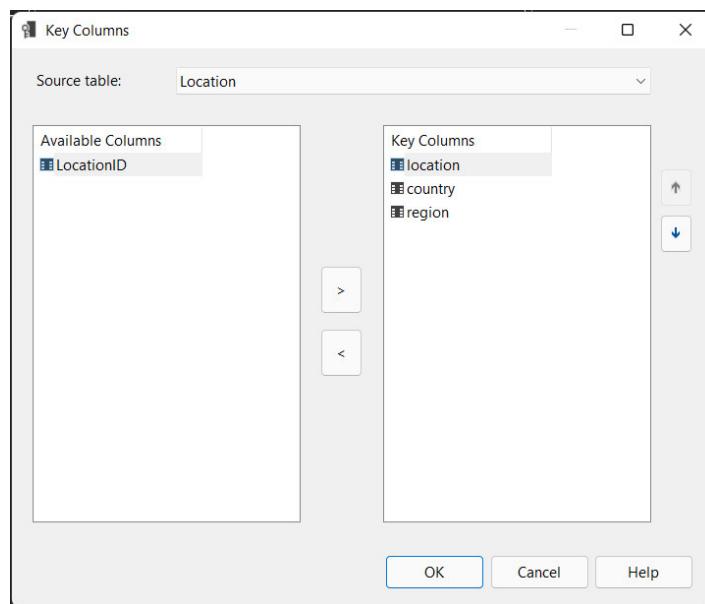
Hình 169. Thực hiện chọn thuộc tính phân cấp cho StartYear

Bước 7: Tạo tên dòng của thuộc tính *StartYear*.

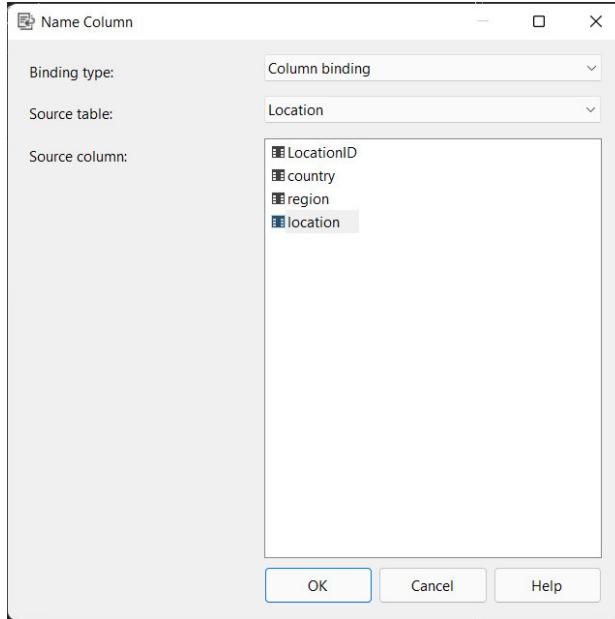


Hình 170. Tạo tên cho thuộc tính *StartYear*

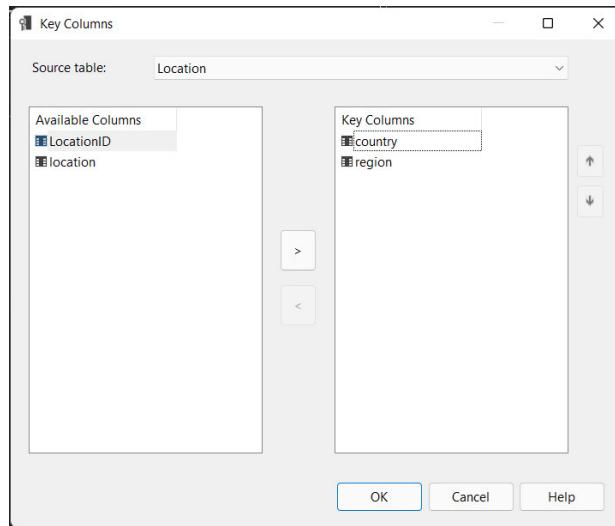
Bước 8: Vào Attributes, chỉnh khóa dòng và tên dòng thuộc tính ***Location***.
(Cấp nhỏ hơn: ***Region, Country***)



Hình 171. Thực hiện chọn thuộc tính phân cấp cho *Location*

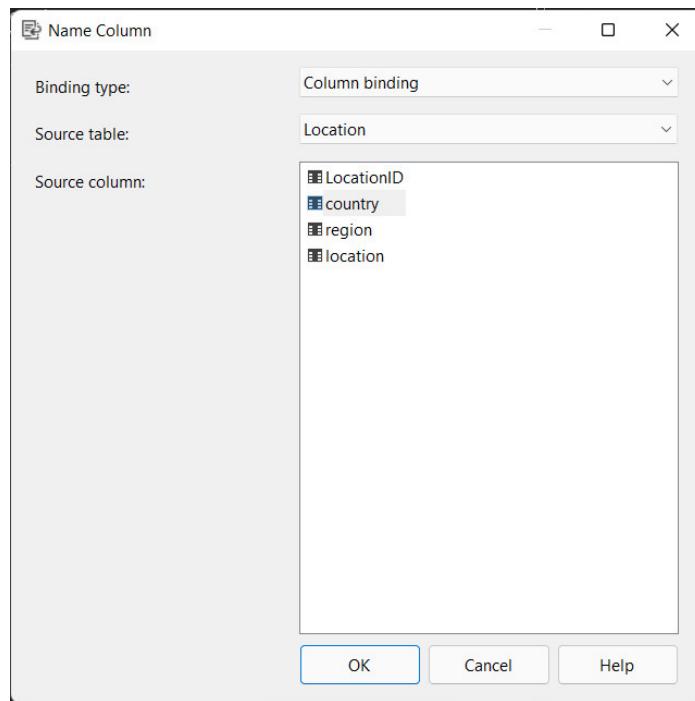
Bước 9: Tạo tên dòng của thuộc tính **Location**

Hình 171 Tạo tên cho thuộc tính Location

Bước 10: Vào Attributes, chỉnh khóa dòng và tên dòng thuộc tính **Country**. (Cấp nhỏ hơn: **Region**)

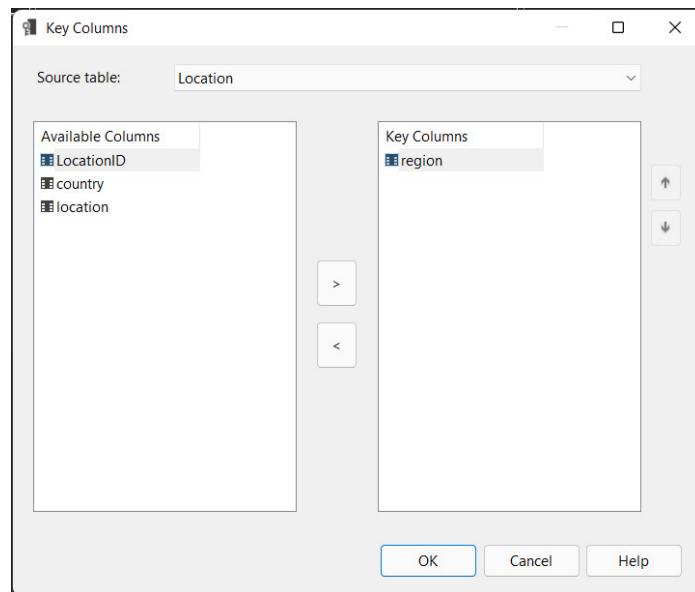
Hình 172. Thực hiện chọn thuộc tính phân cấp cho Country

Bước 11: Tạo tên dòng của thuộc tính **Country**



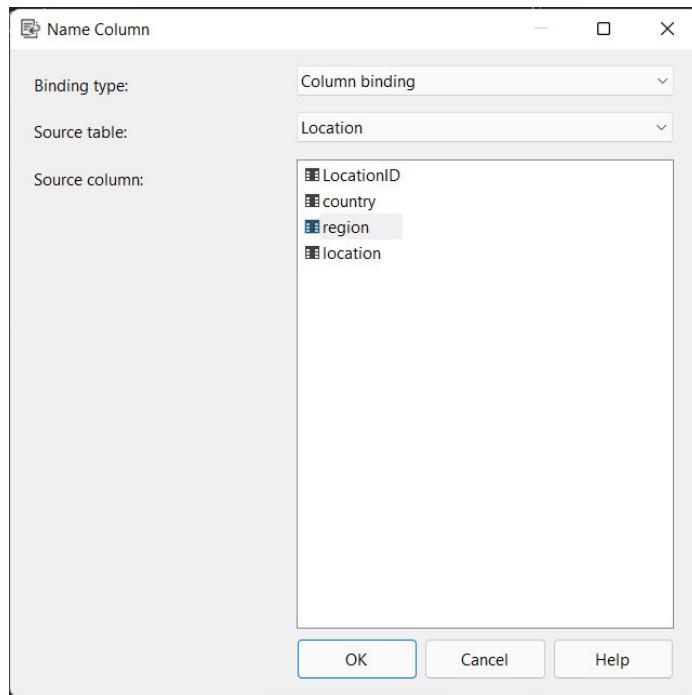
Hình 173. Tạo tên cho thuộc tính Country

Bước 12: Vào Attributes, chỉnh khóa dòng và tên dòng thuộc tính **Region**.



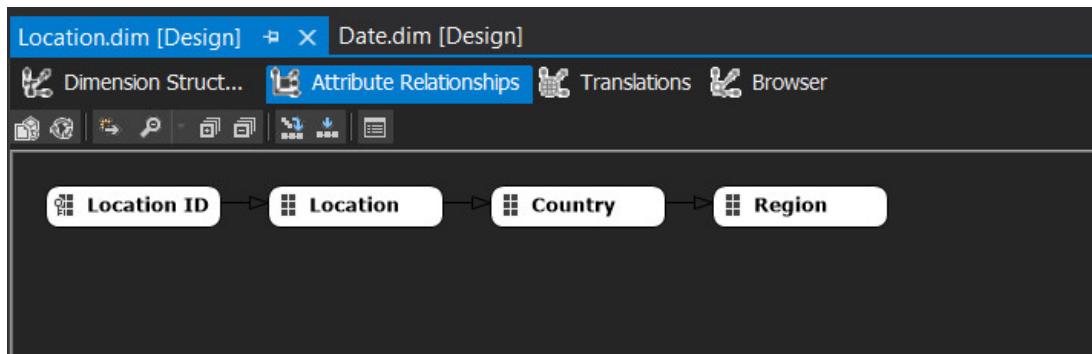
Hình 174. Thực hiện chọn thuộc tính phân cấp cho Region

Bước 13: Tạo tên dòng của thuộc tính **Region**.

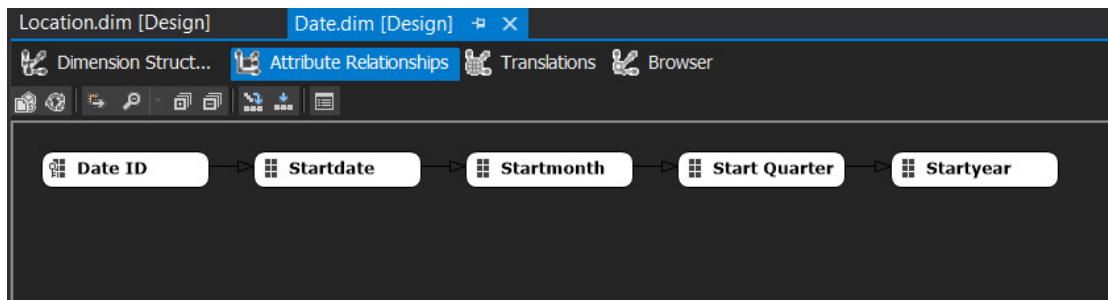


Hình 175. Tạo tên cho thuộc tính Region

Bước 14: Tại tab **Attribute Relationships** của dim LOCATION và dim DATE để liên kết các mối quan hệ phân cấp:

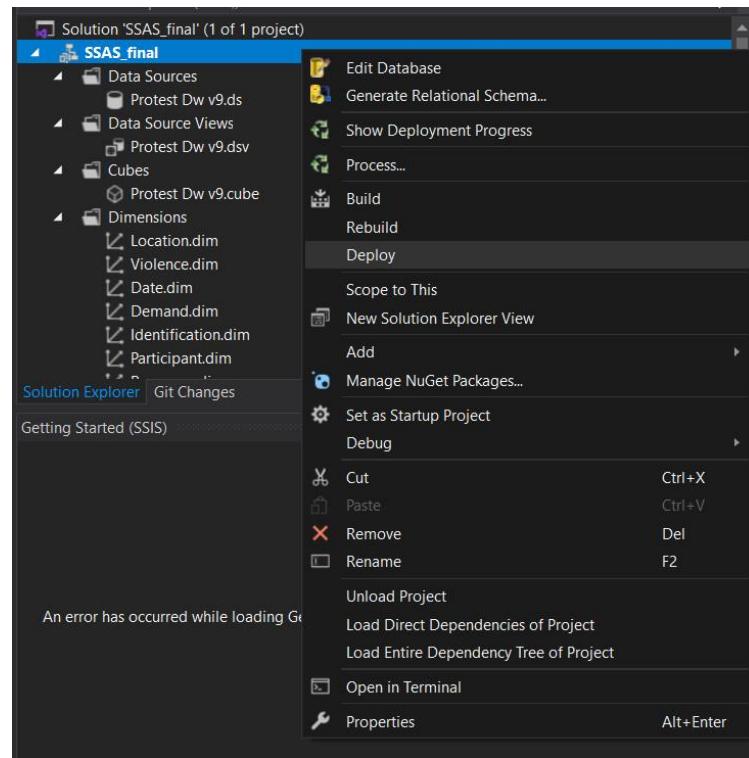


Hình 176.1 Phân cấp thuộc tính theo địa điểm (location)



Hình 176.2 Phân cấp thuộc tính theo thời điểm (date)

Bước 15: Tại thư mục Project > Chọn **Deploy** để chạy Project



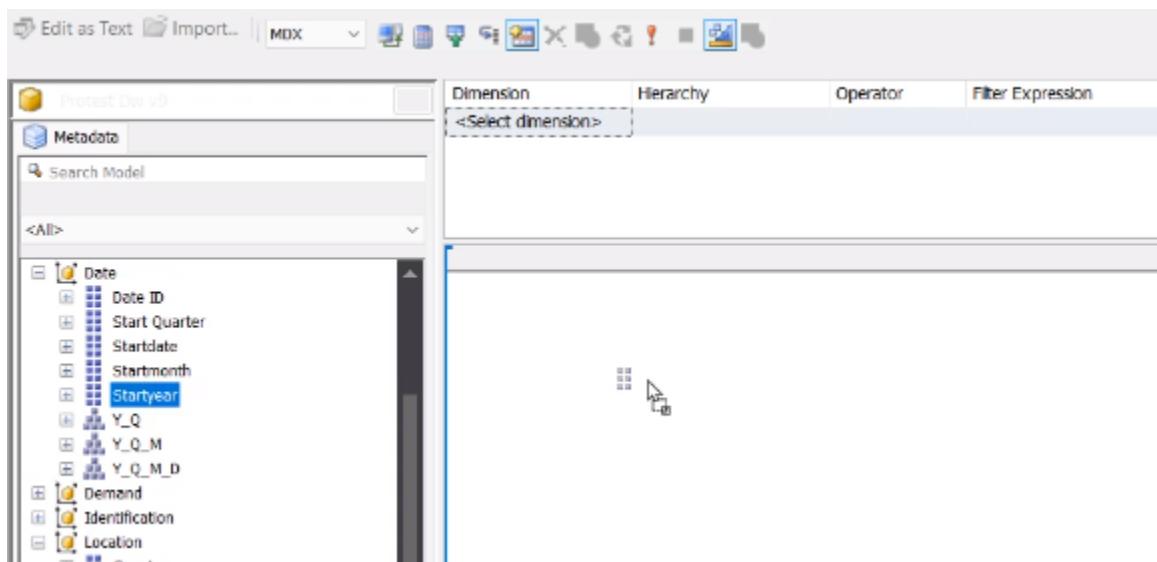
Hình 177. Bước khởi chạy project

2. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ SSAS TRÊN CÁC KHỐI CUBE

2.1 Cho biết tổng số cuộc biểu tình ở Russia theo từng năm.

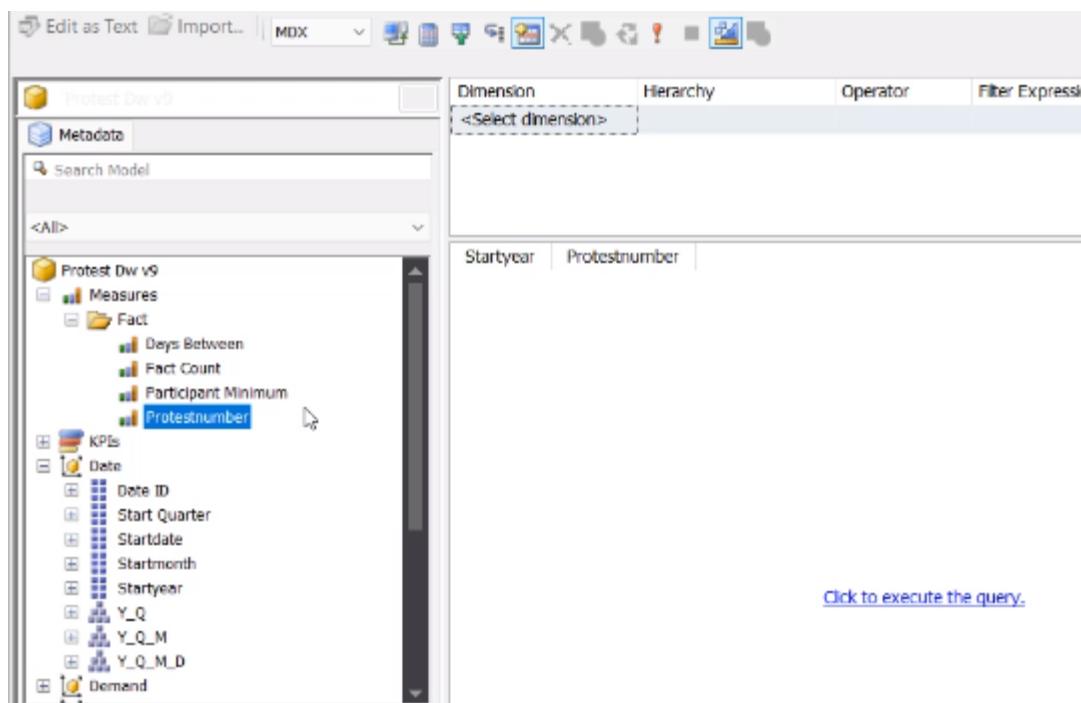
- Quá trình thực hiện:

Bước 1: Lấy thuộc tính ‘*StartYear*’ của bảng *Dim_Date* sang vùng thực thi câu truy vấn



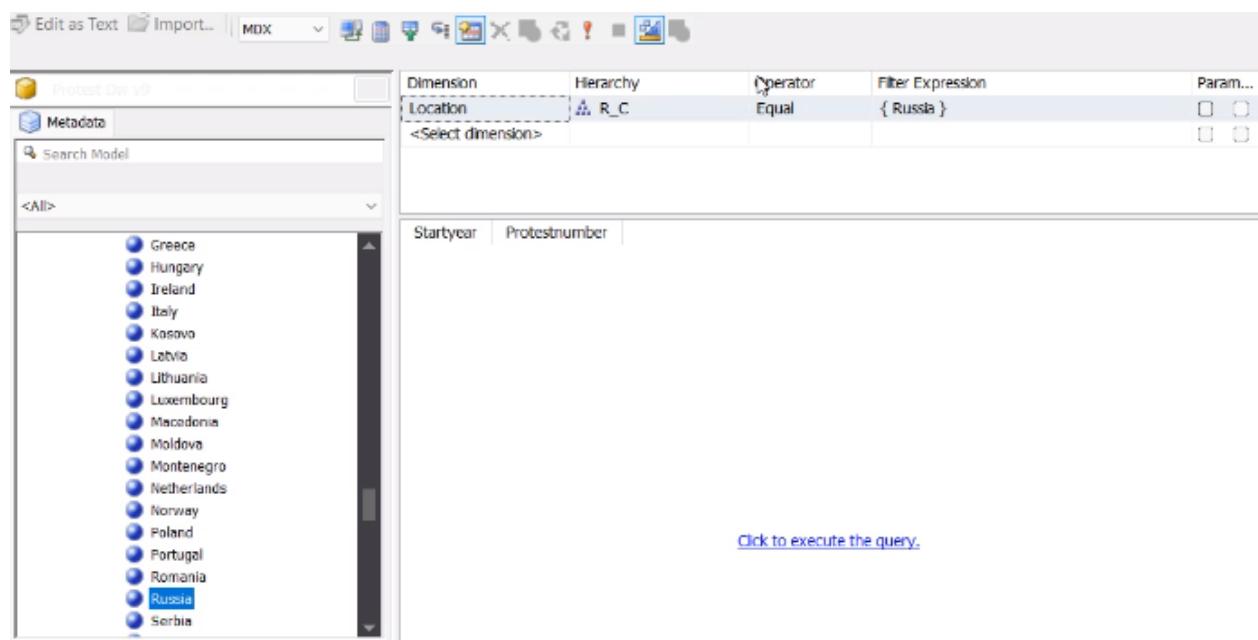
Hình 178. Thực hiện kéo thuộc tính ‘*StartYear*’ để truy vấn

Bước 2: Lấy độ đo ‘*ProtestNumber*’ của bảng *Fact* sang vùng thực thi câu truy vấn



Hình 179. Thực hiện kéo độ đo ‘ProtestNumber’ để truy vấn

Bước 3: Lấy thuộc tính ‘Country’ của bảng **Dim_Location** sang vùng thực làm điều kiện. Lấy điều kiện lọc là {Russia}



Hình 180. Thực hiện kéo thuộc tính ‘Country’ để làm điều kiện

- Kết quả câu truy vấn:

The screenshot shows a filter configuration window at the top with the following details:

Dimension	Hierarchy	Operator	Filter Expression
Location	R_C	Equal	{ Russia }
<Select dimension>			

Below the filter window is a data grid with two columns: Startyear and Protestnumber. The data is as follows:

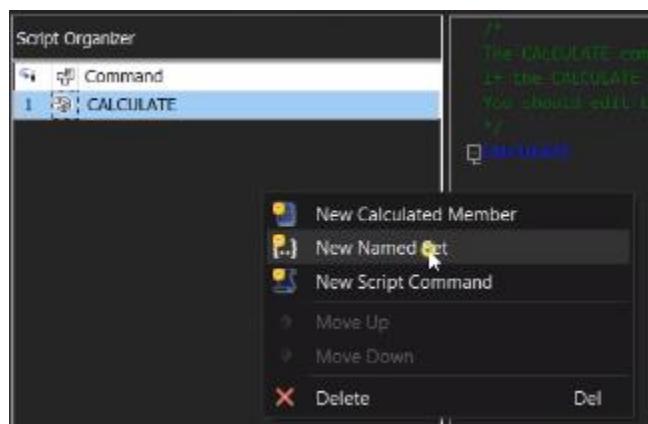
Startyear	Protestnumber
1992	66
1993	28
1994	15
1995	15
1996	21
1997	1
1998	28
1999	3
2000	6
2001	6

Hình 181. Kết quả truy vấn câu 1

2.2 Cho biết thông tin cuộc biểu tình, nước xảy ra biểu tình có tổng số lượng người ước tính tham gia biểu tình cao nhất trong quý 3/1990.

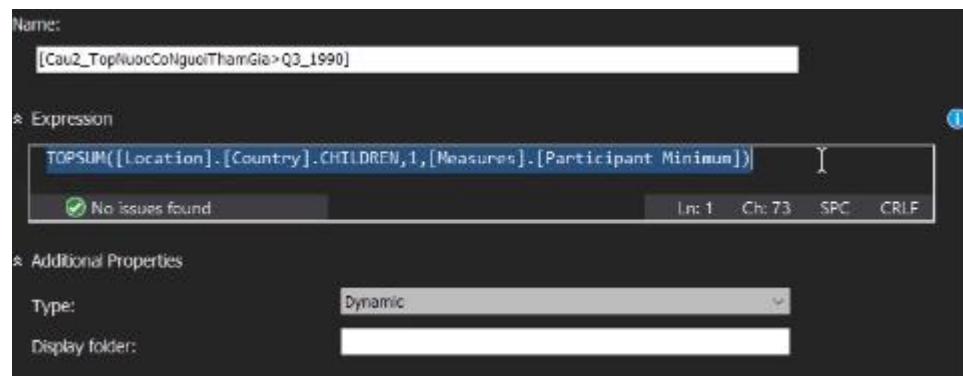
- Quá trình thực hiện:

Bước 1: Tạo một *Name set* mới tại *Script Organizer*



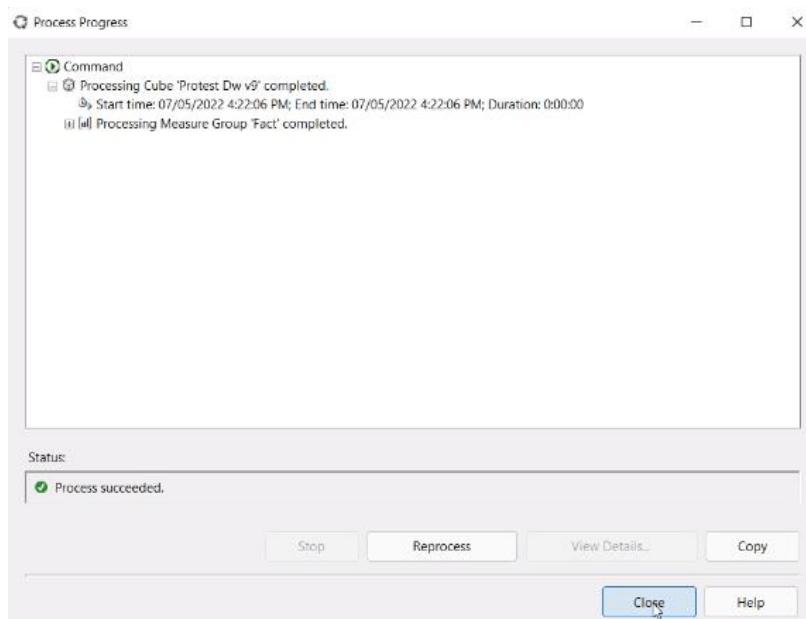
Hình 182. Thực hiện tạo một name set mới

Bước 2: Đặt lại tên cho *Name set* và viết câu truy vấn tại ô *Expression*



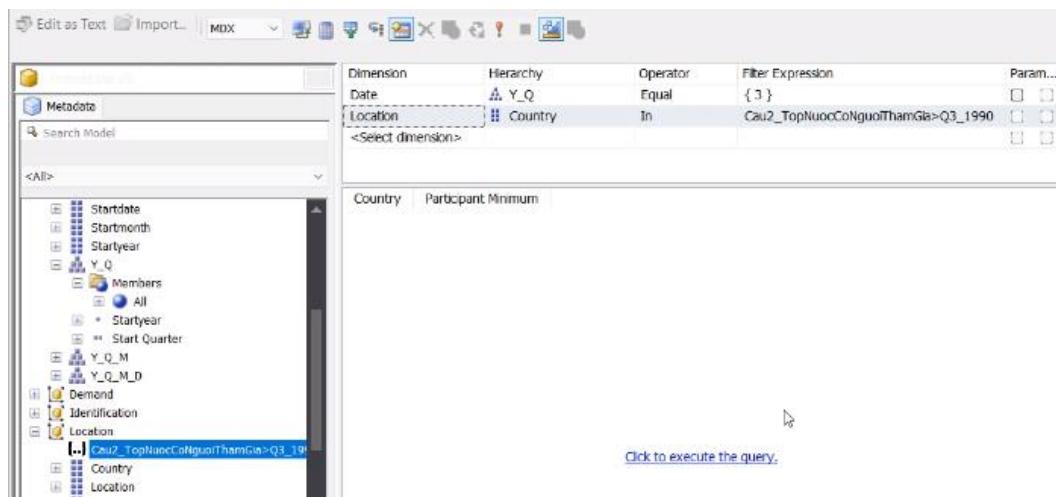
Hình 183. Viết câu lệnh truy vấn tại ô Expression

Bước 3: Tại file **.cube** của project > Thực hiện khởi chạy (**Process**)



Hình 184. Kết quả sau khi thực hiện chạy khởi

Bước 4: Tại mục **Browser** của cube, thực hiện kéo thả thuộc tính **Country**, **Participant Minimum**, điều kiện thời gian và name set vừa tạo ở bước trên vào cửa sổ truy vấn.



Hình 185. Thực hiện thao tác truy vấn câu 2

- Kết quả câu truy vấn:

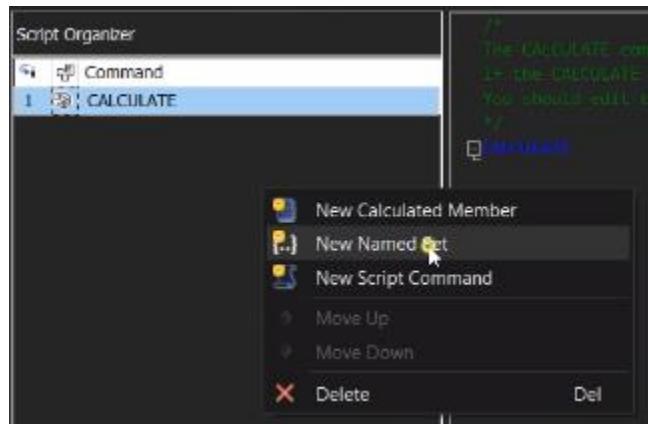
Dimension	Hierarchy	Operator	Filter Expression	Param...
Date	Y_Q	Equal	{3}	
Location	Country	In	Cau2_TopNuocCoNguoiThamGia>Q3_1990	
<Select dimension>				
Country Participant Minimum				
Venezuela 72150				

Hình 186. Kết quả truy vấn câu 2

2.3 Cho biết tên quốc gia có tổng số lượng người tham gia > 500000 tại Châu Á.

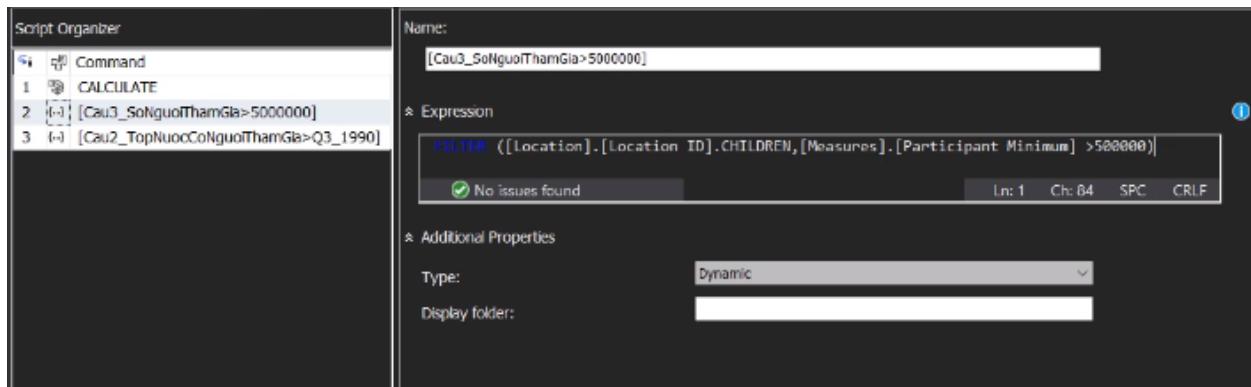
- Quá trình thực hiện:

Bước 1: Tạo một *Name set* mới tại *Script Organizer*



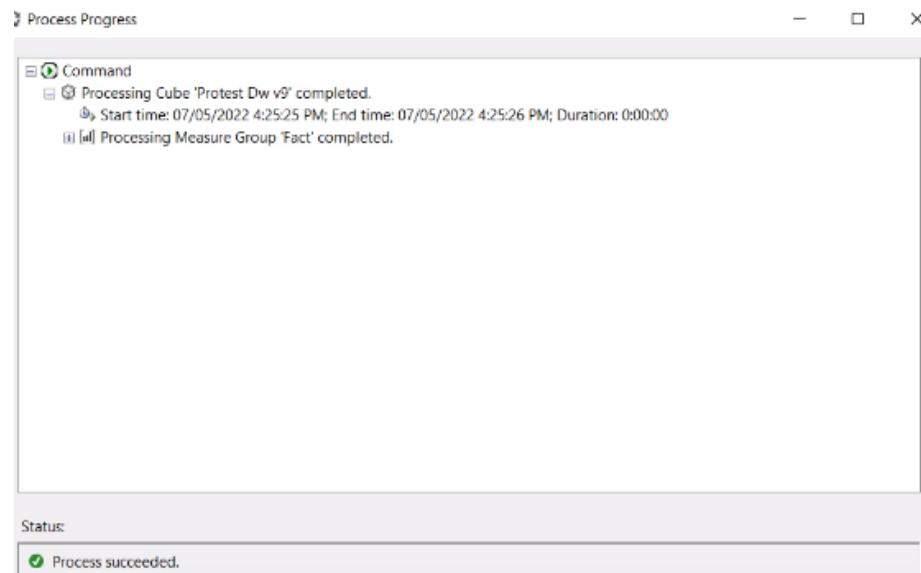
Hình 187. Thực hiện tạo một name set mới

Bước 2: Đặt lại tên cho *Name set* và viết câu truy vấn tại ô *Expression*



Hình 188. Viết câu lệnh truy vấn tại ô Expression

Bước 3: Tại file *.cube* của project > Thực hiện khởi chạy (*Process*)



Hình 189. Kết quả sau khi thực hiện chạy khởi

Bước 4: Tại mục **Browser** của cube, thực hiện kéo thả thuộc tính **LocationID**, **Country**, **Participant Minimum** và name set về điều kiện người tham gia vừa tạo ở bước trên vào cửa sổ truy vấn.

The screenshot shows the MDX query editor with the following query:

```

SELECT
    [Location].[Location ID].Members,
    [Country].[Country].Members,
    [Participant Minimum].[Participant Minimum].Members
FROM [Protestnumber]
WHERE
    [Location].[Location ID].<Select dimension> IN { [Location].[Location ID].<Filter>[Cau2_SoNguoithamGia]>5000000 }
    AND [Country].[Country].<Filter>[Cau2_SoNguoithamGia]>5000000
    AND [Participant Minimum].[Participant Minimum].<Filter>[Cau2_TopNuocCoNguoithamGia]>Q3_199

```

The browser pane on the left shows the cube structure with nodes like 'Cau2_TopNuocCoNguoithamGia>Q3_199' and 'Cau2_SoNguoithamGia>5000000' selected.

Hình 190. Thực hiện thao tác truy vấn câu 3

- Kết quả câu truy vấn:

Dimension	Hierarchy	Operator	Filter Expression	Param...
Location	R_C	Equal	{ Asia }	<input type="checkbox"/> <input type="checkbox"/>
Location	Location ID	In	Cau3_SoNgoiThamGia>5000000	<input type="checkbox"/> <input type="checkbox"/>
<Select dimension>				

Location ID	Country	Participant Minimum
152	Banglades...	8310005
624	China	3002530
1624	India	5270670
1746	Indonesia	2626900
2128	Japan	641950
2561	Malaysia	518858
2897	Nepal	1182703
3108	Pakistan	2123750
3306	Philippines	1811850
3601	South ...	6560407
3807	Taiwan	6454620
3855	Thailand	2687090

Hình 191. Kết quả truy vấn câu 3

2.4 Theo từng tháng, quý, năm liệt kê số lần diễn ra biểu tình tại các nước ở Châu Á.

- Quá trình thực hiện:

Tại mục **Browser** của cube, thực hiện kéo thả thuộc tính **Country**, **StartYear**, **StartQuarter** của bảng Date, **ProtestNumber** và điều kiện Châu lục (Region) vào cửa sổ truy vấn.

Dimension	Hierarchy	Operator	Filter Expression	Param...
Location	R_C	Equal	{ Asia }	<input type="checkbox"/> <input type="checkbox"/>
<Select dimension>				

Country	Startyear	Start Quarter	Startmonth	Partic...
---------	-----------	---------------	------------	-----------

All

 Africa

 Asia

 Central America

 Europe

 MENA

 North America

 Oceania

 South America

 Unknown

[Click to execute the query.](#)

Hình 192. Thực hiện thao tác truy vấn câu 4

- Kết quả câu truy vấn:

Dimension	Hierarchy	Operator	Filter Expression	Para
Location	R_C	Equal	{ Asia }	
<Select dimension>				
Country	Startyear	Start Quarter	Startmonth	Protestnumber
Afghan...	1991	2	5	1
Afghan...	1997	4	12	1
Afghan...	1998	1	3	1
Afghan...	2002	4	11	1
Afghan...	2005	2	5	1
Afghan...	2005	4	10	2
Afghan...	2011	1	1	1
Afghan...	2011	2	5	5
Afghan...	2011	4	11	4
Afghan...	2012	1	2	1
Afghan...	2013	2	5	1
Afghan...	2014	2	6	10
Afghan...	2014	3	7	11
Afghan...	2015	1	3	3
Afghan...	2015	2	4	3
Afghan...	2015	4	11	9
Afghan...	2016	1	3	1
Afghan...	2016	2	5	2

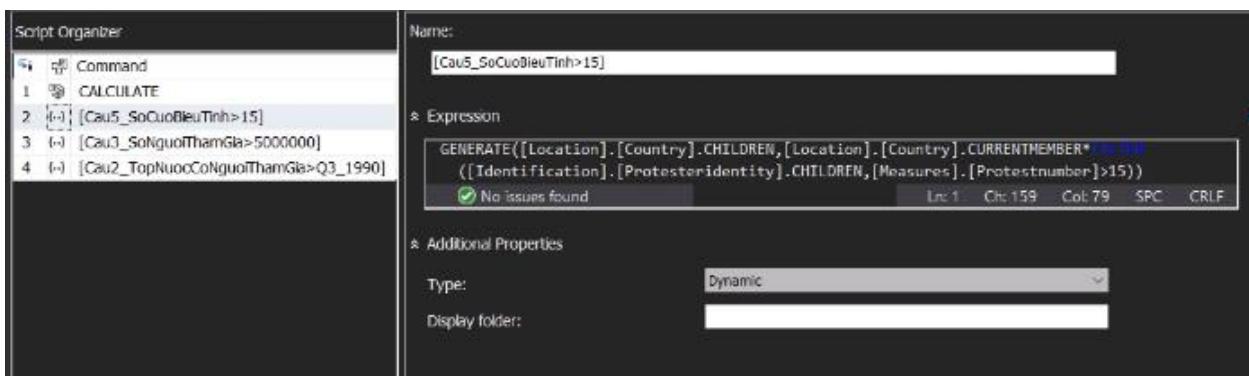
Hình 193. Kết quả truy vấn câu 4

2.5 Cho biết thông tin nhóm người biểu tình và tổng số cuộc biểu tình của cuộc biểu tình đó lớn hơn 15 được đặt theo từng quốc gia.

- Quá trình thực hiện:

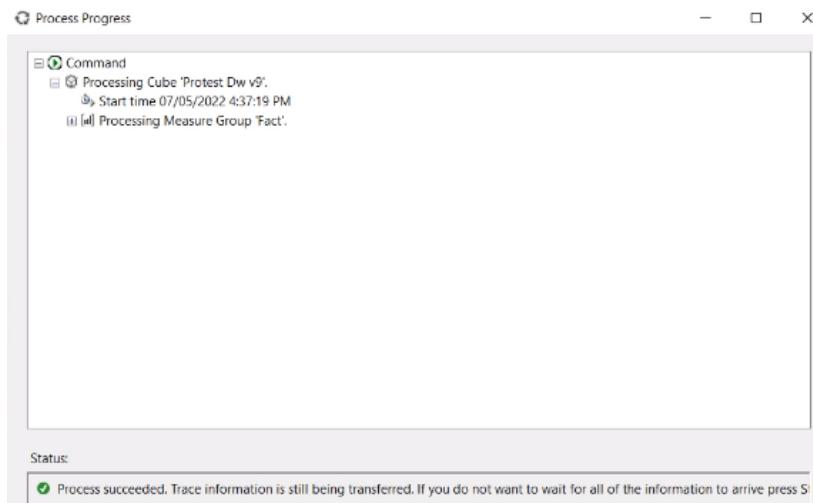
Bước 1: Tạo một *Name set* mới tại *Script Organizer*

Bước 2: Đặt lại tên cho *Name set* và viết câu truy vấn tại ô *Expression*



Hình 194. Viết câu lệnh truy vấn tại ô Expression

Bước 3: Tại file **.cube** của project > Thực hiện khởi chạy (**Process**)



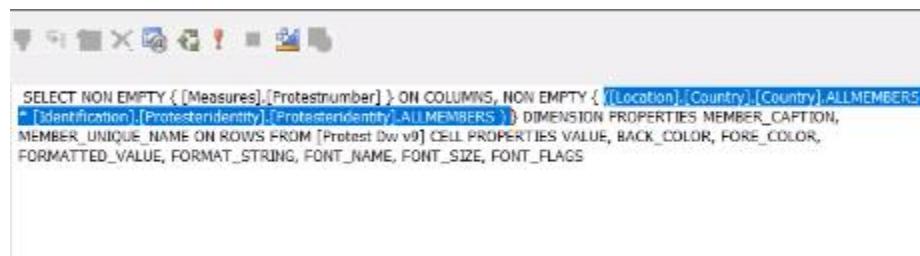
Hình 195. Kết quả sau khi thực hiện chạy khởi

Bước 4: Tại mục **Browser** của cube, thực hiện kéo thả thuộc tính **Country**, **ProtesterIdentity** và **ProtestNumber** vào cửa sổ truy vấn.

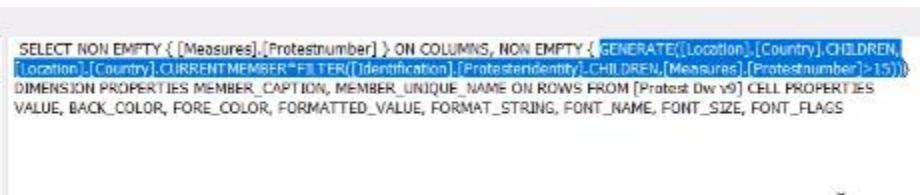
Country	ProtesterIdentity	Protestnumber
Afghan...	11	
Afghani...	ethnic hazaras	5
Afghani...	hazaras	16
Afghani...	human rights a...	2
Afghani...	national islamic f...	2
Afghani...	protesters	31
Afghani...	protesters, sup...	3
Afghani...	protesters, wo...	2
Afghani...	protesters, you...	5
Afghani...	residents	7
Afghani...	students	1
Afghani...	supporters of t...	11
Afghani...	supporters of t...	4
Afghani...	talban supporters	1
Afghani...	university stude...	1
Afghani...	uprising for cha...	5
Afghani...	womens groups	1
Albania	albanians	88
Albania	ethnic greeks	2
Albania	followers of the ...	2
Albania	former politcal ...	1
Albania	leaders and sup...	1
Albania	opposition	162

Hình 196. Thực hiện thao tác truy vấn câu 5

Bước 5: Chọn chức năng ‘**Design Mode**’, chỉnh sửa câu truy vấn thêm điều kiện về số lần biểu tình > Thực hiện câu truy vấn



```
SELECT NON EMPTY { [Measures].[Protestnumber] } ON COLUMNS, NON EMPTY { ([Location].[Country].[Country].ALLMEMBERS * [Identification].[Protesteridentity].[Protesteridentity].ALLMEMBERS) } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [Protest Dw v9] CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

Hình 197. Câu truy vấn 5 ban đầu


```
SELECT NON EMPTY { [Measures].[Protestnumber] } ON COLUMNS, NON EMPTY { GENERATE([Location].[Country].CHILDREN, [Location].[Country].CURRENTMEMBER*FILTER([Identification].[Protesteridentity].CHILDREN,[Measures].[Protestnumber]>15)) } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [Protest Dw v9] CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

Hình 198. Câu truy vấn 5 sau khi chỉnh sửa

- **Kết quả câu truy vấn:**

Country	Protesteridentity	Protestnumber
Afghanistan	hazaras	16
Afghanistan	protesters	31
Albania	albanians	88
Albania	opposition	162
Albania	protesters	28
Albania	students	63
Algeria	judicial activists	58
Algeria	lawyers	22
Algeria	muslim fundamentalists	22
Algeria	policemen	18
Algeria	pro democracy protesters	1741
Algeria	protesters	94
Algeria	protesters; pro democracy	54
Algeria	residents	20

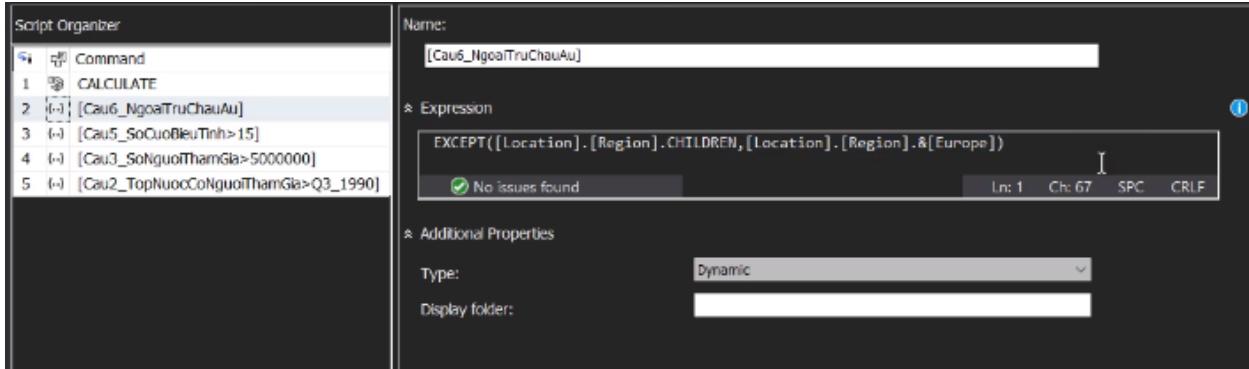
Hình 199. Kết quả truy vấn câu 5

2.6 Với từng châu lục liệt kê số ngày biểu tình theo từng năm, trừ Châu Âu.

- **Quá trình thực hiện:**

Bước 1: Tạo một *Name set* mới tại *Script Organizer*

Bước 2: Đặt lại tên cho *Name set* và viết câu truy vấn tại ô *Expression*



Hình 200. Viết câu lệnh truy vấn tại ô Expression

Bước 3: Tại file *.cube* của project > Thực hiện khởi chạy (*Process*)



Hình 201. Kết quả sau khi thực hiện chạy khởi

Bước 4: Tại mục *Browser* của cube, thực hiện kéo thả thuộc tính *Region*, *StartYear*, *DayBetween* và name set vừa tạo ở bước trên vào cửa sổ truy vấn.

The screenshot shows the Microsoft Analysis Services (SSAS) MDX query editor. On the left, there's a tree view of the metadata model with nodes like Y_Q_M, Y_Q_M_D, Demand, Identification, Location, and various sub-nodes. The main area has a table for defining filters:

Dimension	Hierarchy	Operator	Filter Expression	Param...
Location	# Region	In	Cau6_NgoaiTruChauAu	
<Select dimension>				

Below this is a results grid with columns: Region, Startyear, Days Between. The first row shows the filter applied. A link "Click to execute the query." is at the bottom.

Hình 202. Thực hiện thao tác truy vấn câu 6

- Kết quả câu truy vấn:

The screenshot shows the results of the query from Figure 202. The table has three columns: Region, Startyear, and Days Between. The data is as follows:

Region	Startyear	Days Between
Africa	1990	125
Africa	1991	157
Africa	1992	99
Africa	1993	115
Africa	1994	106
Africa	1995	69
Africa	1996	84
Africa	1997	167
Africa	1998	356
Africa	1999	72
Africa	2000	139
Africa	2001	122
Africa	2002	129
Africa	2003	98
Africa	2004	171
Africa	2005	155
Africa	2006	106
Africa	2007	100
Africa	2008	98
Africa	2009	136

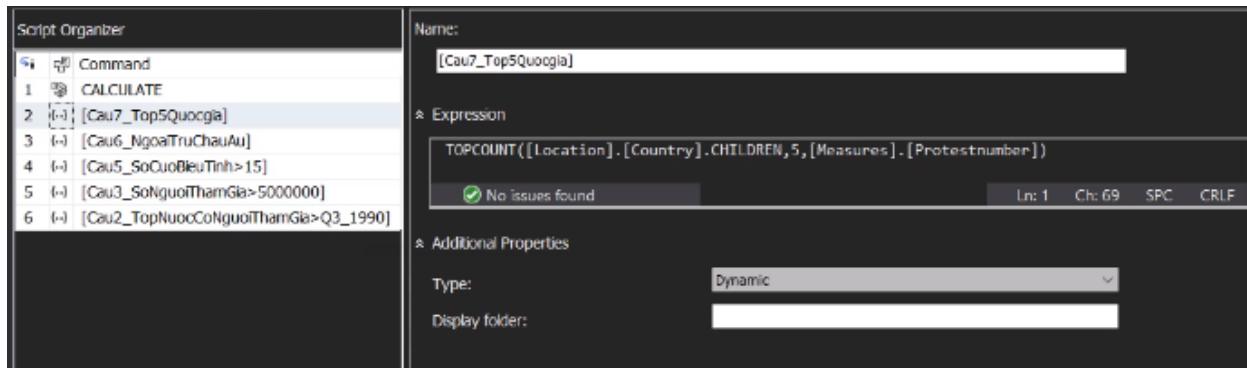
Hình 203. Kết quả truy vấn câu 6

2.7 Đưa ra top 5 quốc gia có số cuộc biểu tình diễn ra nhiều nhất trong năm 2017.

- Quá trình thực hiện:

Bước 1: Tạo một *Name set* mới tại *Script Organizer*

Bước 2: Đặt lại tên cho *Name set* và viết câu truy vấn tại ô *Expression*



Hình 204. Viết câu lệnh truy vấn tại ô Expression

Bước 3: Tại file *.cube* của project > Thực hiện khởi chạy (*Process*)

Bước 4: Tại mục *Browser* của cube, thực hiện kéo thả thuộc tính *Country*, *ProtestNumber*, điều kiện thời gian và name set vừa tạo ở bước trên vào cửa sổ truy vấn.

Dimension	Hierarchy	Operator	Filter Expression
Location	Y_Q	Equal	{2017}
<Select dimension>			Cau7_Top5Quocgia

Hình 205. Thực hiện thao tác truy vấn câu 7

- Kết quả câu truy vấn:

Country	Protestnumber
Dominican Republic	300
Germany	1081
Nigeria	595
Romania	561
Spain	300

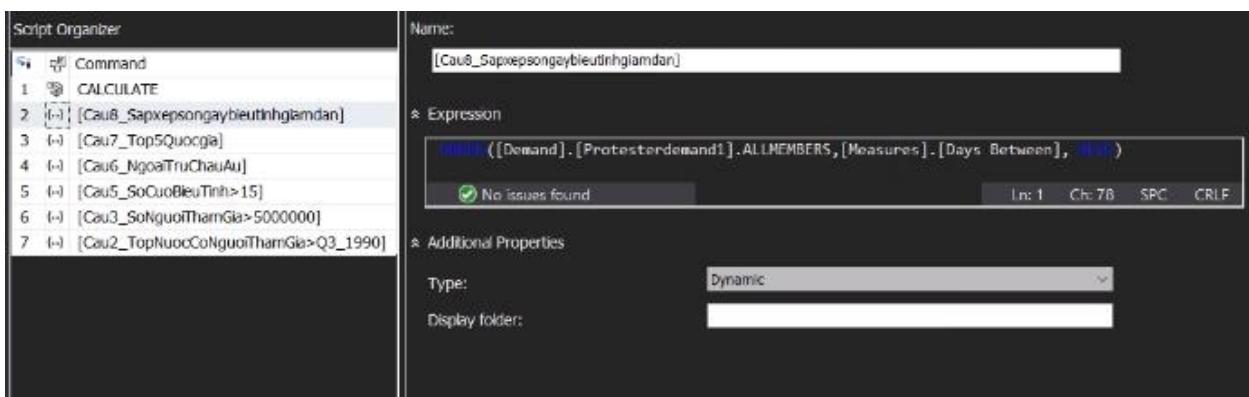
Hình 206. Kết quả truy vấn câu 7

2.8 Thống kê số cuộc biểu tình, số người tham gia và số ngày biểu tình của mục đích biểu tình thứ nhất, sắp xếp số ngày biểu tình theo thứ tự giảm dần.

- Quá trình thực hiện:

Bước 1: Tạo một *Name set* mới tại *Script Organizer*

Bước 2: Đặt lại tên cho *Name set* và viết câu truy vấn tại ô *Expression*



Hình 207. Viết câu lệnh truy vấn tại ô Expression

Bước 3: Tại file *.cube* của project > Thực hiện khởi chạy (*Process*)

Bước 4: Tại mục *Browser* của cube, thực hiện kéo thả thuộc tính *ProtesterDemand1*, *ProtestNumber*, *Participant Minimum*, *Day Between* vào cửa sổ truy vấn.

Dimension	Hierarchy	Operator	Filter Expression
<Select dimension>			
Protesterdemand1	Protestnumber	Participant Minimum	Days Between
	3	5000	1
labor wage dispute	13875	23656565	6310
land farm issue	5485	834067	2299
police brutality	6495	3158457	1638
political behavior, ...	81650	82108092	23238
price increases, ta...	8190	8506046	2531
removal of politician	8559	9596209	2754
social restrictions	2724	1117130	967

Hình 208. Thực hiện thao tác truy vấn câu 8

Bước 5: Chọn chức năng ‘**Design Mode**’, chỉnh sửa câu truy vấn thêm điều kiện về sắp xếp dữ liệu > Thực hiện câu truy vấn

```
SELECT NON EMPTY { [Measures].[Protestnumber], [Measures].[Participant Minimum], [Measures].[Days Between] } ON COLUMNS, NON EMPTY { {[Demand].[Protesterdemand1].ALLMEMBERS} } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [Protest Dw v9] CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

Hình 209. Câu truy vấn 8 ban đầu

```
SELECT NON EMPTY { [Measures].[Protestnumber], [Measures].[Participant Minimum], [Measures].[Days Between] } ON COLUMNS, NON EMPTY { ORDER([Demand].[Protesterdemand1].ALLMEMBERS,[Measures].[Days Between], DESC) } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [Protest Dw v9] CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

Hình 210. Câu truy vấn 8 sau khi chỉnh sửa

- Kết quả câu truy vấn:

Protesterdemand1 (null)	Protestnumber	Participant Minimum	Days Between
political behavior, process	81650	82108092	23238
labor wage dispute	13875	23656565	6310
removal of politician	8559	9596209	2754
price increases, tax policy	8190	8506046	2531
land farm issue	5485	834067	2299
police brutality	6495	3158457	1638
social restrictions	2724	1117130	967
3	5000	1	

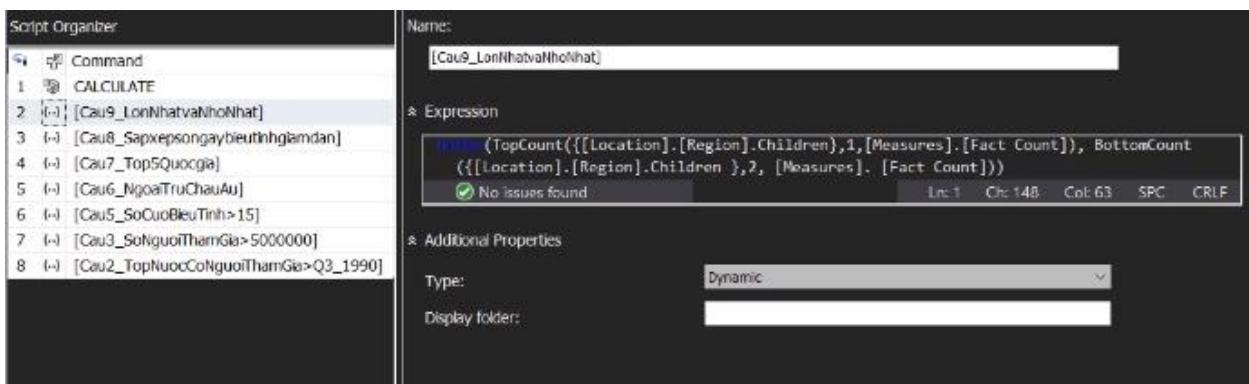
Hình 211. Kết quả truy vấn câu 8

2.9 Cho biết quốc gia có số ngày biểu tình cao nhất và quốc gia có số lần biểu tình thấp nhất.

- Quá trình thực hiện:

Bước 1: Tạo một *Name set* mới tại *Script Organizer*

Bước 2: Đặt lại tên cho *Name set* và viết câu truy vấn tại ô *Expression*



Hình 212. Viết câu lệnh truy vấn tại ô Expression

Bước 3: Tại file *.cube* của project > Thực hiện khởi chạy (*Process*)

Bước 4: Tại mục *Browser* của cube, thực hiện kéo thả thuộc tính *Region*, *FactCount* vào cửa sổ truy vấn.

Region	Fact Count
Africa	3184
Asia	3126
Central Amer...	451
Europe	4994
MENA	1260
North America	527
Oceania	38
South Amer...	1659

Hình 213. Thực hiện thao tác truy vấn câu 9

Bước 5: Chọn chức năng ‘**Design Mode**’, chỉnh sửa câu truy vấn thêm điều kiện top 2 > Thực hiện câu truy vấn

```

SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS, NON EMPTY { ([Location].[Region].[Region].ALLMEMBERS) }
DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [Protest Dw v9] CELL PROPERTIES
VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS

```

Hình 214.1 Câu truy vấn 9 ban đầu

```

SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS, NON EMPTY { Union(TopCount{([Location].[Region].Children),1,
[Measures].[Fact Count]),BottomCount{([Location].[Region].Children ),2, [Measures].[Fact Count])} } DIMENSION PROPERTIES
MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [Protest Dw v9] CELL PROPERTIES VALUE, BACK_COLOR,
FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS

```

Hình 214.2 Câu truy vấn 9 sau khi chỉnh sửa

- Kết quả câu truy vấn:

Region	Fact Count
Europe	4994
Oceania	38

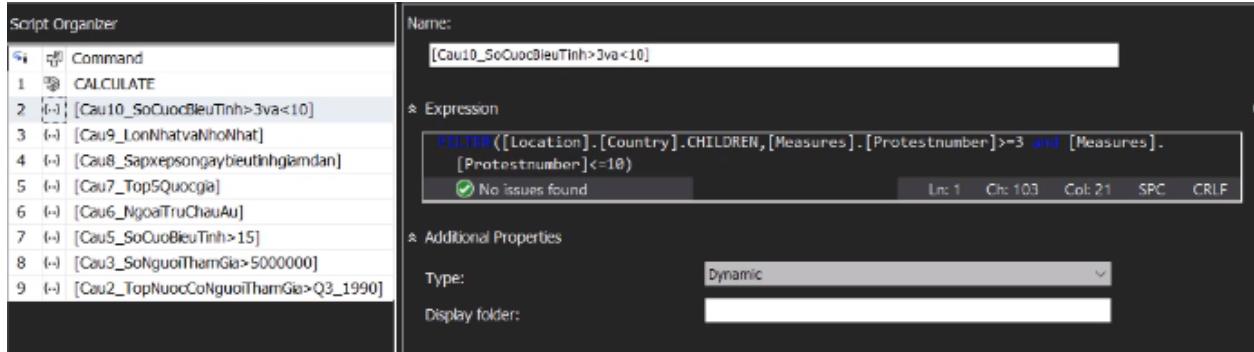
Hình 215. Kết quả truy vấn câu 9

2.10 Liệt kê những quốc gia có số cuộc biểu tình từ 3 đến 10.

- **Quá trình thực hiện:**

Bước 1: Tạo một *Name set* mới tại *Script Organizer*

Bước 2: Đặt lại tên cho *Name set* và viết câu truy vấn tại ô *Expression*



Hình 216. Viết câu lệnh truy vấn tại ô Expression

Bước 3: Tại file *.cube* của project > Thực hiện khởi chạy (*Process*)

Bước 4: Tại mục *Browser* của cube, thực hiện kéo thả thuộc tính *Country*, *ProtestNumber* và name set vừa tạo vào cửa sổ truy vấn.

Dimension	Hierarchy	Operator	Filter Expression	Param...
Location	Country	In	Cau10_SoCuocBieuTinh>3va<10	<input type="checkbox"/> <input type="checkbox"/>
<Select dimension>				

Country Protestnumber

[Click to execute the query.](#)

Hình 217. Thực hiện thao tác truy vấn câu 10

- Kết quả câu truy vấn:

Country	Protestnumber
Cape Verde	3
Czechoslovakia	10
Equatorial Guinea	4
Eritrea	3
Germany West	3
Luxembourg	4
North Korea	9
Norway	7
Serbia and Monte...	3
Turkmenistan	4
United Arab Emirate	3

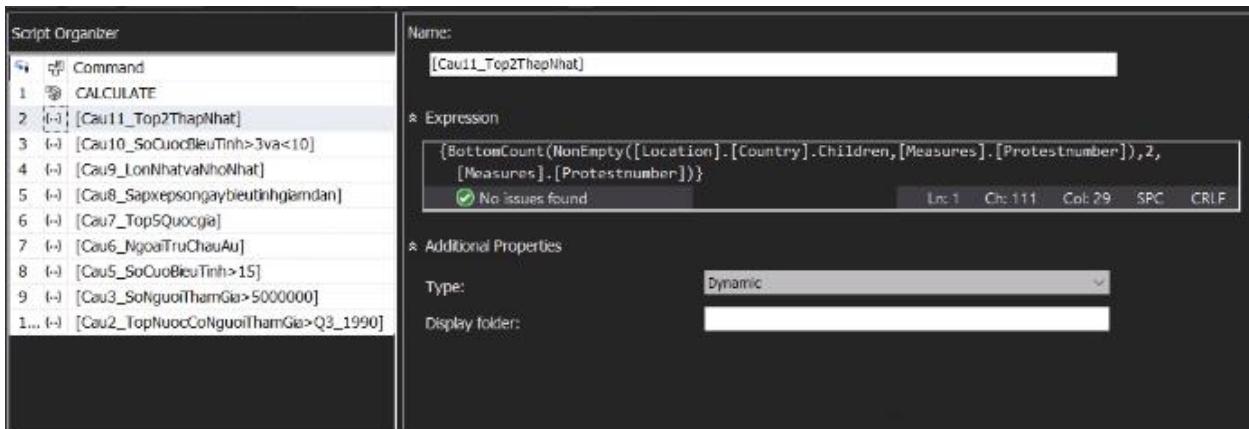
Hình 218. Kết quả truy vấn câu 10

2.11 Xuất ra top 2 những thành phố thuộc Châu Phi có số lần biểu tình thấp nhất trong năm 2018.

- Quá trình thực hiện:

Bước 1: Tạo một *Name set* mới tại *Script Organizer*

Bước 2: Đặt lại tên cho *Name set* và viết câu truy vấn tại ô *Expression*



Hình 219. Viết câu lệnh truy vấn tại ô Expression

Bước 3: Tại file *.cube* của project > Thực hiện khởi chạy (*Process*)

Bước 4: Tại mục **Browser** của cube, thực hiện kéo thả thuộc tính **Country**, **ProtestNumber** và name set vừa tạo ở bước trên vào cửa sổ truy vấn.

Dimension	Hierarchy	Operator	Filter Expression
Location	Region	Equal	{ Africa }
Date	Startyear	Equal	{ 2018 }
Location	Country	In	Cau11_Top2ThapNhat

Country	Protestnumber
Somalia	1
South Africa	1

Hình 220. Thực hiện thao tác truy vấn câu

- Kết quả câu truy vấn:

Country	Protestnumber
Somalia	1
South Africa	1

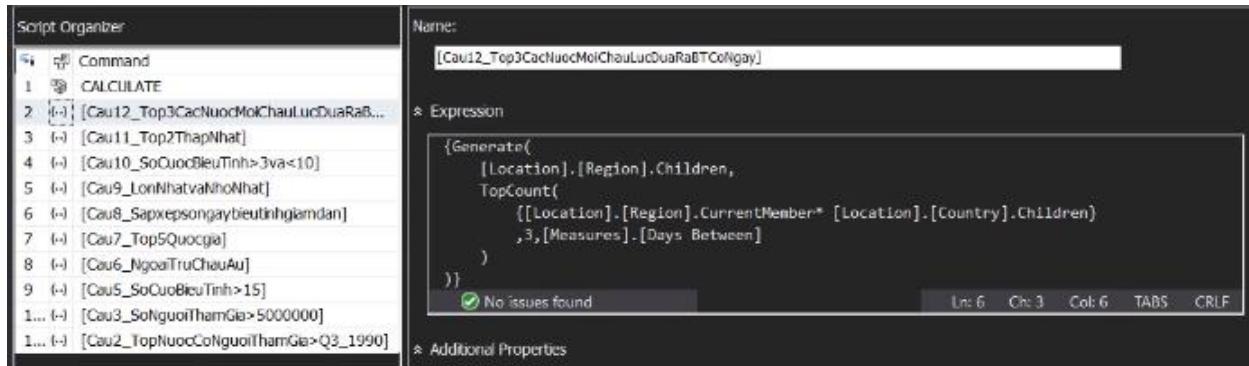
Hình 221. Kết quả truy vấn câu 11

2.12 Với mỗi châu lục, đưa ra 3 quốc gia có tổng số ngày biểu tình cao nhất.

- Quá trình thực hiện:

Bước 1: Tạo một **Name set** mới tại **Script Organizer**

Bước 2: Đặt lại tên cho **Name set** và viết câu truy vấn tại ô **Expression**



Hình 222. Viết câu lệnh truy vấn tại ô Expression

Bước 3: Tại file **.cube** của project > Thực hiện khởi chạy (**Process**)

Bước 4: Tại mục **Browser** của cube, thực hiện kéo thả thuộc tính **Country**, **Region**, **Days Between** vào cửa sổ truy vấn.

Region	Country	Days Between
Africa	Angola	24
Africa	Benin	65
Africa	Botswana	43
Africa	Burkina...	84
Africa	Burundi	124
Africa	Camero...	68
Africa	Cape V...	2
Africa	Central ...	72
Africa	Chad	17
Africa	Comoros	96
Africa	Congo ...	41
Africa	Congo ...	87
Africa	Djibouti	12
Africa	Equato...	4
Africa	Eritrea	5
Africa	Ethiopia	53

Hình 223. Thực hiện thao tác truy vấn câu 12

Bước 5: Chọn chức năng ‘**Design Mode**’, chỉnh sửa câu truy vấn thêm điều kiện về sắp xếp dữ liệu > Thực hiện câu truy vấn

```
SELECT NON EMPTY { [Measures].[Days Between] } ON COLUMNS, NON EMPTY { {[Location].[Region].[Region].[ALLMEMBERS]} * {[Location].[Country].[Country].[ALLMEMBERS]} } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS
FROM [Protest Dw v9] CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING,
FONT_NAME, FONT_SIZE, FONT_FLAGS
```

Hình 224.1 Câu truy vấn 12 ban đầu

```

SELECT NON EMPTY { [Measures].[Days Between] } ON COLUMNS, NON EMPTY {{Generate(
    [Location].[Region].Children,
    TopCount(
        {[Location].[Region].CurrentMember} * [Location].[Country].Children
        ,3,[Measures].[Days Between]
    )
)}}
} DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE, NAME ON ROWS FROM [Protest Dw v9] CELL PROPERTIES
VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS

```

Hình 224.2 Câu truy vấn 12 sau khi chỉnh sửa

- Kết quả câu truy vấn:

Region	Country	Days Between
Africa	Namibia	1033
Africa	Kenya	451
Africa	South ...	298
Asia	India	1423
Asia	Thailand	1228
Asia	China	893
Central America	Nicaragua	479
Central America	Guatem...	202
Central America	Honduras	171
Europe	Greece	1295
Europe	France	1263
Europe	United ...	1154
MENA	Morocco	1120
MENA	Syria	702

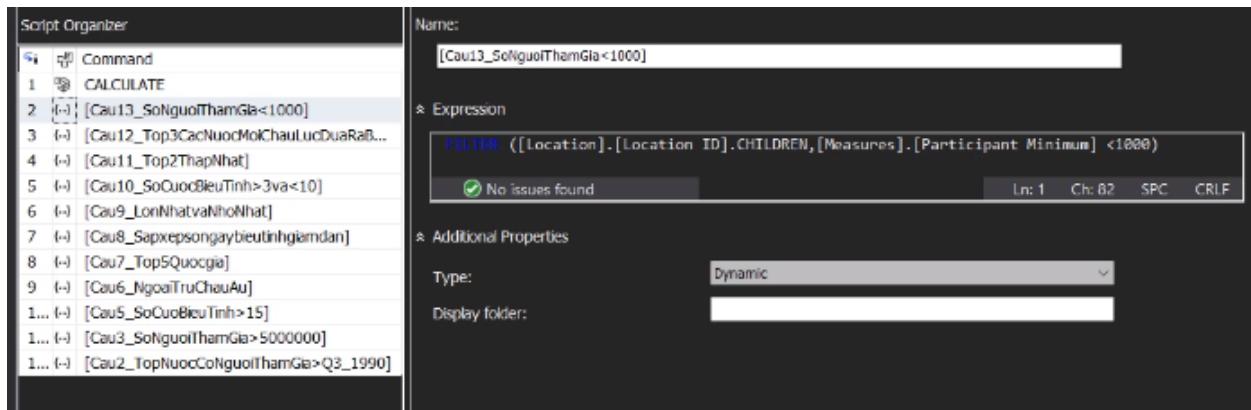
Hình 225. Kết quả truy vấn câu 12

2.13 Liệt kê những cuộc biểu tình ở các quốc gia tại Bắc Mỹ, xảy ra từ năm 1998 đến 2002 với số người tham gia < 1000

- Quá trình thực hiện:

Bước 1: Tạo một *Name set* mới tại *Script Organizer*

Bước 2: Đặt lại tên cho *Name set* và viết câu truy vấn tại ô *Expression*



Hình 226. Viết câu lệnh truy vấn tại ô Expression

Bước 3: Tại file **.cube** của project > Thực hiện khởi chạy (**Process**)

Bước 4: Tại mục **Browser** của cube, thực hiện kéo thả thuộc tính **Country**, **LocationID**, **StartDate**, **ParticipantMinimum** và name set vừa tạo ở bước trên vào cửa sổ truy vấn.

Dimension	Hierarchy	Operator	Filter Expression
Date	# Startyear	Equal	{ 1998, 1999, 2000 }
Location	# Region	Equal	{ North America }
Location	# Location ID	In	Cau13_SoNguoiThamGia<1000

Hình 227. Thực hiện thao tác truy vấn câu

- Kết quả câu truy vấn:

Location ID	Country	Startdate	Participant Minimum
915	Cuba	1999-06-15	50
964	Dominican Republic	1998-03-11	100
964	Dominican Republic	1998-04-28	50
964	Dominican Republic	1998-07-06	50
964	Dominican Republic	1998-10-22	100
964	Dominican Republic	1999-03-15	100
964	Dominican Republic	1999-03-19	100
964	Dominican Republic	2000-03-30	100
2120	Jamaica	1998-09-24	100
2120	Jamaica	1999-04-19	100
2120	Jamaica	1999-06-14	100

Hình 228. Kết quả truy vấn câu 12

2.14 **Liệt kê những quốc gia có các cuộc biểu tình diễn ra với hình thức ôn hòa vào năm 2019, cũng như số ngày diễn ra cuộc biểu tình đó.**

- **Quá trình thực hiện:**

Bước 1: Lấy thuộc tính *StartDate*, *Country*, *DaysBetween*, điều kiện thời gian và điều kiện về *ProtesterViolence* vùng thực thi câu truy vấn

The screenshot shows the Analysis Manager interface with a query editor and a navigation pane. The query editor displays the following MDX code:

```

Edit as Text Import... MDX 
Dimension Hierarchy Operator Filter Expression
Date Startyear Equal { 2019 }
Violence Protesterviolence Equal { peaceful protest }

<Select dimension>

```

The navigation pane on the left shows the cube structure:

- Response ID
- Stateresponse1
- Stateresponse2
- Stateresponse3
- Stateresponse4
- Stateresponse5
- Stateresponse6
- Stateresponse7
- Violence
 - Protesterviolence
 - Members
 - Protesterviolence
 - Member Properties
 - peaceful protest
 - violent protest
 - Unknown
 - Violence ID

Hình 229. Thực hiện kéo các thuộc tính để truy vấn

- Kết quả câu truy vấn:

Dimension	Hierarchy	Operator	Filter Expression
Date	# Startyear	Equal	{ 2019 }
Violence	# Protest/violence	Equal	{ peaceful protest }
<Select dimension>			
Startdate	Country	Days Between	
2019-01-01	India	1	
2019-01-01	Peru	1	
2019-01-02	Madaga...	1	
2019-01-03	Bangkak...	1	
2019-01-03	Madaga...	1	
2019-01-05	Hungary	1	
2019-01-05	Serbia	1	
2019-01-06	Sudan	1	
2019-01-07	Canada	2	
2019-01-07	United ...	1	
2019-01-08	Canada	1	
2019-01-08	Sudan	4	
2019-01-08	Thailand	1	
2019-01-10	Belgium	1	
2019-01-10	Congo ...	1	
2019-01-10	Mongolia	1	

Hình 230. Kết quả truy vấn câu 14

2.15 Cho biết số ngày diễn ra những cuộc biểu tình có sử dụng bạo lực vũ trang cũng như phản ứng đầu tiên của chính phủ đối với cuộc biểu tình đó tại Mexico, sắp xếp theo số ngày biểu tình tăng dần.

- Quá trình thực hiện:

Bước 1: Tạo một *Name set* mới tại *Script Organizer*

Bước 2: Đặt lại tên cho *Name set* và viết câu truy vấn tại ô *Expression*

The screenshot shows the Script Organizer interface. On the left, a list of scripts is visible, including 'Cau13_SoNguoiThamGia<1000', 'Cau12_Top3CaocNuocNoiChauLucDuaRaBTCoNg...', 'Cau11_Top2ThapNhat', 'Cau10_SoCuocBieuTinh>3va<10', 'Cau9_LonNhatVaNhoNhat', 'Cau8_Sapxepsongaybieutinhgiamdan', 'Cau7_Top5Quocgia', 'Cau6_NgoaiTruChauAu', 'Cau14_songaybieutinh>3', 'Cau5_SoCuoBieuTinh>15', 'Cau15_BieuTinhBaolucNgayBieuTinhTangDan' (which is selected and highlighted in blue), 'Cau3_SoNguoiThamGia>5000000', and 'Cau2_TopNuocCoNguoiThamGia>Q3_1990'. On the right, a detailed view of the selected script 'Cau15_BieuTinhBaolucNgayBieuTinhTangDan' is shown. The 'Name' field contains the script name. The 'Expression' field contains the MDX query: `(([Response].[Stateresponse1].ALLMEMBERS,[Measures].[Days Between],))` with a warning icon. The 'Additional Properties' section shows 'Type: Dynamic' and an empty 'Display folder' field.

Hình 231. Viết câu lệnh truy vấn tại ô Expression

Bước 3: Tại file **.cube** của project > Thực hiện khởi chạy (**Process**)

Bước 4: Tại mục **Browser** của cube, thực hiện kéo thả thuộc tính **StateResponse1, DaysBetween**, điều kiện địa điểm và name set vừa tạo ở bước trên vào cửa sổ truy vấn.

Dimension	Hierarchy	Operator	Filter Expression
Location	Country	Equal	{ Mexico }
Violence	Protestviolence	Equal	{ violent protest }

Hình 232. Thực hiện thao tác truy vấn câu 15

Bước 5: Chọn chức năng ‘**Design Mode**’, chỉnh sửa câu truy vấn thêm điều kiện > Thực hiện câu truy vấn

```

SELECT NON EMPTY { [Measures].[Days Between] } ON COLUMNS, NON EMPTY { ([Response].[StateResponse1].[StateResponse1].[AllMembers]) } DIMENSION
  PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM ( SELECT ( { ([Violence].[Protestviolence].&[Violent protest]) } ) ON COLUMNS FROM (
  SELECT { ([Location].[Country].&[Mexico]&[North America]) } ON COLUMNS FROM [Protest Dw v9])) WHERE ( [Location].[Country].&[Mexico]&[North America],
  [Violence].[Protestviolence].&[violent protest] ) CELL_PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME,
  FONT_SIZE, FONT_FLAGS
  
```

Hình 233.1 Câu truy vấn 15 ban đầu

```

SELECT NON EMPTY { [Measures].[Days Between] } ON COLUMNS, NON EMPTY { ORDER([Response].[Stateresponse1].ALLMEMBERS,[Measures].[Days Between],ASC) } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM ( SELECT { { [Violence].[Protesterviolence],[violent protest] } } ON COLUMNS FROM { SELECT { { [Location].[Country]&[Mexico]&[North America] } } ON COLUMNS FROM [Protest_Inv9] } WHERE { [Location].[Country]&[Mexico]&[North America], [Violence].[Protesterviolence],[violent protest] } ) CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING,
FONT_NAME, FONT_SIZE, FONT_FLAGS
  
```

Hình 233.2 Câu truy vấn 15 sau khi chỉnh sửa

- Kết quả câu truy vấn:

Stateresponse1	Days Between
(null)	472
shootings	2
accommodation	10
ignore	54
crowd dispersal	91
arrests	116
beatings	199

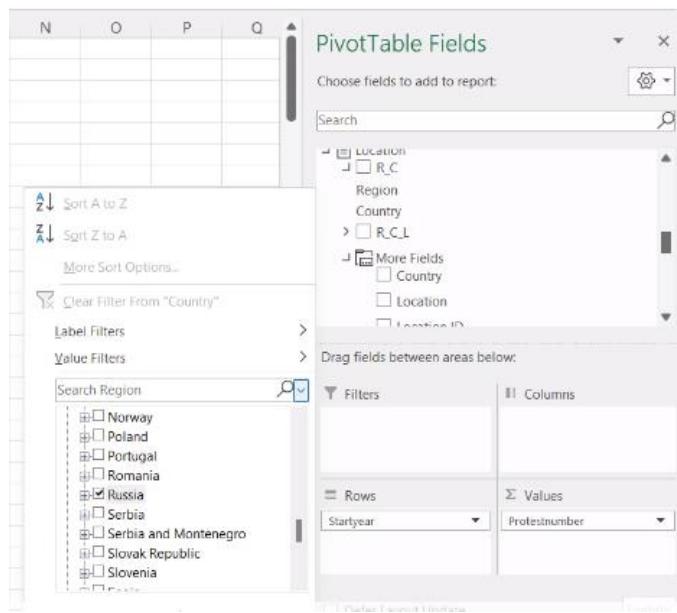
Hình 234. Kết quả truy vấn câu 15

3. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ PIVOT EXCEL

3.1 Cho biết tổng số cuộc biểu tình ở Russia theo từng năm.

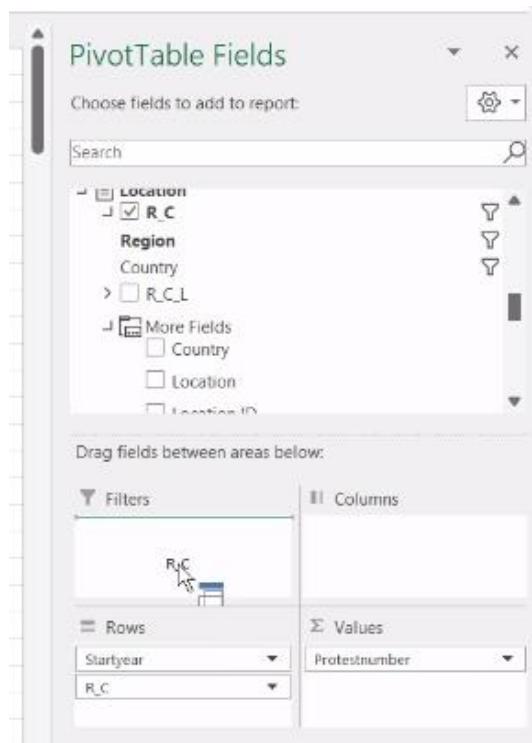
- Quá trình thực hiện:

Bước 1: Từ bảng PivotTable Field, chọn thuộc tính ‘Country’ với dữ liệu là ‘Russia’



Hình 236. Chọn thuộc tính ‘Country’-Russia từ PivotTable Fields

Bước 2: Sau khi chọn được điều kiện là ‘Russia’ thì kéo dữ liệu lên cửa sổ **Filters**



Hình 237. Kéo dữ liệu điều kiện lên cửa sổ Filters

- Kết quả câu truy vấn:

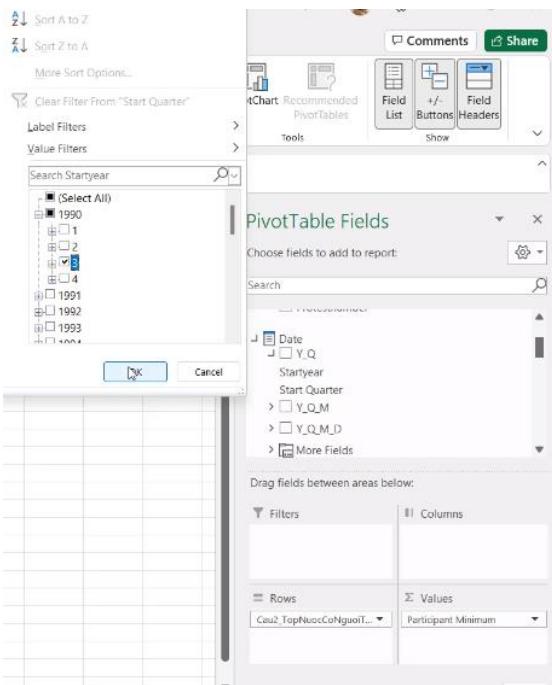
	A	B
1	R_C	Russia
2		
3	Row Labels	Protestnumber
4	1992	66
5	1993	28
6	1994	15
7	1995	15
8	1996	21
9	1997	1
10	1998	28
11	1999	3
12	2000	6
13	2001	6
14	2002	1
15	2003	1
16	2004	15

Hình 238. Kết quả truy vấn câu 1

3.2 Cho biết thông tin cuộc biểu tình, nước xảy ra biểu tình có tổng số lượng người ước tính tham gia biểu tình cao nhất trong quý 3/1990.

- Quá trình thực hiện:

Từ bảng **PivotTable Field**, chọn thuộc tính **Country, Participant Minimum**, name set cũng như điều kiện thời gian vào quý 3/2019



Hình 239. Thao tác tại PivotTable Fields

- Kết quả câu truy vấn:

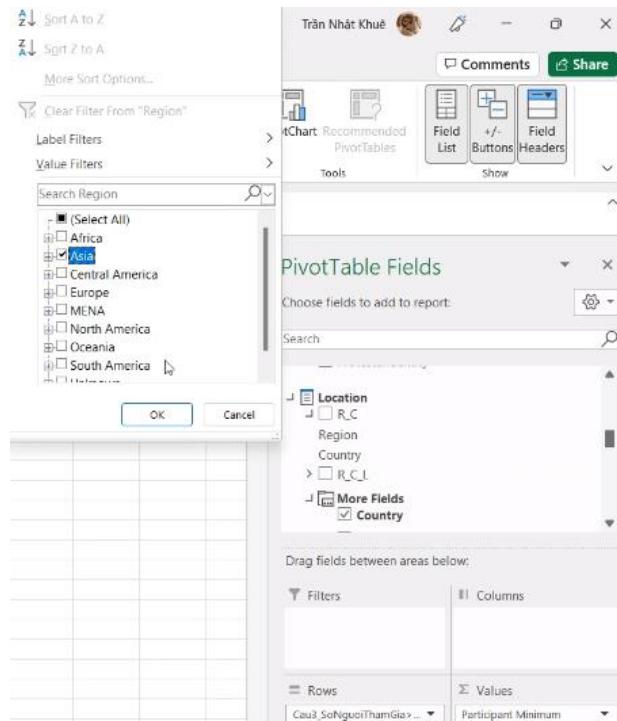
A	B
1 Row Labels	Participant Minimum
2 Venezuela	
3 1990	72150
4	

Hình 240. Kết quả truy vấn câu 2

3.3 Cho biết tên quốc gia có tổng số lượng người tham gia >500000 tại Châu Á.

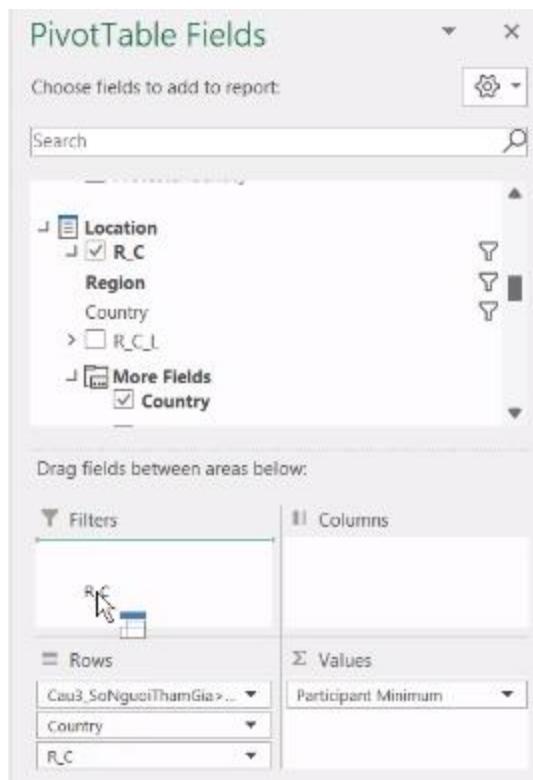
- Quá trình thực hiện:

Bước 1: Từ bảng PivotTable Field, chọn thuộc tính phân cấp **Country**, **Region** và điều kiện name set **ParticipantMinimum**



Hình 241. Thao tác tại PivotTable Fields

Bước 2: Sau khi chọn được điều kiện là ‘Asia’ thì kéo dữ liệu lên cửa sổ *Filters*



Hình 242. Kéo dữ liệu điều kiện lên cửa sổ Filters

- Kết quả câu truy vấn:

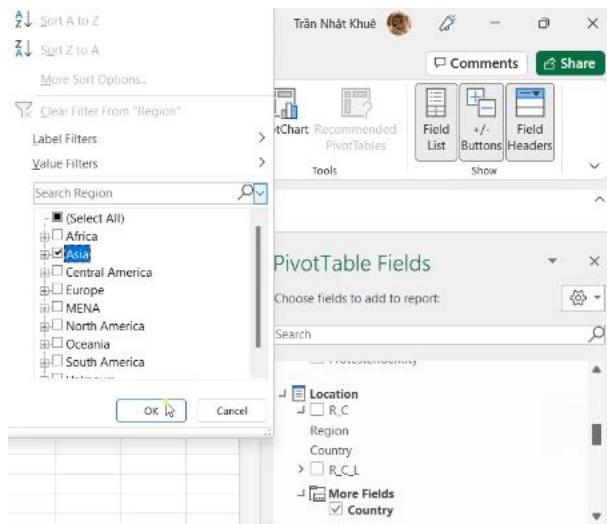
A	B
1 R_C	Asia
2	
3 Row Labels	Participant Minimum
4 152	
5 Bangladesh	8310005
6 624	
7 China	3002530
8 1624	
9 India	5270670
10 1746	
11 Indonesia	2626900
12 2128	
13 Japan	641950
14 2561	
15 Malaysia	518858
16 2897	
17 Nepal	1182703
18 3108	
19 Pakistan	2123750
20 3306	
21 Philippines	1811850
22 3601	
23 South Korea	6560407
24 3807	
25 Taiwan	6454620

Hình 243. Kết quả truy vấn câu 3

3.4 Theo từng tháng, quý, năm liệt kê số lần diễn ra biểu tình tại các nước ở Châu Á.

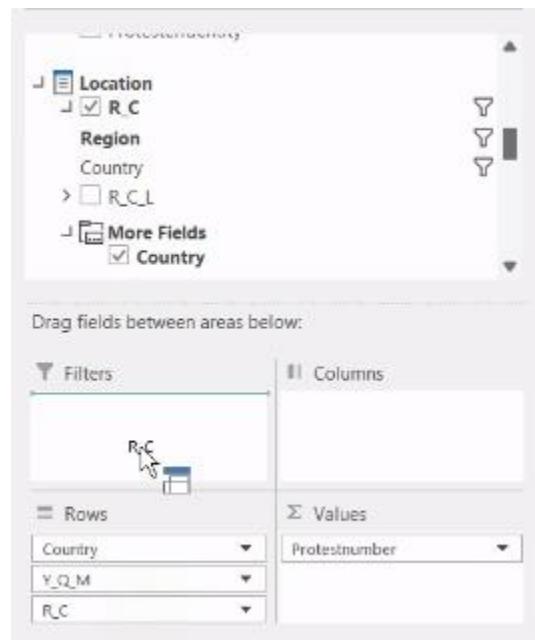
- Quá trình thực hiện:

Bước 1: Từ bảng PivotTable Field, chọn thuộc tính thời gian (*StartMonth*, *StartYear*, *StartQuarter*), *Country*, *Region* và *ParticipantMinimum*



Hình 244. Thao tác tại PivotTable Fields

Bước 2: Sau khi chọn được điều kiện là ‘Asia’ thì kéo dữ liệu lên cửa sổ **Filters**



Hình 245. Kéo dữ liệu điều kiện lên cửa sổ Filters

- Kết quả câu truy vấn:

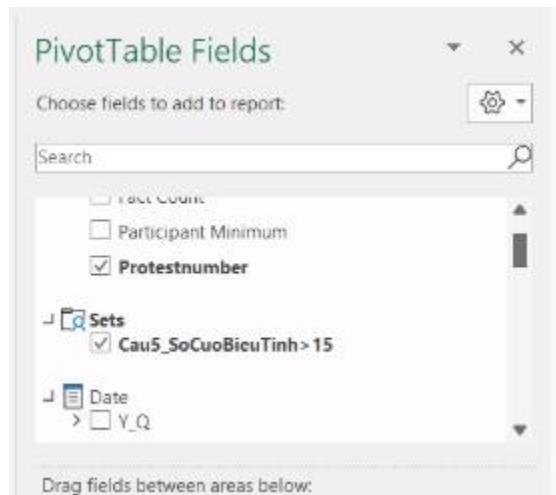
	A	B
1	R_C	Asia
2		
3	Row Labels	Protestnumber
4	Afghanistan	
5	1991	
6	2	
7	5	1
8	1997	1
9	1998	1
10	2002	1
11	2005	3
12	2011	10
13	2012	1
14	2013	1
15	2014	21
16	2015	15
17	2016	10
18	2017	6
19	2018	36
20	2019	1
21	Bangladesh	
22	1990	55
23	1992	21
24	1994	91
25	1995	91
26	1996	21
27	1997	91

Hình 246. Kết quả truy vấn câu 4

3.5 Cho biết thông tin nhóm người biểu tình và tổng số cuộc biểu tình của cuộc biểu tình đó lớn hơn 15 được đặt theo từng quốc gia.

- **Quá trình thực hiện:**

Từ bảng PivotTable Field, chọn thuộc tính **Protestunumber** và name set điều kiện được tạo ở Visual Studio



Hình 247. Thao tác tại PivotTable Fields

- Kết quả câu truy vấn:

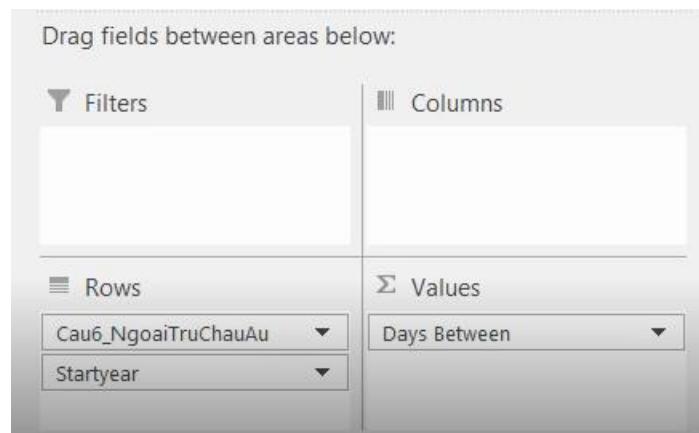
A	B
1 Row Labels	Protestnumber
2 Afghanistan	
3 hazaras	16
4 protesters	31
5 Albania	
6 albanians	88
7 opposition	162
8 protesters	28
9 students	63
10 Algeria	
11 judicial activists	58
12 lawyers	22
13 muslim fundamentalists	22
14 policemen	18
15 pro democracy activists; hirak	1741
16 protesters	94
17 protesters; pro democracy activists	54
18 residents	20
19 students	64
20 students; pro democracy activists; hirak	168

Hình 248. Kết quả truy vấn câu 5

3.6 Với từng châu lục liệt kê số ngày biểu tình theo từng năm, trừ Châu Âu.

- Quá trình thực hiện:

Từ bảng **PivotTable Field**, chọn thuộc tính **Startyear**, **DayBetween** và name set đã tạo ở Visual Studio



Hình 249. Thao tác tại PivotTable Fields

- Kết quả câu truy vấn:

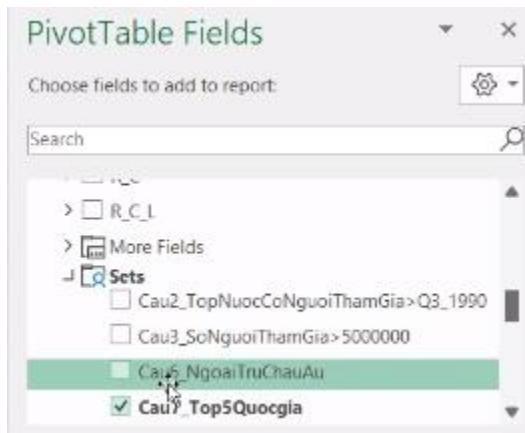
Row Labels	Days Between
Africa	
1990	125
1991	157
1992	99
1993	115
1994	106
1995	69
1996	84
1997	167
1998	356
1999	72
2000	139
2001	122
2002	129
2003	98
2004	171
2005	155
2006	106
2007	100
2008	98

Hình 250. Kết quả truy vấn câu 6

3.7 Đưa ra top 5 quốc gia có số cuộc biểu tình diễn ra nhiều nhất trong năm 2017.

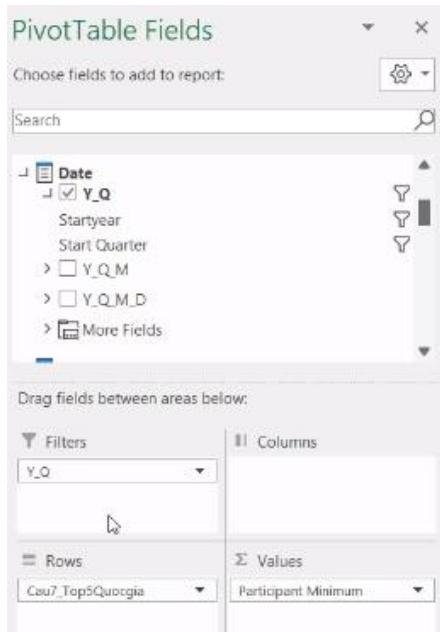
- Quá trình thực hiện:

Bước 1: Từ bảng **PivotTable Field**, chọn thuộc tính **Country**, **ProtestNumber**, điều kiện thời gian và name set điều kiện đã tạo ở Visual Studio



Hình 251. Thao tác tại PivotTable Fields

Bước 2: Sau khi chọn được điều kiện thời gian 2017 thì kéo dữ liệu lên cửa sổ **Filters**



Hình 252. Kéo dữ liệu điều kiện lên cửa sổ Filters

- Kết quả câu truy vấn:

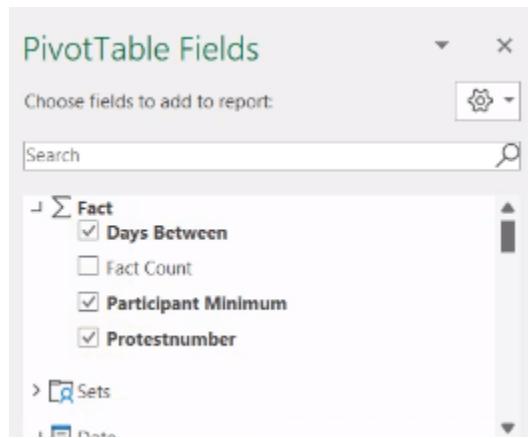
A	B
Y_Q	2017
Row Labels	Protestnumber
Dominican Republic	300
Germany	1081
Nigeria	595
Romania	561
Spain	300

Hình 253. Kết quả truy vấn câu 7

3.8 Thống kê số cuộc biểu tình, số người tham gia và số ngày biểu tình của mục đích biểu tình thứ nhất, sắp xếp số ngày biểu tình theo thứ tự giảm dần.

- Quá trình thực hiện:

Từ bảng **PivotTable Field**, chọn thuộc tính **DayBetween**, **Participant Minimum**, **ProtestNumber** và name set điều kiện đã tạo tại Visual Studio. Thực hiện sắp xếp dữ liệu từ cao đến thấp tại thanh công cụ Excel



Hình 254. Thao tác tại PivotTable Fields

- Kết quả câu truy vấn:

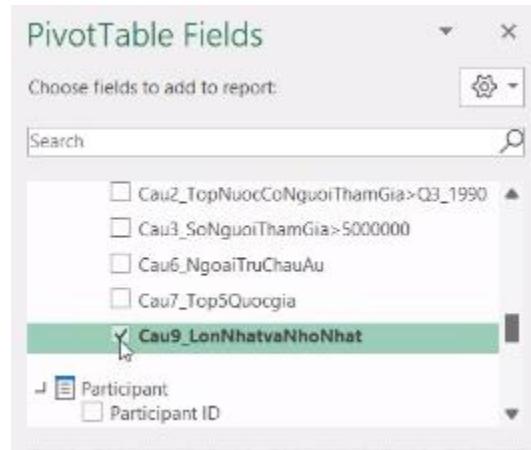
A	B	C	D
Row Labels	Protestnumber	Participant Minimum	Days Between
political behavior, process	81650	82108092	23238
labor wage dispute	13875	23656565	6310
removal of politician	8559	9596209	2754
price increases, tax policy	8190	8506046	2531
land farm issue	5485	834067	2299
police brutality	6495	3158457	1638
social restrictions	2724	1117130	967
	3	5000	1
Grand Total	126981	128981566	39738

Hình 256. Kết quả truy vấn câu 8

3.9 Cho biết quốc gia có số ngày biểu tình cao nhất và quốc gia có số lần biểu tình thấp nhất.

- Quá trình thực hiện:

Từ bảng PivotTable Field, chọn thuộc tính **Country**, **DayBetween**, **ProtestNumber** và name set điều kiện đã tạo ở Visual Studio



Hình 257. Thao tác tại PivotTable Fields

- Kết quả câu truy vấn:

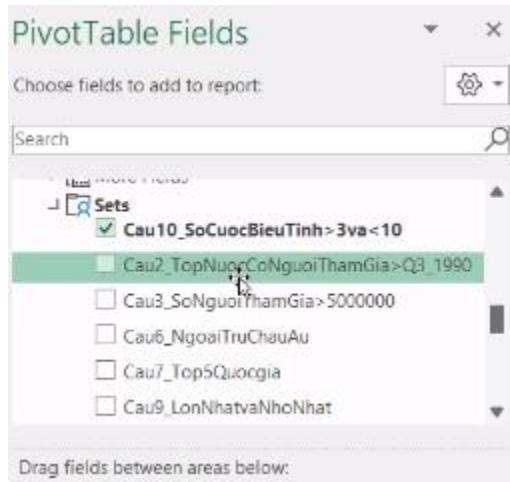
A	B
Row Labels	Fact Count
2 Europe	4994
3 Oceania	38

Hình 258. Kết quả truy vấn câu 9

3.10 Liệt kê những quốc gia có số cuộc biểu tình từ 3 đến 10.

- Quá trình thực hiện:

Bước 1: Từ bảng PivotTable Field, chọn thuộc tính **Country**, **Protest** và name set đã tạo ở Visual Studio



Hình 259. Thao tác tại PivotTable Fields

- Kết quả câu truy vấn:

A	B
Row Labels	Protestnumber
Cape Verde	3
Czechoslovakia	10
Equatorial Guinea	4
Eritrea	3
Germany West	3
Luxembourg	4
North Korea	9
Norway	7
Serbia and Montenegro	3
Turkmenistan	4
United Arab Emirate	3

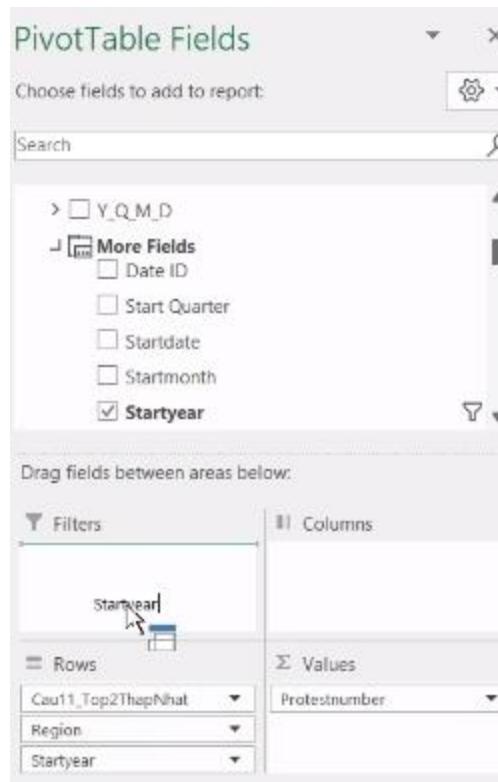
Hình 260. Kết quả truy vấn câu 10

3.11 Xuất ra top 2 những thành phố thuộc Châu Phi có số lần biểu tình thấp nhất trong năm 2018.

- Quá trình thực hiện:

Bước 1: Từ bảng PivotTable Field, chọn thuộc tính **Country**, **ProtestNumber**, **StartYear**

Bước 2: Sau khi chọn được điều kiện thì kéo dữ liệu lên cửa sổ **Filters**



Hình 261. Kéo dữ liệu điều kiện lên cửa sổ Filters

- Kết quả câu truy vấn:

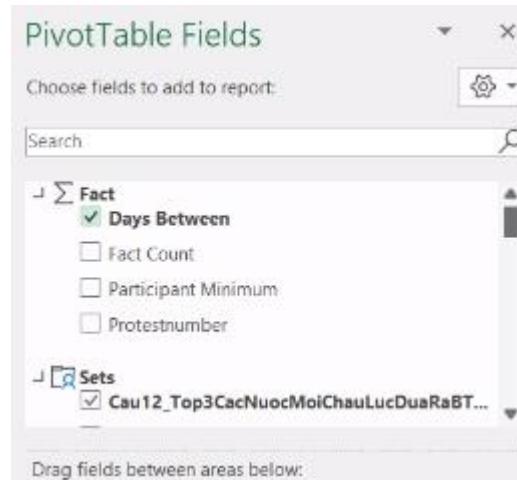
	A	B
1	Startyear	2018
2	Region	Africa
3		
4	Row Labels	Protestnumber
5	Somalia	1
6	South Africa	1
7		

Hình 262. Kết quả truy vấn câu 11

3.12 Với mỗi châu lục, đưa ra 3 quốc gia có tổng số ngày biểu tình cao nhất.

- Quá trình thực hiện:

Từ bảng PivotTable Field, chọn thuộc tính **Country**, **Region**, **DaysBetween** và name set đã tạo ở Visual Studio



Hình 263. Thao tác tại PivotTable Fields

- Kết quả câu truy vấn:

A	B
Row Labels	Days Between
Africa	
Kenya	451
Namibia	1033
South Africa	298
Asia	
China	893
India	1423
Thailand	1228
Central America	
Guatemala	202
Honduras	171
Nicaragua	479
Europe	
France	1263
Greece	1295
United Kingdom	1154
MENA	
Egypt	448
Morocco	1120
Syria	702

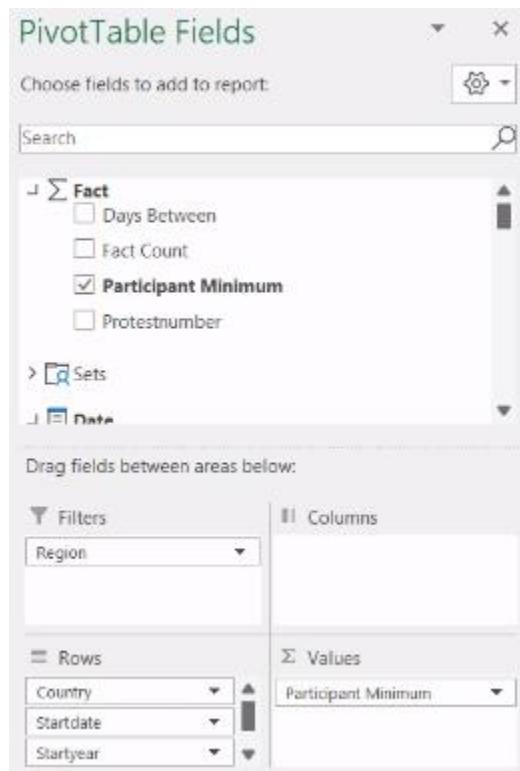
Hình 264. Kết quả truy vấn câu 12

3.13 Liệt kê những cuộc biểu tình ở các quốc gia tại Bắc Mĩ, xảy ra từ năm 1998 đến 2002 với số người tham gia < 1000

- Quá trình thực hiện:

Bước 1: Từ bảng PivotTable Field, chọn thuộc tính **Country**, **LocationID**, **StartDate** và name set đã tạo ở Visual Studio

Bước 2: Sau khi chọn được điều kiện là ‘North America’ thì kéo dữ liệu lên cửa sổ **Filters**



Hình 265. Kéo dữ liệu điều kiện lên cửa sổ Filters

- Kết quả câu truy vấn:

	A	B
4	Row Labels	Participant Minimum
5	■ Cuba	
6	■ 1999-06-15	
7	1999	
8	915	50
9	■ Dominican Republic	
10	■ 1998-03-11	
11	1998	
12	964	100
13	■ 1998-04-28	
14	1998	
15	964	50
16	■ 1998-07-06	
17	1998	
18	964	50
19	■ 1998-10-22	
20	1998	
21	964	100
22	■ 1999-03-15	
23	1999	
24	964	100

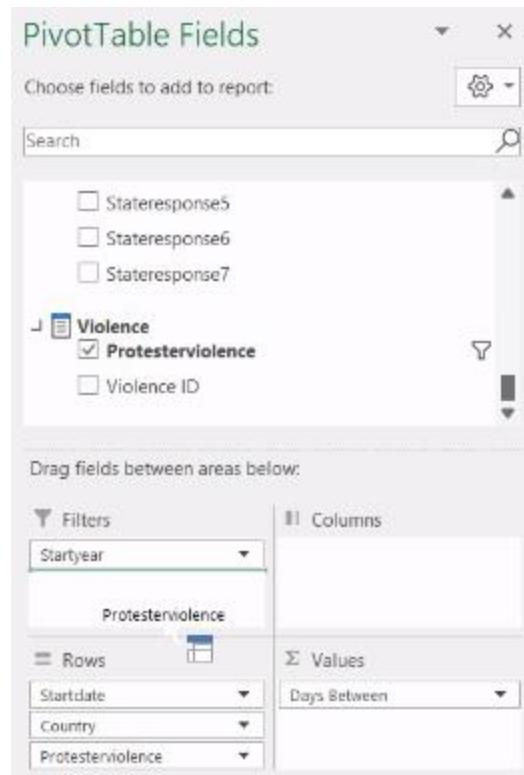
Hình 267. Kết quả truy vấn câu 13

3.14 Liệt kê những quốc gia có các cuộc biểu tình diễn ra với hình thức ôn hòa vào năm 2019, cũng như số ngày diễn ra cuộc biểu tình đó.

- Quá trình thực hiện:

Bước 1: Từ bảng **PivotTable Field**, chọn thuộc tính

Bước 2: Sau khi chọn được điều kiện **ProtesterViolence** và **StartYear** thì kéo dữ liệu lên cửa sổ **Filters**



Hình 268. Kéo dữ liệu điều kiện lên cửa sổ Filters

- Kết quả câu truy vấn:

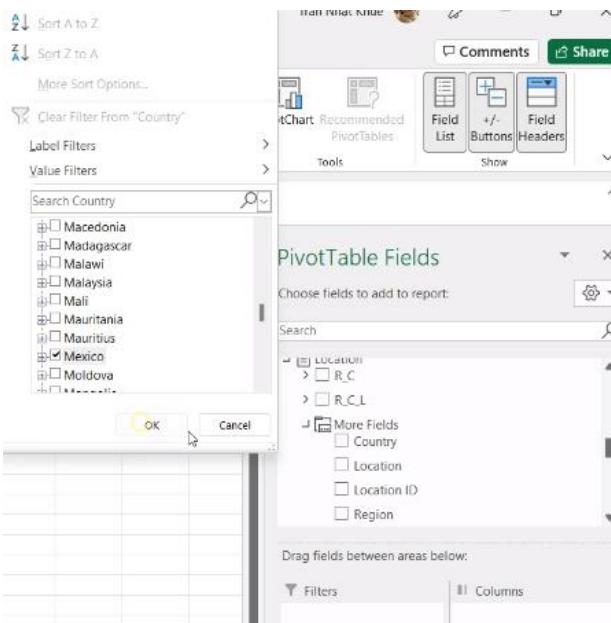
A	B
1 Startyear	2019
2 Protesterviolence	peaceful protest
3	
4 Row Labels	Days Between
5 □ 2019-01-01	
6 India	1
7 Peru	1
8 □ 2019-01-02	
9 Madagascar	1
10 □ 2019-01-03	
11 Bangladesh	1
12 Madagascar	1
13 □ 2019-01-05	
14 Hungary	1
15 Serbia	1
16 □ 2019-01-06	
17 Sudan	1
18 □ 2019-01-07	
19 Canada	2
20 United Kingdom	1

Hình 269. Kết quả truy vấn câu

3.15 Cho biết số ngày diễn ra những cuộc biểu tình có sử dụng bạo lực vũ trang cũng như phản ứng đầu tiên của chính phủ đối với cuộc biểu tình đó tại Mexico, sắp xếp theo số ngày biểu tình tăng dần.

- Quá trình thực hiện:

Từ bảng PivotTable Field, chọn thuộc tính *StateResponse1*, *DaysBetween*, *Country*, *ProtesterViolence*



Hình 270. Thao tác tại PivotTable Fields

- Kết quả câu truy vấn:

A	B
Country	Mexico
Protesterviolence	violent protest
Row Labels	Days Between
shootings	2
accommodation	10
ignore	54
crowd dispersal	91
arrests	116
beatings	199
Grand Total	472

Hình 271. Kết quả truy vấn câu 15

4. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG NGÔN NGỮ MDX

4.1 Cho biết tổng số cuộc biểu tình ở Russia theo từng năm.

- Câu lệnh MDX:

```
SELECT NON EMPTY { [Measures].[Protestnumber] } ON COLUMNS,
NON EMPTY {[Date].[Startyear].Members} ON Rows
From [Protest Dw v9]
Where [Location].[R_C].[Country].&[Russia]&[Europe];
```

- Kết quả truy vấn:

	Protestnumber
All	1150
1992	66
1993	28
1994	15
1995	15
1996	21
1997	1
1998	28
1999	3
2000	6
2001	6
2002	1
2003	1
2004	15
2005	78
2006	10

Hình 272. Kết quả truy vấn câu 1

4.2 Cho biết thông tin cuộc biểu tình, nước xảy ra biểu tình có tổng số lượng người ước tính tham gia biểu tình cao nhất trong quý 3/1990.

- Câu lệnh MDX:

```
Select NON EMPTY {[Measures].[Participant Minimum]} on Columns ,
{TOPSUM([Location].[Country].CHILDREN,1,[Measures].[Participant Minimum])} on Rows
From [Protest Dw v9]
Where [Date].[Y_Q].[Start Quarter].&[3]&[1990];
```

- Kết quả truy vấn:

Messages		Results
		Participant Minimum
Venezuela		72150

Hình 273. Kết quả truy vấn câu 2

4.3 Cho biết tên những quốc gia mà có tổng số lượng người tham gia >500000 tại Châu Á.

- Câu lệnh MDX:

```
Select {[Measures].[Participant Minimum]} on columns,
{FILTER ({[Location].[Location
ID].CHILDREN* [Location].[Country].Children},[Measures].[Participant Minimum] > 500000)} on Rows
From [Protest Dw v9]
Where [Location].[R_C].[Region].&[Asia];
```

- Kết quả truy vấn:

		Participant Minimum
152	Bangladesh	8310005
624	China	3002530
1624	India	5270670
1746	Indonesia	2626900
2128	Japan	641950
2561	Malaysia	518858
2897	Nepal	1182703
3108	Pakistan	2123750
3306	Philippines	1811850
3601	South Korea	6560407
3807	Taiwan	6454620
3855	Thailand	2687090

Hình 274. Kết quả truy vấn câu 3

4.4 Theo từng tháng, quý, năm liệt kê số lần diễn ra biểu tình tại các nước ở Châu Á.

- Câu lệnh MDX:

```
Select Non empty{[Location].[Country].Children*[Measures].[Protestnumber]} on Columns,
{DrillDownLevel(
    DrillDownLevel(
        DrillDownLevel([Date].[Y_Q_M])
    )
)}
} on Rows
From [Protest Dw v9]
where {[Location].[R_C].[Region].&[Asia]};
```

- Kết quả truy vấn:

	Afghanistan	Bangladesh	Bhutan	Cambodia	China	India	Indonesia	Japan	Kazakhstan	Kyrgyzstan	Laos	Malaysia	Mongolia	Myanmar	Pr
	Protestnumber														
All	108	4978	2	142	3841	1328	764	169	266	2014	2	693	179	512	
1990	(null)	55	1	(null)	6	55	1	15	1	(null)	(null)	(null)	78	6	
1	(null)	(null)	(null)	(null)	1	10	1	1	(null)	(null)	(null)	(null)	36	(null)	
2	(null)	(null)	(null)	(null)	1	3	1	(null)	(null)	(null)	(null)	(null)	3	(null)	
3	(null)	(null)	(null)	(null)	(null)	4	(null)	1	(null)	(null)	(null)	(null)	30	(null)	
2	(null)	3	(null)	(null)	5	5	(null)	(null)	(null)	(null)	1	(null)	42	(null)	
4	(null)	(null)	(null)	(null)	2	(null)	9	(null)							
5	(null)	(null)	(null)	(null)	(null)	5	(null)	(null)	(null)	(null)	(null)	(null)	33	(null)	
6	(null)	3	(null)	(null)	3	(null)	(null)	(null)	(null)	(null)	1	(null)	(null)	(null)	3
7	(null)	3	1	(null)	(null)	40	(null)	(null)	1	(null)	(null)	(null)	(null)	(null)	3

Hình 275. Kết quả truy vấn câu 4

4.5 Cho biết thông tin nhóm người biểu tình và tổng số cuộc biểu tình của cuộc biểu tình đó lớn hơn 15 được đặt theo từng quốc gia.

- Câu lệnh MDX:

```
Select {[Measures].[Protestnumber]} on Columns,
{Generate(
    [Location].[Country].Children,
    [Location].[Country].CurrentMember *
    Filter(
        [Identification].[Protesteridentity].CHILDREN,[Measures].[Protestnumber]>15)
    )
} on Rows
From [Protest Dw v9];
```

- Kết quả truy vấn:

		Protestnumber
Afghanistan	hazaras	16
Afghanistan	protesters	31
Albania	albanians	88
Albania	opposition	162
Albania	protesters	28
Albania	students	63
Algeria	judicial activists	58
Algeria	lawyers	22
Algeria	muslim fundamentalists	22
Algeria	policemen	18
Algeria	pro democracy activists; hi...	1741
Algeria	protesters	94
Algeria	protesters; pro democracy ...	54
Algeria	residents	20
Algeria	students	64
Algeria	students; pro democracy a...	168

Hình 276. Kết quả truy vấn câu 5

4.6 Với từng châu lục liệt kê số ngày biểu tình theo từng năm, trừ Châu Âu.

- Câu lệnh MDX:

```
Select {[Date].[Startyear].Members} on columns,
```

```

Non empty
{Except(
    {[Location].[Region].Children},
    {[Location].[Region].&[Europe]}
)
} on Rows
From [Protest Dw v9]
where {[Measures].[Days Between]};
```

- Kết quả truy vấn:

	All	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Africa	5120	125	157	99	115	106	69	84	167	356	72	139	122	129	98	171
Asia	9346	277	82	152	65	126	131	137	275	228	143	458	434	233	175	305
Central America	1159	14	8	13	20	29	79	59	48	19	34	25	20	18	28	24
MENA	4453	20	39	40	34	37	23	16	9	26	44	19	28	33	21	31
North America	1952	139	81	85	161	32	47	30	56	19	39	20	9	13	39	14
Oceania	121	4	43	(null)	1	(null)	1	1	3	1	(null)	2	18	(null)	(null)	2
South America	7209	138	347	205	203	223	192	327	362	272	342	173	146	279	233	236

Hình 277. Kết quả truy vấn câu 6

4.7 Đưa ra top 5 quốc gia có số cuộc biểu tình diễn ra nhiều nhất trong năm 2017.

- Câu lệnh MDX:

```

select {[Measures].[Protestnumber]} on columns,
NON EMPTY
    TOPCOUNT([Location].[Country].CHILDREN,5,[Measures].[Protestnumber])
on rows
from [Protest Dw v9]
where {[Date].[Y_Q].[Startyear].&[2017]};
```

- Kết quả truy vấn:

	Protestnumber
Germany	1081
Nigeria	595
Romania	561
Dominican Republic	300
Spain	300

Hình 278. Kết quả truy vấn câu 7

4.8 Thống kê số cuộc biểu tình, số người tham gia và số ngày biểu tình của mục đích biểu tình thứ nhất, sắp xếp số ngày biểu tình theo thứ tự giảm dần.

- Câu lệnh MDX:

```
select {[Measures].[Protestnumber],[Measures].[Participant Minimum],[Measures].[Days Between]} on columns,
NON EMPTY
    ORDER([Demand].[Protesterdemand1].ALLMEMBERS,[Measures].[Days Between], DESC)
on rows
from [Protest Dw v9];
```

- Kết quả truy vấn:

	Protestnumber	Participant Minimum	Days Between
All	126981	128981566	39738
political behavior, process	81650	82108092	23238
labor wage dispute	13875	23656565	6310
removal of politician	8559	9596209	2754
price increases, tax policy	8190	8506046	2531
land farm issue	5485	834067	2299
police brutality	6495	3158457	1638
social restrictions	2724	1117130	967
	3	5000	1

Hình 279. Kết quả truy vấn câu 8

4.9 Cho biết quốc gia có số ngày biểu tình cao nhất và quốc gia có số lần biểu tình thấp nhất.

- Câu lệnh MDX:

```
Select {[Measures].[Fact Count]} on columns,
NON EMPTY
    Union(TopCount({[Location].[Region].Children},1,
        [Measures].[Fact Count]),
        BottomCount({[Location].[Region].Children },2,
        [Measures].[Fact Count])) on rows
From [Protest Dw v9];
```

- Kết quả truy vấn:

Fact Count	
Europe	4994
Oceania	38

Hình 280. Kết quả truy vấn câu 9

4.10 Liệt kê những quốc gia có số cuộc biểu tình từ 3 đến 10.

- Câu lệnh MDX:
 - Cách 1

```
Select {[Measures].[Protestnumber]} on columns,
NON EMPTY
  FILTER([Location].[Country].CHILDREN,[Measures].[Protestnumber]>=3 and
[Measures].[Protestnumber]<=10)
on rows
from [Protest Dw v9];
```

- Cách 2

```
Select {[Measures].[Protestnumber]} on Columns,
NON EMPTY
  {intersect(
    {FILTER([Location].[Country].CHILDREN,[Measures].[Protestnumber]>=3)}
    ,{FILTER([Location].[Country].CHILDREN,[Measures].[Protestnumber]<=10)})
  }
on rows
from [Protest Dw v9];
```

- Kết quả truy vấn:

	Protestnumber
Cape Verde	3
Czechoslovakia	10
Equatorial Guinea	4
Eritrea	3
Germany West	3
Luxembourg	4
North Korea	9
Norway	7
Serbia and Montenegro	3
Turkmenistan	4
United Arab Emirate	3

Hình 281. Kết quả truy vấn câu 10

4.11 Xuất ra top 2 những thành phố thuộc Châu Phi có số lần biểu tình thấp nhất trong năm 2018.

- Câu lệnh MDX:

```
select {[Measures].[Protestnumber]} on columns,
NON EMPTY
    BottomCount(NonEmpty([Location].[Country].Children,[Measures].[Protestnumber]),2
,[Measures].[Protestnumber])
on rows
from [Protest Dw v9]
where {[Location].[R_C].[Region]&[Africa],[Date].[Y_Q].[Startyear]&[2018])};
```

- Kết quả truy vấn:

	Protestnumber
South Africa	1
Somalia	1

Hình 282. Kết quả truy vấn câu 11

4.12 Với mỗi châu lục, đưa ra 3 quốc gia có tổng số ngày biểu tình cao nhất.

- Câu lệnh MDX:

```
Select {[Measures].[Days Between]} on Columns,
```

```

{Generate(
    [Location].[Region].Children,
    TopCount(
        {[Location].[Region].CurrentMember* [Location].[Country].Children}
        ,3,[Measures].[Days Between]
    )
)} on Rows
From [Protest Dw v9];

```

- Kết quả truy vấn:

		Days Between
Africa	Namibia	1033
Africa	Kenya	451
Africa	South Africa	298
Asia	India	1423
Asia	Thailand	1228
Asia	China	893
Central America	Nicaragua	479
Central America	Guatemala	202
Central America	Honduras	171
Europe	Greece	1295
Europe	France	1263
Europe	United Kingdom	1154
MENA	Morocco	1120
MENA	Syria	702
MENA	Egypt	448
North America	Mexico	807

Hình 283. Kết quả truy vấn câu 12

4.13 Liệt kê những cuộc biểu tình ở các quốc gia tại Bắc Mĩ, xảy ra từ năm 1998 đến 2002 với số người tham gia < 1000

- Câu lệnh MDX:

```

select {[Measures].[Participant Minimum]} on columns,
NON EMPTY
    [Date].[Startdate].CHILDREN*
    FILTER ([Location].[Country].CHILDREN,[Measures].[Participant Minimum] <1000)
on rows
from [Protest Dw v9]

```

```

where
{([Date].[Y_Q].[Startyear].&[1998]:[Date].[Y_Q].[Startyear].&[2000], [Location].[R_C].[Region].&[North America])};

```

- **Kết quả truy vấn:**

		Participant Minimum
1998-03-11	Dominican Republic	100
1998-04-28	Dominican Republic	50
1998-07-06	Dominican Republic	50
1998-09-24	Jamaica	100
1998-10-22	Dominican Republic	100
1999-03-15	Dominican Republic	100
1999-03-19	Dominican Republic	100
1999-04-19	Jamaica	100
1999-06-14	Jamaica	100
1999-06-15	Cuba	50
2000-03-30	Dominican Republic	100

Hình 284. Kết quả truy vấn câu 13

4.14 Liệt kê những quốc gia có các cuộc biểu tình diễn ra với hình thức ôn hòa vào năm 2019, cũng như số ngày diễn ra cuộc biểu tình đó.

- **Câu lệnh MDX:**

```

select {[Measures].[Days Between]} on columns,
NON EMPTY
    [Date].[Startdate].CHILDREN*
    {[Location].[Country].CHILDREN} on rows
from [Protest Dw v9]
where {[Violence].[Protesterviolence].&[peaceful
protest],[Date].[Startyear].&[2019])};

```

- **Kết quả truy vấn:**

		Days Between
2019-01-01	India	1
2019-01-01	Peru	1
2019-01-02	Madagascar	1
2019-01-03	Bangladesh	1
2019-01-03	Madagascar	1
2019-01-05	Hungary	1
2019-01-05	Serbia	1
2019-01-06	Sudan	1
2019-01-07	Canada	2
2019-01-07	United Kingdom	1
2019-01-08	Canada	1
2019-01-08	Sudan	4
2019-01-08	Thailand	1
2019-01-10	Belgium	1
2019-01-10	Congo Kinshasa	1
2019-01-10	Mongolia	1

Hình 285. Kết quả truy vấn câu 14

4.15 Cho biết số ngày diễn ra những cuộc biểu tình có sử dụng bạo lực vũ trang cũng như phản ứng đầu tiên của chính phủ đối với cuộc biểu tình đó tại Mexico, sắp xếp theo số ngày biểu tình tăng dần.

- **Câu lệnh MDX:**

```
select {[Measures].[Days Between]} on columns,
NON EMPTY
    ORDER([Response].[Stateresponse1].ALLMEMBERS,[Measures].[Days Between], ASC) on
rows
From [Protest Dw v9]
where {[[Violence].[Protesterviolence].&[violent
protest],[Location].[Country].&[Mexico]&[North America])}
```

- **Kết quả truy vấn:**

	Days Between
All	472
shootings	2
accomodation	10
ignore	54
crowd dispersal	91
arrests	116
beatings	199

Hình 286. Kết quả truy vấn câu 15

CHƯƠNG 4: QUY TRÌNH LẬP BÁO CÁO (REPORT)

1. QUÁ TRÌNH TẠO BÁO CÁO BẰNG CÔNG CỤ VISUAL STUDIO 2019 (SSRS)

1.2 Khởi tạo project SSRS

- Quá trình thực hiện:

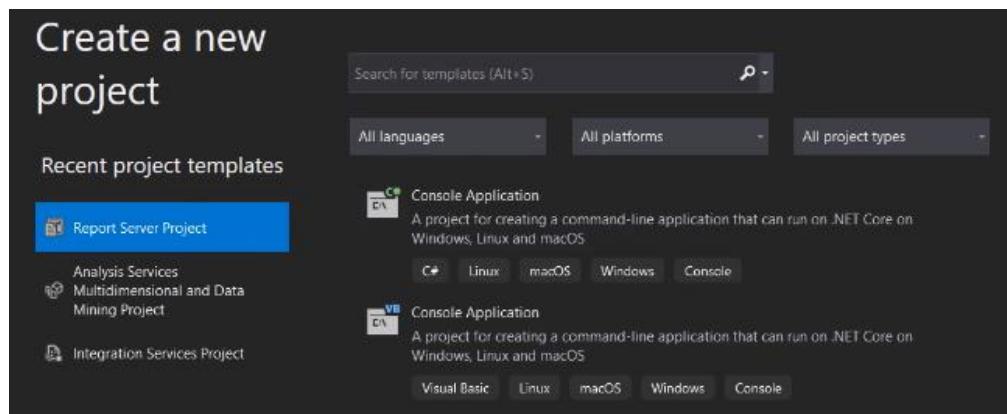
Bước 1: Cài đặt các công cụ cần thiết để thực hiện SSRS

- File SQLReportingService.exe:

Cài đặt Service SSRS tạo server deploy các báo cáo

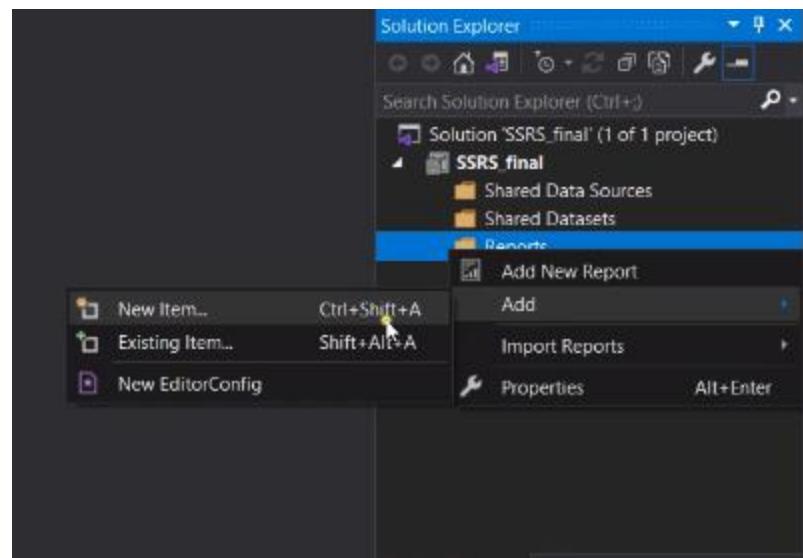
<https://www.microsoft.com/en-us/download/details.aspx?id=100122>

Bước 2: Tại Visual Studio, chọn *File* -> *Create New Projects* > Tìm ‘*Report Server Project*’ > Đặt tên và chọn vị trí lưu project



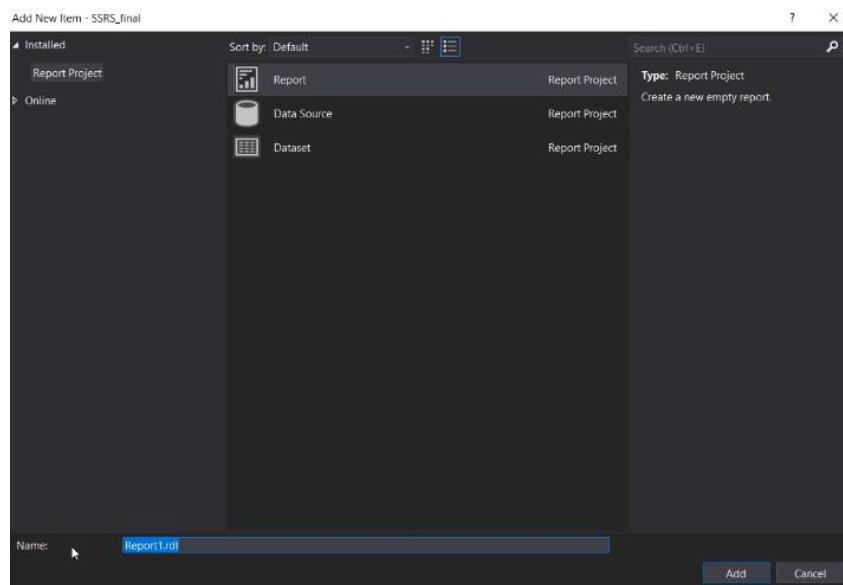
Hình 287. Khởi tạo project SSRS

Bước 3: Tại thư mục *Reports* > *Add* > *New Item*



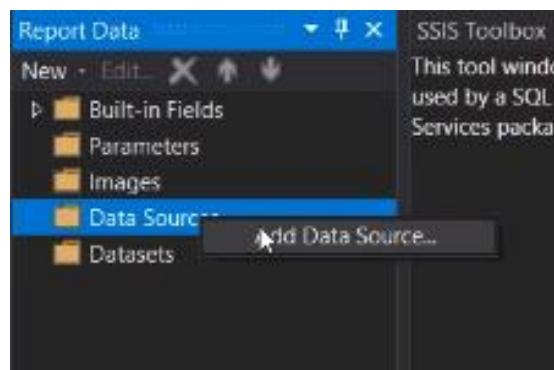
Hình 288. Tạo file báo cáo (report)

Bước 4: Chọn ‘Report’ tại màn hình vừa xuất hiện và đặt tên cho file báo cáo



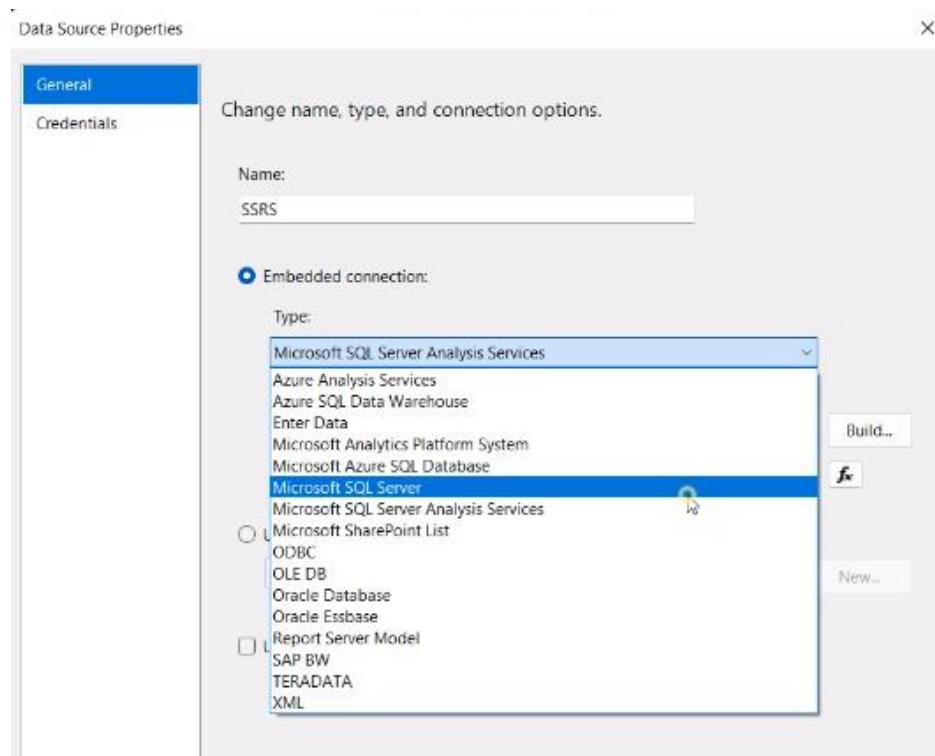
Hình 289. Thực hiện chọn loại báo cáo (type report)

Bước 5: Thực hiện thêm dữ liệu nguồn vào project (Add Data Source...)



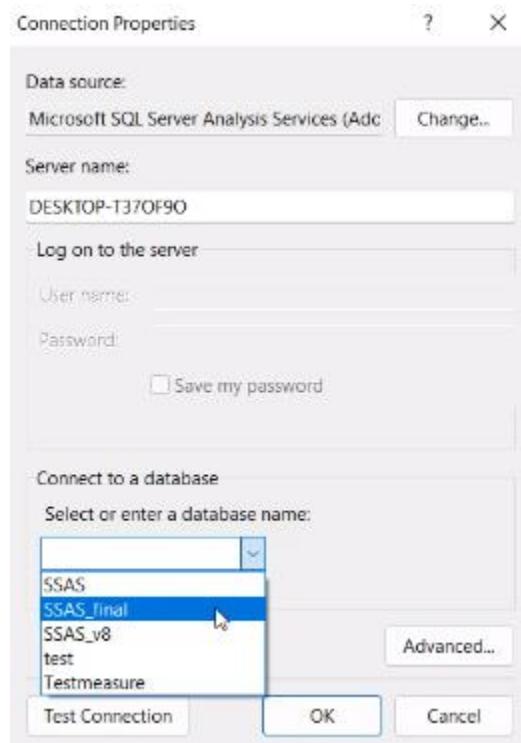
Hình 290. Thực hiện thêm dữ liệu nguồn

Bước 6: Sau khi cửa sổ hiện lên, đặt tên cho dữ liệu nguồn và chọn loại kết nối => **Microsoft SQL Server**



Hình 291. Thực hiện thêm dữ liệu nguồn

Bước 7: Chọn nút *Build* > chọn *Change* để thay điều chỉnh loại kết nối > Chọn *Microsoft SQL Server Analysis Services* > Dán tên server của máy và chọn database > *OK*



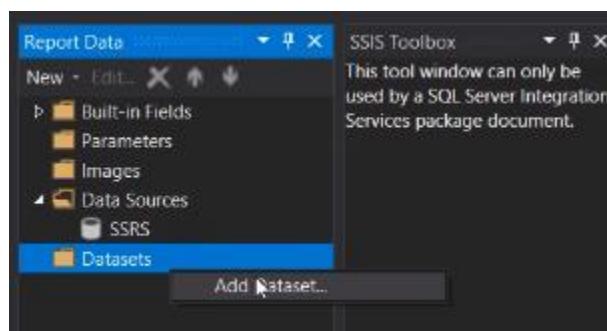
Hình 292. Thực hiện chỉnh loại kết nối cơ sở dữ liệu

Bước 8: Tại mục *Credentials* > Chọn *Use Windows Authentication*



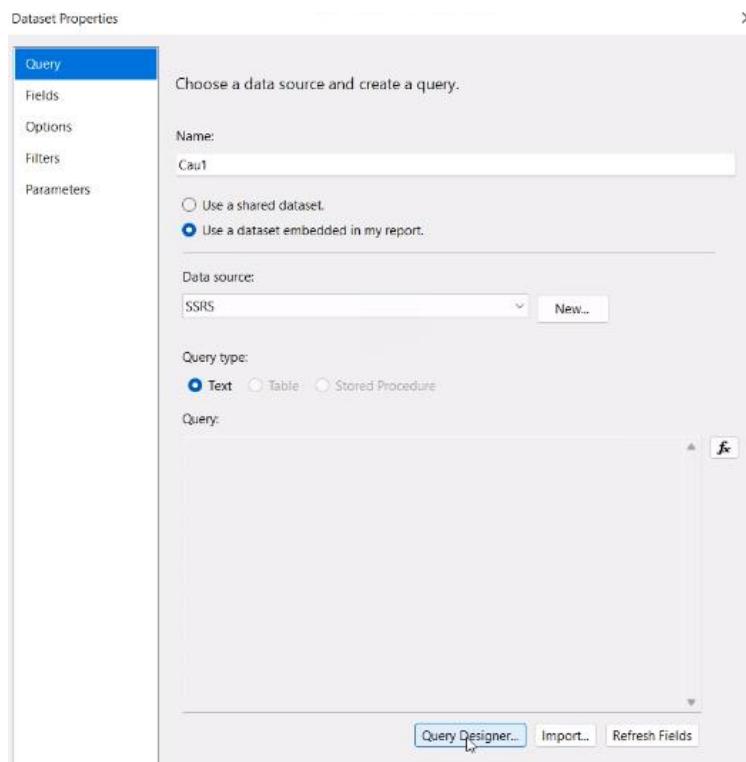
Hình 293. Thay đổi thông tin xác thực để kết nối dữ liệu nguồn

Bước 9: Tại thư mục **Datasets** > **Add Datasets**



Hình 294. Thực hiện tạo Dataset

Bước 10: Tại tab **Query** > Đặt tên câu truy vấn > Chọn dữ liệu nguồn (*data source*) > Chọn **Query Designer**



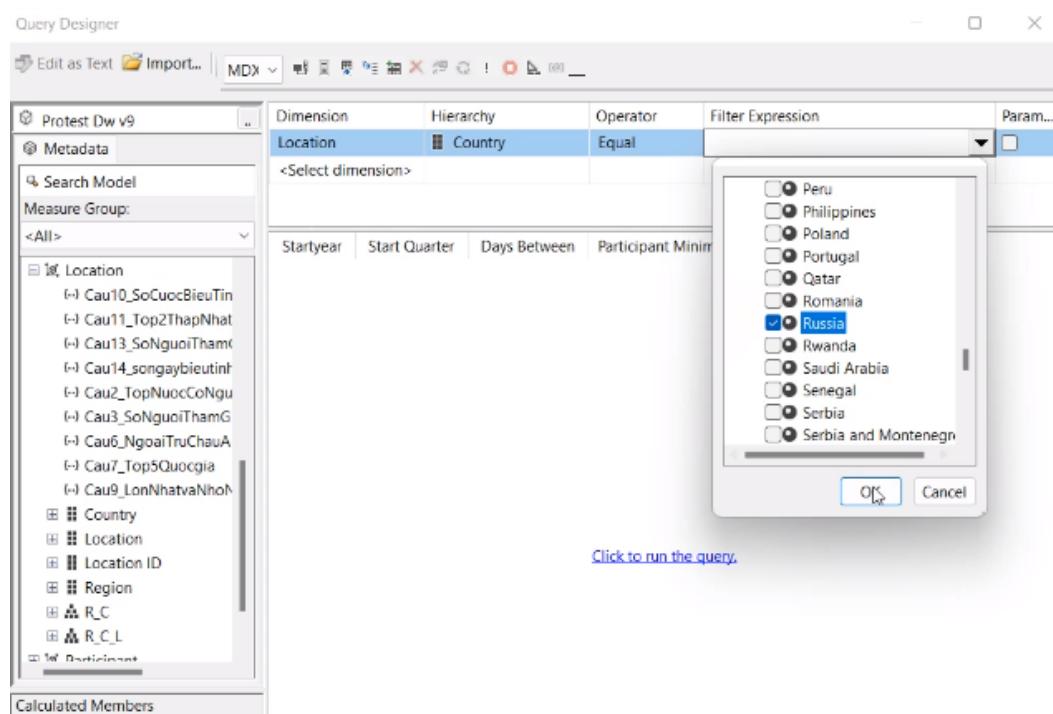
Hình 295. Thực hiện tạo thông tin cho câu truy vấn tạo báo cáo

1.3 Thực hiện tạo report trên Visual Studio

1.3.1 Báo cáo thống kê số cuộc biểu tình, số ngày biểu tình, số người tham gia biểu tình tại Nga từ 1992 đến 1999

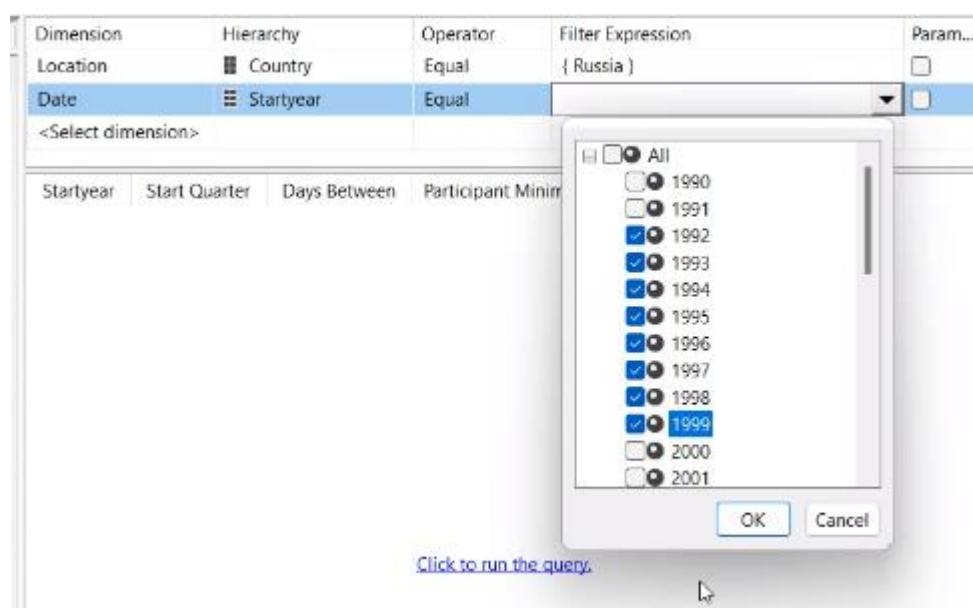
Bước 1: Tại cửa sổ Query Desgin, kéo thả các thuộc tính liên quan đến câu truy vấn 1 ‘*Start Year*, *StartQuarter*, *DaysBetween*, *ProtestNumber*, *ParticipantMinimum*’ vào vùng thực thi truy vấn.

Bước 2: Kéo thả thuộc tính Location và chọn điều kiện về đất nước ‘*Russia*’



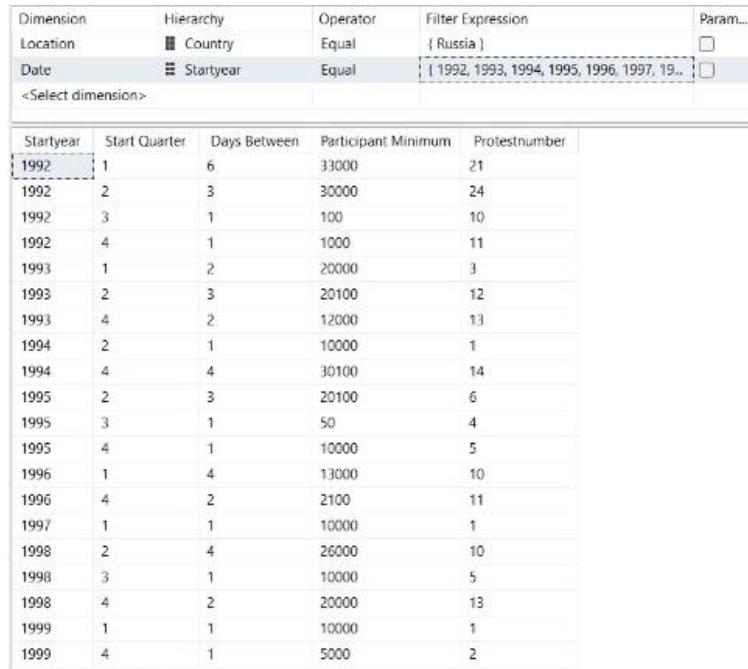
Hình 296. Điều chỉnh dimension cho điều kiện địa điểm ở câu 1

Bước 3: Kéo thả thuộc tính Location và chọn điều kiện về đất nước ‘**Russia**’



Hình 297. Điều chỉnh dimension cho điều kiện thời gian ở câu 1

Bước 4: Chọn ‘**Run the query**’ để xem hiện kết quả truy vấn



The screenshot shows a data analysis interface. At the top, there is a configuration table for dimensions:

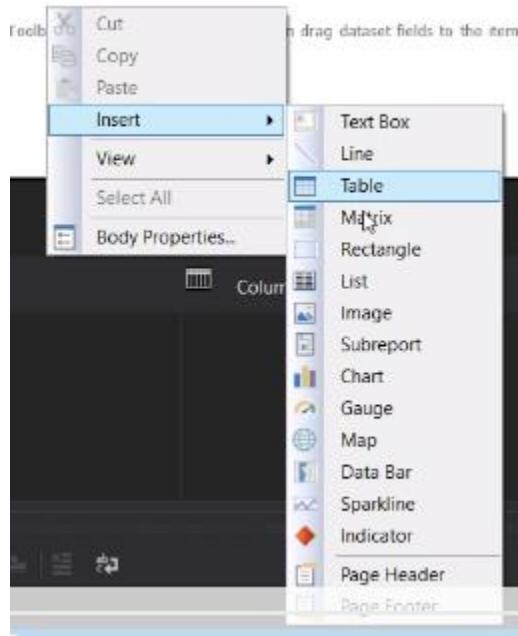
Dimension	Hierarchy	Operator	Filter Expression	Param...
Location	Country	Equal	{ Russia }	<input type="checkbox"/>
Date	Startyear	Equal	{ 1992, 1993, 1994, 1995, 1996, 1997, 19... }	<input type="checkbox"/>
<Select dimension>				

Below this is a data table with the following columns: Startyear, Start Quarter, Days Between, Participant Minimum, and Protestnumber. The data is as follows:

Startyear	Start Quarter	Days Between	Participant Minimum	Protestnumber
1992	1	6	33000	21
1992	2	3	30000	24
1992	3	1	100	10
1992	4	1	1000	11
1993	1	2	20000	3
1993	2	3	20100	12
1993	4	2	12000	13
1994	2	1	10000	1
1994	4	4	30100	14
1995	2	3	20100	6
1995	3	1	50	4
1995	4	1	10000	5
1996	1	4	13000	10
1996	4	2	2100	11
1997	1	1	10000	1
1998	2	4	26000	10
1998	3	1	10000	5
1998	4	2	20000	13
1999	1	1	10000	1
1999	4	1	5000	2

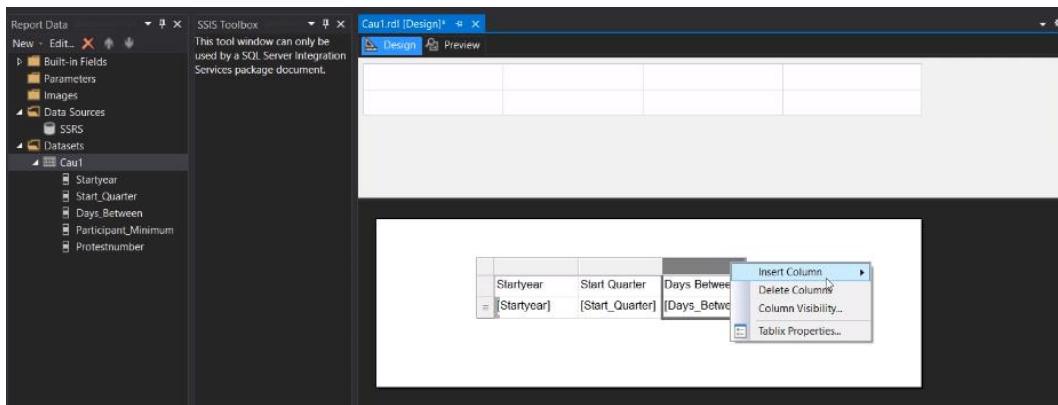
Hình 298. Kết quả chạy truy vấn câu 1

Bước 5: Nhấn chuột phải tại màn hình **Design > Insert > Table**



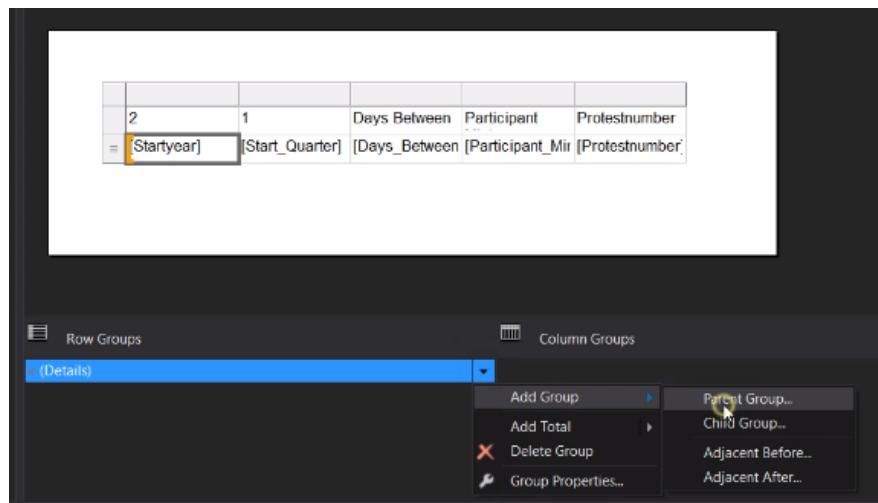
Hình 299. Thực hiện tạo bảng báo cáo

Bước 6: Kéo thả các thuộc tính đã chọn ở bước trên vào bảng màn hình *Desgin* > thêm cột mới cho bảng **Insert Column** để thêm các thuộc tính còn lại



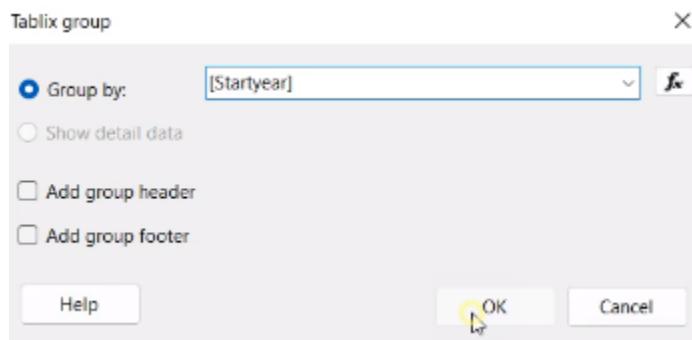
Hình 300. Thực hiện kéo thả thuộc tính vào bảng báo cáo

Bước 7: Thay đổi tên thuộc tính *StartYear*, *StartQuarter*. Chọn ô thuộc tính *StartYear* > Tại thư mục **Row Groups** > **Add Group** > **Parent Group**



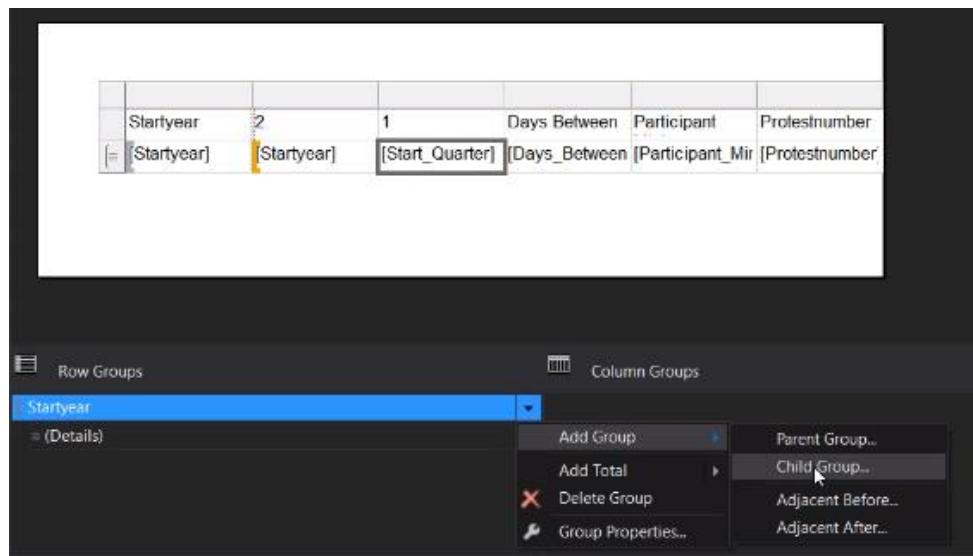
Hình 301. Thêm thuộc tính bằng ‘Parent Group’

Bước 8: Chọn thuộc tính mình muốn sử dụng ‘*Group by*’



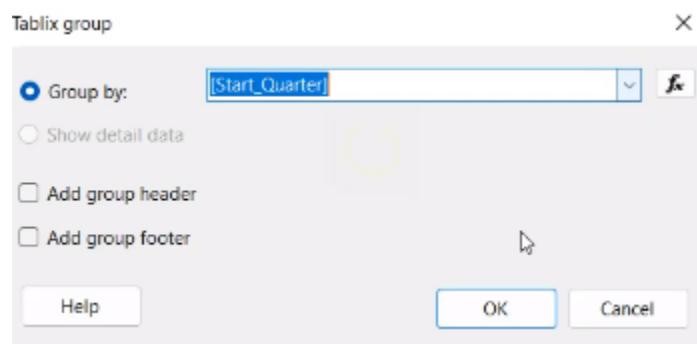
Hình 302. Chọn thuộc tính StartYear để sử dụng ‘*Group by*’

Bước 9: Chọn ô thuộc tính *StartQuarter* > Tại thư mục *Row Groups* > *Add Group* > *Child Group*



Hình 303. Thêm thuộc tính bằng ‘Child Group’

Bước 10: Chọn thuộc tính mình muốn sử dụng ‘**Group by**’



Hình 304. Chọn thuộc tính StartQuarter để sử dụng ‘Group by’

Bước 11: Xóa 2 cột thuộc tính StartQuater, StartYear cũ

Bước 12: Chọn ô thuộc tính *StartQuarter > Add Total > After*



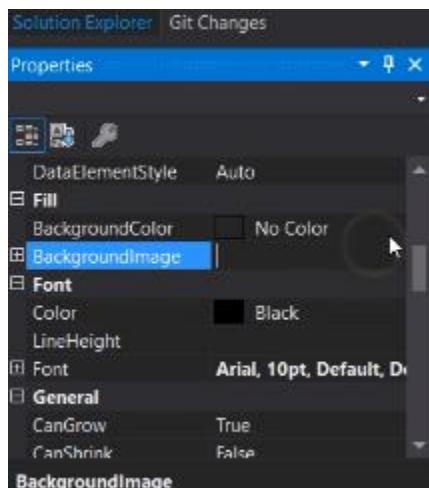
Hình 305. Tạo thêm hàng tổng cho 3 thuộc tính

Bước 13: Thay đổi tên các ô

	Năm	Quý	Số ngày biểu	Số người tham	Số cuộc biểu
[Startyear]	[Start_Quarter]		Days_Between [Participant_Mir	[Protestnumber]	
	Tổng cộng		[Sum(Days_Bet	[Sum(Participar	[Sum(Protestnu
	Tổng cộng		[Sum(Days_Bet	[Sum(Participar	[Sum(Protestnu

Hình 306. Thực hiện đổi tên cột

Bước 14: Thiết kế lại bảng thông qua cửa sổ Properties



Hình 307. Cửa sổ Properties

Kết quả:

Thống kê số cuộc biểu tình, số ngày biểu tình, số người tham gia biểu tình tại Nga từ 1992 đến 1999

Ngày lập báo cáo: 5/13/2022 10:02:13 PM

Năm	Quý	Số ngày biểu tình	Số người tham gia	Số cuộc biểu tình
1992	1	6	33000	21
	2	3	30000	24
	3	1	100	10
	4	1	1000	11
Tổng cộng		11	64100	66
1993	1	2	20000	3
	2	3	20100	12
	4	2	12000	13
	Tổng cộng		7	52100
1994	2	1	10000	1
	4	4	30100	14
	Tổng cộng		5	40100
1995	2	3	20100	6
	3	1	50	4
	4	1	10000	5
	Tổng cộng		5	30150
1996	1	4	13000	10
	4	2	2100	11
	Tổng cộng		6	15100
1997	1	1	10000	1
	Tổng cộng		1	10000
1998	2	4	26000	10
	3	1	10000	5
	4	2	20000	13
	Tổng cộng		7	56000
1999	1	1	10000	1
	4	1	5000	2
	Tổng cộng		2	15000
Tổng cộng		44	282550	177

Hình 308. Báo cáo kết quả truy vấn câu 1

1.3.2 Báo cáo thống kê số người tham gia biểu tình tại Nam Mĩ và Bắc Mĩ trong năm 2019

Bước 1: Sau khi tạo thêm file dataset cho câu 2, tại cửa sổ Query Desgin, kéo thả các thuộc tính liên quan đến câu truy vấn 2 ‘**Region, Country, ParticipantMinimum**’ vào vùng thực thi truy vấn.

Bước 2: Kéo thả thuộc tính Location, Date theo điều kiện câu truy vấn là tại ‘NorthAmerica, South America’ và thời gian 2019

Dimension	Hierarchy	Operator	Filter Expression	Param...
Location	Region	Equal	{ North America, South America }	<input type="checkbox"/>
Date	Startyear	Equal		<input type="checkbox"/>
<Select dimension>				

Region	Country	Participant Minimum
North America	Canada	2250
North America	Cuba	200
North America	Haiti	59300
North America	Mexico	2400
South America	Argenti...	100
South America	Bolivia	90250
South America	Brazil	10150
South America	Chile	90100
South America	Colom...	50000
South America	Ecuador	65250
South America	Paraguay	2000
South America	Peru	4000
South America	Venezu...	32450

Hình 309. Điều chỉnh dimension ở câu 2

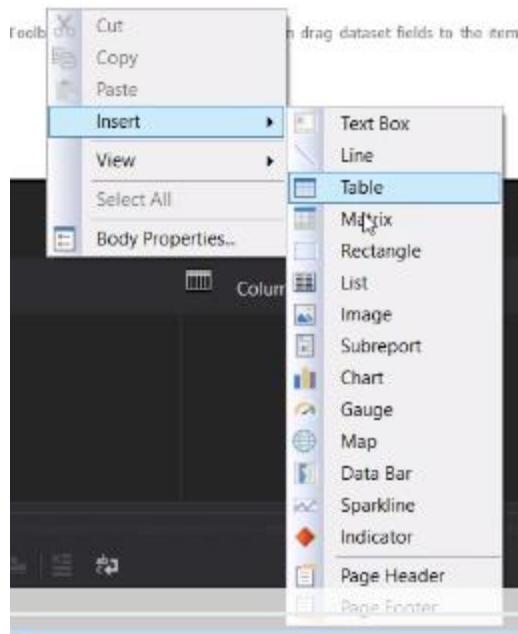
Bước 3: Chọn ‘Run the query’ để xem hiện kết quả truy vấn

Dimension	Hierarchy	Operator	Filter Expression	Param...
Location	Region	Equal	{ North America, South America }	<input type="checkbox"/>
Date	Startyear	Equal	{ 2019 }	<input type="checkbox"/>
<Select dimension>				

Region	Country	Participant Minimum
North America	Canada	2250
North America	Cuba	200
North America	Haiti	59300
North America	Mexico	2400
South America	Argenti...	100
South America	Bolivia	90250
South America	Brazil	10150
South America	Chile	90100
South America	Colom...	50000
South America	Ecuador	65250
South America	Paraguay	2000
South America	Peru	4000
South America	Venezu...	32450

Hình 310. Kết quả chạy truy vấn câu 2

Bước 4: Nhấn chuột phải tại màn hình **Design > Insert > Table**



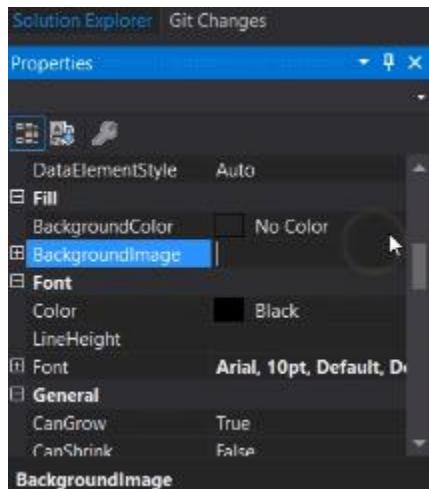
Hình 311. Thực hiện tạo bảng báo cáo

Bước 5: Kéo thả các thuộc tính đã chọn ở bước trên vào bảng màn hình **Desgin >** thêm cột mới cho bảng **Insert Column** để thêm các thuộc tính còn lại. Thực hiện thêm các thuộc tính Group by tương tự như câu truy vấn thứ nhất.

	Châu lục	Quốc gia	Số người tham
[=]	[Region]	[Country]	[Participant_Minin
	Tổng cộng		[Sum(Participant_
	Tổng cộng		[Sum(Participant_

Hình 312. Thực hiện kéo thả và điều chỉnh thuộc tính cho bảng

Bước 6: Thiết kế lại bảng thông qua cửa sổ Properties



Hình 313. Cửa sổ Properties

Kết quả:

**Thống kê số người tham gia biểu tình tại Nam Mĩ và Bắc Mĩ
trong 2019**

Ngày lập báo cáo: 5/13/2022 10:12:34 PM

Châu lục	Quốc gia	Số người tham gia
North America	Canada	2250
	Cuba	200
	Haiti	59300
	Mexico	2400
	Tổng cộng	64150
South America	Argentina	100
	Bolivia	90250
	Brazil	10150
	Chile	90100
	Colombia	50000
	Ecuador	65250
	Paraguay	2000
	Peru	4000
	Venezuela	32450
	Tổng cộng	344300
Tổng cộng		408450

Hình 314. Báo cáo kết quả truy vấn câu 2

1.3.3 Báo cáo Thống kê số người tham gia biểu tình ở MENA trong năm 2019 - 2020 (theo quý năm)

Bước 1: Sau khi tạo thêm file dataset cho câu 3, tại cửa sổ Query Design, kéo thả các thuộc tính liên quan đến câu truy vấn 3 ‘*StartYear*, *StartQuarter*, *StartMonth*, *Region*, *Country*, *ParticipantMinumum*’ vào vùng thực thi truy vấn.

Bước 2: Kéo thả thuộc tính Location, Date theo điều kiện câu truy vấn số 3 như địa điểm tại ‘Central America’, khoảng thời gian 2019-2020

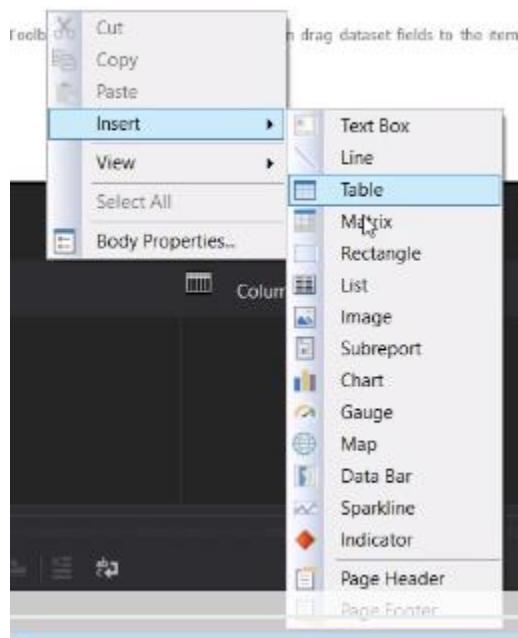
Bước 3: Chọn ‘*Run the query*’ để xem hiện kết quả truy vấn

The screenshot shows the Microsoft Power BI Query Designer interface. On the left, there's a navigation pane with 'Protest Dw v9' selected. Under 'Measures', 'Fact' is expanded, showing 'Days Between', 'Fact Count', 'Participant Minimu', and 'Protestnumber'. Under 'Date', 'Start Year' is selected. The main area displays a table with the following data:

Startyear	Start Quarter	Startmonth	Region	Country	Participant Minimum
2019	1	1	Central America	Guatemala	100
2019	1	2	Central America	Nicaragua	50
2019	1	3	Central America	Nicaragua	150
2019	2	4	Central America	Honduras	15000
2019	2	4	Central America	Nicaragua	200
2019	2	5	Central America	Honduras	10000
2019	2	6	Central America	Honduras	10100
2019	3	7	Central America	Guatemala	150
2019	3	8	Central America	Guatemala	2000
2019	3	9	Central America	Nicaragua	100
2019	4	10	Central America	Honduras	6000

Hình 315. Kết quả chạy truy vấn câu 3

Bước 4: Nhấn chuột phải tại màn hình *Design* > *Insert* > *Table*



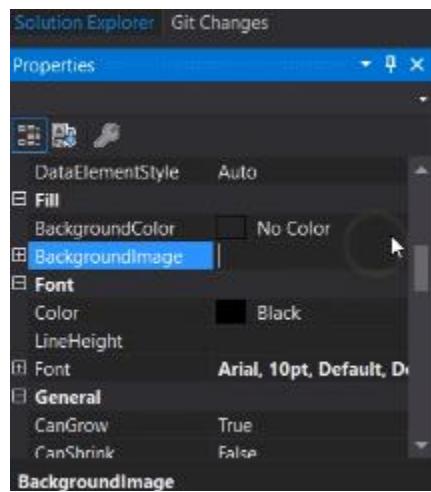
Hình 316. Thực hiện tạo bảng báo cáo

Bước 5: Kéo thả các thuộc tính đã chọn ở bước trên vào bảng màn hình *Desgin* > thêm cột mới cho bảng ***Insert Column*** để thêm các thuộc tính còn lại. Thực hiện thêm các thuộc tính Group by tương tự như câu truy vấn thứ nhất.

[Region]				
Năm	Quý	Tháng	[Country]	Tổng
[Startyear]	[Start_Quarter]	[Startmonth]	[Sum(Participar	[Sum(Participar
		Tổng	[Sum(Participar	[Sum(Participar
		Tổng	[Sum(Participar	[Sum(Participar
Tổng			[Sum(Participar	[Sum(Participar

Hình 317. Thực hiện kéo thả và điều chỉnh thuộc tính cho bảng

Bước 6: Thiết kế lại bảng thông qua cửa sổ Properties



Hình 318. Cửa sổ Properties

Kết quả:

**Thống kê số người tham gia biểu tình ở Central America
trong năm 2019 - 2020 (theo tháng quý năm)**

Ngày lập báo cáo: 5/13/2022 10:52:30 PM

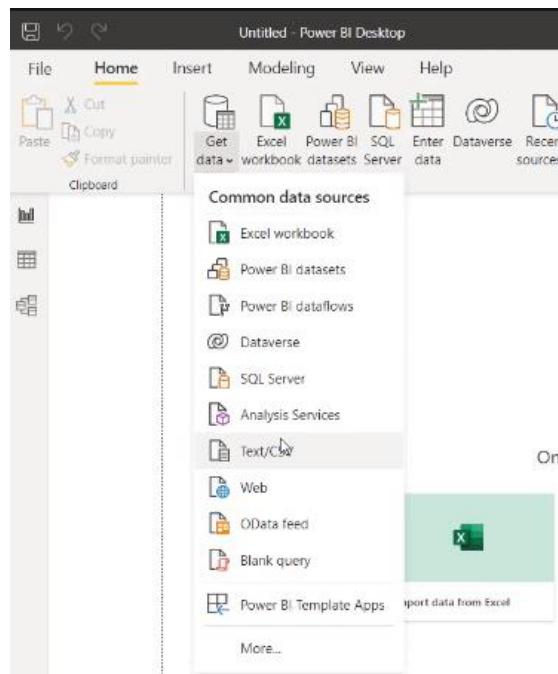
Central America						
Năm	Quý	Tháng	Guatemala	Honduras	Nicaragua	Tổng
2019	1	1	100			100
		2			50	50
		3			150	150
		Tổng	100		200	300
	2	4		15000	200	15200
		5		10000		10000
		6		10100		10100
		Tổng		35100	200	35300
	3	7	150			150
		8	2000			2000
		9			100	100
		Tổng	2150		100	2250
	4	10		6000		6000
		Tổng		6000		6000
		Tổng	2250	41100	500	43850
Tổng			2250	41100	500	43850

Hình 319. Báo cáo kết quả truy vấn câu 3

2. QUÁ TRÌNH TẠO BÁO CÁO BẰNG CÔNG CỤ POWER BI

2.1 Nhập dữ liệu nguồn (Data Source)

Bước 1: Mở Power *BI Desktop*, tại mục *Home > Get Data > Analysis Service* để import dữ liệu đã xử lý qua quá trình SSIS và SSAS



Hình 320. Thực hiện import data source vào Power BI

Bước 2: Tại màn hình vừa hiện lên, dán tên server của SQL Server máy mình > Nhập tên Database mình muốn sử dụng

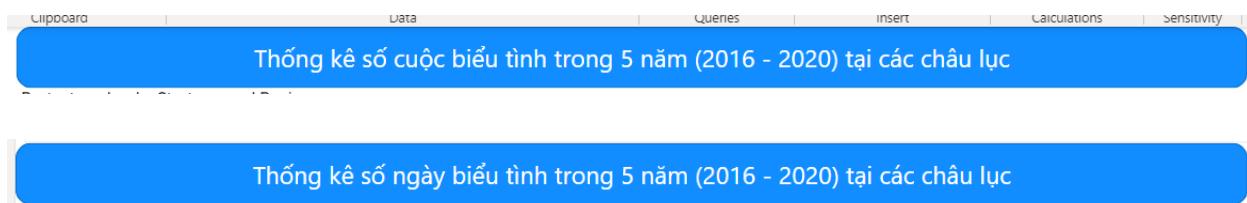


Hình 321. Điện thông tin đến cơ sở dữ liệu nguồn

2.2 Các bước thực hiện

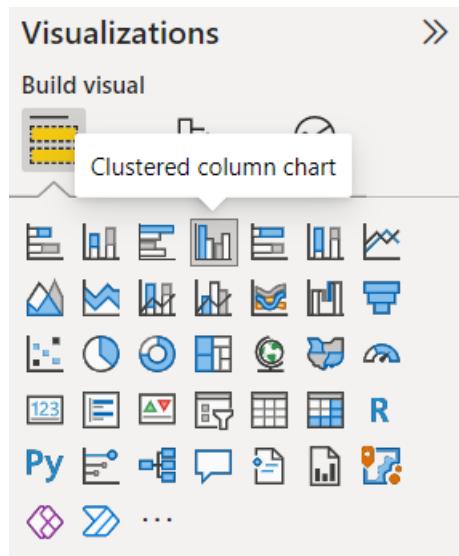
2.2.1 Thống kê số cuộc biểu tình, số ngày biểu tình trong 5 năm (2016 – 2020) tại các châu lục

Bước 1: Tạo header cho Report



Hình 322. Thêm Header cho Report 1 (Power BI)

Bước 2: Tại cửa sổ *Visualizations* > Chọn biểu đồ ‘*Clustered column chart*’



Hình 323. Chọn kiểu báo cáo (đường và cột)

Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Region, Protestnumber, Startyear)

- Tại thuộc tính Startyear chọn 5 năm từ 2016 đến 2020

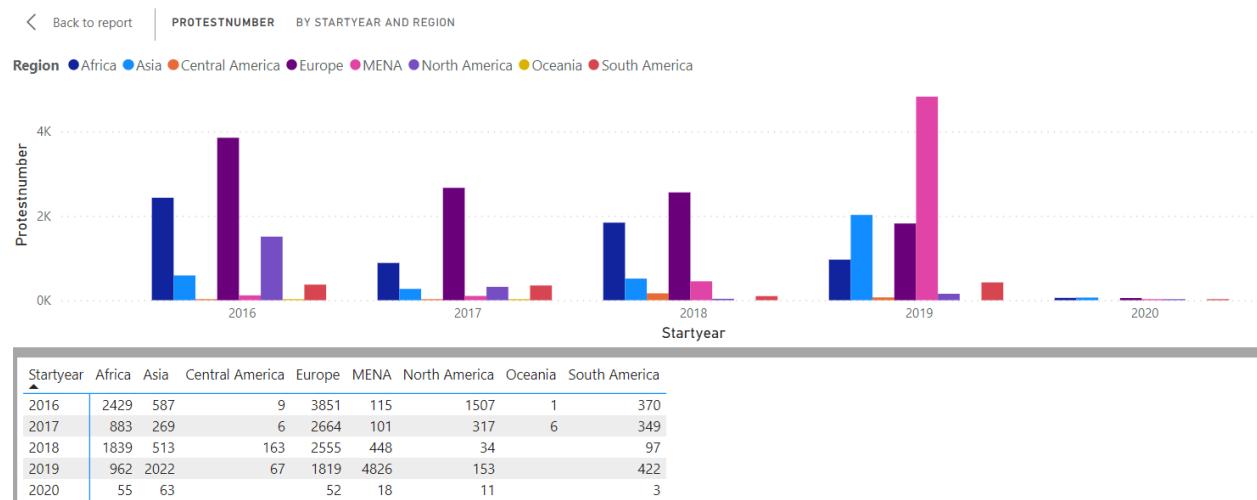
The image shows two identical filter panels side-by-side. Each panel has a header section with fields like 'Protestnumber' and 'Region'. Below this is a 'Startyear' section where 'is 2016, 2017, 2018, 2...' is selected. Underneath is a 'Filter type' dropdown set to 'Basic filtering' with a search bar. The main area contains a list of years from 2015 to 2020, with 2016, 2017, 2018, and 2019 checked, and 2020 also checked.

Hình 324. Chọn các thuộc tính truy vấn câu 2.2.1

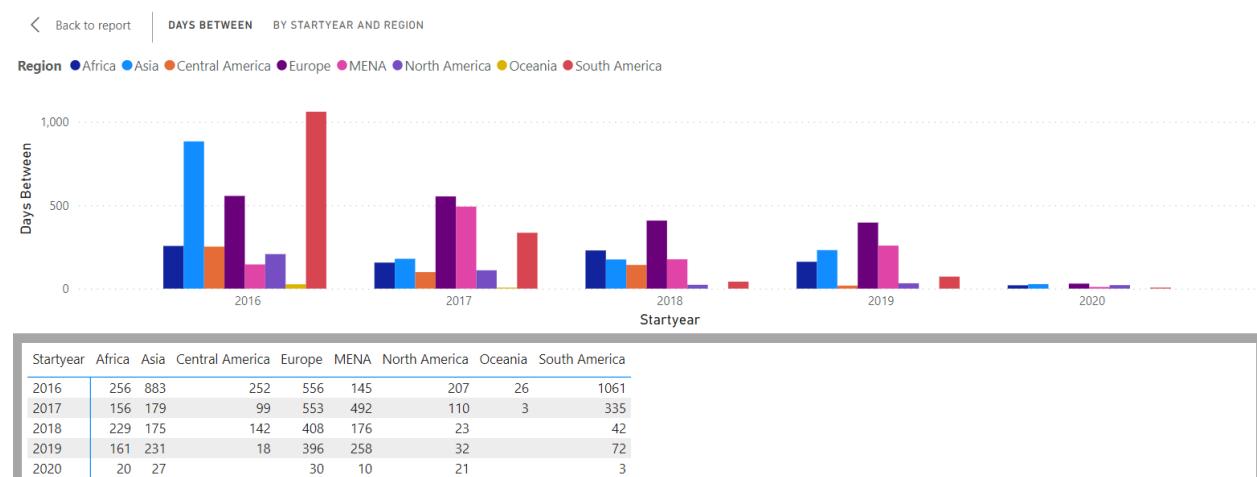
The image shows two identical visualization configuration panels side-by-side. Each panel has an 'X-axis' section labeled 'Startyear' and a 'Y-axis' section labeled 'Protestnumber'. Below these are sections for 'Legend' (labeled 'Region'), 'Small multiples', and an 'Add data fields here' button. The panels are identical in structure and content.

Hình 325. Các thuộc tính để biểu diễn cho biểu đồ

Bước 4: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

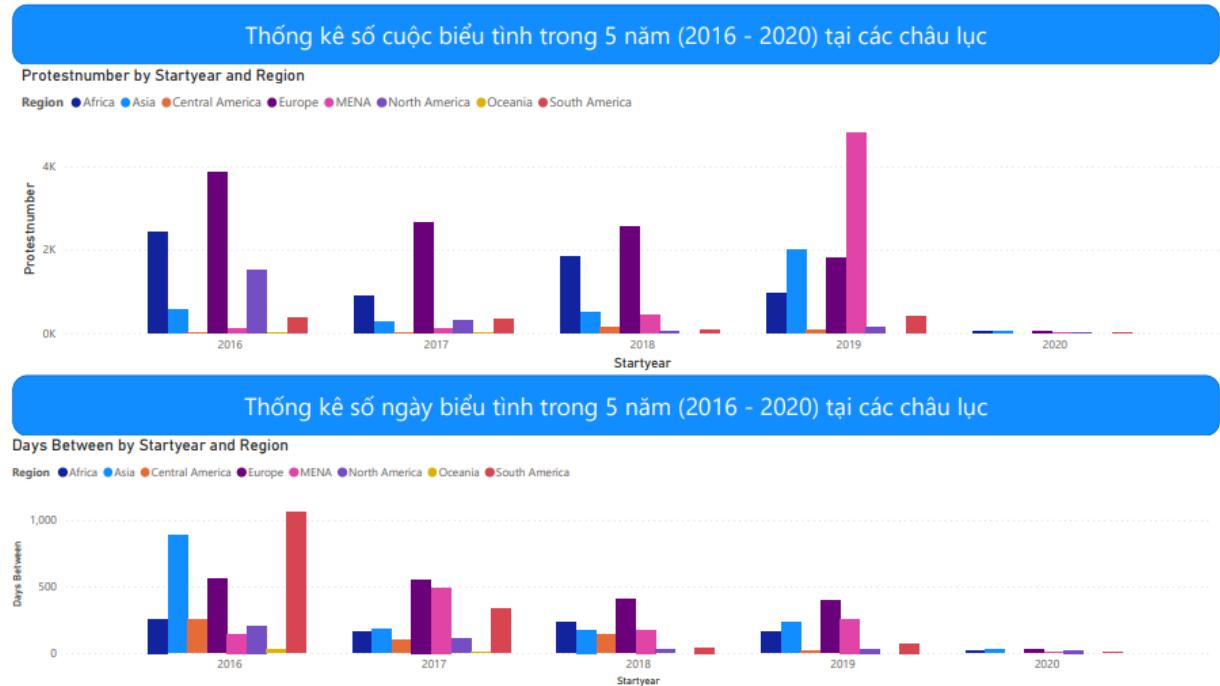


Hình 326. Biểu đồ cột liên cụm thể hiện số cuộc biểu tình tại các châu lục trong 2016-2020



Hình 327. Biểu đồ cột liên cụm thể hiện số ngày biểu tình tại các châu lục trong 2016-2020

- Kết quả hoàn chỉnh:

*Hình 328. Report 1*

Bảng thống kê số cuộc biểu tình, số ngày biểu tình tại các châu lục trong 5 năm 2016 - 2020

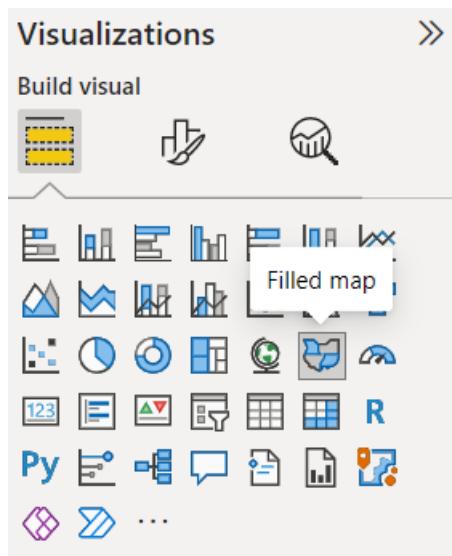
2.2.2 Thống kê số người tham gia các cuộc biểu tình ôn hòa tại MENA

Bước 1: Tạo header cho Report

Thống kê số người tham gia các cuộc biểu tình ôn hòa tại MENA

Hình 329. Thêm Header cho Report 2 (Power BI)

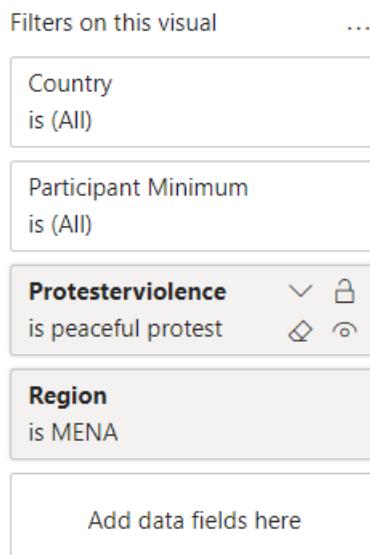
Bước 2: Tại cửa sổ **Visualizations** > Chọn biểu đồ ‘Filled map’



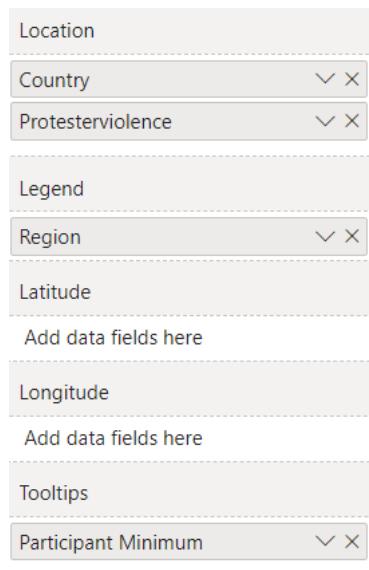
Hình 330. Chọn kiểu báo cáo

Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Country, Region, Participant Minimum, Protestviolence)

- Tại thuộc tính Region chọn MENA
- Tại thuộc tính Protestviolence chọn peaceful protest (biểu tình ôn hòa)



Hình 331. Chọn các thuộc tính truy vấn câu 2.2.2



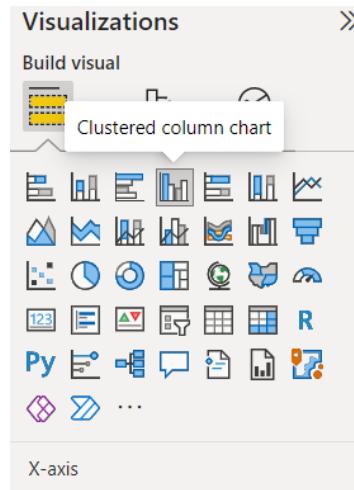
Hình 332. Các thuộc tính để biểu diễn cho biểu đồ

Bước 4: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:



Hình 333. Biểu đồ dạng bản đồ thể hiện quốc gia xảy ra cuộc biểu tình ôn hòa tại MENA

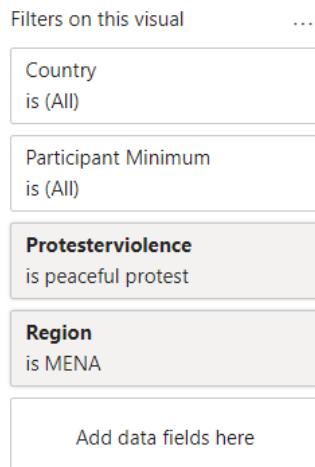
Bước 5: Tạo biểu đồ thứ hai - Tại cửa sổ **Visualizations** > Chọn biểu đồ ‘**Clustered column chart**’



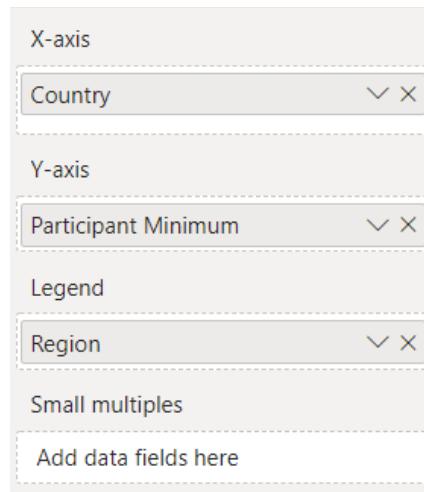
Hình 334. Chọn kiểu biểu đồ

Bước 6: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Country, Region, Participant Minimun, Protestviolence)

- Tại thuộc tính Region chọn MENA
- Tại thuộc tính Protestviolence chọn peaceful protest (biểu tình ôn hòa)

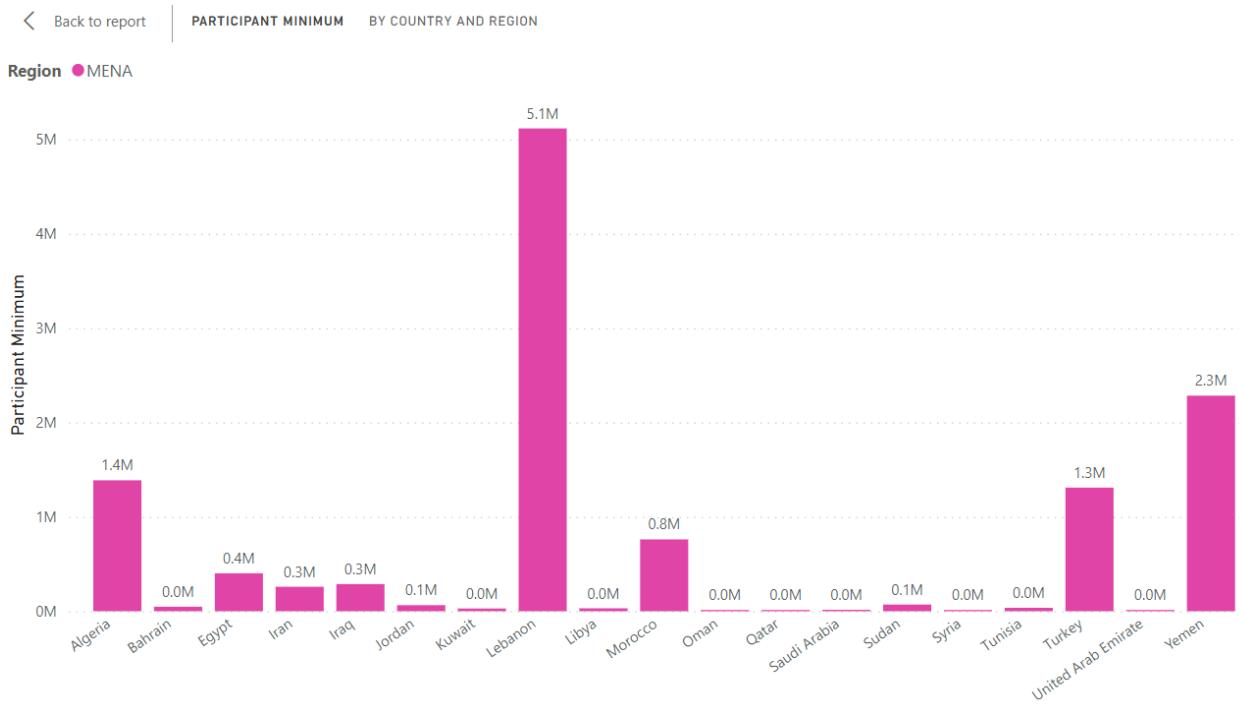


Hình 335. Các thuộc tính để biểu diễn cho biểu đồ



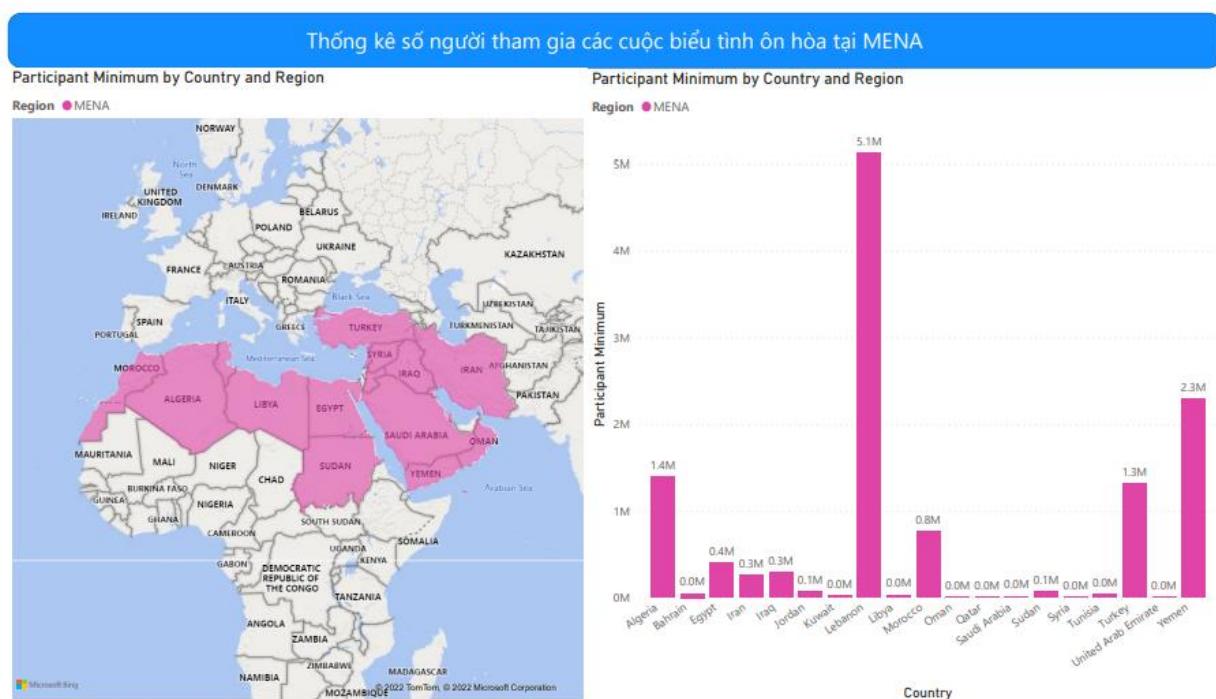
Hình 336. Các thuộc tính để biểu diễn cho biểu đồ

Bước 7: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:



Hình 337. Biểu đồ dạng cột thể hiện quốc gia xảy ra cuộc biểu tình ôn hòa tại MENA

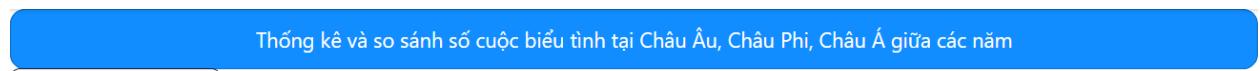
○ *Kết quả hoàn chỉnh:*



Hình 338. Thống kê số người tham gia các cuộc biểu tình ôn hòa tại MENA

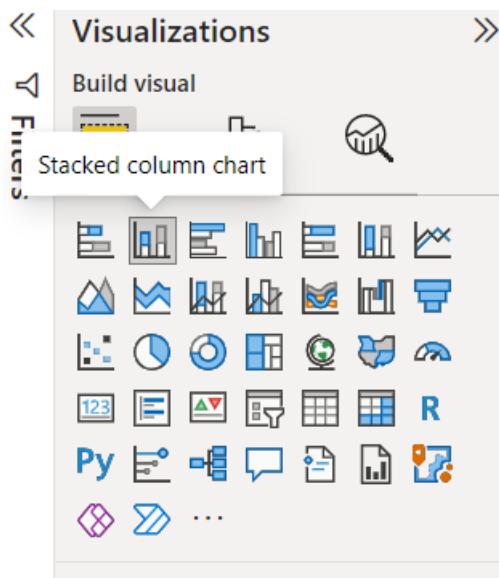
2.2.3 Thống kê và so sánh số cuộc biểu tình tại Châu Âu, Châu Phi và Châu Á giữa các năm

Bước 1: Tạo header cho Report



Hình 339. Thêm Header cho Report 3 (Power BI)

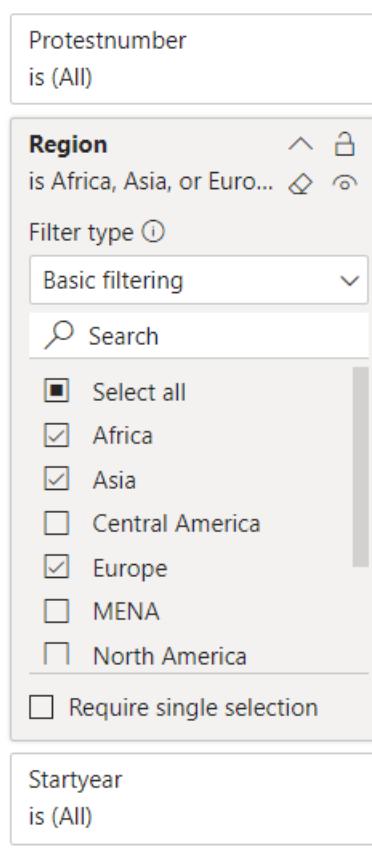
Bước 2: Tại cửa sổ **Visualizations** > Chọn biểu đồ ‘Stacked column chart’



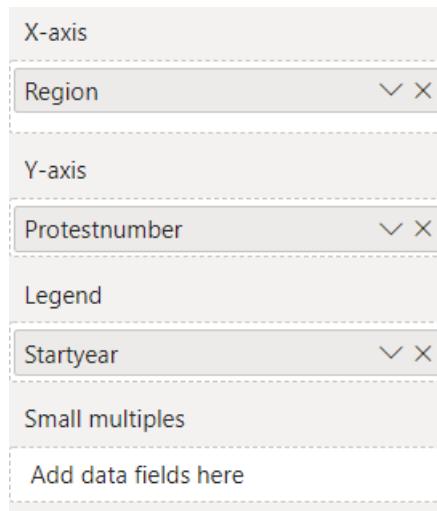
Hình 340. Chọn kiểu báo cáo

Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Startyear, Region, Protestnumber)

- Tại thuộc tính Region chọn Châu Á, Châu Âu, Châu Phi



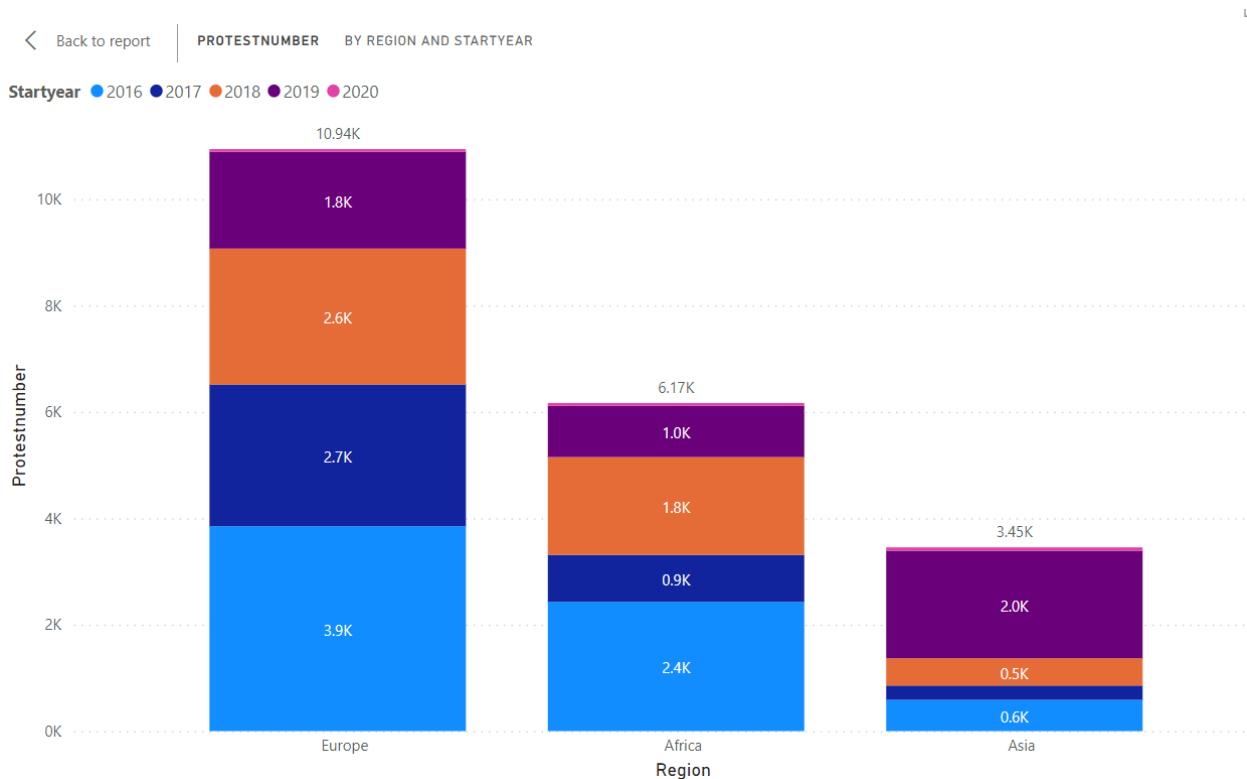
Hình 341. Chọn các thuộc tính truy vấn câu 2.2.3



Hình 342. Các thuộc tính để biểu diễn cho biểu đồ

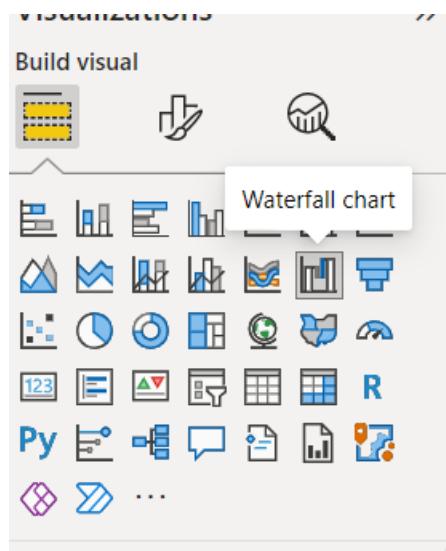
Bước 4: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

* (biểu đồ lấy ví dụ 5 năm 2016 -2020)



Hình 344. Biểu đồ dạng cột chồng thể hiện số cuộc biểu tình diễn ra tại Châu Âu, Châu Phi, Châu Á qua các năm

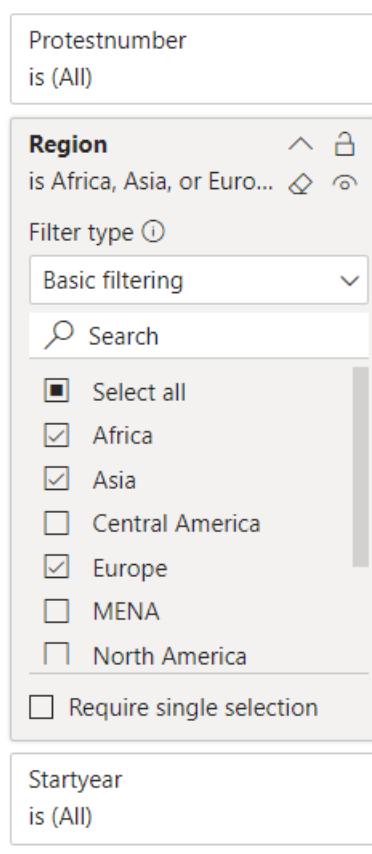
Bước 5: Tạo biểu đồ thứ hai - Tại cửa sổ **Visualizations** > Chọn biểu đồ ‘Waterfall chart’



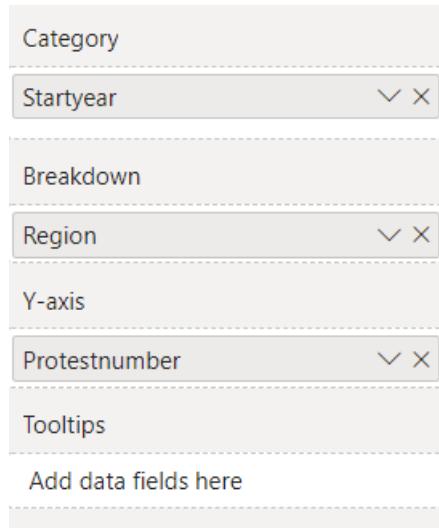
Hình 345. Chọn kiểu biểu đồ

Bước 6: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Startyear, Region, Protestnumber)

- Tại thuộc tính Region chọn Châu Á, Châu Âu, Châu Phi



Hình 346. Chọn các thuộc tính truy vấn câu 2.2.3



Hình 347. Các thuộc tính để biểu diễn cho biểu đồ

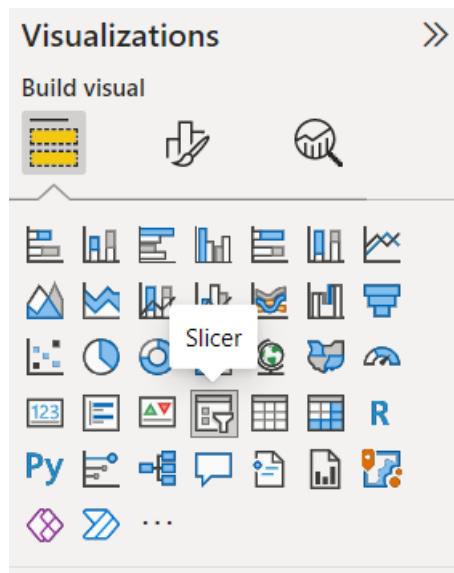
Bước 7: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

* (biểu đồ lấy ví dụ 5 năm 2016 -2020)



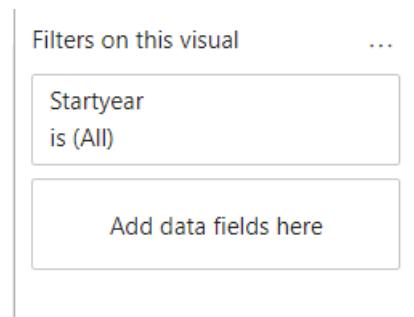
Hình 348. Biểu đồ dạng thác nước so sánh sự tăng giảm số cuộc biểu tình tại Châu Âu, Châu Phi, Châu Á với 2 năm liền kề nhau

Bước 8: Tại cửa sổ **Visualizations** > Chọn biểu đồ ‘**Stacked column chart**’



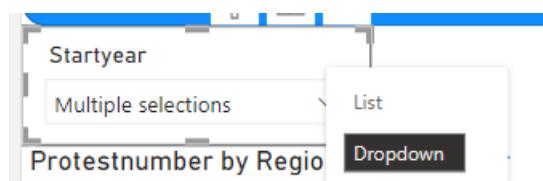
Hình 349. Chọn kiểu báo cáo

Bước 9: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Startyear)



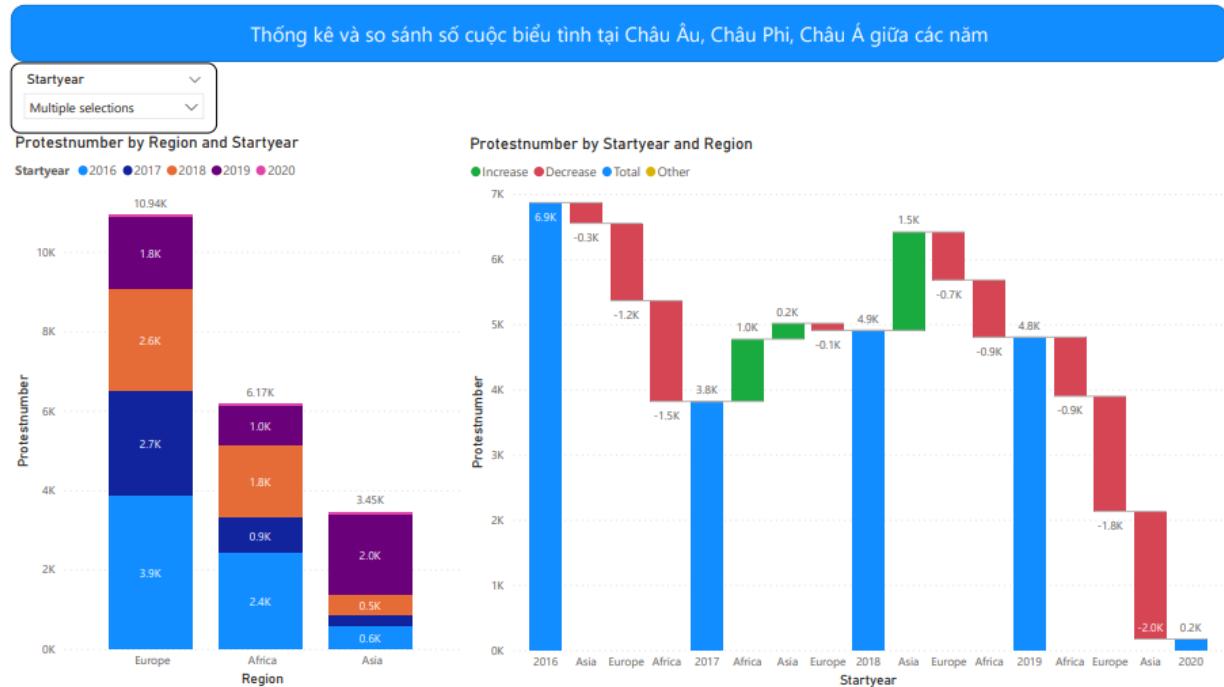
Hình 350. Chọn thuộc tính cho bộ lọc

Bước 10: Chọn kiểu cho bộ lọc là Dropdown



Hình 351. Chọn loại cho bộ lọc

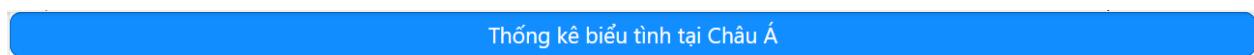
- Kết quả hoàn chỉnh:
- * (biểu đồ lấy ví dụ 5 năm 2016 -2020)



Hình 352. Thống kê và so sánh số cuộc biểu tình tại Châu Âu, Châu Phi và Châu Á giữa các năm

2.2.4 Thống kê biểu tình tại Châu Á

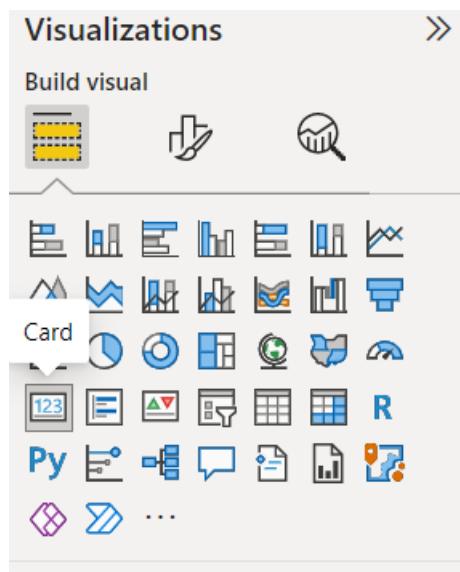
Bước 1: Tạo header cho Report



Hình 353. Thêm Header cho Report 4 (Power BI)

Bước 2: Tại cửa sổ *Visualizations* > Chọn biểu đồ ‘Card’ –

- Ta kéo thả 3 biểu đồ Card



Hình 354. Chọn kiểu báo cáo

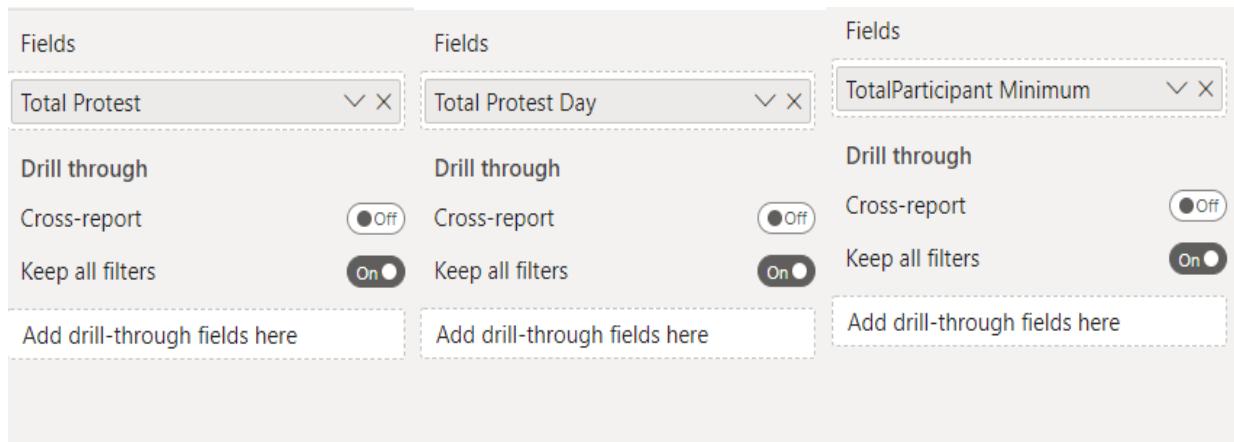
Bước 3: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Region, Fact count¹, Days Between², Participant Minimun³)

- Tại thuộc tính Region chọn Châu Á (Asia)
- Fact Count đổi tên thành Total Protets
- Days Between đổi tên thành Total Protest Day
- Participant Minimun đổi tên thành Total Participant Minimun

The screenshot shows the 'Fields' pane with three columns of filters:

- Column 1 (Left):** Filters on this visual
Region: is Asia
Total Protest: is (All)
Add data fields here
- Column 2 (Middle):** Filters on this visual
Region: is Asia
Total Protest Day: is (All)
Add data fields here
- Column 3 (Right):** Filters on this visual
Region: is Asia
TotalParticipant Minim...: is (All)
Add data fields here

Hình 355. Chọn các thuộc tính truy vấn câu 2.2.4



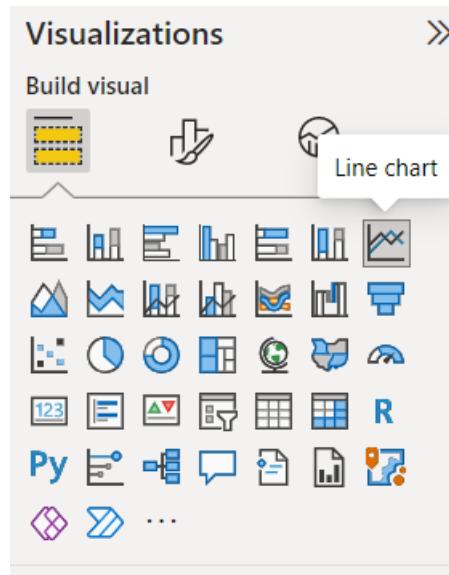
Hình 356. Các thuộc tính để biểu diễn cho biểu đồ

Bước 4: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:



Hình 357. Biểu đồ dạng thẻ thể hiện số cuộc biểu tình diễn ra, số ngày diễn ra, số người tham gia biểu tình tại Châu Á

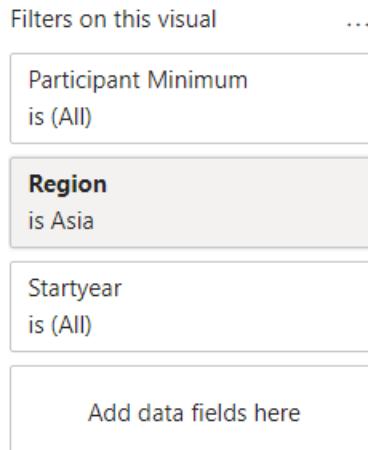
Bước 5: Tạo biểu đồ thứ hai - Tại cửa sổ **Visualizations** > Chọn biểu đồ ‘Line Chart’



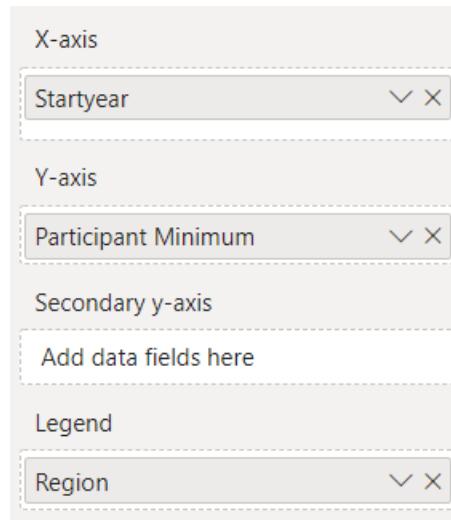
Hình 358. Chọn kiểu biểu đồ

Bước 6: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Startyear, Region, Participant Minimum)

- Tại thuộc tính Region chọn Châu Á



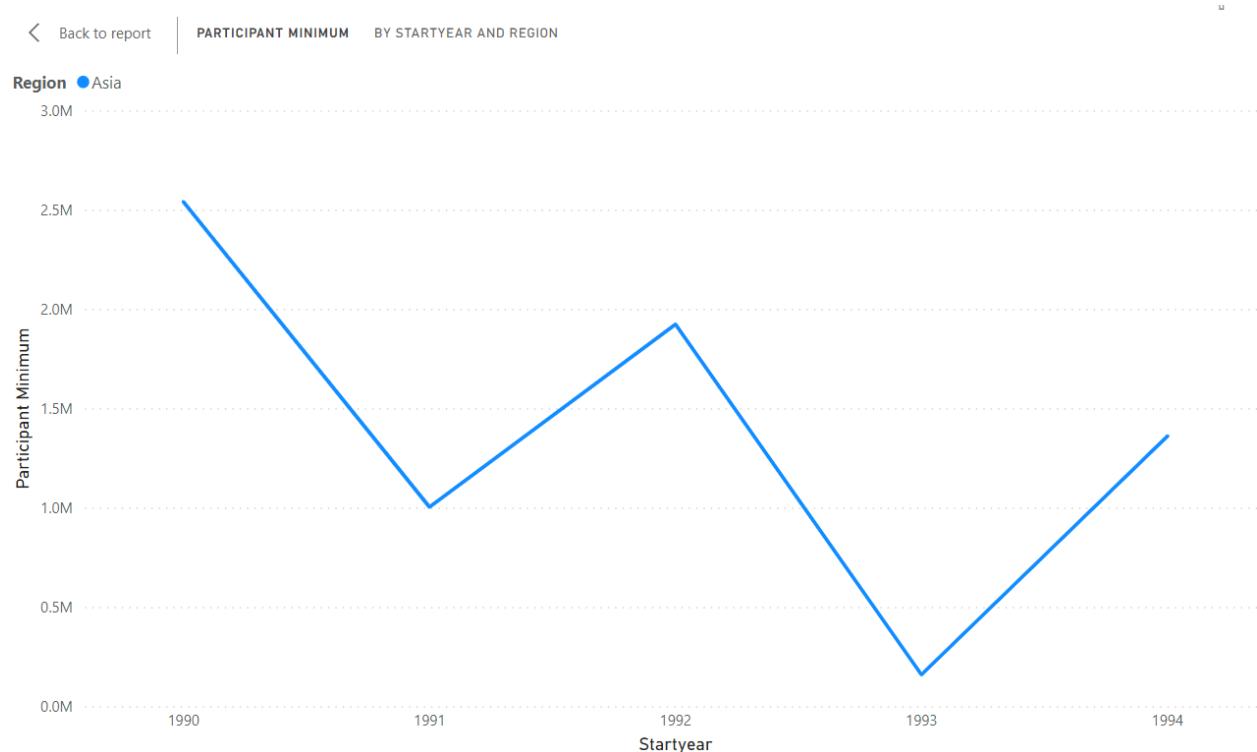
Hình 359. Chọn các thuộc tính truy vấn câu 2.2.4



Hình 360. Các thuộc tính để biểu diễn cho biểu đồ

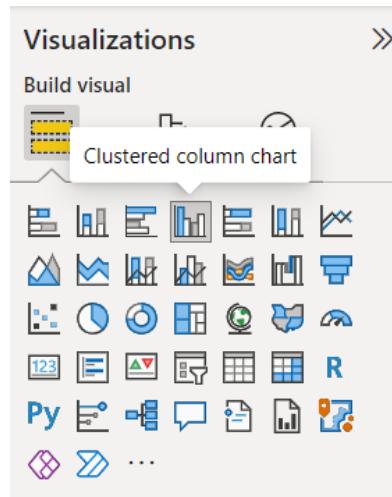
Bước 7: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

*(Biểu đồ lấy ví dụ từ 1990 – 1994)



Hình 361. Biểu đồ dạng đường thể hiện sự thay đổi của số người tham gia biểu tình tại Châu Á qua các năm

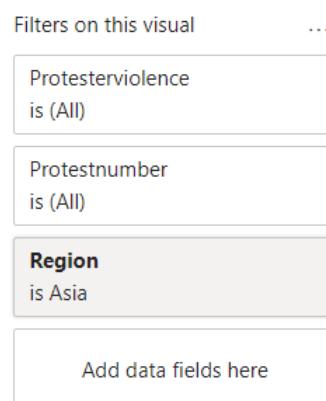
Bước 8: Tại cửa sổ **Visualizations** > Chọn biểu đồ ‘**Clustered column chart**’



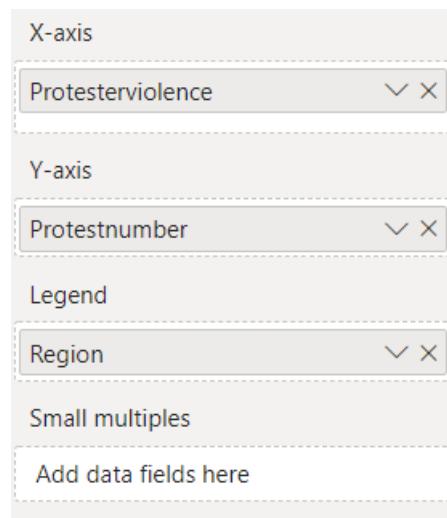
Hình 362. Chọn kiểu báo cáo

Bước 9: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Protest Violence, Region, Participant Minimun)

- Tại thuộc tính Region chọn Châu Á



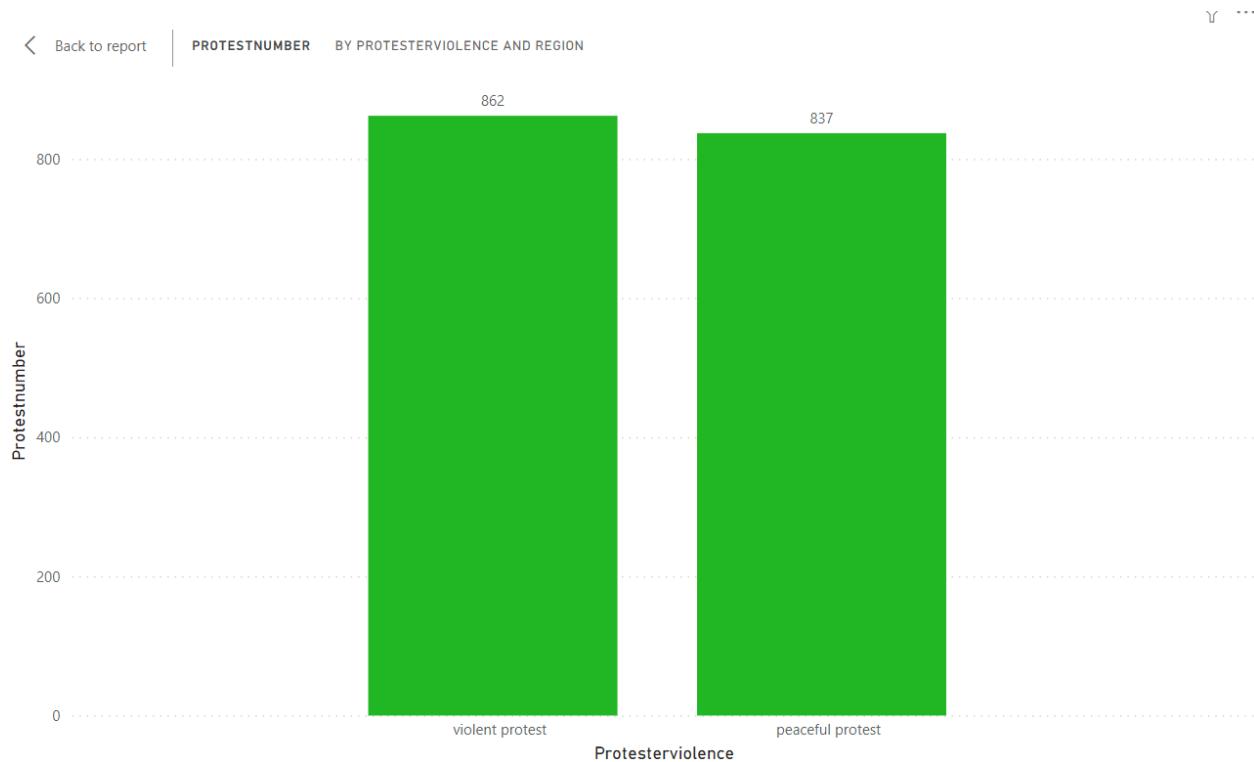
Hình 363. Chọn thuộc tính cho bộ lọc



Hình 364. Các thuộc tính để biểu diễn cho biểu đồ

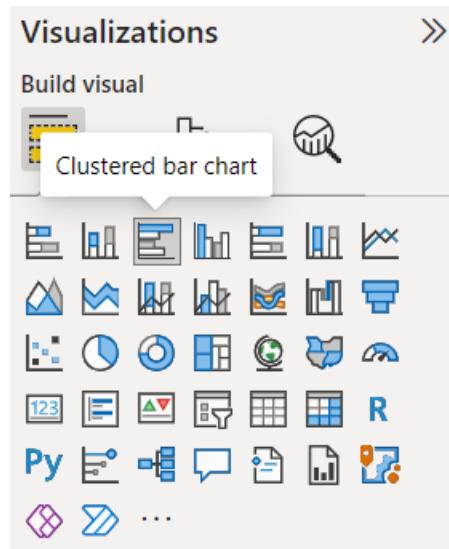
Bước 10: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

*(Biểu đồ lấy ví dụ từ 1990 – 1994)



Hình 365. Biểu đồ dạng cột liên cụm thể hiện số người tham gia ở 2 loại biểu tình ôn hòa và bạo lực vũ trang tại Châu Á

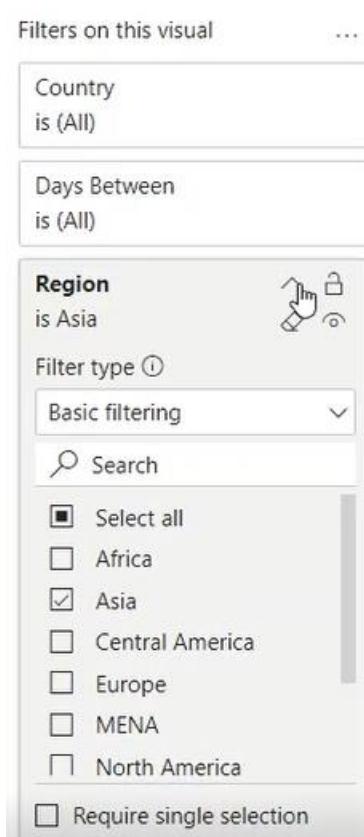
Bước 11: Tại cửa sổ *Visualizations* > Chọn biểu đồ ‘Clustered bar chart’



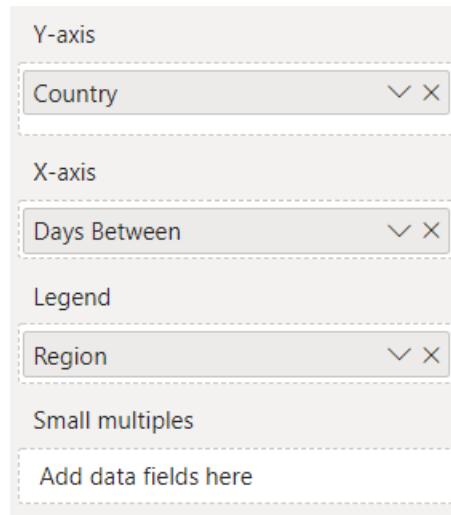
Hình 366. Chọn kiểu báo cáo

Bước 10: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Protest Violence, Region, Participant Minimun)

- Tại thuộc tính Region chọn Châu Á



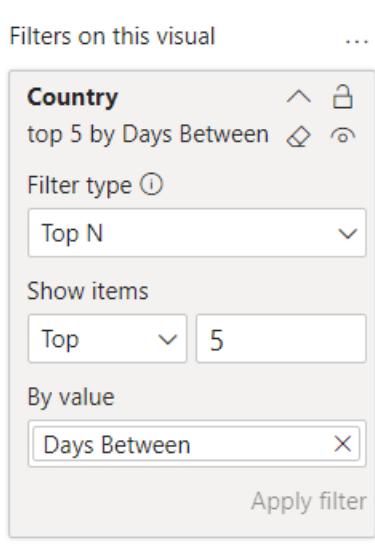
Hình 367. Chọn thuộc tính cho bộ lọc



Hình 368. Các thuộc tính để biểu diễn cho biểu đồ

Bước 12: Tạo điều kiện lấy ra Top 5 quốc gia có số ngày biểu tình cao nhất tại Châu Á

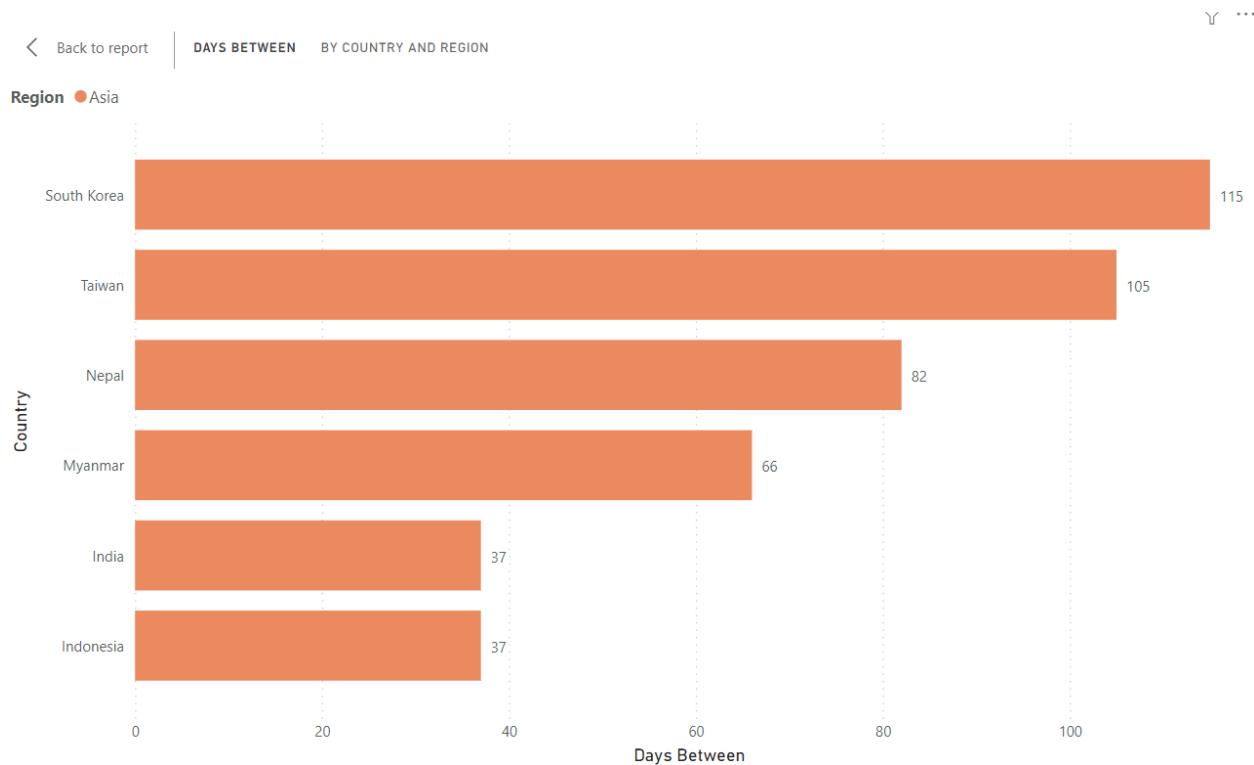
- Tại Filter type chọn Top N
- Show items: chọn loại. Ở đây chọn là Top và thứ tự là 5
- By value: kéo thả thuộc tính Days Between từ cửa sổ Fields



Hình 369. Tạo điều kiện

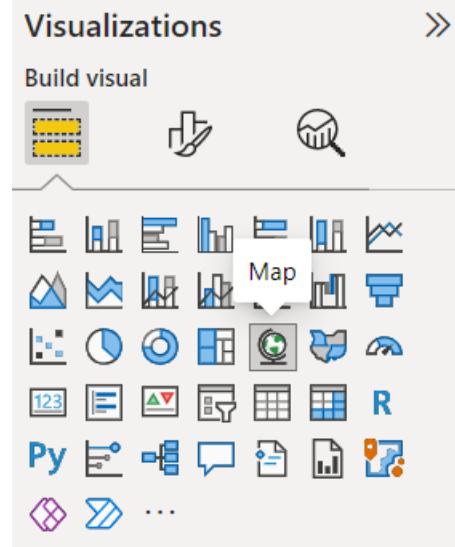
Bước 13: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

*(Biểu đồ lấy ví dụ từ 1990 – 1994)



Hình 370. Biểu đồ dạng thanh chồng thẻ hiện giá trị Top 5 quốc gia có số ngày biểu tình cao nhất tại Châu Á

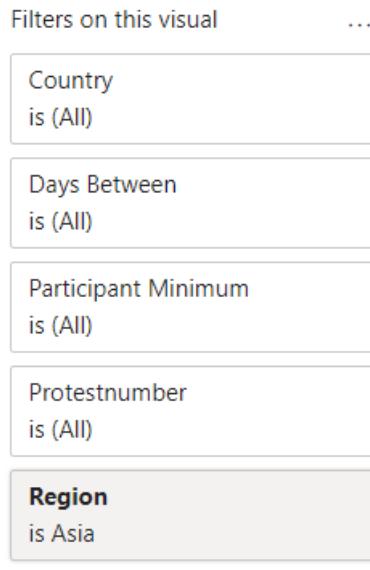
Bước 14: Tại cửa sổ *Visualizations* > Chọn biểu đồ ‘Map’



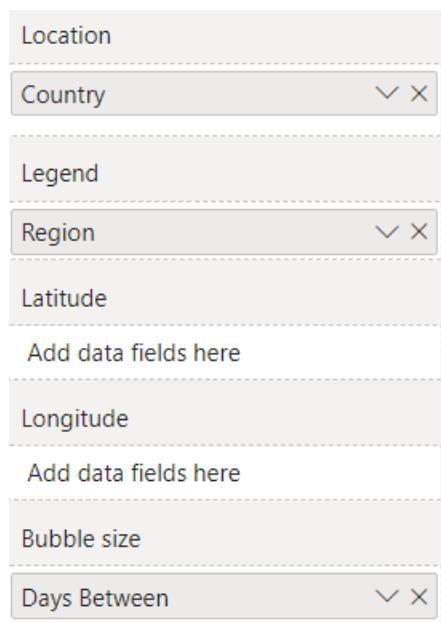
Hình 371. Chọn kiểu báo cáo

Bước 15: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Country, Region, Participant Minimum, Days Between, Protestnumber)

- Tại thuộc tính Region chọn Châu Á



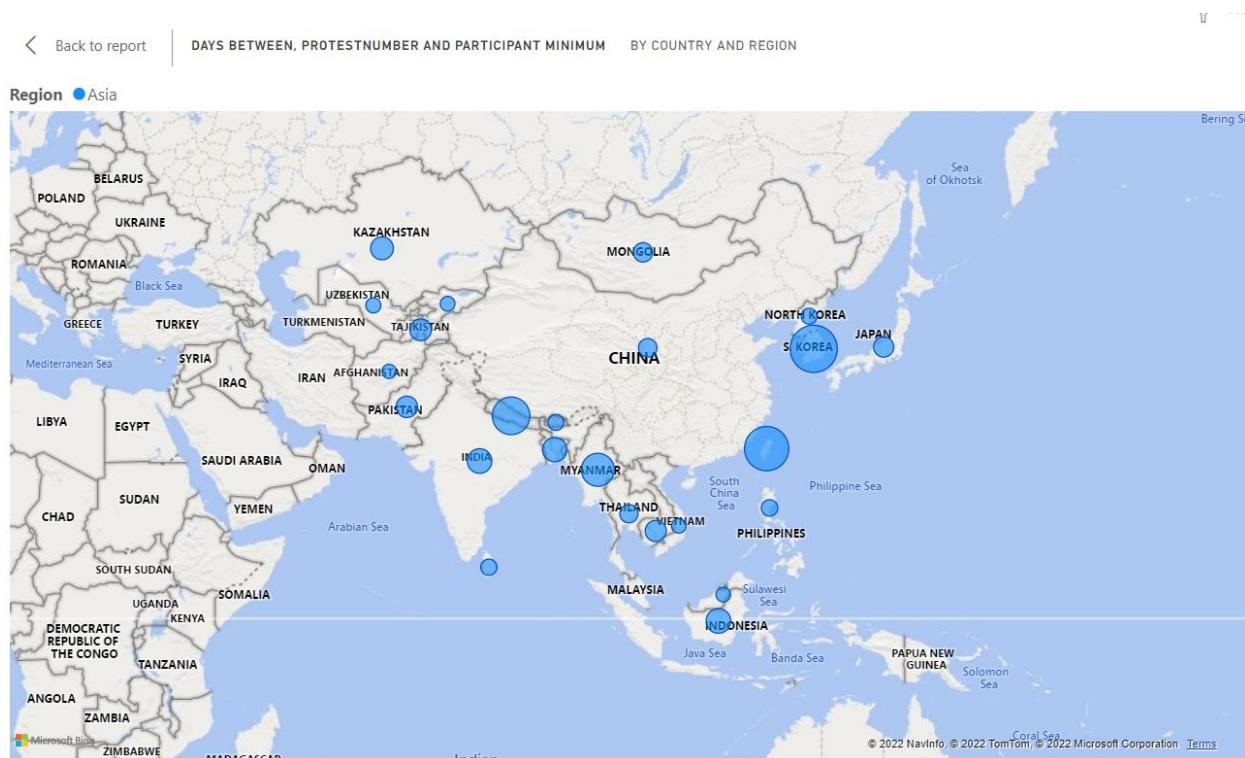
Hình 372. Chọn thuộc tính cho bộ lọc



Hình 373. Các thuộc tính để biểu diễn cho biểu đồ

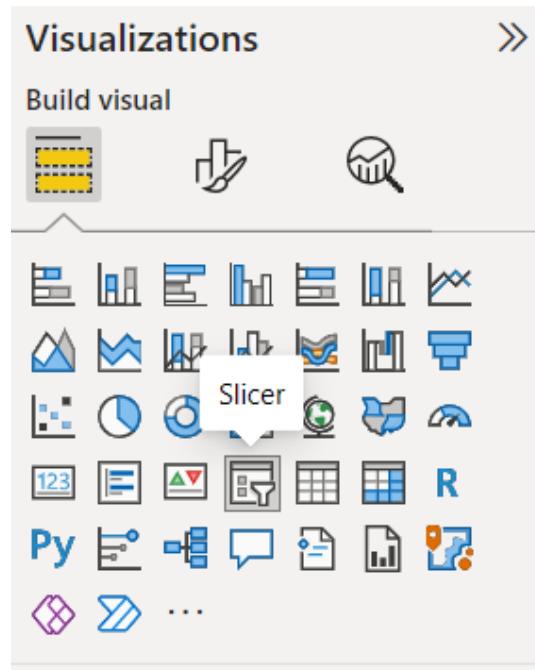
Bước 16: Sau khi đã kéo thả các thuộc tính cần thiết vào biểu đồ thì biểu đồ xuất ra kết quả:

**(Biểu đồ lấy ví dụ từ 1990 – 1994)*



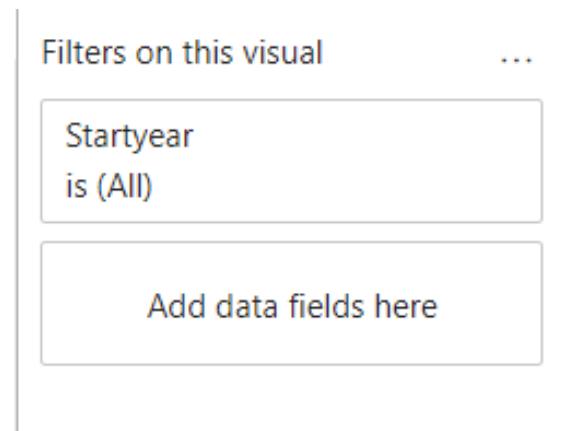
Hình 374. Biểu đồ dạng bản đồ thể hiện số cuộc biểu tình, số người tham gia biểu tình, số ngày biểu tình theo vị trí các quốc gia tại Châu Á

Bước 7: Tại cửa sổ **Visualizations** > Chọn biểu đồ ‘Stacked column chart’



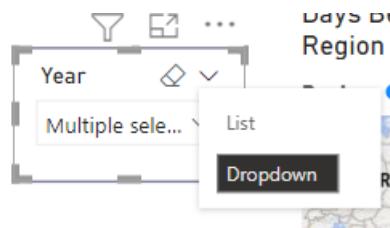
Hình 375. Chọn kiểu báo cáo

Bước 18: Tại cửa sổ Fields > Chọn các thuộc tính cần truy vấn (Startyear)



Hình 376. Chọn thuộc tính cho bộ lọc

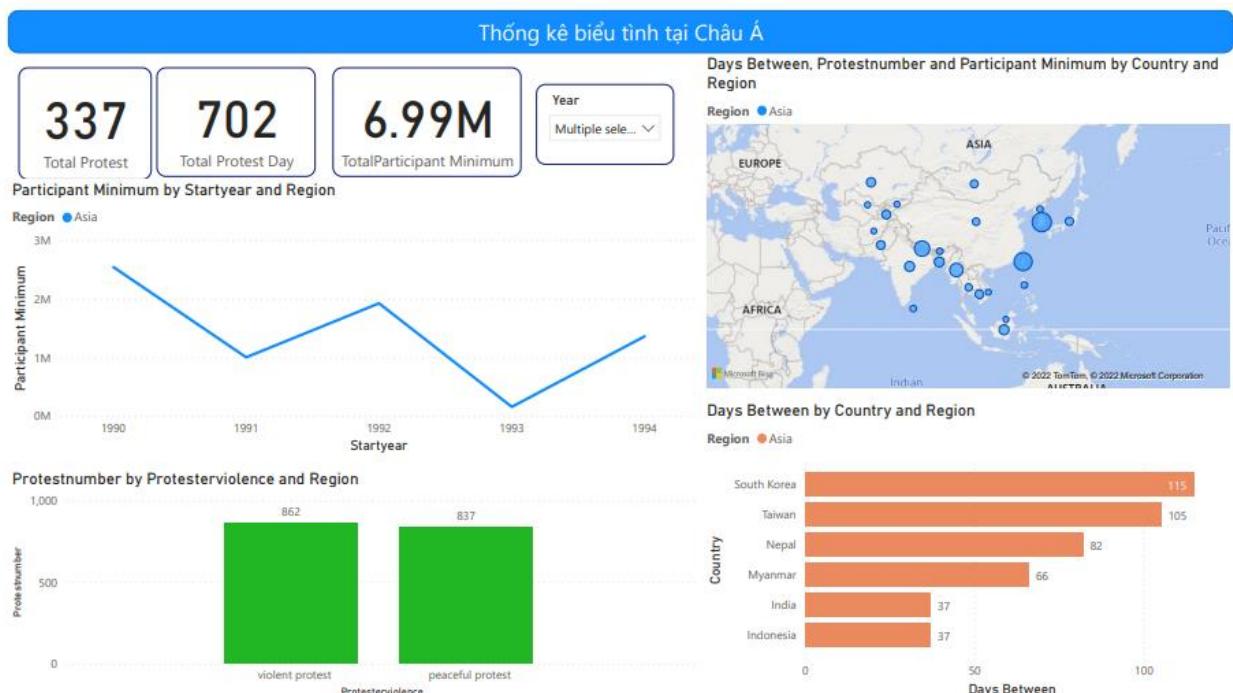
Bước 19: Chọn kiểu cho bộ lọc là Dropdown



Hình 377. Chọn loại cho bộ lọc

- Kết quả hoàn chỉnh:

* (biểu đồ lấy ví dụ 5 năm 2016 -2020)



Hình 378. Thống kê chi tiết biểu tình tại Châu Á qua các năm

CHƯƠNG 5: QUY TRÌNH KHAI THÁC DỮ LIỆU

1. TỔNG QUAN DỮ LIỆU

1.1 Giới thiệu dữ liệu

Tên Dataset: Các chỉ số cá nhân của bệnh tim (Personal Key Indicators of Heart Disease)

Theo như trang [CDC](#), bệnh về tim là một trong những nguyên nhân dẫn đến cái chết tại Mỹ. Hầu hết những người Mỹ (47%) có ít nhất 1 trong 3 yếu tố gây ra căn bệnh tim như: huyết áp cao, mỡ trong máu, hút thuốc. Các nhân tố khác có thể kể đến như: tình trạng béo phì, tiểu đường, không thường xuyên vận động hoặc uống rượu bia quá nhiều. Việc phát hiện và ngăn chặn những yếu tố gây bệnh tim đó đem lại ảnh hưởng rất lớn trong y tế.

Kho dữ liệu được lấy từ CDC và hầu hết là từ ‘Hệ thống giám sát các yếu tố rủi ro hành vi’ - the Behavioral Risk Factor Surveillance System (BRFSS). Hàng năm, họ tiến hành các khảo sát bằng điện thoại để tập hợp thông tin tình trạng của những người dân ở Mỹ, với hơn 400 000 người thực hiện khảo sát trong mỗi năm và là một trong hệ thống khảo sát sức khỏe lớn nhất thế giới.

Kho dữ liệu được cập nhật gần nhất vào ngày 15/02/2022, bao gồm 401,958 dòng và 18 cột dữ liệu.

Link dataset: <https://shorturl.ae/OeJZV>

KAMIL PYTLAK · UPDATED 3 MONTHS AGO

474

New Notebook

Download (3 MB)

Personal Key Indicators of Heart Disease

2020 annual CDC survey data of 400k adults related to their health status

Data Code (96) Discussion (9) Metadata

About Dataset

Key Indicators of Heart Disease

2020 annual CDC survey data of 400k adults related to their health status

What topic does the dataset cover?

According to the [CDC](#), heart disease is one of the leading causes of death for people of most races in the US

Usability 10.00

License CC0: Public Domain

Expected update frequency Annually

Hình 379. Trang web của kho dữ liệu

1.2 Mô tả các thuộc tính

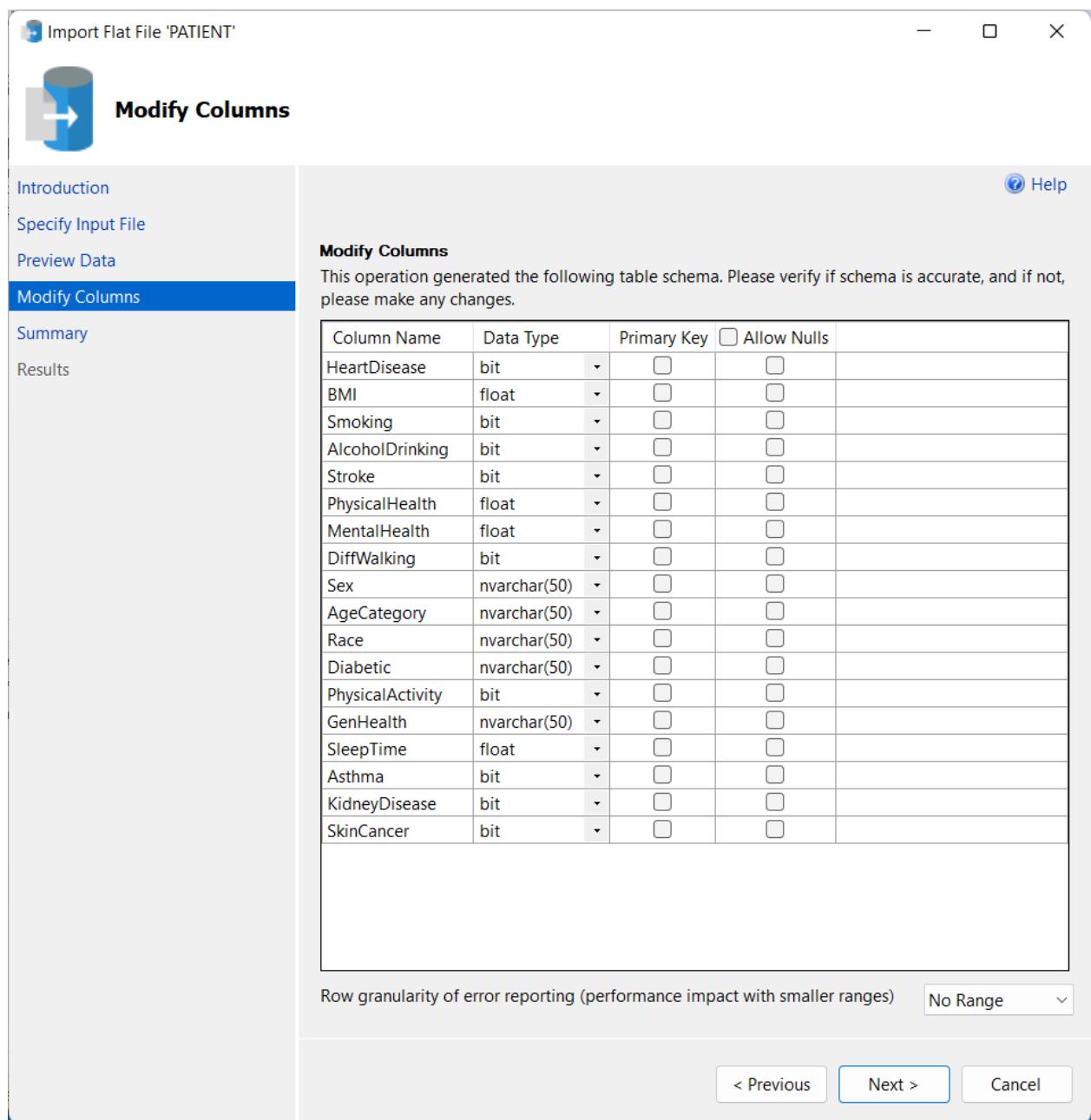
STT	Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa
1	HeartDisease	Boolean	Người thực hiện khảo sát có mang bệnh tim mạch vành hoặc bị nhồi máu cơ tim
2	BMI	Text	Body Mass Index (BMI) <i>Chỉ số khói cơ thể, chỉ số thể trọng, là một công cụ thường được sử dụng để đo lượng mỡ trong cơ thể.</i>
3	Smoking	Text	Có hút ít nhất 100 điếu thuốc? (5 gói = 100 điếu)

4	AlcoholDrinking	Text	Có phải người thích uống rượu bia? (uống hơn 14 lần trong một tuần đối với đàn ông hoặc hơn 7 lần đối với phụ nữ)
5	Stroke	Text	Đã bị đột quỵ?
6	PhysicalHealth	Text	Trong 30 ngày vừa qua, có bao nhiêu ngày bị thương/ bệnh vật lý?
7	MentalHealth	Text	Trong 30 ngày vừa qua, có bao nhiêu ngày tâm lý cảm thấy không tốt?
8	DiffWalking	Text	Có gặp khó khăn khi đi lại hay khi leo cầu thang?
9	Sex	Text	Giới tính người được phỏng vấn
10	AgeCategory	Text	13 nhóm tuổi: 18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-older.
11	Race	Text	Dân tộc
12	Diabetic	Text	Có bị tiểu đường không?
13	PhysicalActivity	Text	Người thực hiện phỏng vấn có vận động, tập thể dục nhiều hơn công việc của họ hay không? (đối với người lớn)

14	GenHealth	Text	Bạn thấy sức khỏe tổng quát của mình như thế nào? (Poor, Fair, Good, Very Good, Excellent)
15	SleepTime	Text	Bạn thường ngủ trong bao lâu?
16	Asthma	Text	Bạn có bị hen suyễn không?
17	KidneyDisease	Text	Ngoại trừ các bệnh như sỏi thận, nhiễm trùng đường tiêu, hay tiêu không tự chủ được thì bạn đã bị bệnh nào về thận chưa?
18	SkinCancer	Text	Bạn có bị ung thư?

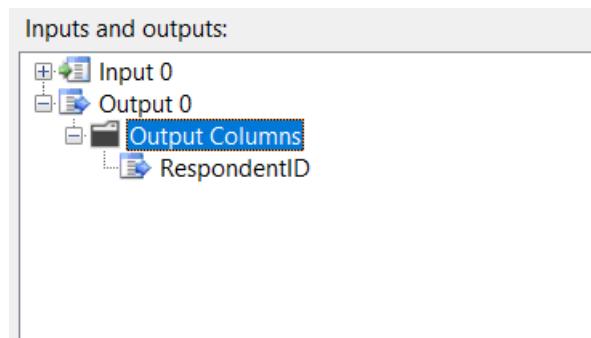
2. QUÁ TRÌNH THỰC HIỆN SSIS, SSAS

Bước 1: Tại SQL Server Studio, thực hiện tạo database và nhập dữ liệu gốc vào bằng ‘Import Flat Files’ như quá trình tích hợp dữ liệu vào kho đã trình bày tại mục Chương 2.

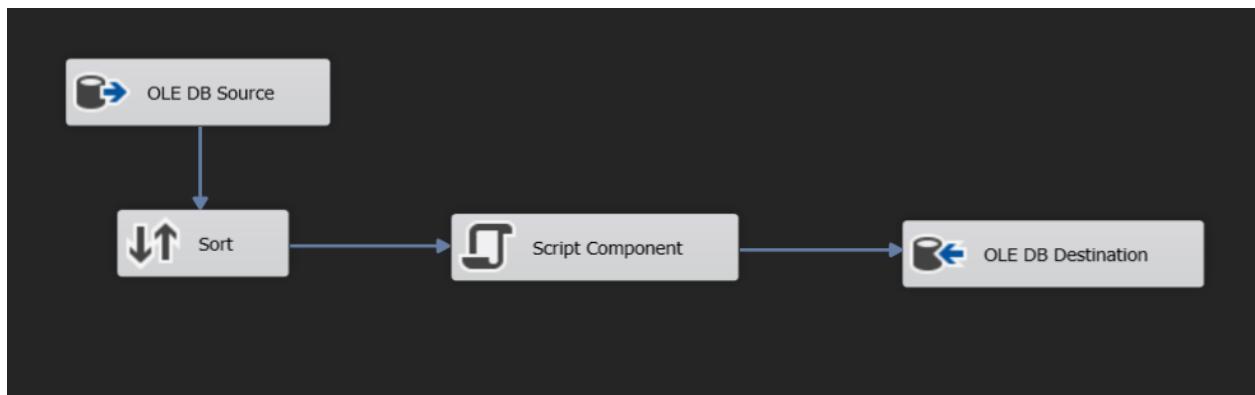


Hình 380. Bước thay đổi kiểu dữ liệu cho các cột thuộc tính

Bước 2: Tại Visual Studio 2019, tạo project SSIS > thực hiện quá trình SSIS để làm sạch dữ liệu và đưa vào kho dữ liệu mới. Tạo khóa chính là ID tăng tự động (RespondentID)

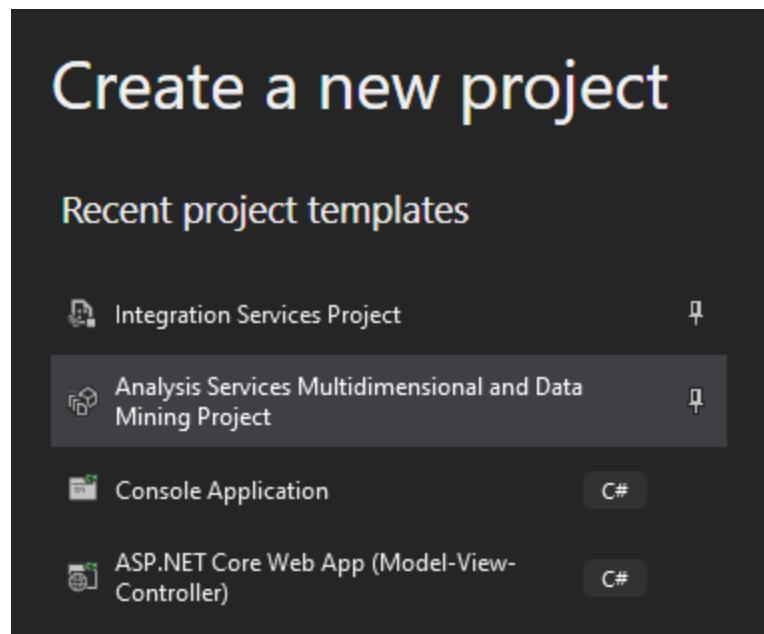


Hình 381. Tạo ID tự động tăng (RespondentID)



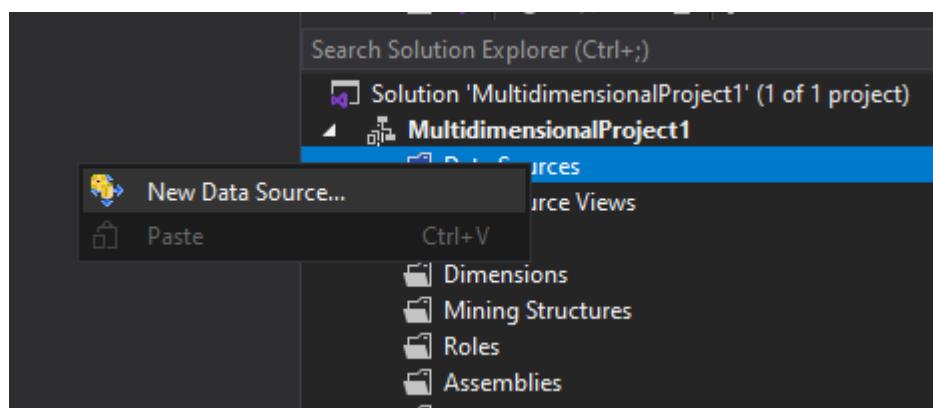
Hình 382. Quá trình làm sạch dữ liệu

Bước 3: Quay lại Visual Studio 2019, tạo mới project SSAS để thực hiện data mining > **Create a new project** > Chọn ‘**Analysis Service Multidimensional and Data Mining Project**’ > Đặt tên và lưu đường dẫn lưu thư mục làm việc



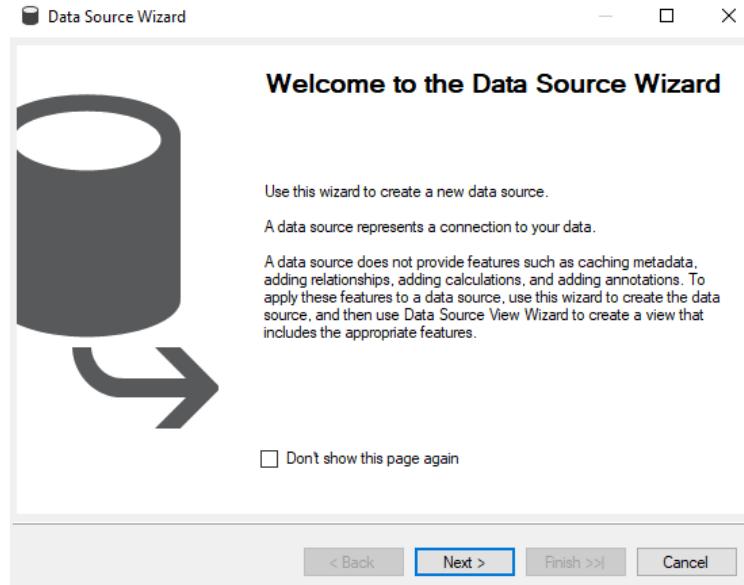
Hình 383. Thực hiện tạo project SSAS mới

Bước 4: Tại thư mục *Data Sources*, chọn *New Data Sources*

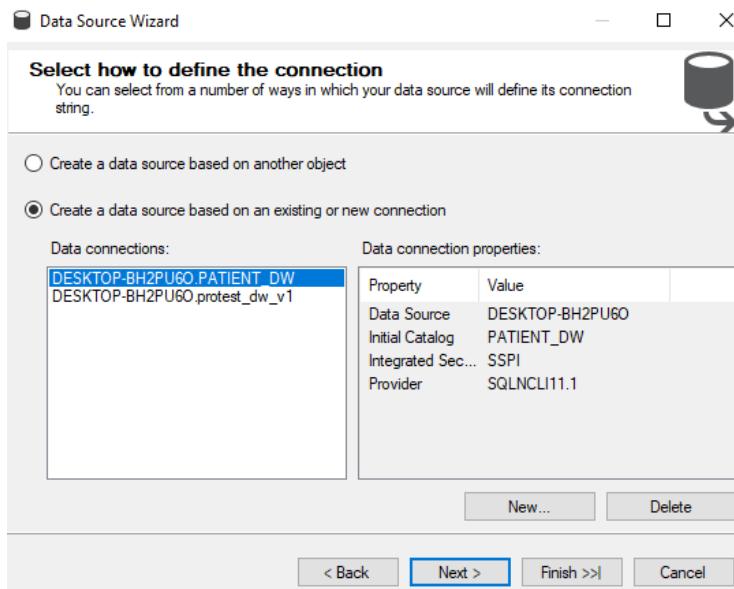


Hình 384. Thực hiện nhập dữ liệu nguồn

Bước 5: Tại màn hình *Data Source Wizard > Next >* chọn ‘*Create a data source based on an existing or new connection*’ > chọn database đích đã tạo ở bước 1, 2

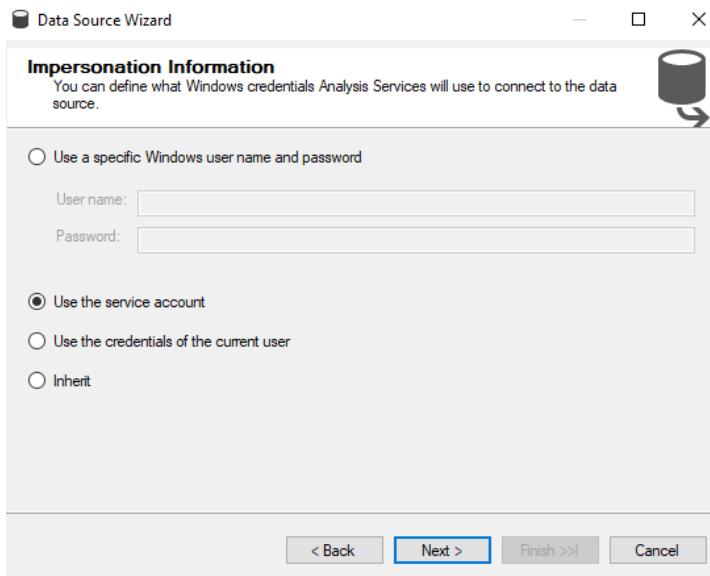


Hình 385. Màn hình Data Source Wizard



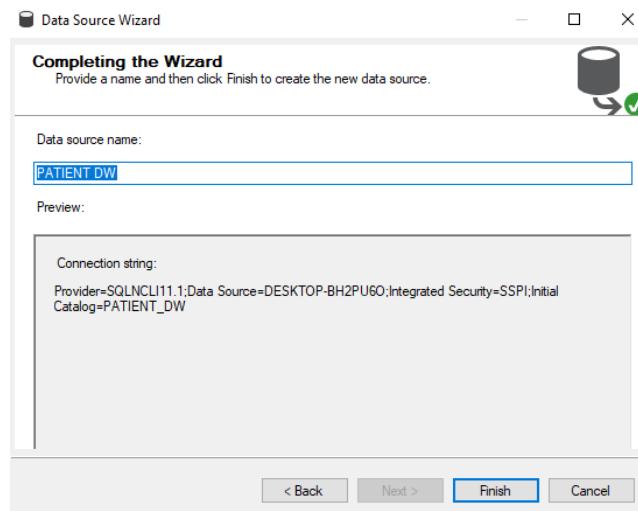
Hình 386. Chọn database đã tạo ở bước 1, 2

Bước 6: Chọn 'Use the service account'



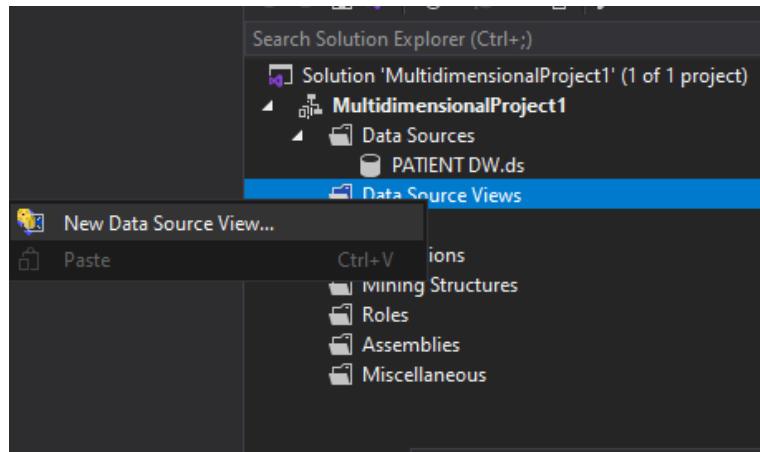
Hình 387. Bước chọn tài khoản kết nối đến dữ liệu nguồn

Bước 6: Đặt tên cho dữ liệu nguồn > Finish



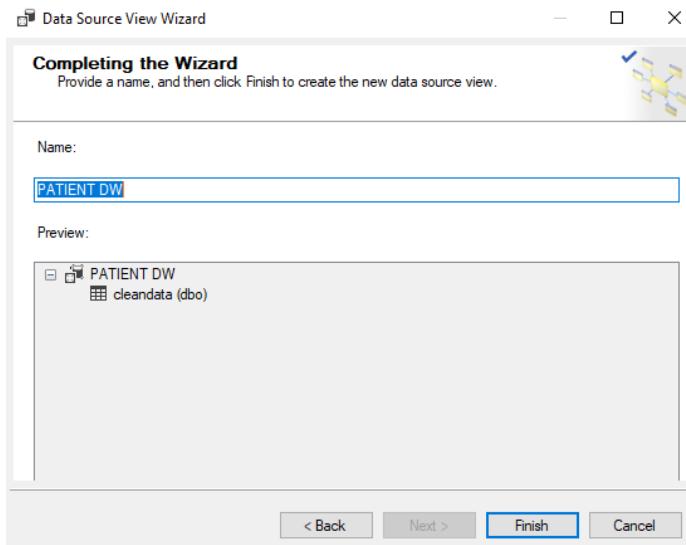
Hình 361. Bước chọn tài khoản kết nối đến dữ liệu nguồn

Bước 7: Tại thư mục Data Source View > thực hiện tạo mới 'New Data Source View'



Hình 388. Bước tạo Data Source View

Bước 7: Chọn dữ liệu nguồn vừa tạo và đặt tên cho data source view > chọn Finish > Xong quá trình nhập dữ liệu cho việc khai thác dữ liệu.

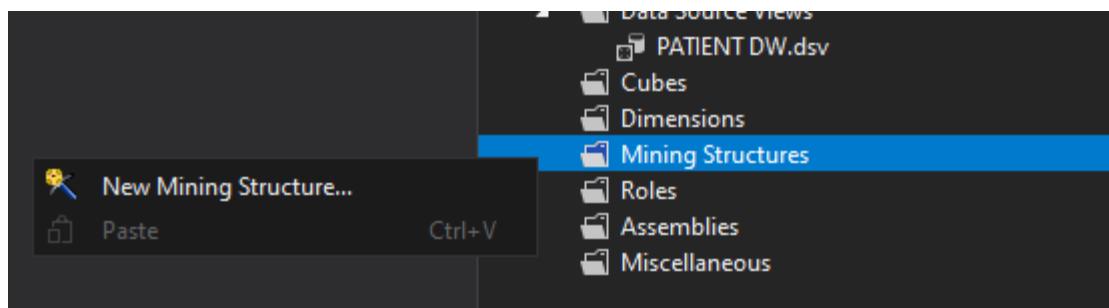


Hình 389. Bước tạo Data Source View

3. QUÁ TRÌNH THỰC HIỆN KHAI THÁC DỮ LIỆU BẰNG THUẬT TOÁN CÂY QUYẾT ĐỊNH

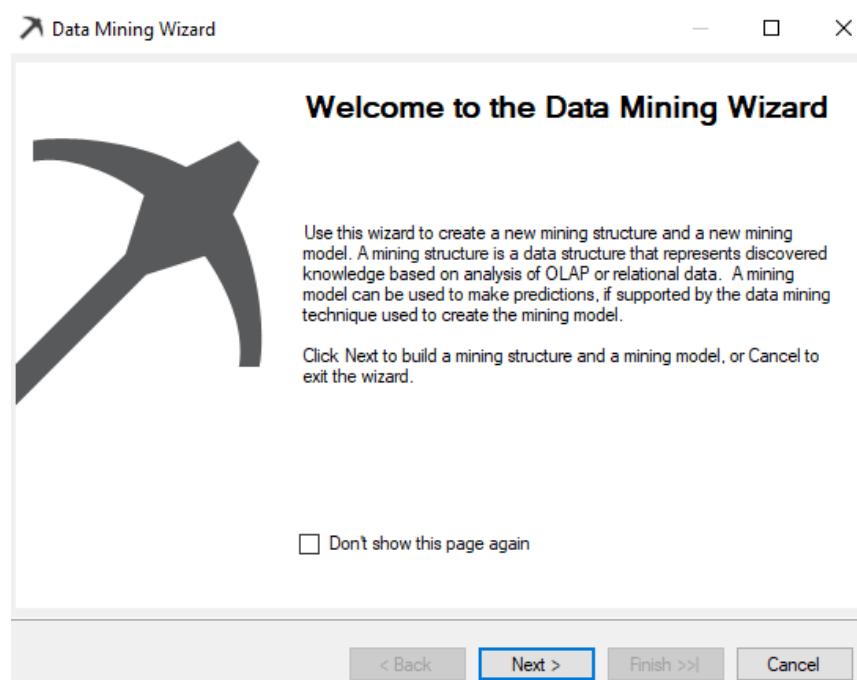
3.1 Tạo và deploy cấu trúc thuật toán cây quyết định

Bước 1: Chọn tạo mới thuật toán khai thác dữ liệu ‘New Mining Structure’



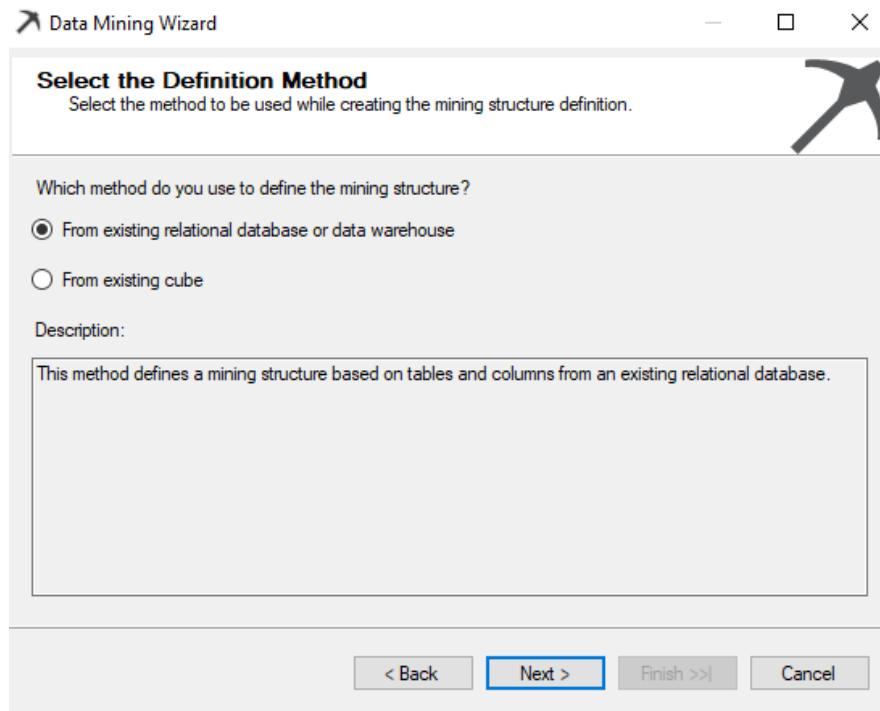
Hình 390. Bước tạo Data Source View

Bước 2: Tại màn hình Data Mining Wizard, chọn Next



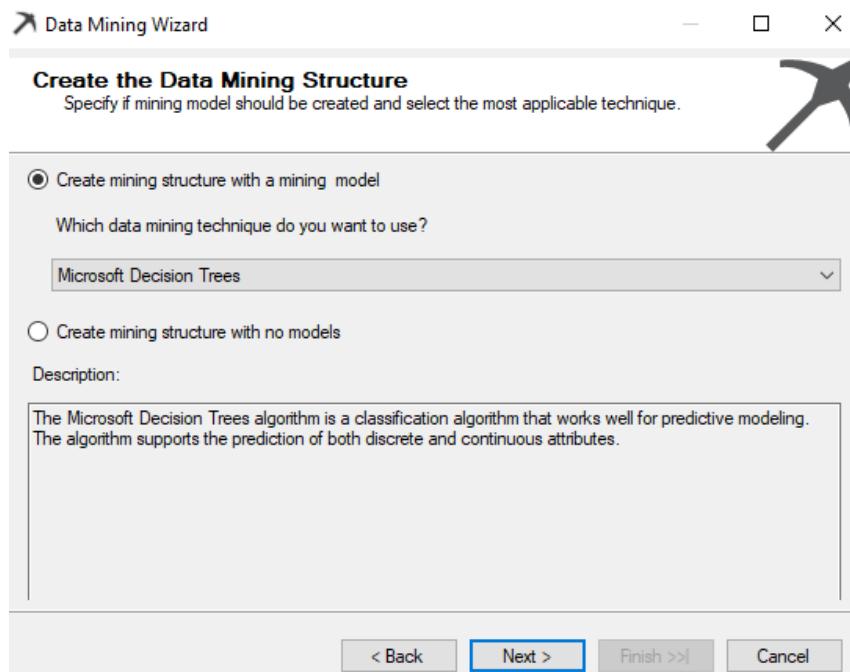
Hình 391. Màn hình Data Mining Wizard

Bước 3: Tại màn hình ‘Select the Definition Method’, chọn ‘*From existing relational database hoặc data warehouse*’



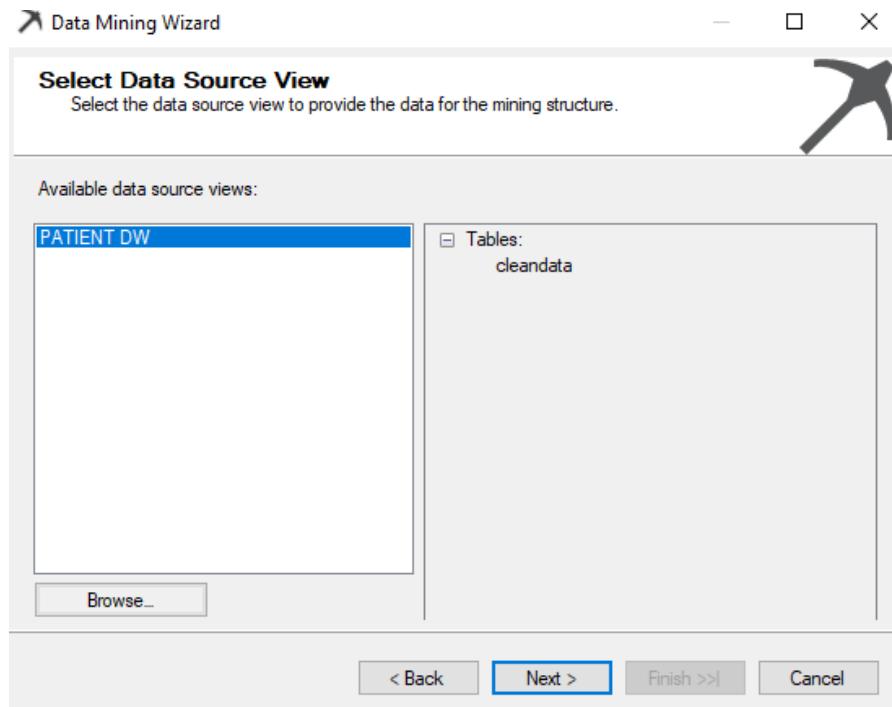
Hình 392. Chọn phương thức định nghĩa cấu trúc khai thác

Bước 4: Chọn thuật toán cây quyết định – Microsoft Decision Trees



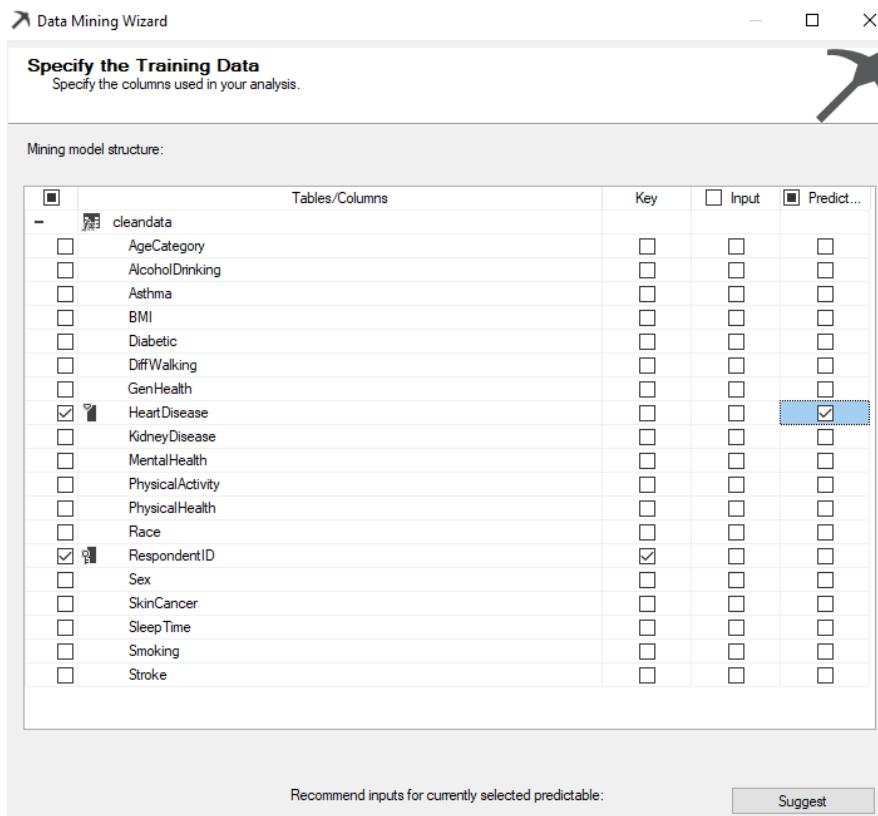
Hình 393. Chọn thuật toán khai thác dữ liệu

Bước 5: Chọn data source view vừa tạo ở các bước trên.



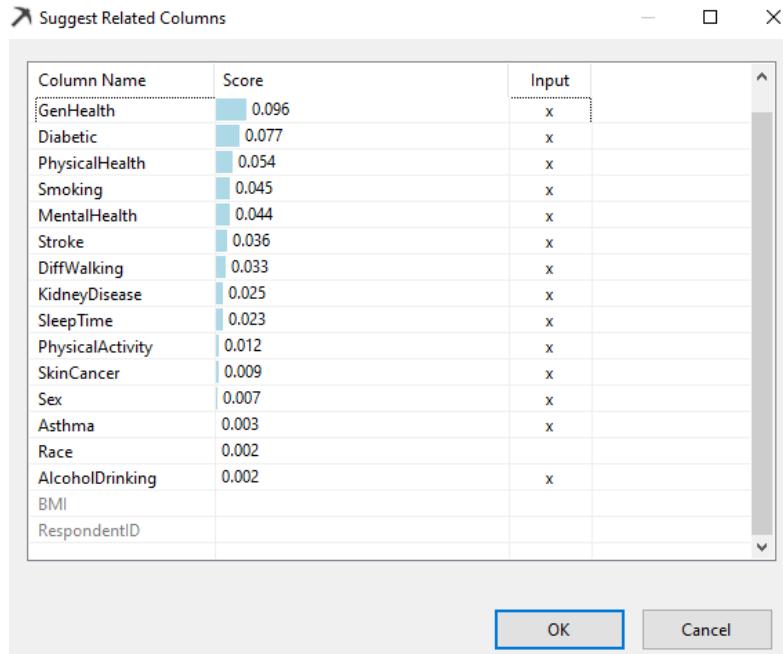
Hình 394. Chọn dữ liệu nguồn

Bước 6: Chọn thuộc tính khóa (RespondentID) và thuộc tính dự đoán bệnh tim (HeartDisease)

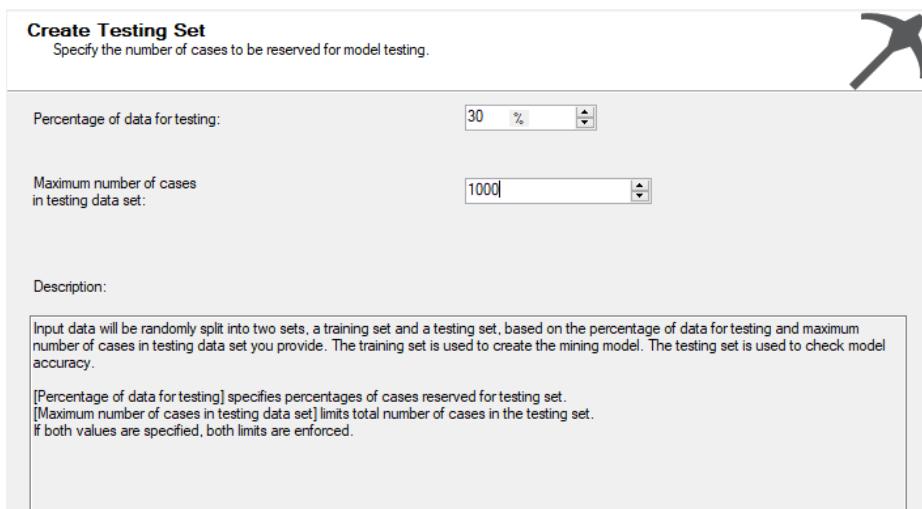


Hình 395. Chọn thuộc tính khóa và thuộc tính dự đoán

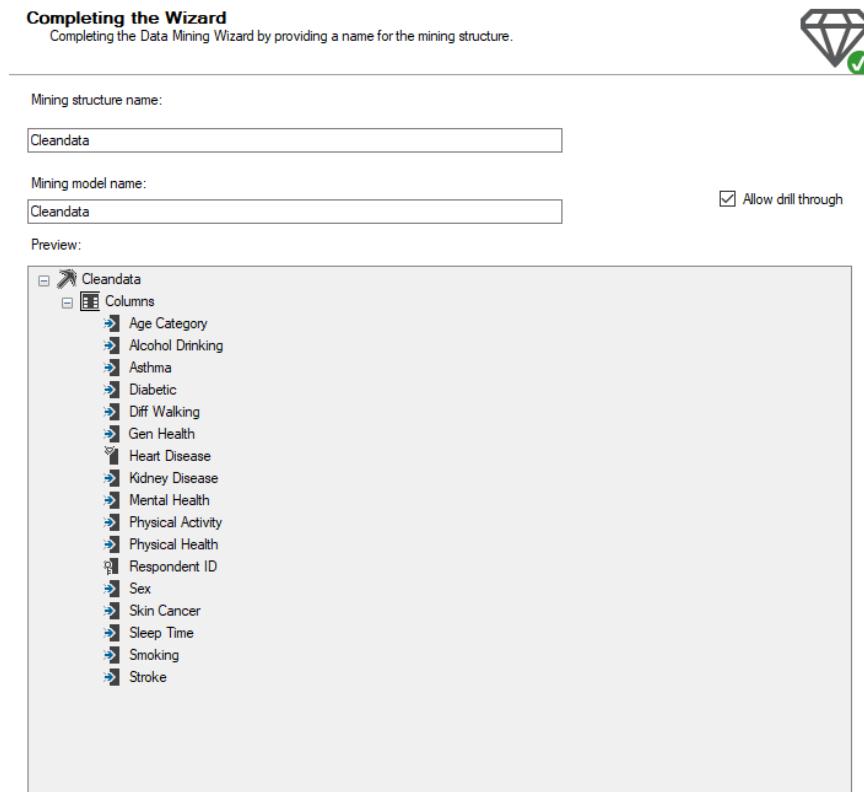
Bước 7: Chọn ‘Suggest’ để hệ thống đưa ra các input gợi ý trên các thuộc tính đã chọn. Chọn các input mình muốn sử dụng.

*Hình 396. Chọn thuật toán khai thác dữ liệu*

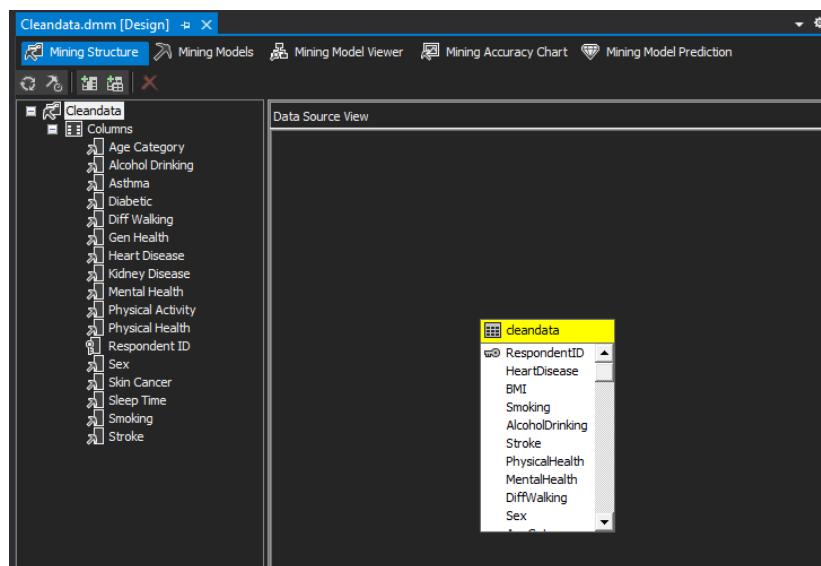
Bước 7: Tạo bộ test với 30% data, và test ít nhất 1000 dòng dữ liệu

*Hình 397. Chọn thuật toán khai thác dữ liệu*

Bước 8: Hoàn thành Wizard bằng việc đặt tên và chọn ‘Allow drill through’

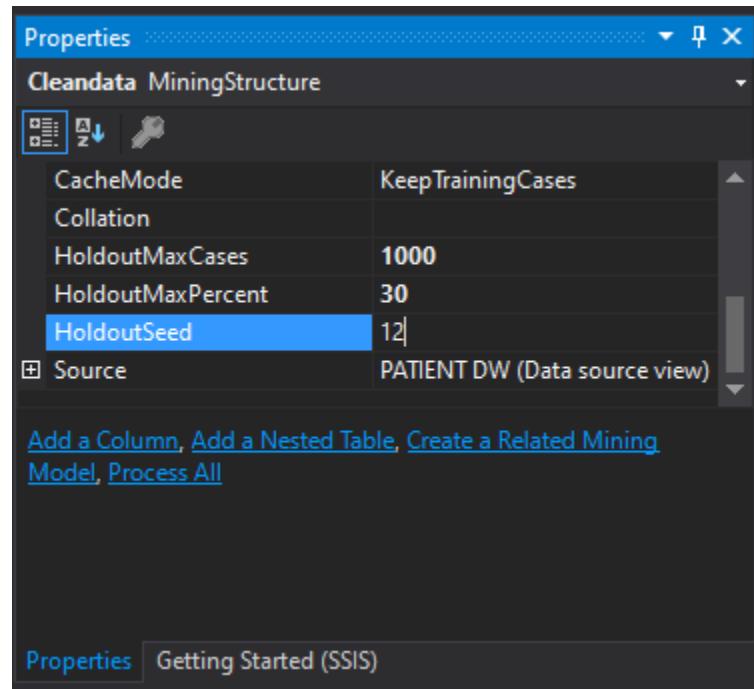


Hình 398. Chọn thuật toán khai thác dữ liệu



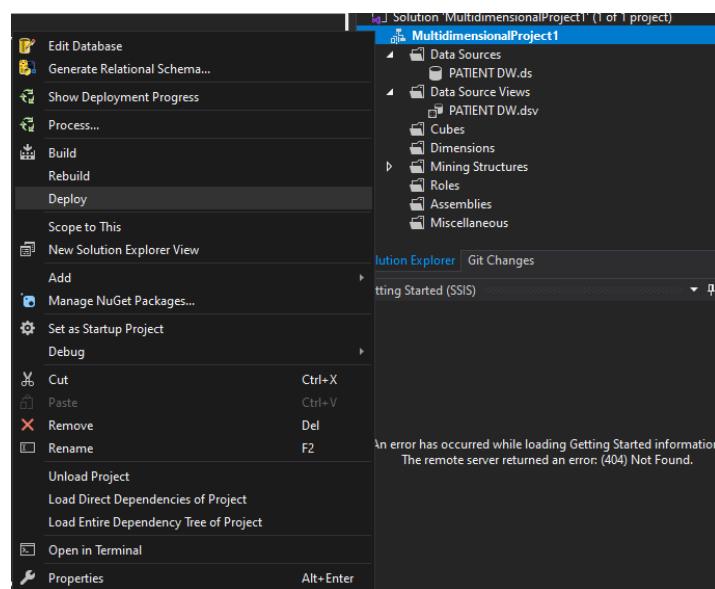
Hình 399. Kết quả sau khi tạo cấu trúc khai thác

Bước 9: Chỉnh sửa *HoldoutSeed* thành 12.

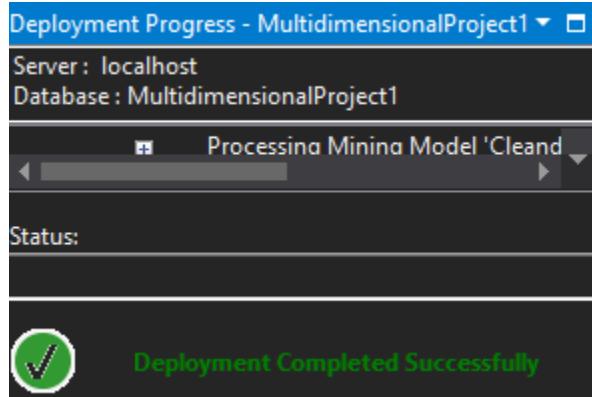


Hình 400. Điều chỉnh properties

Bước 10: Thực hiện chạy project > Deploy

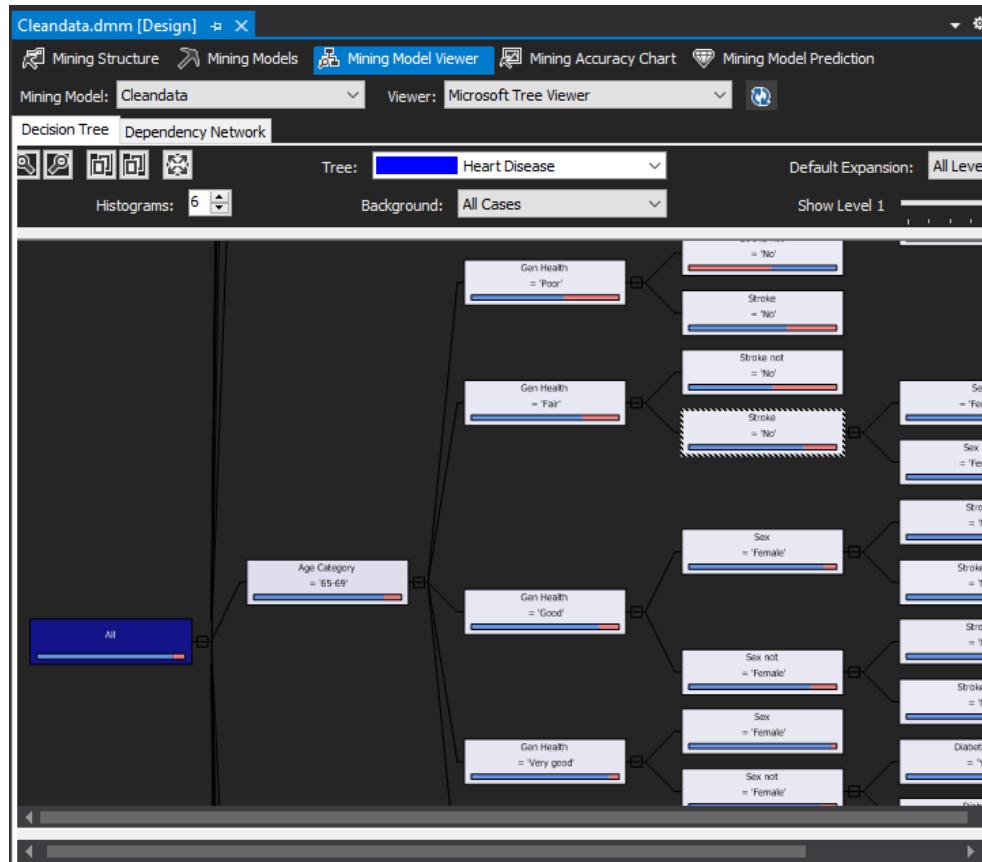


Hình 401. Kết quả sau khi tạo cấu trúc khai thác



Hình 402. Kết quả sau khi deploy project

Bước 11: Thực hiện chạy project > Deploy

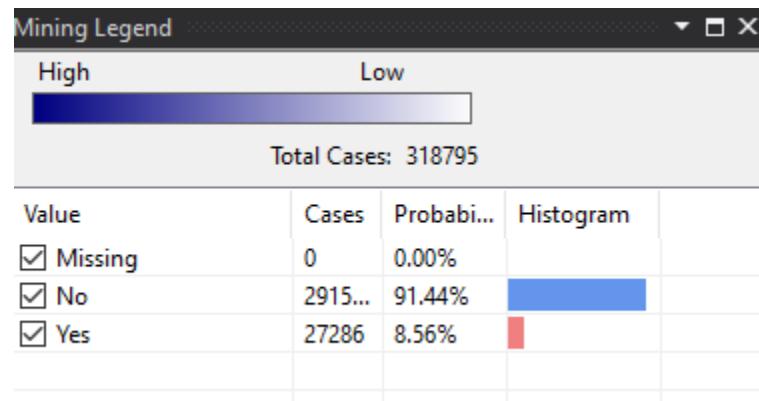


Hình 403. Kết quả sau khi deploy project

3.2 Phân tích và đưa ra tập luật

- Xem **Mining Model Viewer** của thuật toán Microsoft Decision Trees

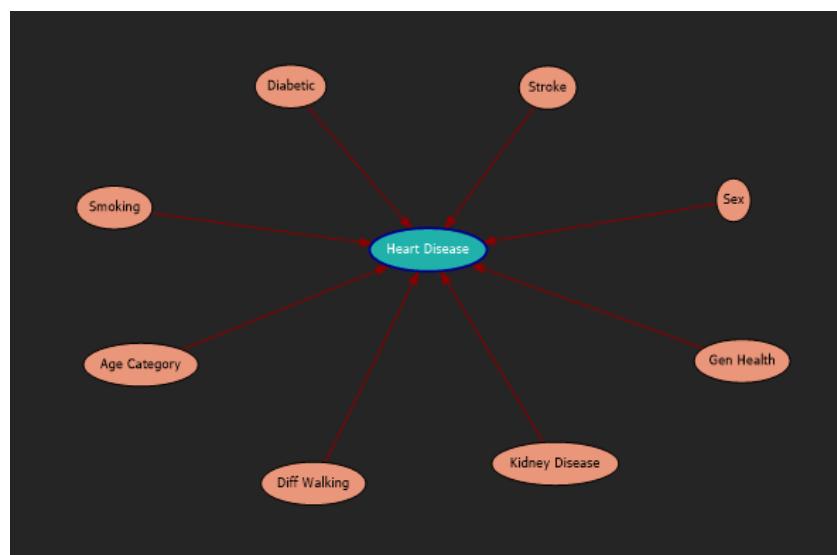
- Với thuộc tính dự đoán là HeartDisease.
 - o Tỷ lệ bị bệnh tim (Yes) là 8.56%
 - o Tỷ lệ không bị bệnh tim (No) là 91.44%



Hình 404. Kết quả của Mining Model Viewer

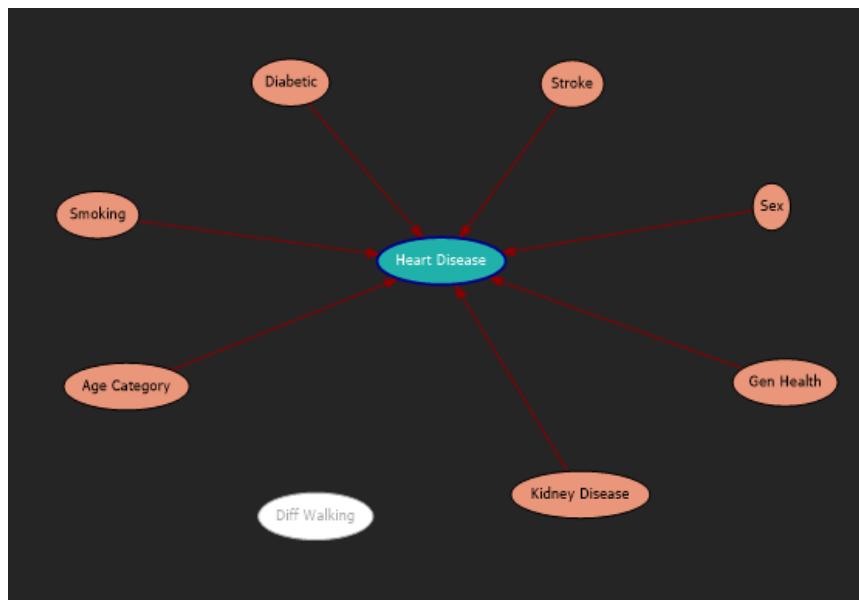
- Tại tab **Dependency Network** của thuật toán Microsoft Decision Tree cho thấy những thuộc tính có ảnh hưởng tới việc bị bệnh tim.

Level 8:



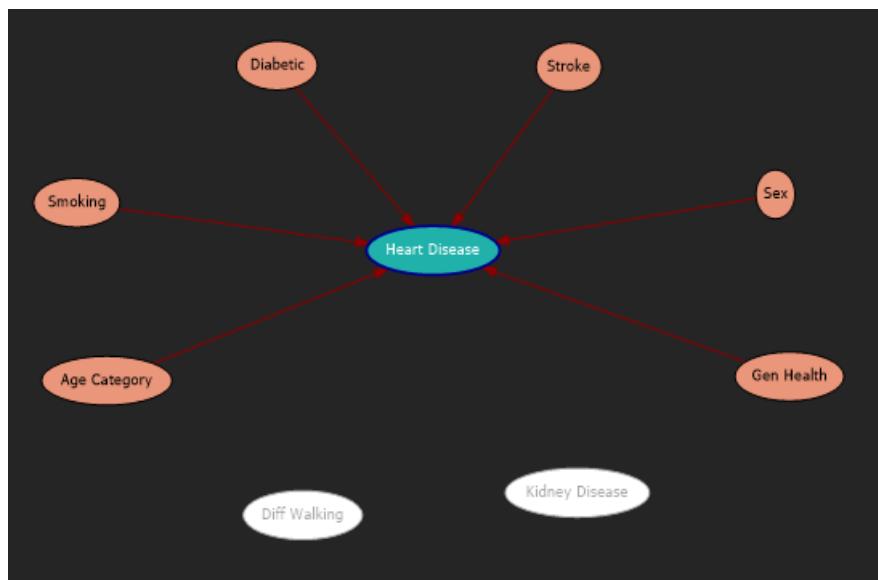
Hình 405. Mức phụ thuộc cấp 8

Level 7:



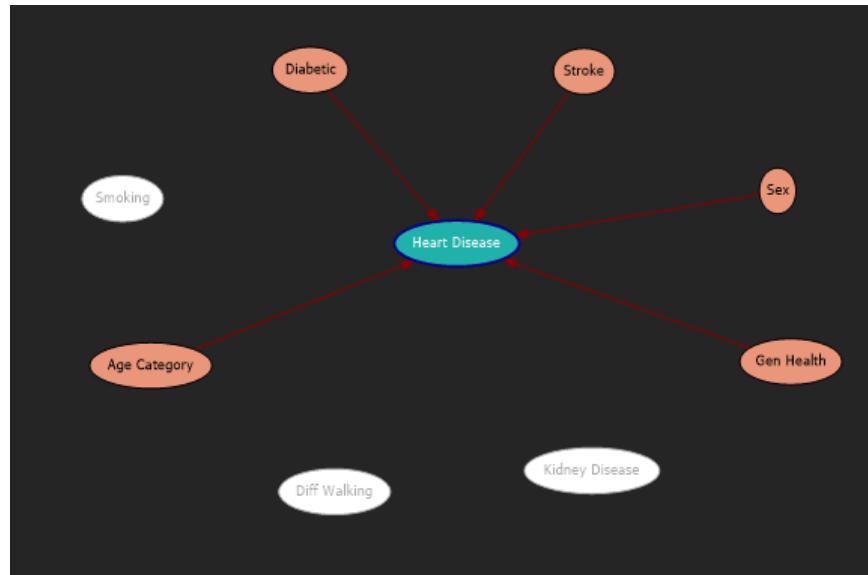
Hình 406. Mức phụ thuộc cấp 7

Level 6:



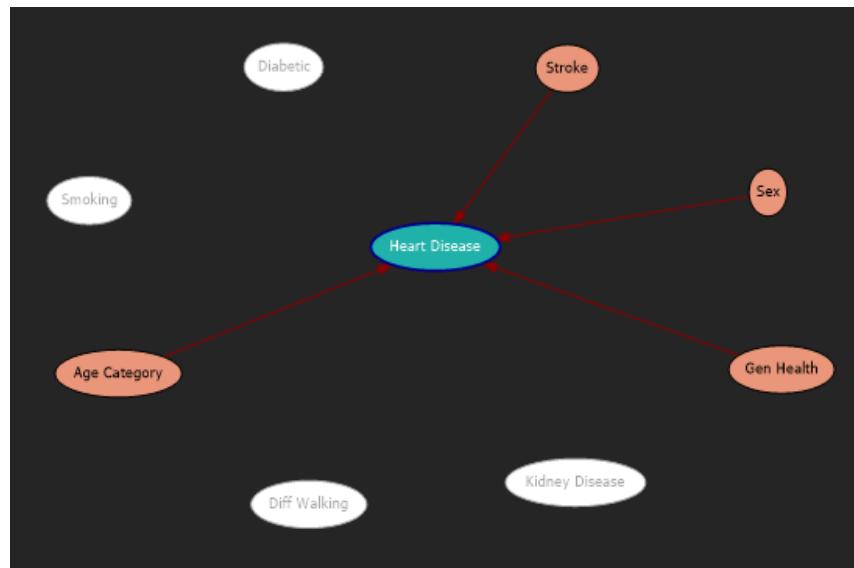
Hình 407. Mức phụ thuộc cấp 6

Level 5:



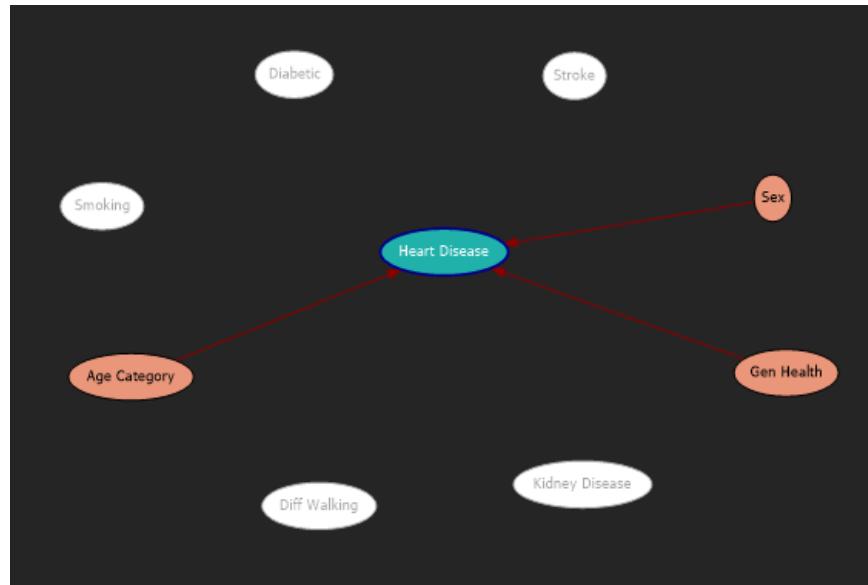
Hình 408. Mức phụ thuộc cấp 5

Level 4:



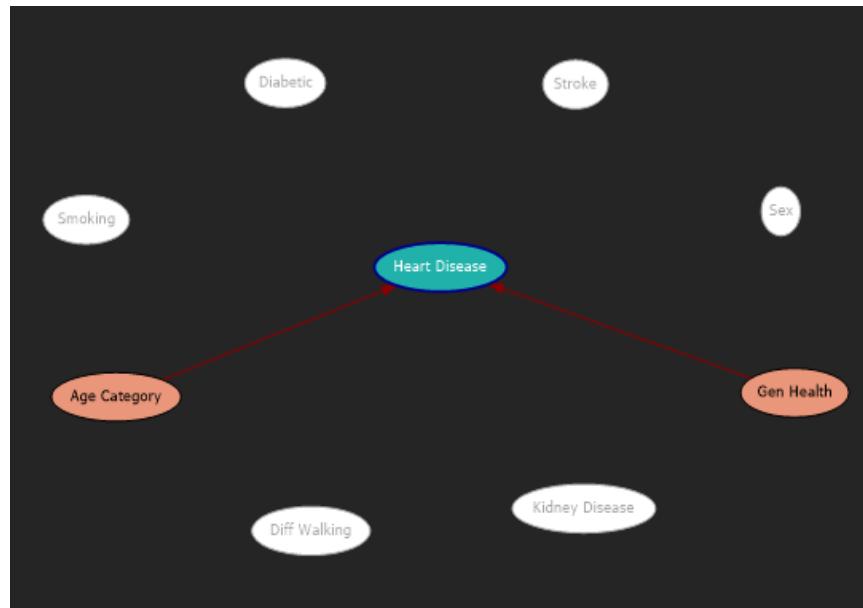
Hình 409. Mức phụ thuộc cấp 4

Level 3:



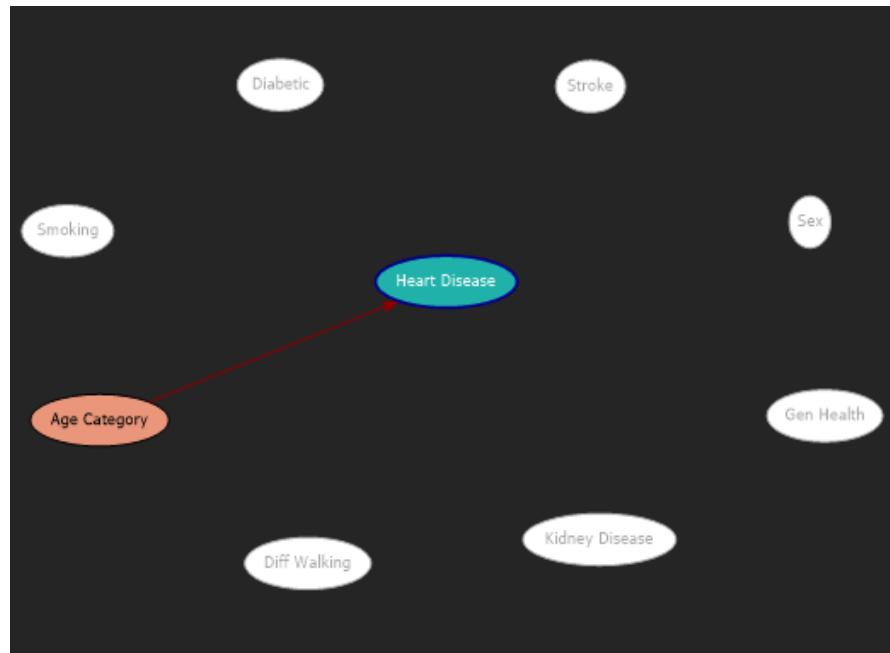
Hình 410. Mức phụ thuộc cấp 3

Level 2:



Hình 411. Mức phụ thuộc cấp 2

Level 1:

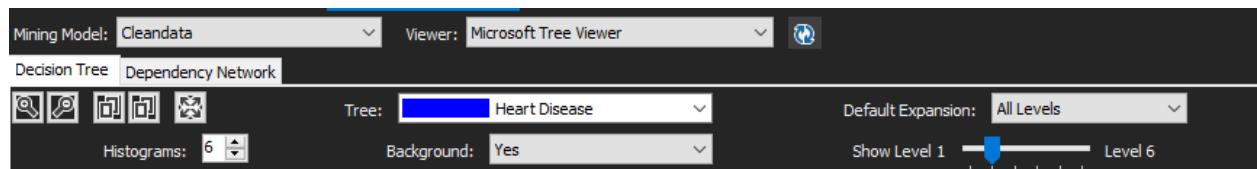


Hình 412. Mức phụ thuộc cấp 1

a, Dự đoán nhóm người sẽ mắc bệnh tim

Chọn điều kiện lọc với Background là ‘Yes’ (Có mắc bệnh tim).

Với mức 2 là Age Category nhằm dự đoán nhóm tuổi người bị mắc bệnh



Hình 413. Dự đoán nhóm tuổi người mắc bệnh

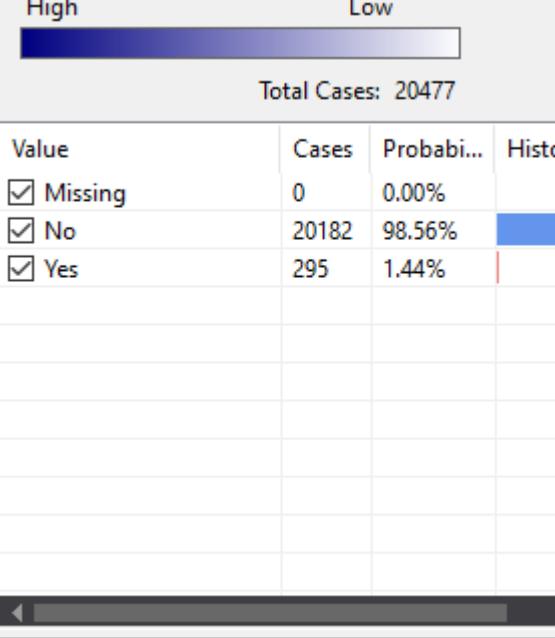
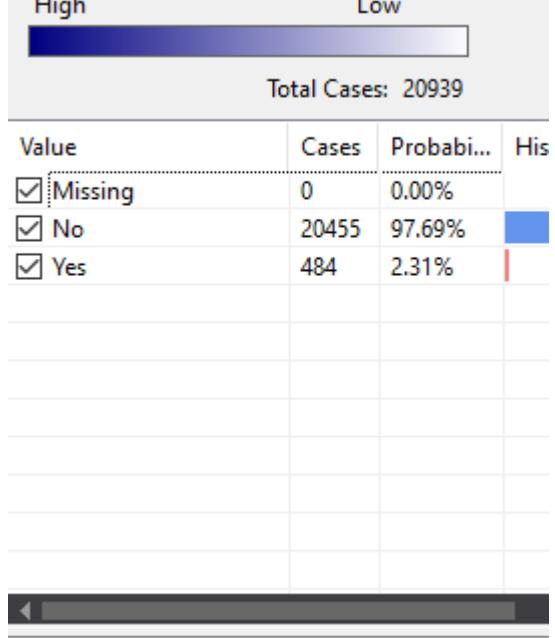


Hình 414. Cây quyết định mức 2

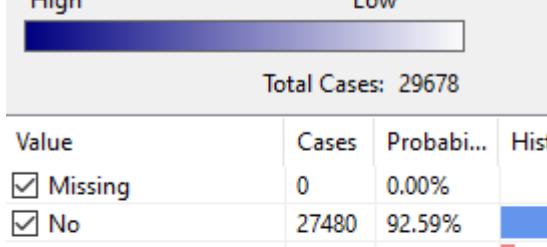
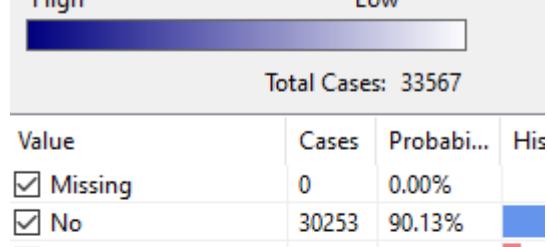
Ta có kết quả dự đoán mắc bệnh tim theo từng nhóm tuổi như sau:

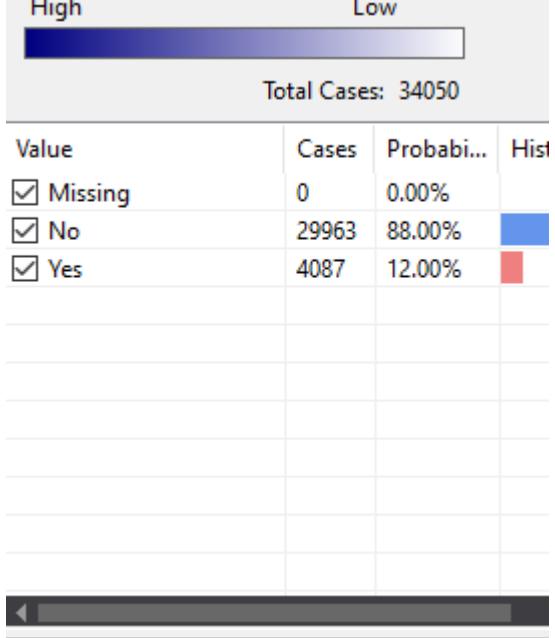
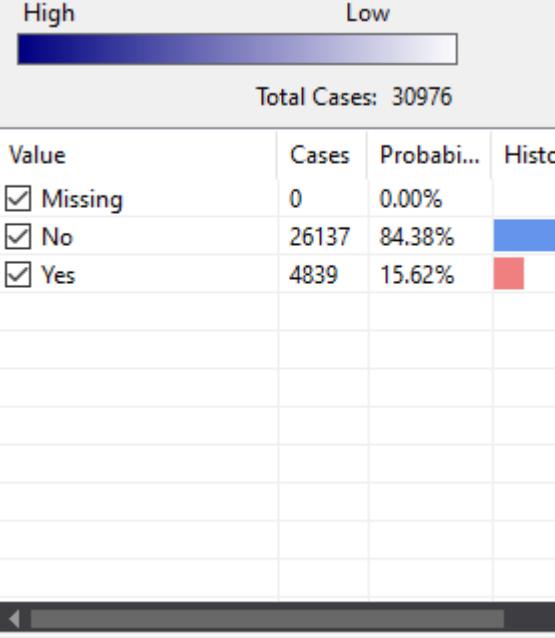
'Age Category'	Hình kết quả Mining	Phân tích																
18-24	<p>High Low Total Cases: 21003</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probability</th> <th>Histogram</th> </tr> </thead> <tbody> <tr> <td>Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td>No</td> <td>20873</td> <td>99.38%</td> <td>Blue bar</td> </tr> <tr> <td>Yes</td> <td>130</td> <td>0.62%</td> <td>Red bar</td> </tr> </tbody> </table> <p>Age Category = '18-24'</p>	Value	Cases	Probability	Histogram	Missing	0	0.00%		No	20873	99.38%	Blue bar	Yes	130	0.62%	Red bar	<p>Cases: 21003</p> <p>Tỉ lệ mắc bệnh tim nhóm từ 18-24 tuổi:</p> <ul style="list-style-type: none"> No: 99.38% Yes: 0.62%
Value	Cases	Probability	Histogram															
Missing	0	0.00%																
No	20873	99.38%	Blue bar															
Yes	130	0.62%	Red bar															

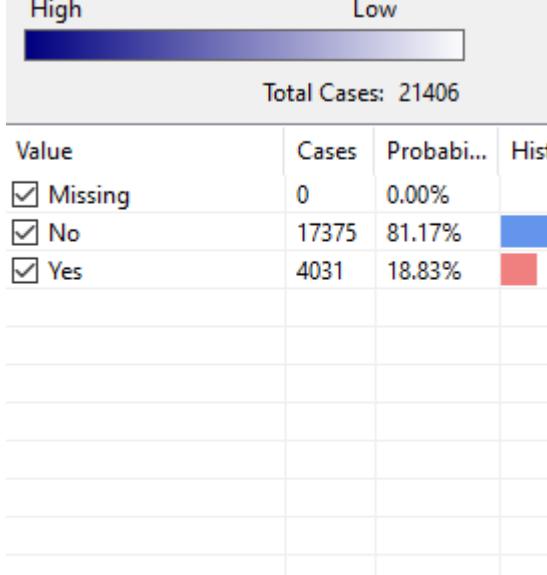
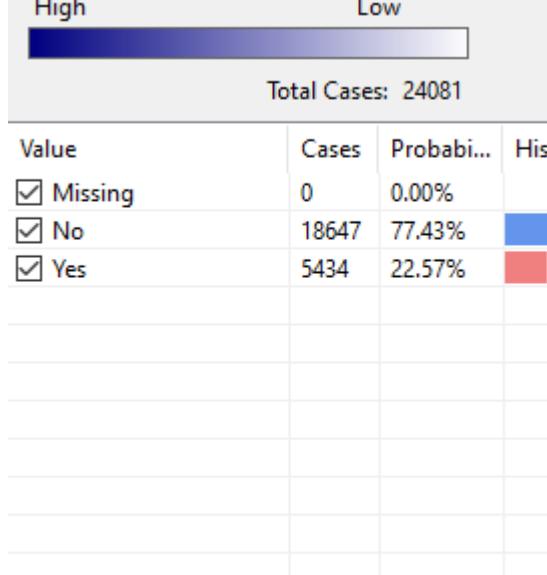
25-29	<p>High Low</p> <p>Total Cases: 16903</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probabi...</th> <th>Histo...</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>16770</td> <td>99.21%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>133</td> <td>0.79%</td> <td></td> </tr> </tbody> </table> <p>Age Category = '25-29'</p>	Value	Cases	Probabi...	Histo...	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	16770	99.21%		<input checked="" type="checkbox"/> Yes	133	0.79%		<p>Cases: 16903</p> <p>Tỉ lệ mắc bệnh tim nhóm từ 25-29 tuổi:</p> <p>No: 99.21%</p> <p>Yes: 0.79%</p>
Value	Cases	Probabi...	Histo...															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	16770	99.21%																
<input checked="" type="checkbox"/> Yes	133	0.79%																
30-34	<p>High Low</p> <p>Total Cases: 18697</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probabi...</th> <th>Histo...</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>18471</td> <td>98.79%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>226</td> <td>1.21%</td> <td></td> </tr> </tbody> </table> <p>Age Category = '30-34'</p>	Value	Cases	Probabi...	Histo...	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	18471	98.79%		<input checked="" type="checkbox"/> Yes	226	1.21%		<p>Cases: 18697</p> <p>Tỉ lệ mắc bệnh tim nhóm từ 30-34 tuổi:</p> <p>No: 98.79%</p> <p>Yes: 1.21%</p>
Value	Cases	Probabi...	Histo...															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	18471	98.79%																
<input checked="" type="checkbox"/> Yes	226	1.21%																

35-39	 <p>Total Cases: 20477</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probability</th> <th>Histogram</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>20182</td> <td>98.56%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>295</td> <td>1.44%</td> <td></td> </tr> </tbody> </table> <p>Age Category = '35-39'</p>	Value	Cases	Probability	Histogram	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	20182	98.56%		<input checked="" type="checkbox"/> Yes	295	1.44%		Cases: 20477 Tỉ lệ mắc bệnh tim nhóm từ 25-29 tuổi: No: 99.21% Yes: 0.79%
Value	Cases	Probability	Histogram															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	20182	98.56%																
<input checked="" type="checkbox"/> Yes	295	1.44%																
40-44	 <p>Total Cases: 20939</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probability</th> <th>Histogram</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>20455</td> <td>97.69%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>484</td> <td>2.31%</td> <td></td> </tr> </tbody> </table> <p>Age Category = '40-44'</p>	Value	Cases	Probability	Histogram	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	20455	97.69%		<input checked="" type="checkbox"/> Yes	484	2.31%		Cases: 20939 Tỉ lệ mắc bệnh tim nhóm từ 40-44 tuổi: No: 97.69% Yes: 2.31%
Value	Cases	Probability	Histogram															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	20455	97.69%																
<input checked="" type="checkbox"/> Yes	484	2.31%																

45-49	 <p>Total Cases: 21722</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probability</th> <th>Hi</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>20983</td> <td>96.60%</td> <td>█</td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>739</td> <td>3.40%</td> <td>█</td> </tr> </tbody> </table> <p>Age Category = '45-49'</p>	Value	Cases	Probability	Hi	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	20983	96.60%	█	<input checked="" type="checkbox"/> Yes	739	3.40%	█	Cases: 21722 Tỉ lệ mắc bệnh tim nhóm từ 45-49 tuổi: No: 96.60% Yes: 3.40%
Value	Cases	Probability	Hi															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	20983	96.60%	█															
<input checked="" type="checkbox"/> Yes	739	3.40%	█															
50-54	 <p>Total Cases: 25296</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probability</th> <th>Hi</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>23917</td> <td>94.55%</td> <td>█</td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>1379</td> <td>5.45%</td> <td>█</td> </tr> </tbody> </table> <p>Age Category = '50-54'</p>	Value	Cases	Probability	Hi	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	23917	94.55%	█	<input checked="" type="checkbox"/> Yes	1379	5.45%	█	Cases: 25296 Tỉ lệ mắc bệnh tim nhóm từ 50-54 tuổi: No: 94.55% Yes: 5.45%
Value	Cases	Probability	Hi															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	23917	94.55%	█															
<input checked="" type="checkbox"/> Yes	1379	5.45%	█															

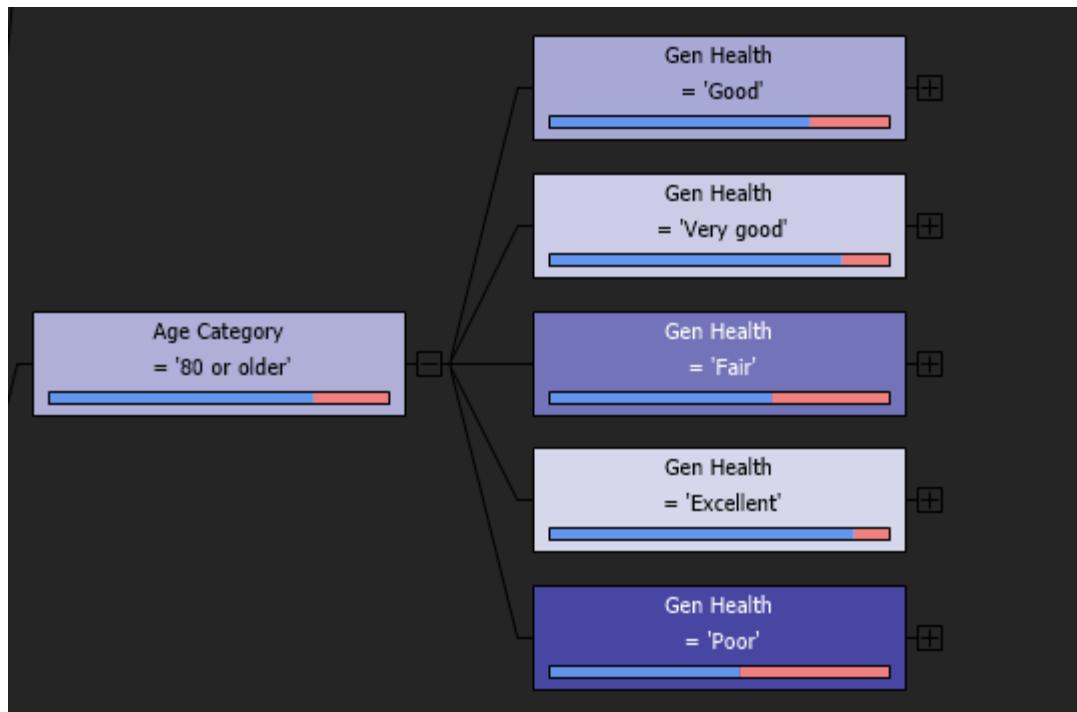
55-59	 <p>Total Cases: 29678</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probabi...</th> <th>Hist</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>27480</td> <td>92.59%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>2198</td> <td>7.41%</td> <td></td> </tr> </tbody> </table> <p>Age Category = '55-59'</p>	Value	Cases	Probabi...	Hist	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	27480	92.59%		<input checked="" type="checkbox"/> Yes	2198	7.41%		Cases: 29678 Tỉ lệ mắc bệnh tim nhóm từ 55-59 tuổi: No: 92.59% Yes: 7.41%
Value	Cases	Probabi...	Hist															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	27480	92.59%																
<input checked="" type="checkbox"/> Yes	2198	7.41%																
60-64	 <p>Total Cases: 33567</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probabi...</th> <th>Hist</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>30253</td> <td>90.13%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>3314</td> <td>9.87%</td> <td></td> </tr> </tbody> </table> <p>Age Category = '60-64'</p>	Value	Cases	Probabi...	Hist	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	30253	90.13%		<input checked="" type="checkbox"/> Yes	3314	9.87%		Cases: 33567 Tỉ lệ mắc bệnh tim nhóm từ 45-49 tuổi: No: 90.13% Yes: 9.87%
Value	Cases	Probabi...	Hist															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	30253	90.13%																
<input checked="" type="checkbox"/> Yes	3314	9.87%																

65-69	 <p>Total Cases: 34050</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probability</th> <th>Histo</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>29963</td> <td>88.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>4087</td> <td>12.00%</td> <td></td> </tr> </tbody> </table> <p>Age Category = '65-69'</p>	Value	Cases	Probability	Histo	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	29963	88.00%		<input checked="" type="checkbox"/> Yes	4087	12.00%		Cases: 34050 Tỉ lệ mắc bệnh tim nhóm từ 65-69 tuổi: No: 88.00% Yes: 12.00%
Value	Cases	Probability	Histo															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	29963	88.00%																
<input checked="" type="checkbox"/> Yes	4087	12.00%																
70-74	 <p>Total Cases: 30976</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probability</th> <th>Histo</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>26137</td> <td>84.38%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>4839</td> <td>15.62%</td> <td></td> </tr> </tbody> </table> <p>Age Category = '70-74'</p>	Value	Cases	Probability	Histo	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	26137	84.38%		<input checked="" type="checkbox"/> Yes	4839	15.62%		Cases: 30976 Tỉ lệ mắc bệnh tim nhóm từ 70-74 tuổi: No: 84.38% Yes: 15.62%
Value	Cases	Probability	Histo															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	26137	84.38%																
<input checked="" type="checkbox"/> Yes	4839	15.62%																

75-79	 <p>Total Cases: 21406</p> <table border="1" data-bbox="453 333 1000 487"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probabi...</th> <th>Hist</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>17375</td> <td>81.17%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>4031</td> <td>18.83%</td> <td></td> </tr> </tbody> </table> <p>Age Category = '75-79'</p>	Value	Cases	Probabi...	Hist	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	17375	81.17%		<input checked="" type="checkbox"/> Yes	4031	18.83%		Cases: 21406 Tỉ lệ mắc bệnh tim nhóm từ 75-79 tuổi: No: 81.17% Yes: 18.83%
Value	Cases	Probabi...	Hist															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	17375	81.17%																
<input checked="" type="checkbox"/> Yes	4031	18.83%																
80 or older	 <p>Total Cases: 24081</p> <table border="1" data-bbox="453 1087 1000 1241"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probabi...</th> <th>Hist</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>18647</td> <td>77.43%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>5434</td> <td>22.57%</td> <td></td> </tr> </tbody> </table> <p>Age Category = '80 or older'</p>	Value	Cases	Probabi...	Hist	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	18647	77.43%		<input checked="" type="checkbox"/> Yes	5434	22.57%		Cases: 24081 Tỉ lệ mắc bệnh tim nhóm từ 80 tuổi trở lên: No: 77.43% Yes: 22.57%
Value	Cases	Probabi...	Hist															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	18647	77.43%																
<input checked="" type="checkbox"/> Yes	5434	22.57%																

- Ta thấy phần trăm tỉ lệ mắc bệnh tim ở nhóm người già từ 80 tuổi trở đi (Age Category = ‘80 or older’) cao và khả quan nhất nên ta

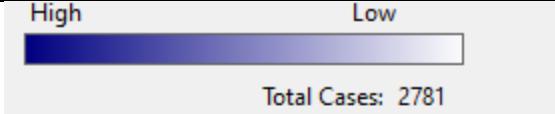
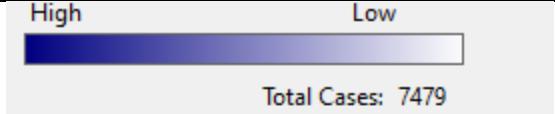
quyết định đi theo nhánh Age Category = ‘80 or older’ và tiếp tục là dự đoán mức tiếp theo.

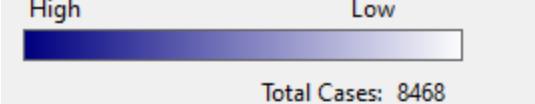
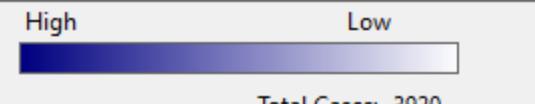


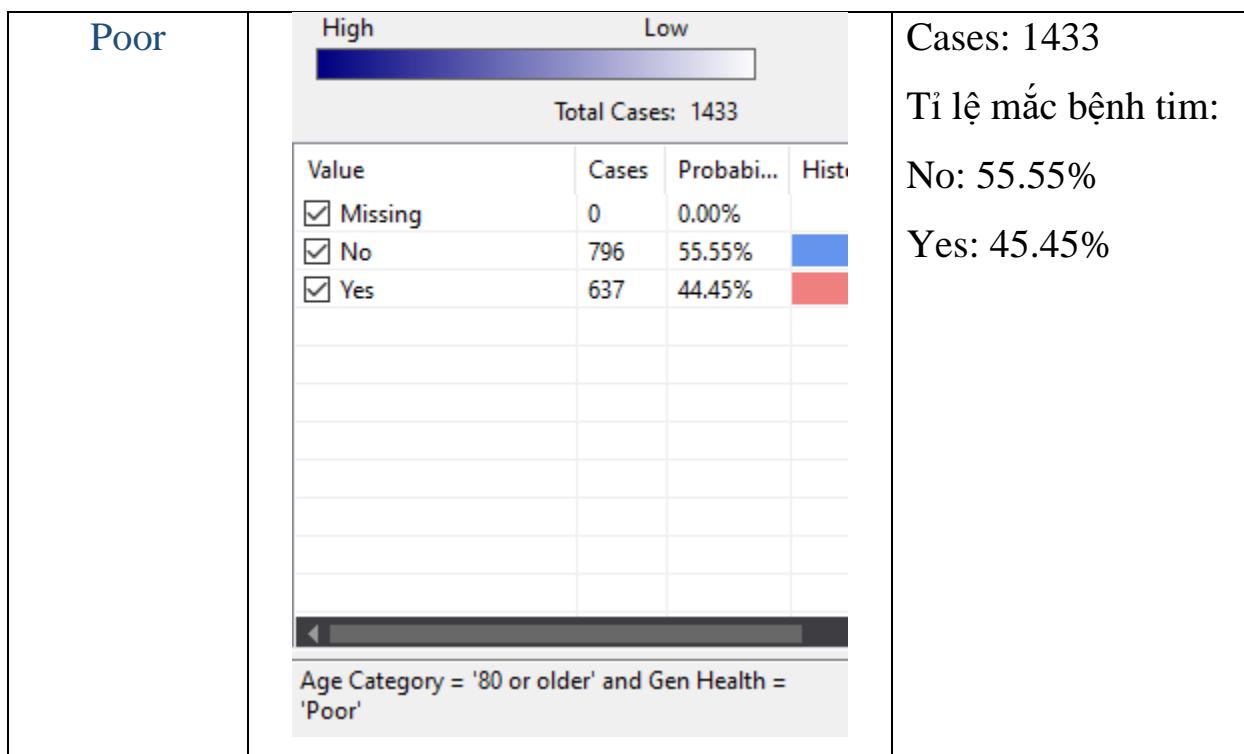
Hình 415. Cây quyết định mức 3

Ta có kết quả dự đoán mắc bệnh tim theo từng nhóm tuổi như sau:

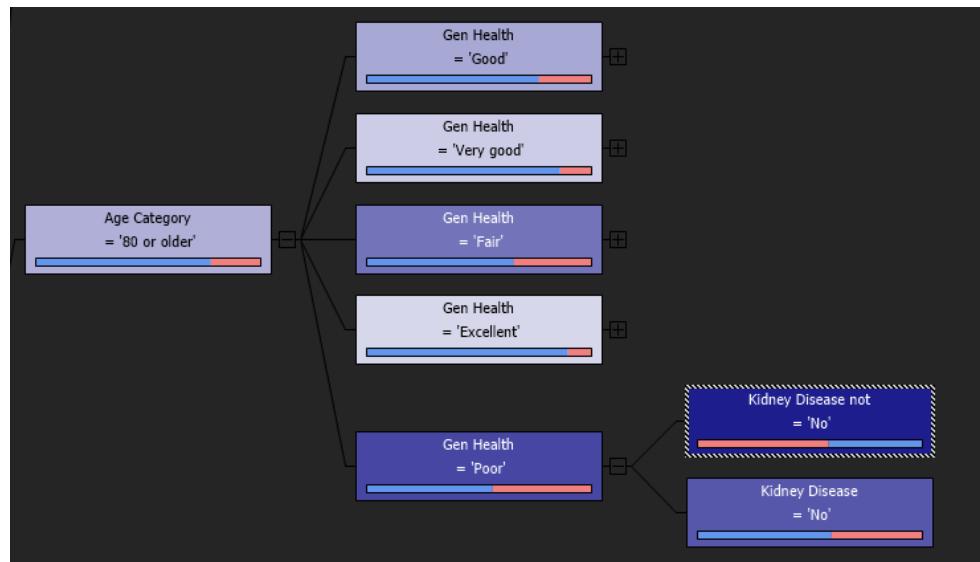
‘Gen Health’	Hình kết quả Mining	Phân tích
--------------	---------------------	-----------

Excellent	 <table border="1" data-bbox="474 316 1029 485"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probabi...</th> <th>Hist</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>2480</td> <td>89.17%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>301</td> <td>10.83%</td> <td></td> </tr> </tbody> </table> <p data-bbox="474 792 1029 908">Age Category = '80 or older' and Gen Health = 'Excellent'</p>	Value	Cases	Probabi...	Hist	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	2480	89.17%		<input checked="" type="checkbox"/> Yes	301	10.83%		Cases: 2781 Tỉ lệ mắc bệnh tim: No: 89.17% Yes: 10.83%
Value	Cases	Probabi...	Hist															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	2480	89.17%																
<input checked="" type="checkbox"/> Yes	301	10.83%																
Very Good	 <table border="1" data-bbox="474 1134 1029 1303"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probabi...</th> <th>Hist</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>6386</td> <td>85.38%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>1093</td> <td>14.62%</td> <td></td> </tr> </tbody> </table> <p data-bbox="474 1615 1029 1731">Age Category = '80 or older' and Gen Health = 'Very good'</p>	Value	Cases	Probabi...	Hist	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	6386	85.38%		<input checked="" type="checkbox"/> Yes	1093	14.62%		Cases: 7479 Tỉ lệ mắc bệnh tim: No: 83.58% Yes: 14.62%
Value	Cases	Probabi...	Hist															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	6386	85.38%																
<input checked="" type="checkbox"/> Yes	1093	14.62%																

Good	 <p>Total Cases: 8468</p> <table border="1"> <thead> <tr> <th>Value</th><th>Cases</th><th>Probabi...</th><th>Hist</th></tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td><td>0</td><td>0.00%</td><td></td></tr> <tr> <td><input checked="" type="checkbox"/> No</td><td>6436</td><td>76.00%</td><td></td></tr> <tr> <td><input checked="" type="checkbox"/> Yes</td><td>2032</td><td>24.00%</td><td></td></tr> </tbody> </table> <p>Age Category = '80 or older' and Gen Health = 'Good'</p>	Value	Cases	Probabi...	Hist	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	6436	76.00%		<input checked="" type="checkbox"/> Yes	2032	24.00%		<p>Cases: 8468</p> <p>Tỉ lệ mắc bệnh tim:</p> <p>No: 76.00%</p> <p>Yes: 24.00%</p>
Value	Cases	Probabi...	Hist															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	6436	76.00%																
<input checked="" type="checkbox"/> Yes	2032	24.00%																
Fair	 <p>Total Cases: 3920</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probabi...</th> <th>Hist</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>2549</td> <td>65.02%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>1371</td> <td>34.98%</td> <td></td> </tr> </tbody> </table> <p>Age Category = '80 or older' and Gen Health = 'Fair'</p>	Value	Cases	Probabi...	Hist	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	2549	65.02%		<input checked="" type="checkbox"/> Yes	1371	34.98%		<p>Cases: 3920</p> <p>Tỉ lệ mắc bệnh tim:</p> <p>No: 65.02%</p> <p>Yes: 34.98%</p>
Value	Cases	Probabi...	Hist															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	2549	65.02%																
<input checked="" type="checkbox"/> Yes	1371	34.98%																

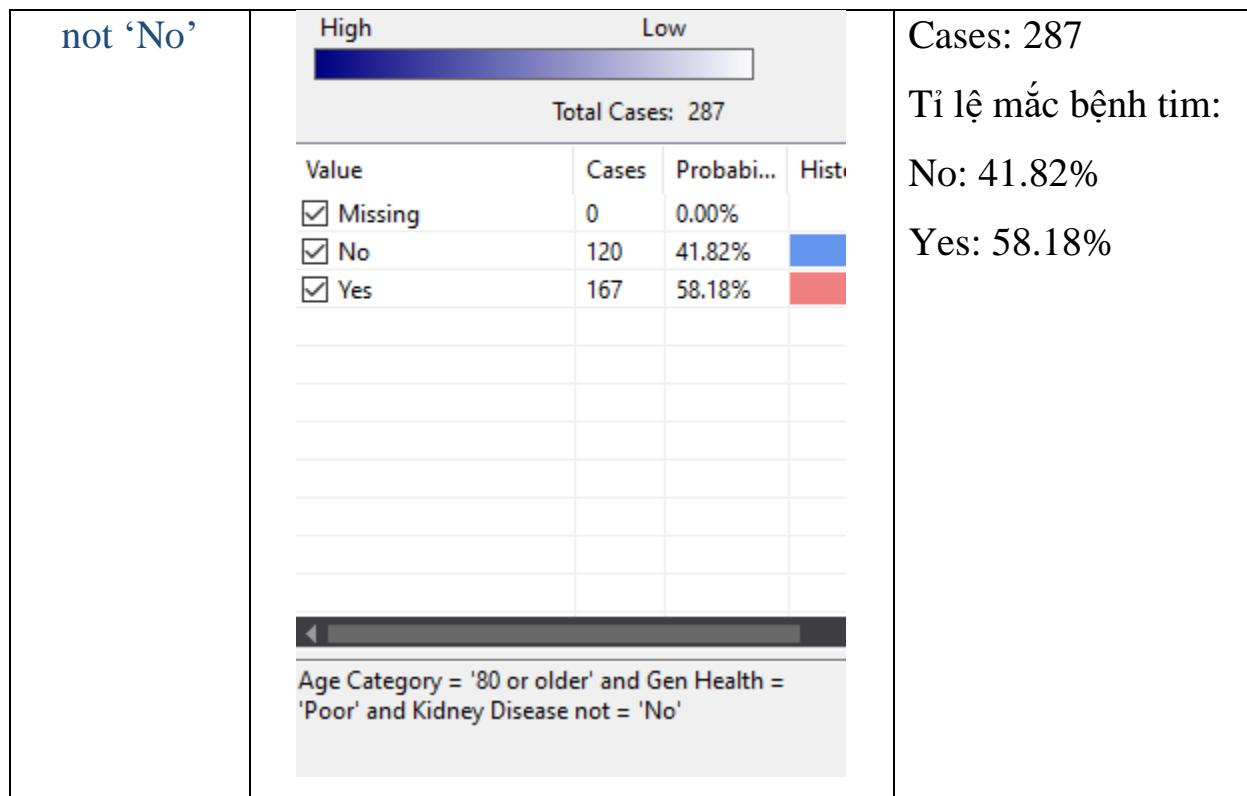


- Ta thấy phần trăm tỉ lệ mắc bệnh tim ở nhóm người già từ 80 tuổi trở đi có sức khỏe kém (Age Category = ‘80 or older’ and Gen Health = ‘Poor’) cao và khả quan nhất nên ta quyết định đi theo nhánh Age Category = ‘80 or older’ và Gen Health = ‘Poor’ và tiếp tục là dự đoán mức tiếp theo.



Hình 416. Cây quyết định mức 4

‘Kidney Disease’	Hình kết quả Mining	Phân tích																
No	<p style="text-align: center;">High Low</p> <p style="text-align: center;">Total Cases: 1146</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probabi...</th> <th>Histo...</th> </tr> </thead> <tbody> <tr> <td><input checked="" type="checkbox"/> Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> No</td> <td>676</td> <td>58.99%</td> <td></td> </tr> <tr> <td><input checked="" type="checkbox"/> Yes</td> <td>470</td> <td>41.01%</td> <td></td> </tr> </tbody> </table> <p style="text-align: center;">Age Category = '80 or older' and Gen Health = 'Poor' and Kidney Disease = 'No'</p>	Value	Cases	Probabi...	Histo...	<input checked="" type="checkbox"/> Missing	0	0.00%		<input checked="" type="checkbox"/> No	676	58.99%		<input checked="" type="checkbox"/> Yes	470	41.01%		<p>Cases: 1146</p> <p>Tỉ lệ mắc bệnh tim:</p> <p>No: 58.99%</p> <p>Yes: 41.01%</p>
Value	Cases	Probabi...	Histo...															
<input checked="" type="checkbox"/> Missing	0	0.00%																
<input checked="" type="checkbox"/> No	676	58.99%																
<input checked="" type="checkbox"/> Yes	470	41.01%																

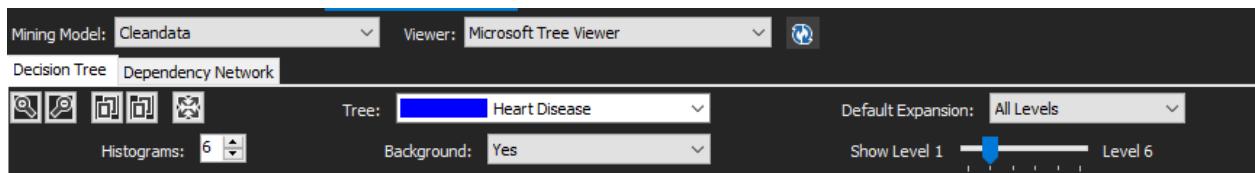


- ⇒ Ta thấy tỉ lệ phần trăm của việc mắc bệnh về thận (Kidney Disease = not 'No') là cao nhất và khả quan nhất.
- ⇒ **Rút ra tập luật:** Những người thực hiện khảo sát tại Mỹ, nếu là người từ 80 tuổi trở lên (Age Category = '80 or older'), sức khỏe yếu (Gen Health = 'Poor') và mắc bệnh về thận (Kidney disease = 'Yes') thì có khả năng mắc bệnh về tim (nhồi máu cơ tim hoặc bệnh về động mạch vành) cao hơn các nhóm tuổi còn lại.

b, Dự đoán nhóm người sẽ không mắc bệnh tim

Chọn điều kiện lọc với Background là 'No' (Không mắc bệnh tim).

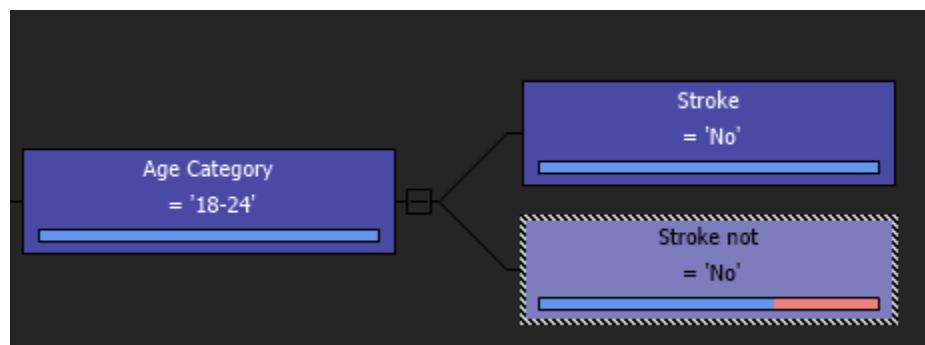
Với mức 2 là Age Category nhằm dự đoán nhóm tuổi người bị mắc bệnh



Hình 417. Dự đoán nhóm tuổi người mắc bệnh

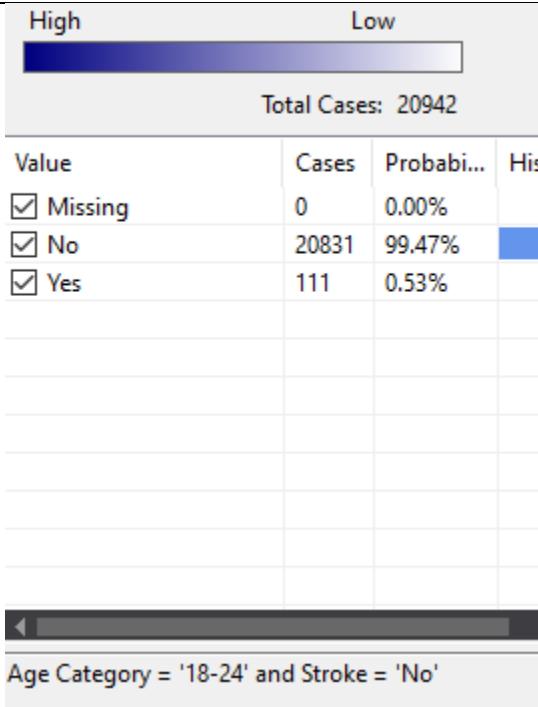
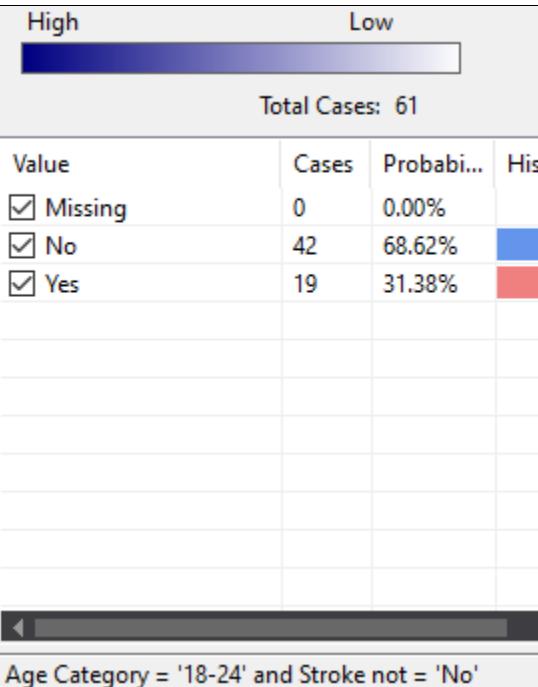
Thực hiện xét tỉ lệ phần trăm không mắc bệnh tương tự như phần a.

Ta thấy nhóm tuổi 18-24% là có tính khả dụng cao nhất, tỉ lệ không mắc bệnh cao nhất. Tiếp tục xét mức 3 về độ tách.



Hình 418. Dự đoán nhóm tuổi người không mắc bệnh mức 3

‘Stroke’	Hình kết quả Mining	Phân tích
----------	---------------------	-----------

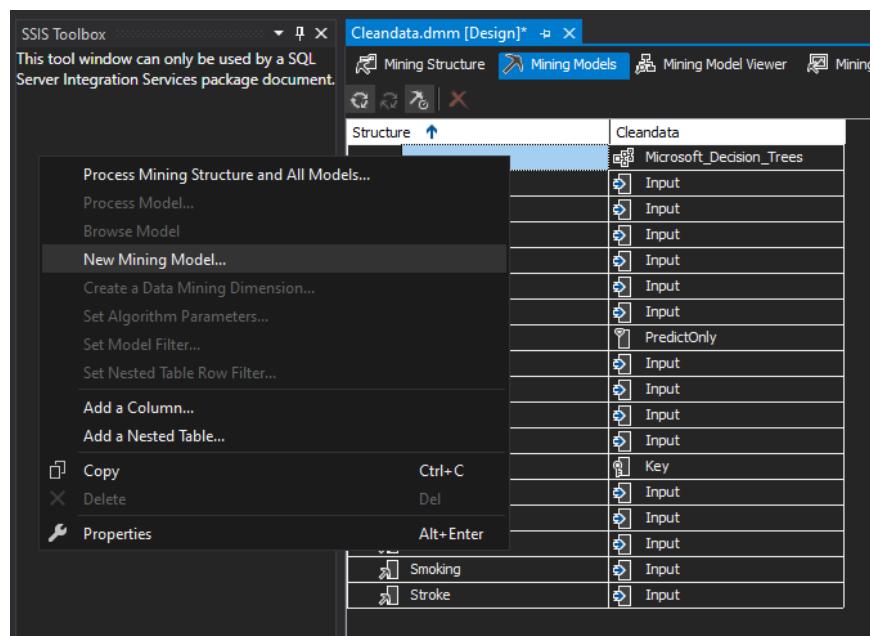
No	 <p>Total Cases: 20942</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probability</th> <th>Histogram</th> </tr> </thead> <tbody> <tr> <td>Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td>No</td> <td>20831</td> <td>99.47%</td> <td></td> </tr> <tr> <td>Yes</td> <td>111</td> <td>0.53%</td> <td></td> </tr> </tbody> </table> <p>Age Category = '18-24' and Stroke = 'No'</p>	Value	Cases	Probability	Histogram	Missing	0	0.00%		No	20831	99.47%		Yes	111	0.53%		<p>Cases: 20942</p> <p>Tỉ lệ mắc bệnh tim:</p> <p>No: 99.47%</p> <p>Yes: 0.53%</p>
Value	Cases	Probability	Histogram															
Missing	0	0.00%																
No	20831	99.47%																
Yes	111	0.53%																
not 'No'	 <p>Total Cases: 61</p> <table border="1"> <thead> <tr> <th>Value</th> <th>Cases</th> <th>Probability</th> <th>Histogram</th> </tr> </thead> <tbody> <tr> <td>Missing</td> <td>0</td> <td>0.00%</td> <td></td> </tr> <tr> <td>No</td> <td>42</td> <td>68.62%</td> <td></td> </tr> <tr> <td>Yes</td> <td>19</td> <td>31.38%</td> <td></td> </tr> </tbody> </table> <p>Age Category = '18-24' and Stroke not = 'No'</p>	Value	Cases	Probability	Histogram	Missing	0	0.00%		No	42	68.62%		Yes	19	31.38%		<p>Cases: 61</p> <p>Tỉ lệ mắc bệnh tim:</p> <p>No: 68.62%</p> <p>Yes: 31.38%</p>
Value	Cases	Probability	Histogram															
Missing	0	0.00%																
No	42	68.62%																
Yes	19	31.38%																

⇒ **Rút ra tập luật:** Những người thực hiện khảo sát tại Mỹ, nếu là người từ 18 -24 (Age Category = ‘18-24’), không bị/ chưa từng bị đột quy thì có khả năng không bị mắc bệnh về tim (nhồi máu cơ tim hoặc bệnh về động mạch vành) hơn các nhóm tuổi còn lại.

4. QUÁ TRÌNH THỰC HIỆN KHAI THÁC DỮ LIỆU BẰNG THUẬT TOÁN CLUSTERING VÀ NAÏVE BAYES

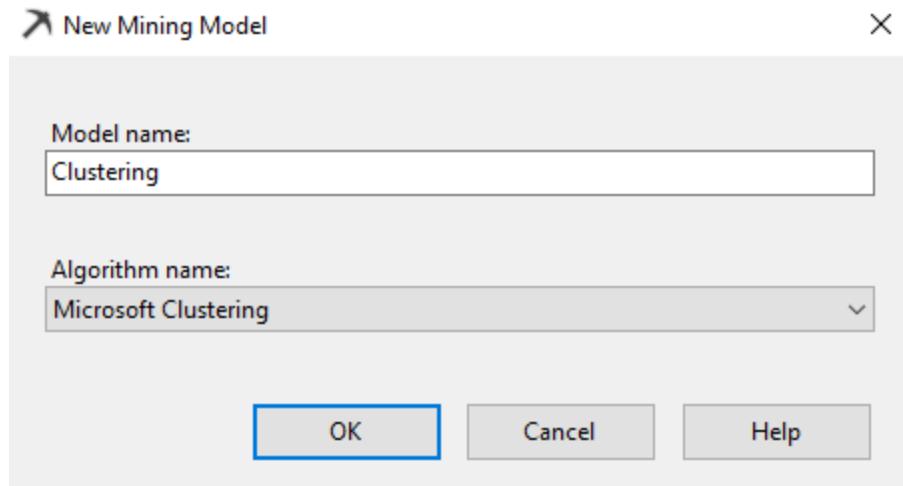
4.1 Tạo và deploy cấu trúc thuật toán Clustering và Naïve Bayes

Bước 1: Thực hiện tạo thuật toán mới (Clustering)



Hình 419. Thực hiện tạo mô hình khai thác

Bước 2: Đặt tên mô hình và chọn thuật toán:

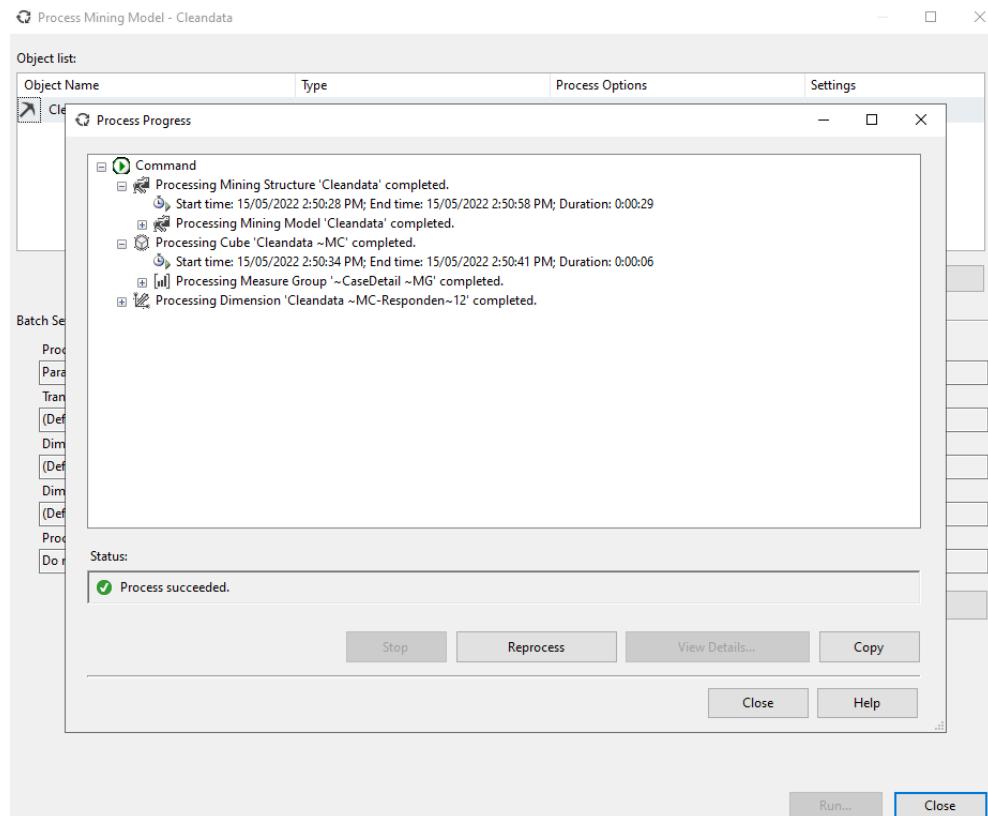


Hình 420. Chọn thuật toán khai thác

A screenshot of the 'Cleandata.dmm [Design]' window in Microsoft SQL Server Data Mining. The title bar shows 'Cleandata.dmm [Design]'. The tabs at the top are 'Mining Structure', 'Mining Models' (which is selected), 'Mining Model Viewer', 'Mining Accuracy Chart', and 'Mining Model Prediction'. Below the tabs is a toolbar with icons for refresh, save, and delete. The main area is a grid table with three columns: 'Structure' (containing 'Cleandata'), 'Cleandata' (containing nodes like 'Microsoft_Decision_Trees', 'Input', 'PredictOnly', etc.), and 'Clustering' (containing nodes like 'Microsoft_Clustering', 'Input', 'Key', etc.). The table lists numerous fields such as Age Category, Alcohol Drinking, Asthma, Diabetic, Diff Walking, Gen Health, Heart Disease, Kidney Disease, Mental Health, Physical Activity, Physical Health, Respondent ID, Sex, Skin Cancer, Sleep Time, Smoking, and Stroke.

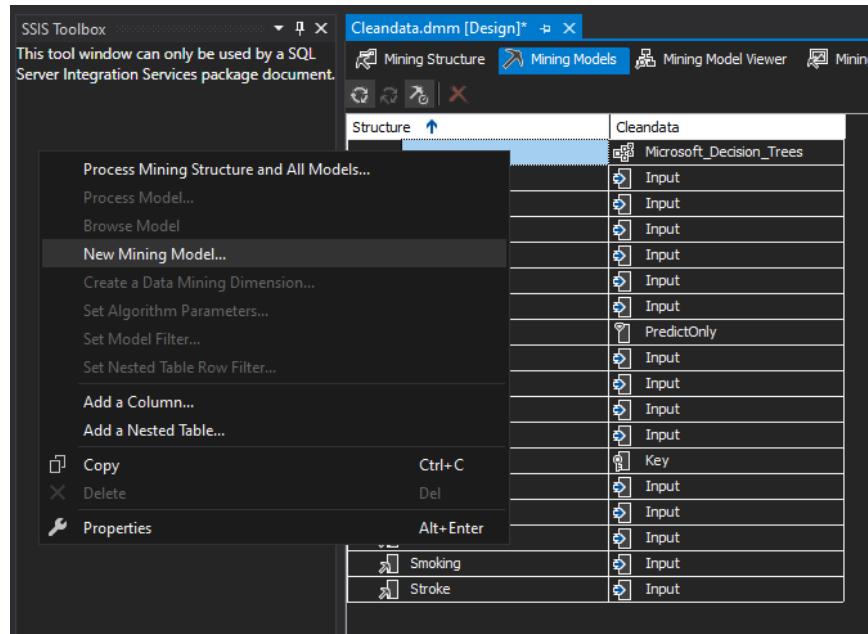
Hình 421. Mining Model sau khi tạo thuật toán Clustering

Bước 2: Tiến hành chạy project



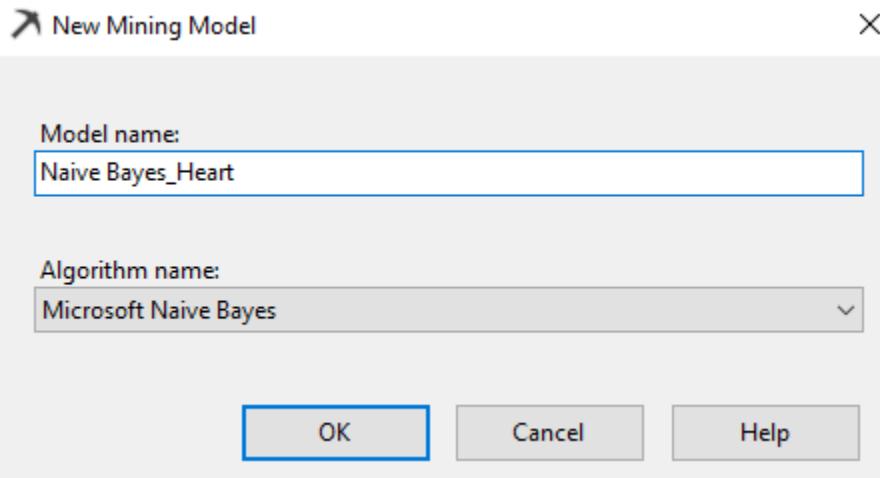
Hình 423. Kết quả sau khi chạy project

Bước 3: Thực hiện tạo thuật toán mới (Naïve Bayes)



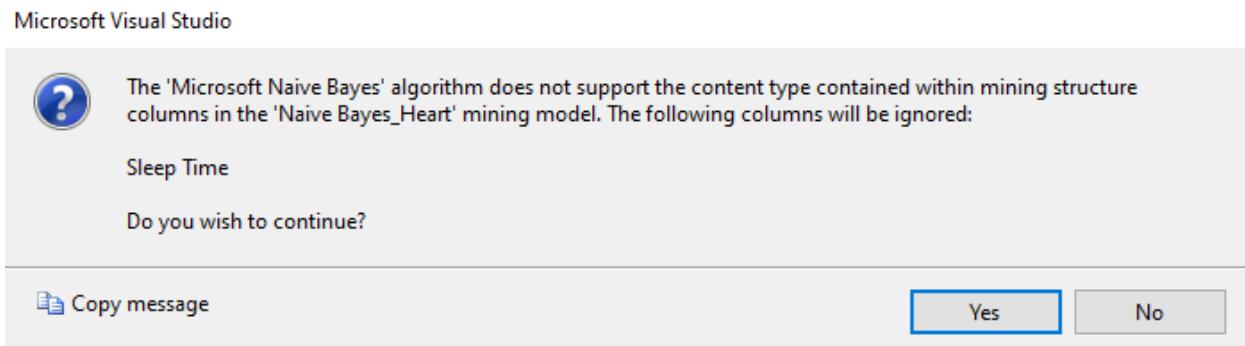
Hình 424. Thực hiện tạo mô hình khai thác

Bước 4: Tiến hành chạy project

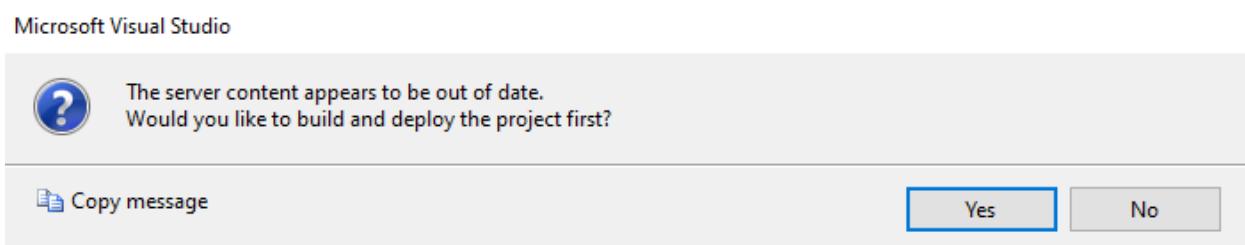


Hình 425. Chọn thuật toán khai thác

Bước 5: Tiến hành chạy project



Hình 426. Thông báo thuật toán sẽ bỏ qua thuộc tính Sleep Time

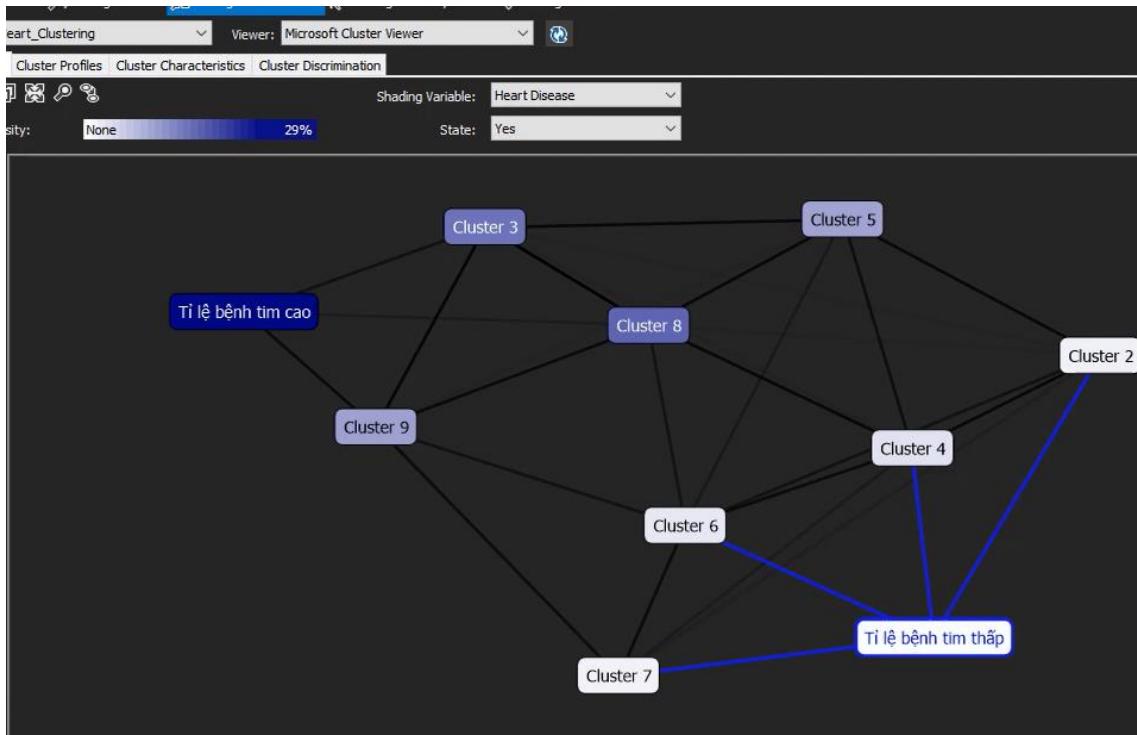


Hình 427. Thông báo chạy project

4.2 Phân tích và đưa ra tập luật

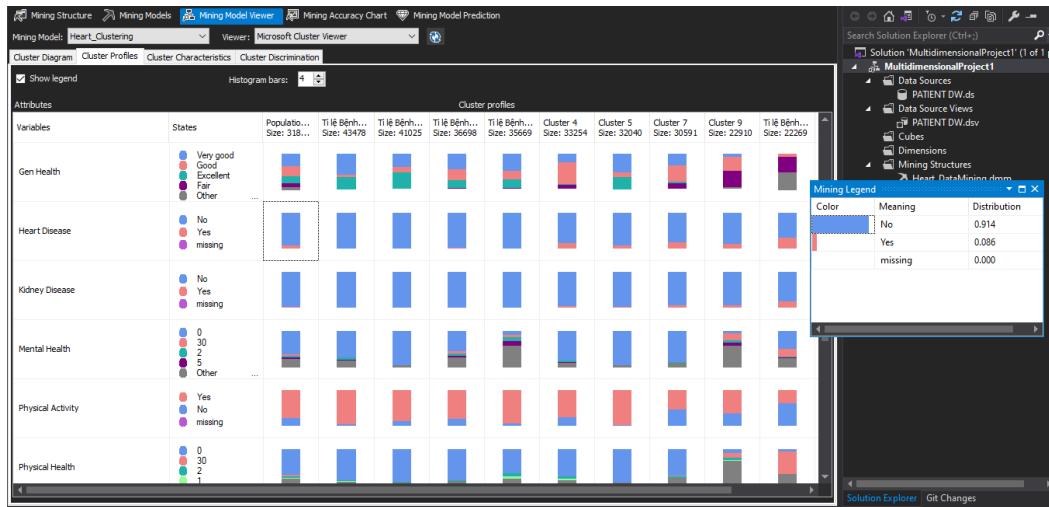
Cluster Diagram của thuật toán với:

- Shading Variable là '**HeartDisease**', value là 'yes' (mắc bệnh tim).
- Cluster đậm màu nhất: Khả năng bị bệnh tim cao nhất
- Cluster nhạt màu nhất: Khả năng bị bệnh tim thấp nhất



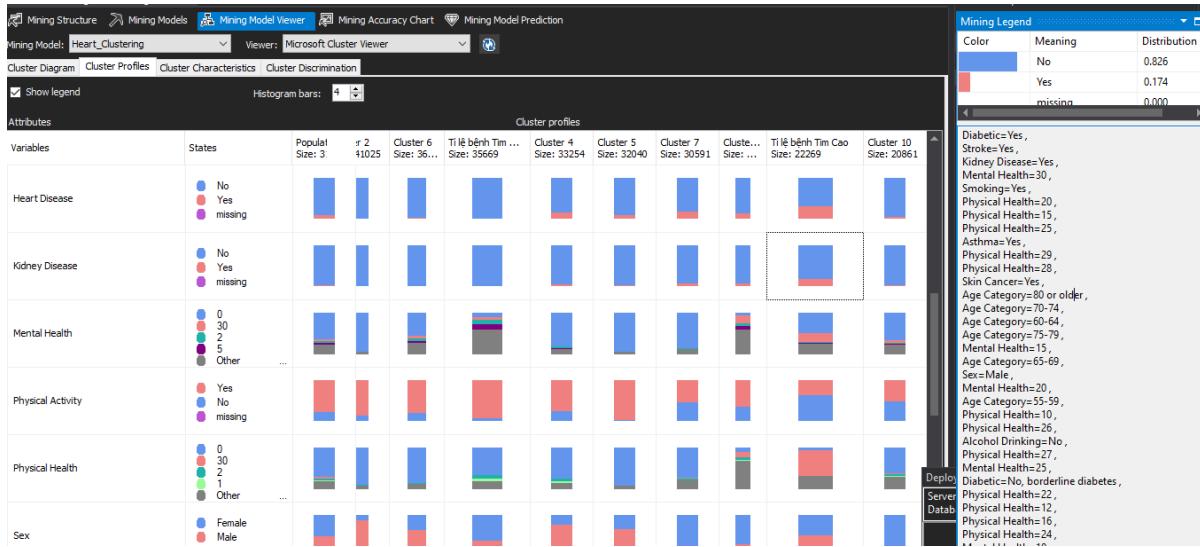
Hình 428. Mô hình Clustering với State = Yes (Mắc bệnh tim)

Tại tab ‘Cluster Profile’:



Hình 429. Cluster Profile

a. Đối với người mắc bệnh tim

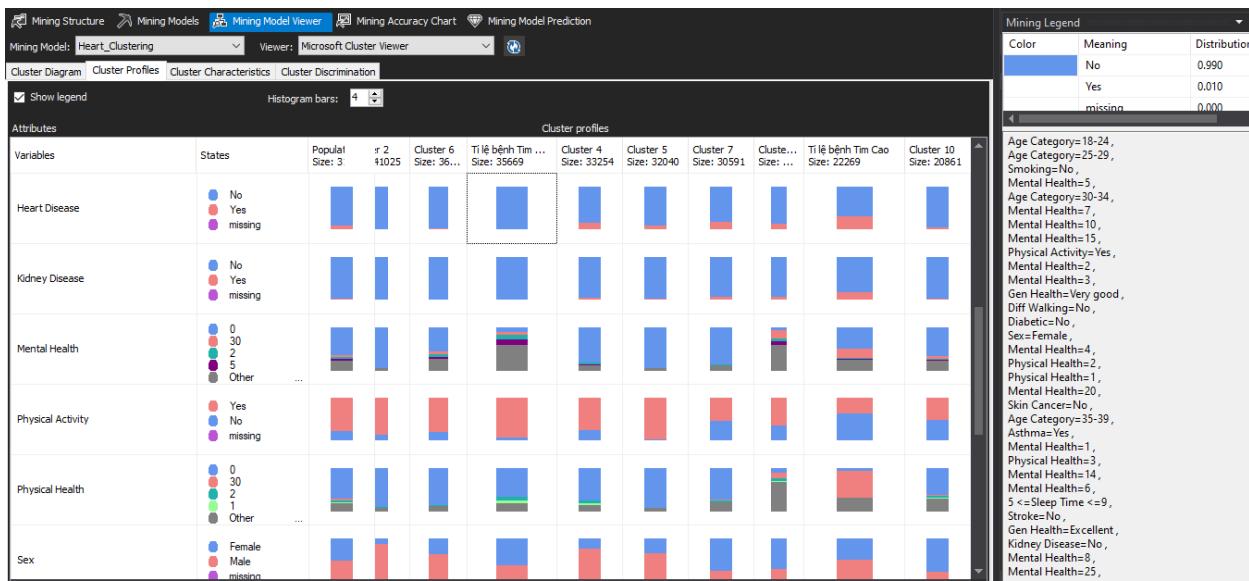


Hình 430. Mining Legend của Cluster ‘Tỉ lệ bệnh tim cao’

Với **Cluster Profiles**, Cluster tên ‘**Tỉ lệ bệnh tim Cao**’ cho ta thấy tập lục của những người được phỏng vấn bị bệnh tim chiếm 17.4%

Tập lục: Những người có độ tuổi từ 60 tuổi trở lên, giới tính nam (sex), mắc bệnh tiểu đường (Diabetic), bị đột quy (stroke), mắc bệnh về thận (kidney disease), cảm thấy không khỏe về mặt thể chất từ 15 ngày trở lên, mắc bệnh về da (Skin cancer) sẽ có tỉ lệ mắc bệnh tim cao hơn.

b. Đối với người không mắc bệnh tim



Hình 431. Mining Legend của Cluster ‘Tỉ lệ bệnh tim thấp’

Với **Cluster Profiles**, Cluster tên ‘**Tỉ lệ bệnh tim thấp**’ cho ta thấy tập lục của những người được phỏng vấn bị bệnh tim chiếm 99%

Tập lục: Những người có độ tuổi từ 18-34, giới tính nữ (sex), sức khỏe tổng quát tốt, không gặp khó khăn trong đi lại (Diff Walking), không mắc bệnh tiểu đường (Diabetic), không bị đột quy (stroke), không mắc bệnh về thận (kidney disease), không hút thuốc (Smoking), không mắc bệnh về da (Skin cancer), giấc ngủ đủ từ 5 đến 9 tiếng (Sleep Time), có tập thể dục, vận động (Physical Activity) có tỉ lệ mắc bệnh tim cao.



Hình 432. So sánh điểm giữa hai cluster ‘tỉ lệ bệnh tim cao’ và ‘tỉ lệ bệnh tim thấp’
(Clustering)

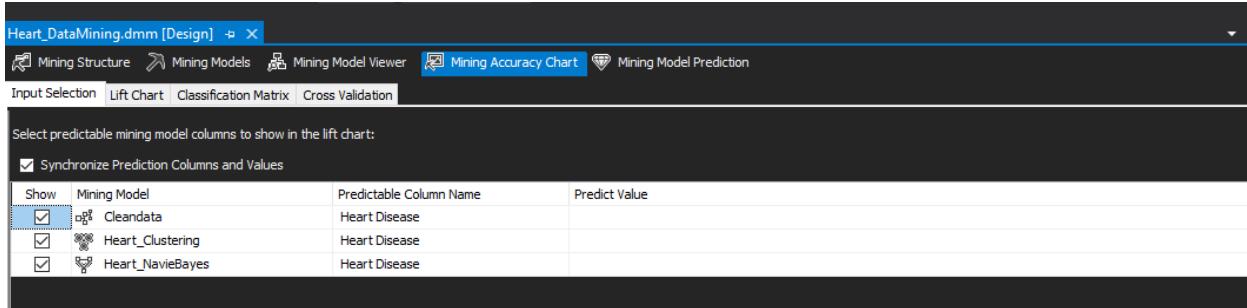


Hình 433. So sánh điểm giữa 2 thuộc tính ‘yes’ và ‘no’ của HeartDisease (Naïve Bayes)

5. SO SÁNH GIỮA CÁC THUẬT TOÁN BẰNG ĐỒ THỊ LIFT

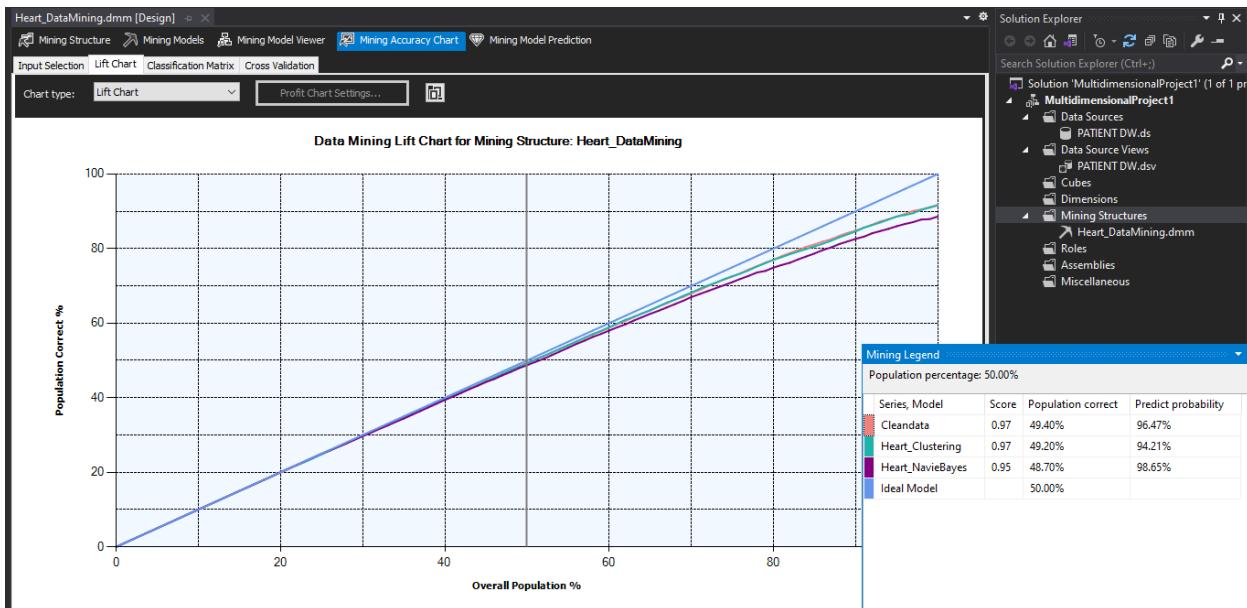
Tại tab Mining Accuracy Chart, chọn các thuật toán muốn so sánh.

- Cleandata: thuật toán cây quyết định
- Heart_Clustering: thuật toán clustering
- Heart_Naive Bayes: thuật toán Naïve Bayes



Hình 434. Lựa chọn thuật toán tại 'Input Selection'

Tại tab nhỏ ‘Lift chart’, hệ thống sẽ hiển thị biểu đồ so sánh



- Cleandata: thuật toán cây quyết định đạt độ chính xác 96.47%
- Heart_Clustering: thuật toán clustering đạt độ chính xác 94.21%
- Heart_Naive Bayes: thuật toán Naïve Bayes đạt độ chính xác 98.65%

CHƯƠNG 6: TÀI LIỆU THAM KHẢO

[1] Download SQL Server Data Tools (SSDT) for Visual Studio

<https://docs.microsoft.com/en-us/sql/ssdt/download-sql-server-data-tools-ssdt?view=sql-server-ver15>

[2] Tài liệu do giảng viên cung cấp trên course.