

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN



**BÁO CÁO ĐỒ ÁN MÔN HỌC
KHO DỮ LIỆU VÀ OLAP**

ĐỀ TÀI:

**PHÂN TÍCH DỮ LIỆU CÔNG TRÌNH THI CÔNG ĐƯỜNG BỘ
HOA KỲ (2016 - 2021)**

Nhóm sinh viên thực hiện:

Lê Trần Anh Quí MSSV: 21520094

Nguyễn Hữu Phụng MSSV: 21522481

TP. HỒ CHÍ MINH, THÁNG 5 NĂM 2024

Ths. Đỗ Thị Minh Phụng

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

....., ngày tháng năm 2023

Người nhận xét

(Ký tên và ghi rõ họ tên)

Ths. Đỗ Thị Minh Phụng

LỜI CẢM ƠN

Nhóm 39 chúng em xin gửi lời tri ân sâu sắc đến Ban Giám hiệu, thầy cô Trường Đại học Công nghệ Thông tin - Đại học Quốc gia TP.HCM và Khoa Hệ thống thông tin đã tạo điều kiện cho chúng em học tập và rèn luyện, cung cấp nền tảng kiến thức vững chắc để hoàn thành bài tập môn học.

Đặc biệt, nhóm em xin bày tỏ lòng biết ơn sâu sắc nhất đến Thạc sĩ Đỗ Thị Minh Phụng - người đã tận tình hướng dẫn, truyền đạt kiến thức, góp ý và chỉnh sửa bài tập giúp chúng em hoàn thành xuất sắc báo cáo môn học. Sự tận tâm và nhiệt huyết của thầy cô đã tiếp thêm động lực cho chúng em trong suốt quá trình học tập.

Cuối cùng, nhóm em xin gửi lời cảm ơn đến các thành viên trong nhóm. Nhờ sự chăm chỉ, kiên nhẫn và tinh thần trách nhiệm cao của mỗi cá nhân, chúng em đã cùng nhau hoàn thành bài tập đúng thời hạn và đạt được kết quả như mong muốn.

Dù đã cố gắng hết sức, chúng em nhận thức được vẫn còn nhiều thiếu sót trong bài tập. Mong thầy cô góp ý để chúng em hoàn thiện hơn trong tương lai.

Chúc thầy cô luôn dồi dào sức khỏe, thành công và hạnh phúc trong cuộc sống!

MỤC LỤC

CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI-----	8
1.1. Lý do chọn đề tài: -----	8
1.2. Giới thiệu nguồn dữ liệu: -----	8
1.2.1. Tên bộ dữ liệu: -----	8
1.2.2. Đơn vị cung cấp: -----	8
1.2.3. Mô tả số dòng, cột, thời gian thu thập: -----	9
1.2.4. Nguồn tải dataset: -----	9
1.2.5. Hướng chủ đề của kho dữ liệu: -----	9
1.2.6. Mô tả chi tiết các cột thuộc tính trong dataset: -----	9
1.2.7. Lọc dữ liệu: -----	13
1.3. Thiết kế kho dữ liệu: -----	16
1.3.1. Lược đồ kho dữ liệu: -----	16
1.3.2. Mô tả bảng Fact: -----	17
1.3.3. Mô tả chi tiết các bảng Dimension:-----	18
1.4. Các câu truy vấn -----	21
CHƯƠNG 2. XÂY DỰNG KHO DỮ LIỆU (SSIS)-----	24
2.1. Tạo kết nối và thiết lập kết nối: -----	24
2.1.1. Tạo project: -----	24
2.1.2. Tạo cơ sở dữ liệu và thiết lập kết nối:-----	25
2.1.3. Thiết lập kết nối đến database US_ROAD_WH:-----	29
2.2. Import và làm sạch dữ liệu -----	31
2.2.1. Chuẩn bị dữ liệu gốc, import dữ liệu gốc: -----	31
2.2.2. Quá trình làm sạch dữ liệu-----	35
2.3. Quá trình tạo các bảng Dimension và Fact_Raw: -----	50
2.3.1. Bảng Dim_Time -----	53
2.3.2. Bảng Dim_Weather-----	63
2.3.3. Bảng Dim_Location -----	68
2.3.4. Bảng Dim_POI -----	73

2.3.5. Bảng Fact -----	78
2.4. Merge Fact_Raw và Dim_Location thành Fact1-----	82
2.5. Merge Fact1 với Dim_Weather vào Fact2-----	99
2.6. Merge Fact2 với Dim_POI vào Fact -----	113
2.7. Tạo khóa ngoại từ bảng Fact đến các Dimension -----	129
2.8. Chạy dự án SSIS -----	132
2.8.1. Khởi chạy SSIS -----	132
2.8.2. Kiểm tra dữ liệu các bảng: -----	137
CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU-----	141
3.1. Chuẩn bị các công cụ-----	141
3.2. Tạo mới Project SSAS -----	141
3.3. Xác định dữ liệu nguồn (Data Sources) -----	142
3.4. Xác định Data Source Views -----	148
3.5. Xây dựng khối (Cubes) và xác định các độ đo (Measures) -----	154
3.6. Xác định các chiều (Dimension) -----	161
3.7. Xác định các độ đo (Measures) -----	163
3.7.1. Đổi tên và thuộc tính các độ đo ban đầu -----	163
3.7.2. Tạo các độ đo mới-----	166
3.7.3. Tổng kết độ đo -----	170
3.8. Phân cấp trong bảng chiều-----	170
3.8.1. Phân cấp bảng Dim_Time-----	170
3.8.2. Phân cấp bảng Dim_Location -----	192
3.9. Thực hiện 10 câu truy vấn – Quá trình phân tích dữ liệu bằng thao tác tay trên các khối CUBE-----	201
3.9.1. Câu truy vấn 1 -----	201
3.9.2. Câu truy vấn 2 -----	205
3.9.3. Câu truy vấn 3 -----	207
3.9.4. Câu Truy vấn 4 -----	211
3.9.5. Câu truy vấn 5 -----	213
3.9.6. Câu truy vấn 6 -----	216

3.9.7. Câu truy vấn 7 -----	218
3.9.8. Câu truy vấn 8 -----	221
3.9.9. Câu truy vấn 9 -----	223
3.9.10. Câu truy vấn 10 -----	225
3.9.11. Câu truy vấn 11 -----	227
3.9.12. Câu truy vấn 12: -----	228
3.9.13. Câu truy vấn 13-----	230
3.9.14. Câu truy vấn 14-----	233
3.9.15. Câu truy vấn 15-----	236
3.10. Quá trình phân tích dữ liệu bằng công cụ Pivot Excel -----	238
3.10.1. Câu truy vấn 1: -----	239
3.10.2. Câu truy vấn 2: -----	240
3.10.3. Câu truy vấn 3: -----	243
3.10.4. Câu truy vấn 4: -----	245
3.10.5. Câu truy vấn 5: -----	247
3.10.6. Câu truy vấn 6: -----	249
3.10.7. Câu truy vấn 7: -----	251
3.10.8. Câu truy vấn 8: -----	253
3.10.9. Câu truy vấn 9: -----	256
3.10.10. Câu truy vấn 10: -----	258
3.10.11. Câu truy vấn 11: -----	260
3.10.12. Câu truy vấn 13: -----	263
3.10.13. Câu truy vấn 14: -----	265
3.10.14. Câu truy vấn 15: -----	267
3.11. Thực hiện 10 câu truy vấn – Quá trình phân tích dữ liệu bằng ngôn ngữ truy vấn MDX-----	269
3.11.1. Câu truy vấn 1 -----	269
3.11.2. Câu truy vấn 2 -----	270
3.11.3. Câu truy vấn 3 -----	272
3.11.4. Câu truy vấn 4 -----	274
3.11.5. Câu truy vấn 5 -----	275

3.11.6. Câu truy vấn 6 -----	276
3.11.7. Câu truy vấn 7 -----	277
3.11.8. Câu truy vấn 8 -----	279
3.11.9. Câu truy vấn 9 -----	280
3.11.10. Câu truy vấn 10 -----	282
3.11.11. Câu truy vấn 11 -----	284
3.11.12. Câu truy vấn 12 -----	286
3.11.13. Câu truy vấn 13 -----	287
3.11.14. Câu truy vấn 14 -----	289
3.11.15. Câu truy vấn 15 -----	291
CHƯƠNG 4. QUÁ TRÌNH XÂY DỰNG BÁO CÁO VÀ SỬ DỤNG CÔNG CỤ POWER BI-----	293
4.1. Tạo project SSRS và thêm các Datasource và Dataset-----	293
4.2. Tạo các báo cáo từ project SSRS -----	300
4.2.1. Báo cáo 1-----	300
4.2.2. Báo cáo 2-----	329
4.2.3. Báo cáo 3-----	343
4.3. Tạo báo cáo bằng POWER BI -----	358
4.3.1. Kết nối Power BI với SQL Server -----	358
4.3.2. Thực hiện Report bằng Power BI -----	363
CHƯƠNG 5. KHAI THÁC DỮ LIỆU-----	383
5.1. Chuẩn bị dữ liệu: -----	383
5.2. Đỗ dữ liệu cho Data Mining bằng công cụ SSIS -----	384
5.3. Thực hiện Data Mining bằng SSAS:-----	390
5.4. Phân tích các mô hình khai phá dữ liệu:-----	400
5.4.1. Mô hình cây quyết định:-----	400
5.4.2. Mô hình Gom cụm:-----	407
5.5. Đánh giá độ chính xác của các mô hình thuật toán bằng Mining Accuracy Chart:-----	412
TÀI LIỆU THAM KHẢO-----	417

CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI

1.1. Lý do chọn đề tài:

Hệ thống giao thông đóng vai trò thiết yếu cho sự phát triển kinh tế và xã hội của bất kỳ quốc gia nào. Tuy nhiên, hệ thống này thường xuyên chịu ảnh hưởng bởi các công trình xây dựng và đóng cửa đường bộ, dẫn đến tắc nghẽn, ùn ú, và tăng chi phí vận tải. Việc quản lý hiệu quả các hoạt động này là một thách thức lớn đối với các nhà quản lý giao thông.

- + Kho dữ liệu "US Road Construction and Closures (2016-2021)" cung cấp thông tin cập nhật về các công trình xây dựng và đường bộ tại Hoa Kỳ trong 6 năm gần đây. Đây là một nguồn tài nguyên quý giá cho các nhà nghiên cứu, hoạch định chính sách và người dân quan tâm đến lĩnh vực giao thông.

Với tính cập nhật, khả năng tiếp cận và ứng dụng cao, nhóm chúng em tin kho dữ liệu này có thể giúp giải quyết các vấn đề giao thông hiệu quả, góp phần thúc đẩy phát triển kinh tế và xã hội.

1.2. Giới thiệu nguồn dữ liệu:

1.2.1. Tên bộ dữ liệu:

- Tên dataset: US Road Construction and Closures (2016 - 2021)
- Ngày cập nhật lần cuối: 28/03/2022

1.2.2. Đơn vị cung cấp:

- Đây là bộ dữ liệu toàn quốc về các sự kiện xây dựng và đóng cửa đường, bao gồm 49 tiểu bang của Hoa Kỳ. Các sự kiện xây dựng trong bộ dữ liệu này có thể là bất kỳ công việc đường bộ nào, từ sửa chữa via hè đến các dự án quan trọng có thể mất vài tháng để hoàn thành. Dữ liệu được thu thập từ tháng 1 năm 2016 đến tháng 12 năm 2021, sử dụng nhiều API cung cấp dữ liệu sự cố (hoặc sự kiện) lưu lượng truy cập trực tuyến. Các API này phát dữ liệu giao thông được thu thập bởi nhiều thực

thể khác nhau, chẳng hạn như Bộ Giao thông Vận tải Hoa Kỳ và tiểu bang, cơ quan thực thi pháp luật, camera giao thông và cảm biến giao thông trong mạng lưới đường bộ.

1.2.3. Mô tả số dòng, cột, thời gian thu thập:

- Bộ dữ liệu gồm 6170627 dòng với 47 cột thuộc tính.
- Thời gian thu thập: từ 2016 đến 2021

1.2.4. Nguồn tải dataset:

- Link tải: [US Road Construction and Closures \(2016 - 2021\) \(kaggle.com\)](#)

1.2.5. Hướng chủ đề của kho dữ liệu:

- Kho dữ liệu được xây dựng với hướng chủ đề là Giao thông vận tải và Công trình xây dựng.

1.2.6. Mô tả chi tiết các cột thuộc tính trong dataset:

STT	Tên thuộc tính	Ý nghĩa	Kiểu dữ liệu
1	ID	Mã định danh dòng dữ liệu (Mã sự kiện)	Varchar(50)
2	Severity	Mức độ nghiêm trọng của việc xây dựng ảnh hưởng đến giao thông nơi mà thực hiện xây dựng, sửa chữa. Mức độ được thể hiện bằng số từ 1 đến 4 cho thấy ít ảnh hưởng nhất cho đến ảnh hưởng nhiều nhất	Int
3	Start_Time	Hiển thị thời gian bắt đầu của công trình xây dựng theo múi giờ địa phương.	DATETIME

4	End_Time	Hiển thị thời gian kết thúc của công trình xây dựng theo múi giờ địa phương.	DATETIME
5	Start_Lat	Hiển thị vĩ độ trong tọa độ GPS của điểm bắt đầu.	Varchar(50)
6	Start_Lng	Hiển thị kinh độ trong tọa độ GPS của điểm bắt đầu.	Varchar(50)
7	End_Lat	Hiển thị vĩ độ trong tọa độ GPS của điểm cuối.	Varchar(50)
8	End_Lng	Hiển thị kinh độ trong tọa độ GPS của điểm cuối.	Varchar(50)
9	Distance	Độ dài của phần đường bị ảnh hưởng bởi công trình xây dựng.	decimal(18, 10)
10	Description	Hiển thị mô tả do con người cung cấp về sự kiện.	Varchar(255)
11	Number	Số đường của địa chỉ.	Int
12	Street	Tên đường của địa chỉ.	Varchar(255)
13	Side	Phía trái/phải đường của địa chỉ.	varchar(50)
14	City	Thành phố (trong cột địa chỉ).	varchar(255)
15	County	Quận.	varchar(255)
16	State	Tiểu bang.	varchar(50)
17	Zipcode	Hiển thị mã zip trong trường địa chỉ.	Int
18	Country	Quốc gia.	varchar(50)

19	Timezone	Múi giờ dựa trên vị trí nơi xảy ra công trình (phía đông, trung tâm, v.v.).	Decimal
20	Airport_Code	Chỉ định một trạm thời tiết dựa trên sân bay gần nhất với vị trí của diễn ra công trình.	Int
21	Weather_Timestamp	Hiển thị thời điểm ghi lại quan sát thời tiết (theo múi giờ địa phương).	DATETIME
22	Temperature	Hiển thị nhiệt độ (tính bằng Fahrenheit).	Decimal
23	Wind_Chill	Hiển thị gió lạnh (ở Fahrenheit).	Decimal
24	Humidity	Hiển thị độ ẩm (tính theo tỷ lệ phần trăm).	Decimal
25	Pressure	Hiển thị áp suất không khí (tính bằng inch).	Decimal
26	Visibility	Hiển thị tầm nhìn (tính bằng dặm).	Int
27	Wind_Direction	Hiển thị hướng gió.	Varchar(50)
28	Wind_Speed	Hiển thị tốc độ gió (tính bằng dặm mỗi giờ).	Decimal
29	Precipitation	Hiển thị lượng mưa tính bằng inch, nếu có.	Decimal
30	Weather_Condition	Hiển thị tình trạng thời tiết (mưa, tuyết, giông bão, sương mù, v.v.)	Varchar(255)
31	Amenity	Cho biết gần đó có sự tiện nghi nào không.	Bool

32	Bump	Cho biết có gờ giảm tốc ở gần đó không.	Bool
33	Crossing	Cho biết có đường băng qua (giao nhau) ở gần đó không.	Bool
34	Give_Way	Cho biết có biển báo nhường đường ở gần đó không.	Bool
35	Junction	Cho biết có giao lộ (nút giao thông) ở vị trí gần đó không.	Bool
36	No_Exit	Cho biết sự hiện diện của no_exit ở vị trí gần đó không.	Bool
37	Railway	Cho biết sự hiện diện của đường sắt ở một vị trí gần đó không.	Bool
38	Roundabout	Cho biết sự hiện diện của bùng binh ở một vị trí gần đó không.	Bool
39	Station	Cho biết sự hiện diện của trạm ở vị trí gần đó không.	Bool
40	Stop	Cho biết sự hiện diện của biển báo dừng ở vị trí gần đó không.	Bool
41	Traffic_Calming	Cho biết sự hiện diện của biển giảm tốc độ ở một vị trí gần đó không.	Bool
42	Traffic_Signal	Cho biết sự hiện diện của đèn giao thông ở vị trí gần đó không.	Bool
43	Turning_Loop	Cho biết sự hiện diện của vòng xoay ở vị trí gần đó không.	Bool

44	Sunrise_Sunset	Hiển thị giai đoạn của ngày (tức là ban ngày hoặc ban đêm) dựa trên lặn/mọc của mặt trời. Dựa trên thời gian bắt đầu của sự kiện.	Varchar(50)
45	Civil_Twilight	Hiển thị khoảng thời gian trong ngày (tức là ngày hoặc đêm) dựa trên chang vang. Điều này dựa trên thời gian bắt đầu của sự kiện.	Varchar(50)
46	Nautical_Twilight	Hiển thị khoảng thời gian trong ngày (tức là ngày hoặc đêm) dựa trên hoàng hôn hải lý. Điều này dựa trên thời gian bắt đầu của sự kiện.	Varchar(50)
47	Astronomical_Twilight	Hiển thị khoảng thời gian trong ngày (tức là ngày hoặc đêm) dựa trên hoàng hôn thiên văn. Điều này dựa trên thời gian bắt đầu của sự kiện.	Varchar(50)

1.2.7. Lọc dữ liệu:

Vì dataset **US Road Construction and Closures (2016 - 2021)** khá lớn (lên tới 2,6GB với hơn 6 triệu dòng dữ liệu) nên laptop của nhóm chúng em khó có thể xử lý hết được nên nhóm chúng em quyết định sẽ lấy dữ liệu từ năm 2017 tới năm 2019 để xây dựng cho kho dữ liệu. Nhóm chúng em sẽ sử dụng python để lọc dữ liệu.

1.2.7.1. Import các thư viện cần thiết và file dataset:

```

import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
import seaborn as sns
import scipy

df = pd.read_csv("US_Constructions_Dec21.csv", index_col = False)

pd.set_option('display.max_columns', None)

df.shape

(6170627, 47)

df.head()

```

	ID	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Description	Number	Street	Side	Gi
0	C-1	4	2019-04-05 16:00:00.000000000	2020-09-29 11:53:57.000000000	32.838360	-93.152378	32.850740	-93.164388	1.103497	Construction on LA-534 WB near EDMONDS LOOP Ro...	4200.0	Highway 534	R	Haynesvi
1	C-2	2	2021-11-12 07:59:00.000000000	2021-11-12 08:22:30.000000000	30.221331	-92.008625	30.216642	-92.003809	0.433173	Slow traffic on US-90 E from US-167/Louisiana —	1098.0	Evangeline Trwy	R	Lafayef
2	C-3	2	2021-10-12 07:17:30.000000000	2021-10-12 09:18:55.000000000	39.653153	-104.910224	39.653120	-104.913838	0.192266	Slow traffic on CO-30 from S Tamarac Dr (E Ham...	6779.0	E Hampden Ave	R	Dens
3	C-4	4	2021-02-10 02:46:10.000000000	2021-02-17 23:59:00.000000000	33.961506	-118.029339	33.961919	-118.029082	0.032112	Closed road from Whittier to College Ave due t...	13585.0	Whittier Blvd	L	Whitt
4	C-5	2	2020-09-24 15:58:00.000000000	2020-09-25 21:04:54.000000000	40.008734	-79.599696	40.022822	-79.595703	0.996057	Construction on US-119 NB near SAMPSON ST Alto...	1144.0	Schley St	R	Connellsvi

1.2.7.2. Tạo thêm cột thuộc tính "Duration":

- + Cột “Duration” tính khoảng thời gian thực hiện các công trình xây dựng để phục vụ tạo kho dữ liệu. Nhóm chúng em sẽ lấy thời gian kết thúc xây dựng (End_Time) trừ cho thời gian bắt đầu xây dựng (Start_Time). Sau đó, dùng hàm total_seconds() để lấy ra khoảng thời gian đó (tính bằng giây).
- + Định dạng lại thời gian.

Ths. Đỗ Thị Minh Phụng

	ID	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Description	Number	Street	Side	City
0	C-1	4	2019-04-05 16:00:00	2020-09-29 11:53:57	32.838360	-93.152378	32.850740	-93.164388	1.103497	Construction on LA-534 WB near EDMONDS LOOP Rd...	4200.0	Highway 534	R	Haynesville
1	C-2	2	2021-11-12 07:59:00	2021-11-12 08:22:30	30.221331	-92.008625	30.216642	-92.003809	0.433173	Slow traffic on US-90 E from US-167/Louisiana...	1098.0	Evangeline Trwy	R	Lafayette
2	C-3	2	2021-10-12 07:17:30	2021-10-12 09:18:55	39.653153	-104.910224	39.653120	-104.913838	0.192266	Slow traffic on CO-30 from S Tamarac Dr (E Ham...	6779.0	E Hampden Ave	R	Denver
3	C-4	4	2021-02-10 02:46:10	2021-02-17 23:59:00	33.961506	-118.029339	33.961919	-118.029082	0.032112	Closed road from Whittier to College Ave due t...	13585.0	Whittier Blvd	L	Whittier
4	C-5	2	2020-09-24 15:58:00	2020-09-25 21:04:54	40.008734	-79.599696	40.022822	-79.595703	0.996057	Construction on US-119 NB near SAMPSON ST Allo...	1144.0	Schley St	R	Connellsville

+ Thêm cột Duration.

	df['Start_Time'] = pd.to_datetime(df['Start_Time'],format ='%Y-%m-%d %H:%M:%S')	df['End_Time'] = pd.to_datetime(df['End_Time'],format ='%Y-%m-%d %H:%M:%S')	df['Duration'] = (df['End_Time'] - df['Start_Time']).dt.total_seconds().astype(int)
<code>df</code>			
xit Railway Roundabout Station Stop Traffic_Calming Traffic_Signal Turning_Loop Sunrise_Sunset Civil_Twilight Nautical_Twilight Astronomical_Twilight Duration			
Ise False False False False False False False Day Day Day Day Day 46900437			
Ise False False False False False False False Day Day Day Day Day 1410			
Ise False False True False False False False Day Day Day Day Day 7285			
Ise False False False False False False False Night Night Night Night Night 681170			
Ise False False False False False False False Day Day Day Day Day 104814			

1.2.7.3. Lấy ra các thuộc tính cần thiết để tạo kho dữ liệu:

Ths. Đỗ Thị Minh Phụng

	ID	Severity	Start_Time	End_Time	Duration	Distance(mi)	Street	Side	City	County	...	Roundabout	Station	Stop	Traffic_Calming
0	C-1	4	2019-04-05 16:00:00	2020-09-29 11:53:57	46900437	1.103497	Highway 534	R	Haynesville	Claiborne	...	False	False	False	False
1	C-2	2	2021-11-12 07:59:00	2021-11-12 08:22:30	1410	0.433173	SW Evangeline Trwy	R	Lafayette	Lafayette	...	False	False	False	False
2	C-3	2	2021-10-12 09:18:55	2021-10-12 09:18:55	7285	0.192266	Hampden Ave	R	Denver	Denver	...	False	True	False	False
3	C-4	4	2021-02-10 02:46:10	2021-02-17 23:59:00	681170	0.032112	Whittier Blvd	L	Whittier	Los Angeles	...	False	False	False	False
4	C-5	2	2020-09-24 15:58:00	2020-09-25 21:04:54	104814	0.996057	Schley St	R	Connellsville	Fayette	...	False	False	False	False
...	
6170622	C-6293072	2	2020-01-10 10:39:00	2021-10-19 01:02:59	55952639	0.644691	US-1 S	R	Feasterville Trevose	Bucks	...	False	False	False	False

1.2.7.4. Lọc dữ liệu trong khoảng thời gian từ 2017 đến 2019:

	ID	Severity	Start_Time	End_Time	Duration	Distance(mi)	Street	Side	City	County	...	Roundabout	Station	Stop	Traffic_Calming
15	C-16	2	2019-11-06 09:04:01	2019-11-10 09:03:37	345576	0.118220	Van Dyke Ave	L	Center Line	Macomb	...	False	False	False	False
17	C-18	4	2018-10-08 04:29:57	2018-10-08 06:29:45	7188	0.219182	North Innerbelt W	R	Columbus	Franklin	...	False	False	False	False
27	C-28	4	2017-08-03 21:41:25	2017-08-04 03:47:06	21941	0.106878	Market St	L	Lemoyne	Cumberland	...	False	False	False	False
28	C-29	4	2019-05-14 16:35:59	2019-05-14 21:35:49	17990	0.235121	Midland Blvd	R	Nampa	Canyon	...	False	False	False	False

- + Lưu dataset lại thành file csv “USRoadConstructionandClosures(2017 - 2019).csv”:

```
df1_filtered.to_csv('USRoadConstructionandClosures(2017 - 2019).csv', index=False)
```

1.3. Thiết kế kho dữ liệu:

1.3.1. Lược đồ kho dữ liệu:



1.3.2. Mô tả bảng Fact:

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa/ghi chú
1	id	Varchar(50)	PK	Mã định danh dòng dữ liệu (Mã sự kiện)
2	severity	Int		Mức độ nghiêm trọng của việc xây dựng ảnh hưởng đến giao thông nơi mà thực hiện xây dựng, sửa chữa. Mức độ được thể hiện bằng số từ 1 đến 4 cho thấy ít ảnh hưởng nhất cho đến ảnh hưởng nhiều nhất

3	duration	Int		Khoảng thời gian diễn ra sự kiện xây dựng/đóng cửa đường bộ
4	distance	decimal(18, 10)		Độ dài của phần đường bị ảnh hưởng bởi công trình xây dựng.
5	location_ID	Varchar(50)	FK	Mã địa điểm
6	weather_ID	Varchar(50)	FK	Mã thời tiết
7	time_ID	Varchar(50)	FK	Mã thời gian
8	POI_ID	Varchar(50)	FK	Mã POI

1.3.3. Mô tả chi tiết các bảng Dimension:

1.3.3.1. Dim_Location

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa/ghi chú
1	location_ID	Varchar(50)	PK	Mã địa điểm
2	street	Varchar(255)		Tên đường của địa chỉ.
3	side	varchar(50)		Phía trái/phải đường của địa chỉ
4	city	varchar(255)		Thành phố (trong cột địa chỉ)
5	county	varchar(255)		Quận

6	state	varchar(50)		Tiêu bang
---	-------	-------------	--	-----------

1.3.3.2. Dim_POI

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa/ghi chú
1	POI_ID	Varchar(50)	PK	Mã POI
2	amenity	int		Cho biết gần đó có sự tiện nghi nào không.
3	bump	int		Cho biết có gờ giảm tốc ở gần đó không.
4	crossing	int		Cho biết có đường băng qua (giao nhau) ở gần đó không.
5	give_way	int		Cho biết có biển báo nhường đường ở gần đó không.
6	junction	int		Cho biết có giao lộ (nút giao thông) ở vị trí gần đó không.
7	no_exit	int		Cho biết sự hiện diện của no_exit ở vị trí gần đó không.
8	railway	int		Cho biết sự hiện diện của đường sắt ở một vị trí gần đó không.
9	roundabout	int		Cho biết sự hiện diện của bùng binh ở một vị trí gần đó không.

10	station	int		Cho biết sự hiện diện của trạm ở vị trí gần đó không.
11	stop	int		Cho biết sự hiện diện của biển báo dừng ở vị trí gần đó không.
12	traffic_calming	int		Cho biết sự hiện diện của biển giảm tốc độ ở một vị trí gần đó không.
13	traffic_signal	int		Cho biết sự hiện diện của đèn giao thông ở vị trí gần đó không.
14	turning_loop	int		Cho biết sự hiện diện của vòng xoay ở vị trí gần đó không.

1.3.3.3. Dim_Time

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa/ghi chú
1	time_ID	Varchar(50)	PK	Mã thời gian
2	start_year	int		Năm bắt đầu
3	start_month	int		Tháng bắt đầu
4	start_day	int		Ngày bắt đầu
5	end_year	int		Năm kết thúc
6	end_month	int		Tháng kết thúc
7	end_day	int		Ngày kết thúc

8	start_hour	Int		Hiển thị giờ bắt đầu sự kiện xây dựng/dóng cửa đường bộ
9	start_minute	Int		Hiển thị phút bắt đầu sự kiện xây dựng/dóng cửa đường bộ
10	end_hour	Int		Hiển thị giờ kết thúc sự kiện xây dựng/dóng cửa đường bộ
11	end_minute	Int		Hiển thị phút kết thúc sự kiện xây dựng/dóng cửa đường bộ

1.3.3.4. Dim_Weather

STT	Tên thuộc tính	Kiểu dữ liệu	Ràng buộc	Ý nghĩa/ghi chú
1	weather_ID	Varchar(50)	PK	Mã thời tiết
2	temperature	decimal(18, 2)		Hiển thị nhiệt độ (tính bằng Fahrenheit).
3	humidity	decimal(18, 2)		Hiển thị độ ẩm (tính theo tỷ lệ phần trăm).
4	wind_speed	decimal(18, 2)		Hiển thị tốc độ gió (tính bằng dặm một giờ).
5	weather_condition	varchar(255)		Hiển thị tình trạng thời tiết (mưa, tuyết, giông bão, sương mù, v.v.)

1.4. Các câu truy vấn

Ths. Đỗ Thị Minh Phụng

Câu 1: Thống kê số lượng công trình làm ảnh hưởng tới đường bộ xảy ra tại 4 quận Kent, Newport, Providence và Washington của bang Rhode Island (RI) trong 3 năm liên tục 2017 – 2019.

Câu 2: Thống kê tổng số công trình, mức độ ảnh hưởng trung bình, chiều dài trung bình và thời gian thi công trung bình của đoạn đường mà có công trình thi công ảnh hưởng tới đường bộ theo các tháng trong năm 2019.

Câu 3: Thống kê với 5 loại điều kiện thời tiết xảy ra nhiều nhất có tổng công trình được thực hiện nhiều nhất, theo từng tháng, năm và tính tổng chiều dài các đoạn đường có công trình thi công đó.

Câu 4: Thống kê tổng số công trình (Fact Count) trên 6 tiểu bang (State): California (CA), New York (NY), Florida (FL), Ohio (OH), Texas (TX) và Washington (WA) có thời gian bắt đầu 2017 và kết thúc 2019.

Câu 5: Thống kê số công trình và trung bình độ dài đoạn đường thi công với nhiệt độ trên 70F và độ ẩm trên 70%.

Câu 6: Thống kê theo từng bang có số lượng các công trình thi công lớn 10000.

Câu 7: Liệt kê 10 tiểu bang bị ảnh hưởng đến giao thông đường bộ bởi các công trình xây dựng nhiều nhất trong 3 năm từ 2017 tới 2019. Với 10 tiểu bang đó, liệt kê tên các con đường và phía thi công có tổng độ dài đoạn đường bị ảnh hưởng bởi các công trình lớn hơn 100 miles.

Câu 8: Với từng thành phố thuộc quận Orange của bang California, hãy cho biết độ dài đường thi công lớn nhất, tổng độ dài thi công và tỉ lệ giữa công trình thi công dài nhất so với tổng độ dài trong năm 2019.

Câu 9: Thống kê tổng số công trình theo từng ngày, tháng, năm.

Câu 10: Liệt kê số công trình thi công có các cơ sở vật chất giao thông gần đó như giao lộ (junction) và bùng binh (roundabout) theo thành phố, quận, bang trong 3 năm 2017- 2019.

Ths. Đỗ Thị Minh Phụng

Câu 11: Cho biết số công trình và mức độ nghiêm trọng trung bình của các công trình đó xảy ra tại bang Texas trong điều kiện thời tiết có liên quan tới bão (Thunderstorm) trong 3 năm 2017 – 2019.

Câu 12: Cho biết tiểu bang có số công trình nhiều nhất và tiểu bang có số công trình ít nhất theo từng năm.

Câu 13: Cho biết tốc độ hoàn thành trung bình của mỗi công trình ở gần vị trí gờ giảm tốc (bump) theo từng bang trong cả 3 năm 2017 – 2019.

Câu 14: Thống kê trung bình mức độ nghiêm trọng của các công trình thi công đường tại các bang ở dưới mức 3 trong năm 2017 nhưng đã lên hơn mức 3 trong năm 2018.

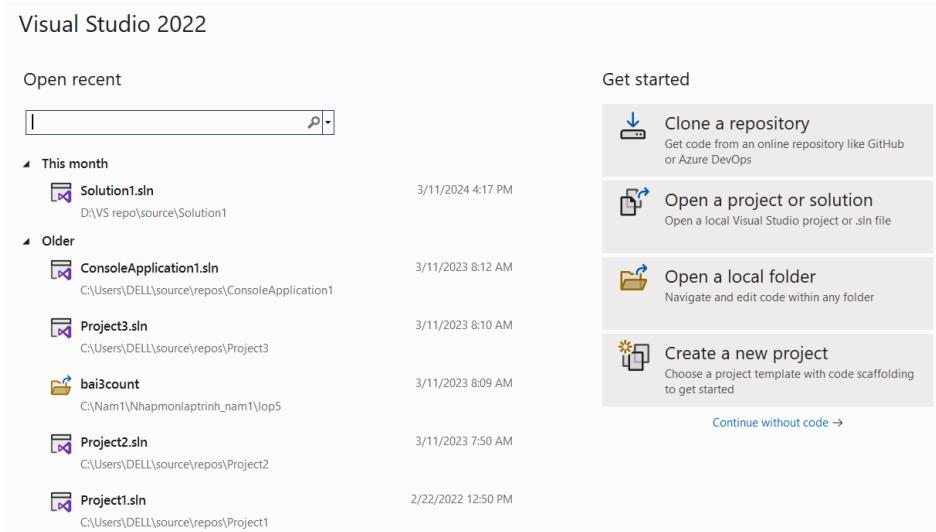
Câu 15: Thống kê thời gian thi công trung bình và khoảng đường thi công trung bình theo từng thành phố trong điều kiện thời tiết khắc nghiệt là có tốc độ gió trên 70mph

CHƯƠNG 2. XÂY DỰNG KHO DỮ LIỆU (SSIS)

2.1. Tạo kết nối và thiết lập kết nối:

2.1.1. Tạo project:

Bước 1: Mở Visual Studio và tạo một project mới (**Create a new project**)



Bước 2: Ở màn hình tiếp theo tìm kiếm từ khóa “**intergration**” và chọn **Integration Service Project**

Ths. Đỗ Thị Minh Phụng

Create a new project

Recent project templates

All languages All platforms All project types

Blank Solution

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

Integration Services Project (Azure-Enabled) **New**
This project may be used for building high performance data integration and workflow solutions that can also be run/debugged on SSIS Platform-as-a-Service (PaaS) in Azure Data Factory.

Integration Services Import Project Wizard **New**
A wizard that assists you in creating a new Integration Services (SSIS) project that is based on an existing one. Import from a project deployment file (.ispac extension) or from an Integration Services catalog on an instance of SQL Server.

Not finding what you're looking for?
[Install more tools and features](#)

Bước 3: Đặt tên project và chọn địa chỉ lưu project đó

Configure your new project

Integration Services Project

Project name

Demo_Olap_SSIS

Location

D:\VS repo\source



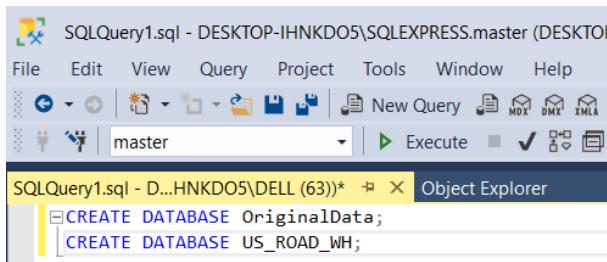
Solution name (1)

Demo_Olap_SSIS

Place solution and project in the same directory

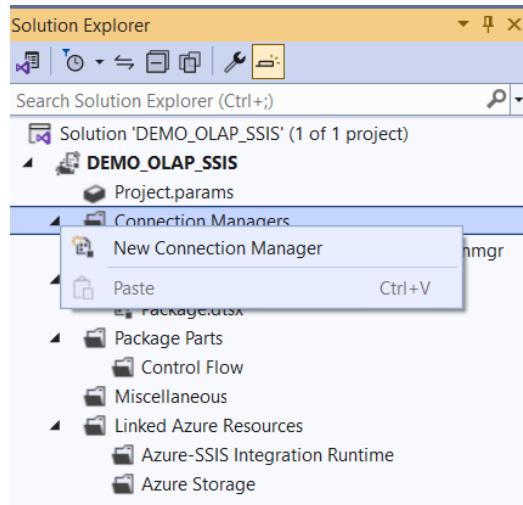
2.1.2. Tạo cơ sở dữ liệu và thiết lập kết nối:

Bước 1: Tạo cơ sở dữ liệu:

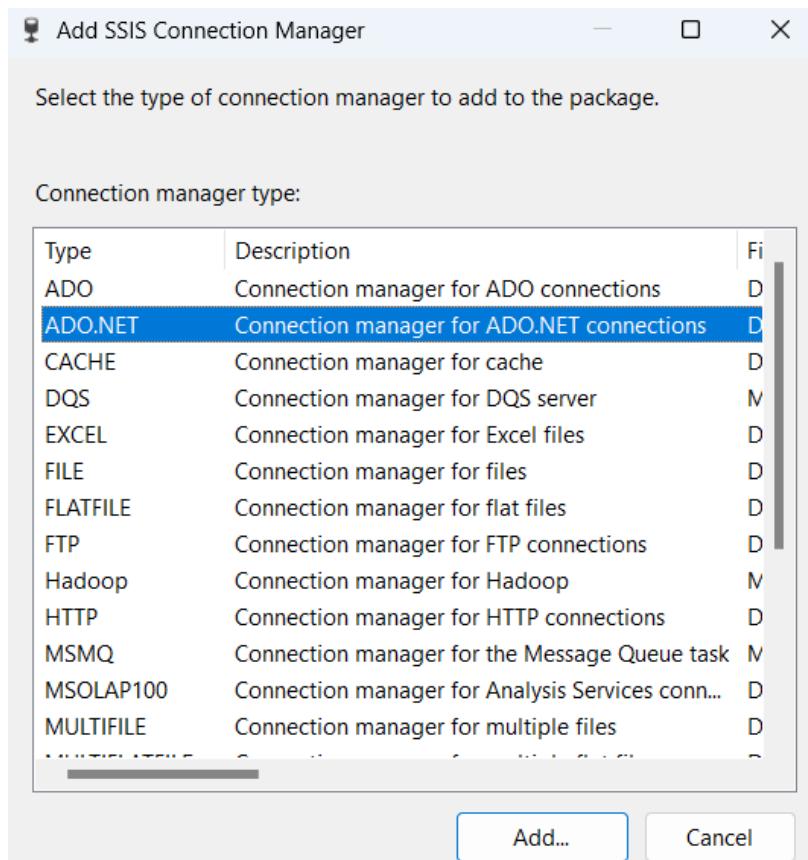


```
SQLQuery1.sql - DESKTOP-IHNKDO5\SQLEXPRESS.master (DESKTOP-IHNKDO5)
File Edit View Query Project Tools Window Help
New Query MDX DMX XMLA
master Execute ✓
SQLQuery1.sql - D...HNKDO5\DELL (63)* Object Explorer
CREATE DATABASE OriginalData;
CREATE DATABASE US_ROAD_WH;
```

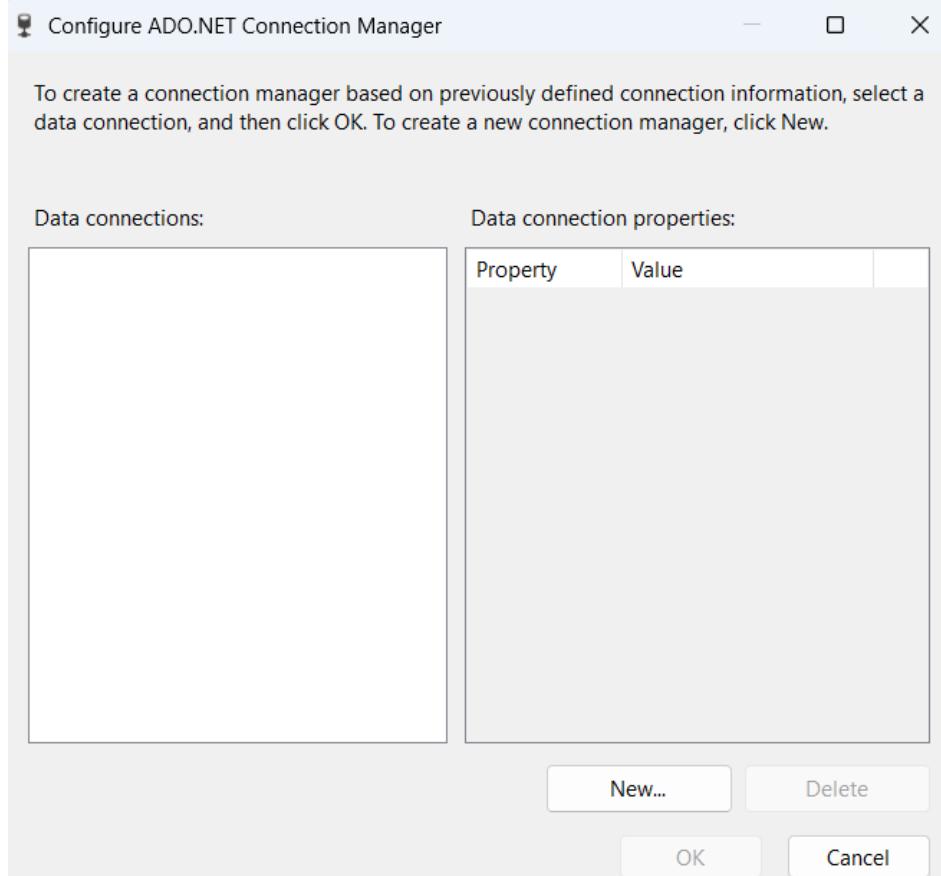
- + **OriginalData:** là database chứa dữ liệu gốc, bao gồm các bảng.
- + **DataGoc:** được import thẳng từ file csv.
- + **Data_Clean:** chứa dữ liệu sử dụng sau khi làm sạch.
- + **DataNull:** chứa dữ liệu Null trong quá trình làm sạch.
- + **Bước 2:** Thiết lập kết nối:
 - + Thiết lập kết nối đến database **OriginalData:**
 - + Click chuột phải vào Connection Manager trong tab Solution Explorer và chọn New Connection Manager.



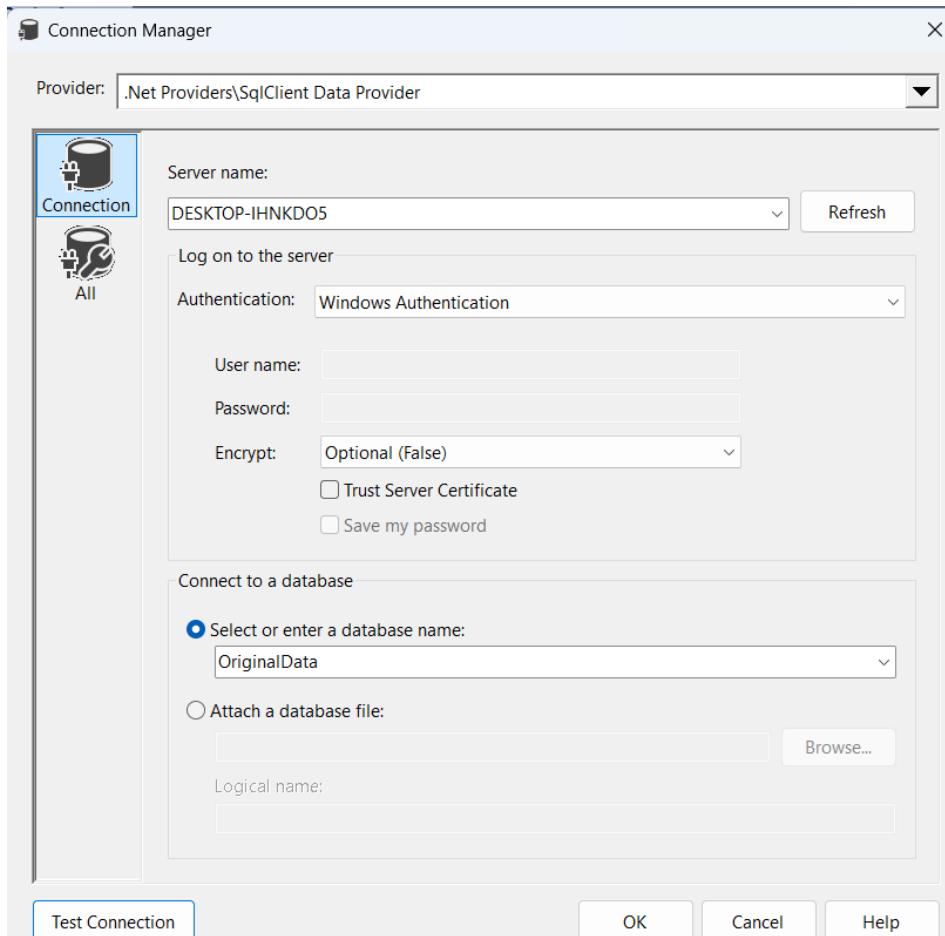
- + Nhấn **Add** thực hiện việc tạo kết nối với database của SQL Server



- + Click **New...**

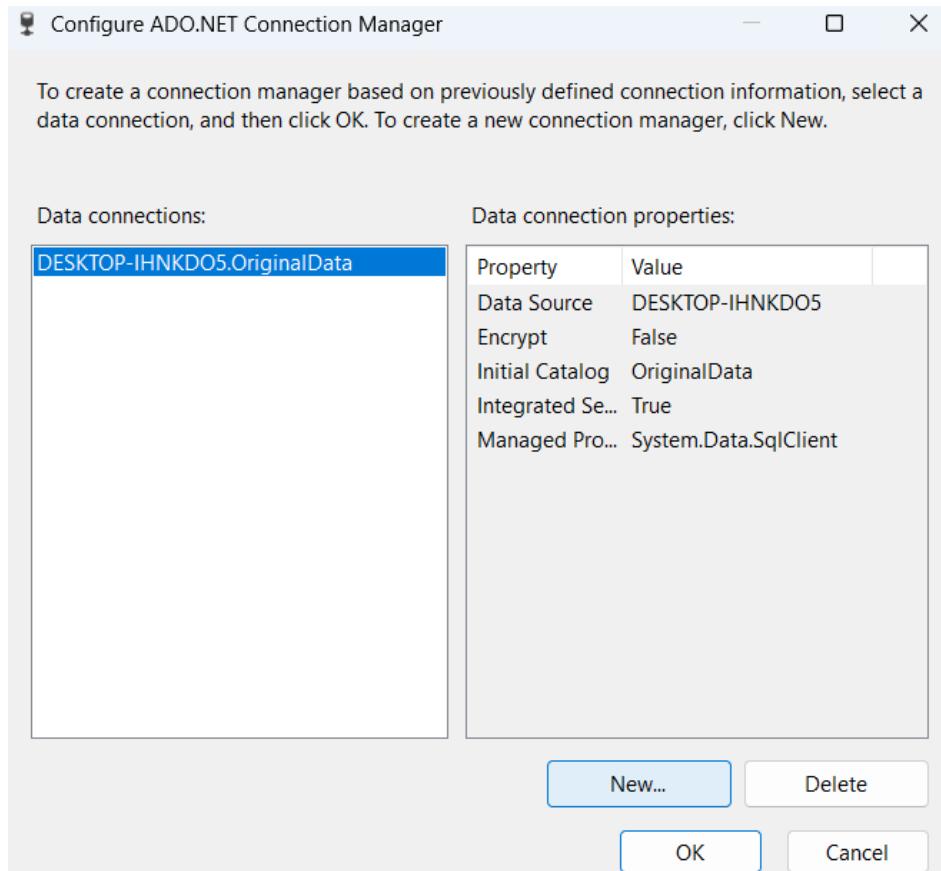


- + Chọn tên server của máy tính trong ô **Server name** và chọn database **OriginalData** ở ô **Connect to a database** và nhấp **OK**

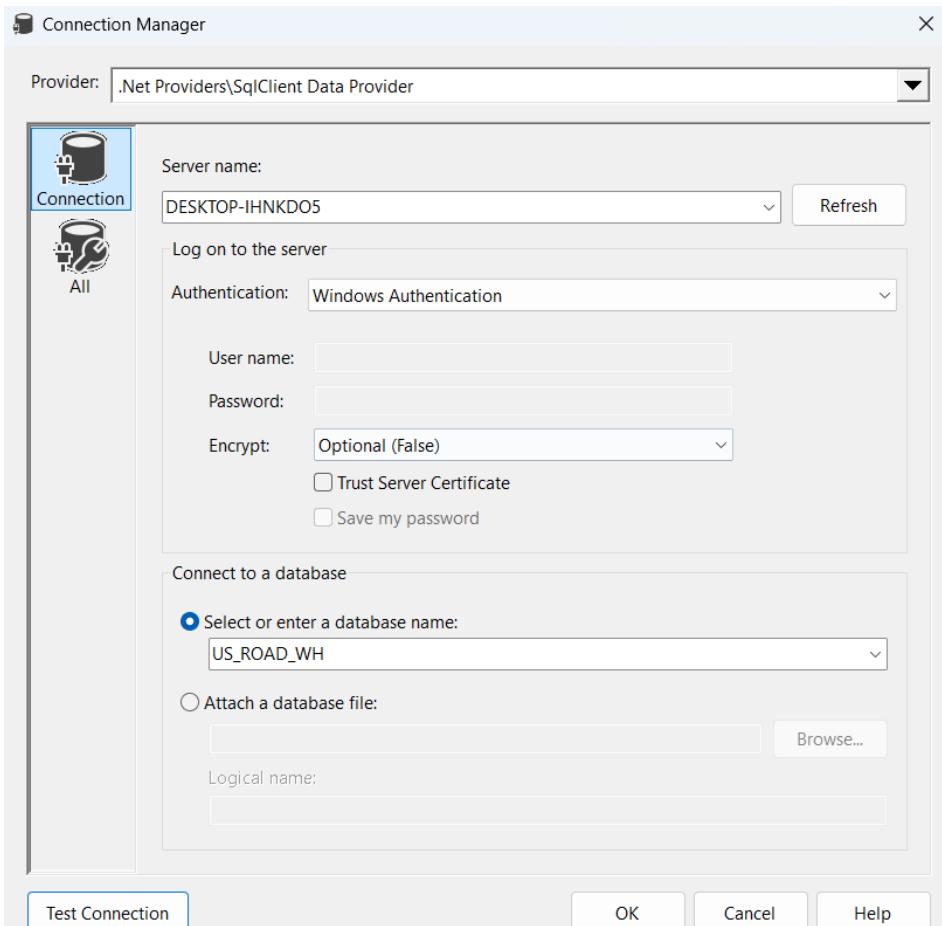


2.1.3. Thiết lập kết nối đến database US_ROAD_WH:

- + **Bước 1:** Nhấn New.. ở cửa sổ tiếp theo hiện lên (**Configure ADO NET Connection Manager**)



Bước 2: Chọn Database US_ROAD_WH trong ô select or enter database name và nhấn OK để hoàn tất



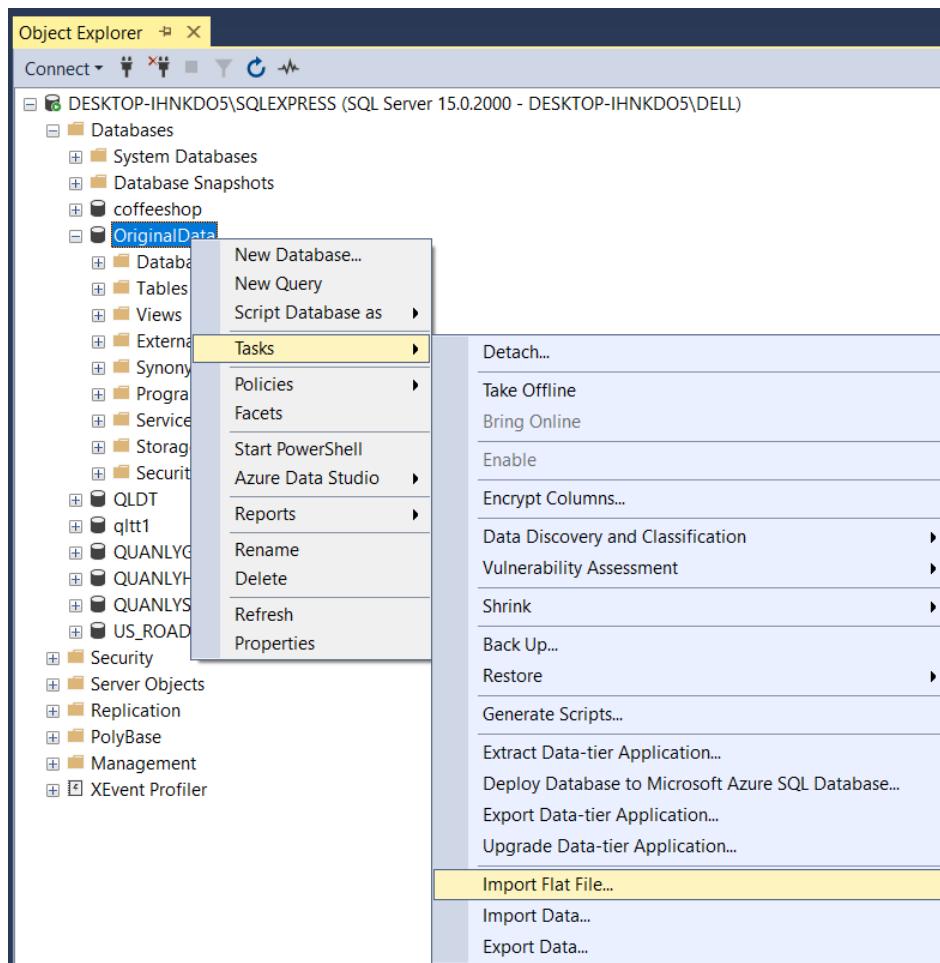
2.2. Import và làm sạch dữ liệu

2.2.1. Chuẩn bị dữ liệu gốc, import dữ liệu gốc:

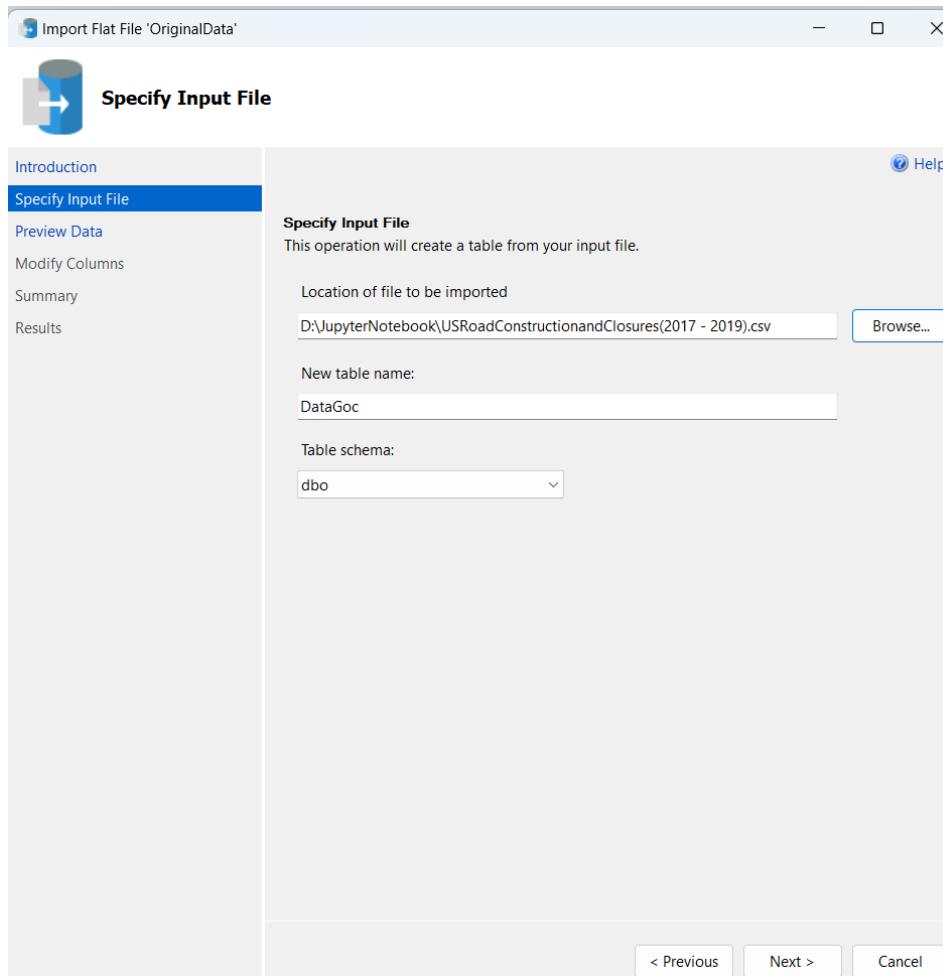
Dữ liệu được chuẩn bị là 1 file csv được thu thập trước và sau đó sẽ tiến hành Import và tạo một bảng chứa toàn bộ dữ liệu gốc này có tên là “**DataGoc**”

Tiếp theo, click chuột phải vào database **OriginalData** -> Chọn **Tasks** -> Chọn **Import Flat File...**

Cửa sổ Import Flat File “OriginalData” hiện lên và nhấn **Next**



Nhập địa chỉ file csv cần tìm vào **Location of file to be imported** và chỉnh sửa tên table thành **DataGoc** rồi nhấn **Next** để sang thẻ **Preview Data** sau đó tiếp tục nhấn **Next** để chuyển sang thẻ **Modify columns**.



Tùy chỉnh kiểu dữ liệu của các cột thuộc tính cũng như khóa chính, allow nulls và tiếp tục nhấn **Next** rồi nhấn **Finish**

Ths. Đỗ Thị Minh Phụng

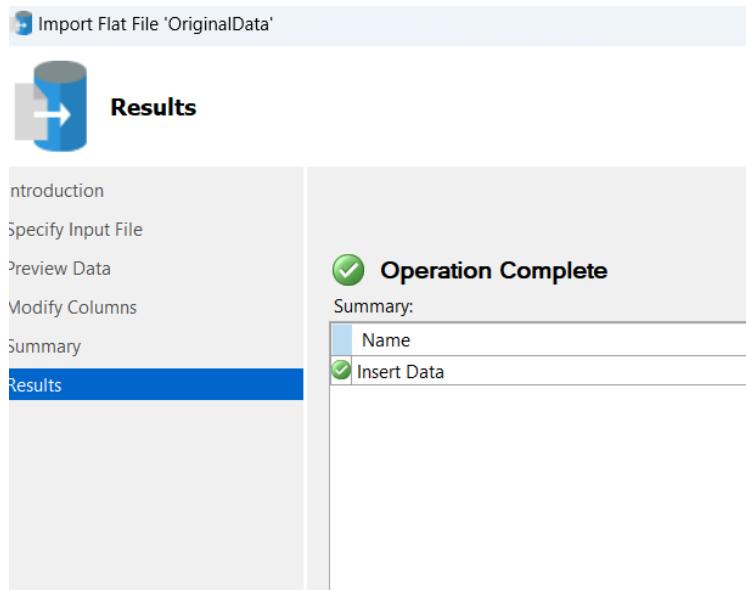
Import Flat File 'OriginalData'

Modify Columns

Introduction Specify Input File Preview Data Modify Columns Summary Results

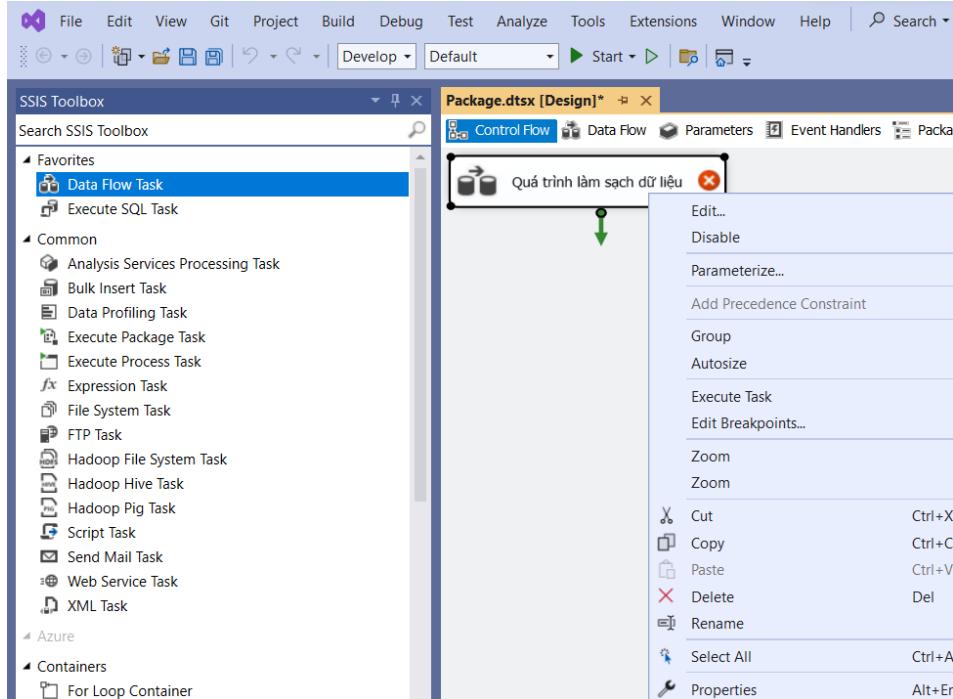
Modify Columns
This operation generated the following table schema. Please verify if schema is accurate, and if not, please make any changes.

Column Name	Data Type	Primary Key	<input type="checkbox"/> Allow Nulls
Side	nvarchar(200)	<input type="checkbox"/>	<input type="checkbox"/>
City	nvarchar(200)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
County	nvarchar(200)	<input type="checkbox"/>	<input type="checkbox"/>
State	nvarchar(200)	<input type="checkbox"/>	<input type="checkbox"/>
Temperature_F	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Humidity	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Wind_Speed_mph	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Weather_Condition	nvarchar(200)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Amenity	bit	<input type="checkbox"/>	<input type="checkbox"/>
Bump	bit	<input type="checkbox"/>	<input type="checkbox"/>
Crossing	bit	<input type="checkbox"/>	<input type="checkbox"/>
Give_Way	bit	<input type="checkbox"/>	<input type="checkbox"/>
Junction	bit	<input type="checkbox"/>	<input type="checkbox"/>
No_Exit	bit	<input type="checkbox"/>	<input type="checkbox"/>
Railway	bit	<input type="checkbox"/>	<input type="checkbox"/>
Roundabout	bit	<input type="checkbox"/>	<input type="checkbox"/>
Station	bit	<input type="checkbox"/>	<input type="checkbox"/>
Stop	bit	<input type="checkbox"/>	<input type="checkbox"/>
Traffic_Calming	bit	<input type="checkbox"/>	<input type="checkbox"/>
Traffic_Signal	bit	<input type="checkbox"/>	<input type="checkbox"/>
Turning_Loop	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>
Sunrise_Sunset	nvarchar(50)	<input type="checkbox"/>	<input checked="" type="checkbox"/>

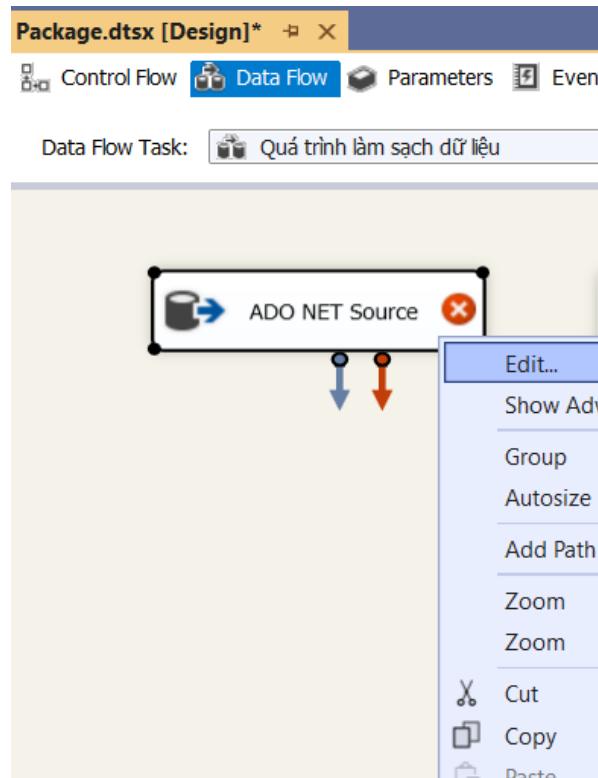


2.2.2. Quá trình làm sạch dữ liệu

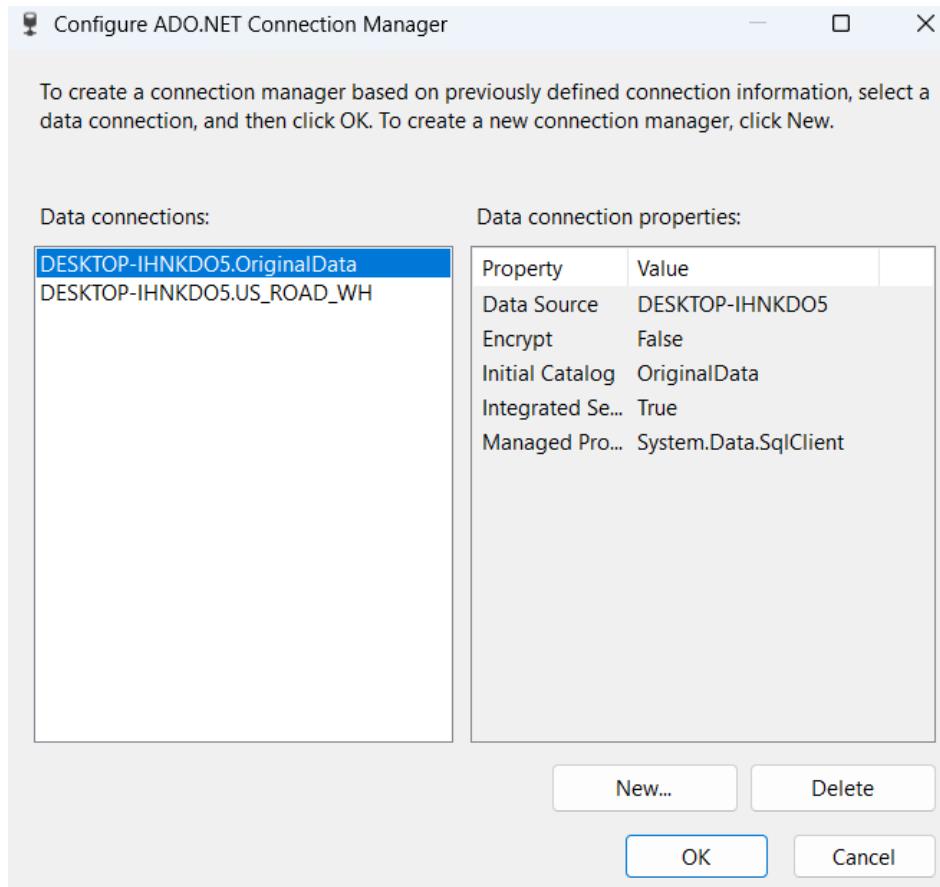
- + Chuyển sang Microsoft Visual Studio
- + Kéo **Data Flow Task** vào màn hình và chỉnh sửa thành “**Quá trình làm sạch dữ liệu**” rồi tiến hành Edit trong Data Flow

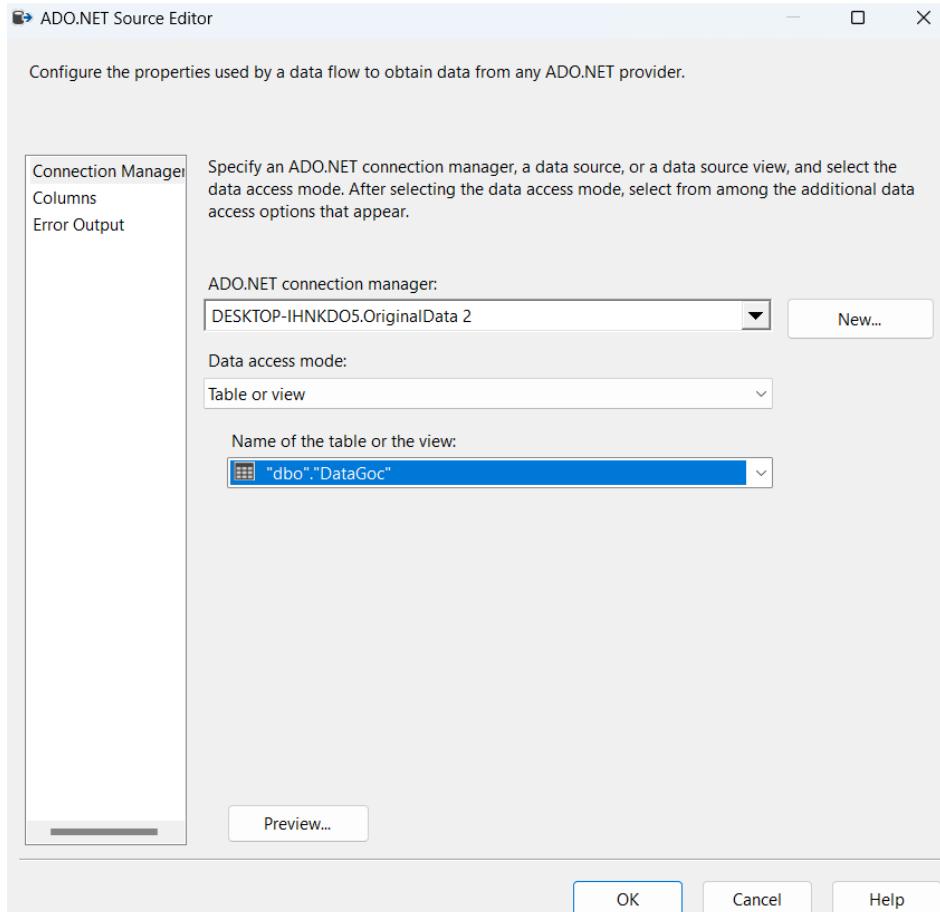


- + Kéo ADO NET Source vào màn hình và chọn **Edit**



- + Sử dụng kết nối tới bảng **DataGoc** trong database **OriginalData**

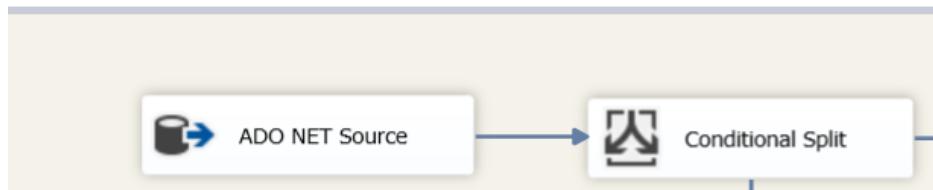




- + Kéo **Conditional Split** vào màn hình rồi nhấn **Edit** để tạo điều kiện tách các hàng chứa giá trị **Null**.

Control Flow Data Flow Parameters Event Handlers Package Explor...

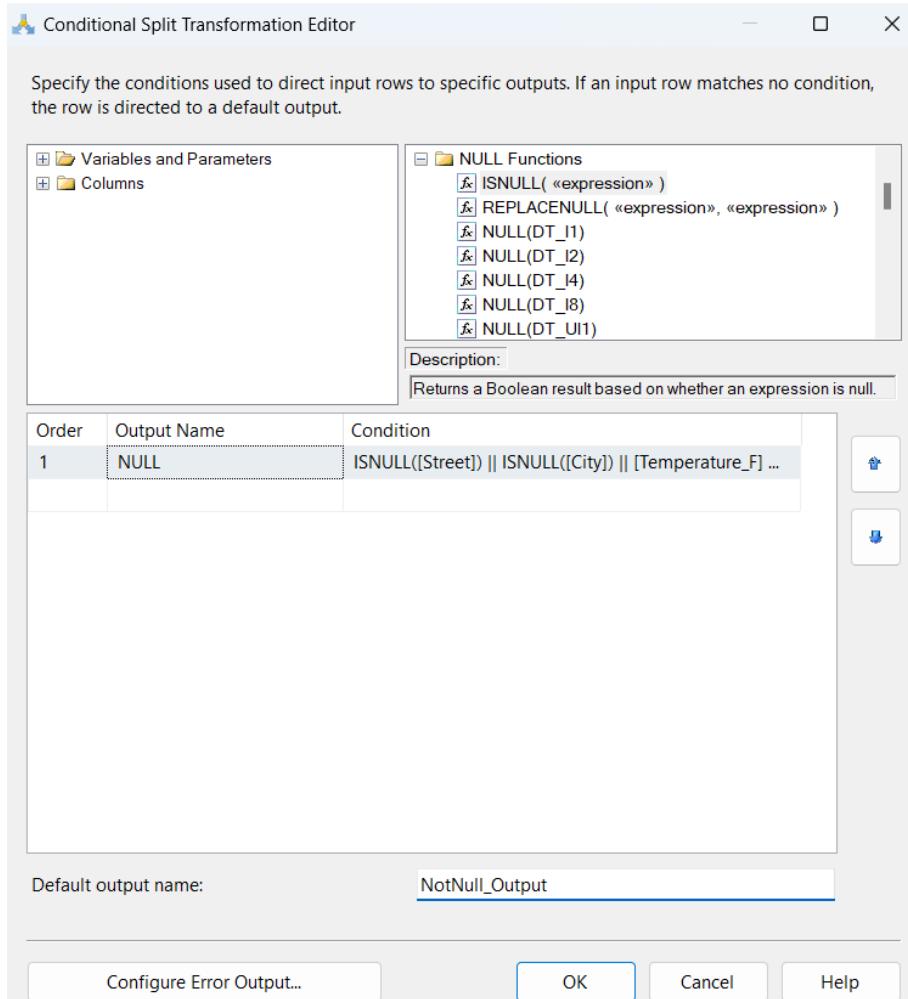
Data Flow Task: Quá trình làm sạch dữ liệu



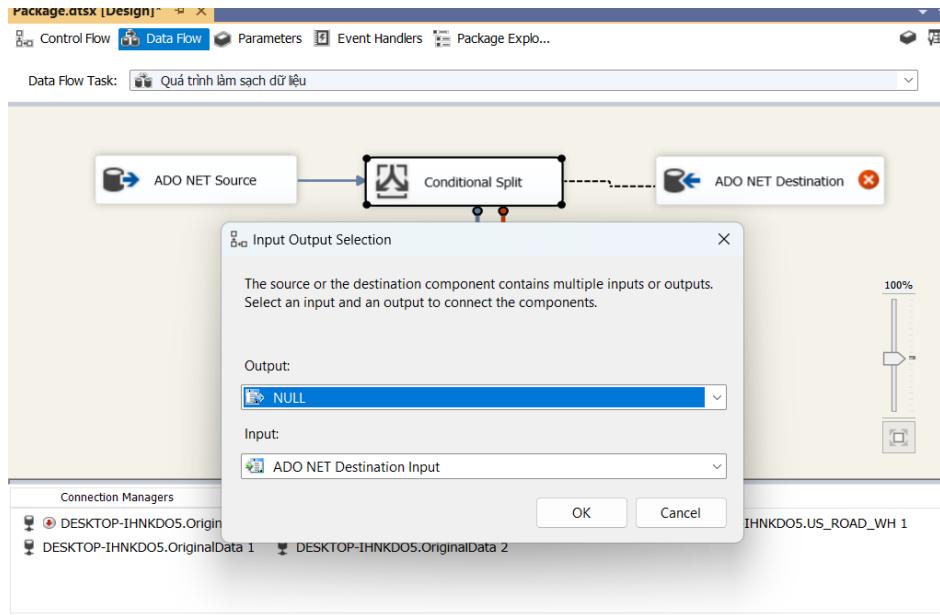
Đoạn mã để lọc giá trị NULL:

```
ISNULL(Street) || ISNULL(City) || ISNULL(Humidity) || ISNULL(Wind_Speed_mph) ||  
ISNULL(Weather_Condition) || REPLACENULL(Distance_mi,0) == 0 ||  
ISNULL(Temperature_F)
```

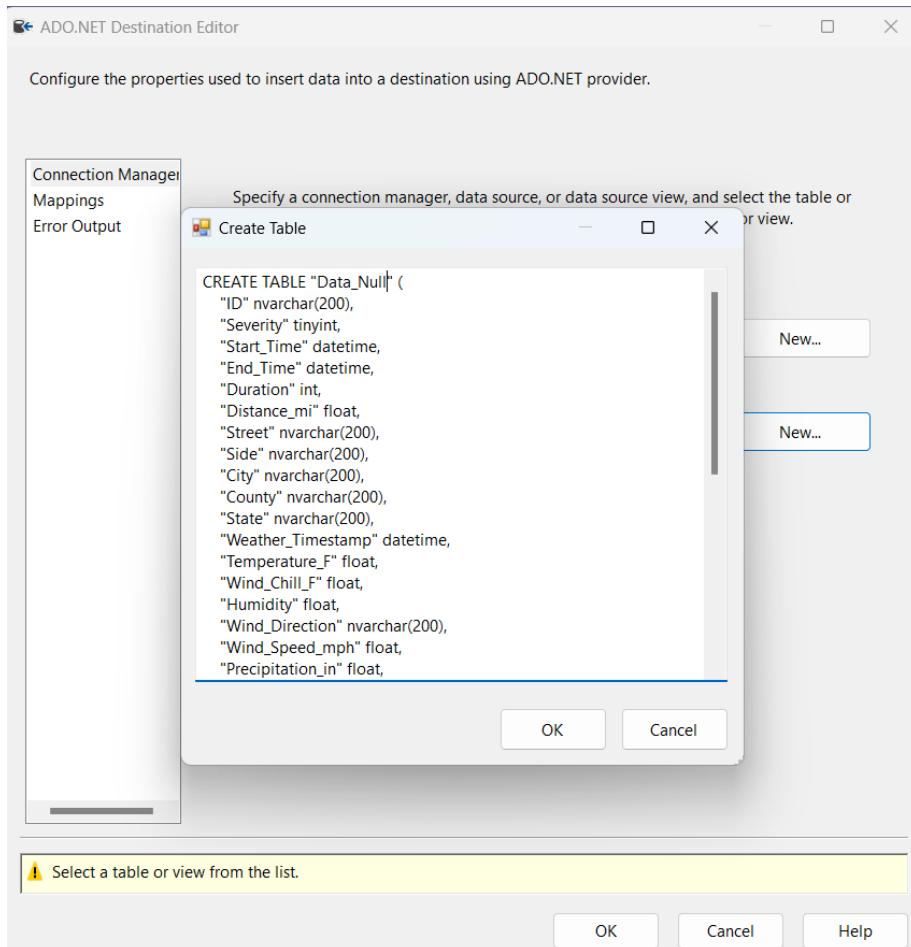
- + Và đặt tên cho output là **NotNull_Output** rồi nhấn **OK**



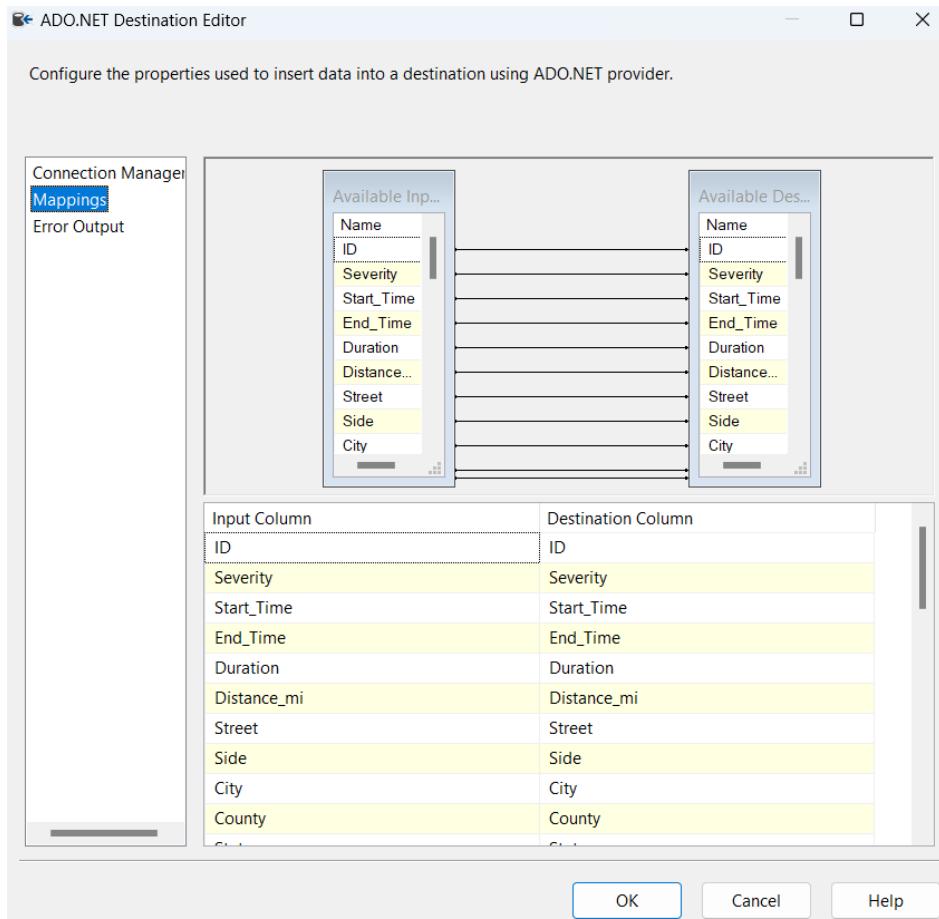
+ Kéo **ADO NET Destination** vào màn hình và chọn **Edit**



- + Kết nối đến database **US_ROAD_WH** và nhán **New..** sau đó chỉnh sửa câu lệnh SQL để tạo bảng **Data_Null**.

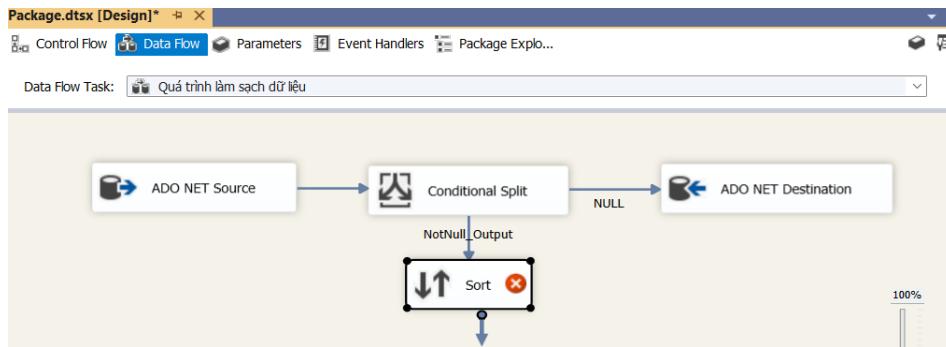


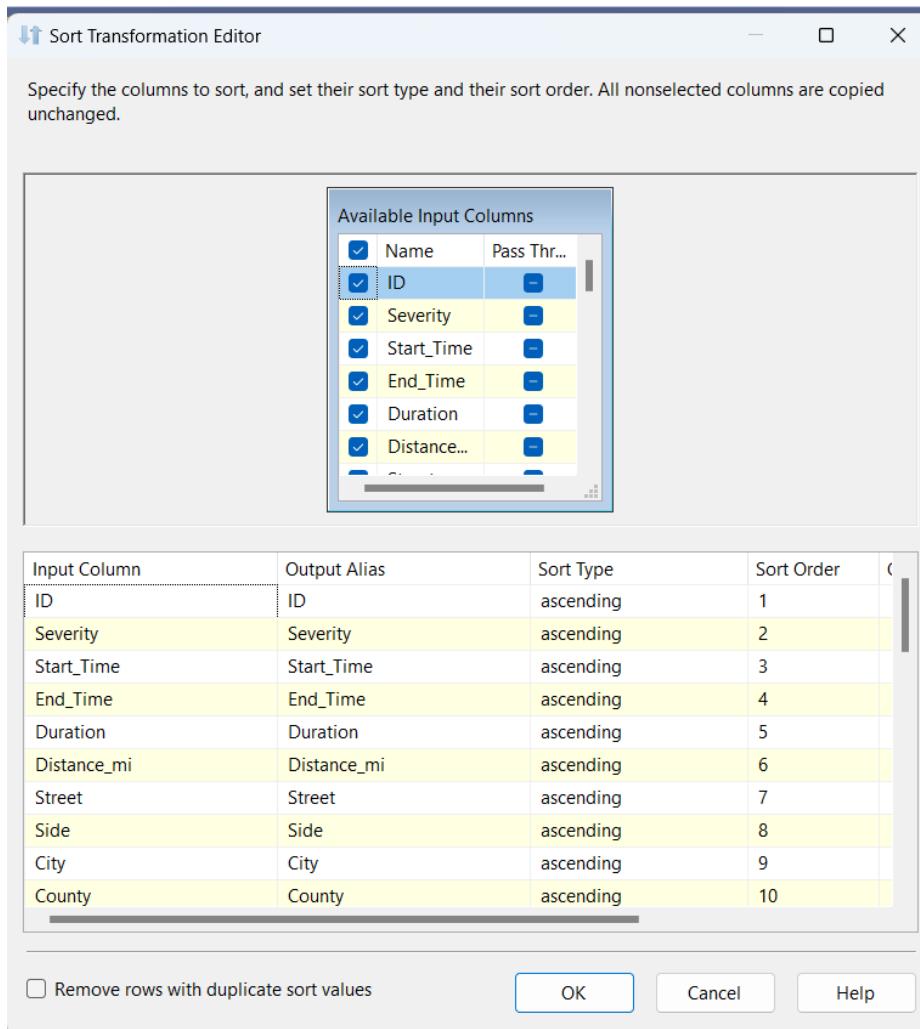
- + Kiểm tra ánh xạ giữa các thuộc tính và nhấn **OK**.



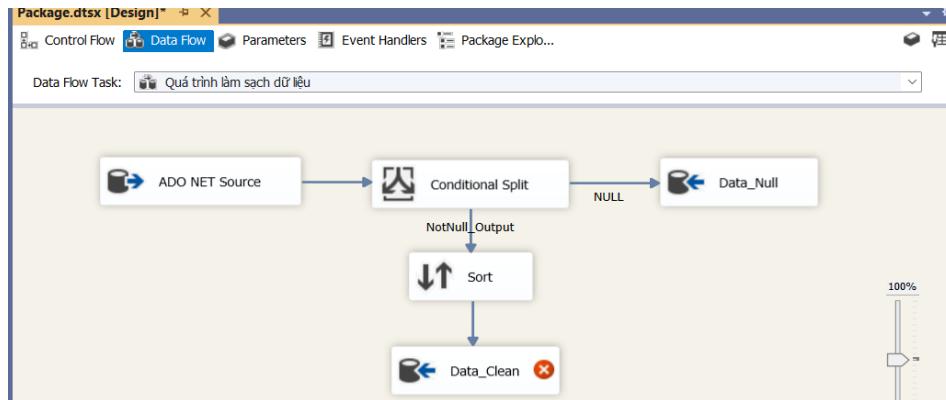
- + Kéo Sort vào màn hình và kéo từ **Conditional Split** đến **Sort** để Sort nhận input **NotNull_Output** từ **Conditional Split** rồi nhấn **Edit**

Ths. Đỗ Thị Minh Phụng

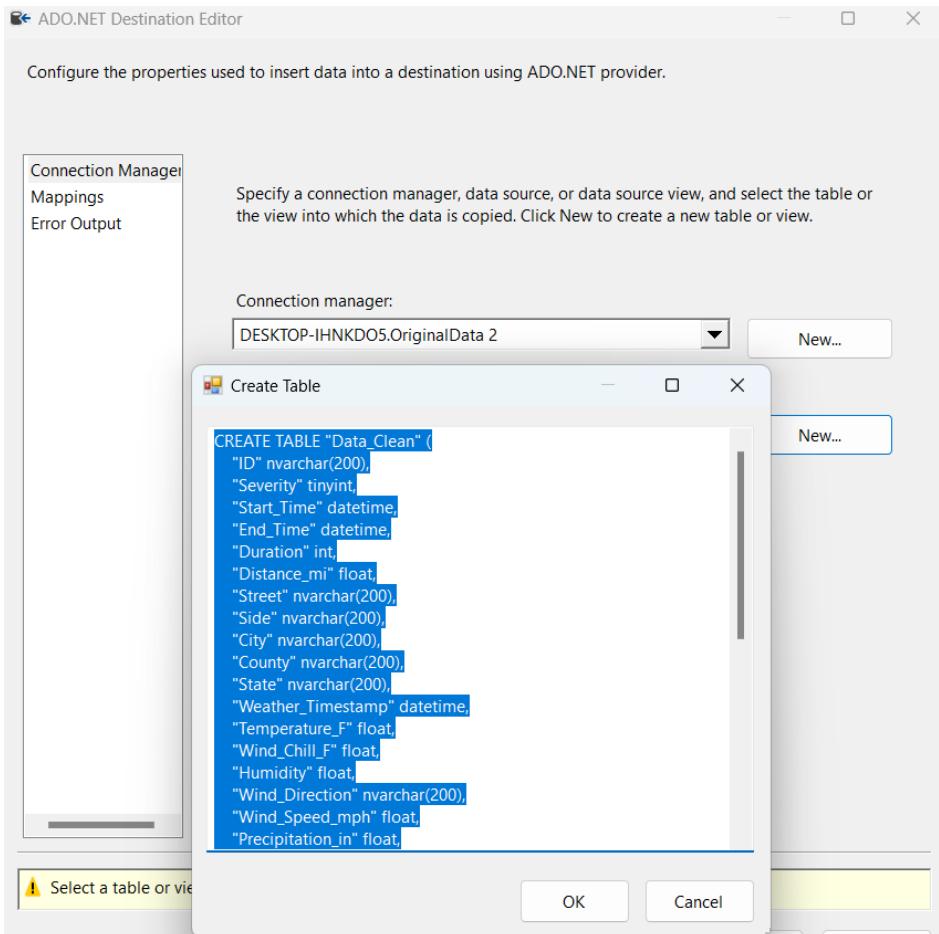




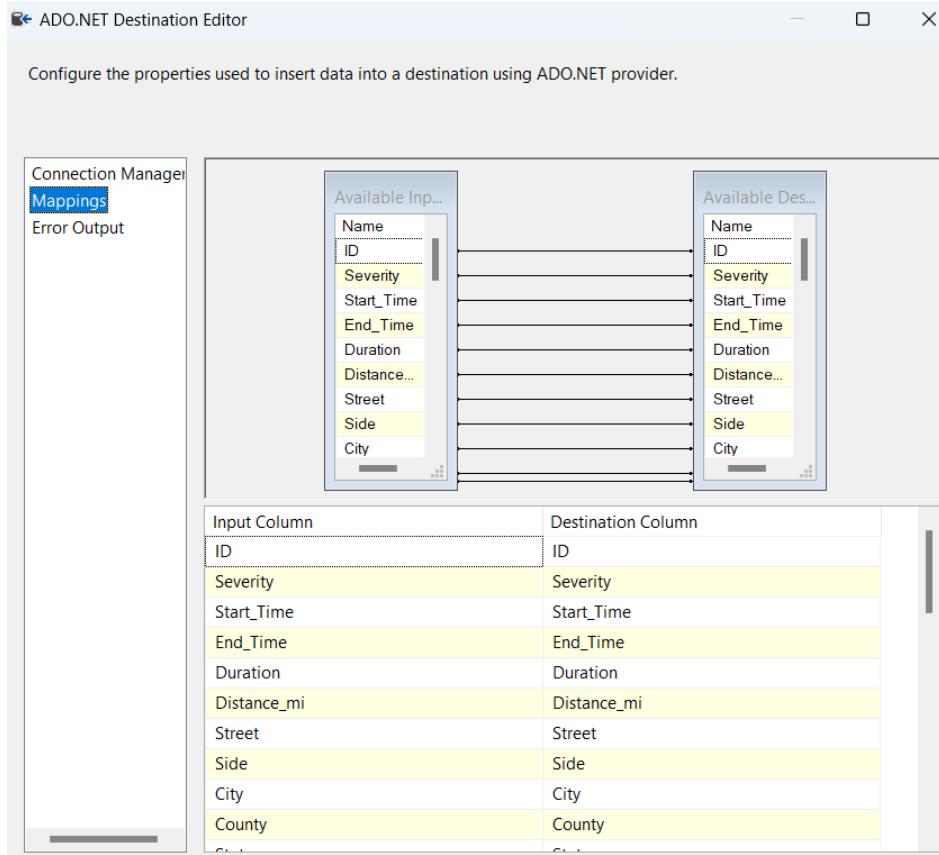
- + Kéo thả để tạo mới một **ADO NET Destination**, đổi tên thành **Data_Clean** rồi nhấn **Edit**



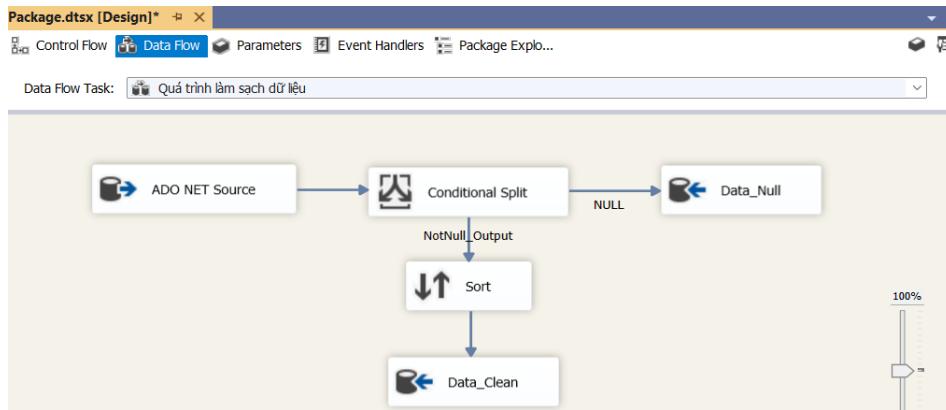
- + Cửa sổ **ADO NET Destination Editor** hiện ra và nhấn **New...** để chỉnh sửa câu lệnh SQL để tạo bảng **Data_Clean**.



- + Kiểm tra ánh xạ các thuộc tính

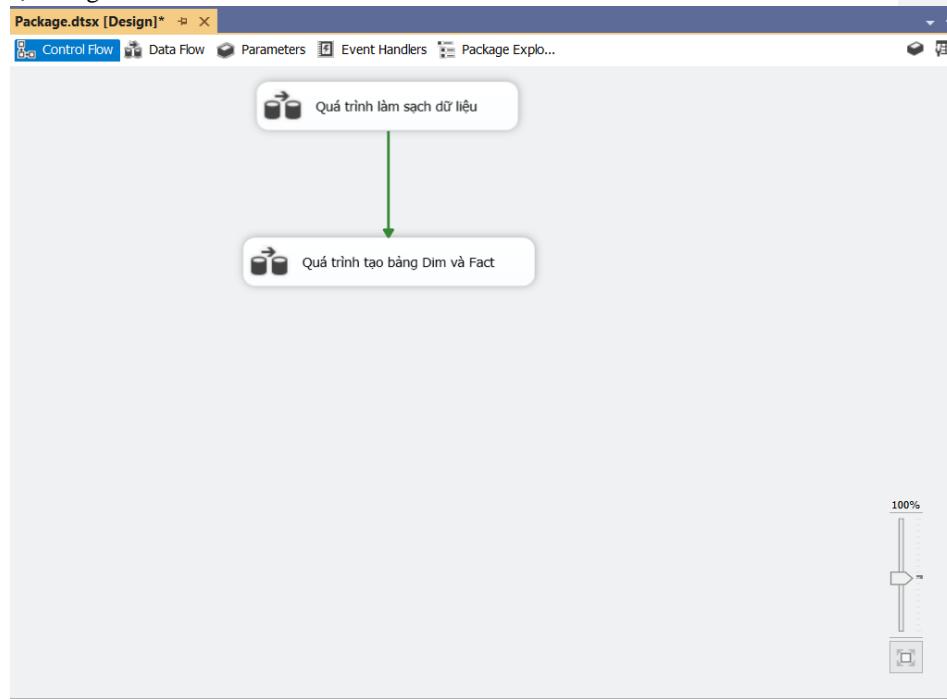


+ **Data Flow** của quá trình làm sạch dữ liệu:

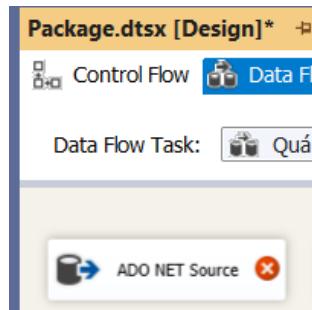


2.3. Quá trình tạo các bảng Dimension và Fact_Raw:

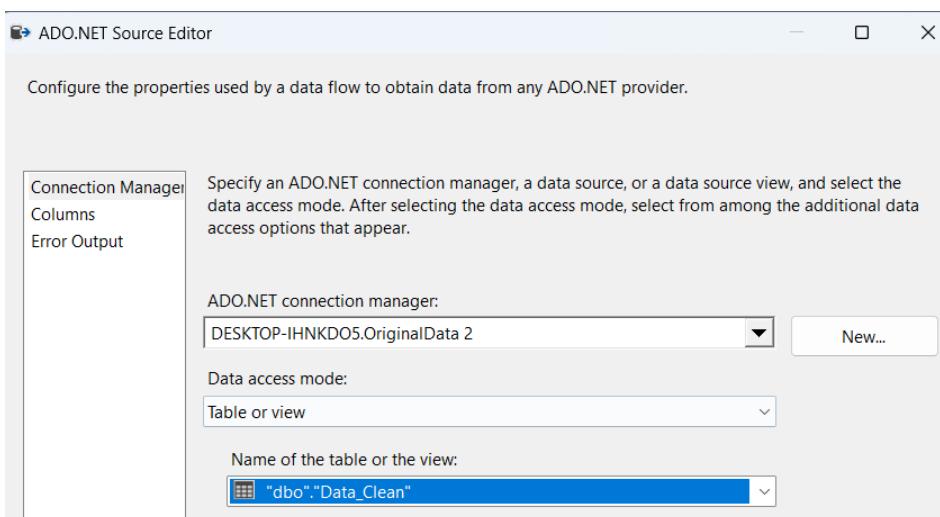
- + Kéo thả và tạo mới **Data Flow Task**. Đặt tên cho **Data Flow Task** đó là Quá trình tạo bảng Dimension và Fact.



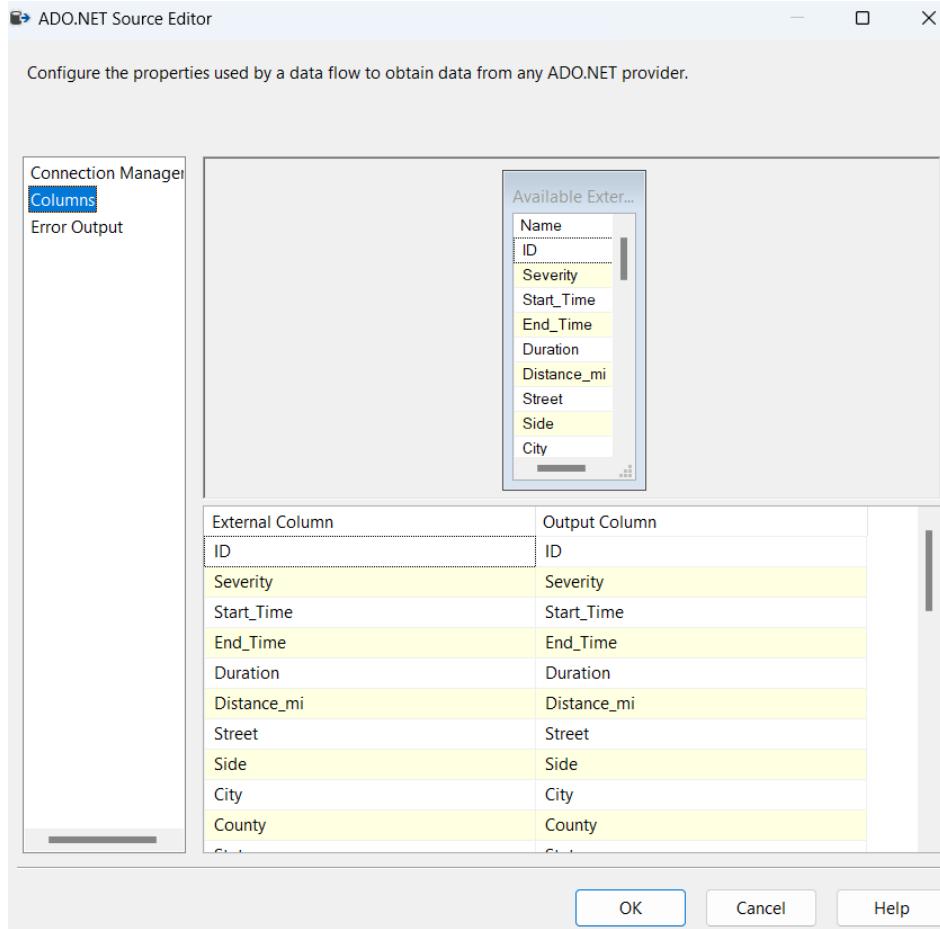
- + Chọn công cụ “ADO NET Source” tại **SSIS Toolbox** vào vùng làm việc **Data Flow Task** mới tạo:



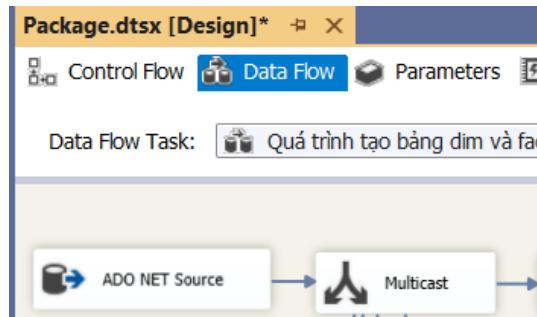
- + Nhấn **Edit** để chọn kết nối tới **database OriginalData** và chọn bảng **CleanData** để làm nguồn dữ liệu



- + Xem lại các cột dữ liệu trong file dữ liệu đã tải lên trong tab **Columns** rồi nhấn **OK**



- + Tạo **Multicast** để phân tán dữ liệu đến các **Dimension** bằng cách kết nối **ADO NET Source** đến **Multicast**

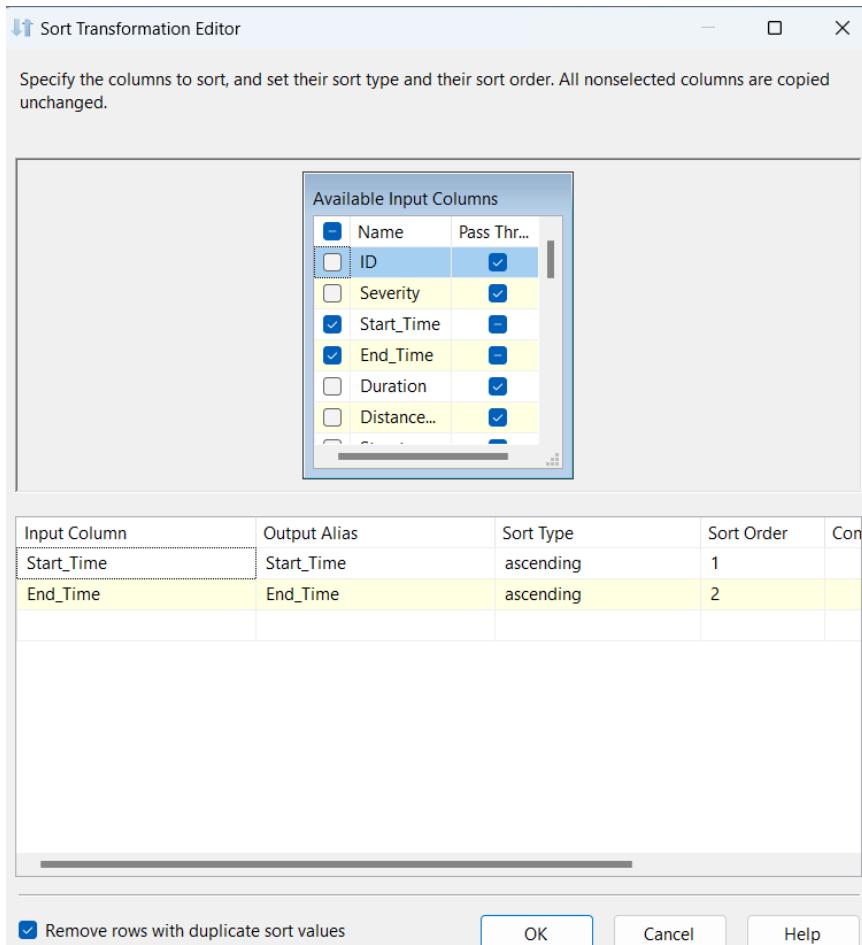


2.3.1. Bảng Dim_Time

- + Kéo Sort vào màn hình để lấy dữ liệu cần thiết cho bảng **Dim_Time**. Nhấn đúp chuột vào Sort để **Edit**



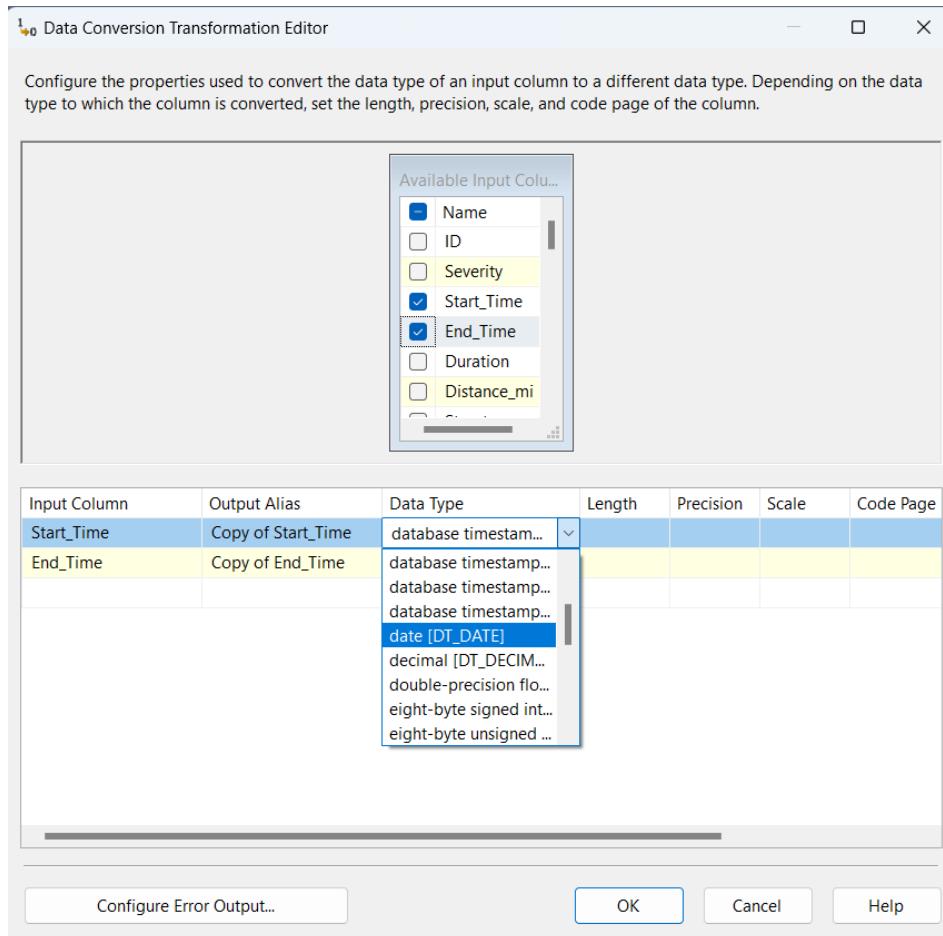
- + Chọn **Start_Time** và **End_Time** để làm thuộc tính cho bảng **Dim_Time**
- + Tick chọn **Remove rows with duplicate sort values** để xóa các dòng dữ liệu trùng lặp rồi nhấn **OK**.

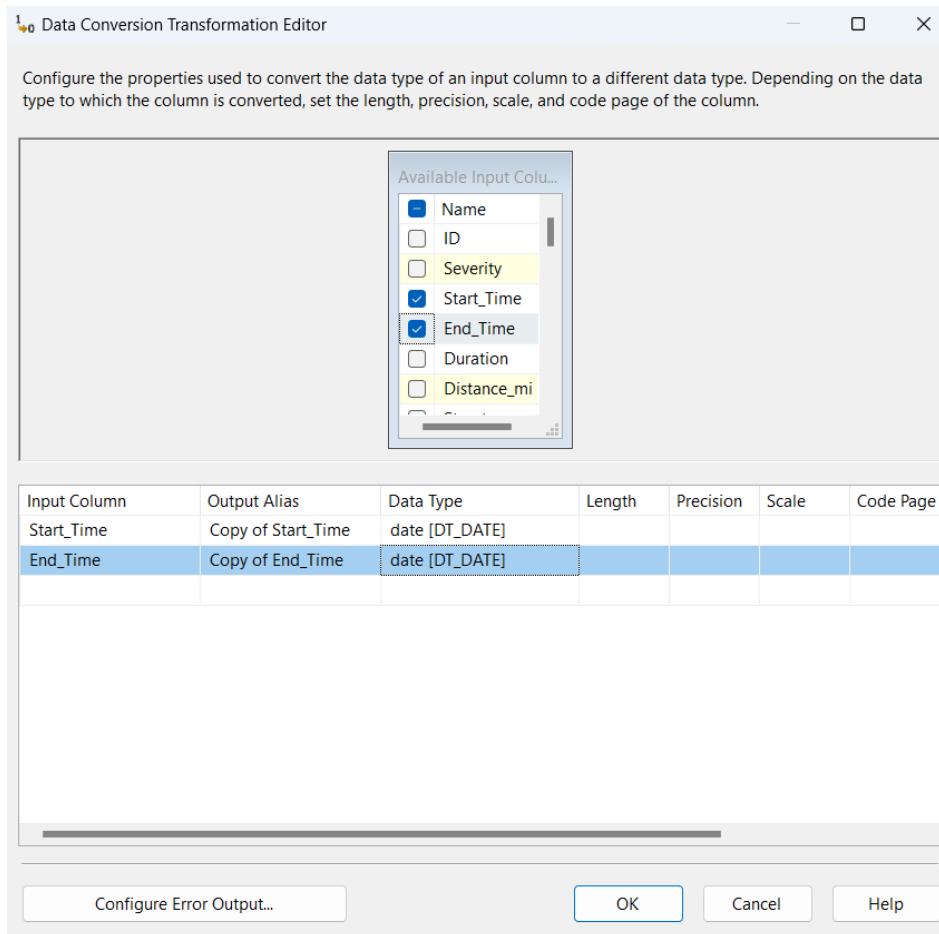


- + Chuyển đổi kiểu dữ liệu từ **String** sang **Date Time** vì kiểu **DateTime** từ file dữ liệu gốc được mặc định là **String**
- + Đổi tên cho **Data Conversion** thành **Change time from string to datetime**
- + Click chuột phải vào **DataConversion** và chọn **Edit**

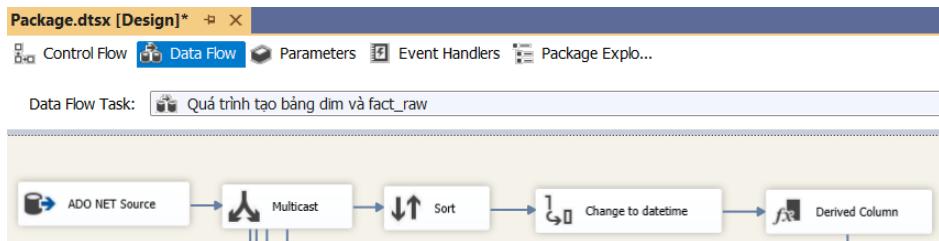


- + Click chọn cột **Start_Time**. Trong cột **Data Type** đổi sang kiểu date [**DT_DATE**] và làm tương tự đối với cột **End_Time**

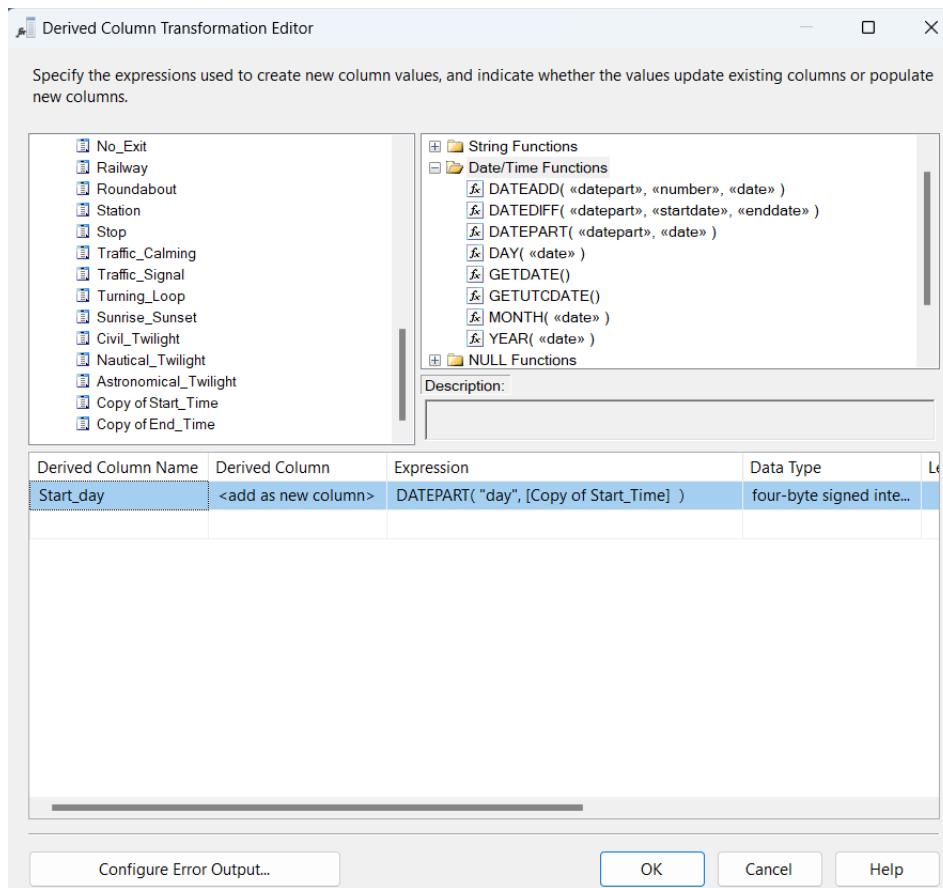


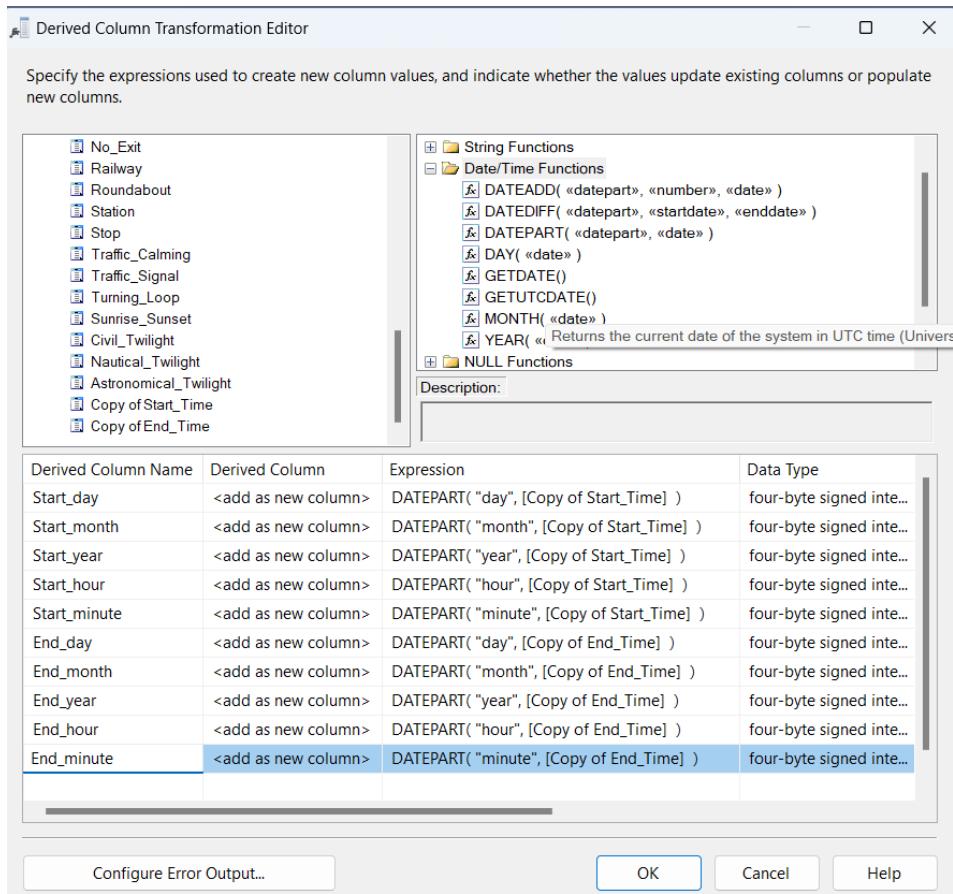


+ Thêm **Derived Column** vào màn hình rồi chọn **Edit** để tiến hành chia cột dữ liệu

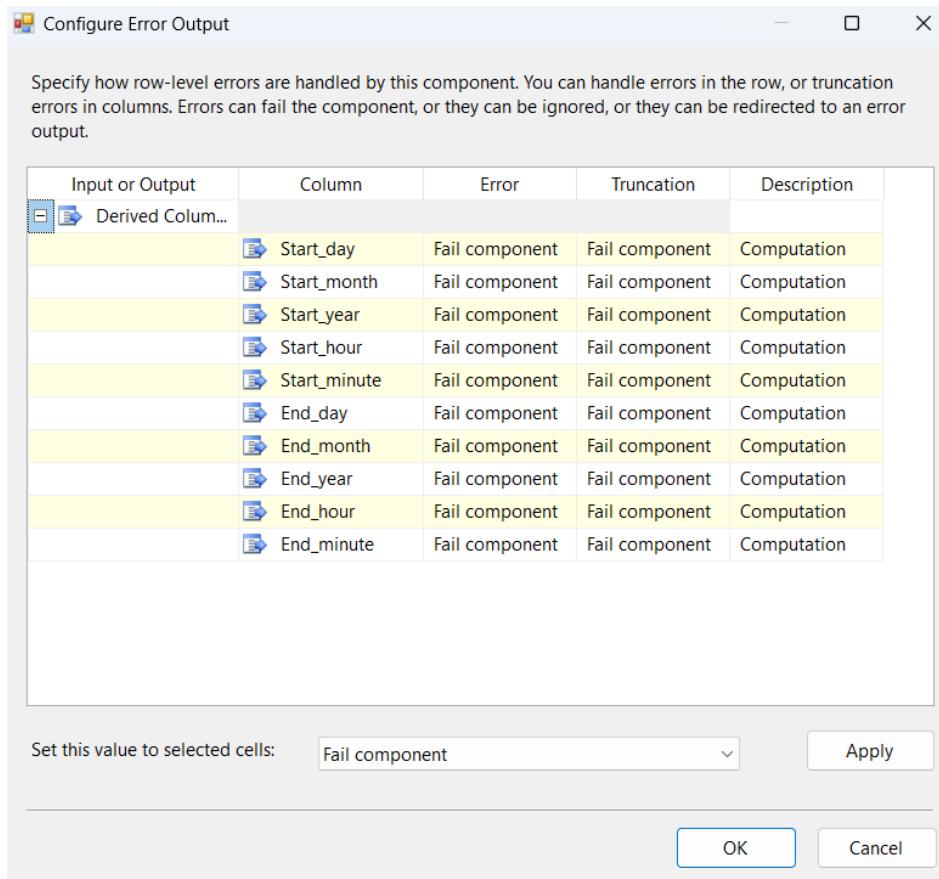


- + Chia dữ liệu từ cột **Copy of Start_Time** và cột **Copy of End_Time** thành các cột **Start_day**, **Start_month**, **Start_year**, **Start_hour**, **Start_minute**, **End_day**, **End_month**, **End_year**, **End_hour**, **End_minute**.
- + Chọn phương thức **DATEPART(<kiểu dữ liệu thời gian>, cột dữ liệu)**. và tiến hành chia như hình bên dưới



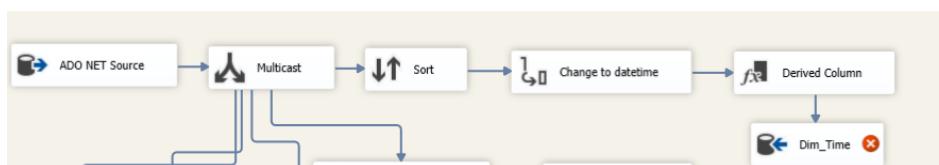


- + Nhấn **Configure Error Output...**
- + Cửa sổ **Configure Error Output** hiện ra và kiểm tra thấy 2 cột **Copy of Start_Time** và **Copy of End_Time** được chia thành các thuộc tính tương ứng với bảng **Dim_Time**.
- + Nhấn **OK** để hoàn tất chia cột.



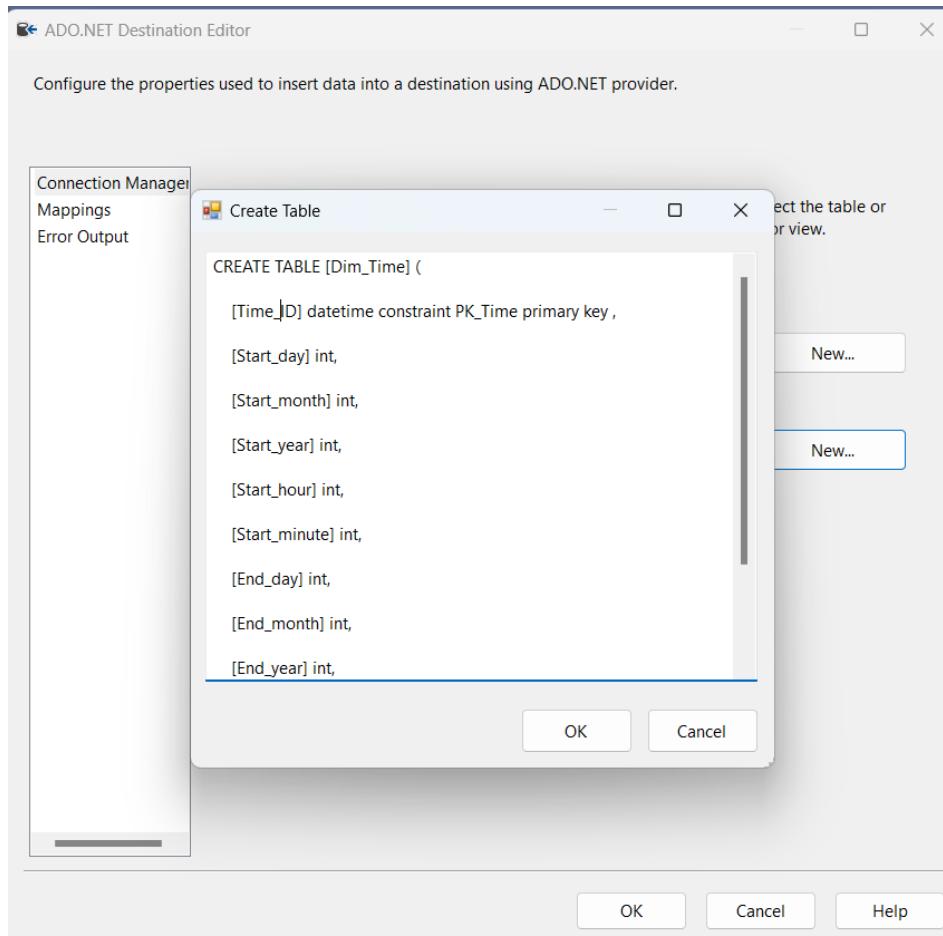
+ Tạo **Dim_Time** từ **ADO NET Destination**.

+ Click chuột phải và chọn **Edit**

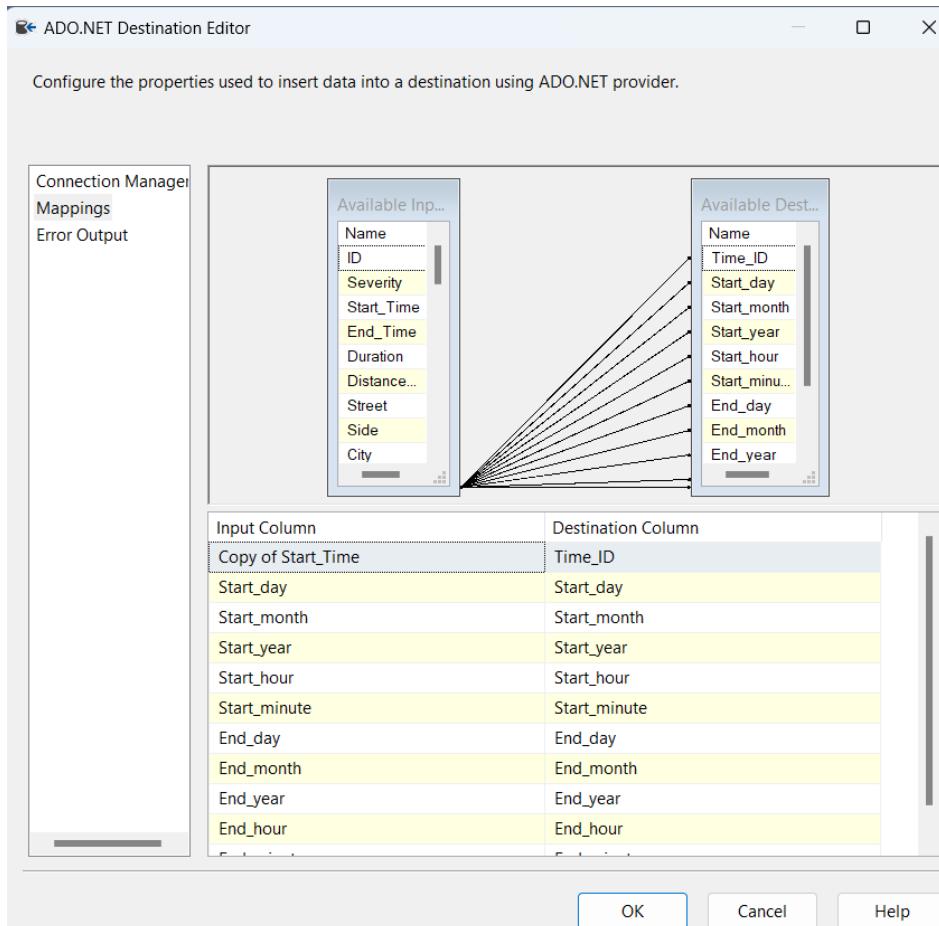


+ Dán đoạn mã SQL như bên dưới vào để tạo bảng **Dim_Time**

```
CREATE TABLE [Dim_Time] (
    [Time_ID] datetime constraint PK_Time primary key ,
    [Start_day] int,
    [Start_month] int,
    [Start_year] int,
    [Start_hour] int,
    [Start_minute] int,
    [End_day] int,
    [End_month] int,
    [End_year] int,
    [End_hour] int,
    [End_minute] int
)
```

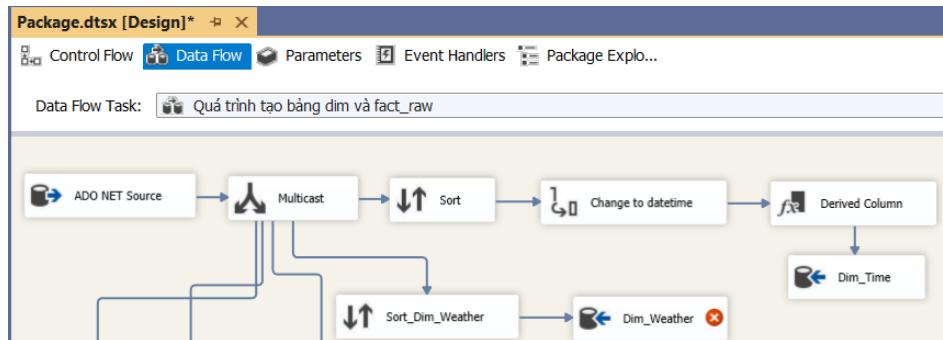


- + Kiểm tra ánh xạ của các thuộc tính trong thẻ **Mapping** rồi nhấn **OK**

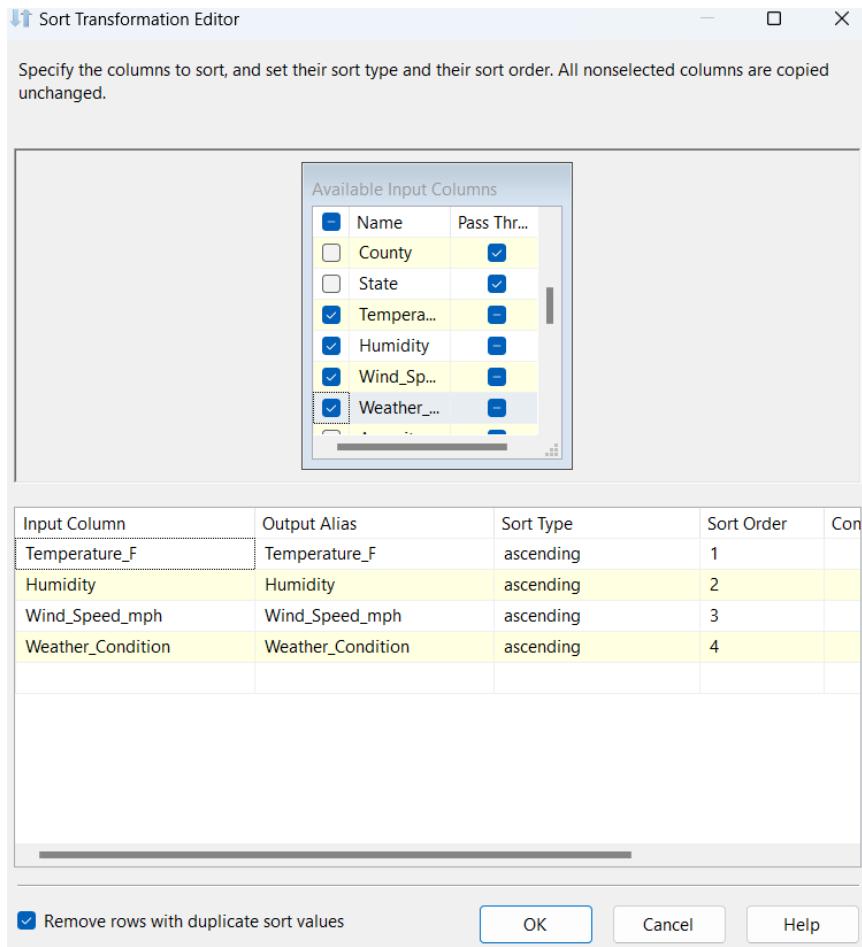


2.3.2. Bảng Dim_Weather

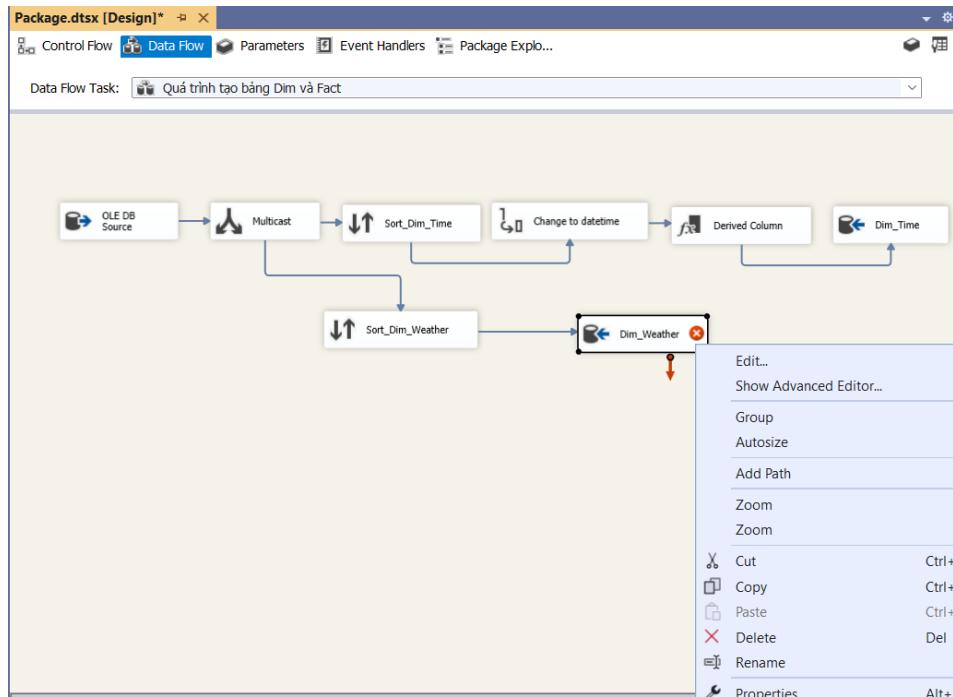
- + Tạo một thành phần Sort để tạo Sort cho Sort_Dim_Weather cho Dim_Weather



- + Double click vào **Sort_Dim_Weather** và chọn các cột **Temperature_F**, **Humidity**, **Wind_Speed_mph**, **Weather_Condition** để đỡ dữ liệu vào **Sort_Dim_Weather**
- + Chọn **Remove rows with duplicate sort values** để loại bỏ các dòng dữ liệu bị trùng lặp. Sau đó nhấn **OK**



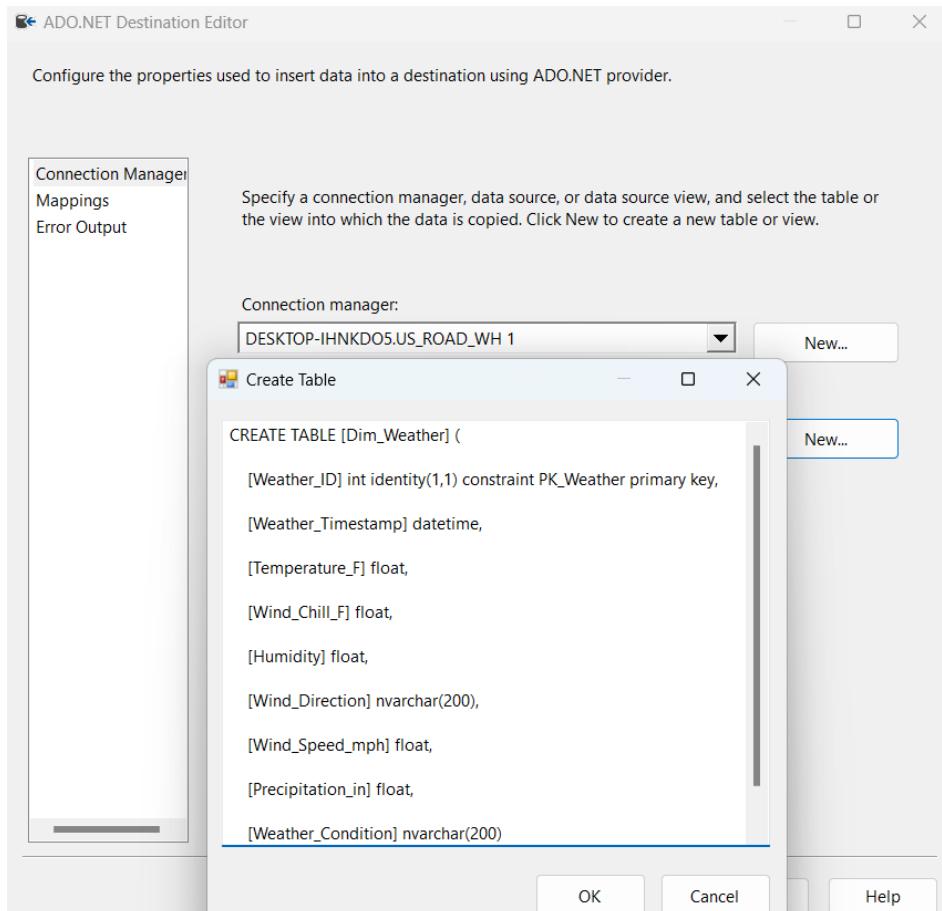
- + Tạo một **ADO NET Destination** để đổ dữ liệu gốc sau khi được xử lý vào trong kho dữ liệu **US_ROAD_WH**. Click chuột phải chọn **Edit**



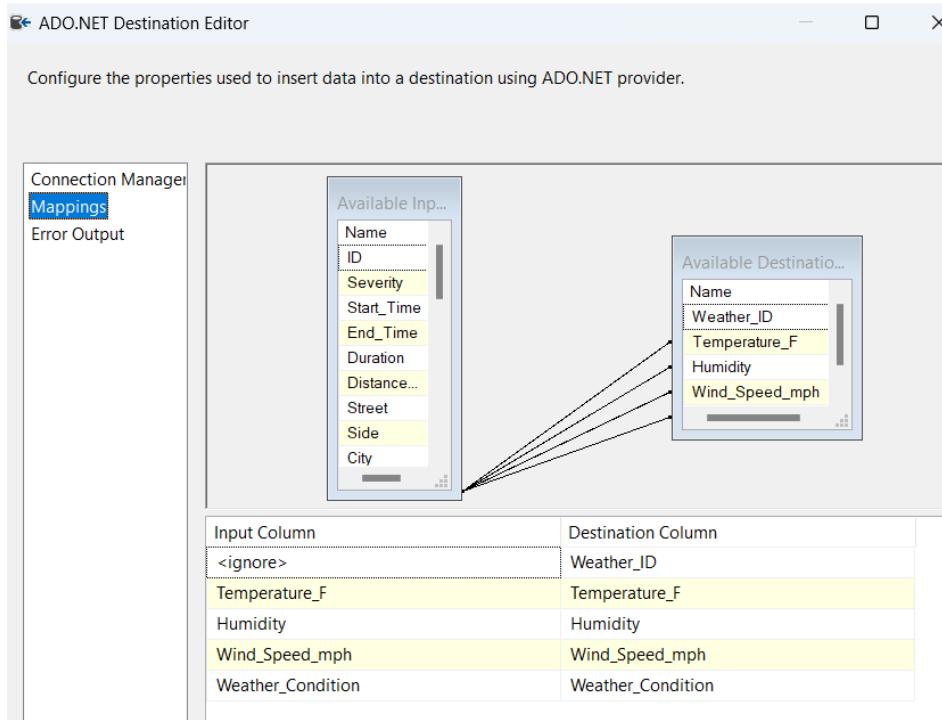
- + Thực hiện tương tự như bảng **Dim_Time** nhưng thay đổi đoạn mã SQL như bên dưới để tạo bảng **Dim_Weather**.

```

CREATE TABLE [Dim_Weather] (
    [Weather_ID] int identity(1,1) constraint PK_Weather primary
key,
    [Temperature_F] float,
    [Humidity] float,
    [Wind_Speed_mph] float,
    [Weather_Condition] nvarchar(200)
)
  
```

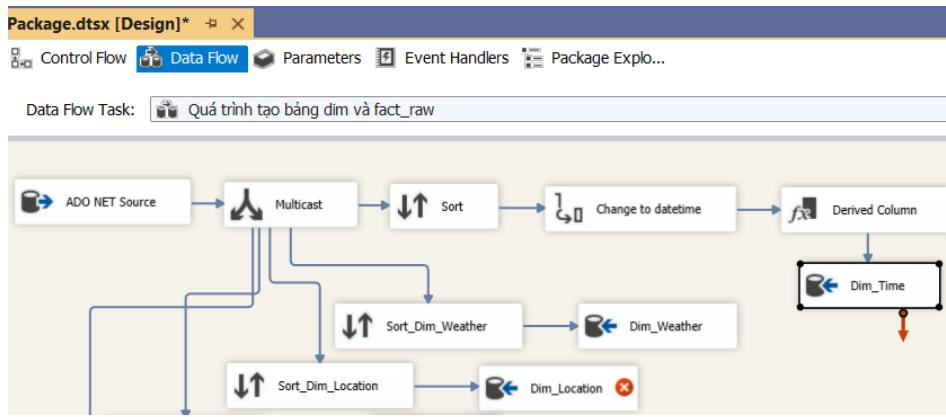


- + Kiểm tra **Mapping** trong tab **Mappings** để kiểm tra ánh xạ giữa các cột dữ liệu rồi nhấn **OK**

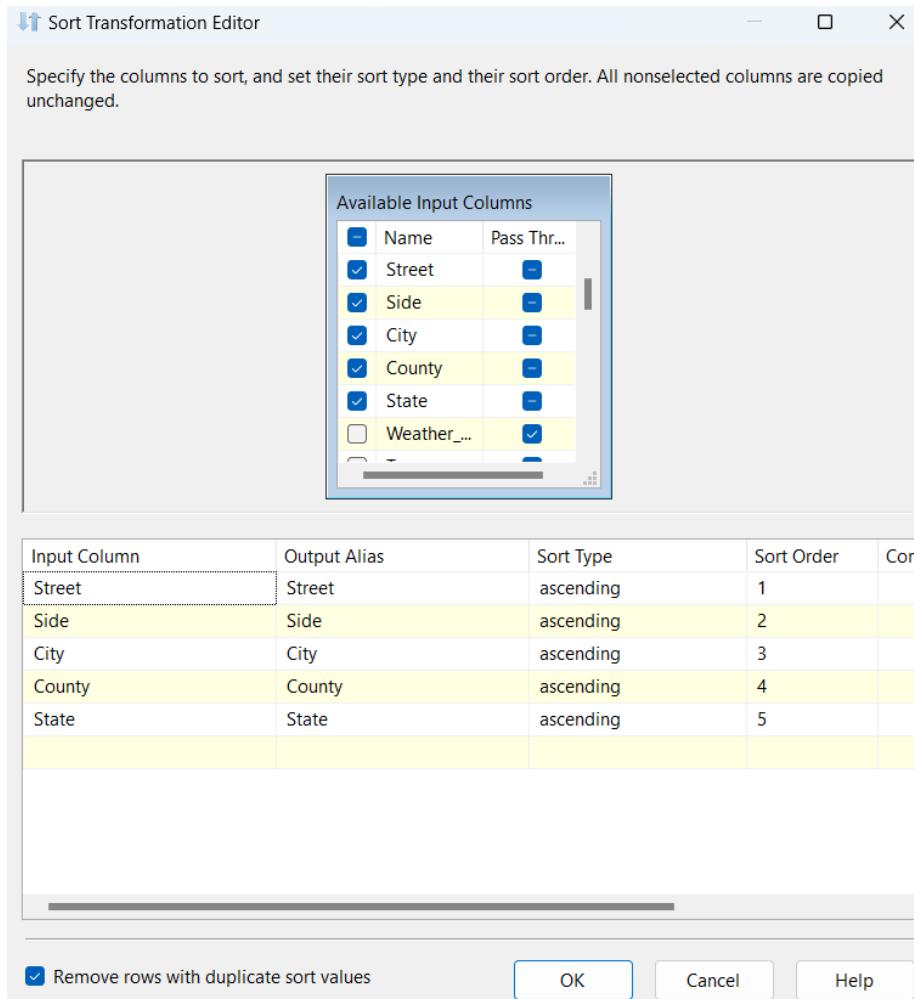


2.3.3. Bảng Dim_Location

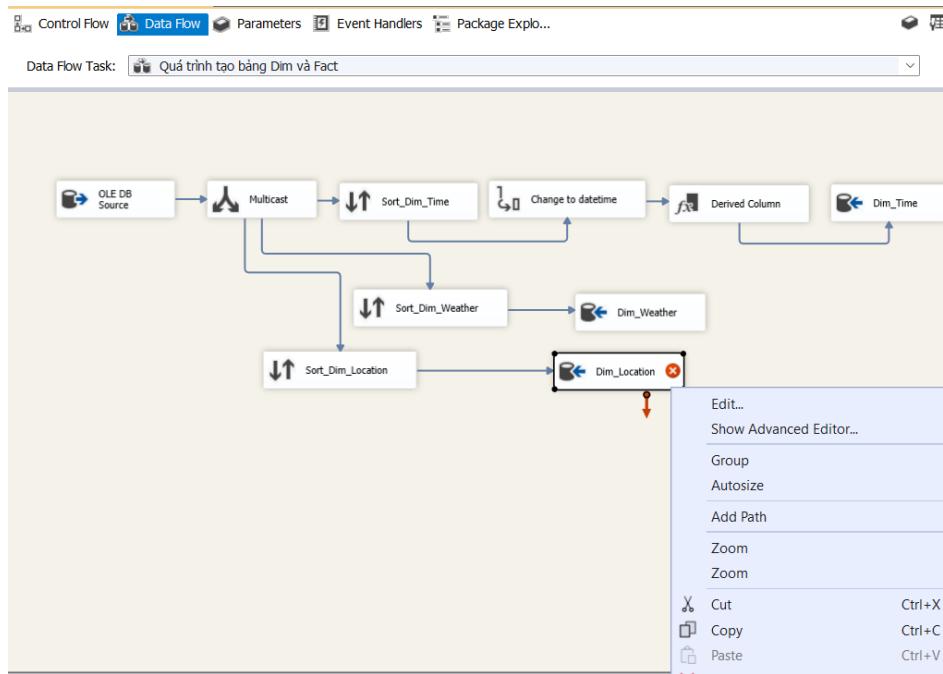
- + Tạo mới Sort cho Sort_Dim_Location để thực hiện việc tạo Dim_Location.
- + Chuột phải vào Sort chọn Edit



- + Chọn các cột thuộc tính: **Street, Side, City, County, State** để đồ dữ liệu vào **Sort_Dim_Location**.
- + Tick chọn **Remove rows with duplicate sort values** để xóa đi các dòng dữ liệu bị trùng lặp và sau đó nhấn **OK**

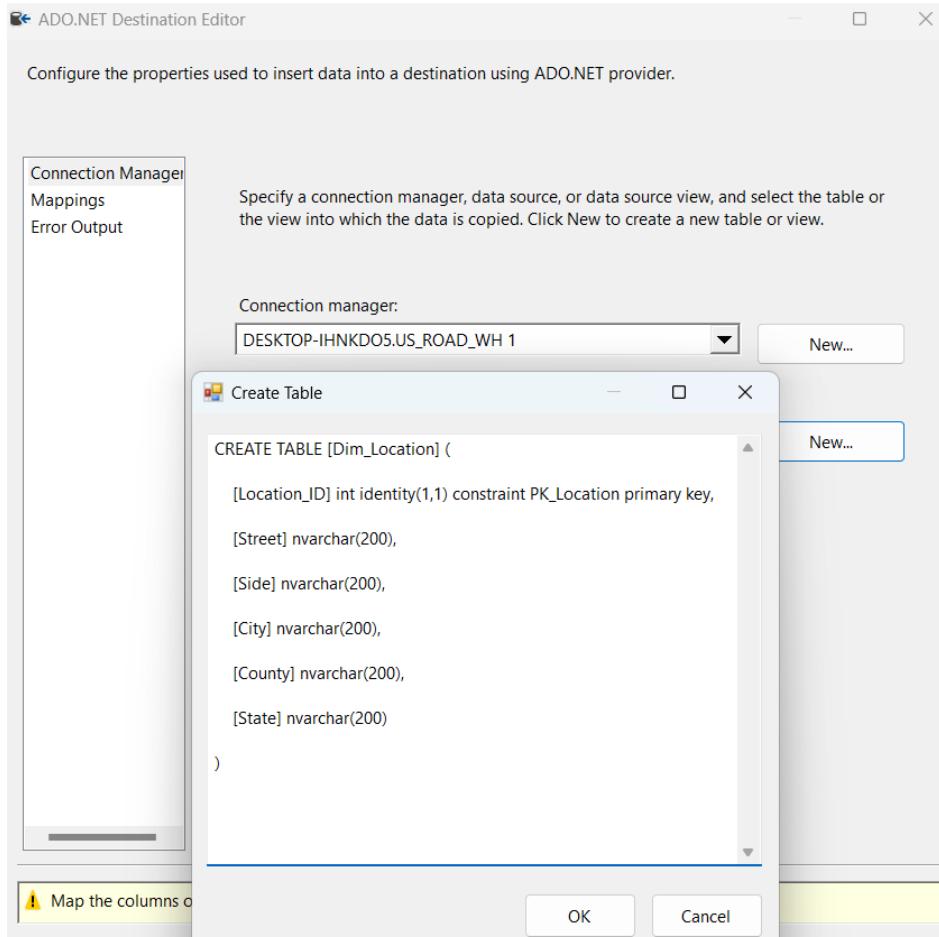


- + Tạo một **ADO NET Destination** để đổ dữ liệu gốc sau khi được xử lý vào trong kho dữ liệu **US_ROAD_WH**. Click chuột phải chọn **Edit**

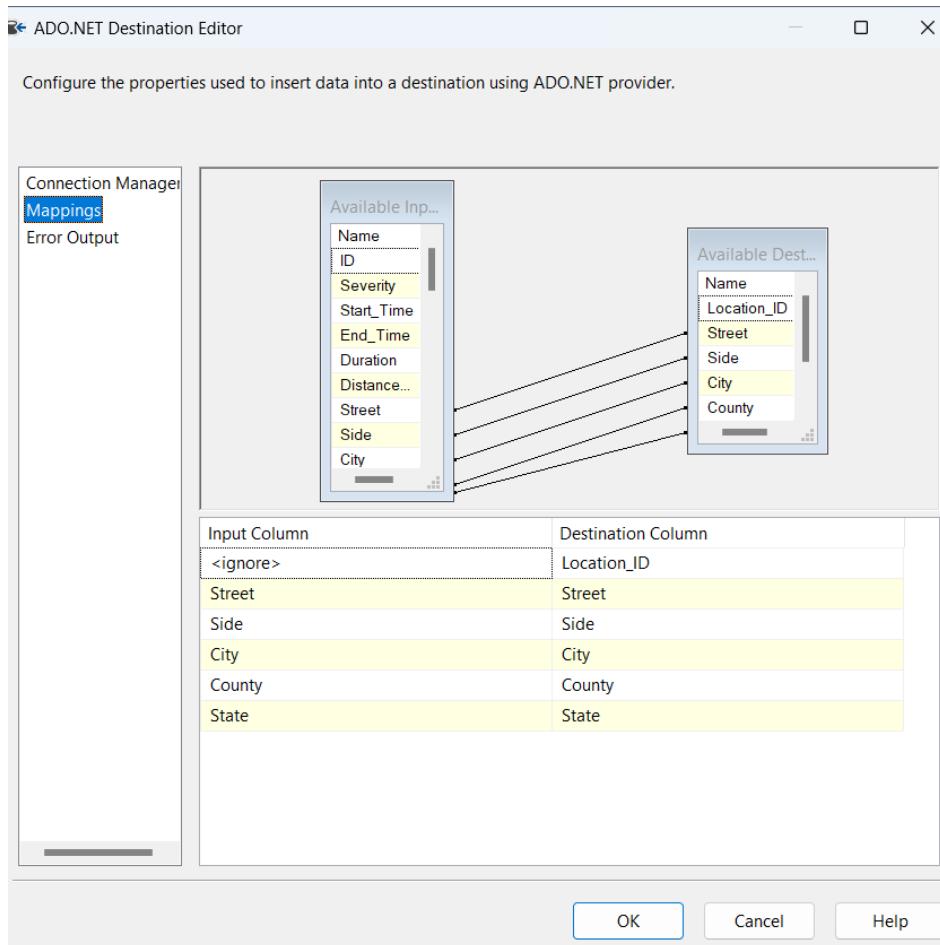


- + Thực hiện tương tự như bảng **Dim_Weather** nhưng thay đổi đoạn mã SQL như sau để tạo ra bảng **Dim_Location**, Sau đó nhấn **OK**

```
CREATE TABLE [Dim_Location] (
    [Location_ID] int identity(1,1) constraint PK_Location
primary key,
    [Street] nvarchar(200),
    [Side] nvarchar(200),
    [City] nvarchar(200),
    [County] nvarchar(200),
    [State] nvarchar(200)
)
```

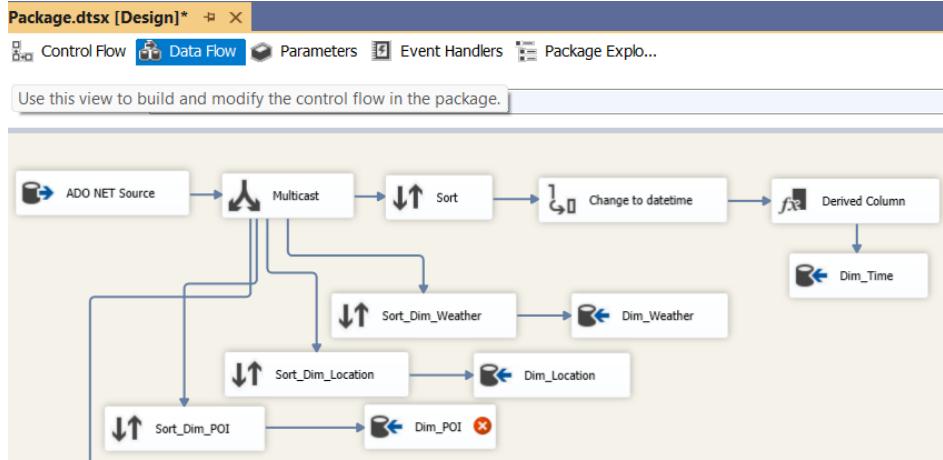


- + Kiểm tra việc ánh xạ các cột thuộc tính trong thẻ **Mappings** và nhấn **OK**

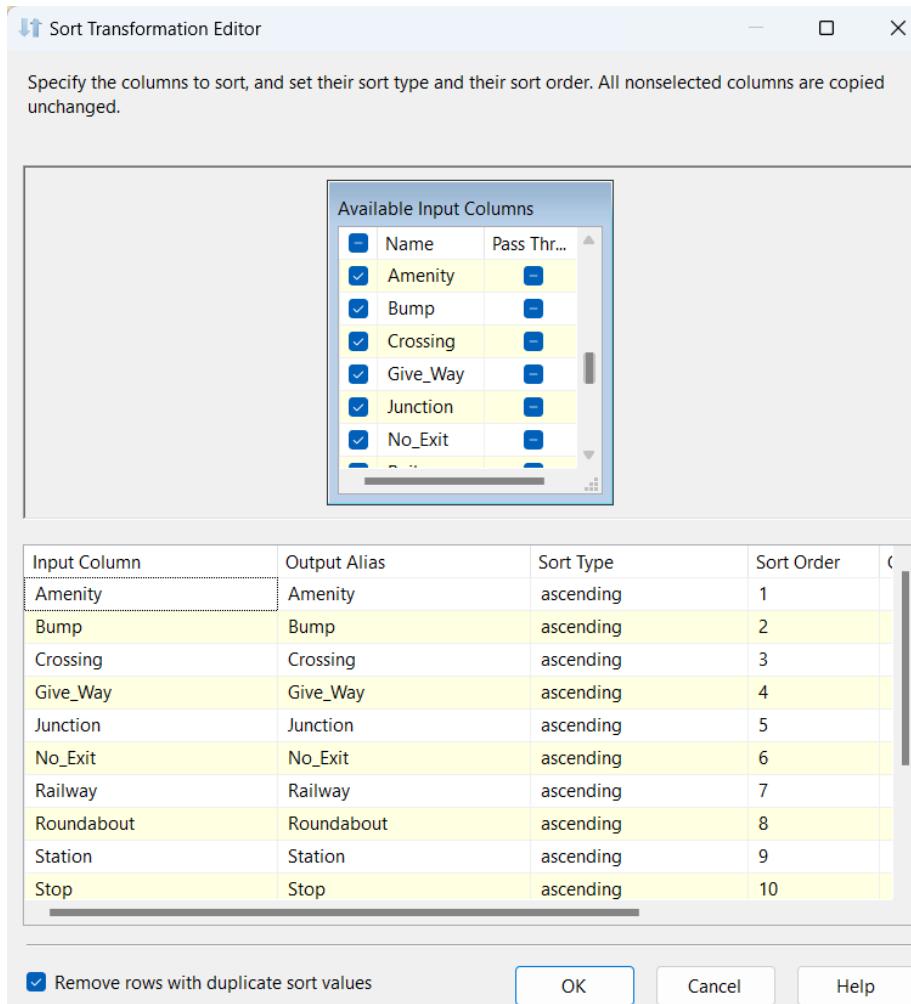


2.3.4. Bảng Dim_POI

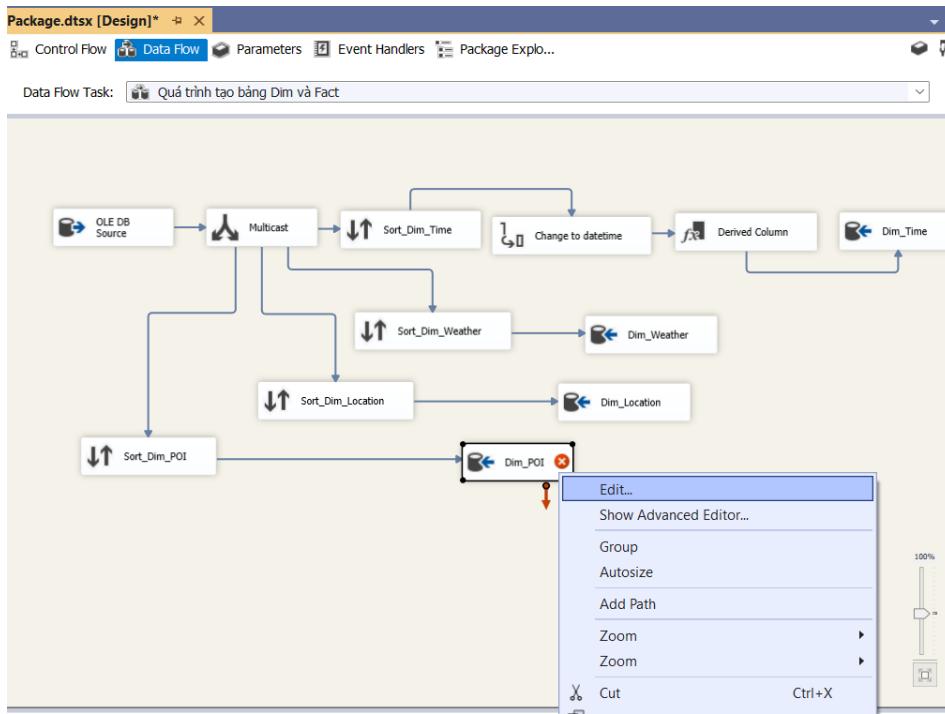
- + Tạo một thành phần Sort mới cho Sort_Dim_POI để tạo bảng Dim_POI
- + Click chuột phải vào Sort chọn Edit



- + Chọn lần lượt các cột **Amenity, Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, Turning_Loop** làm các cột để đồ dữ liệu vào **Sort_Dim_POI**
- + Nhấn chọn **Remove rows with duplicate sort values** để xóa đi các dòng dữ liệu bị trùng lặp và nhấn **OK**.



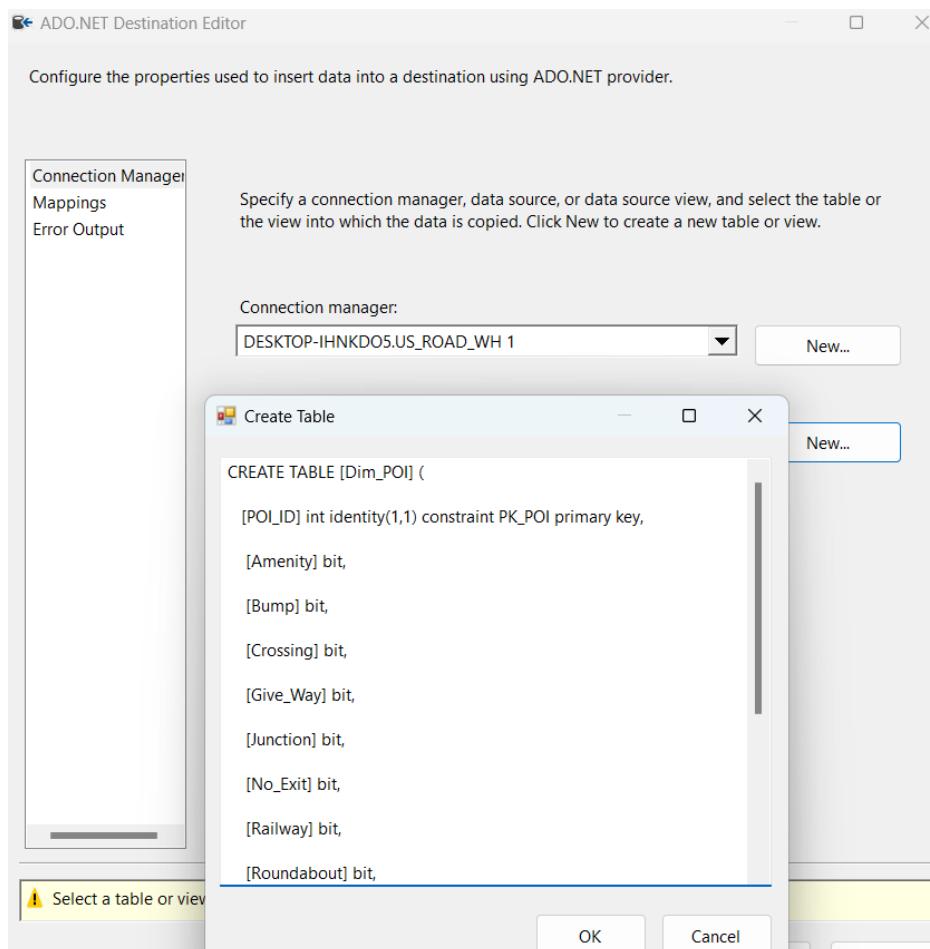
- + Kéo thả và tạo mới một **ADO NET Destination** và đổi tên thành **Dim_POI** để đỗ dữ liệu sau khi xử lý vào kho dữ liệu.
- + Click chuột phải và chọn **Edit**.



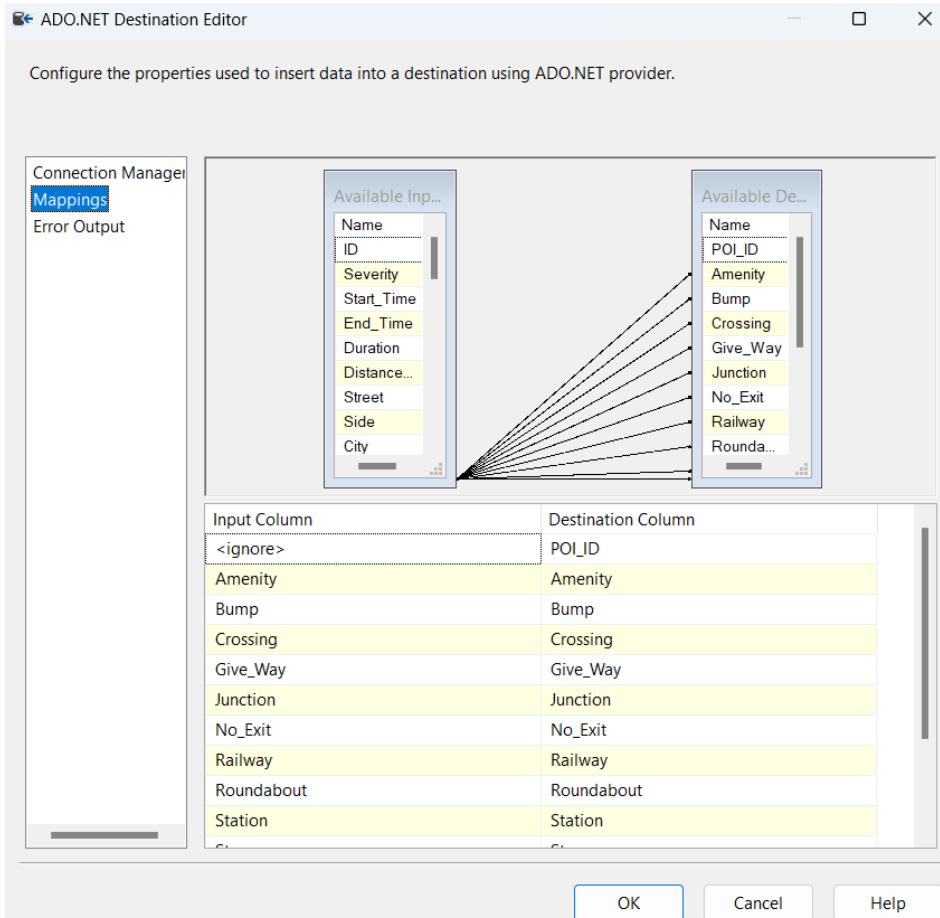
- + Thực hiện tương tự như bảng **Dim_Location** nhưng thay đổi đoạn mã SQL như sau để tạo ra bảng **Dim_POI**, Sau đó nhấn **OK**
- + Nội dung của câu lệnh SQL như sau:

```
CREATE TABLE [Dim_POI] (
    [POI_ID] int identity(1,1) constraint PK_POI primary key,
    [Amenity] bit,
    [Bump] bit,
    [Crossing] bit,
    [Give_Way] bit,
    [Junction] bit,
    [No_Exit] bit,
    [Railway] bit,
```

```
[Roundabout] bit,  
[Station] bit,  
[Stop] bit,  
[Traffic_Calming] bit,  
[Traffic_Signal] bit,  
[Turning_Loop] nvarchar(50)  
)
```

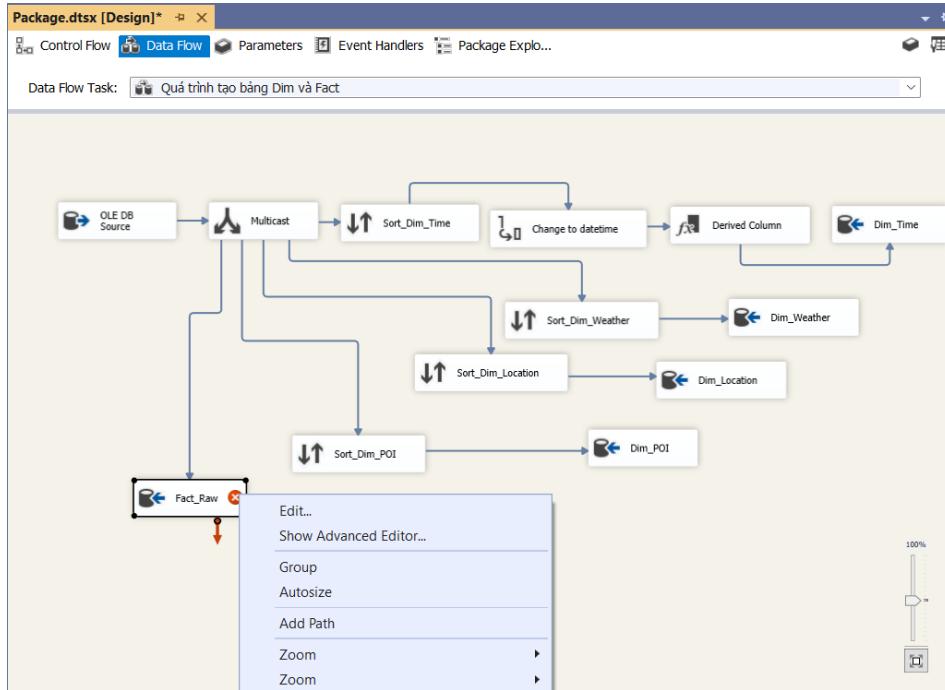


- + Kiểm tra sự ánh xạ của các cột trong thẻ **Mappings** rồi nhấn **OK**



2.3.5. Bảng Fact

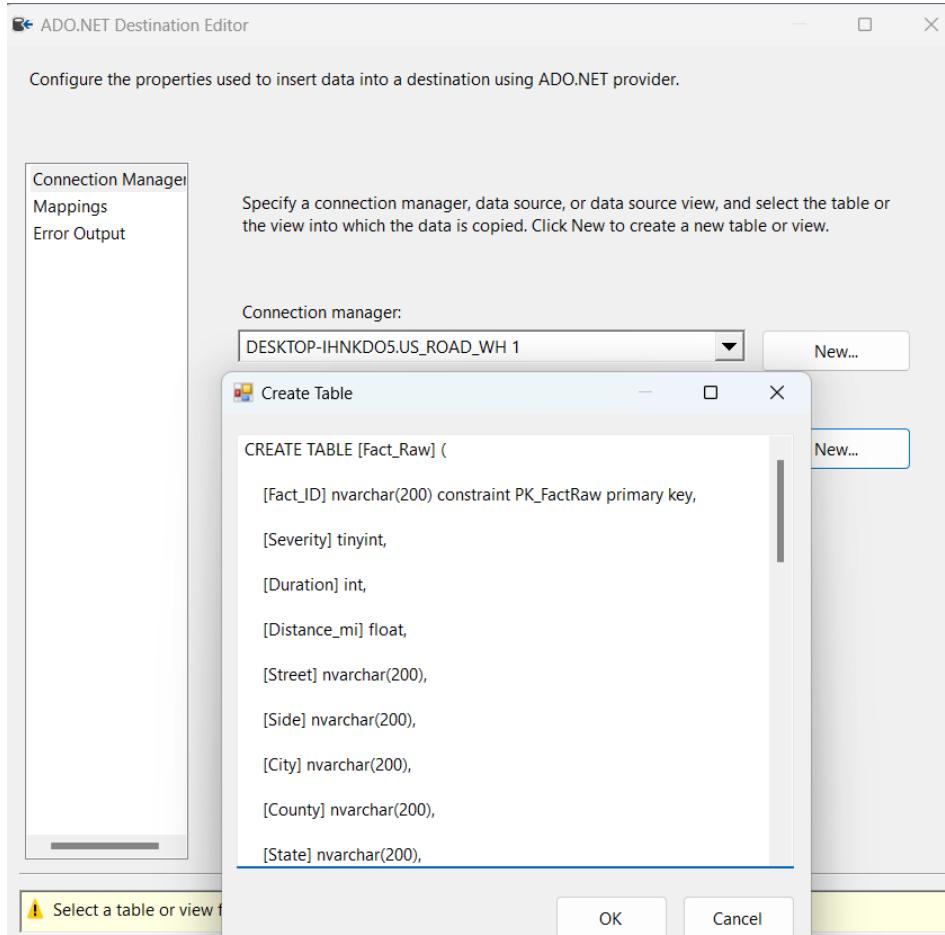
- + Tạo bảng **Fact** bằng cách kéo thả **ADO NET Destination** vào màn hình và đổi tên thành **Fact_Raw**. Sau đó click chuột phải và chọn **Edit**.



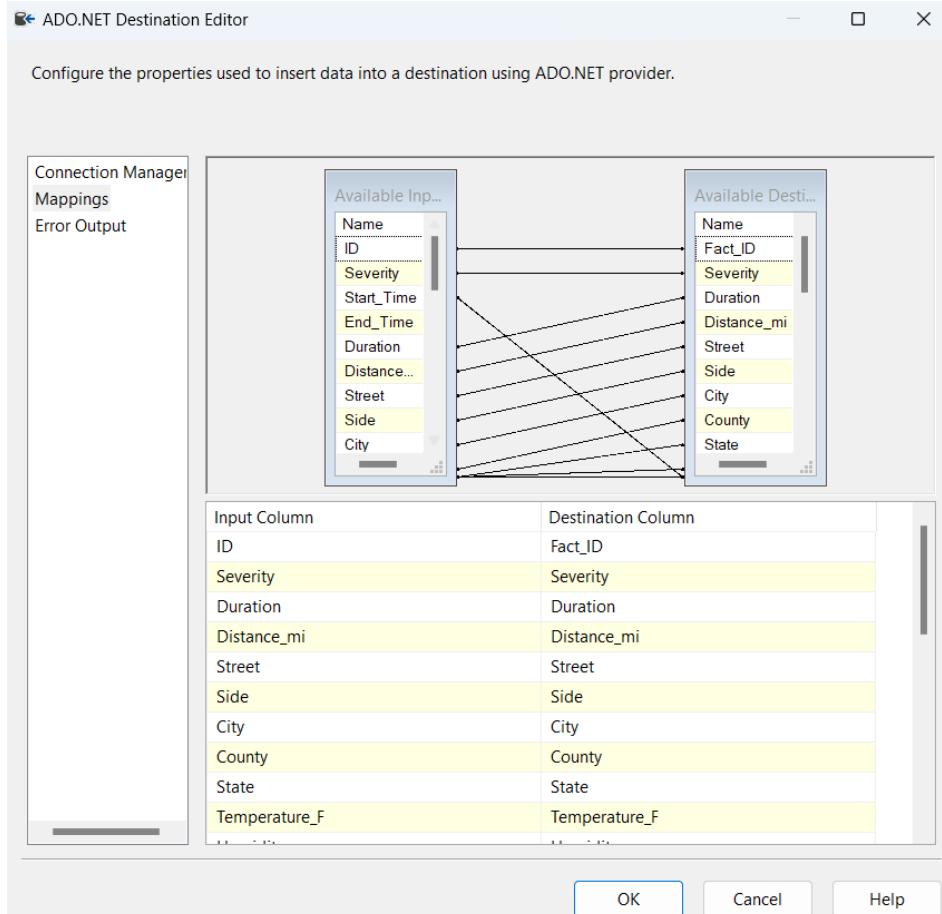
- + Các cột của bảng **Fact_Raw** là các cột từ dữ liệu gốc chứa tất cả các dòng dữ liệu.
- + Nội dung đoạn mã SQL như sau:

```
CREATE TABLE "Fact_Raw" (
    [Fact_ID] nvarchar(200) constraint PK_FactRaw primary key,
    "Severity" tinyint,
    "Duration" int,
    "Distance_mi" float,
    "Street" nvarchar(200),
    "Side" nvarchar(200),
    "City" nvarchar(200),
    "County" nvarchar(200),
    "State" nvarchar(200),
    "Temperature_F" float,
```

```
"Humidity" float,  
"Wind_Speed_mph" float,  
"Weather_Condition" nvarchar(200),  
"Amenity" bit,  
"Bump" bit,  
"Crossing" bit,  
"Give_Way" bit,  
"Junction" bit,  
"No_Exit" bit,  
"Railway" bit,  
"Roundabout" bit,  
"Station" bit,  
"Stop" bit,  
"Traffic_Calming" bit,  
"Traffic_Signal" bit,  
"Turning_Loop" nvarchar(50),  
"Time_ID" datetime  
)
```



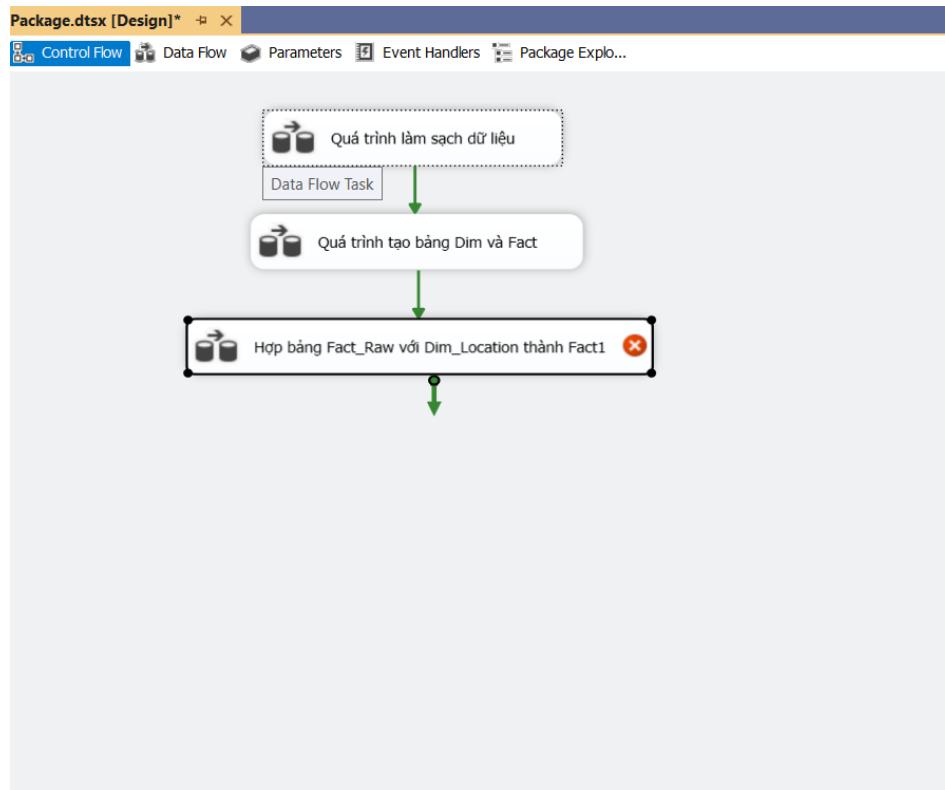
- + Chọn thẻ **Mappings** để kiểm tra việc ánh xạ của các cột thuộc tính. Chọn ánh xạ cột **ID** trong **Input Column** vào cột **Fact_ID** vừa bảng **Fact_Raw**. Sau đó nhấn **OK**



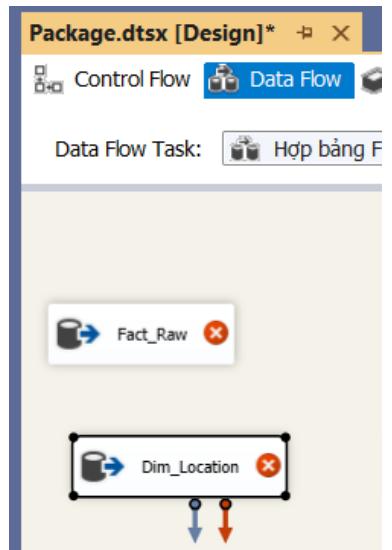
2.4. Merge Fact_Raw và Dim_Location thành Fact1

Ở bước này sẽ loại bỏ các cột dữ liệu trong bảng Fact mà trùng với các cột thuộc tính của bảng Dimension, đồng thời thực hiện thêm khóa ngoại vào bảng Fact để đơn giản hóa quá trình phân tích dữ liệu.

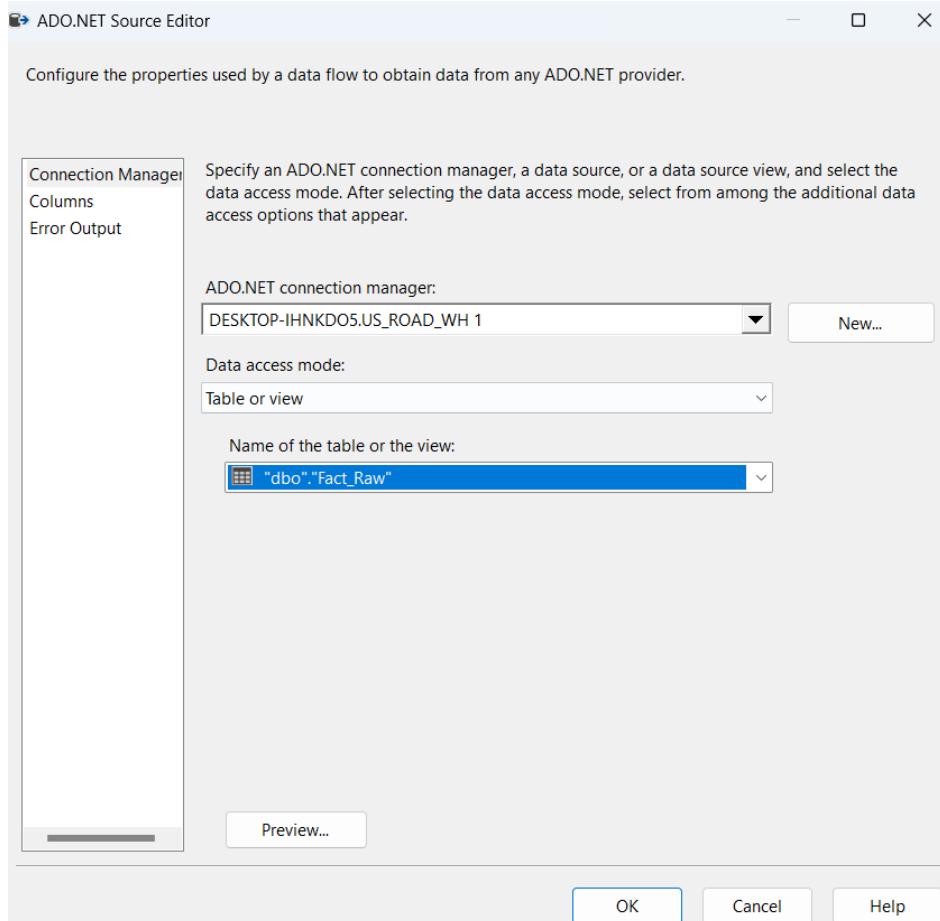
Bước 1: Tạo Data Flow Task bằng cách kéo thả vào thẻ **Control Flow** và đổi tên thành “**Hợp bảng Fact_Raw với Dim_Location thành Fact1**”. Sau đó liên kết với “**Quá trình tạo bảng Dim và Fact**” rồi click chuột phải chọn **Edit**



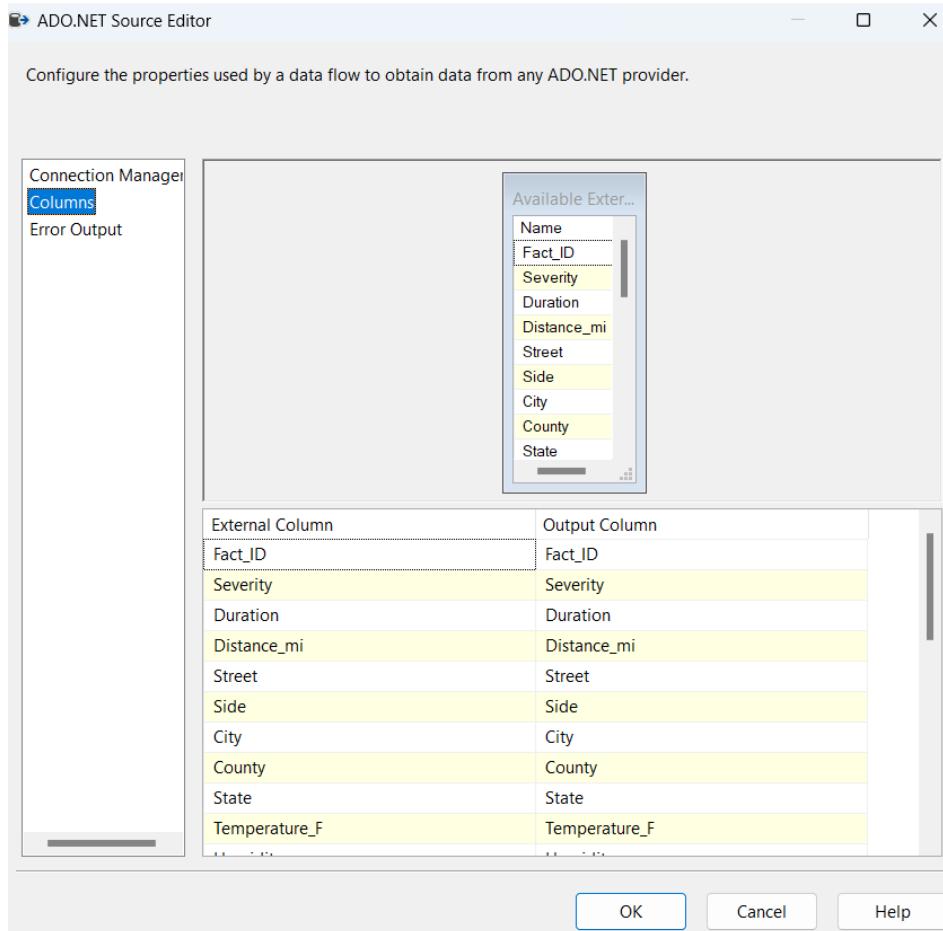
Bước 2: Tạo 2 thành phần ADO NET Source và đổi tên thành **Fact_Raw** và **Dim_Location**



Bước 3: Click chuột phải vào **Fact_Raw**, chọn **Edit** và chọn bảng **Fact_Raw** đã tạo trước đó làm dữ liệu nguồn cho bảng **Fact_Raw** mới này.

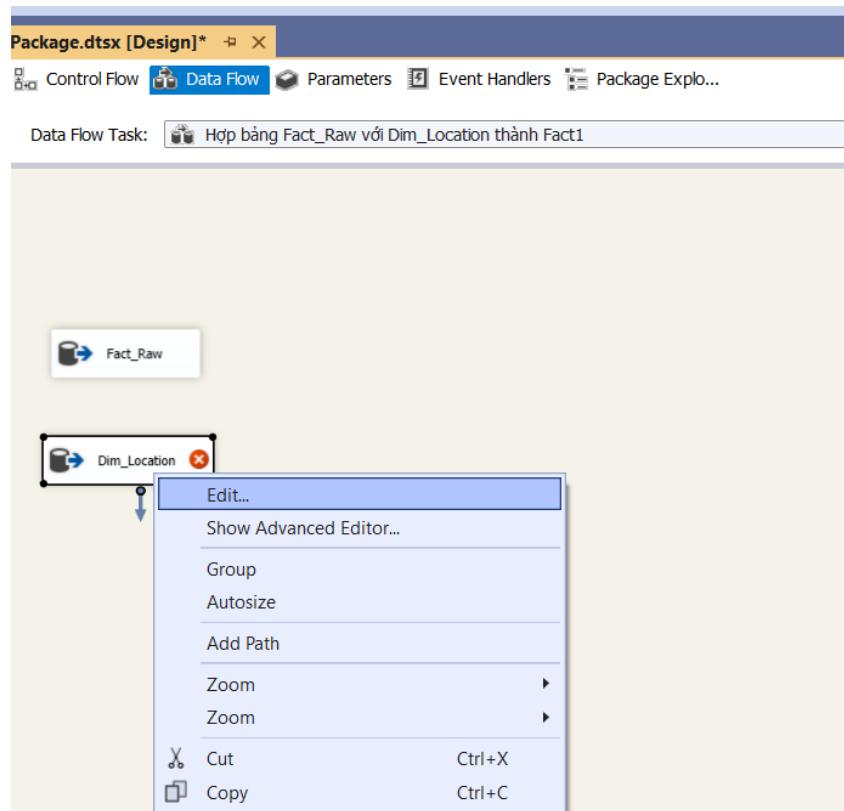


Bước 4: Chọn thẻ **Columns** để xem các cột được ánh xạ. Nhấn **OK**

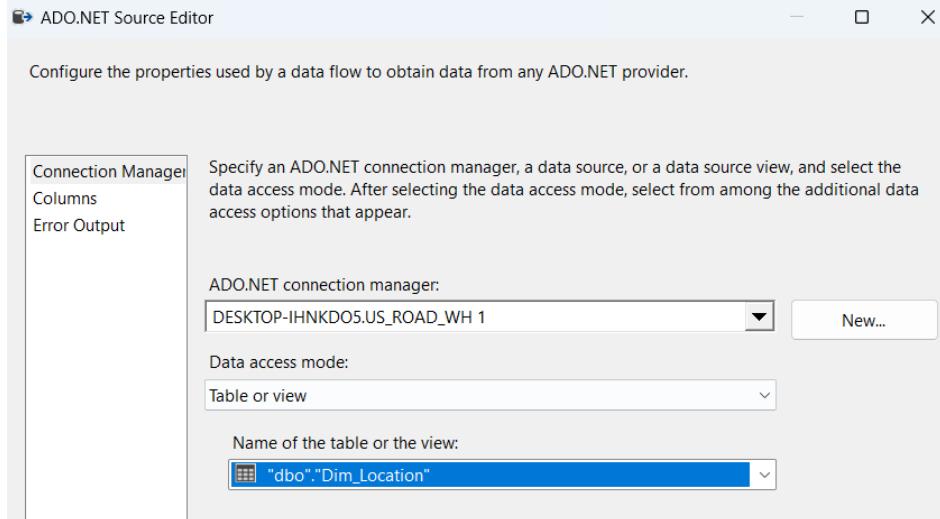


Bước 4:

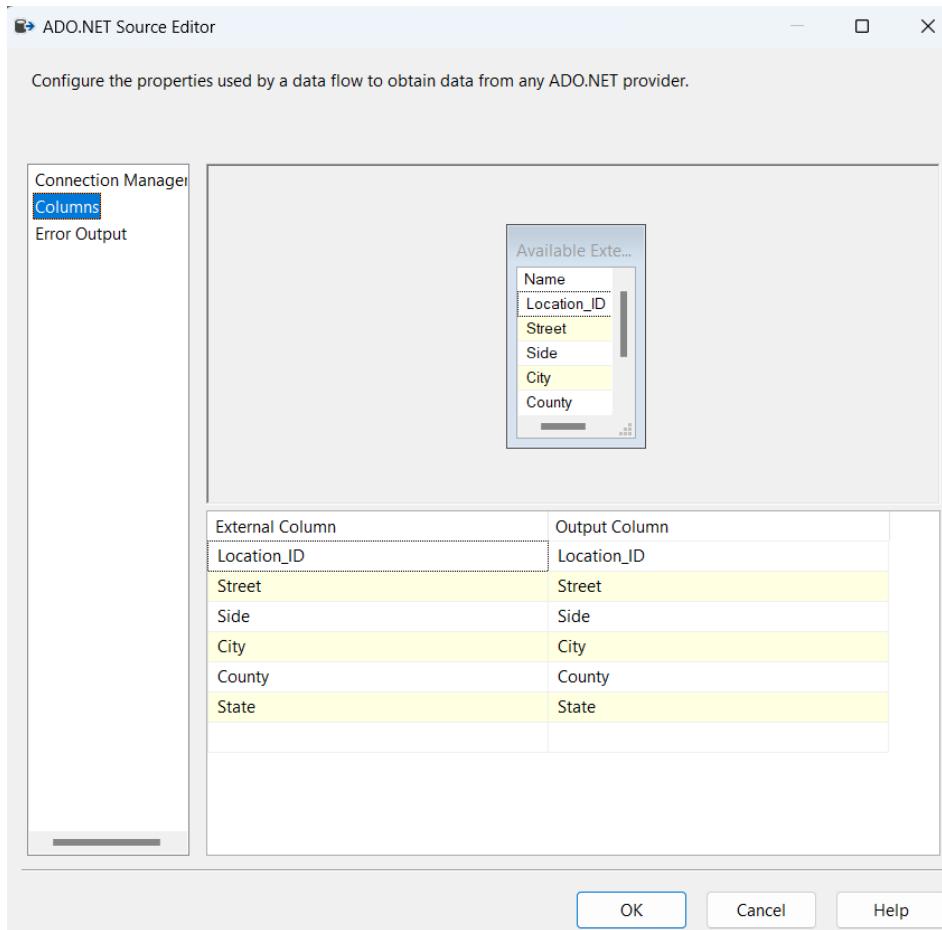
- + Ánh xạ cho bảng **Dim_Location** được thực hiện tương tự.
- + Click chuột phải vào **Dim_Location** chọn **Edit**.



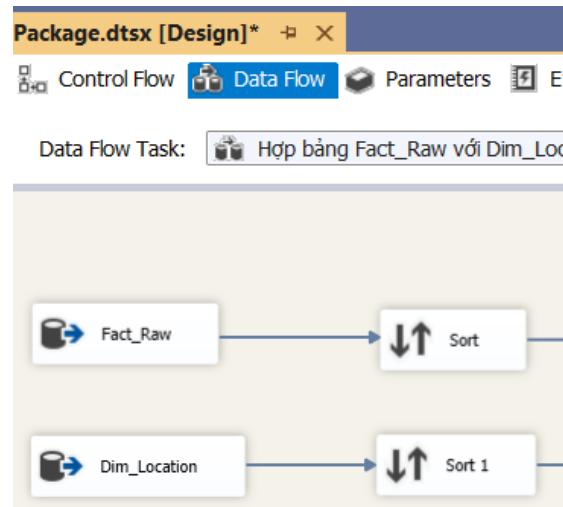
Bước 5: Chọn bảng **Dim_Location** được tạo trước đó để làm dữ liệu nguồn cho bảng **Dim_Location** này



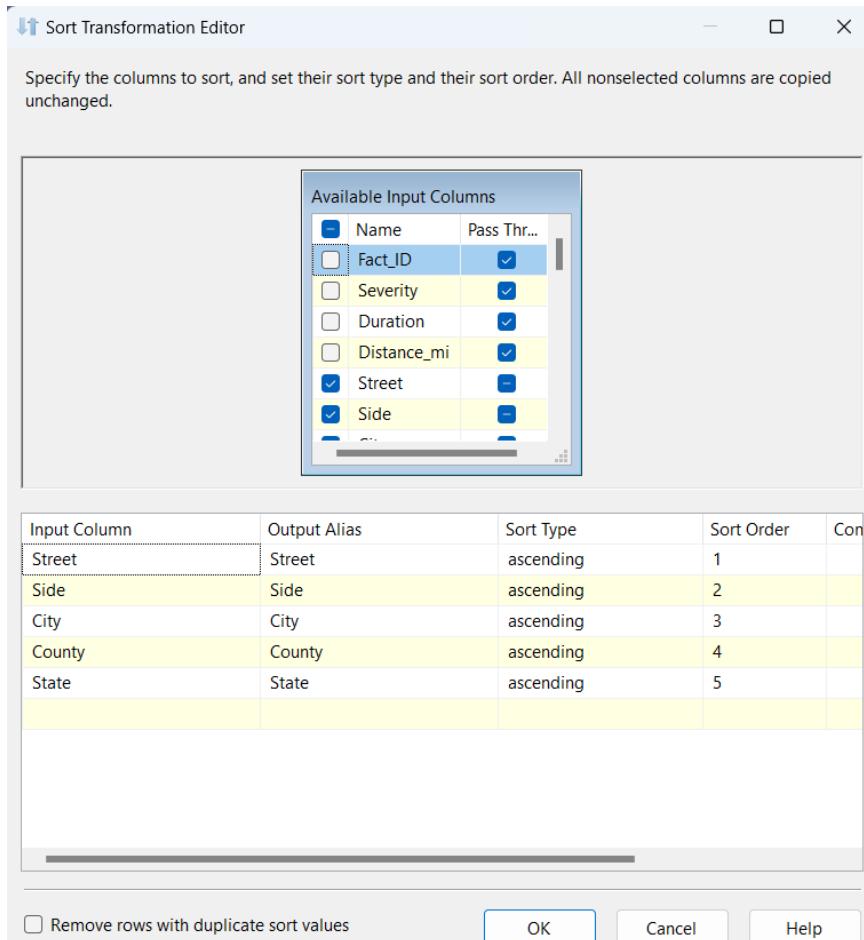
Bước 6: Chuyển sang thẻ **Columns** để xem các cột được ánh xạ vào bảng **Dim_Location**



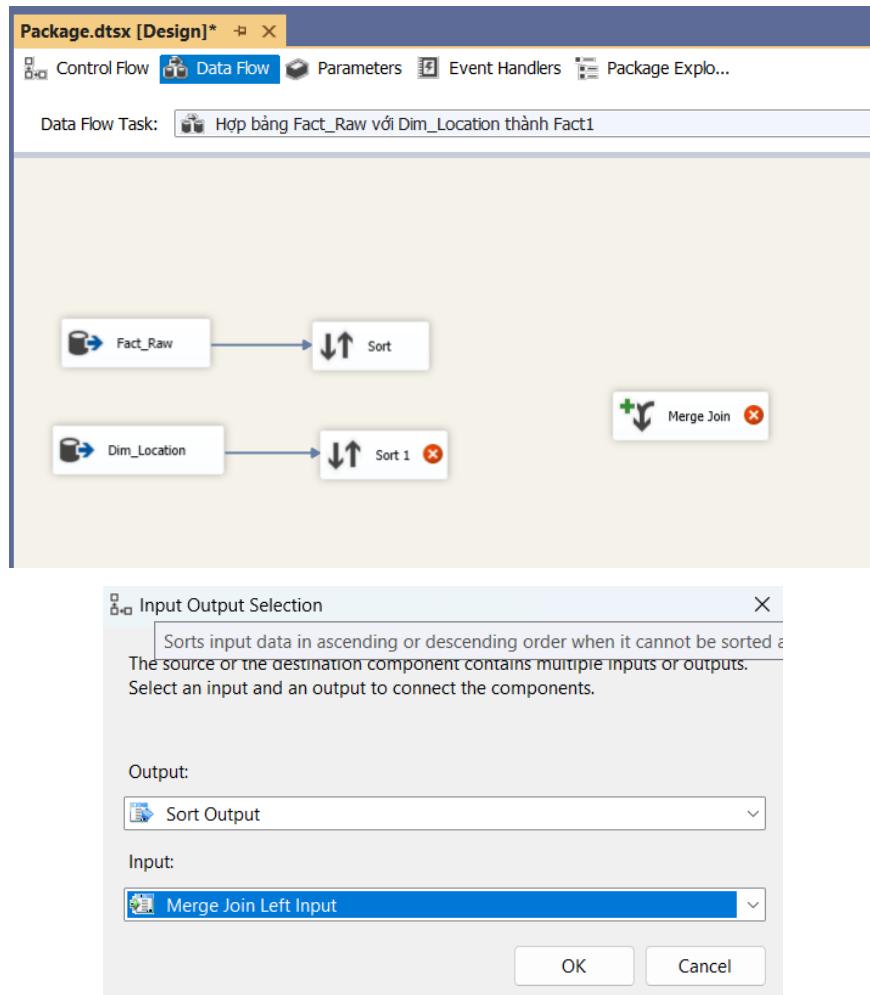
Bước 7: Tạo 2 Sort liên kết với từng Source vừa tạo



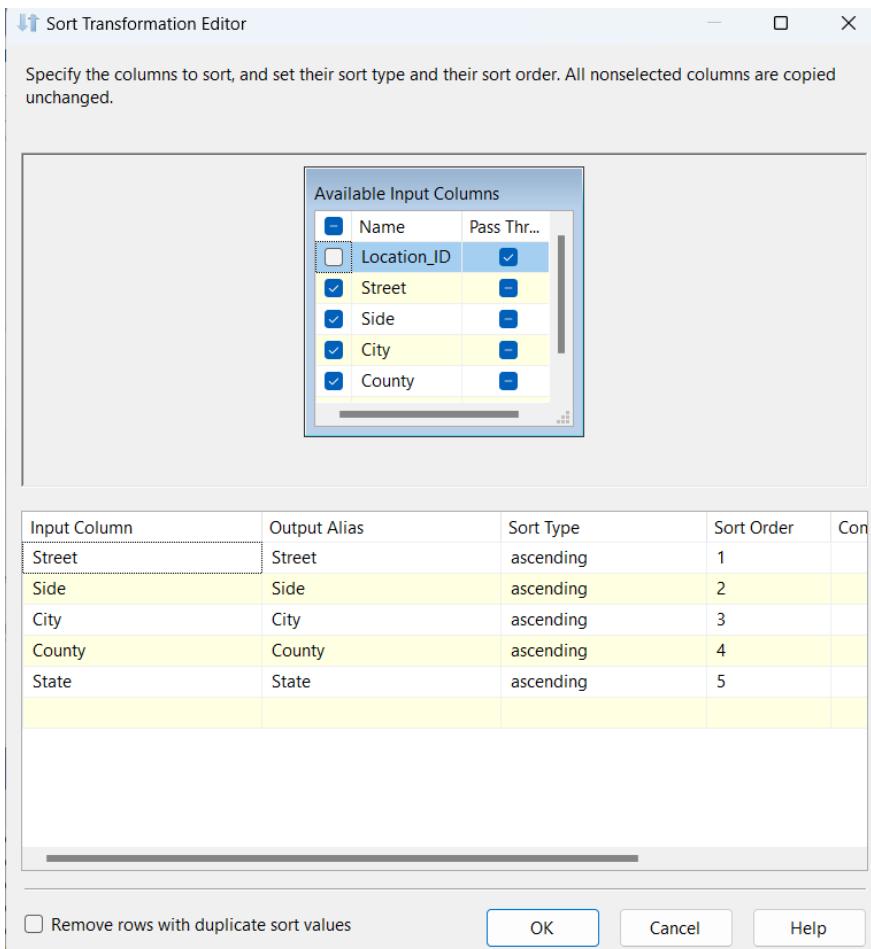
Bước 8: Click chuột phải vào **Sort** và chọn **Edit**. Sau đó chọn các cột **Street, City, County, State** theo thứ tự giống với bảng **Dim_Location** để từng bước chuẩn bị cho quá trình **Merge**.



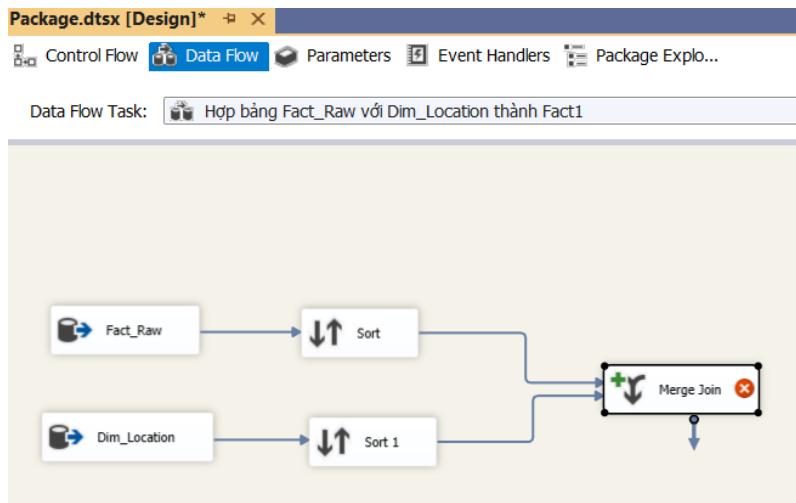
Bước 9: Tạo mới một **Merge Join** và nối **Merge Join** đó với **Sort** và chọn **Input là Merge Join Left Input** để giữ lại toàn bộ các dòng trong bảng **Fact_Raw** bất kể có kết quả khi thực hiện phép kết trái với cột **ID** của bảng **Dim_Location** hay không.



Bước 10: Chọn các cột thuộc tính như hình bên dưới cho thành phần Sort 1

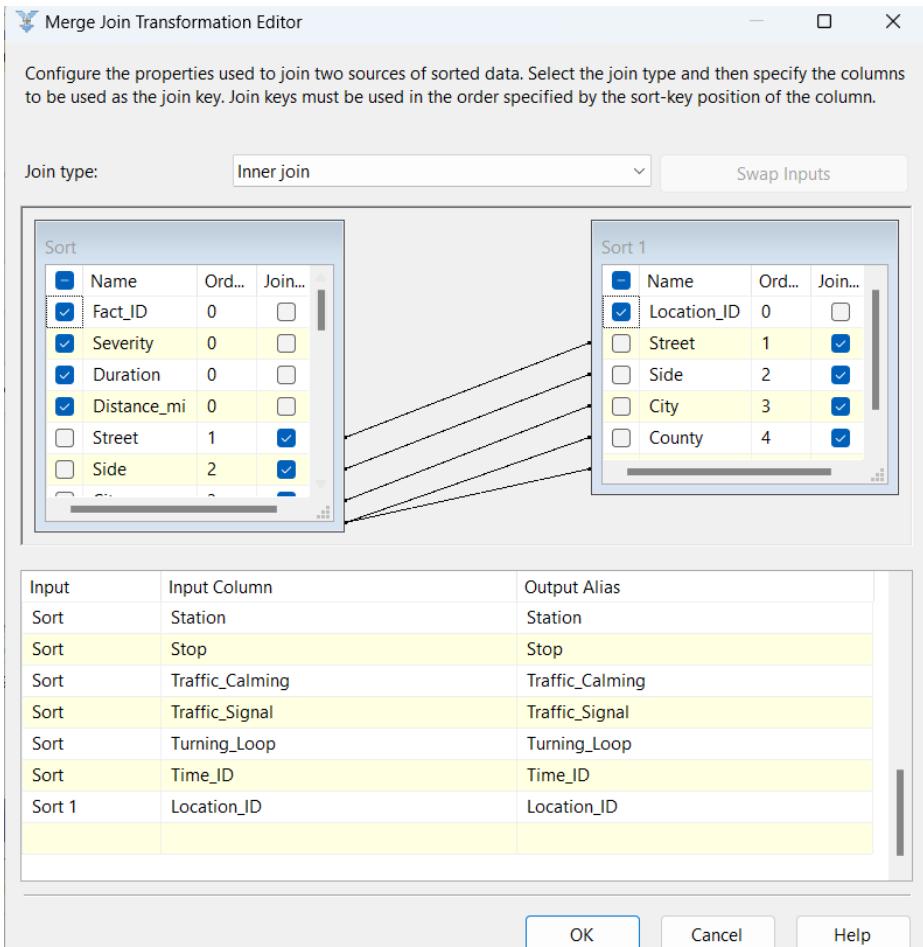


Bước 11: Nối Sort 1 với Merge Join

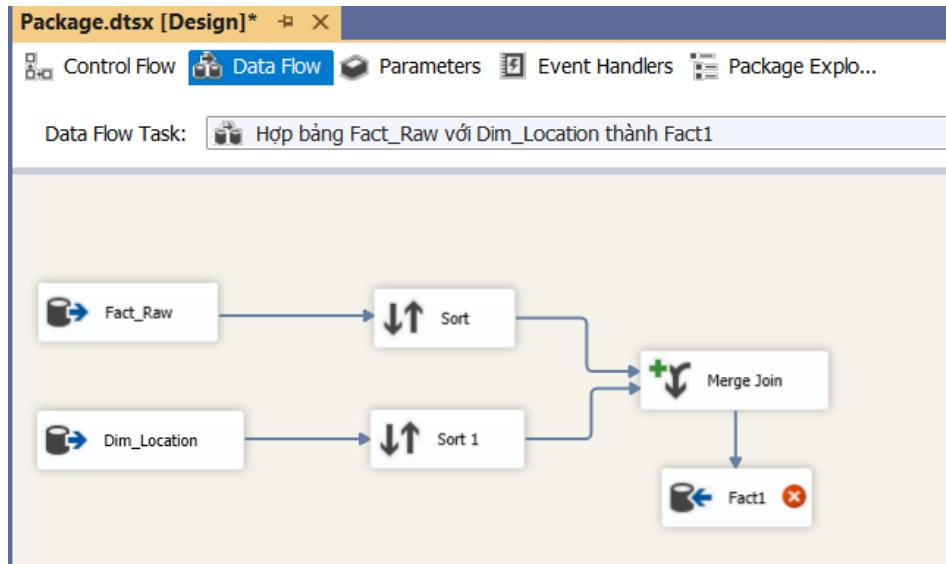


Bước 12:

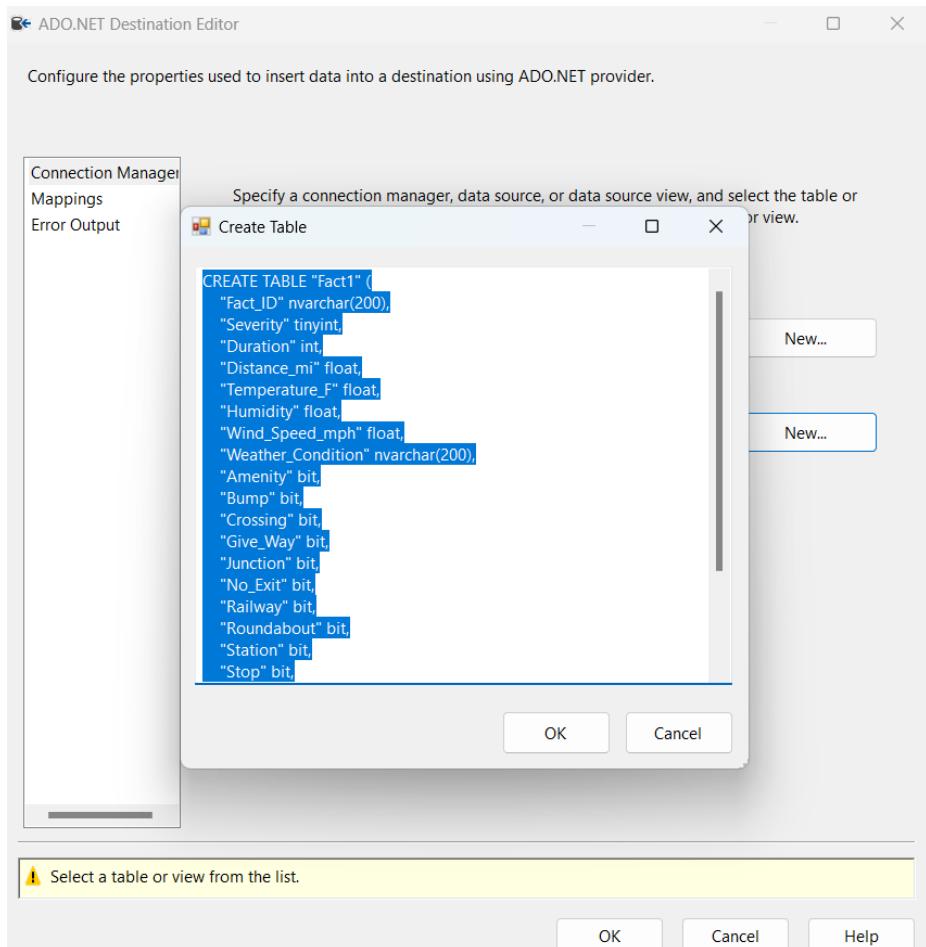
- + Click chuột phải lên **Merge Join** và chọn **Edit**
- + Click chọn các cột thuộc tính của **Sort** nhưng chưa lại các thuộc tính **Street, City, Side, County, State**
- + Chọn **Location_ID** ở **Sort1** để merge vào **Fact_Raw**
- + Sau khi merge xong, bảng **Fact_Raw** không còn 4 thuộc tính của bảng **Dim_Location** nữa và có thêm 1 cột thuộc tính mới là **Location_ID**



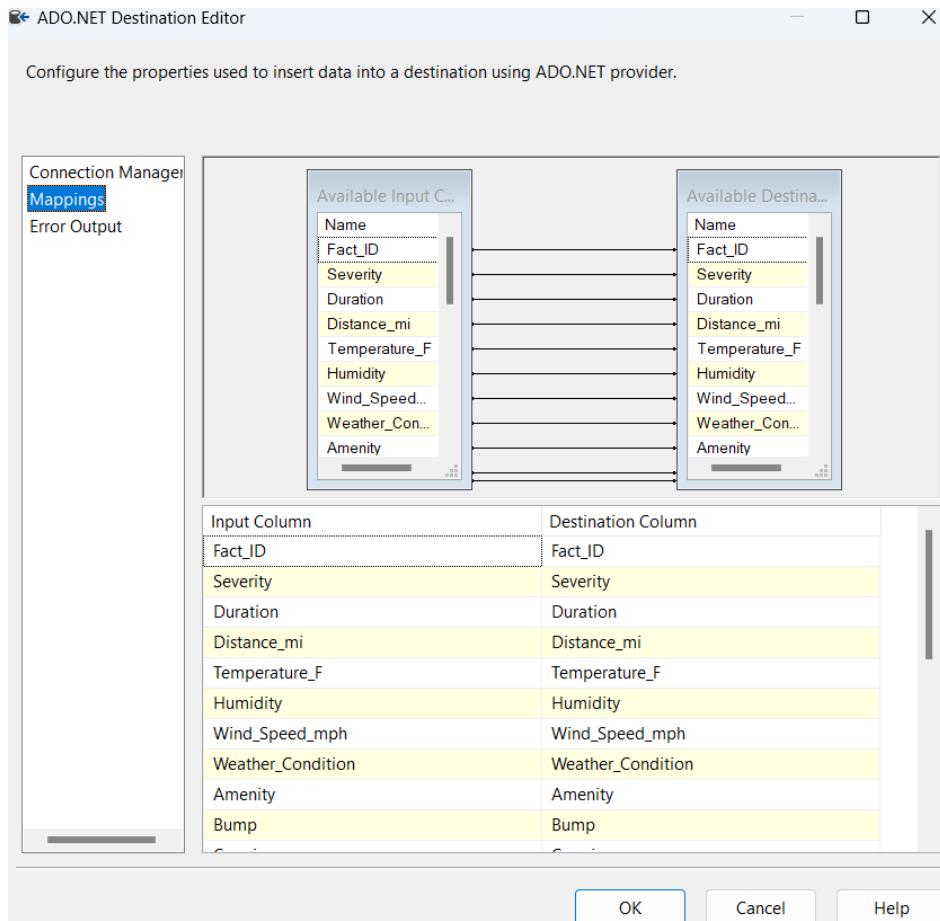
Bước 13: Tạo một ADO NET Destination và đổi tên thành Fact1 để chứa kết quả sau khi Merge

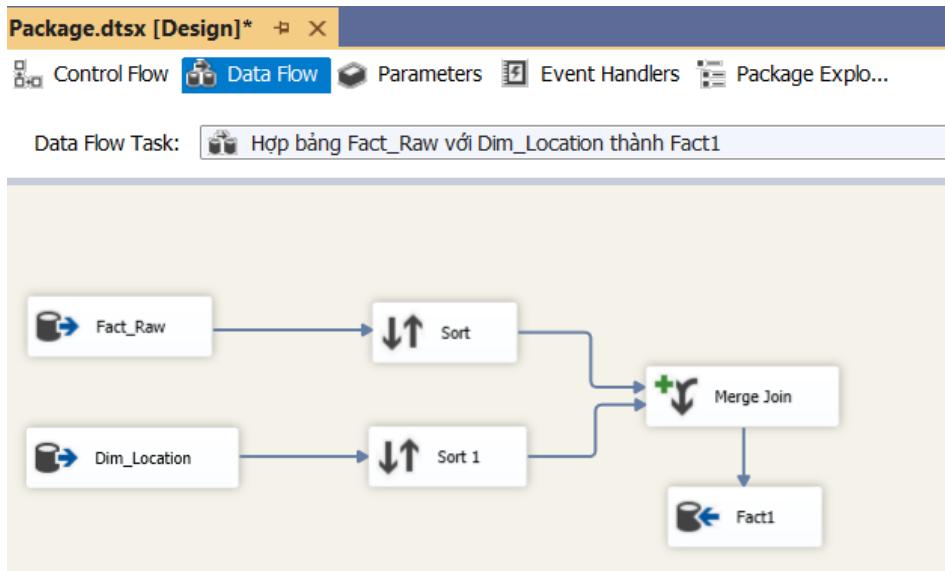


Bước 14: Tạo bảng **Fact1** bằng câu lệnh SQL như bên dưới:



Bước 15: Chọn **Mappings** để kiểm tra ánh xạ các cột thuộc tính

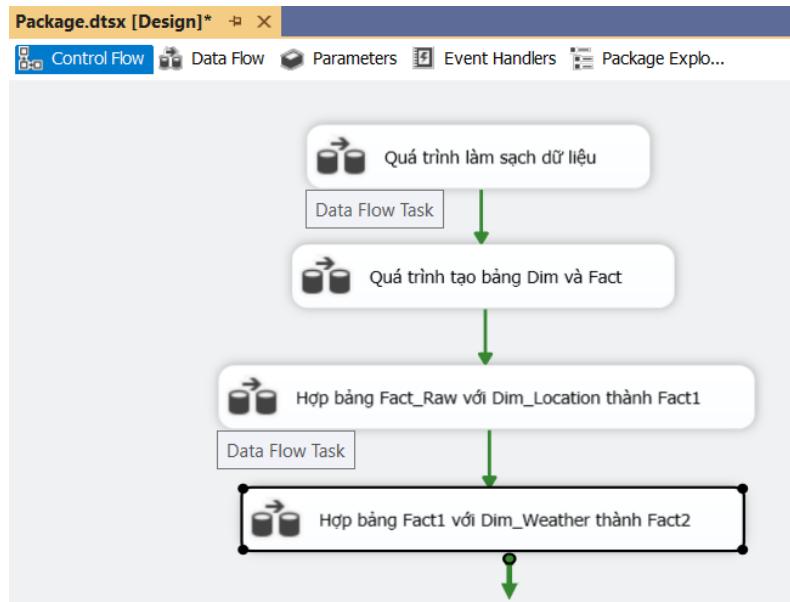




Sau đó tiếp tục ghép bảng **Fact1** với các bảng **Dimension** còn lại

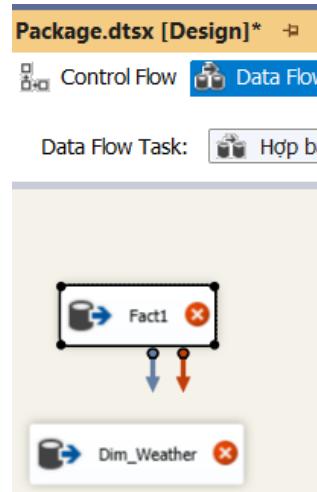
2.5. Merge Fact1 với Dim_Weather vào Fact2

Bước 1: Thêm mới 1 Data Flow Task vào Control Flow và đổi tên thành “**Hợp bảng Fact1 với Dim_Weather thành Fact2**”

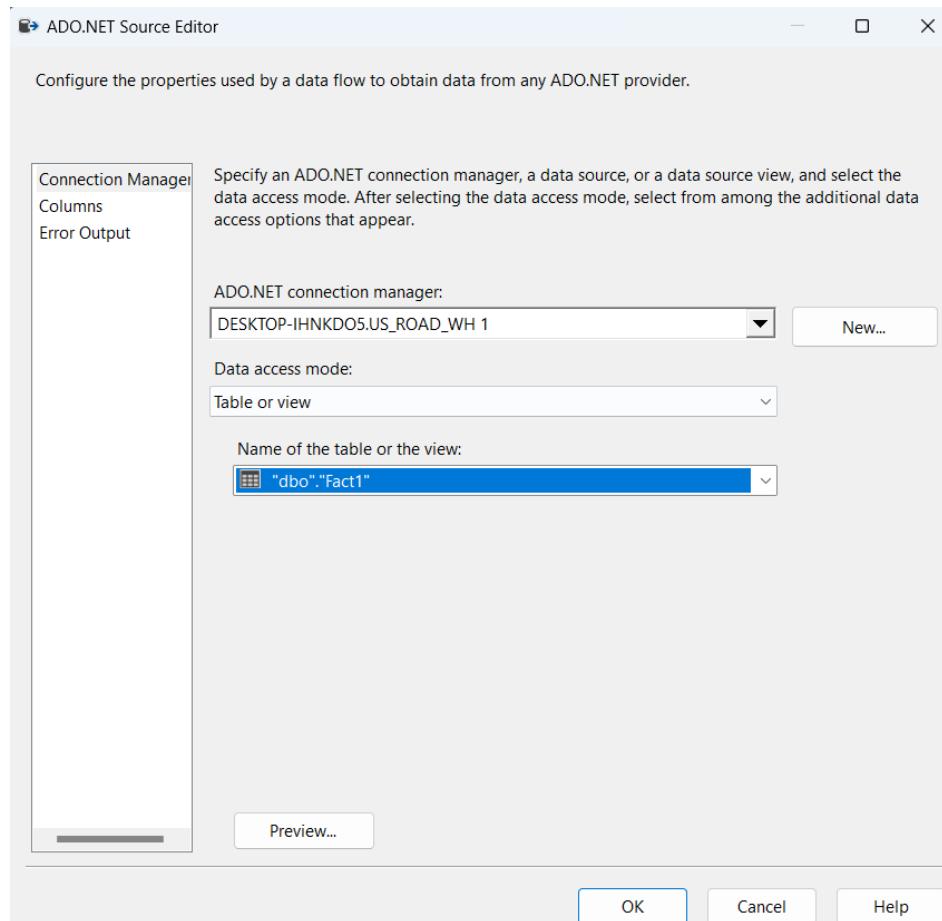


Bước 2: Click chuột phải và chọn **Edit**.

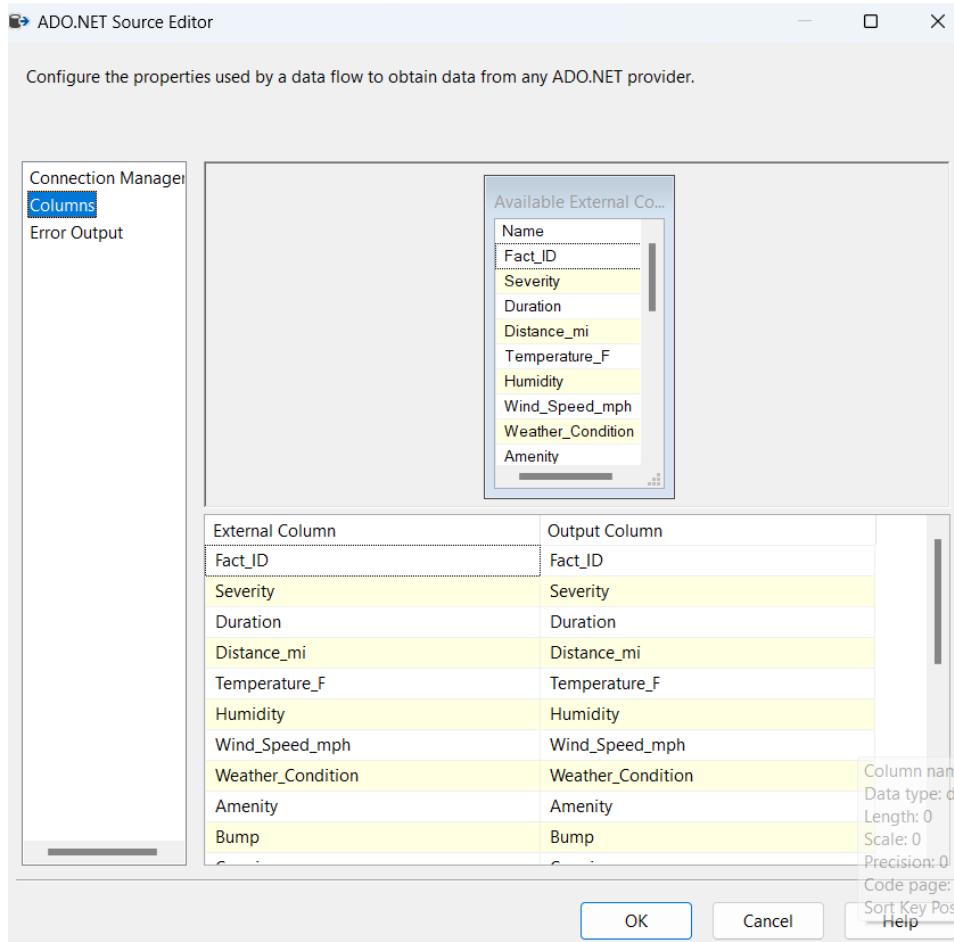
Bước 3: Tạo 2 ADO NET Source bằng cách kéo thả vào màn hình và đổi tên thành **Fact1** và **Dim_Weather**



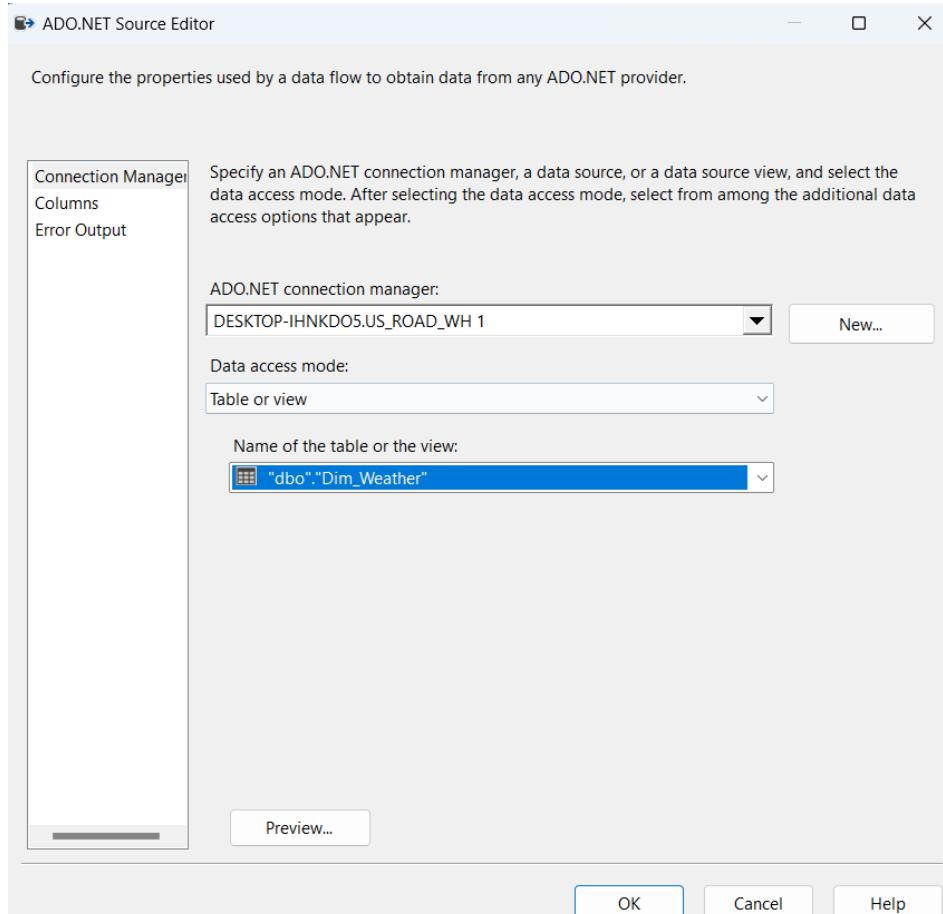
Bước 4: Click chuột phải lên **Fact1** và chọn **Edit**. Tiếp hành chọn bảng **Fact1** là kết quả của quá trình **Merge** ở bước trước làm **data source**.



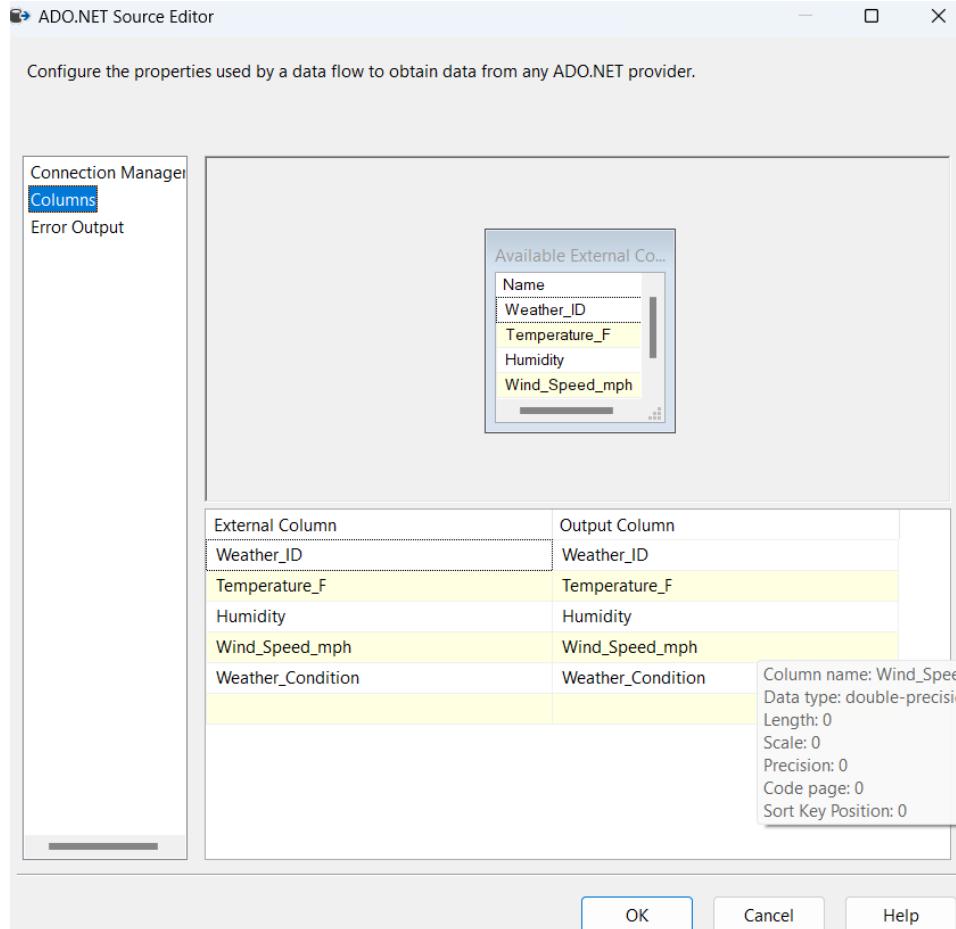
Bước 5: Chọn thẻ **Columns** để chọn các cột thuộc tính cần **Merge** và nhấn **OK**.



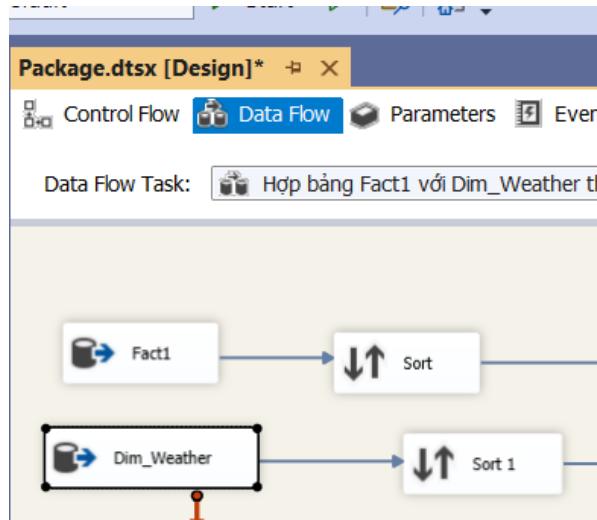
Bước 6: Làm tương tự các bước trên với ADO NET Source cho bảng Dim_Weather



Bước 7: Chọn các cột thuộc tính trong thẻ **Columns** và nhấn **OK**



Bước 8: Tạo 2 Sort tương ứng cho mỗi Data Source.



Bước 9: Click chuột phải lên Sort và chọn **Edit**. Tiếp hành chọn các cột **Temperature_F**, **Humidity**, **Wind_Speed_mph**, **Weather_Condition** với thứ tự giống với bảng **Dim_Weather**.

Sort Transformation Editor

Specify the columns to sort, and set their sort type and their sort order. All nonselected columns are unchanged.

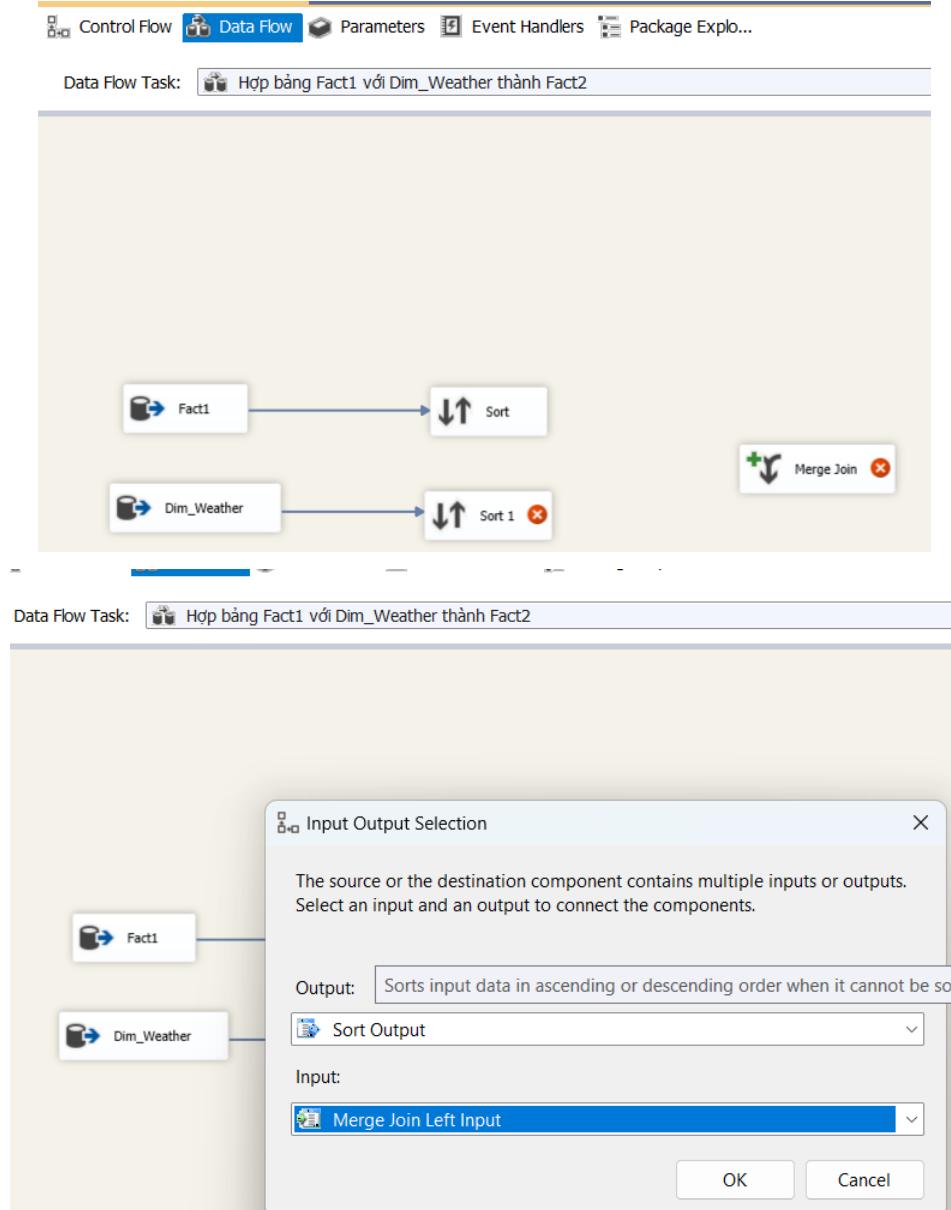
Available Input Columns

Name	Pass Thr...
Distance_mi	<input checked="" type="checkbox"/>
Temperature_F	<input checked="" type="checkbox"/>
Humidity	<input checked="" type="checkbox"/>
Wind_Speed_mph	<input checked="" type="checkbox"/>
Weather_Condition	<input checked="" type="checkbox"/>
Amenity	<input type="checkbox"/>

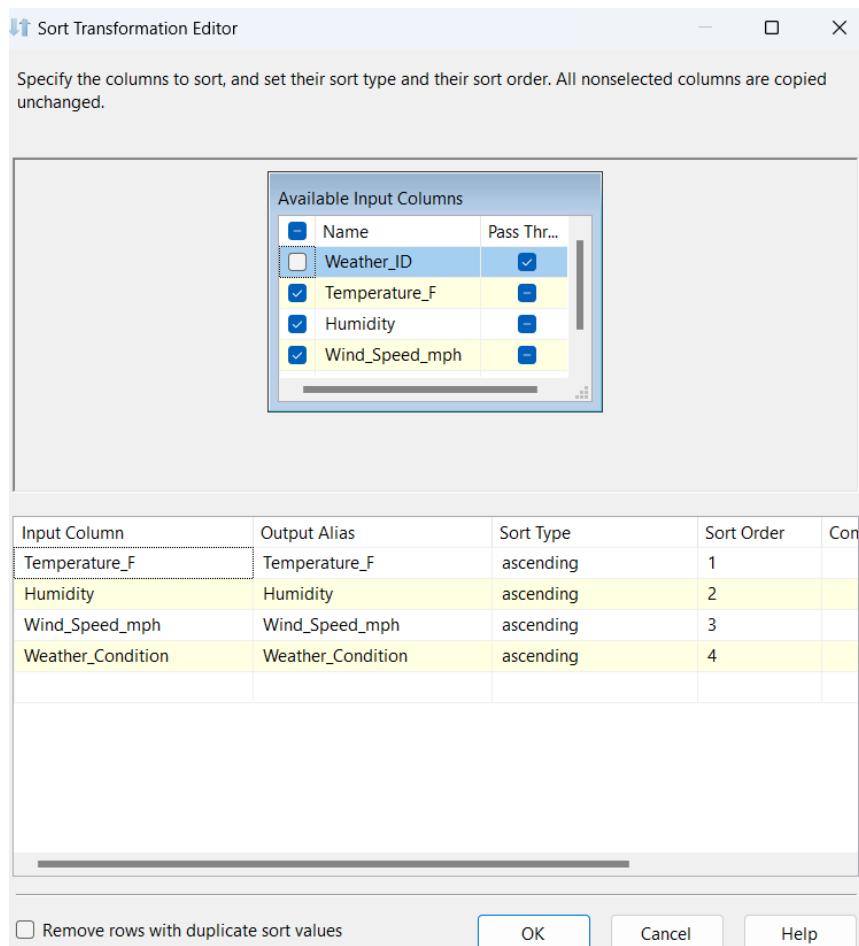
Input Column	Output Alias	Sort Type	Sort Order
Temperature_F	Temperature_F	ascending	1
Humidity	Humidity	ascending	2
Wind_Speed_mph	Wind_Speed_mph	ascending	3
Weather_Condition	Weather_Condition	ascending	4

Bước 10: Thêm một **Merge Join** và nối với **Sort**. Chọn **Merge Join Left Input** để giữ lại toàn bộ các dòng trong bảng **Fact1** bất kể có kết quả khi thực hiện phép kết trái với cột **ID** của bảng **Dim_Weather** hay không.

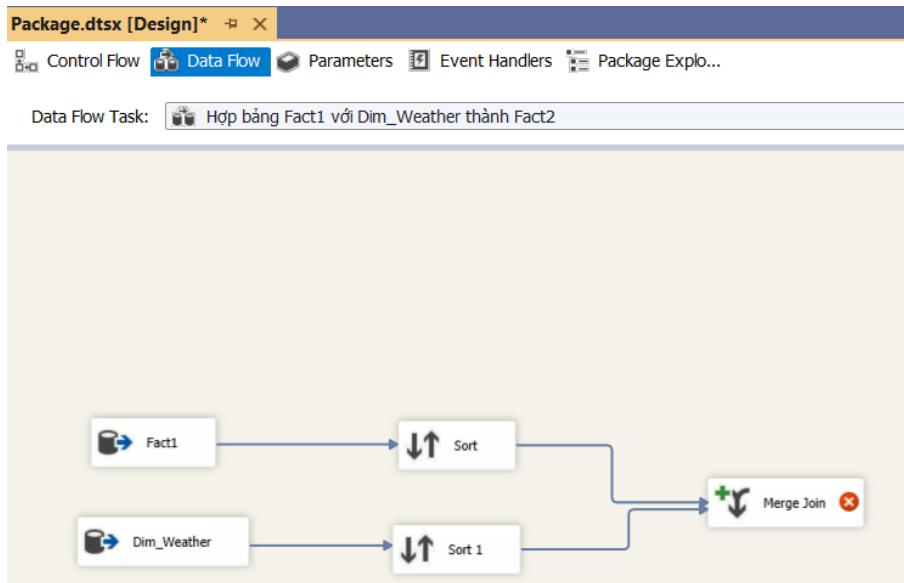
Ths. Đỗ Thị Minh Phụng



Bước 11: Click chuột phải lên **Sort1** và Chọn các cột **Weather_Timestamp**, **Temperature_F**, **Humidity**, **Wind_Direction**, **Wind_Speed_mph**, **Precipitation_in**, **Weather_Condition** và nhấn **OK**.

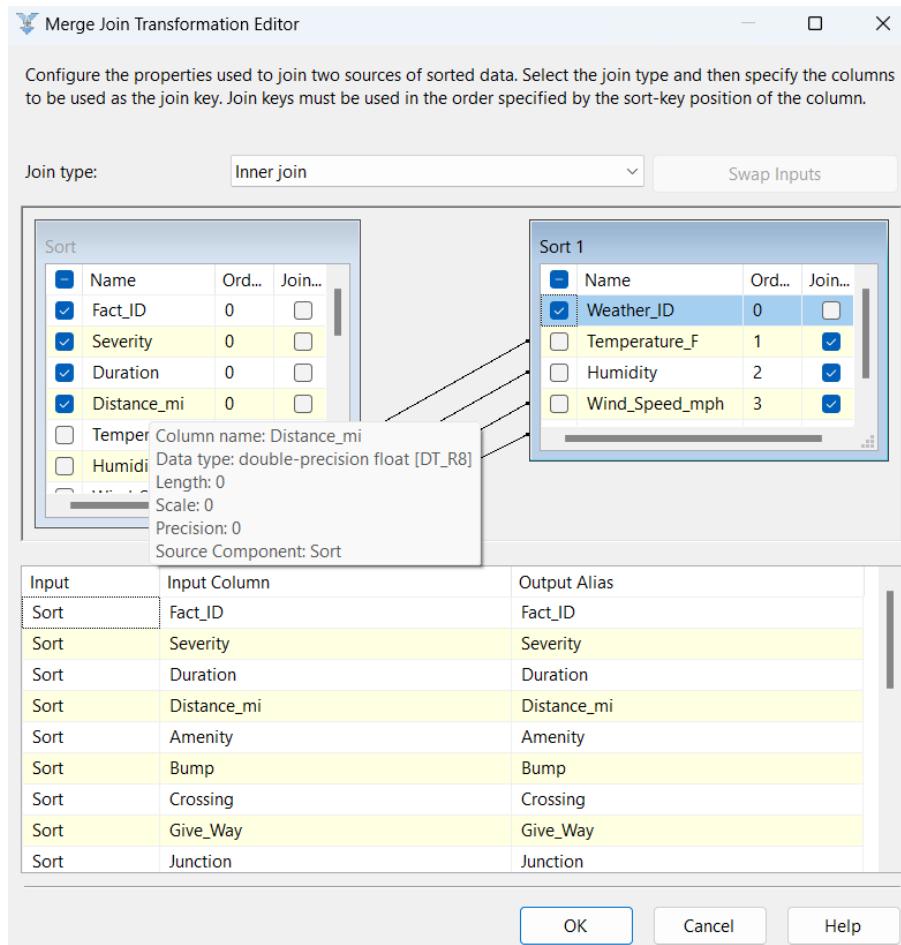


Bước 12: Nối Sort1 với Merge Join



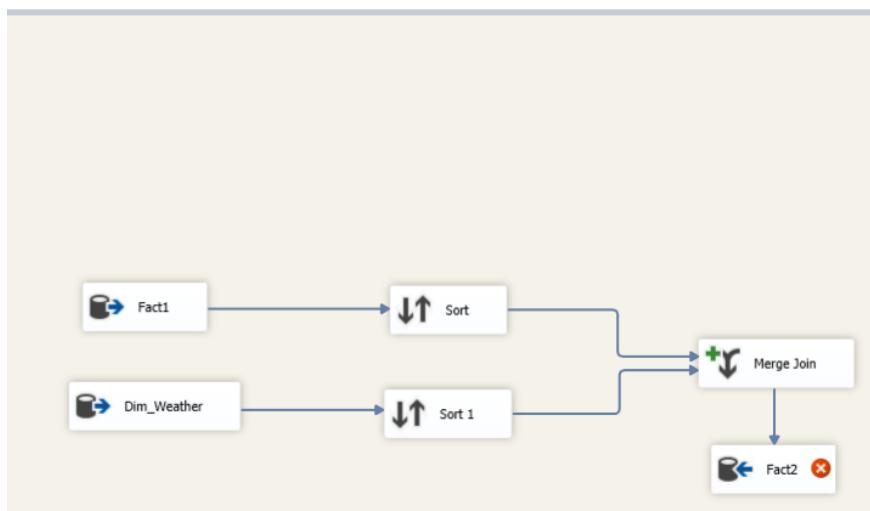
Bước 13:

- + Click chuột phải vào **Merge Join** và chọn **Edit**
- + Tick chọn toàn bộ cột thuộc tính của **Sort** nhưng không lấy thuộc tính **Weather_Timestamp**, **Temperature_F**, **Humidity**, **Wind_Direction**, **Wind_Speed_mph**, **Precipitation_in**, **Weather_Condition**
- + Chọn **Weather_ID** của **Sort1** để tiến hành Merge vào **Fact1**
- + Sau khi merge xong thì bảng **Fact1** không còn các thuộc tính **Weather_Timestamp**, **Temperature_F**, **Humidity**, **Wind_Direction**, **Wind_Speed_mph**, **Precipitation_in**, **Weather_Condition** nữa mà sẽ có thêm 1 thuộc tính mới là **Weather_ID**

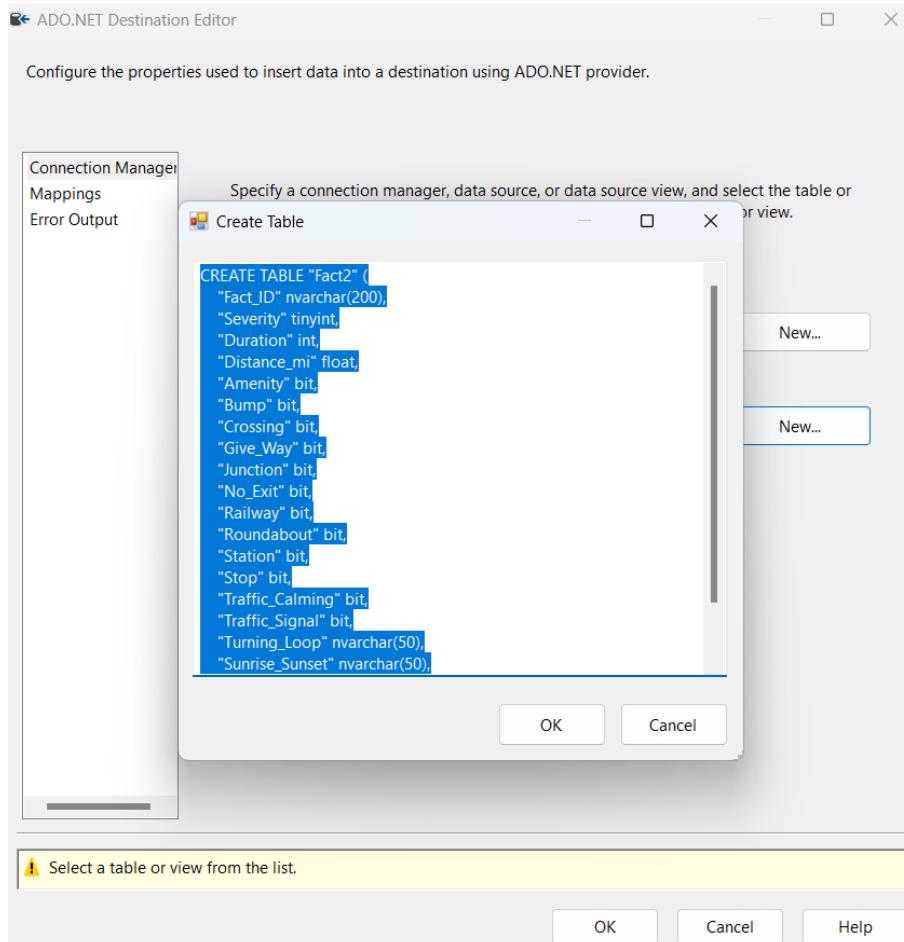


Bước 14: Thêm 1 ADO NET Destination để tạo bảng **Fact2** và chưa toàn bộ kết quả sau khi merge.

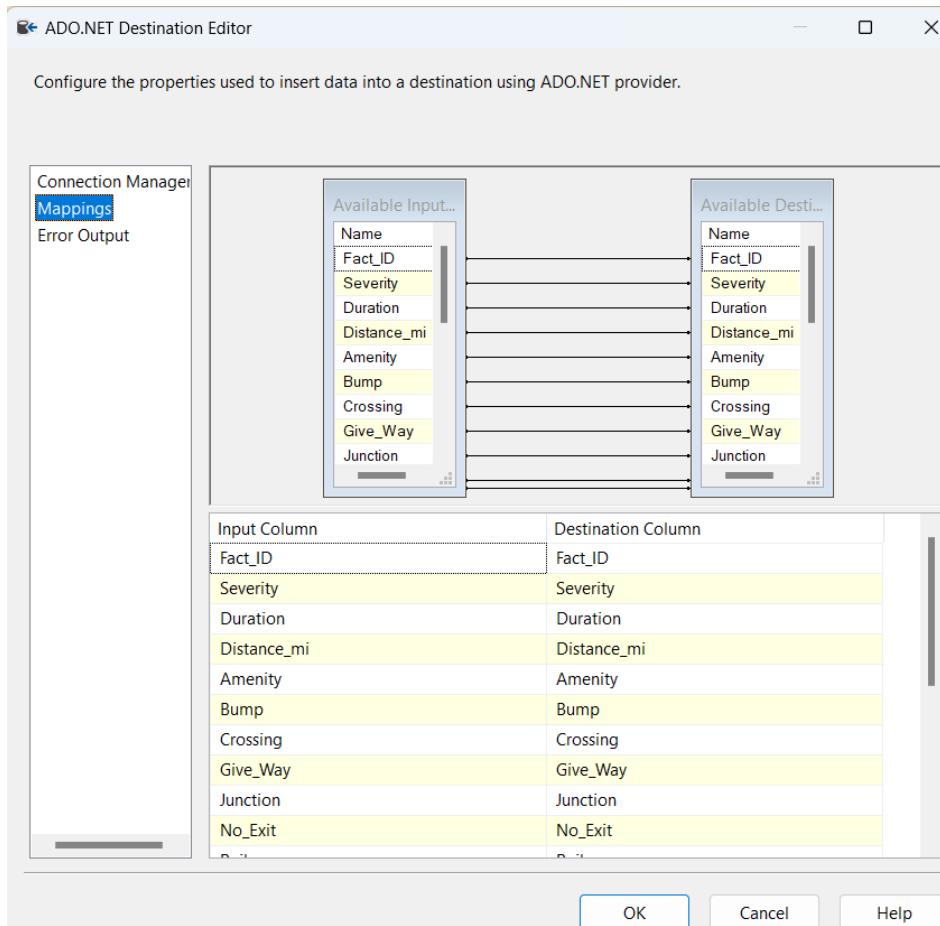
Data Flow Task: Hợp bảng Fact1 với Dim_Weather thành Fact2



Bước 15: Click chuột phải lên **Fact2** và chọn **Edit** sau đó chọn **New...** để hiển thị đoạn mã SQL như hình bên dưới

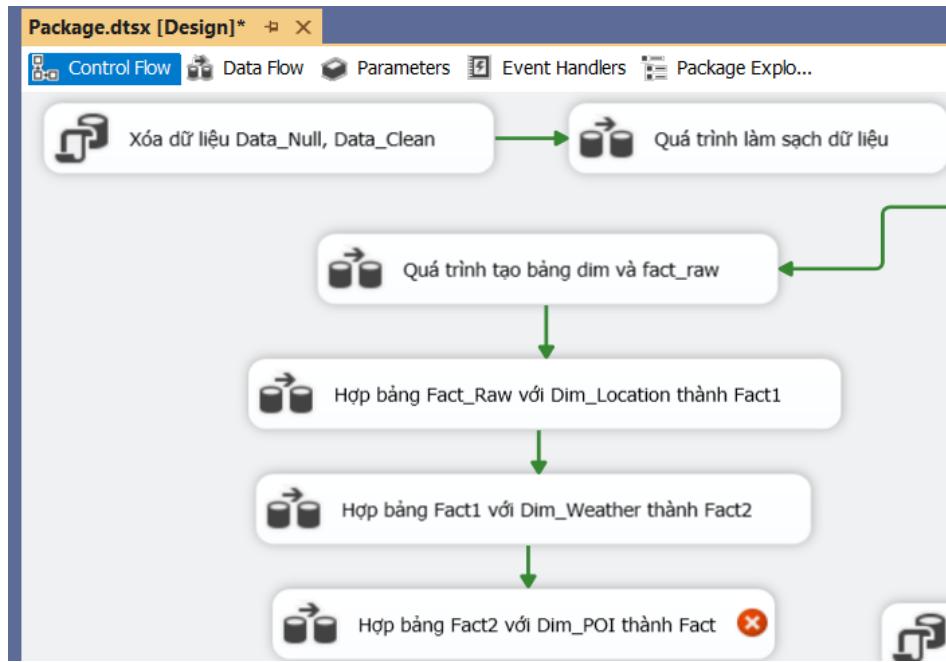


Bước 16: Chọn mục **Mappings** để kiểm tra các ánh xạ của các cột thuộc tính



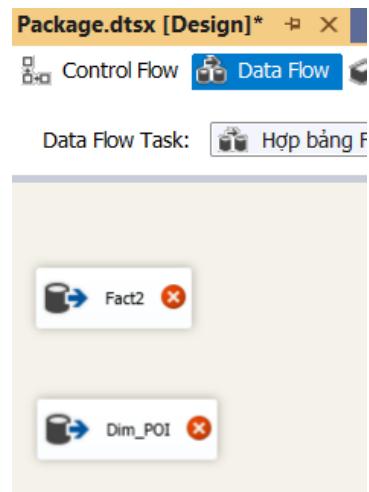
2.6. Merge Fact2 với Dim_POI vào Fact

Bước 1: Thêm một Data Flow Task vào Control Flow và đổi tên thành “Hợp bảng Fact2 với Dim_POI thành Fact”



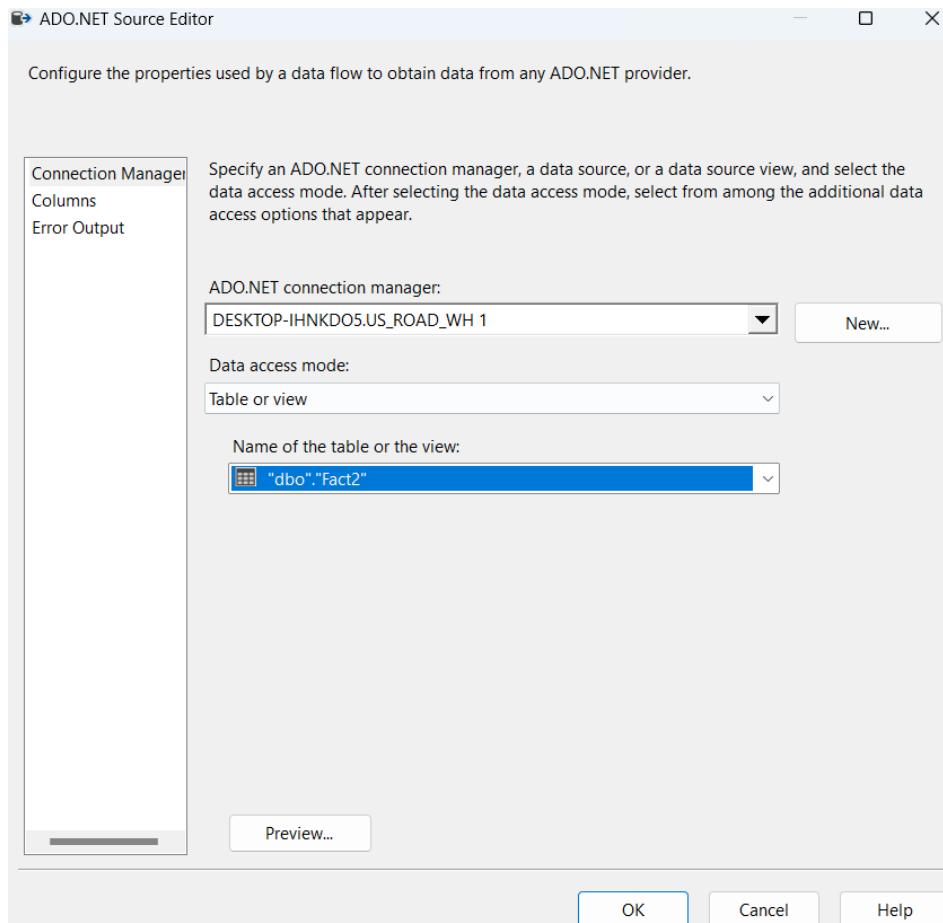
Bước 2:

- + Click chuột phải lên **Data Flow Task** vừa tạo, chọn **Edit**.
- + Thêm 2 **ADO NET Source** vào màn hình và đổi tên thành **Fact2** và **Dim_POI**

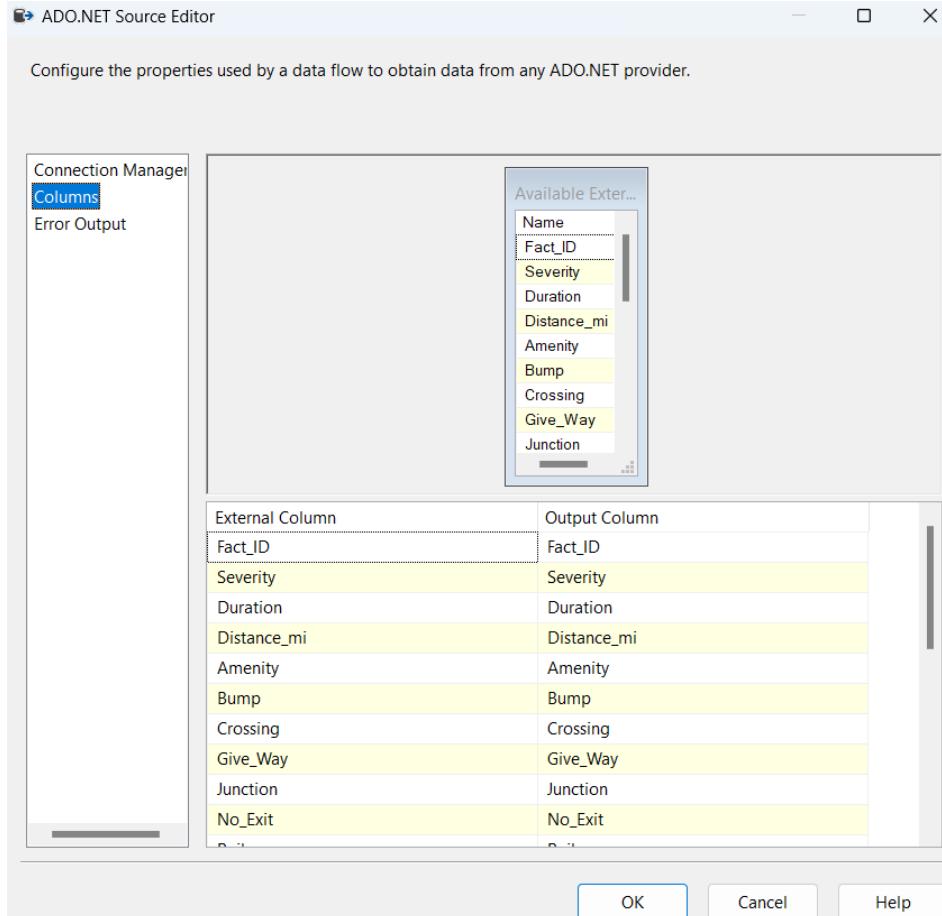


Bước 3:

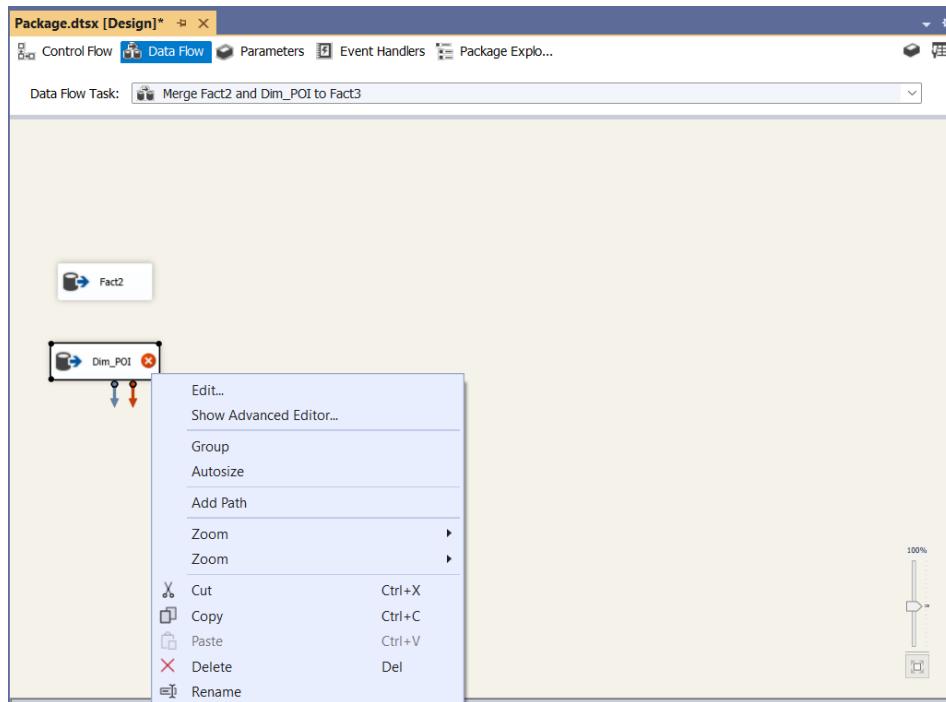
- + Click chuột phải lên **Fact2**, chọn **Edit**. Chọn **Fact2** làm **Data Source** (**Fact2** là kết quả của quá trình merge ở trên)



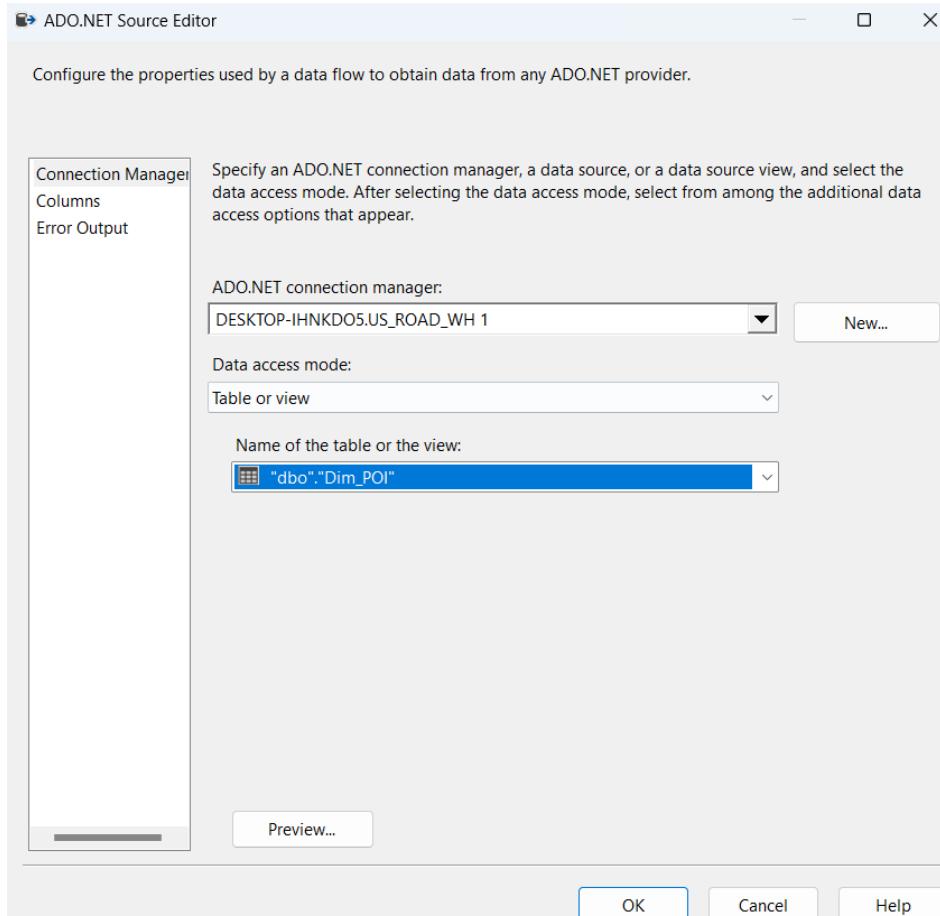
Bước 4: Chọn thẻ **Columns** để kiểm tra các cột thuộc tính được ánh xạ và nhấn **OK**.



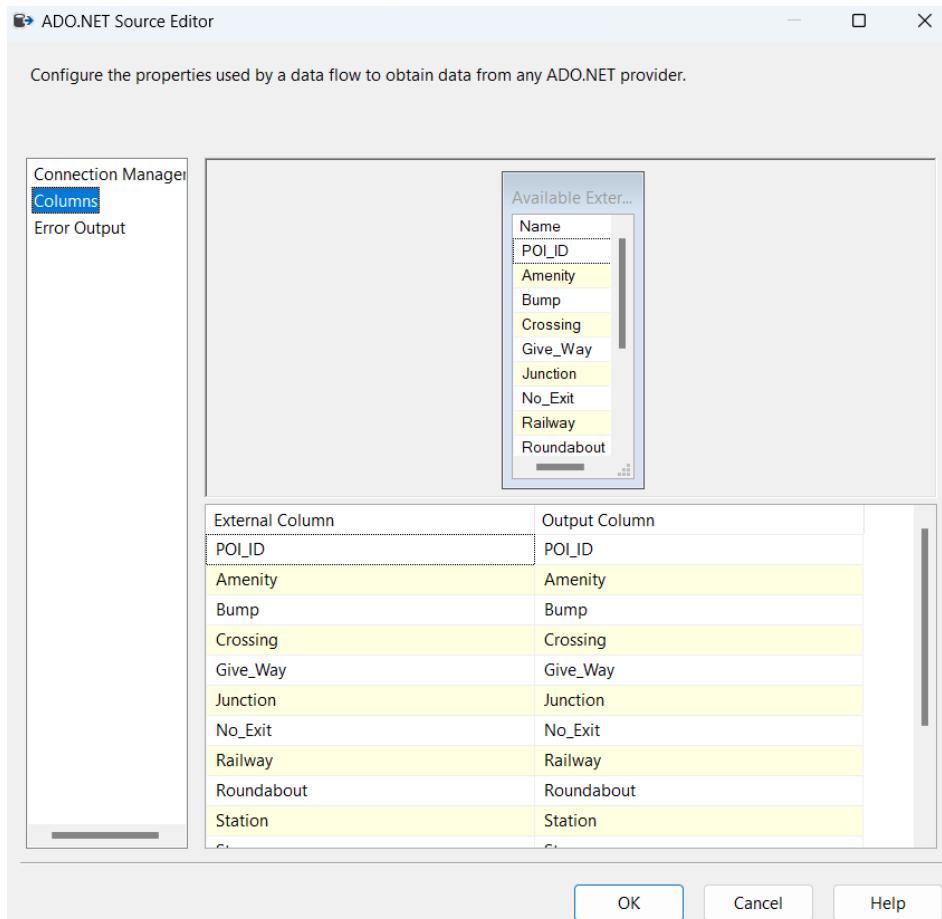
Bước 5: Click chuột phải lên Dim_POI, chọn Edit



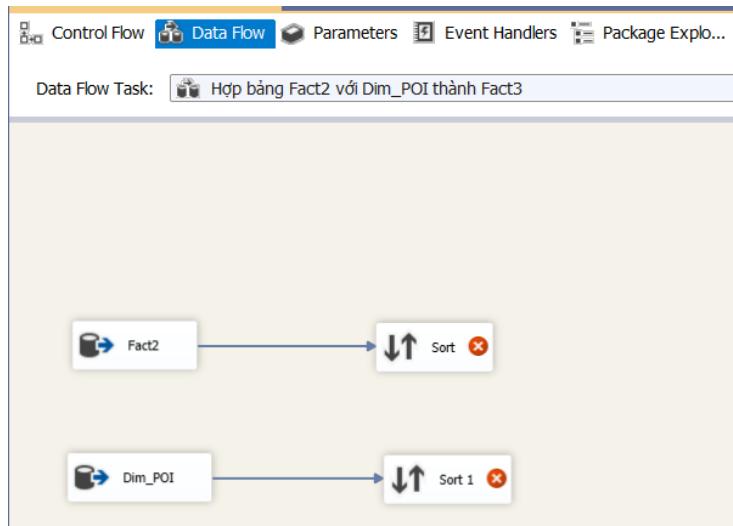
Bước 6: Chọn Dim_POI là Data Source cho Dim_POI



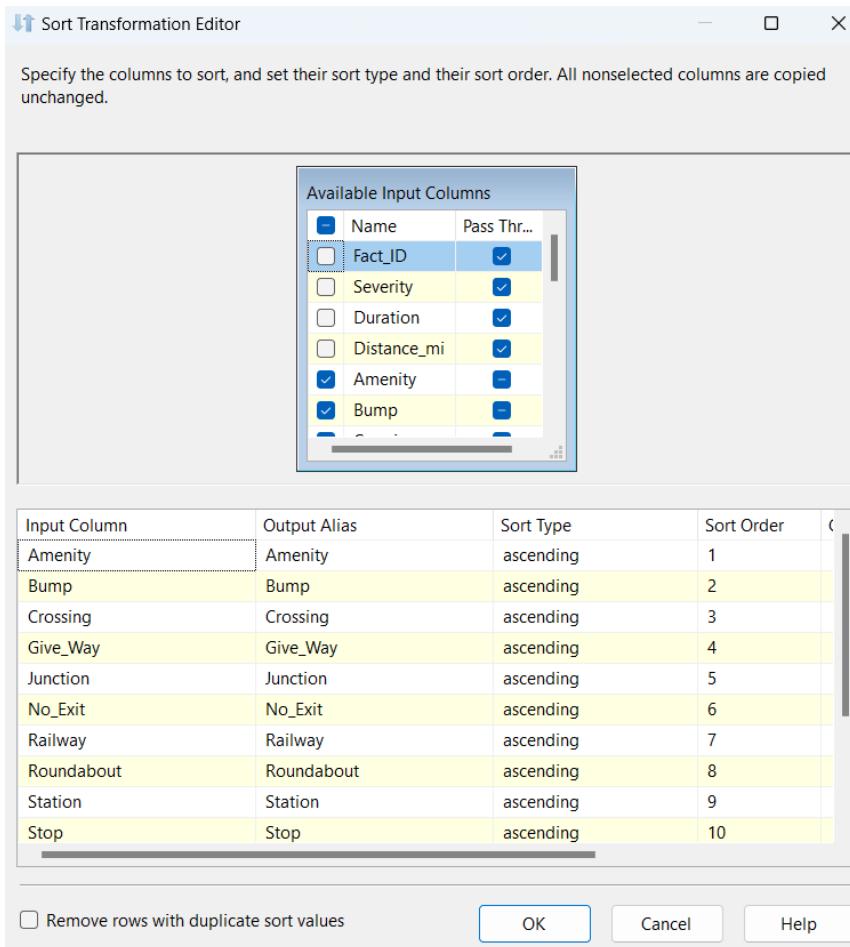
Bước 7: Chuyển sang thẻ **Columns** để kiểm tra các cột thuộc tính được ánh xạ và nhấn **OK**.



Bước 8: Tạo 2 Sort như hình bên dưới ứng với mỗi Source

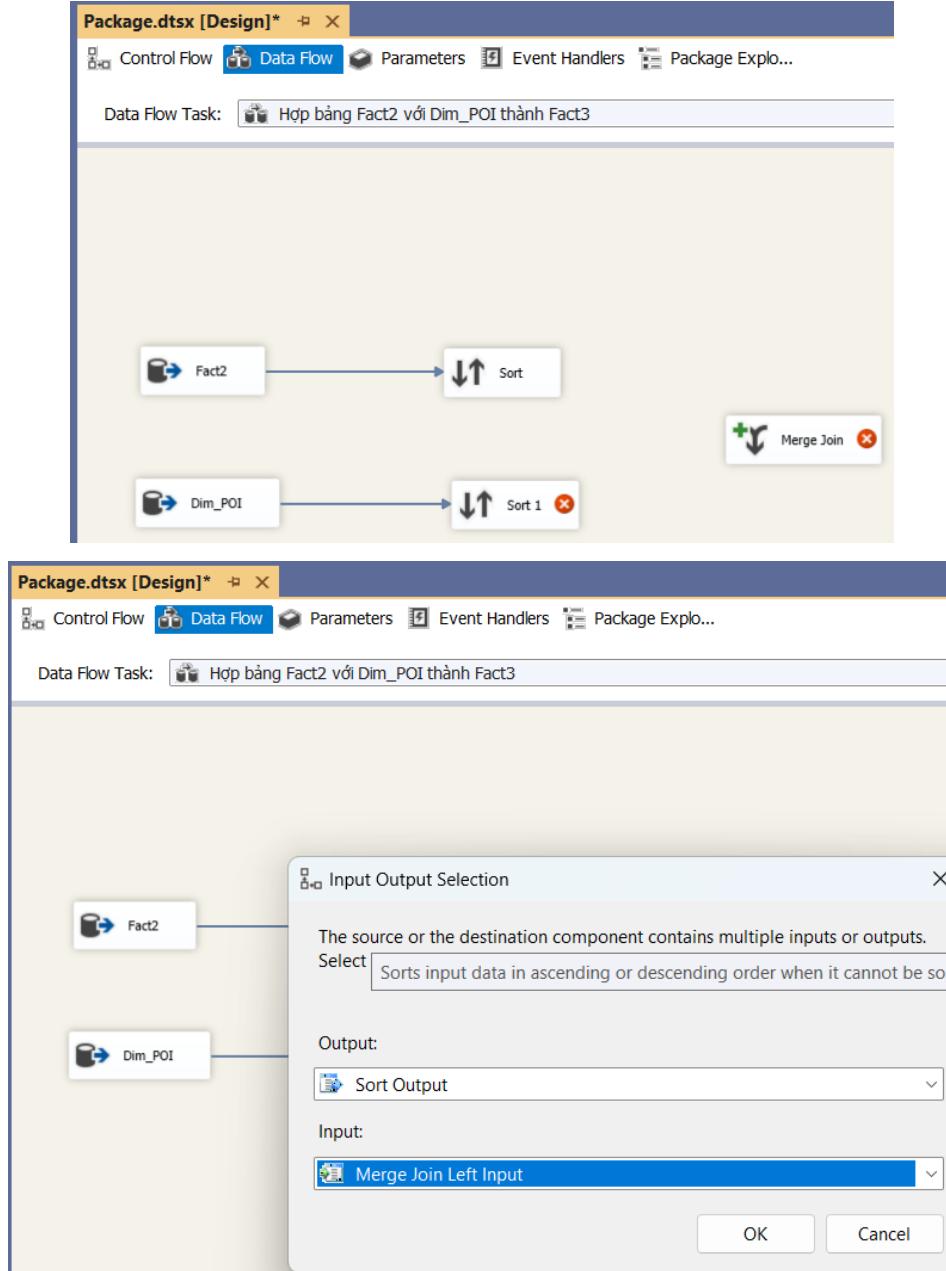


Bước 9: Click chuột phải lên **Sort**, chọn **Edit**. Chọn các cột trong bảng sao cho giống với các cột của bảng **Dim_POI** cả về thứ tự cột để thực hiện quá trình **Merge**

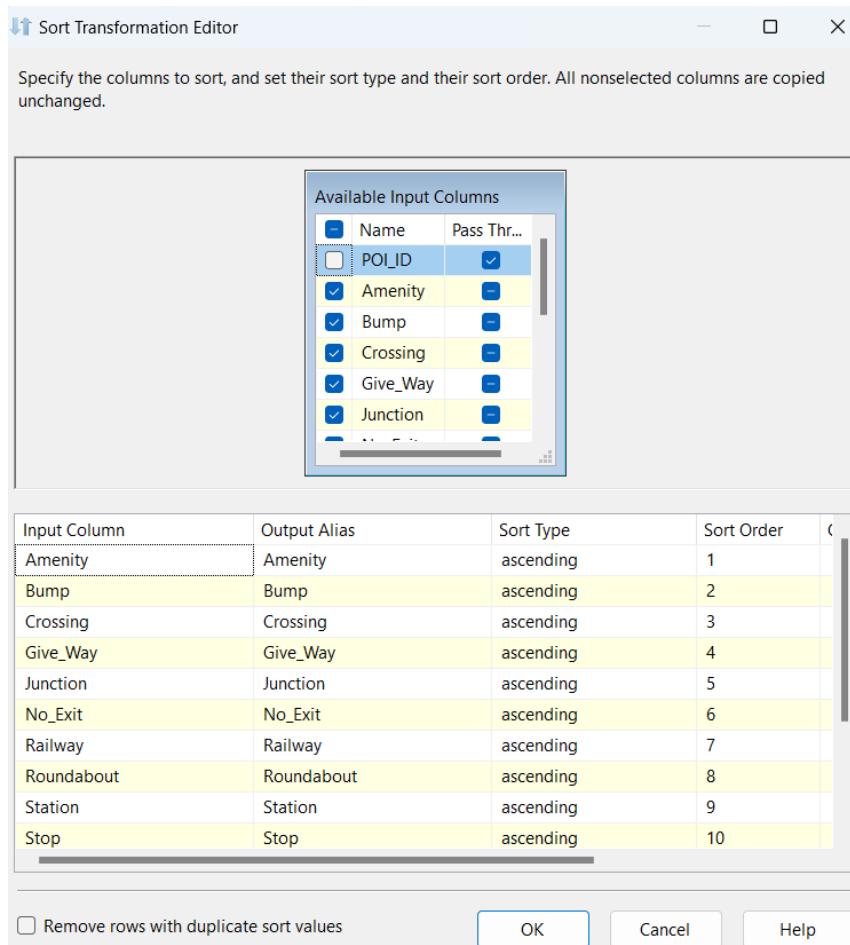


Bước 10: Thêm mới 1 **Merge Join** và nối **Sort** với **Merge Join** được thêm mới đó và chọn **Merge Join Left Input** để giữ lại toàn bộ các dòng trong bảng **Fact2** bất kể có kết quả khi thực hiện phép kết trái với cột **ID** của bảng **Dim_POI** hay không

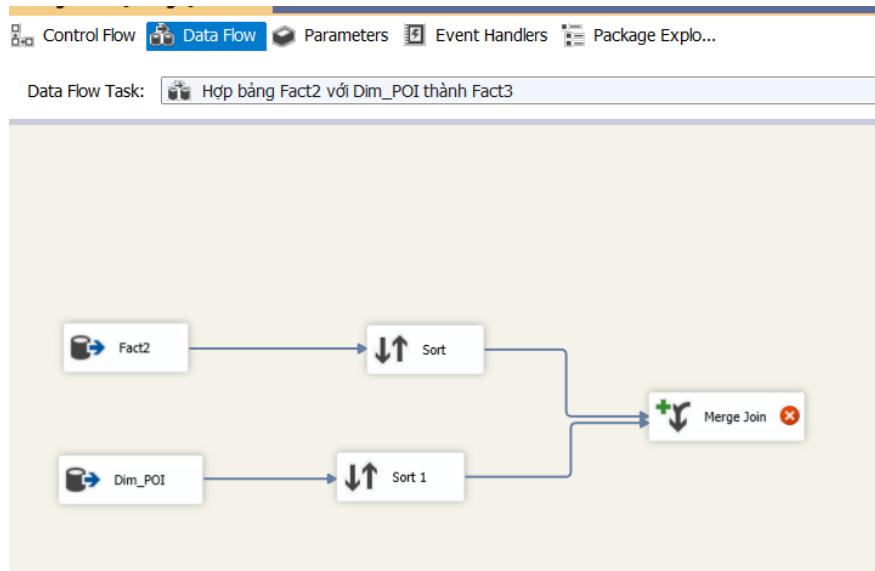
Ths. Đỗ Thị Minh Phụng



Bước 11: Chọn các cột tương ứng cho Sort1 (tương tự như Sort)

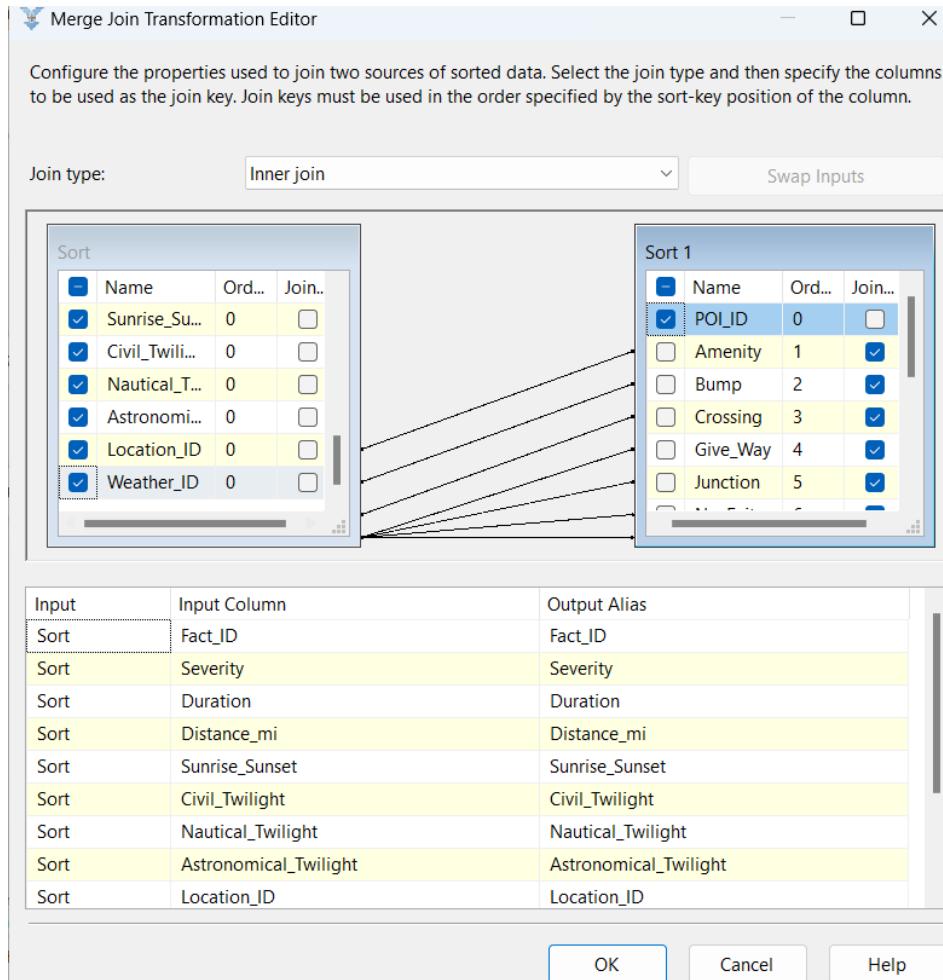


Bước 12: Nối Merge Join với Sort1

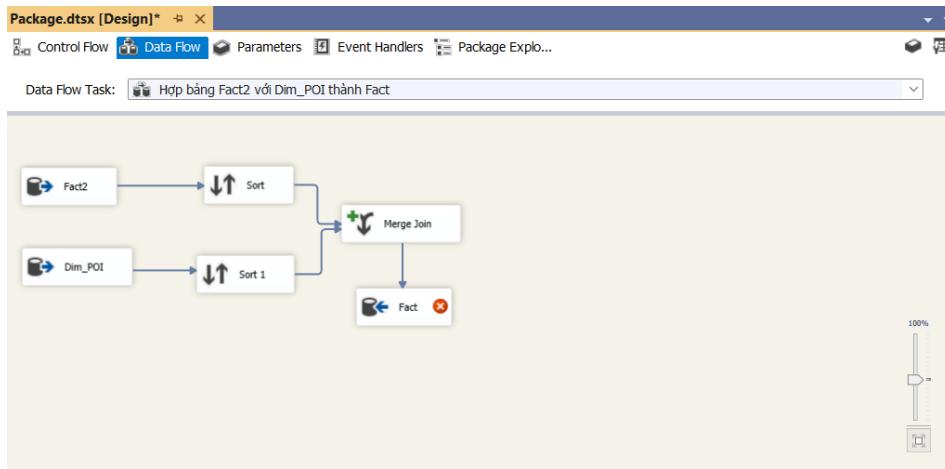


Bước 13:

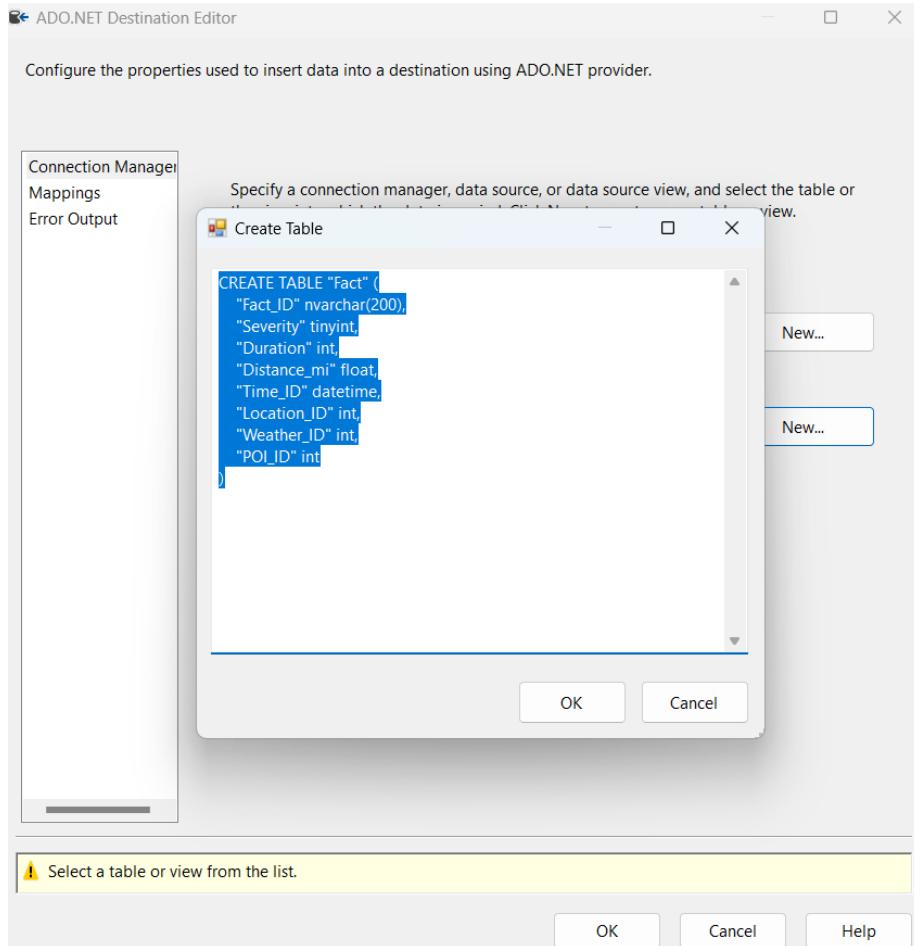
- + Click chuột phải lên Merge Join, chọn **Edit**
- + Tick chọn các cột của Sort nhưng không lấy thuộc tính ở trên vừa làm
- + Chọn **POI_ID** ở Sort1 để tiến hành Merge vào **Fact2**
- + Sau khi **Merge** xong thì bảng **Fact2** không còn các thuộc tính ban nãy nữa mà thay vào đó sẽ có thêm thuộc tính **POI_ID**



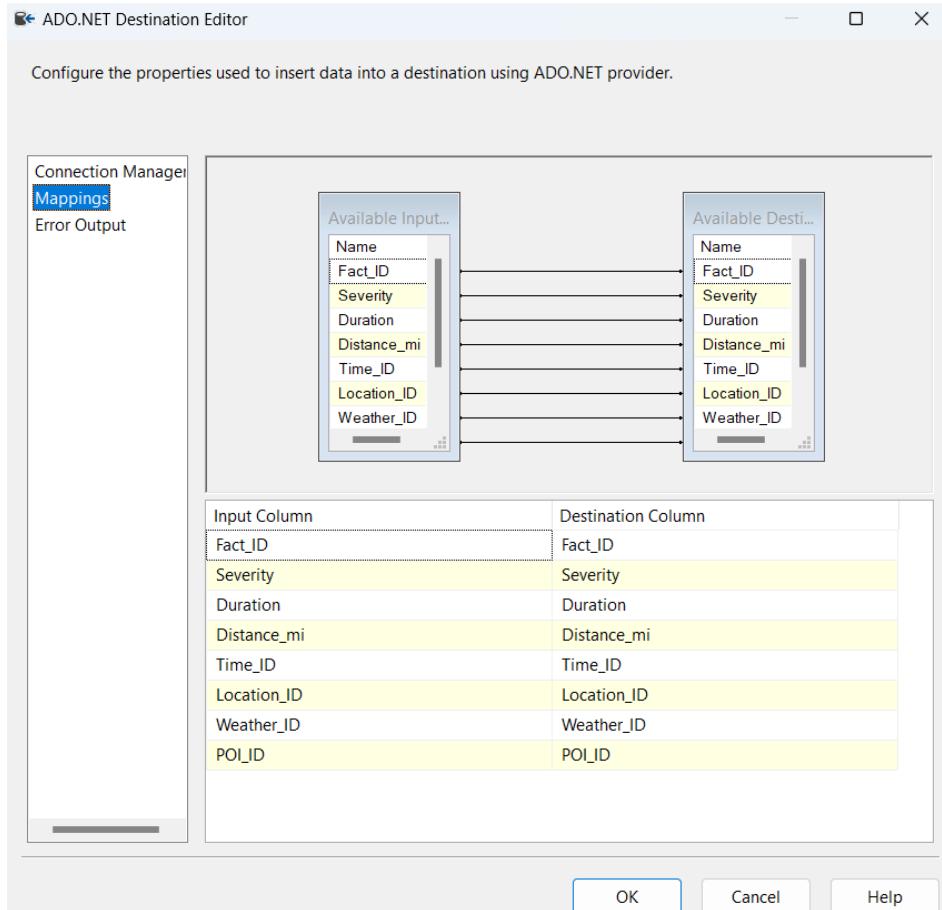
Bước 14: Thêm 1 ADO NET Destination và đổi tên thành Fact để chia toàn bộ những gì đã Merge



Bước 15: Nội dung của câu lệnh SQL để tạo bảng **Fact3** như hình bên dưới

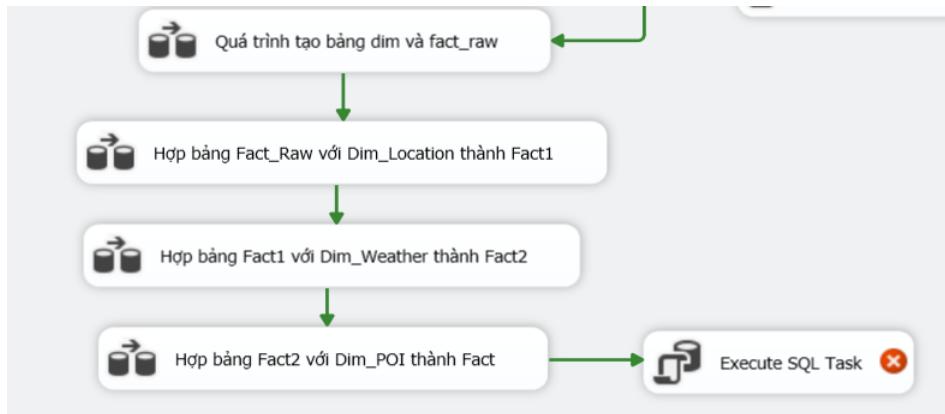


Bước 16: Chuyển sang thẻ **Mappings** để kiểm tra ánh xạ các cột thuộc tính và nhấn **OK**.

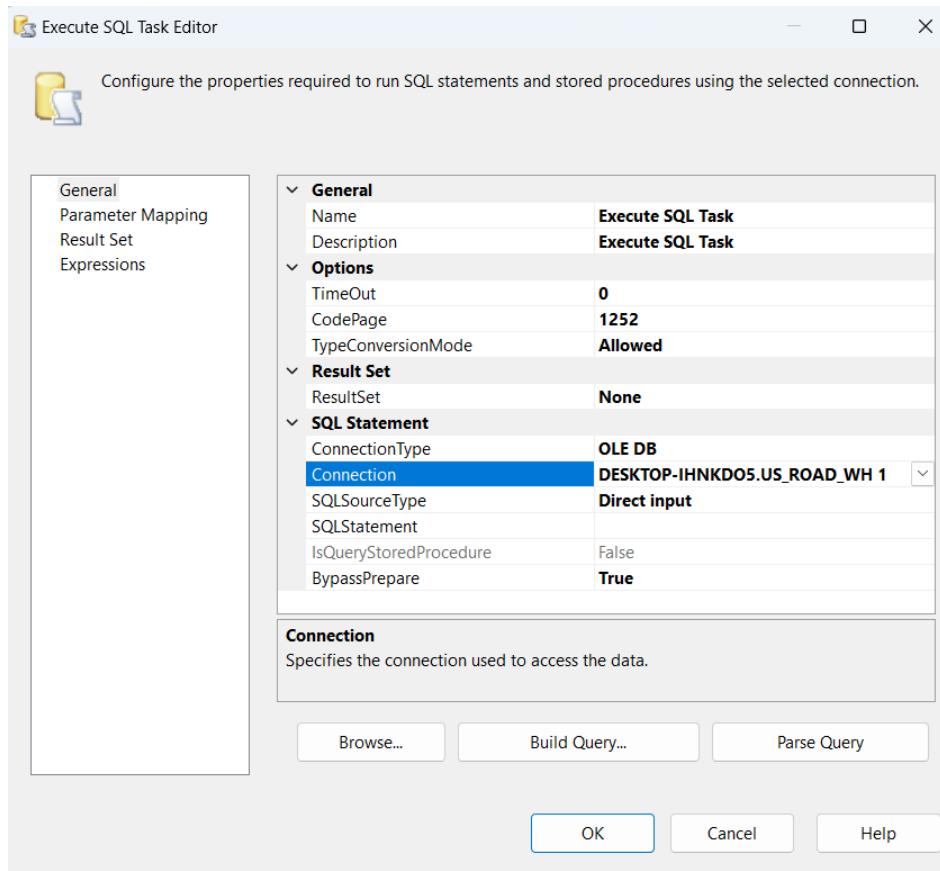


2.7. Tạo khóa ngoại từ bảng Fact đến các Dimension

Bước 1: Trong Control Flow tạo 1 Execute SQL Task để thực thi các câu lệnh SQL tạo các khóa ngoại từ bảng Fact đến các Dimension.



Bước 2: Click chuột phải lên **Execute SQL Task**, chọn **Edit**. Chọn thiết lập kết nối đến **data warehouse** trong ô **connection** của **SQL Server**.

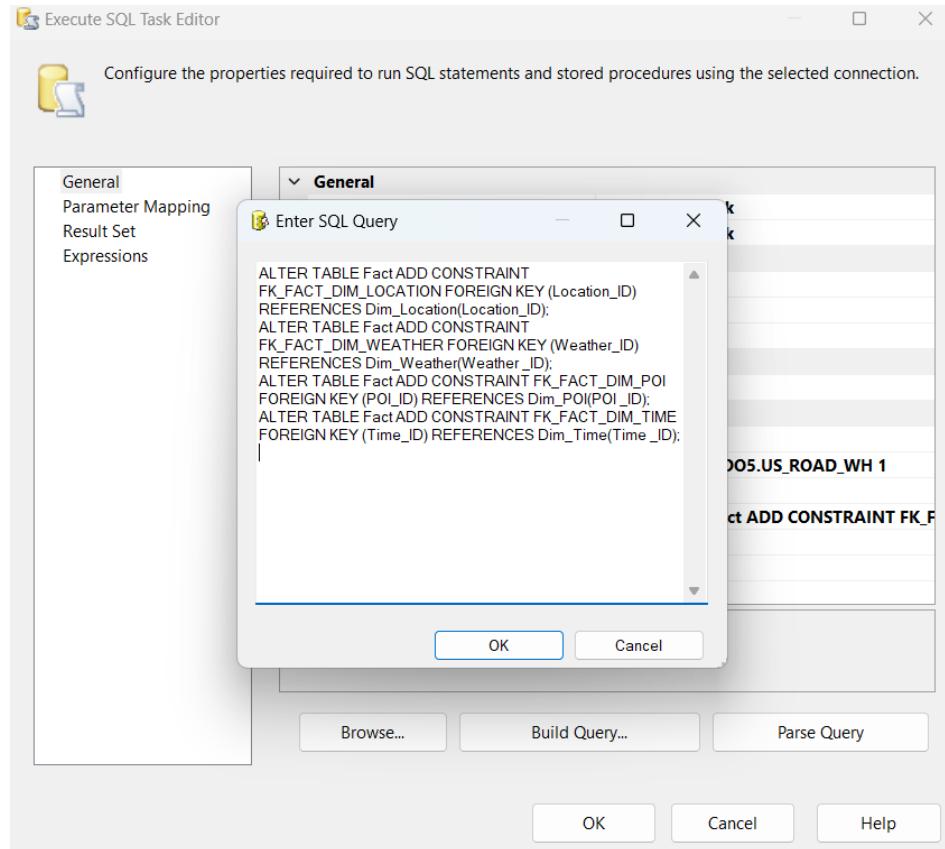


Bước 3: Thêm các câu truy vấn SQL tạo khóa ngoại đến các **Dimension** vào ô **SQL Statement**. Nhấn **OK**.

Nội dung câu truy vấn SQL để tạo khóa ngoại như sau:

```
ALTER TABLE Fact ADD CONSTRAINT FK_FACT_DIM_LOCATION FOREIGN KEY
(Location_ID) REFERENCES Dim_Location(Location_ID);
ALTER TABLE Fact ADD CONSTRAINT FK_FACT_DIM_WEATHER FOREIGN KEY
(Weather_ID) REFERENCES Dim_Weather(Weather_ID);
ALTER TABLE Fact ADD CONSTRAINT FK_FACT_DIM_POI FOREIGN KEY
(POI_ID) REFERENCES Dim_POI(POI_ID);
```

```
ALTER TABLE Fact ADD CONSTRAINT FK_FACT_DIM_TIME FOREIGN KEY  
(Time_ID) REFERENCES Dim_Time(Time_ID);
```

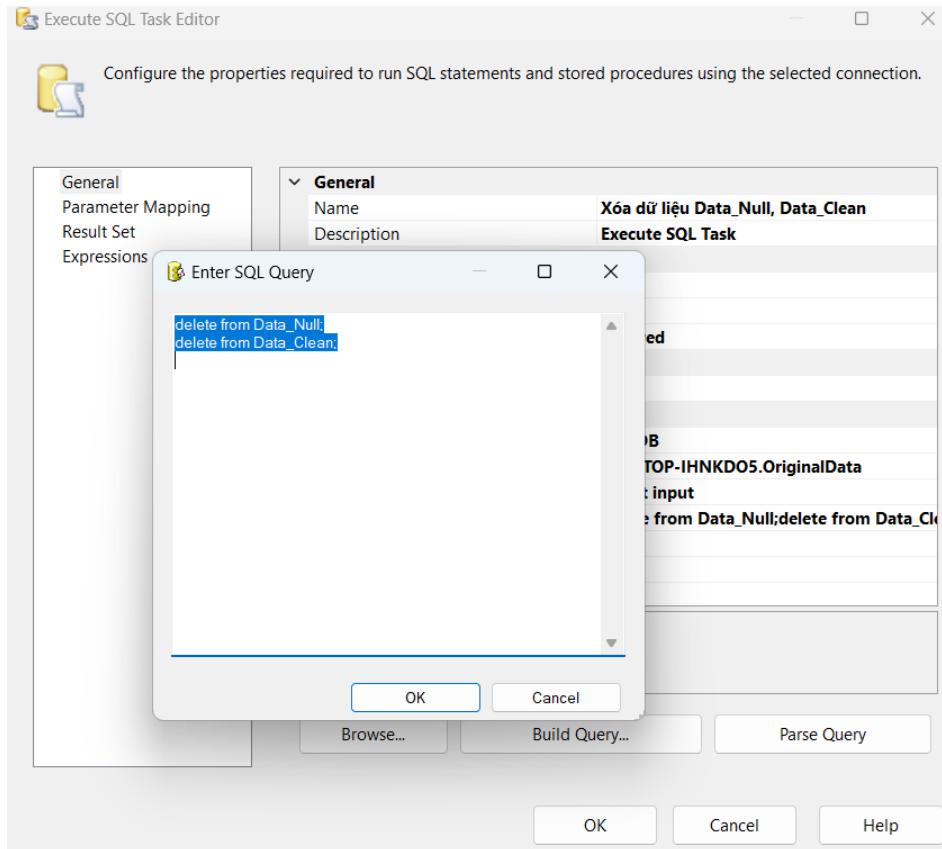


2.8. Chạy dự án SSIS

2.8.1. Khởi chạy SSIS

Bước 1: Thêm vào một **Execute SQL Task** nhằm thực hiện nhiệm vụ xóa các dữ liệu Null và dữ liệu đã được **Clean** trong quá trình chạy SSIS trước đó. Các câu lệnh trong **Execute SQL Task** này là:

```
delete from DataNull;  
delete from Data_Clean;
```



Bước 2: Thêm vào một **Execute SQL Task** nhằm thực hiện nhiệm vụ đảm bảo dữ liệu mới hoàn toàn (không bị chồng chéo dữ liệu cũ) mỗi khi chạy project, trước quá trình tạo bảng **Dimension** và **Fact_Row**. Các câu lệnh trong **Execute SQL Task** này là:

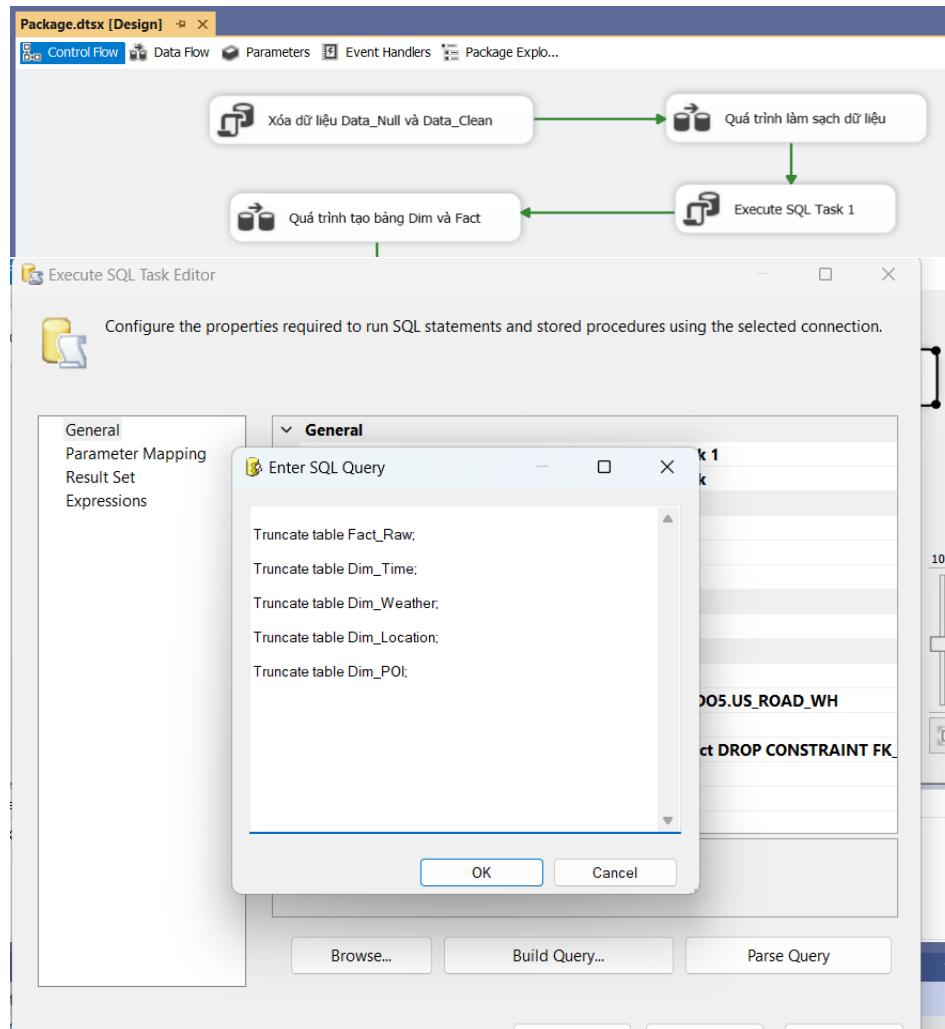
Truncate table Dim_Time;

Truncate table Dim_Weather;

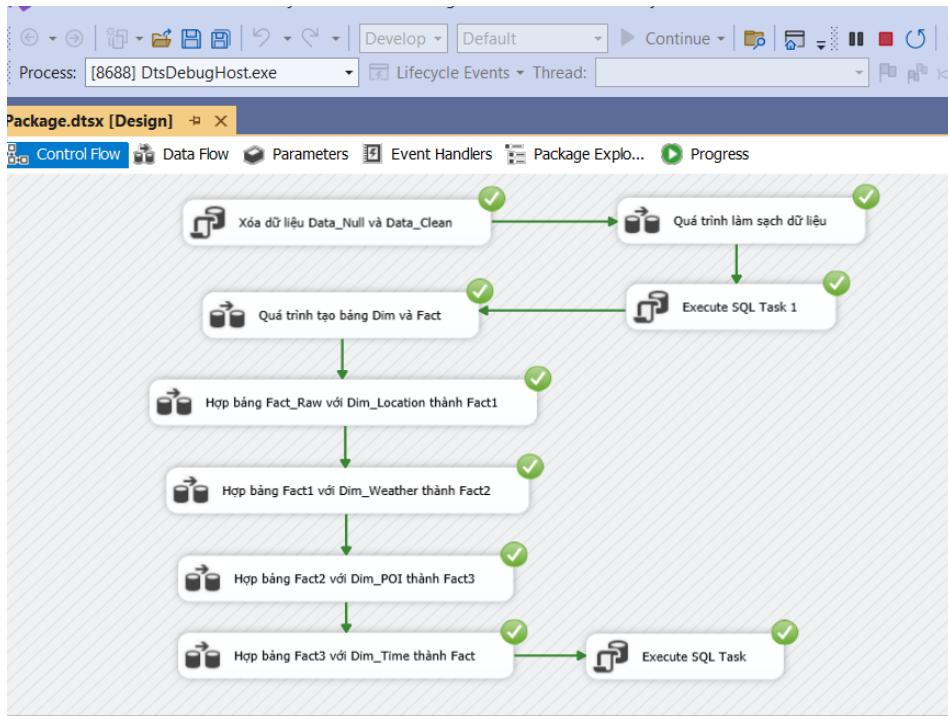
Truncate table Dim_Location;

Truncate table Dim_POI;

Truncate table Fact_Raw;



Bước 3: Nhấn nút Start trên thanh menu để tiến hành chạy project. Đây là kết quả:



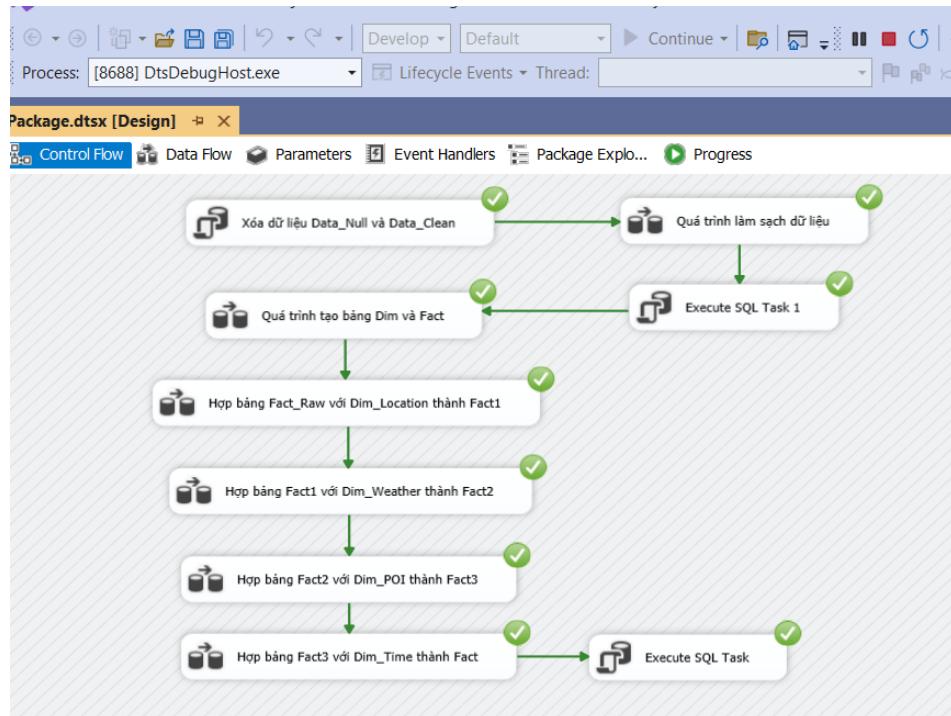
Bước 4: Sau khi đã chạy project lần đầu thành công, khóa ngoại của bảng **Fact** tham chiếu đến các **Dimension** đã được tạo. Ta tiến hành thêm các lệnh SQL để xóa và thêm lại các khóa ngoại vào mỗi lần khởi chạy lại project tiếp theo. Các câu lệnh trong **Execute SQL** Task viết lại này là:

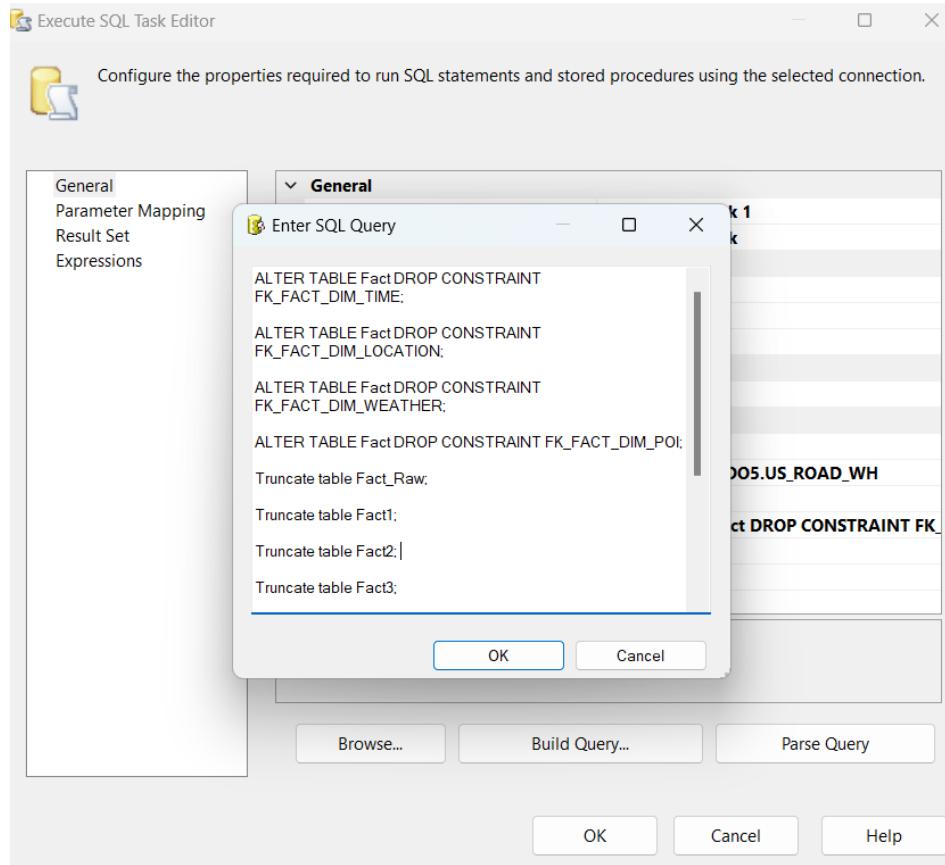
```

ALTER TABLE Fact DROP CONSTRAINT FK_FACT_DIM_TIME;
ALTER TABLE Fact DROP CONSTRAINT FK_FACT_DIM_LOCATION;
ALTER TABLE Fact DROP CONSTRAINT FK_FACT_DIM_WEATHER;
ALTER TABLE Fact DROP CONSTRAINT FK_FACT_DIM_POI;
Truncate table Fact_Raw;
Truncate table Fact1;
Truncate table Fact2;
Truncate table Fact;
Truncate table Dim_Time;
  
```

```
Truncate table Dim_Weather;  
Truncate table Dim_Location;  
Truncate table Dim_POI;
```

Bước 5: Tiến hành chạy lại project.





2.8.2. Kiểm tra dữ liệu các bảng:

2.8.2.1. Kiểm tra dữ liệu bảng Dim_Time:

Ths. Đỗ Thị Minh Phụng

	Time_ID	Start_day	Start_month	Start_year	Start_hour	Start_minute	End_day	End_month	End_year	End_hour	End_minute	Sunrise_Sunset
1	1	1	1	2017	0	1	31	1	2017	23	59	Night
2	2	1	1	2017	0	1	27	5	2017	23	59	Night
3	3	1	1	2017	0	1	30	7	2017	23	59	Night
4	4	1	1	2017	0	1	1	10	2017	23	59	Night
5	5	1	1	2017	0	1	1	12	2017	23	59	Night
6	6	1	1	2017	0	1	31	12	2017	23	59	Night
7	7	1	1	2017	0	1	1	6	2018	23	59	Night
8	8	1	1	2017	1	1	1	1	2017	7	1	Night
9	9	1	1	2017	2	1	1	1	2017	8	1	Night
10	10	1	1	2017	3	1	1	7	2017	2	59	Night
11	11	1	1	2017	6	0	31	10	2018	18	0	Night
12	12	1	1	2017	6	0	1	5	2017	6	0	Night
13	13	1	1	2017	6	42	8	2	2017	23	59	Night
14	14	1	1	2017	8	42	30	6	2017	16	0	Day
15	15	1	1	2017	8	42	2	1	2017	20	42	Day
16	16	1	1	2017	11	45	3	1	2017	13	1	Day
17	17	1	1	2017	11	45	6	1	2017	12	0	Day

2.8.2.2. Kiểm tra dữ liệu bảng Dim_Location:

	Location_ID	Street	Side	City	County	State
1	1	00-199 Yerger Rd	R	Fair Lawn	Bergen	NJ
2	2	0N999 E County Line Rd	R	Maple Park	Kane	IL
3	3	1/2 E River St	L	Ilion	Herkimer	NY
4	4	1/2 Federal St	R	Salem	Essex	MA
5	5	1/2 Henry St	L	Brooklyn	Kings	NY
6	6	1/2 Mile Rd	R	Albion	Calhoun	MI
7	7	1/2 Mile Rd	R	Caledonia	Racine	WI
8	8	1/2 Vanderbilt Ave	L	Brooklyn	Kings	NY
9	9	10 Mile Rd	R	South Lyon	Oakland	MI
10	10	1000N	R	Bridgeport	Lawrence	IL
11	11	100th Ave	L	Hartley	O'Brien	IA
12	12	100th St	L	Oelwein	Fayette	IA
13	13	100th St	L	Union	Putnam	MO
14	14	100th St	L	Urbandale	Polk	IA
15	15	100th St	R	Clinton	Rock	MN
16	16	100th St	R	Dows	Franklin	IA

2.8.2.3. Kiểm tra dữ liệu bảng Dim_Weather:

100 % ▾

Results Messages

	Weather_ID	Temperature_F	Humidity	Wind_Speed_mph	Weather_Condition
1	1	-27	62	9.19999980926514	Clear
2	2	-27	62	13.8000001907349	Haze
3	3	-25.6000003814697	69	8.10000038146973	Clear
4	4	-23.7999992370605	52	12.6999998092651	Clear
5	5	-22	69	6.90000009536743	Clear
6	6	-22	70	11.5	Clear
7	7	-22	77	6.90000009536743	Clear
8	8	-21.7999992370605	75	6.90000009536743	Clear
9	9	-21.5	64	4.59999990463257	Clear
10	10	-20.8999996185303	54	13.8000001907349	Overcast
11	11	-20.8999996185303	66	8.10000038146973	Clear
12	12	-20.2000007629395	52	16.1000003814697	Overcast
13	13	-20.2000007629395	52	25.2999992370605	Partly Cloudy
14	14	-20.2000007629395	60	12.6999998092651	Clear
15	15	-20	60	16.1000003814697	Clear
16	16	-20	74	4.59999990463257	Clear
17	17	-20	79	9.10000038146072	Clear

2.8.2.4. Kiểm tra dữ liệu bảng Dim_POI:

Results Messages

	POI_ID	Amenity	Bump	Crossing	Give_Way	Junction	No_Exit	Railway	Roundabout	Station	Stop	Traffic_Calming	Traffic_Signal	Turning_Loop
1	1	0	0	0	0	0	0	0	0	0	0	0	0	False
2	2	0	0	0	0	0	0	0	0	0	0	0	1	False
3	3	0	0	0	0	0	0	0	0	0	0	1	0	False
4	4	0	0	0	0	0	0	0	0	0	0	1	1	False
5	5	0	0	0	0	0	0	0	0	0	1	0	0	False
6	6	0	0	0	0	0	0	0	0	0	1	0	1	False
7	7	0	0	0	0	0	0	0	0	0	1	1	0	False
8	8	0	0	0	0	0	0	0	0	1	0	0	0	False
9	9	0	0	0	0	0	0	0	0	1	0	0	1	False
10	10	0	0	0	0	0	0	0	0	1	1	0	0	False
11	11	0	0	0	0	0	0	0	0	1	1	0	1	False
12	12	0	0	0	0	0	0	0	0	1	1	1	1	False
13	13	0	0	0	0	0	0	0	1	0	0	0	0	False
14	14	0	0	0	0	0	0	0	1	0	1	0	0	False
15	15	0	0	0	0	0	0	1	0	0	0	0	0	False
16	16	0	0	0	0	0	0	1	0	0	0	0	1	False
17	17	0	0	0	0	0	0	1	0	0	1	0	0	False

2.8.2.5. Kiểm tra dữ liệu bảng Fact:

Ths. Đỗ Thị Minh Phụng

	Fact_ID	Severity	Duration	Distance_mi	Location_ID	Weather_ID	POI_ID	Time_ID
1	C-2403303	2	291666	0.483494132757187	7771	189399	1	3110
2	C-3626978	2	6523200	23.3500003814697	24478	23369	1	3110
3	C-2596386	2	506467	4.94165992736816	21027	44845	1	3110
4	C-4717573	3	18000	1.89999997615814	18819	63343	1	3110
5	C-1581356	2	345572	0.211642786860466	19963	46756	1	3110
6	C-6094458	2	345698	0.203743353486061	24949	31787	1	3110
7	C-1676864	2	46129	0.00348701374605298	9300	2593	1	3110
8	C-1119607	2	604819	0.174397230148315	2946	29826	1	3110
9	C-563122	2	345557	0.318870782852173	38899	61244	1	3110
10	C-5036694	2	14400	0.0755253583192825	15196	58656	1	3110
11	C-4052358	4	345581	0.435437053442001	23301	146648	1	3110
12	C-5418317	2	11527200	0.793246984481812	31630	51176	1	3110
13	C-4665520	2	345622	0.0545668713748455	1777	11670	1	3110
14	C-5403576	2	37323	0.165343135595322	15581	119018	1	3110
15	C-5144701	2	43200	0.0395670831203461	15354	100598	1	3110
16	C-3165276	2	298800	0.110209450125694	13741	97286	1	3110
17	C-760272	2	22400	5.61000099556009	51212	110122	1	3110

CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU

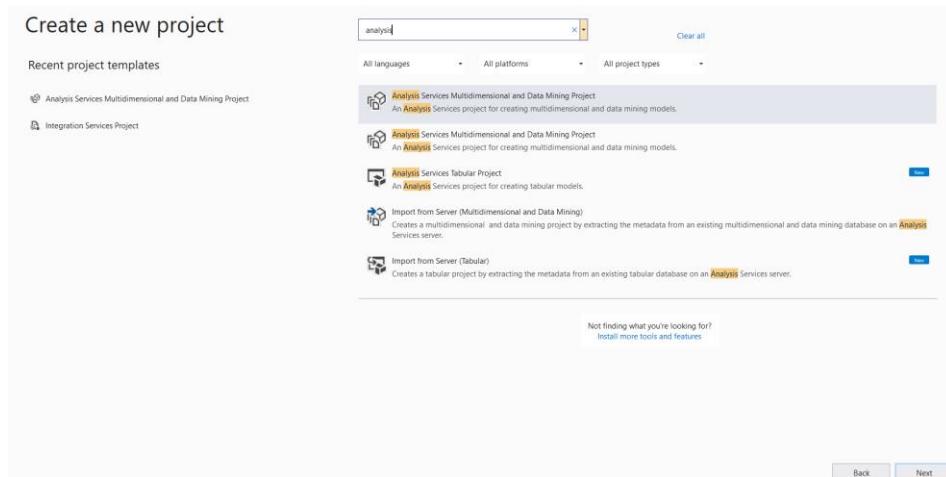
3.1. Chuẩn bị các công cụ

Để tiến hành thực hiện quá trình SSAS, cần phải sử dụng các phần mềm và công cụ sau:

- Microsoft SQL Server đã cài Analysis Services.
- Microsoft Analysis Services Projects

3.2. Tạo mới Project SSAS

Bước 1: Tạo 1 project mới trên **Visual Studio** và tìm kiếm **Analysis Services Multidimensional and Data Mining Project**



Bước 2: Đặt tên cho Project và nhấn **Create**

Configure your new project

Analysis Services Multidimensional and Data Mining Project

Project name

OLAP_SSSA_Team39

Location

D:\VS repo



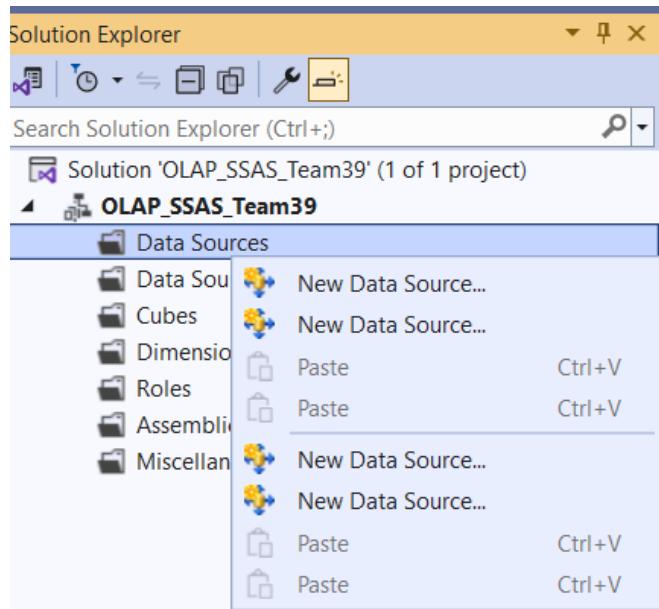
Solution name [\(i\)](#)

OLAP_SSSA_Team39

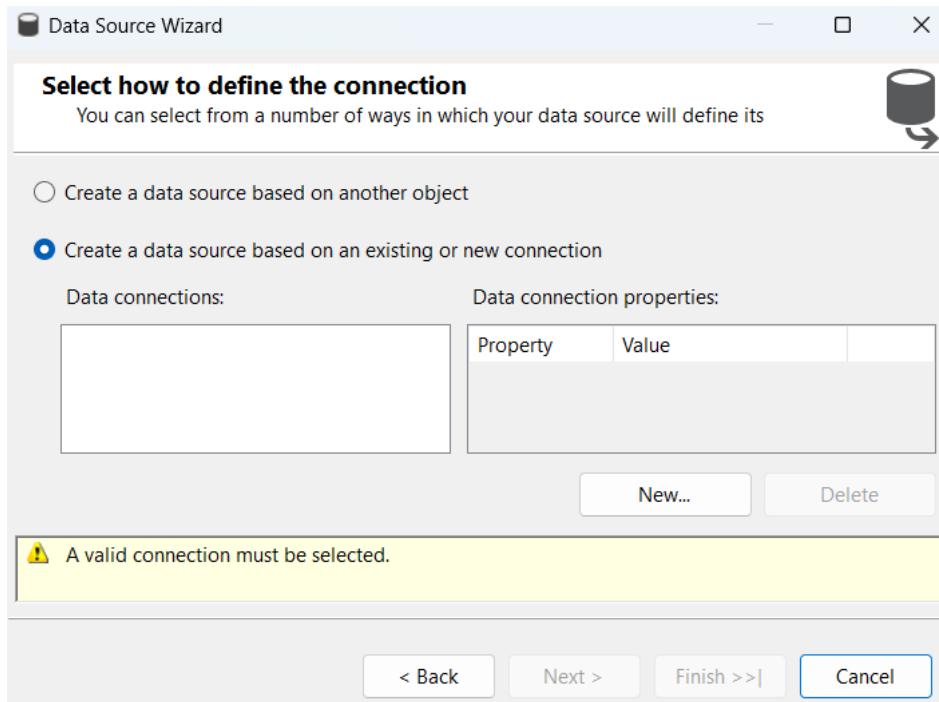
Place solution and project in the same directory

3.3. Xác định dữ liệu nguồn (Data Sources)

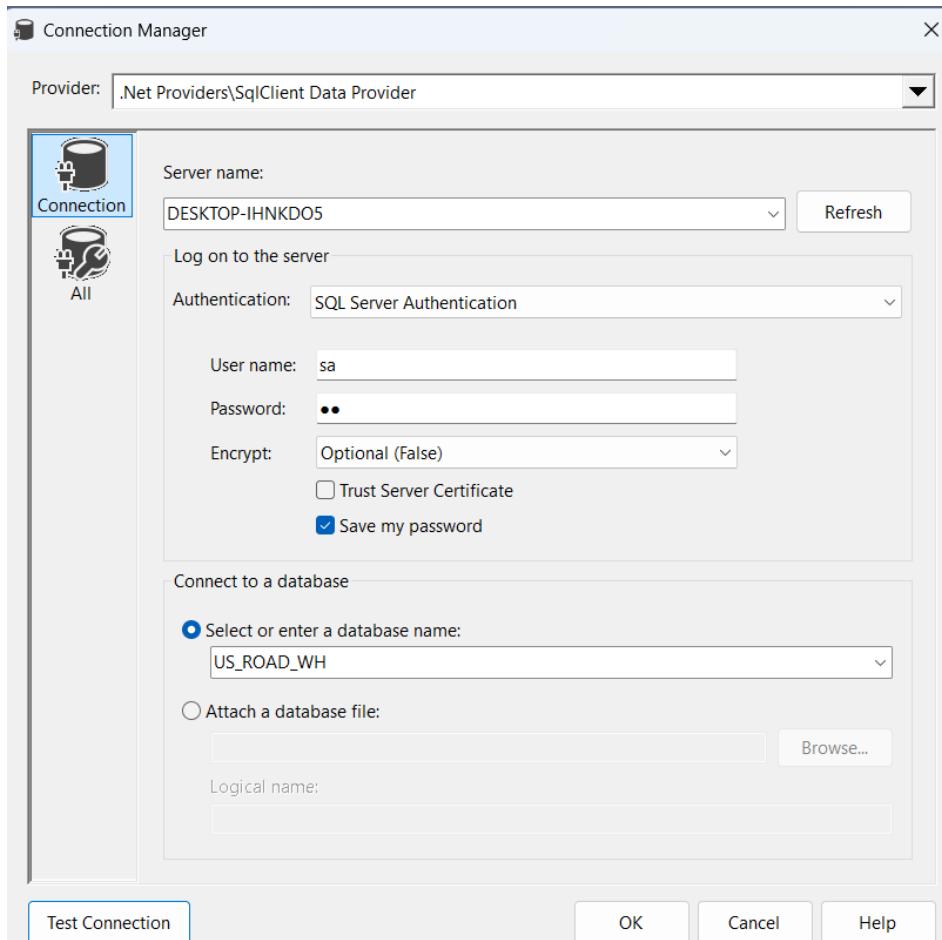
Bước 1: Click chuột phải vào **Data Sources** ở cửa sổ **Solution**, chọn **New Data Sources** và nhấn **Next**



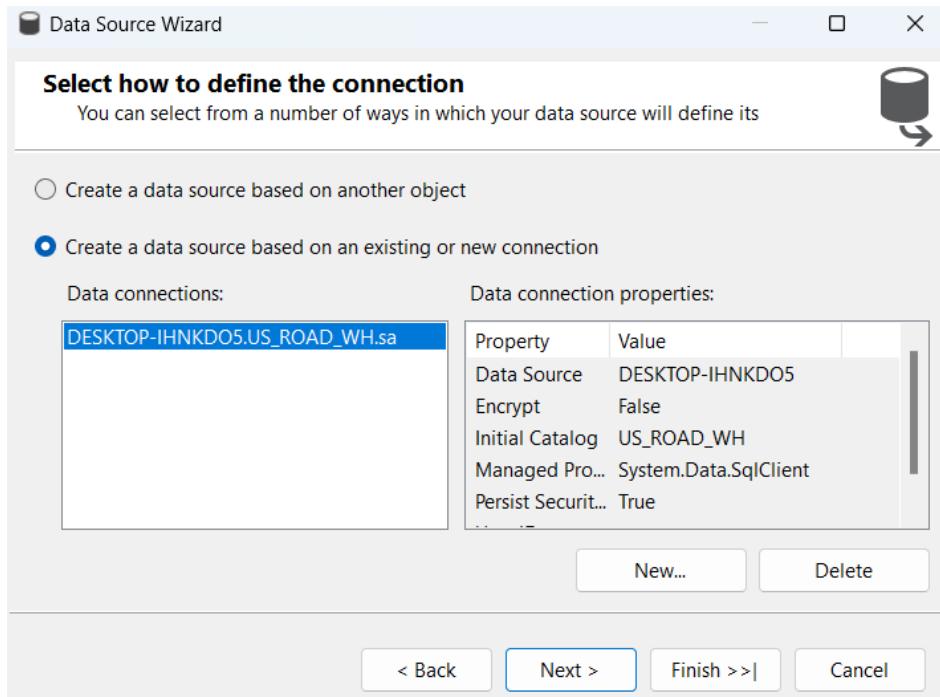
Bước 2: Click chọn “Create a data source based on an existing or new connection” và chọn New...



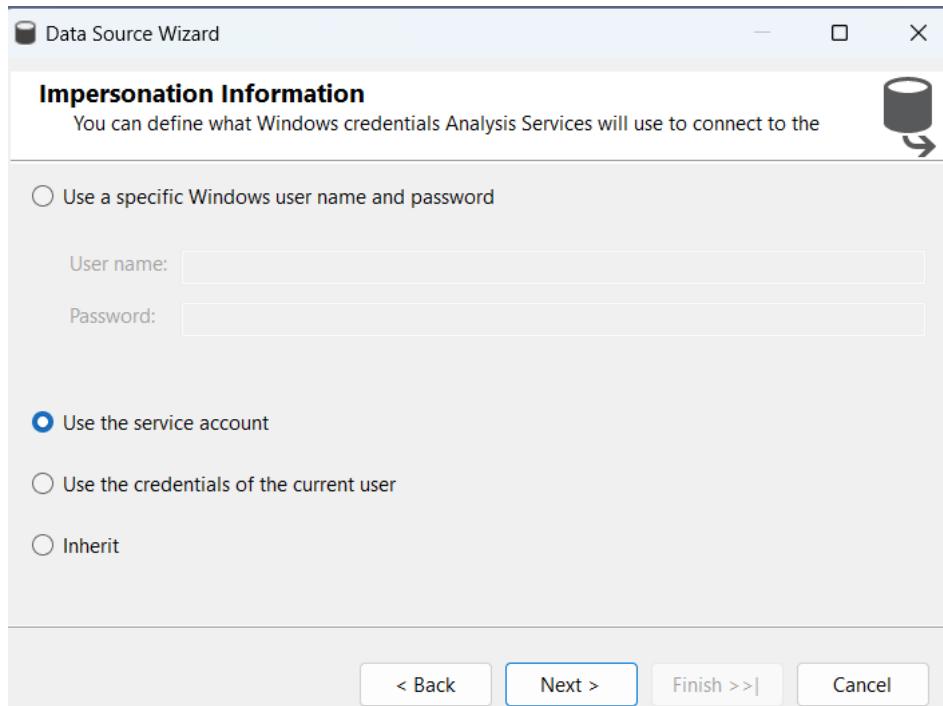
Bước 3: Chọn cơ sở dữ liệu của quá trình SSIS từ trước và nhấn **OK**.



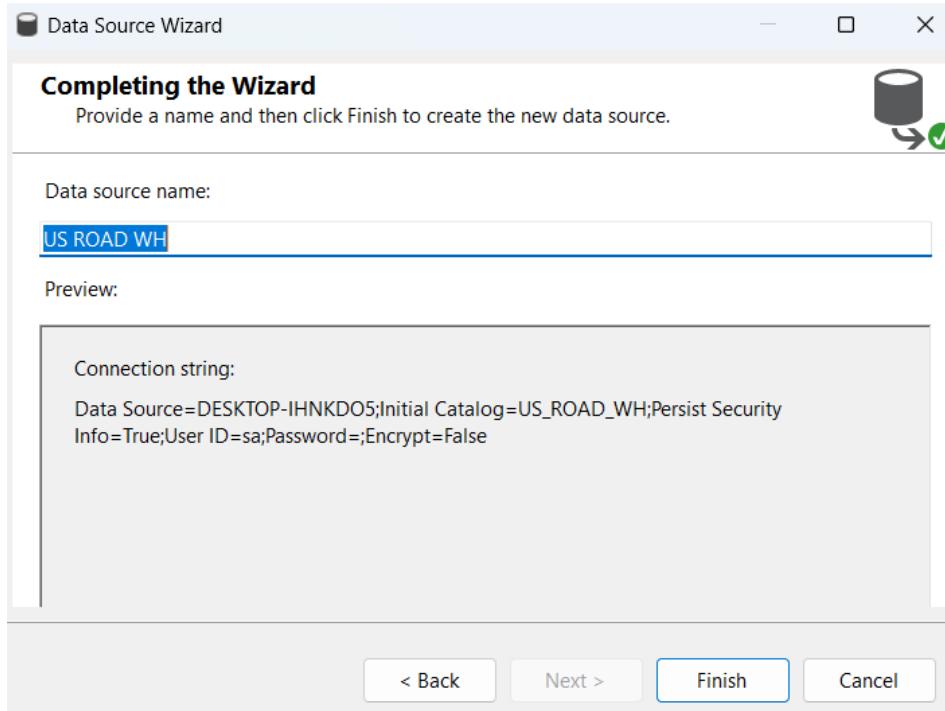
Chọn **Data source** mới tạo và nhấn **Next**



Bước 4: Nhấn “Use the service account”, và nhấn Next

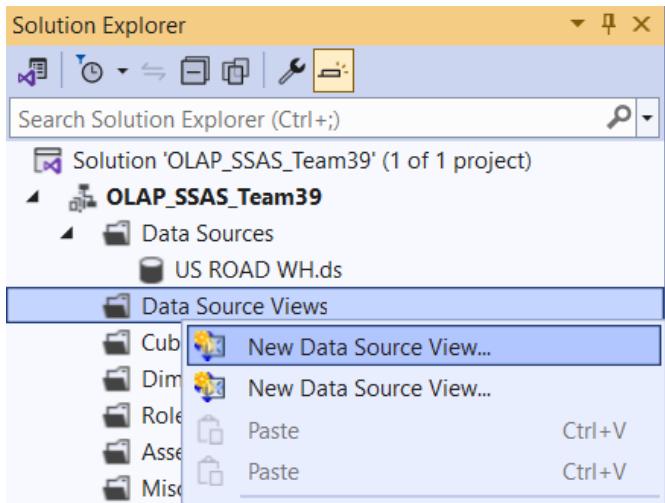


Bước 5: Nhấn **Finish** để hoàn tất quá trình định nghĩa/xác định nguồn dữ liệu



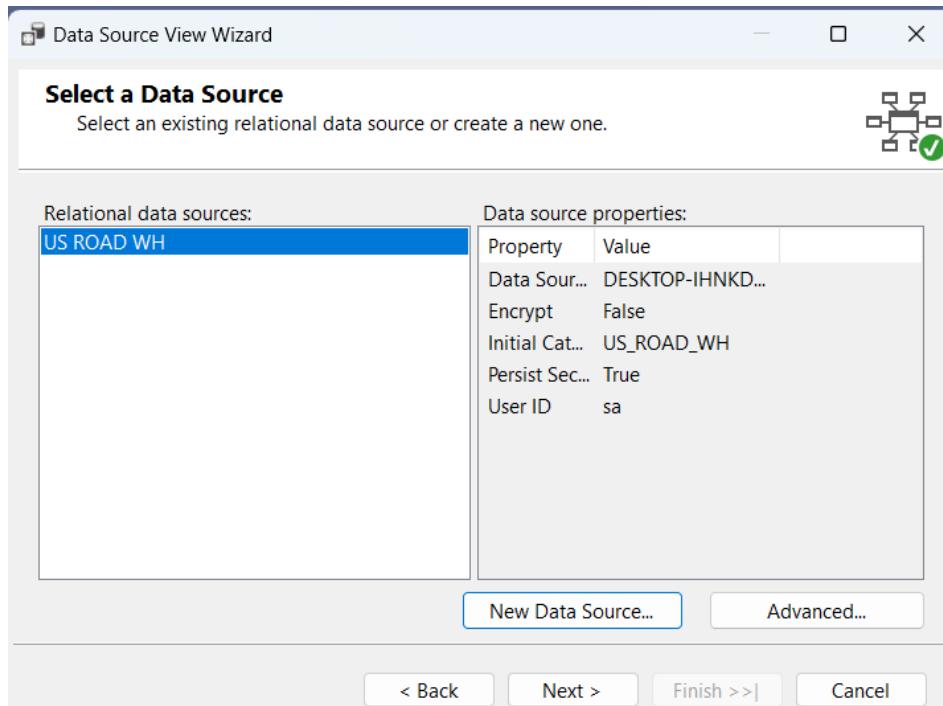
3.4. Xác định Data Source Views

Bước 1: Click chuột phải vào thư mục **Data Source Views** trong cửa sổ **Solution** và trong folder **Data Sources** và chọn **New Data Source Views...**

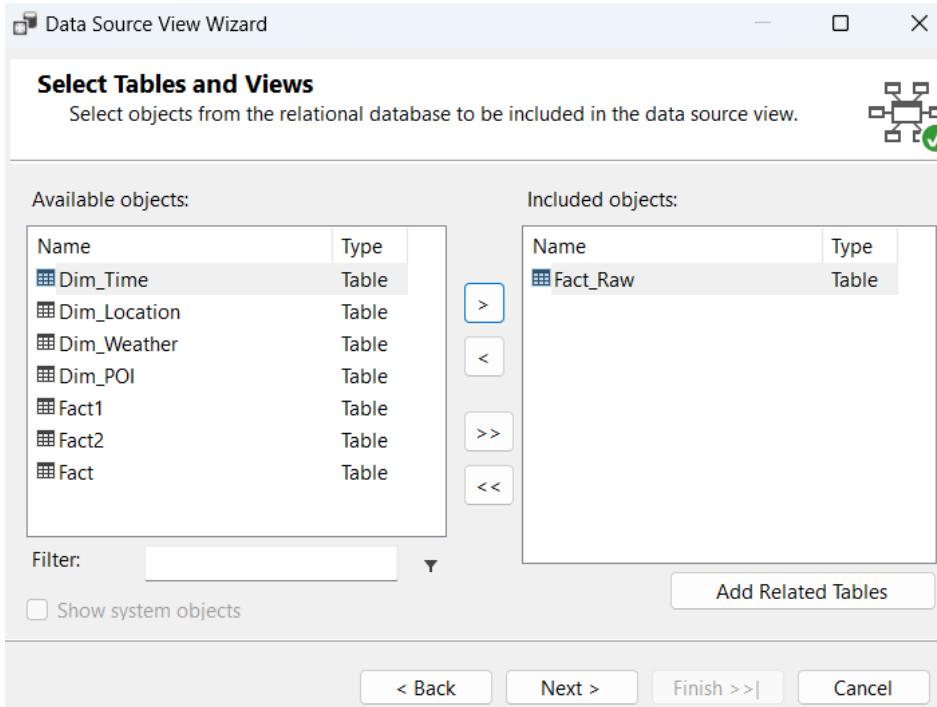


Nhấn Next để tiếp tục

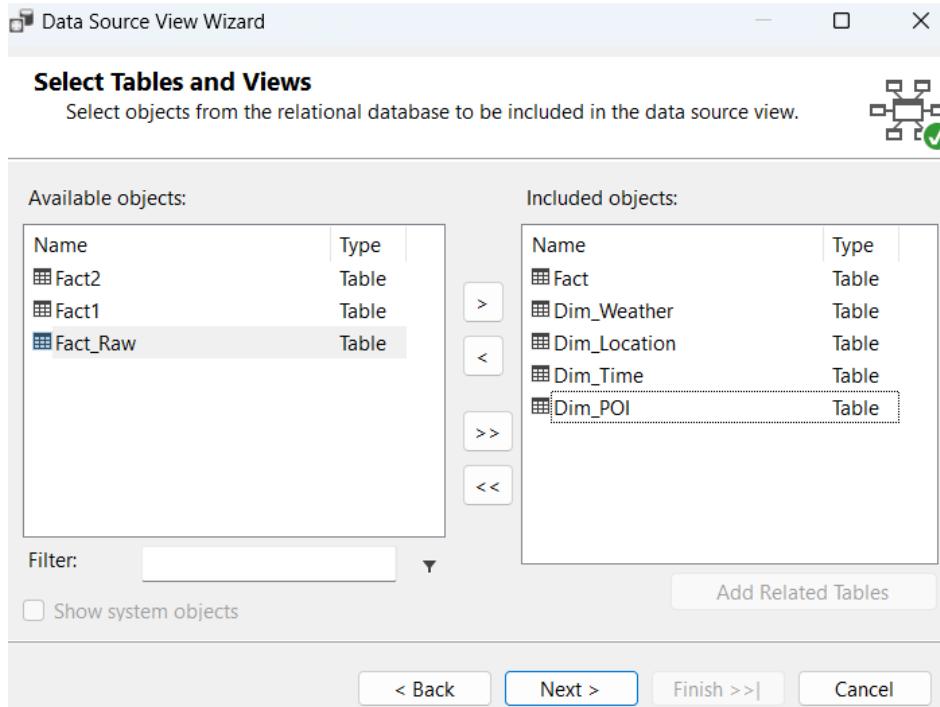
Bước 2: Chọn nguồn vừa tạo và nhấn Next



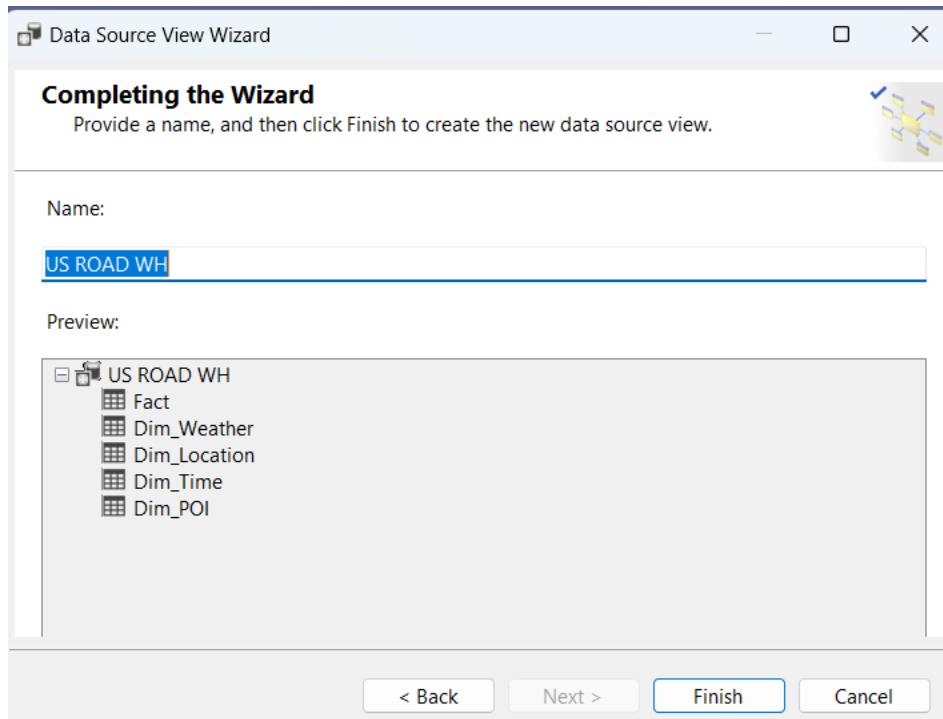
Bước 3: Thêm bảng Fact vào khung Included Objects từ khung Available objects



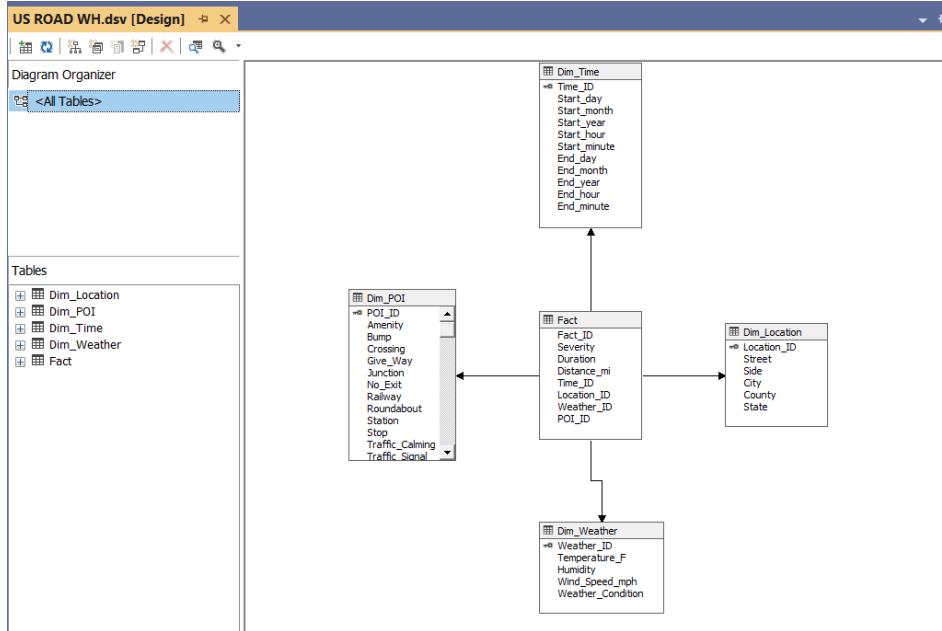
Bước 4: Nhấn **Add Related Tables** để thêm những bảng liên quan với bảng **Fact** và nhấn **Next**.



Bước 5: Nhấn **Finish** để hoàn tất

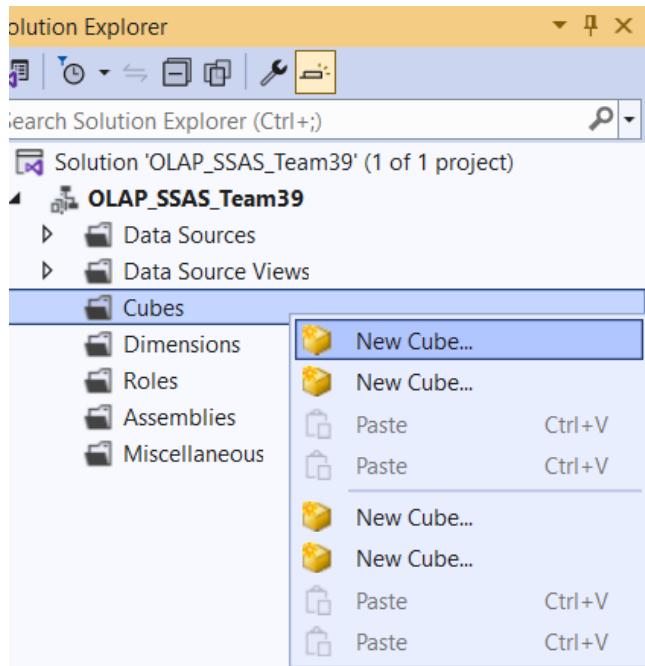


Màn hình sẽ hiển thị **Data Source View** sau khi nhấn **Finish**

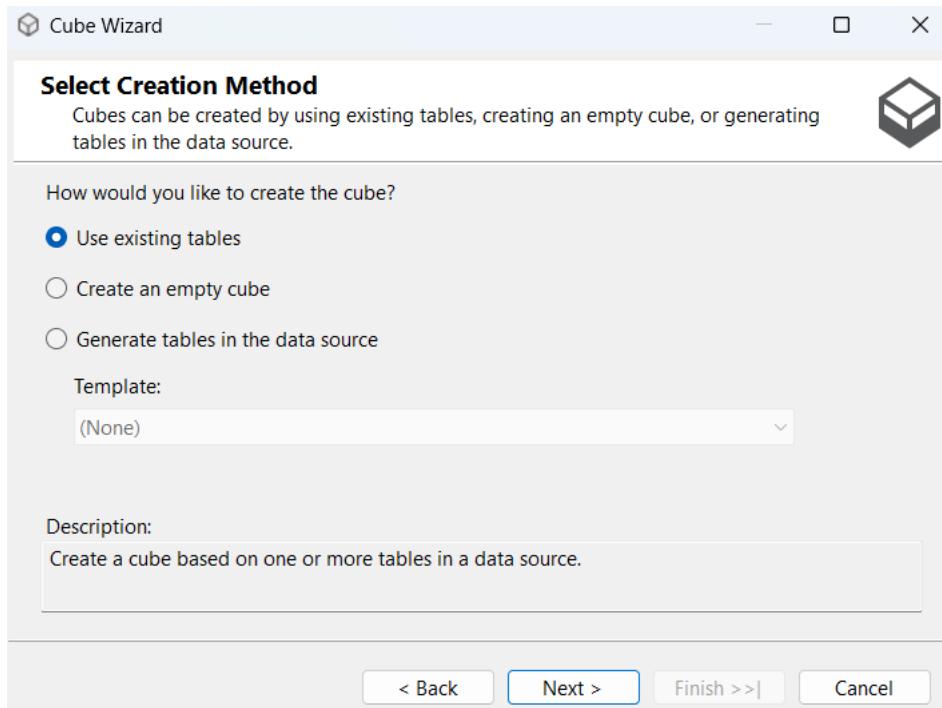


3.5. Xây dựng khối (Cubes) và xác định các độ đo (Measures)

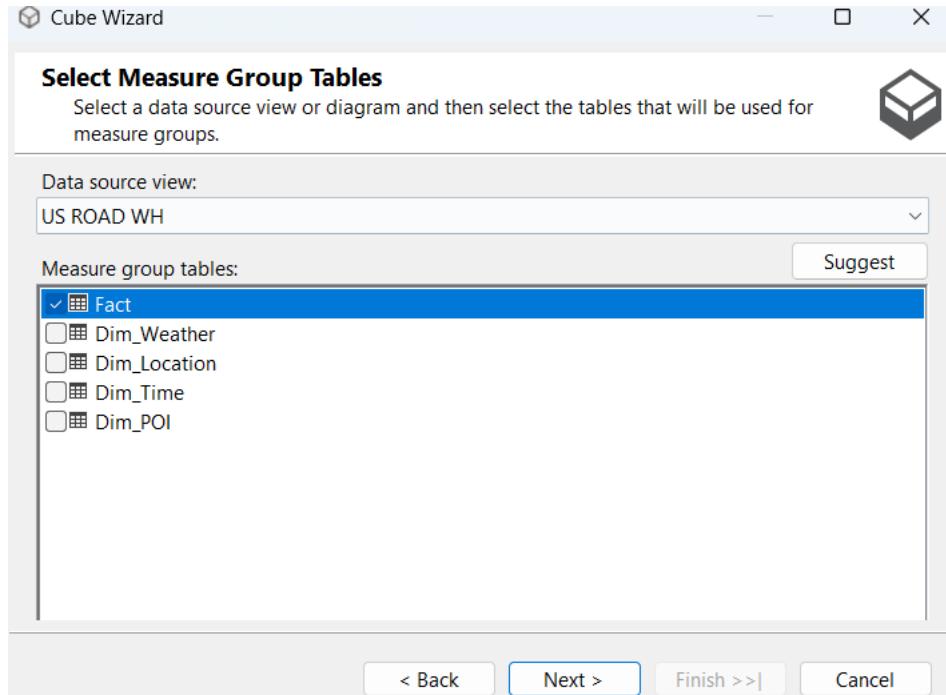
Bước 1: Chuột phải vào **Cubes** trong Folder **Data Source Views** của cửa sổ **Solution**, chọn **New Cube...** và nhấn **Next** để tiếp tục



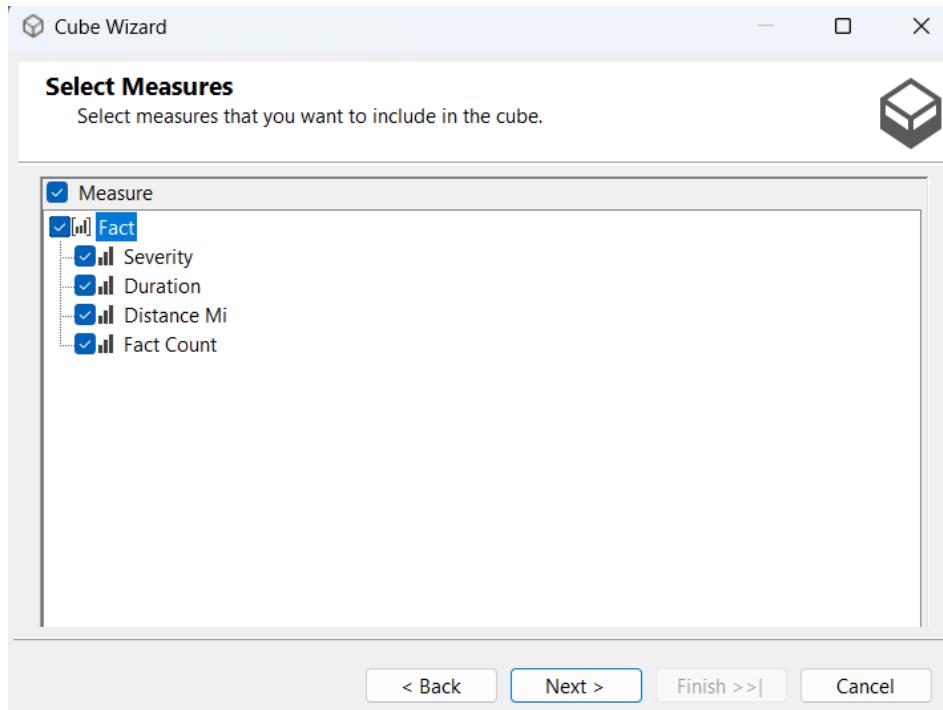
Bước 2: Tick vào ô **Use existing tables** và nhấn **Next**



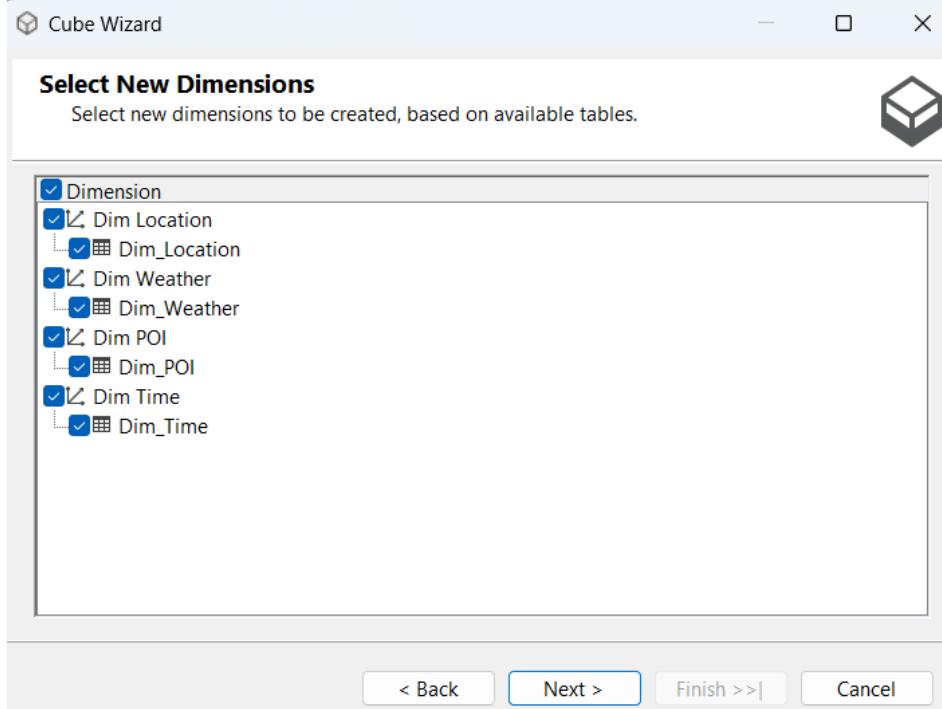
Bước 3: Click vào **Fact** và nhấn **Next** để phân chia các nhóm độ đo **Measure group**



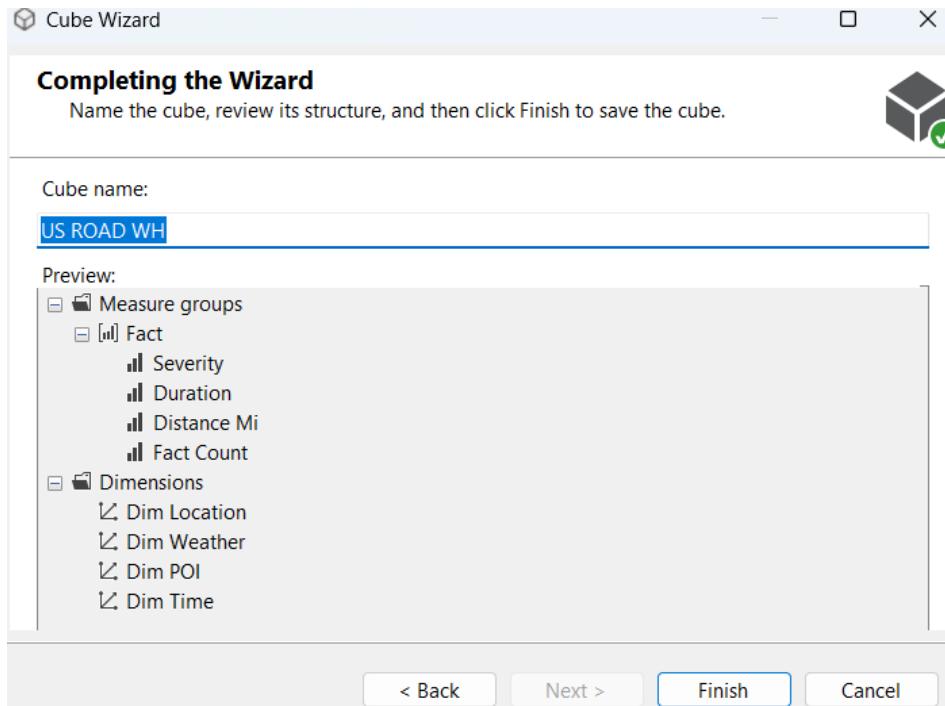
Bước 4: Chọn những thuộc tính độ đo được gợi ý và nhấn **Next**



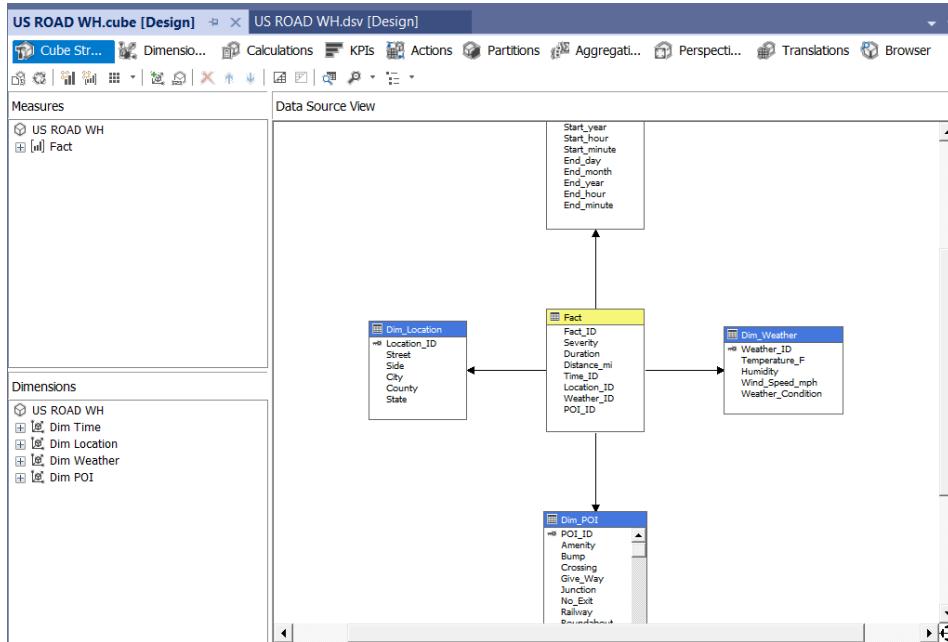
Bước 5: Chọn các bảng **Dimension** và nhấn **Next**



Bước 6: Nhấn finish để hoàn tất xây dựng các khối (**Cubes**) và xác định các độ đo (**Measures**)



Kết quả đạt được như sau

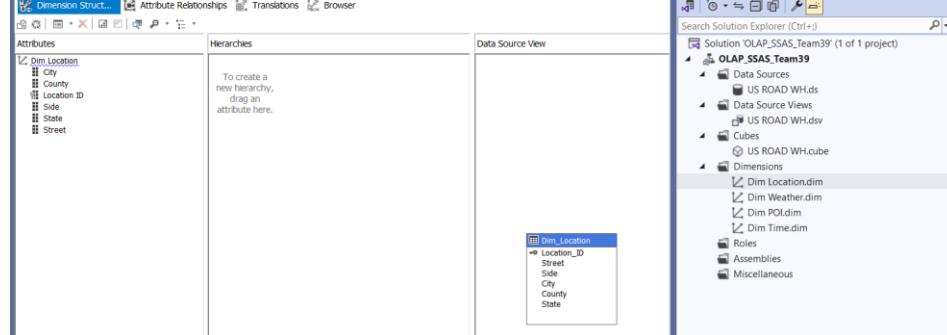


3.6. Xác định các chiều (Dimension)

Bước 1:

- Chọn **Dim Location.dim** trong folder **Dimension** của cửa sổ **Solution**
- Kéo thả các thuộc tính **Street, Side, City, County, State** từ **Data Source View** vào **Attributes**

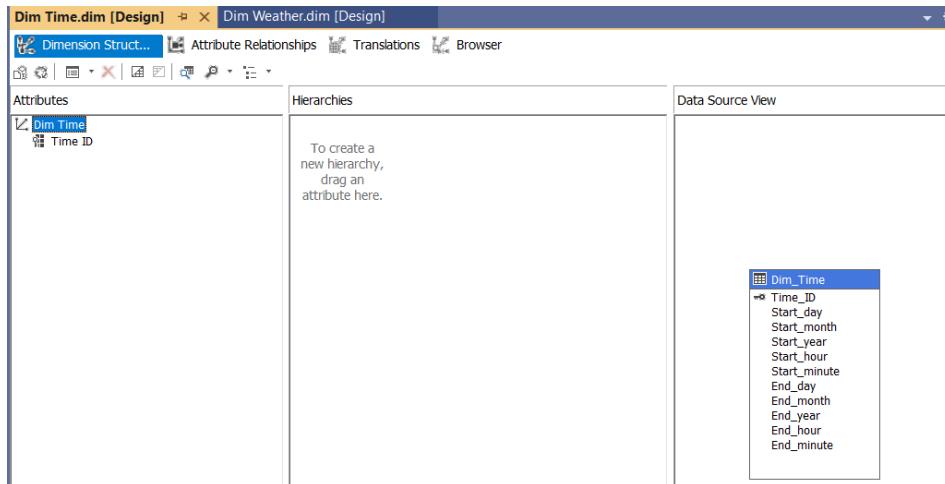
Attributes



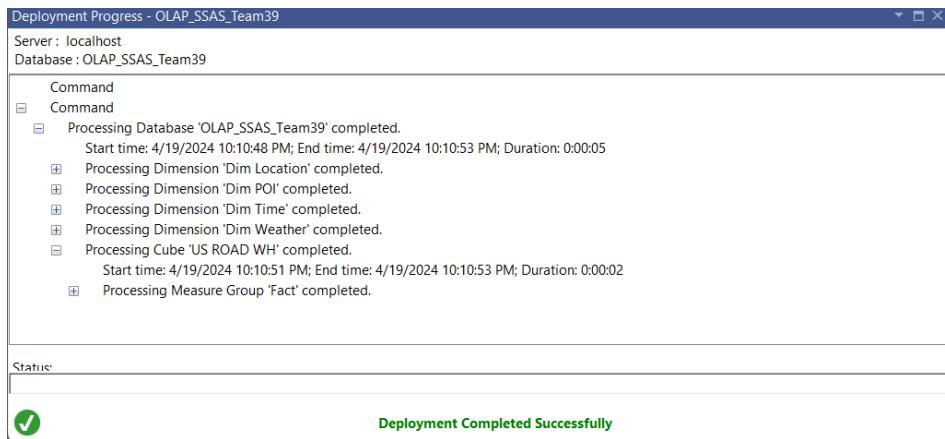
Bước 2: Thực hiện tương tự với các bảng **Dimension** còn lại với các thuộc tính cần thiết

The screenshot shows the 'Dim Weather.dim [Design]' window. The 'Attributes' pane on the left lists 'Dim Weather' with a single attribute 'Weather_ID'. The 'Hierarchies' pane in the center has a placeholder message: 'To create a new hierarchy, drag an attribute here.' The 'Data Source View' pane on the right shows a table named 'Dim_Weather' with five columns: 'Weather_ID', 'Temperature_F', 'Humidity', 'Wind_Speed_mph', and 'Weather_Condition'.

The screenshot shows the 'Dim POI.dim [Design]' window. The 'Attributes' pane on the left lists 'Dim POI' with multiple attributes: Amenity, Bump, Crossing, Give_Way, Junction, No_Exit, POI_ID, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, and Turning_Loop. The 'Hierarchies' pane in the center has a placeholder message: 'To create a new hierarchy, drag an attribute here.' The 'Data Source View' pane on the right shows a table named 'Dim_POI' with the same list of attributes.



Bước 3: Click chọn **Start** để deploy project. Kết quả sẽ hiển thị như sau khi deploy thành công



3.7. Xác định các độ đo (Measures)

3.7.1. Đổi tên và thuộc tính các độ đo ban đầu

Bước 1: Chọn **Show Measures Grid** tại khối vừa tạo để hiển thị các độ đo

	Name	Measure Group	Data Type	Aggregation
Severity	Fact	UnsignedInt	Sum	
Duration		Integer	Sum	
Distance Mi		Double	Sum	
Fact Count		Integer	Count	

Bước 2: Đổi tên và thuộc tính của các độ đo theo các hàm tổng hợp (aggregation) như sau:

- **Severity** (mức độ nghiêm trọng của công trình thi công ảnh hưởng tới đường bộ) đổi thành Average Severity (mức độ nghiêm trọng trung bình của công trình thi công ảnh hưởng tới đường bộ), sau đó đổi Aggregation từ Sum sang Average.
- **Duration** (Khoảng thời gian công trình thi công ảnh hưởng tới đường bộ kéo dài) đổi thành Average Duration (Khoảng thời gian công trình thi công ảnh hưởng tới đường bộ kéo dài trung bình), sau đó đổi Aggregation từ Sum sang Average.
- **Distance** (chiều dài của đoạn đường bị ảnh hưởng bởi công trình thi công) đổi thành Sum Distance (tổng chiều dài của các đoạn đường bị ảnh hưởng bởi công trình thi công).
- **Fact Count** (Số lượng công trình thi công ảnh hưởng tới đường bộ) đổi thành Fact Count (Số lượng công trình thi công ảnh hưởng tới đường bộ xảy ra tại Mỹ) và giữ nguyên Aggregation là Count.

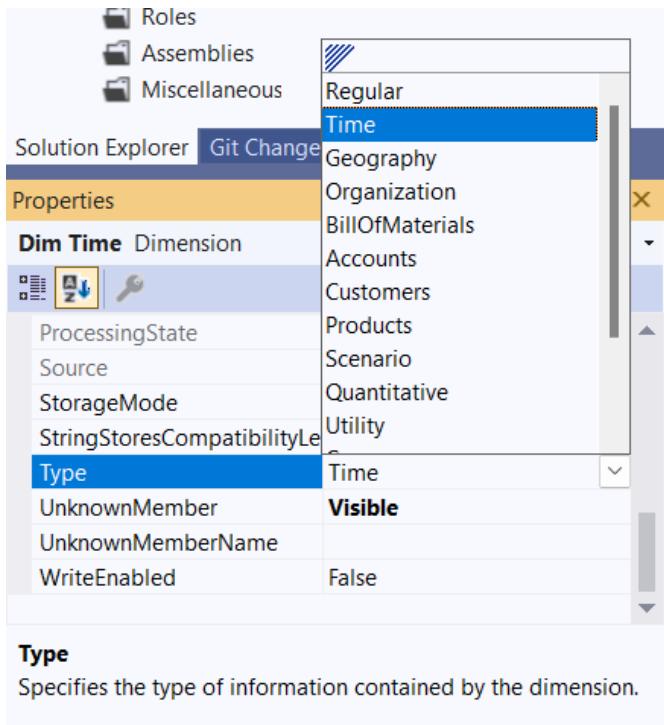
The screenshot shows the 'Measures' table in the 'US ROAD WH.cube [Design]' window. The table has four columns: Name, Measure Group, Data Type, and Aggregation. There are four rows:

	Name	Measure Group	Data Type	Aggregation
1	Average Severity	~~~~~	UnsignedInt	AverageOfC...
2	Average Duration	~~~~~	Integer	AverageOfC...
3	Sum Distance	~~~~~	Double	Sum
4	Fact Count	Fact	Integer	Count

Sau khi đổi tên và thuộc tính các độ đo ban đầu. Một thông báo xuất hiện yêu cầu ta phải có một **time dimension**.

The screenshot shows the same 'Measures' table, but now a tooltip appears over the 'Fact Count' row: "Semi-additive measure requires a time dimension." This indicates that the measure needs a time dimension to be fully additive.

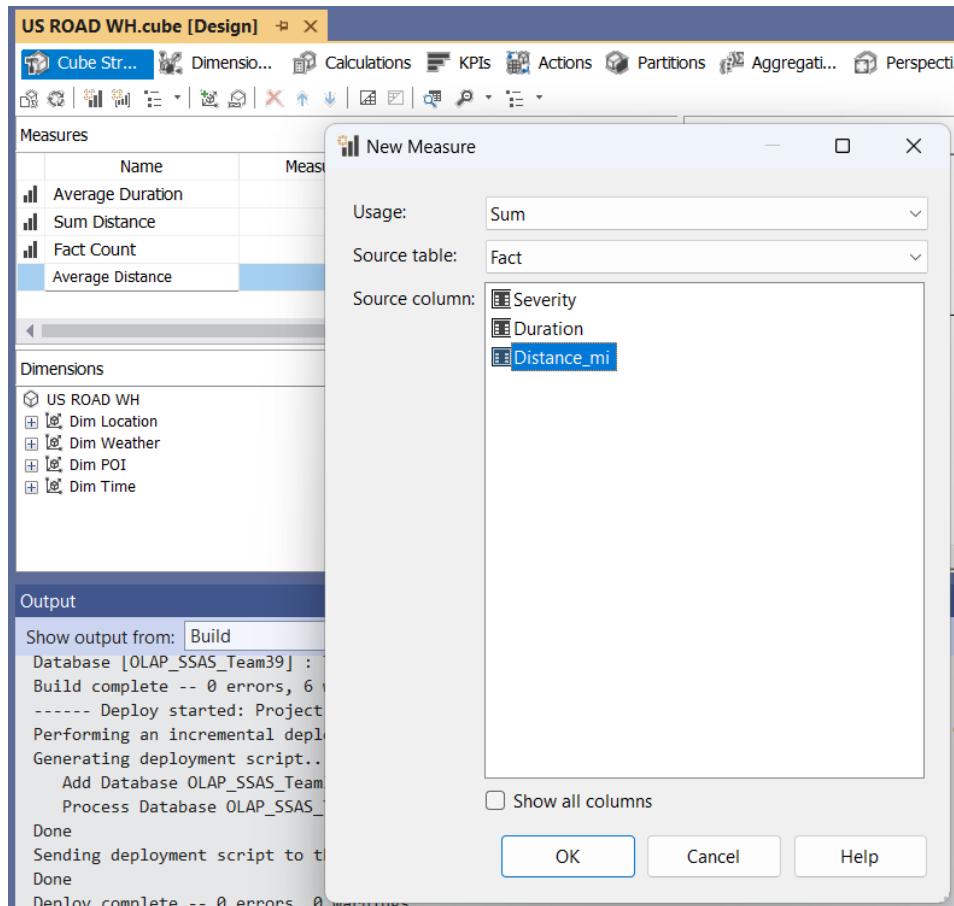
Bước 3: Mở Dim Time.dim Sau đó tiến hành đổi kiểu bảng từ Regular sang Time



3.7.2. Tạo các độ đo mới

Để tạo độ đo mới, chọn **Add new measure...**, tạo thêm 1 độ đo mới:

- Average Distance



- Max distance

US ROAD WH.cube [Design]*

Cube Str... Dimension... Calculations KPIs Actions Partitions Aggregati... Perspec...

Measures

Name	Measure
Sum Distance	
Fact Count	
Average Distance	
Max Distance	

Dimensions

- US ROAD WH
- Dim Location
- Dim Weather
- Dim POI
- Dim Time

Output

Show output from: Build
Database [OLAP_SSAS_Team39] :
Build complete -- 0 errors, 6 warnings
----- Deploy started: Project
Performing an incremental deployment
Generating deployment script...
Add Database OLAP_SSAS_Team39
Process Database OLAP_SSAS_Team39
Done
Sending deployment script to target...

New Measure

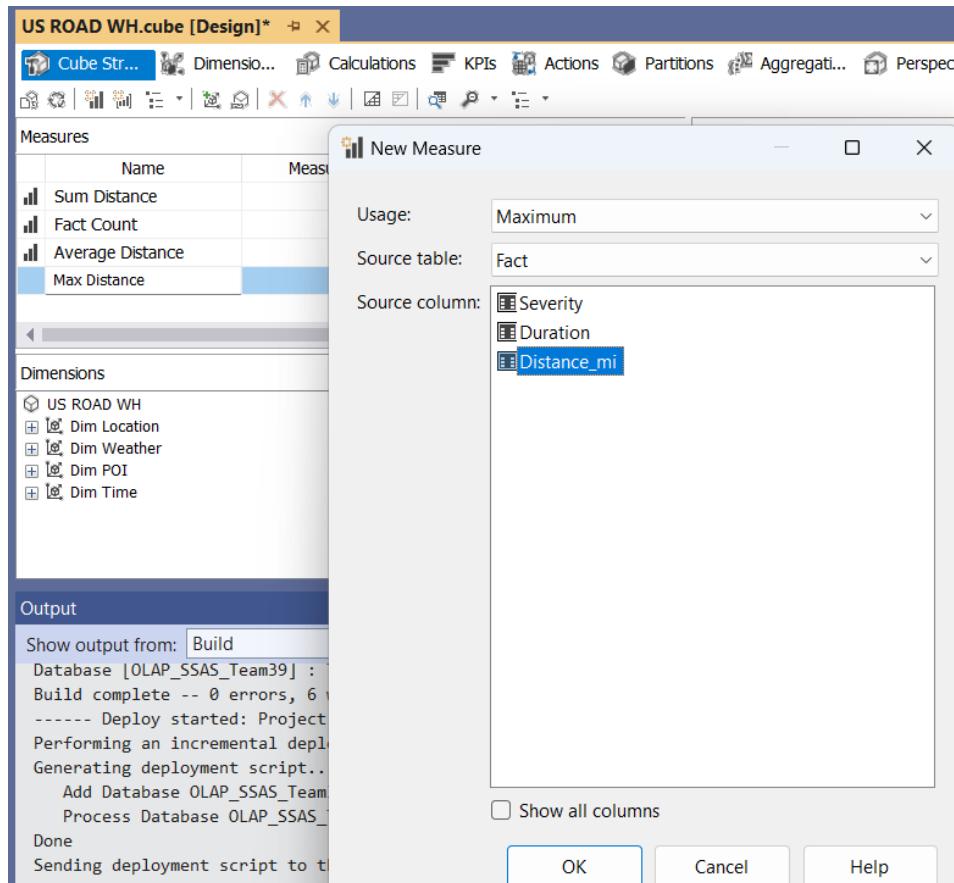
Usage: Maximum

Source table: Fact

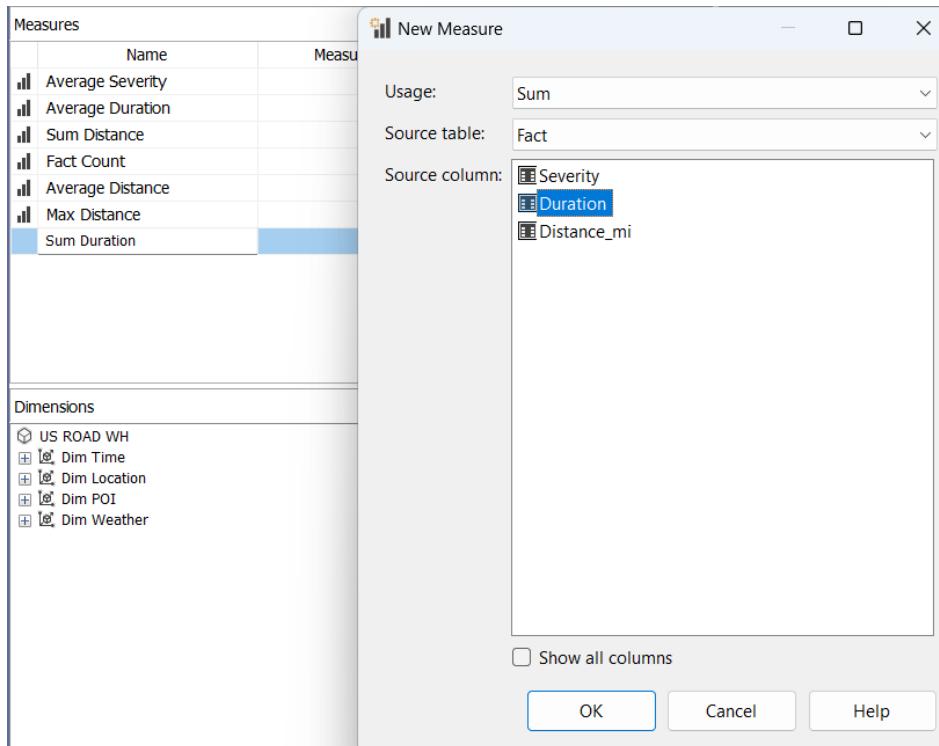
Source column: Severity
Duration
Distance_mi

Show all columns

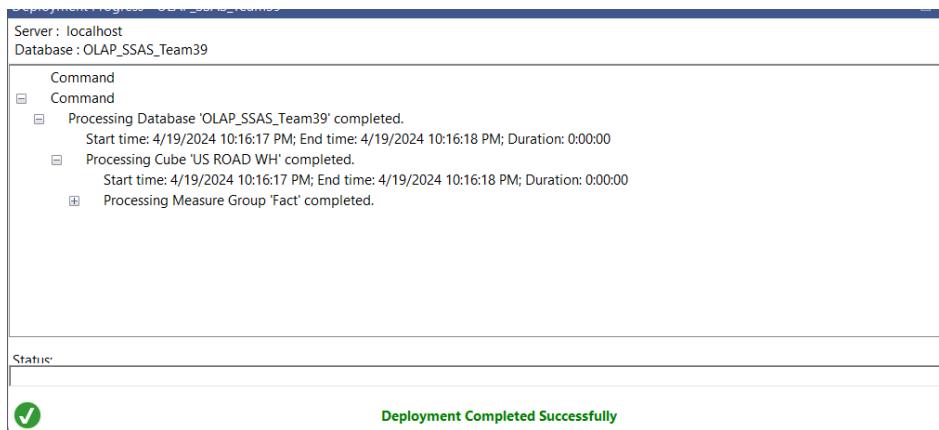
OK Cancel Help



- Sum Duration



Deploy lại project để đảm bảo không xảy ra lỗi



3.7.3. Tổng kết độ đo

STT	Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa
1	Fact Count	Int	Đếm số vụ ảnh hưởng bởi công trình
2	Average Severity	Double	Mức độ nghiêm trọng trung bình của sự ảnh hưởng của công trình
3	Average Duration	Double	Khoảng thời gian diễn ra trung bình sự ảnh hưởng
4	Sum Distance	Double	Tổng chiều dài đoạn đường bị ảnh hưởng
5	Sum Duration	Int	Tổng thời gian đoạn đường bị ảnh hưởng
6	Average Distance	Double	Chiều dài trung bình của các đoạn đường bị ảnh hưởng
7	Max Distance	Double	Chiều dài lớn nhất đoạn đường bị ảnh hưởng

3.8. Phân cấp trong bảng chiều

3.8.1. Phân cấp bảng Dim_Time

Bước 1: Kéo các thuộc tính cần thiết để thực hiện phân cấp qua cửa sổ **Hierarchies**

Dim Time.dim [Design]*

Dimension Struct... Attribute Relationships Translations Browser

Attributes

- Dim Time
 - End Day
 - End Hour
 - End Minute
 - Start Day
 - Start Hour
 - Start Minute
 - Start Month
 - Start Year
 - Time ID

Hierarchies

- Construction_Start
 - Start Year
 - Start Month
 - Start Day
 - Start Hour
 - Start Minute
 - <new level>
- Construction_End
 - End Year
 - End Month
 - End Day
 - End Hour
 - End Minute
 - <new level>

Bước 2: Tạo mối quan hệ trong thẻ Attribute Relationships

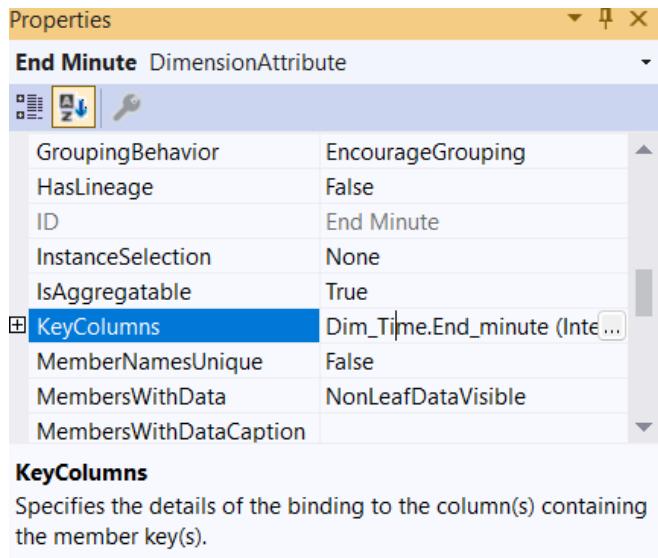
Dim Time.dim [Design]* US ROAD WH.cube [Design]*

Dimension Struct... Attribute Relationships Translations Browser

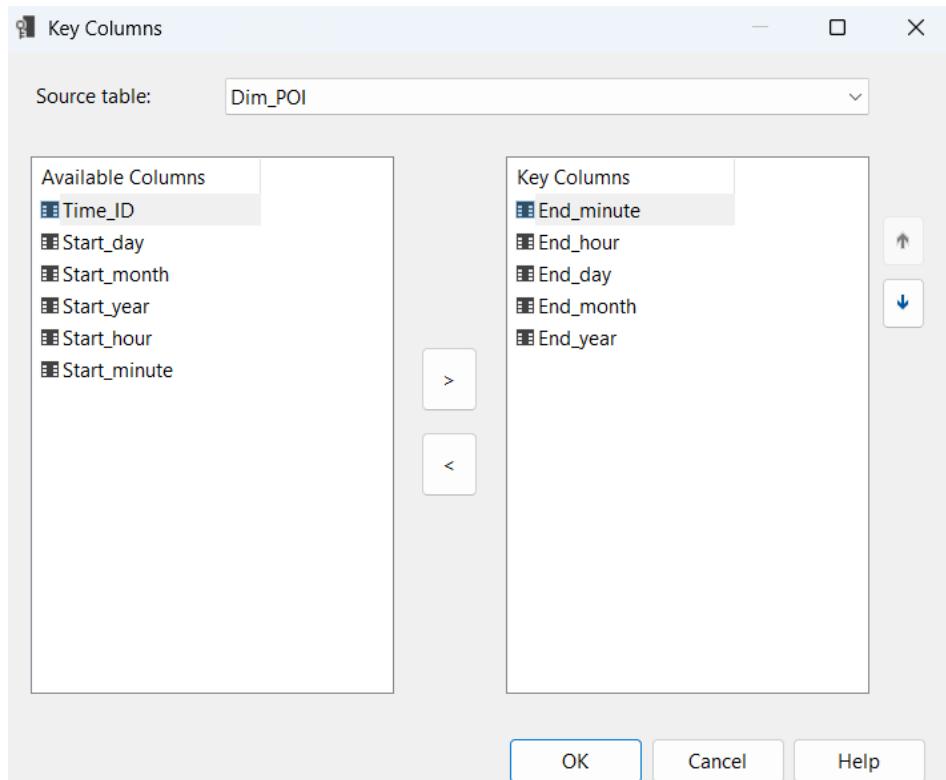
Time ID → End Minute → End Hour → End Day → End Month → End Year

Time ID → Start Minute → Start Hour → Start Day → Start Month → Start Year

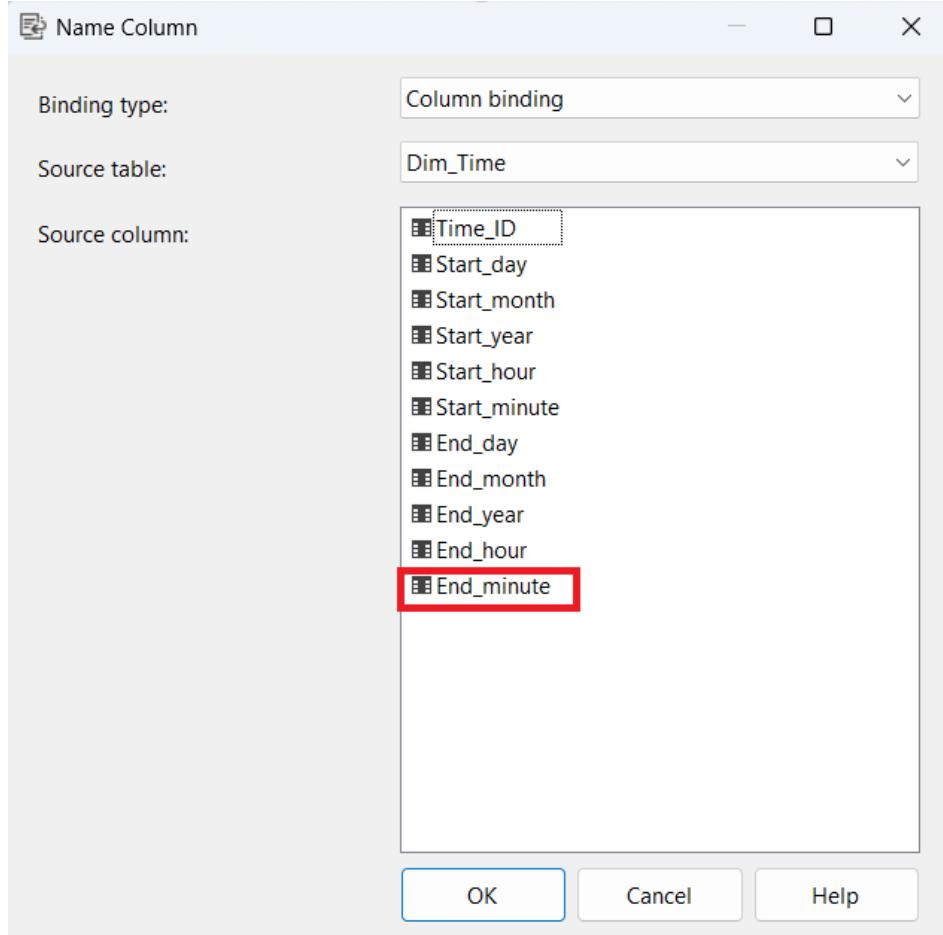
Bước 3: Chỉnh KeyColumns và NameColumn của thuộc tính End Minute trong cửa sổ Properties bằng cách chọn KeyColumns.



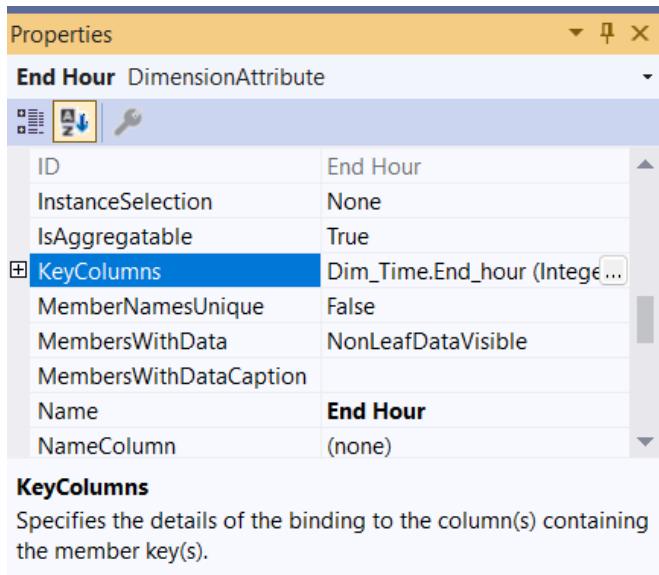
Thêm các thuộc tính cấp cao hơn vào **KeyColumns**, và nhấn **OK**



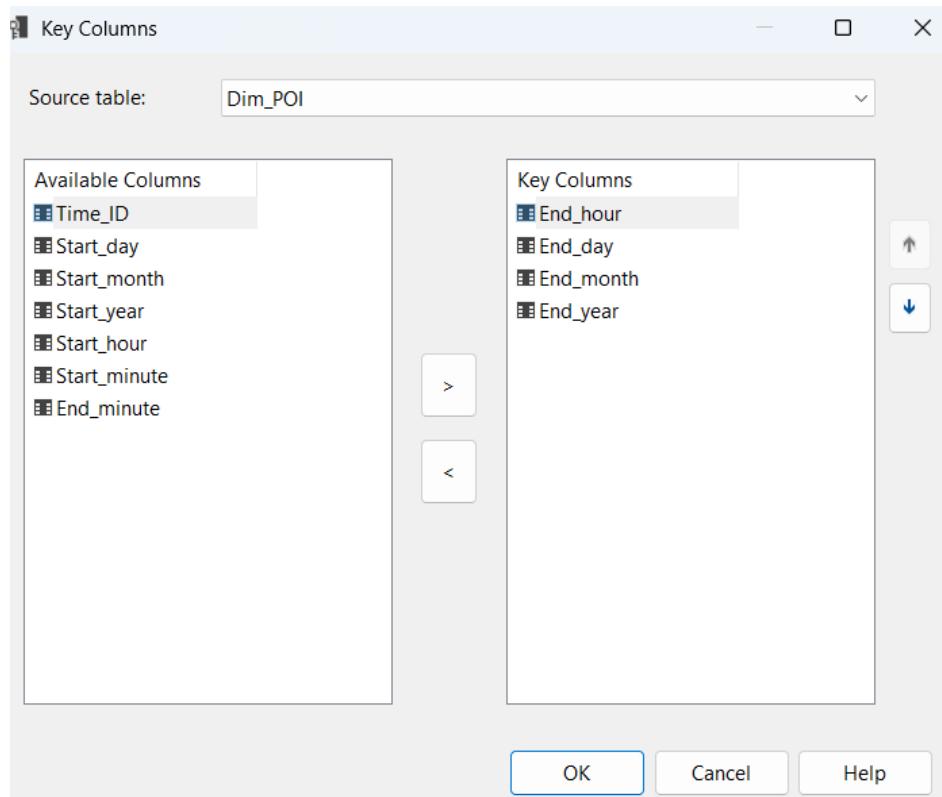
Chon Name Column trong cửa sổ Properties của thuộc tính End Minute



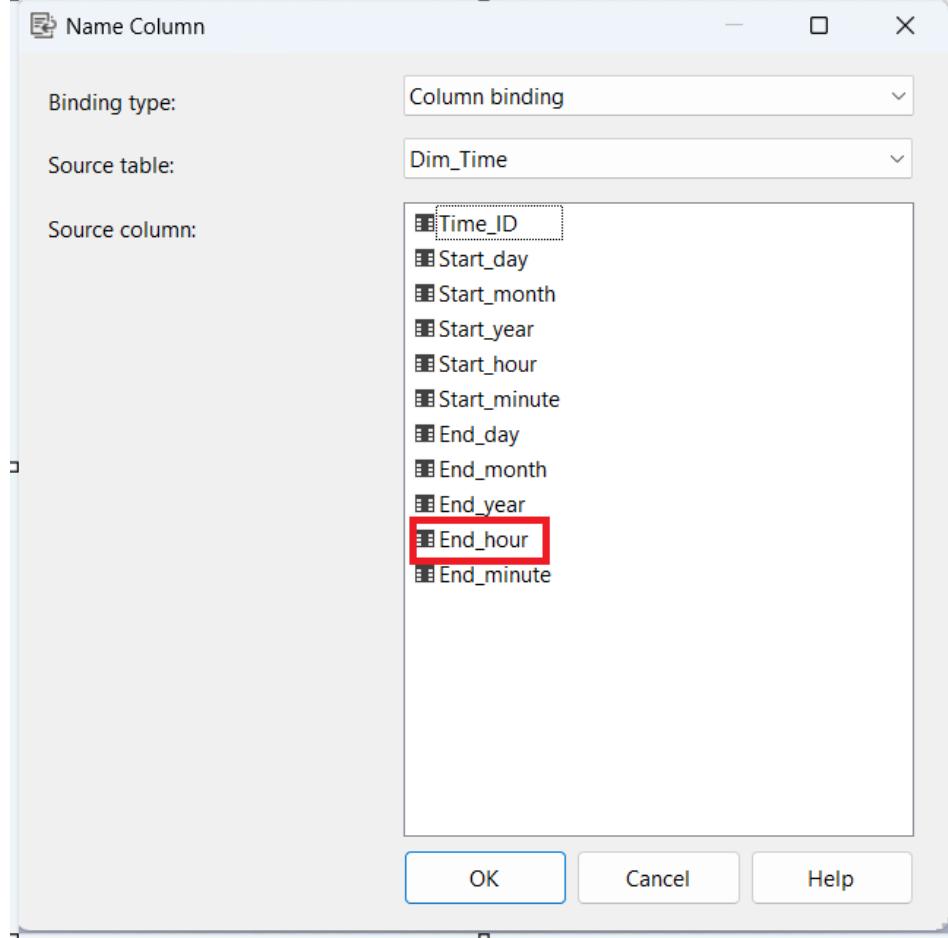
Bước 4: Chỉnh KeyColumns và NameColumn của thuộc tính **End Hour** trong cửa sổ Properties bằng cách chọn KeyColumns.



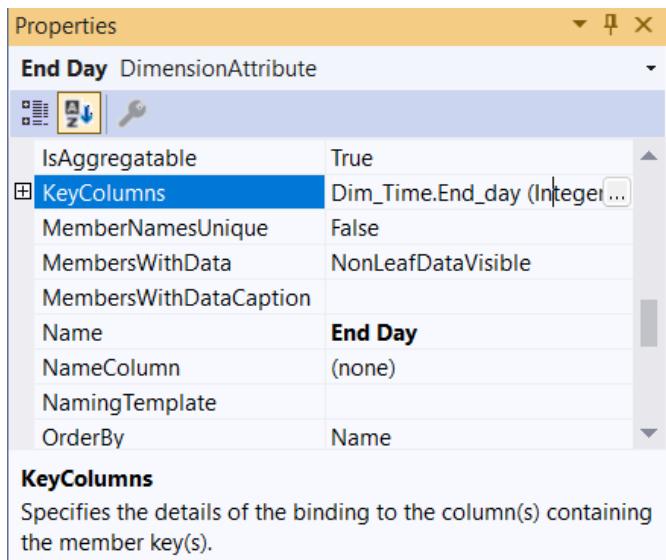
- Thêm các thuộc tính cấp cao hơn của End Hour vào **KeyColumns**, và nhấn **OK**



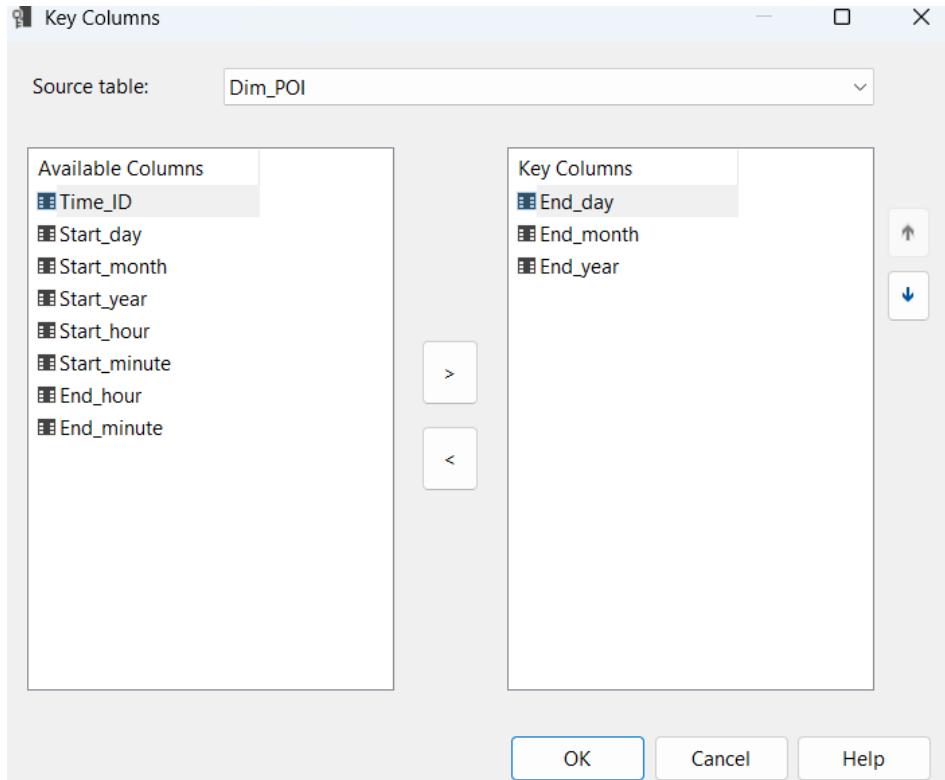
- Tiếp theo chọn **Name Column** trong cửa sổ **Properties** của thuộc tính **End Hour** và chọn **End_hour**



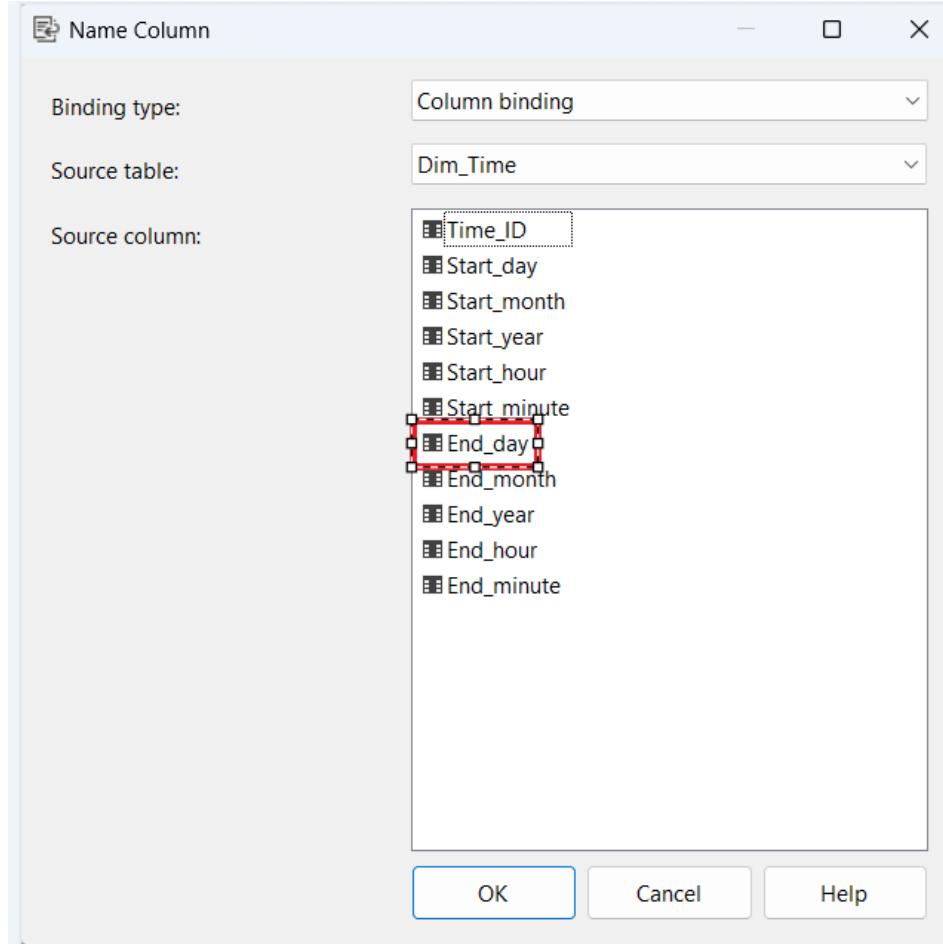
Bước 5: Chỉnh KeyColumns và NameColumn của thuộc tính **End Day** trong cửa sổ Properties bằng cách chọn KeyColumns.



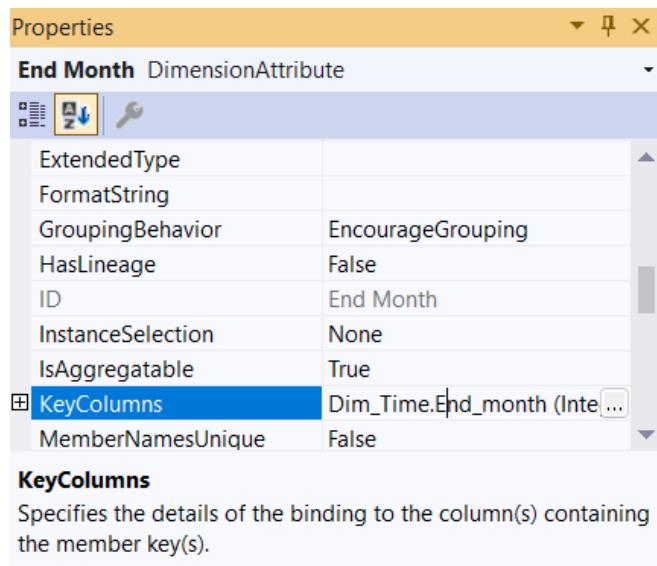
- Thêm các thuộc tính cấp cao hơn vào **KeyColumns**, và nhấn **OK**



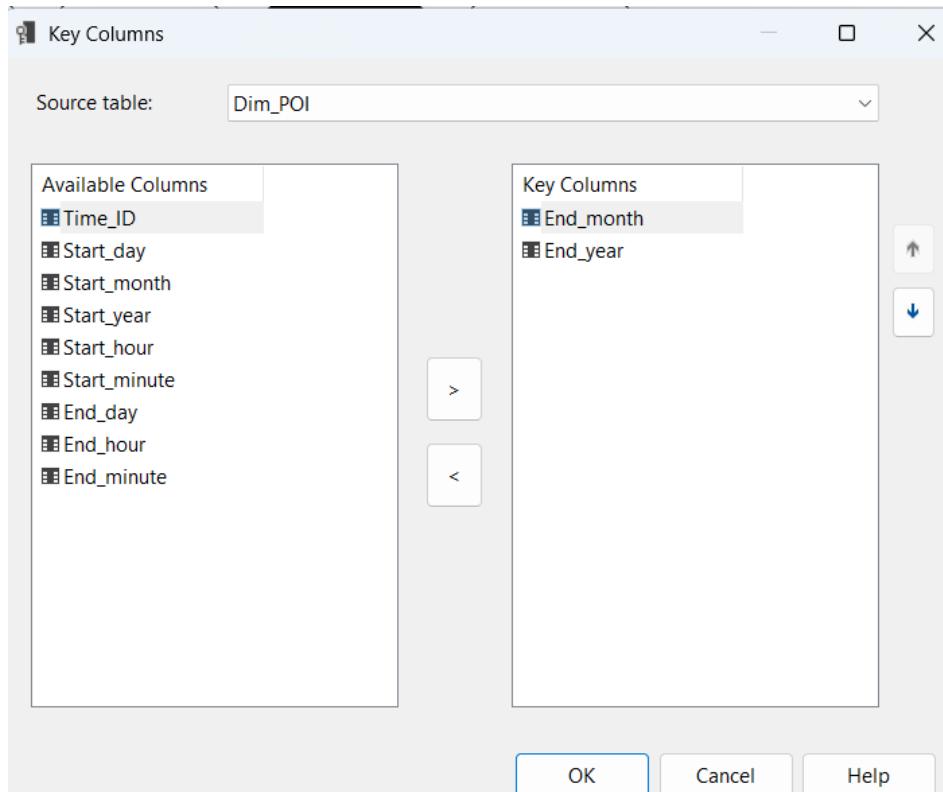
- Tiếp theo chọn **Name Column** trong cửa sổ **Properties** của thuộc tính **End Day** và chọn **End_day**



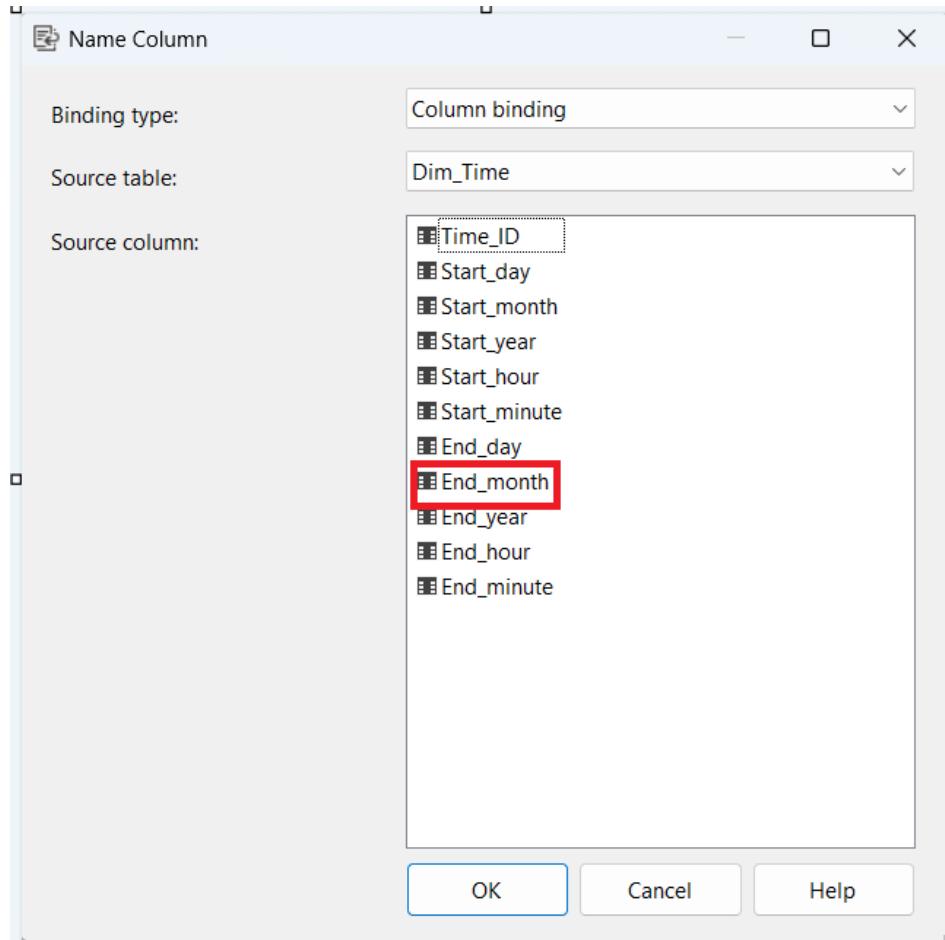
Bước 6 Chỉnh KeyColumns và NameColumn của thuộc tính End Month trong cửa sổ Properties, chọn KeyColumns.



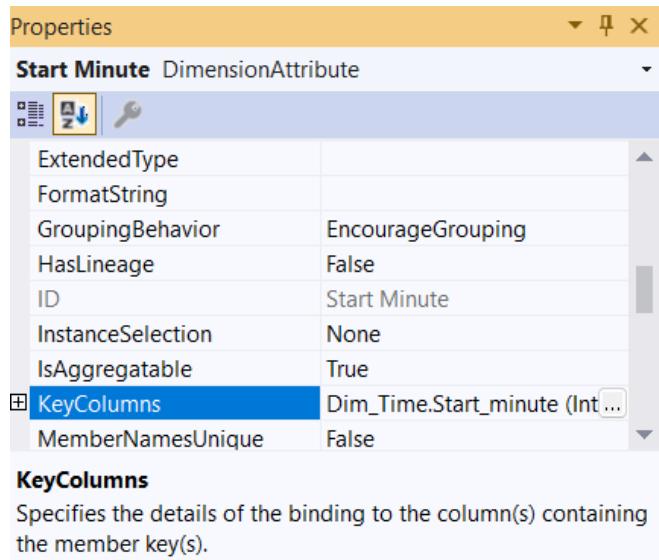
- Thêm các thuộc tính cấp cao hơn vào **KeyColumns**, và nhấn **OK**



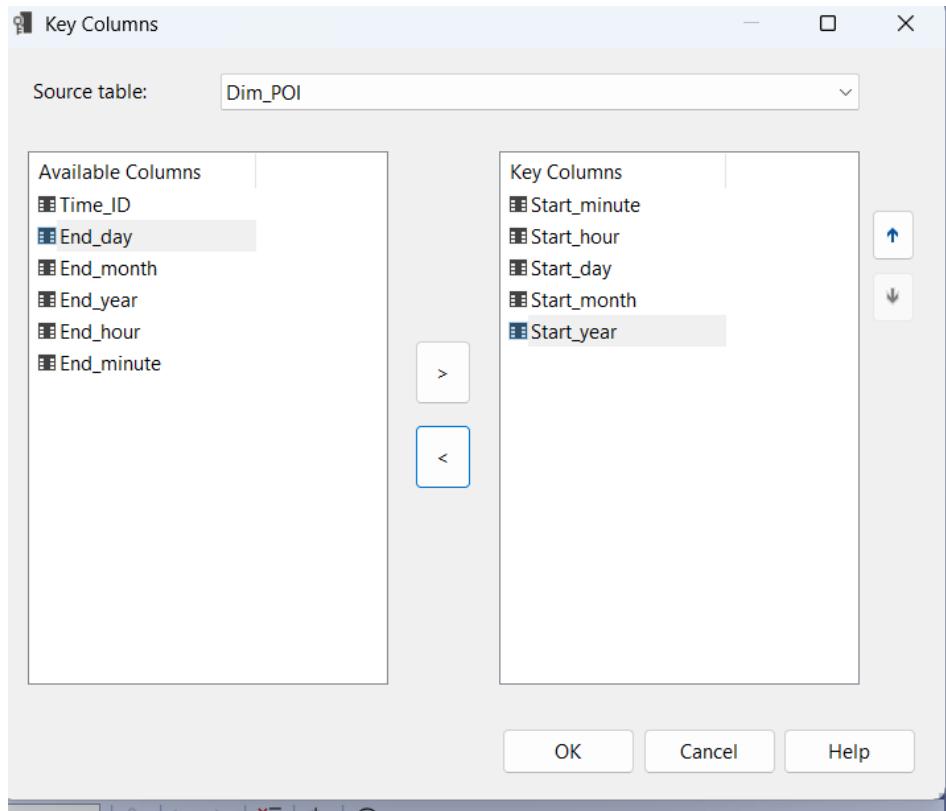
- Tiếp theo chọn **Name Column** trong cửa sổ **Properties** của thuộc tính **End Month** và chọn **End_month**



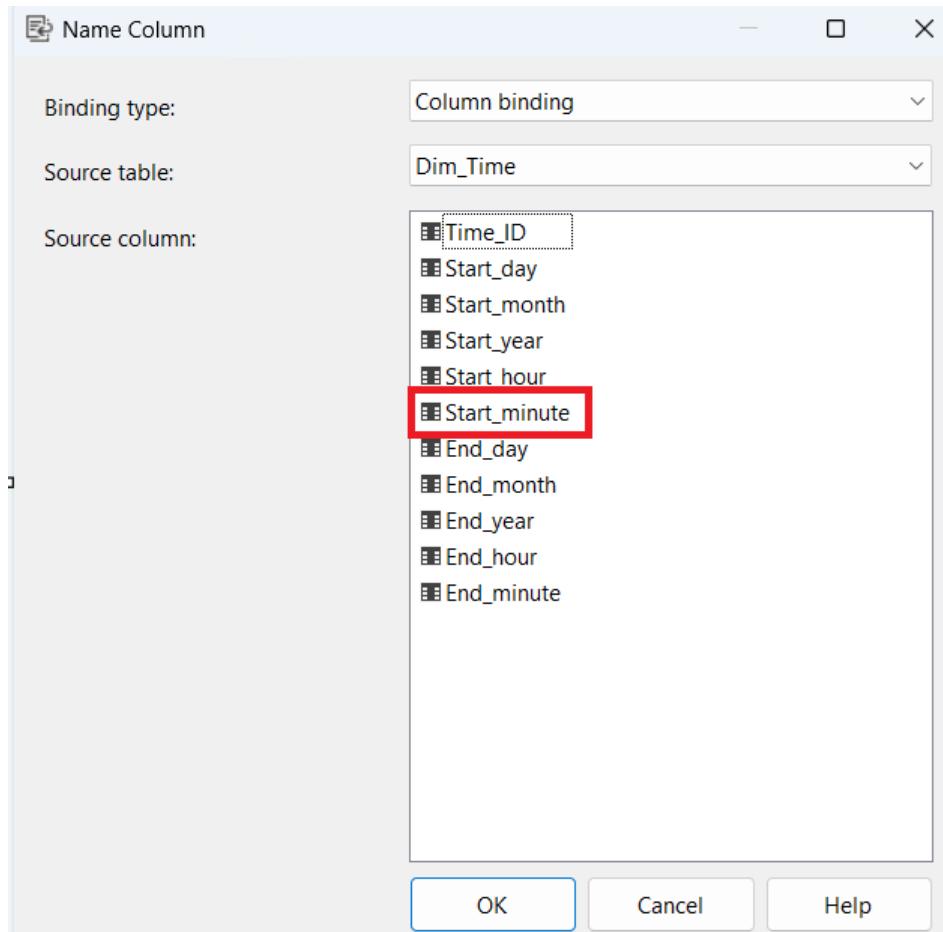
Bước 7: Chỉnh KeyColumns và NameColumn của thuộc tính Start Minute trong cửa sổ Properties, chọn KeyColumns.



- Thêm các thuộc tính cấp cao hơn vào **KeyColumns**, và nhấn **OK**

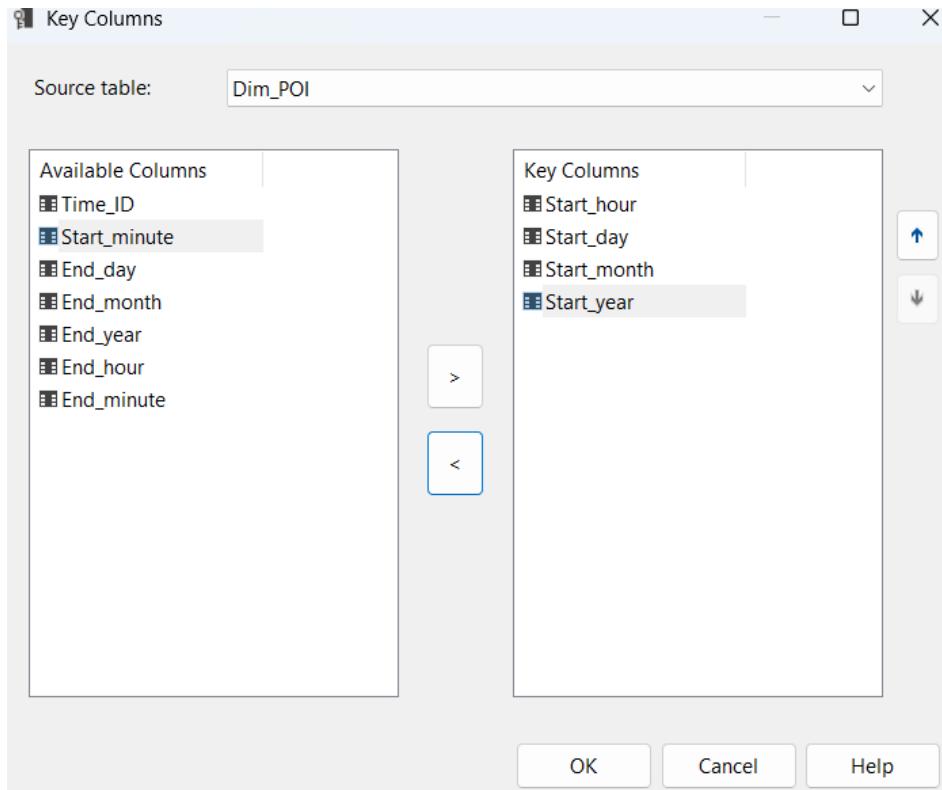


- Tiếp theo chọn **Name Column** trong cửa sổ **Properties** của thuộc tính **Start Minute** và chọn **Start_minute**

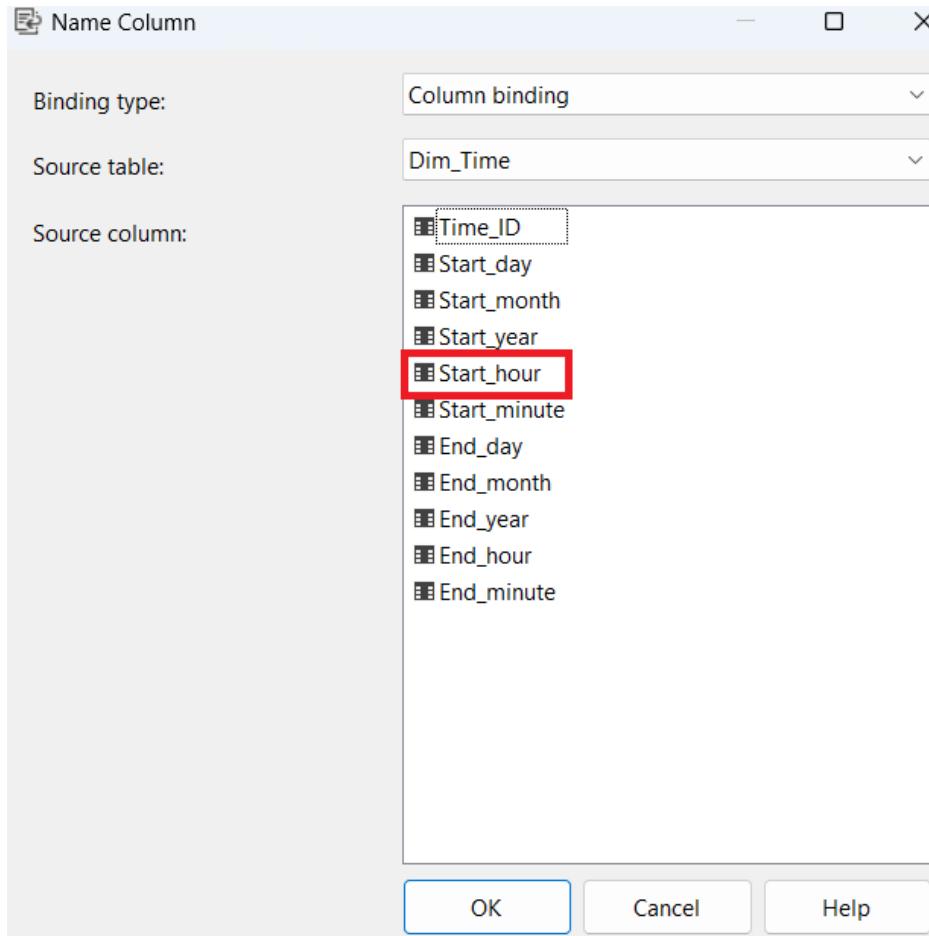


Bước 8: Chỉnh KeyColumns và NameColumn của thuộc tính Start Hour trong cửa sổ Properties, chọn KeyColumns.

- Thêm các thuộc tính cấp cao hơn vào KeyColumns, và nhấp OK

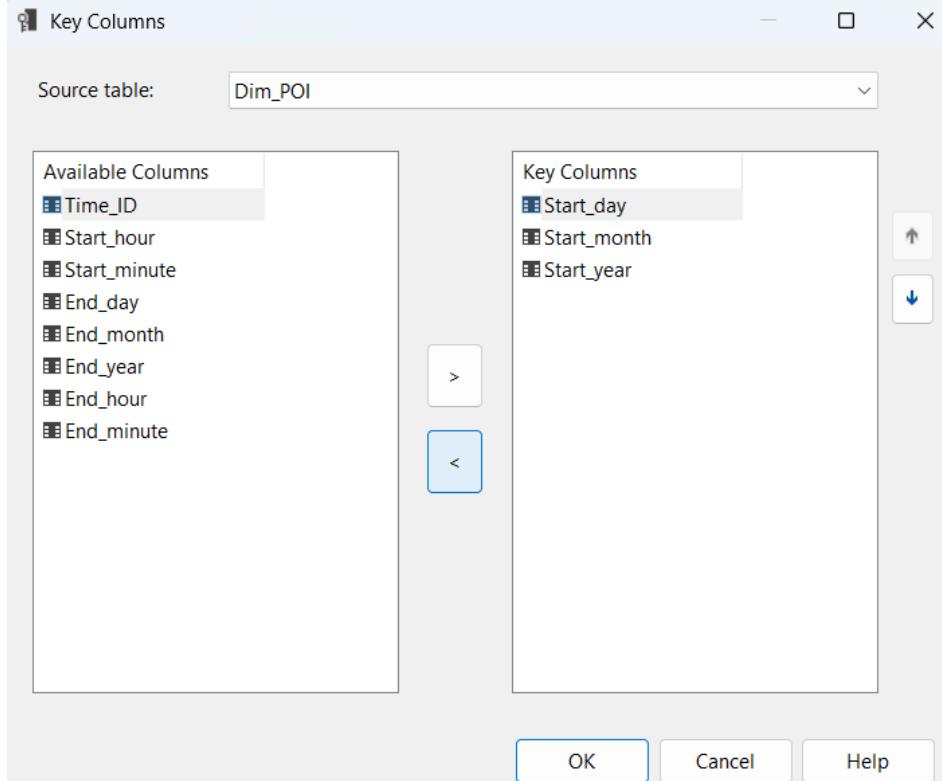


- Tiếp theo chọn **Name Column** trong cửa sổ **Properties** của thuộc tính **Start Hour** và chọn **Start_hour**

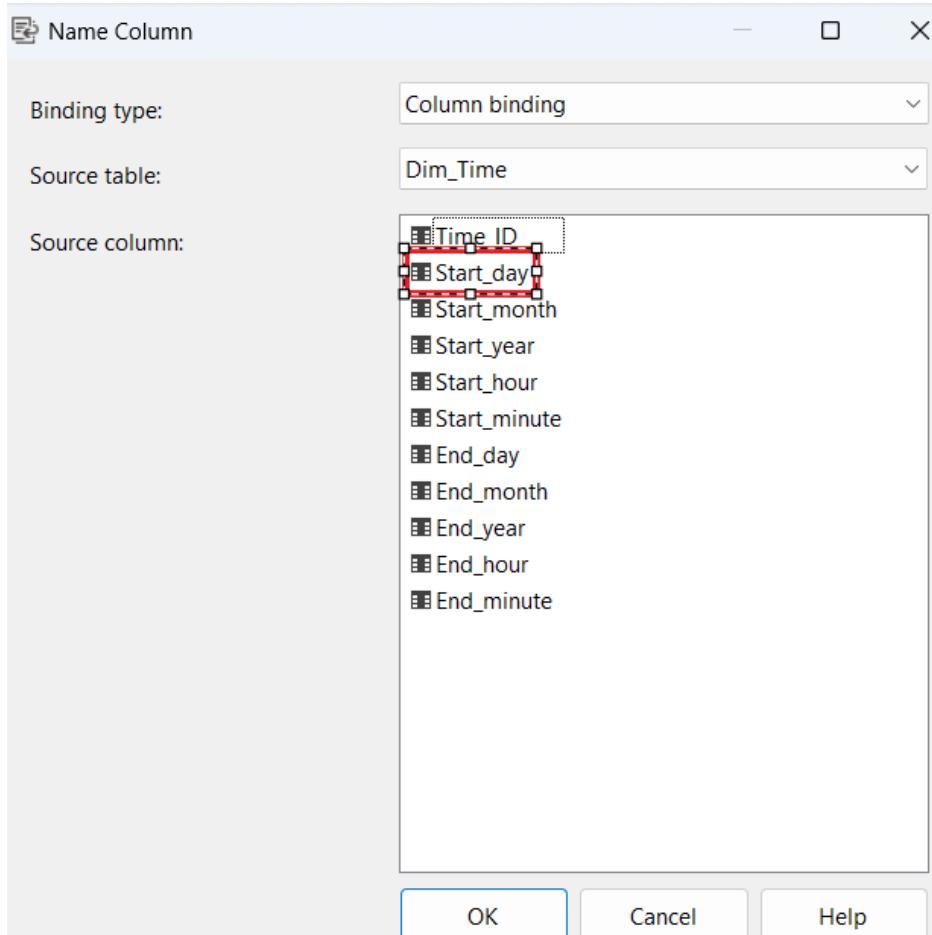


Bước 9: Chỉnh KeyColumns và NameColumn của thuộc tính Start Day trong cửa sổ Properties, chọn KeyColumns.

- Thêm các thuộc tính cấp cao hơn vào KeyColumns, và nhấn OK

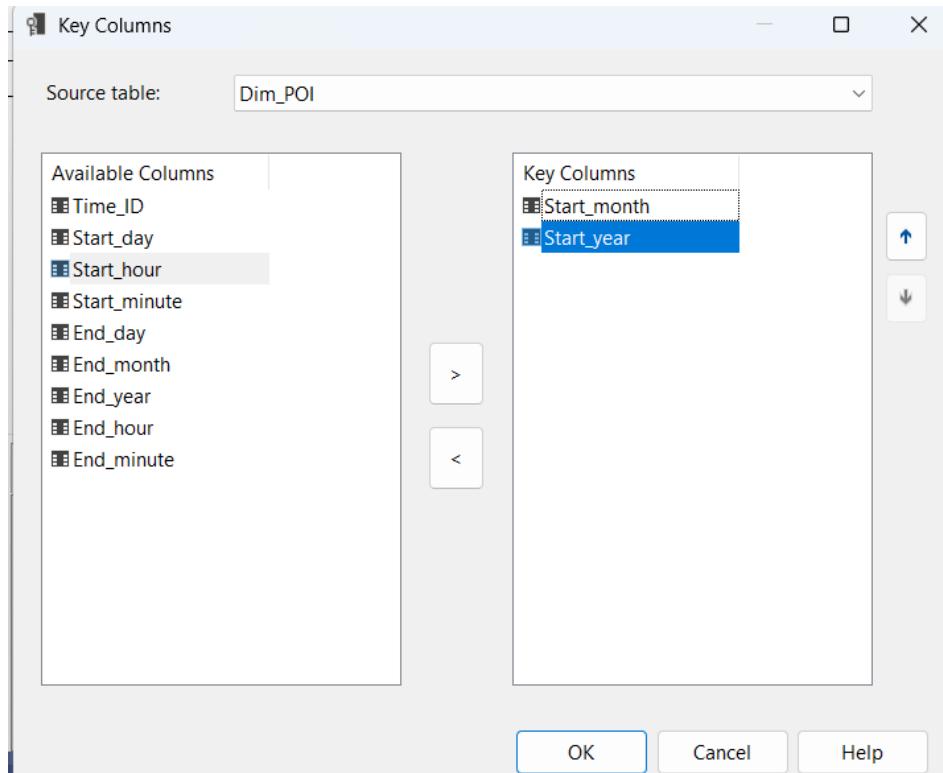


- Tiếp theo chọn **Name Column** trong cửa sổ **Properties** của thuộc tính **Start Day** và chọn **Start_day**

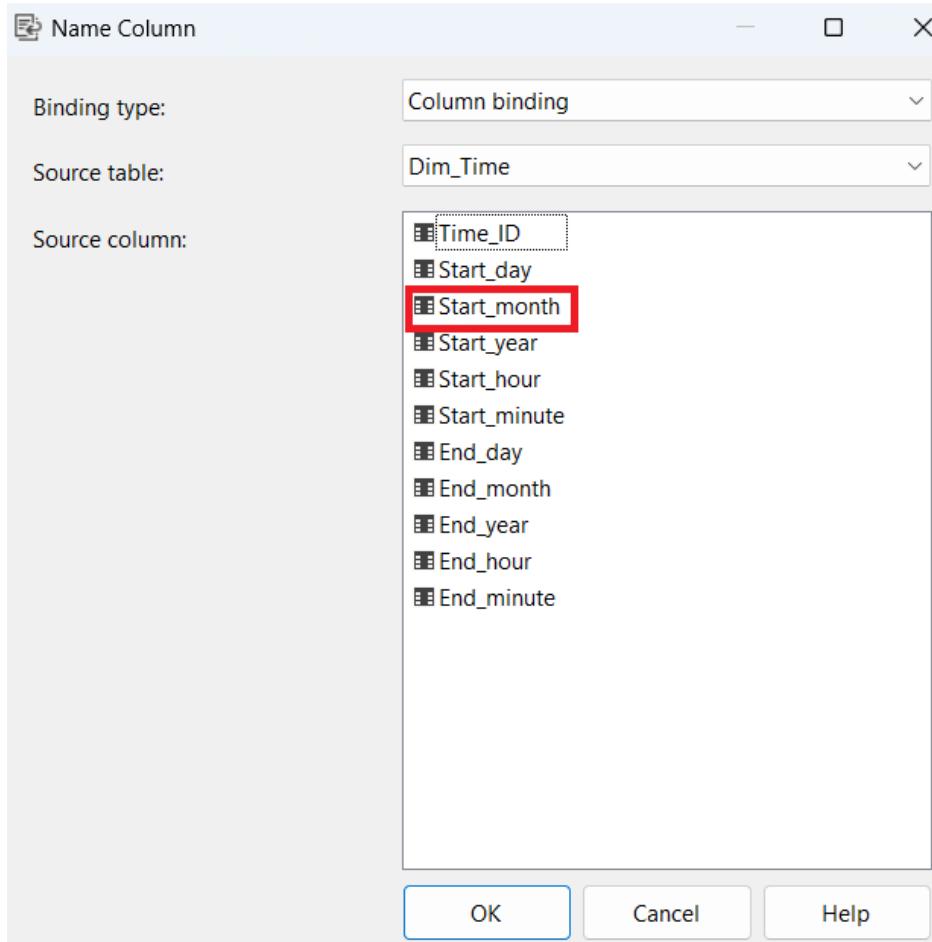


Bước 10: Chỉnh KeyColumns và NameColumn của thuộc tính **Start month** trong cửa sổ Properties, chọn KeyColumns.

- Thêm các thuộc tính cấp cao hơn vào **KeyColumns**, và nhấn **OK**



- Tiếp theo chọn **Name Column** trong cửa sổ **Properties** của thuộc tính **Start Month** và chọn **Start_month**



3.8.2. Phân cấp bảng Dim_Location

Bước 1: Kéo các thuộc tính cần thực hiện phân cấp qua cửa sổ **Hierarchies**

The screenshot shows the SSAS Dimension Designer interface. The title bar says "Dim Location.dim [Design]*" and "US ROAD WH.cube [Design]". The ribbon tabs include "Dimension Struct...", "Attribute Relationships", "Translations", and "Browser".

Attributes pane: Shows the attributes of the Dim Location dimension: City, County, Location ID, Side, State, and Street. The "Street" attribute is selected.

Hierarchies pane: Shows the hierarchy structure. A tooltip says: "To create a new hierarchy, drag an attribute here." The hierarchy levels are: Side, County, State, Street, and City. A placeholder "<new level>" is shown at the bottom.

Bước 2: Sắp xếp lại các thuộc tính cần phân cấp theo đúng thứ tự và đổi tên **Hierarchies** thành **Construction_Location**.

The screenshot shows the SSAS Dimension Designer interface. The title bar says "Dim Location.dim [Design]*" and "US ROAD WH.cube [Design]". The ribbon tabs include "Dimension Struct...", "Attribute Relationships", "Translations", and "Browser".

Attributes pane: Shows the attributes of the Dim Location dimension: City, County, Location ID, Side, State, and Street. The "Street" attribute is selected.

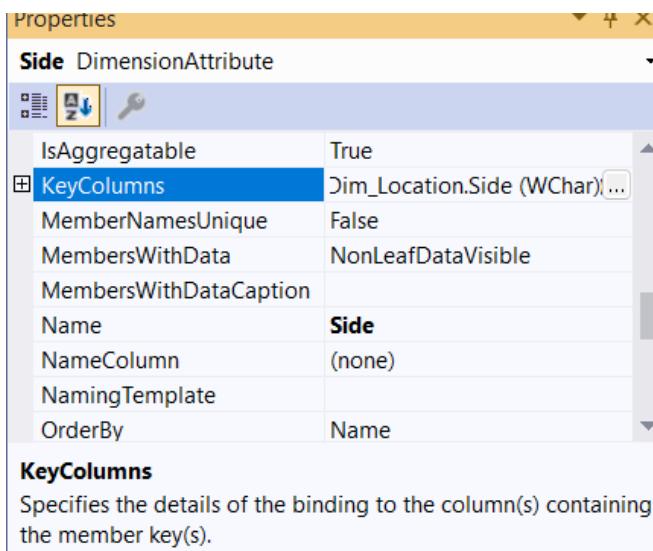
Hierarchies pane: Shows the renamed hierarchy "Construction_Location". The hierarchy levels are: State, County, City, Street, and Side. A placeholder "<new level>" is shown at the bottom.

Bước 3: Tạo mối quan hệ trong thẻ **Attribute Relationships**

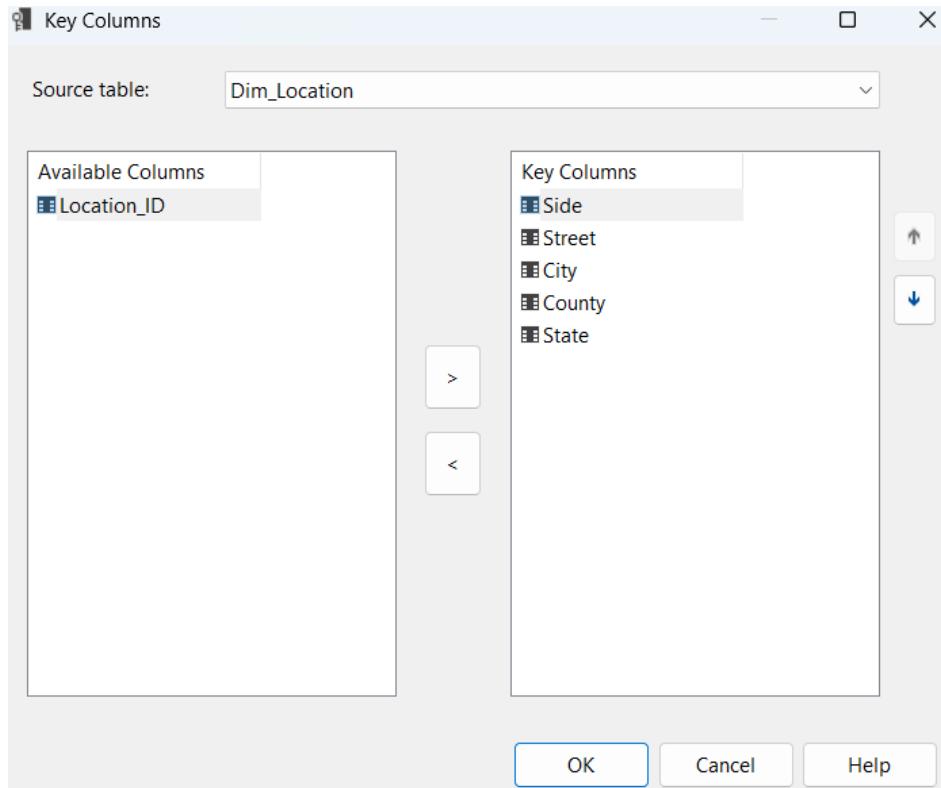
The screenshot shows the SSAS Dimension Designer interface. The title bar says "Dim Location.dim [Design]*" and "Dim Time.dim [Design]". The ribbon tabs include "Dimension Struct...", "Attribute Relationships", "Translations", and "Browser".

Attribute Relationships pane: Displays a relationship diagram where "Location ID" is connected to "Side", which is connected to "Street", which is connected to "City", which is connected to "County", which is connected to "State".

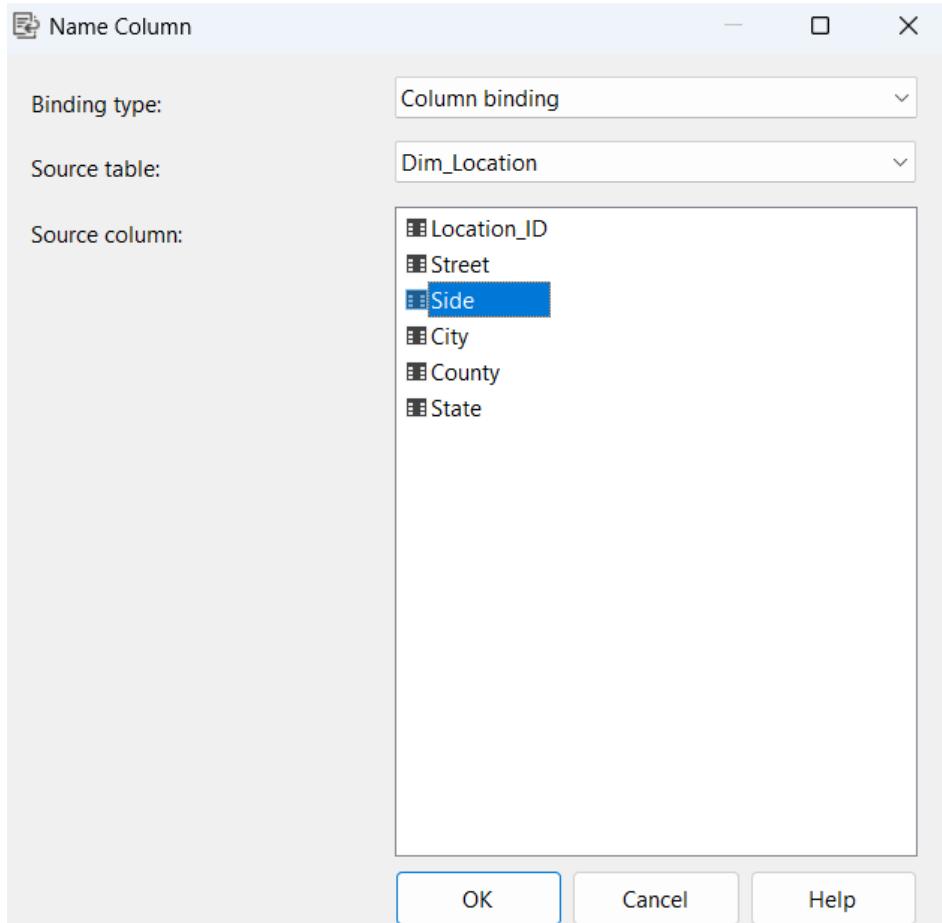
Bước 4: Chỉnh **KeyColumns** và **NameColumn** của thuộc tính **Side**, trong cửa sổ **Properties**, chọn **KeyColumns**.



Thêm các thuộc tính cấp cao hơn vào **KeyColumns**, nhấn chọn **OK** để hoàn tất

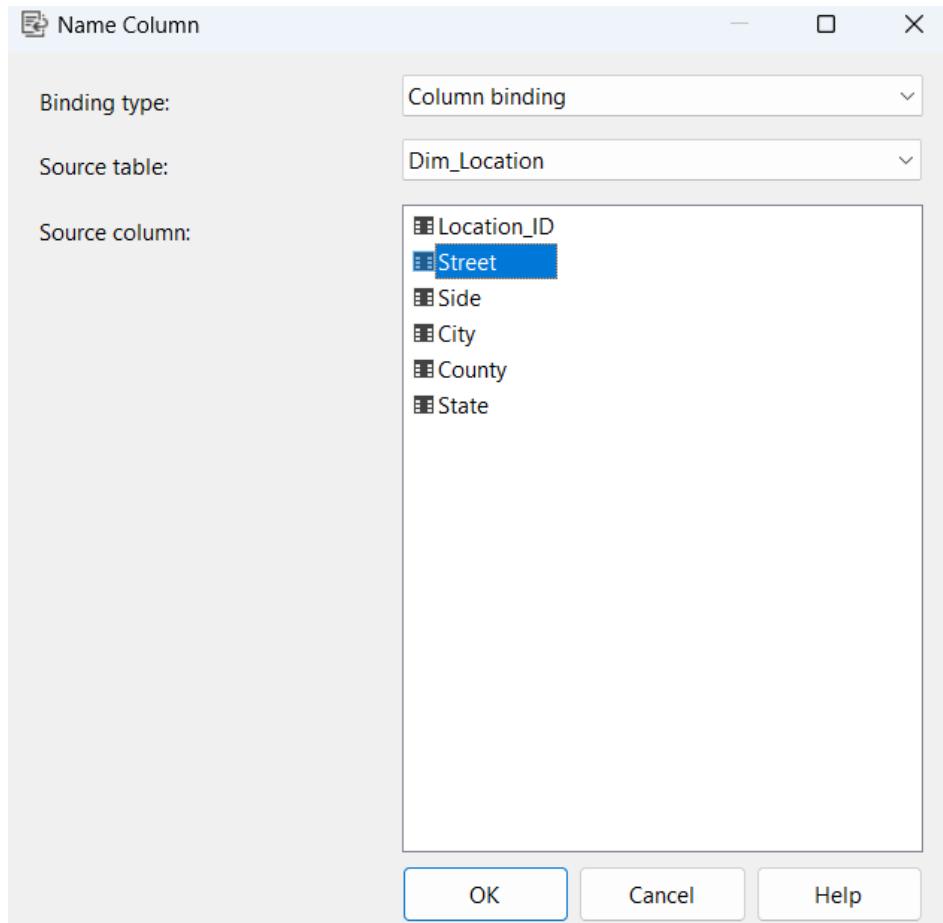


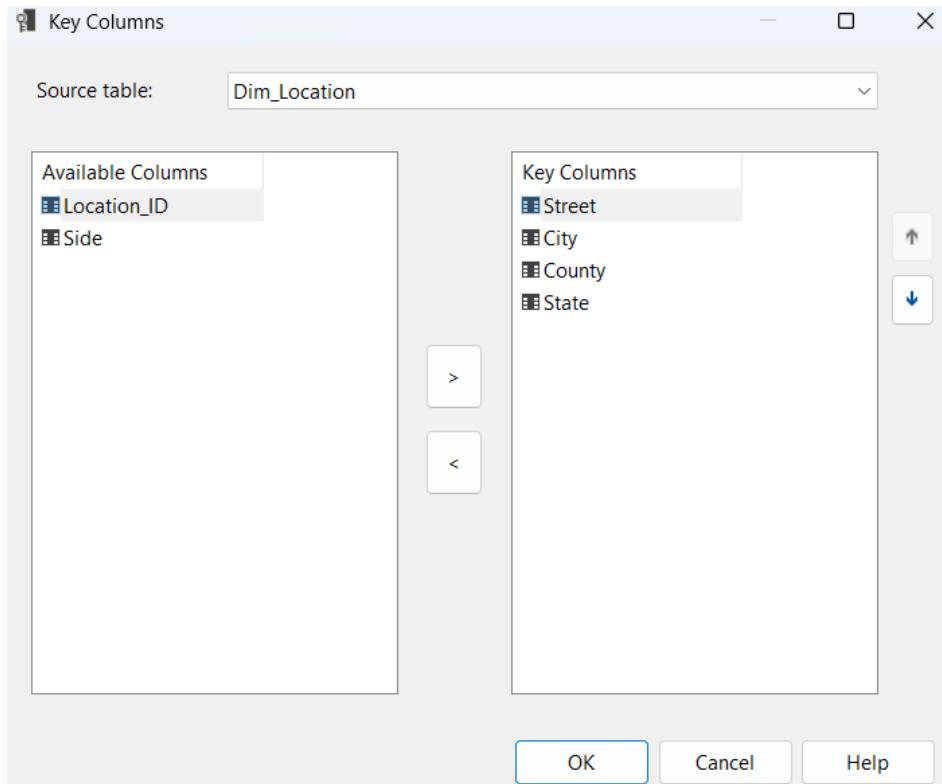
Tại **Properties** của thuộc tính **Side**, chọn **Name Column** và chọn tên thuộc tính là **Side**



Bước 5: Chỉnh KeyColumns và NameColumn của thuộc tính Street, trong cửa sổ Properties, chọn KeyColumns.

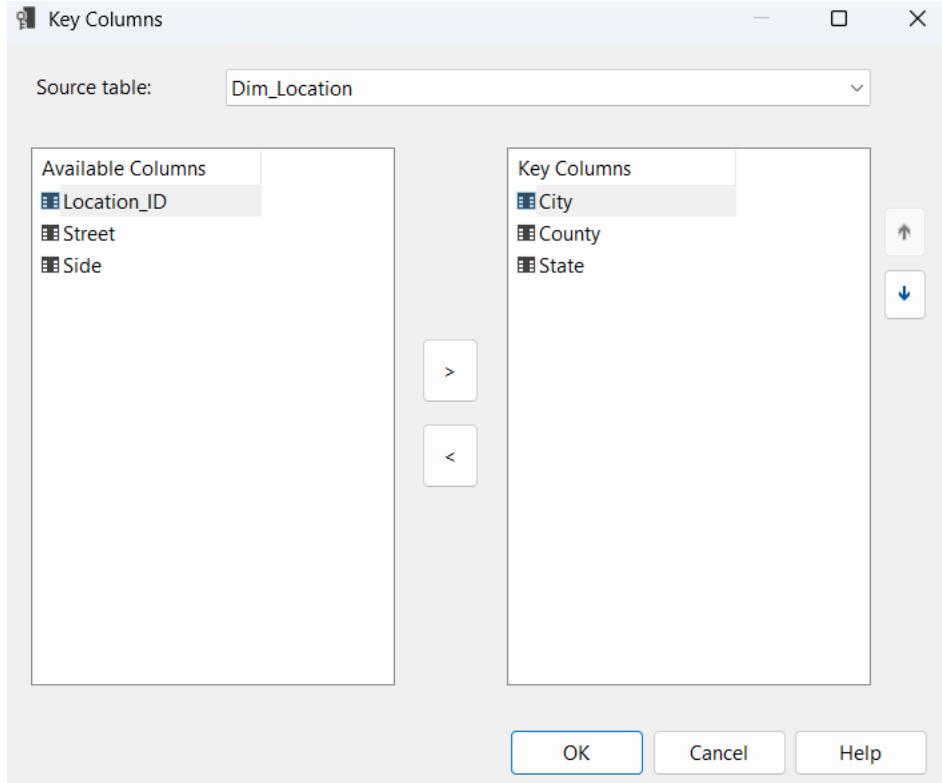
- Thêm các thuộc tính cấp cao hơn vào KeyColumns, nhấn chọn OK để hoàn tất.
- Tại Properties của thuộc tính Side, chọn Name Column và chọn tên thuộc tính là Street.





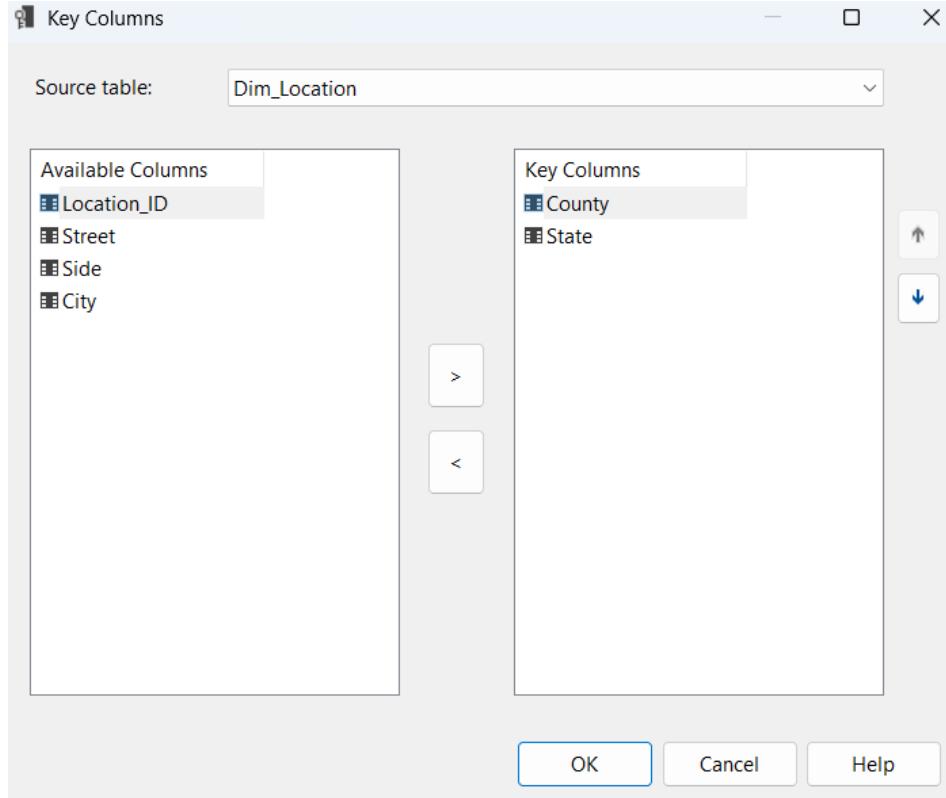
Bước 6: Chỉnh **KeyColumns** và **NameColumn** của thuộc tính **City**, trong cửa sổ **Properties**, chọn **KeyColumns**.

- Thêm các thuộc tính cấp cao hơn vào **KeyColumns**, nhấn chọn **OK** để hoàn tất
- Tại **Properties** của thuộc tính **City**, chọn **Name Column** và chọn tên thuộc tính là **City**

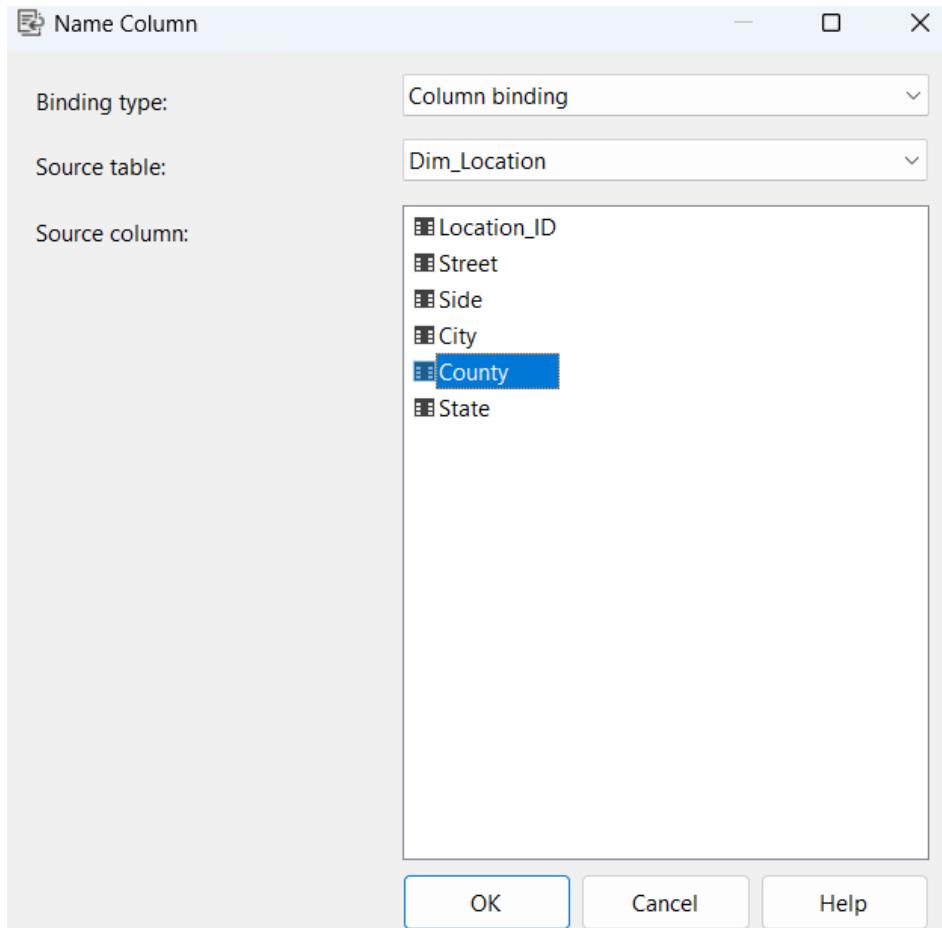


Bước 7: Chỉnh **KeyColumns** và **NameColumn** của thuộc tính **County**, trong cửa sổ **Properties**, chọn **KeyColumns**.

- Thêm các thuộc tính cấp cao hơn vào **KeyColumns**, nhấn chọn **OK** để hoàn tất



- Tại **Properties** của thuộc tính **County**, chọn **Name Column** và chọn tên thuộc tính là **County**



3.9. Thực hiện 10 câu truy vấn – Quá trình phân tích dữ liệu bằng thao tác tay trên các khối CUBE

3.9.1. Câu truy vấn 1

Nội dung câu truy vấn: Thống kê số lượng công trình làm ảnh hưởng tới đường bộ xảy ra tại 4 quận Kent, Newport, Providence và Washington của bang Rhode Island (RI) trong 3 năm liên tục 2017 – 2019.

Ths. Đỗ Thị Minh Phụng

Ý nghĩa: Để biết được trong 4 quận Kent, Newport, Providence và Washington quận nào có số lượng công trình ảnh hưởng tới đường bộ nhiều nhất và ít nhất mỗi năm từ 2017 – 2019.

Kỹ thuật sử dụng: Slice and Dice. Truy vấn này lọc dữ liệu theo vị trí cụ thể (các quận ở Rhode Island) và thời gian cụ thể (2017-2019).

Bước 1: Kéo thả Start_year, Construction_Location vào cửa sổ thực thi câu truy vấn

The screenshot shows a data query interface with the following components:

- Left Panel:** A tree view of the data model.
 - Root node: US ROAD WH
 - Child nodes: Metadata, Search Model, Measure Group, <All>, US ROAD WH, Measures, KPIs.
 - Under US ROAD WH:
 - Dim Location
 - City
 - County
 - Location ID
 - Side
 - State
 - Street
 - Construction_Location
 - Dim POI
 - Dim Time
- Top Bar:** Dimension, Hierarchy, Operator, Filter Expression.
- Bottom Bar:** Start Year, State, County, City, Street, Side.
- Text at bottom right:** Click to execute the query.

Bước 2: Loại bỏ các thuộc tính không cần dùng: City, Street, Side

The screenshot shows the Microsoft SQL Server Analysis Services (SSAS) Dimension Designer. On the left, the 'Dimension Structure' pane displays the 'US ROAD WH' dimension. It includes sections for 'Metadata', 'Measure Group' (set to '<All>'), and 'Calculated Members'. The 'Measures' section lists 'Fact' measures: Average Distance, Average Duration, Average Severity, Fact Count, Max Distance, and Sum Distance. The 'Dimensions' section lists 'Dim Location' with 'City' and 'County' levels. The 'Hierarchy' section at the top right is currently empty, showing '<Select dimension>'. Below the hierarchy are three buttons: 'Start Year', 'State', and 'County'.

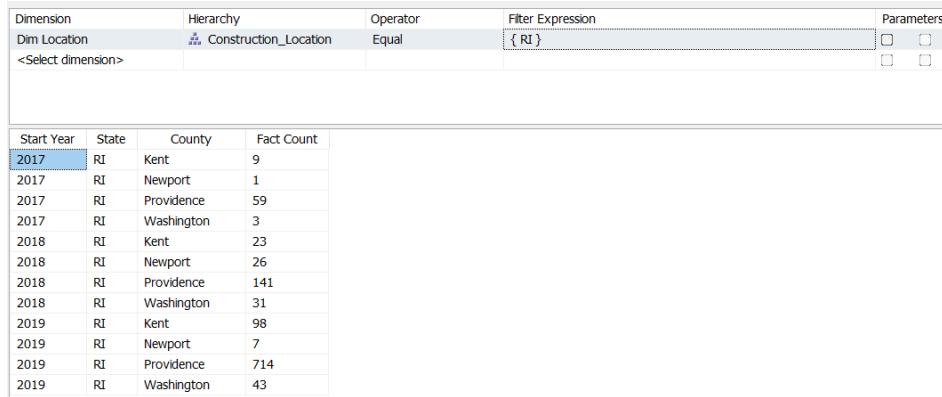
Bước 3: Kéo thả thuộc tính **Fact Count** vào cửa sổ thực thi truy vấn.

The screenshot shows the 'US ROAD WH' dimension structure. Under 'Measures', there is a 'Fact' folder containing several measures: Average Distance, Average Duration, Average Severity, Fact Count (which is selected), Max Distance, and Sum Distance. Under 'Dim Location', there are 'City' and 'County' dimensions. On the right, there is a 'Hierarchy' section with a dropdown menu set to '<Select dimension>'.

Bước 4: Thực hiện lọc **Hierarchy Construction_Location** của **Dim_Location**, chọn ra tiêu bang RI theo điều kiện rồi nhấn **OK**

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim Location	Construction_Location	Equal	{ RI }	
<Select dimension>				
Start Year State County Fact Count				

Bước 5: Click vào **Click to execute the query** để thực thi câu truy vấn và xem kết quả



The screenshot shows a Power BI interface. At the top, there is a filter pane with the following settings:

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim Location	Construction_Location	Equal	{ RI }	<input type="checkbox"/> <input type="checkbox"/>
<Select dimension>				

Below the filter pane is a pivot table with the following data:

Start Year	State	County	Fact Count
2017	RI	Kent	9
2017	RI	Newport	1
2017	RI	Providence	59
2017	RI	Washington	3
2018	RI	Kent	23
2018	RI	Newport	26
2018	RI	Providence	141
2018	RI	Washington	31
2019	RI	Kent	98
2019	RI	Newport	7
2019	RI	Providence	714
2019	RI	Washington	43

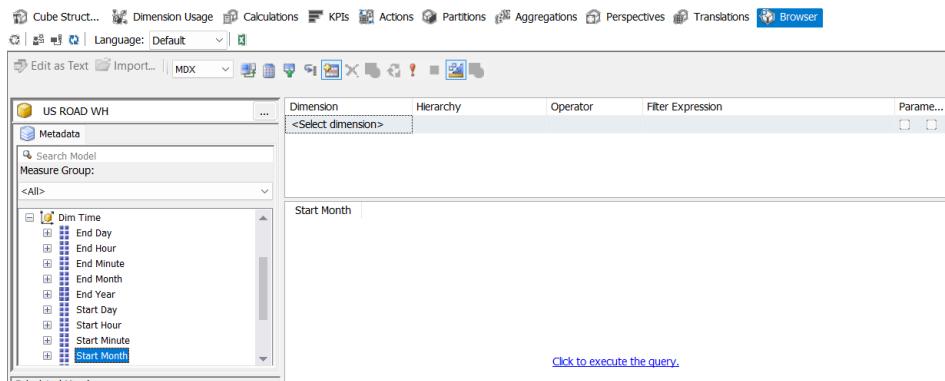
3.9.2. Câu truy vấn 2

Nội dung câu truy vấn: Thống kê tổng số công trình, mức độ ảnh hưởng trung bình, chiều dài trung bình và thời gian thi công trung bình của đoạn đường mà có công trình thi công ảnh hưởng tới đường bộ theo các tháng trong năm 2019.

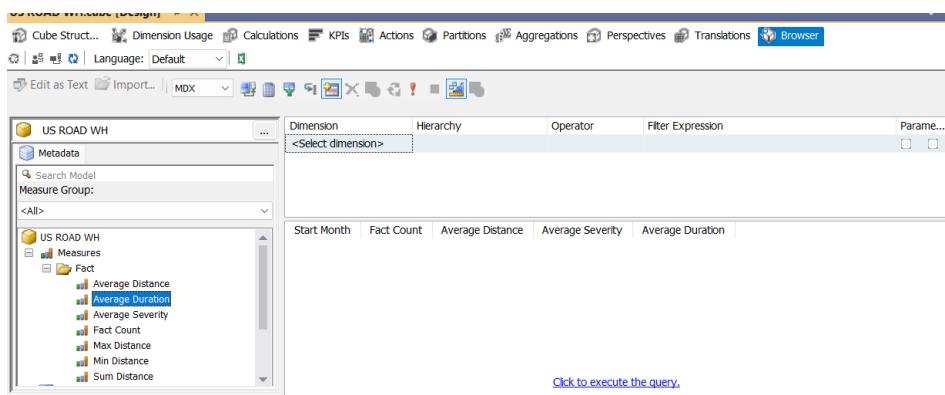
Ý nghĩa: Dùng để đánh giá mức độ xảy ra và mức độ nghiêm trọng của việc công trình thi công ảnh hưởng tới đường bộ theo tháng trong năm 2019. Qua câu truy vấn ta có thể thống kê được tháng nào trong năm thường xảy ra nhiều sự ảnh hưởng bởi công trình nhiều nhất cũng như các mức độ nghiêm trọng.

Kỹ thuật sử dụng: Pivot. Ở đây, dữ liệu được xoay để hiển thị các thông tin tổng hợp theo từng tháng.

Bước 1: Kéo thả thuộc tính **Start Month** vào cửa sổ thực thi câu truy vấn



Bước 2: Kéo các độ đo Fact Count, Average Distance, Average Severity, Average Duration vào cửa sổ thực thi truy vấn



Bước 3: Chọn điều kiện trong Filter Expression là năm 2019 và nhấn Click to execute the query để xem kết quả

Dimension	Hierarchy	Operator	Filter Expression	Param...
Dim Time	Construction_Start	Equal		
<Select dimension>				
Start Month Fact Count Average Distance Average Severity A				
<div style="border: 1px solid #ccc; padding: 10px;"> <input type="checkbox"/> All <input type="checkbox"/> 2017 <input type="checkbox"/> 2018 <input checked="" type="checkbox"/> 2019 <input type="checkbox"/> Unknown </div> <p style="text-align: center;">Click to execute the query</p>				
<input type="button" value="OK"/> <input type="button" value="Cancel"/>				

Dimension	Hierarchy	Operator	Filter Expression																																																																		
Dim Time	Construction_Start	Equal	{ 2019 }																																																																		
<Select dimension>																																																																					
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>Start Month</th> <th>Fact Count</th> <th>Average Severity</th> <th>Average Distance</th> <th>Average Duration</th> </tr> </thead> <tbody> <tr><td>1</td><td>8479</td><td>2.71871682981484</td><td>2.69838917166439</td><td>996630.190706451</td></tr> <tr><td>2</td><td>7333</td><td>2.70448656757125</td><td>2.8743362402352</td><td>959944.899222692</td></tr> <tr><td>3</td><td>10758</td><td>2.84597508830638</td><td>2.4828358236728</td><td>937432.960494516</td></tr> <tr><td>4</td><td>17061</td><td>2.83822753648672</td><td>2.75891873347705</td><td>707547.72258367</td></tr> <tr><td>5</td><td>17696</td><td>2.89862115732369</td><td>2.84304074139709</td><td>595100.788313743</td></tr> <tr><td>6</td><td>15147</td><td>2.86644219977553</td><td>2.57515363355587</td><td>628554.409454017</td></tr> <tr><td>7</td><td>16721</td><td>2.80886310627355</td><td>2.77802939503669</td><td>506609.414209676</td></tr> <tr><td>8</td><td>16563</td><td>2.81452635392139</td><td>2.46104032015158</td><td>372958.001026384</td></tr> <tr><td>9</td><td>14701</td><td>2.80953676620638</td><td>3.09815900058439</td><td>373538.748180396</td></tr> <tr><td>10</td><td>15228</td><td>2.77213028631468</td><td>2.56695560331212</td><td>282845.84489099</td></tr> <tr><td>11</td><td>7290</td><td>2.71838134430727</td><td>2.84743213925293</td><td>251070.105761317</td></tr> <tr><td>12</td><td>5444</td><td>2.76010286554004</td><td>2.67410724301782</td><td>174070.2404482</td></tr> </tbody> </table>					Start Month	Fact Count	Average Severity	Average Distance	Average Duration	1	8479	2.71871682981484	2.69838917166439	996630.190706451	2	7333	2.70448656757125	2.8743362402352	959944.899222692	3	10758	2.84597508830638	2.4828358236728	937432.960494516	4	17061	2.83822753648672	2.75891873347705	707547.72258367	5	17696	2.89862115732369	2.84304074139709	595100.788313743	6	15147	2.86644219977553	2.57515363355587	628554.409454017	7	16721	2.80886310627355	2.77802939503669	506609.414209676	8	16563	2.81452635392139	2.46104032015158	372958.001026384	9	14701	2.80953676620638	3.09815900058439	373538.748180396	10	15228	2.77213028631468	2.56695560331212	282845.84489099	11	7290	2.71838134430727	2.84743213925293	251070.105761317	12	5444	2.76010286554004	2.67410724301782	174070.2404482
Start Month	Fact Count	Average Severity	Average Distance	Average Duration																																																																	
1	8479	2.71871682981484	2.69838917166439	996630.190706451																																																																	
2	7333	2.70448656757125	2.8743362402352	959944.899222692																																																																	
3	10758	2.84597508830638	2.4828358236728	937432.960494516																																																																	
4	17061	2.83822753648672	2.75891873347705	707547.72258367																																																																	
5	17696	2.89862115732369	2.84304074139709	595100.788313743																																																																	
6	15147	2.86644219977553	2.57515363355587	628554.409454017																																																																	
7	16721	2.80886310627355	2.77802939503669	506609.414209676																																																																	
8	16563	2.81452635392139	2.46104032015158	372958.001026384																																																																	
9	14701	2.80953676620638	3.09815900058439	373538.748180396																																																																	
10	15228	2.77213028631468	2.56695560331212	282845.84489099																																																																	
11	7290	2.71838134430727	2.84743213925293	251070.105761317																																																																	
12	5444	2.76010286554004	2.67410724301782	174070.2404482																																																																	

3.9.3. Câu truy vấn 3

Nội dung câu truy vấn: Thống kê với 5 loại điều kiện thời tiết xảy ra nhiều nhất có tổng công trình được thực hiện nhiều nhất, theo từng tháng, năm và tính tổng chiều dài các đoạn đường có công trình thi công đó.

Ý nghĩa: mục đích xác định 5 loại điều kiện thời tiết xảy ra nhiều nhất mà có liên quan đến tổng số công trình thi công đường bộ cao nhất. Dữ liệu sẽ được phân tích theo từng tháng và từng năm, và tính tổng chiều dài các đoạn đường có công trình thi công dưới những điều kiện thời tiết này.

Việc xác định các loại thời tiết thường xuyên gây ra các công trình thi công nhiều nhất cũng sẽ giúp lập kế hoạch phòng ngừa và ứng phó tốt hơn, đảm bảo rằng các biện pháp bảo vệ và quản lý giao thông được triển khai kịp thời, giảm thiểu tác động tiêu cực đến người tham gia giao thông trong những điều kiện thời tiết này.

Kỹ thuật thực hiện: Drill Down đào sâu vào chi tiết về các điều kiện thời tiết, tổng công trình và chiều dài các đoạn đường theo từng tháng và năm.

Bước 1: Tạo NameSet [Top 5 Weather Condition] như hình bên dưới với biểu thức sau:

TOPCOUNT([Dim Weather].[Weather Condition].[Weather Condition],5,[Measures].[Fact Count])

The screenshot shows the SSAS Script Editor interface. In the left pane, there is a 'Script Organizer' with two items: '1 CALCULATE' and '2 [Top 5 Weather Condition]'. The second item is selected. In the main pane, there is a 'Name:' field containing '[Top 5 Weather Condition]' and an 'Expression' field containing 'TOPCOUNT([Dim Weather].[Weather Condition].[Weather Condition],5,[Measures].[Fact Count])'. Below the expression, there is a message 'No issues found' and some status information: Ln: 1 Ch: 91 SPC CRLF. There are also 'Additional Properties' sections for 'Type' (set to 'Dynamic') and 'Display folder'.

Bước 2: Kéo thả các hierarchies có tên là **Construction_Start** vào cửa sổ thực thi câu truy vấn.

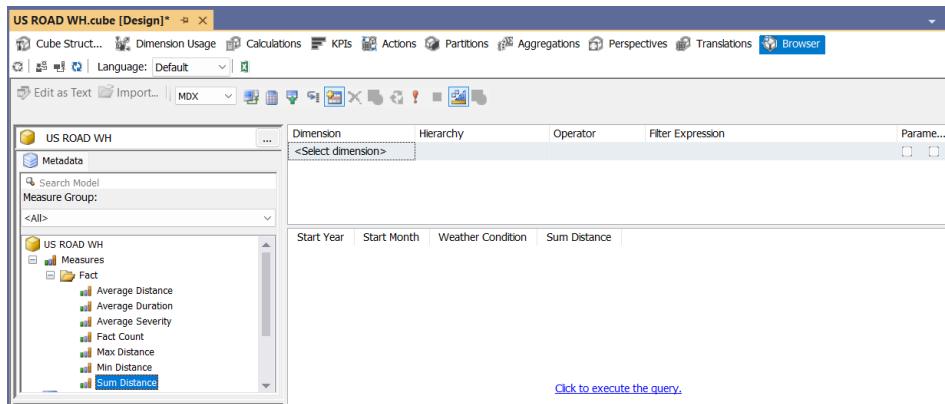
Ths. Đỗ Thị Minh Phụng

Cube Struct... Dimension Usage Calculations KPIs Actions Partitions Aggregations Perspectives Translations Browser
Edit as Text Import... MDX ...
Language: Default
Dimension Hierarchy Operator Filter Expression Parameter
<Select dimension>
Start Year Start Month Start Day Start Hour Start Minute
Click to execute the query.

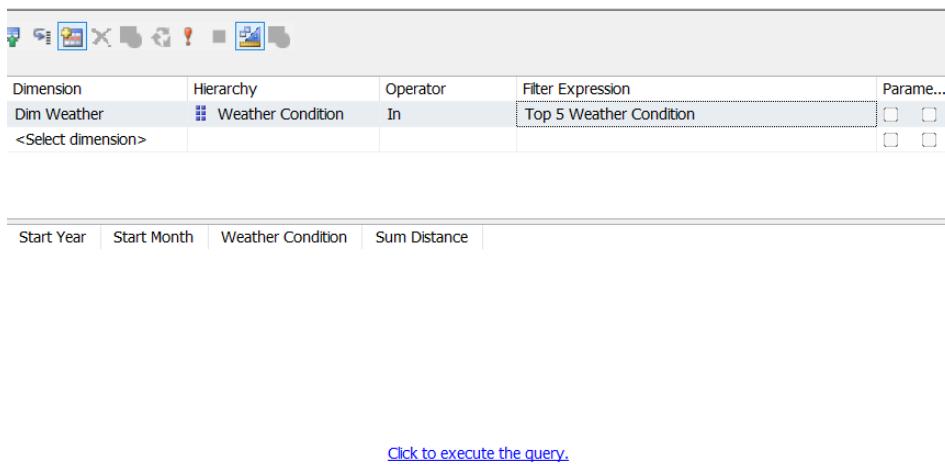
Bước 3: Loại bỏ các thuộc tính không cần dùng để còn các thuộc tính như hình bên dưới

Cube Struct... Dimension Usage Calculations KPIs Actions Partitions Aggregations Perspectives Translations Browser
Edit as Text Import... MDX ...
Language: Default
Dimension Hierarchy Operator Filter Expression Parameter
<Select dimension>
Start Year Start Month
Click to execute the query.

Bước 4: Kéo thả thuộc tính Weather Condition và độ đo Sum Distance vào cửa sổ thực thi câu truy vấn



Bước 5: Chọn NameSet [Top 5 Weather Condition] ở Filter Expression.



Bước 6: Nhấn vào Click to execute the query để thực thi câu truy vấn và xem kết quả

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim Weather	# Weather Condition	In	Top 5 Weather Condition	<input type="checkbox"/> <input type="checkbox"/>
<Select dimension>				
Start Year	Start Month	Weather Condition	Sum Distance	
2017	1	Clear	2367.94597999845	
2017	1	Fair	17.7583822384477	
2017	1	Mostly Cloudy	1078.74510004765	
2017	1	Overcast	3265.10455733849	
2017	1	Partly Cloudy	491.427991696517	
2017	2	Clear	4392.06524147249	
2017	2	Fair	38.8428800514666	
2017	2	Mostly Cloudy	1065.26204045713	
2017	2	Overcast	2172.61909123123	
2017	2	Partly Cloudy	636.432268098579	
2017	3	Clear	4272.53934276033	
2017	3	Fair	68.0551402061246	
2017	3	Mostly Cloudy	1078.39809921663	
2017	3	Overcast	2576.3179291104	
2017	3	Partly Cloudy	940.641004937977	
2017	4	Clear	5500.43487978468	
2017	4	Fair	159.948172128759	
2017	4	Mostly Cloudy	1838.17438358131	
2017	4	Overcast	4214.14708452085	

3.9.4. Câu Truy vấn 4

Nội dung câu truy vấn: Thông kê tổng số công trình (Fact Count) trên 6 tiểu bang (State): California (CA), New York (NY), Florida (FL), Ohio (OH), Texas (TX) và Washington (WA) có thời gian bắt đầu 2017 và kết thúc 2019.

Ý nghĩa câu truy vấn: để hiểu rõ quy mô và sự phân bố của các công trình xây dựng qua các tiểu bang được chọn trong một khoảng thời gian nhất định. Thông tin này sẽ hỗ trợ các nhà quản lý giao thông và các cơ quan lập kế hoạch đánh giá được khối lượng công việc và tác động của các công trình xây dựng đến giao thông tại các tiểu bang quan trọng này.

Kỹ thuật thực hiện: Slice and Dice.

Thực hiện câu truy vấn:

Bước 1: Kéo thả các thuộc tính **State** và **Fact Count** vào cửa sổ thực thi câu truy vấn

Ths. Đỗ Thị Minh Phụng

The screenshot shows a query builder interface with a single filter row. The columns are: Dimension, Hierarchy, Operator, Filter Expression, and Parameter. The Dimension is set to '<Select dimension>'. The Hierarchy is set to 'Construction_Location'. The Operator is set to 'Equal'. The Filter Expression dropdown contains the value '{ CA, FL, NY, OH, WA, TX }'. Below the table, there are tabs for 'State' and 'Fact Count', and a link 'Click to execute the query.'

Bước 2: Kéo thả Dimension [Dim Location] vào dòng đầu tiên của cửa sổ lọc với Hierarchy [Construction_Location], Operator [Equal] và lọc ra các giá trị {CA, FL, NY, OH, WA, TX} ở Filter Expression. Và kéo thả các Dim Time vào tương tự như Dim Location.

The screenshot shows a query builder interface with two filter rows. The first row has the Dimension set to 'Dim Location', Hierarchy to 'Construction_Location', Operator to 'Equal', and Filter Expression to '{ CA, FL, NY, OH, WA, TX }'. The second row has the Dimension set to 'Dim Time', Hierarchy to 'Construction_Start', Operator to 'Equal', and Filter Expression to '{ 2017 }'. The third row has the Dimension set to 'Dim Time', Hierarchy to 'Construction_End', Operator to 'Equal', and Filter Expression to '{ 2019 }'. Below the table, there are tabs for 'State' and 'Fact Count'.

Bước 3: Nhấn **Click to execute the query** để thực thi câu lệnh và được kết quả

State	Fact Count
CA	3
FL	1
OH	8
TX	19
WA	1

3.9.5. Câu truy vấn 5

Nội dung câu truy vấn: Thông kê số công trình và trung bình độ dài đoạn đường thi công với nhiệt độ trên 70F và độ ẩm trên 70%.

Ý nghĩa câu truy vấn: để hiểu rõ quy mô và phân bố của các công trình xây dựng trong điều kiện thời tiết cụ thể. Thông tin này có thể giúp các nhà quản lý giao thông và quy hoạch xác định mức độ ảnh hưởng của điều kiện thời tiết đến quá trình xây dựng, từ đó có thể đưa ra các biện pháp điều chỉnh và cải thiện quản lý dự án.

Ngoài ra, việc tính toán độ dài trung bình của các đoạn đường thi công dưới điều kiện thời tiết nhất định cũng giúp đánh giá hiệu quả và tác động của các công trình xây dựng đến giao thông và môi trường xung quanh.

Kỹ thuật thực hiện: Slice and Dice lọc dữ liệu theo các điều kiện thời tiết cụ thể.

Thực hiện câu truy vấn:

Bước 1: Tạo NameSet [Temperature >70] để lọc ra những công trình có nhiệt độ trên 70 độ F.

```
FILTER( [Dim Weather].[Temperature F].CHILDREN, [Dim Weather].[Temperature F].CurrentMember.MemberValue > 70)
```

Name: [Temperature > 70]

Expression

```
FILTER( [Dim Weather].[Temperature F].CHILDREN,
        [Dim Weather].[Temperature F].CurrentMember.MemberValue > 70)
```

No issues found

Ln: 2 Ch: 70 SPC CRLF

Additional Properties

Type: Dynamic

Display folder:

Bước 2: Tạo NameSet [Humidity >70%] để lọc ra những công trình có độ ẩm trên 70%.

```
FILTER( [Dim Weather].[Humidity].CHILDREN,
        [Dim Weather].[Humidity].CurrentMember.MemberValue > 70)
```

Name: [Humidity > 70%]

Expression

```
FILTER( [Dim Weather].[Humidity].CHILDREN,
        [Dim Weather].[Humidity].CurrentMember.MemberValue > 70)
```

No issues found

Ln: 1 Ch: 1 SPC CRLF

Additional Properties

Type: Dynamic

Display folder:

Bước 3: Kéo thả các thuộc tính Temperature F, Humidity, Fact Count và Average Distance vào cửa sổ thực thi câu truy vấn.

The screenshot shows a user interface for executing a query. At the top, there are four tabs: Temperature F, Humidity, Fact Count, and Average Distance. Below these tabs is a large empty area. In the bottom right corner of this area, there is a blue link labeled "Click to execute the query.". Above this link, the text "Bước 4: Kéo thả Dimension [Dim Weather] vào dòng đầu tiên của cửa sổ lọc với Hierarchy [Tempature F], Operator [In] và kéo NameSet [Tempature > 70] ở Filter Expression. Và kéo thả Dim Weather và thực hiện tương tự với NameSet [Humidity> 70%].".

Dimension	Hierarchy	Operator	Filter Expression	Param...
Dim Weather	# Temperature F	In	Temperature > 70	<input type="checkbox"/> <input type="checkbox"/>
Dim Weather	# Humidity	In	Humidity > 70%	<input type="checkbox"/> <input type="checkbox"/>
<Select dimension>				<input type="checkbox"/> <input type="checkbox"/>

At the very bottom of the interface, there is a row of buttons: Start Year, Temperature F, Humidity, and Average Distance.

Bước 5: Nhấn vào **Click to execute the query** để thực thi câu truy vấn.

Temperature F	Humidity	Fact Count	Average Distance
100.4000015...	71	1	0.00766466045...
70.19999694...	100	1	3.67000007629...
70.19999694...	71	1	1.16754066944...
70.19999694...	73	3	7.73789385954...
70.19999694...	74	3	1.48787854115...
70.19999694...	77	1	0.07610085606...
70.19999694...	82	1	0.32292965054...
70.19999694...	84	1	8.13000011444...
70.19999694...	85	1	14.8000001907...
70.19999694...	86	5	10.2360002487...
70.19999694...	87	4	0.09240888059...
70.19999694...	88	2	2.66487360000...
70.19999694...	89	1	0.32267323136...
70.19999694...	90	1	0.01537467911...
70.19999694...	93	2	1.13536752760...
70.19999694...	94	2	0.08726157993...

3.9.6. Câu truy vấn 6

Nội dung câu truy vấn: Thông kê theo từng bang có số lượng các công trình thi công lớn 10000.

Ý nghĩa câu truy vấn: mục đích đếm số lượng các công trình thi công ở mỗi bang mà có số lượng lớn hơn 10000.

Mục tiêu của việc thống kê này là để xác định các bang có mức độ hoạt động xây dựng cao, đánh giá tình trạng phát triển hạ tầng và xác định các khu vực mà có nhiều công trình xây dựng diễn ra.

Kỹ thuật thực hiện: Drill Down bắt đầu từ tổng số công trình và sau đó đi sâu hơn để thống kê theo từng bang. Pivot hiển thị thông tin theo nhiều chiều khác nhau (số lượng công trình theo bang).

Thực hiện câu truy vấn:

Bước 1: Tạo NameSet [Fact Count > 10000] như hình minh họa với biểu thức như sau:

```
FILTER( [Dim Location].[Construction_Location].[State].ALLMEMBERS,
[Measures].[Fact Count] > 10000)
```

Name: [Fact Count >10000]

Expression:

```
FILTER( [Dim Location].[Construction_Location].[State].ALLMEMBERS,
[Measures].[Fact Count] > 10000)
```

No issues found

Additional Properties:

- Type: Dynamic
- Display folder:

Bước 2: Kéo thả các thuộc tính State, Fact Count vào cửa sổ thực thi truy vấn

Bước 3: Kéo thả NameSet Fact Count > 10000 ở bảng Dim Location vào cửa sổ lọc

Dimension	Hierarchy	Operator	Filter Expression	Param...
Dim Location	Construction_Location	In	Fact Count >10000	
<Select dimension>				

Bước 4: Nhấn **Click to execute the query** để thực thi câu lệnh và được kết quả như sau

State	Fact Count
CA	30750
FL	12539
IL	11439
MI	10060
NY	21328
OH	16232
OR	10343
PA	66502
TX	14773

3.9.7. Câu truy vấn 7

Nội dung câu truy vấn: Liệt kê 10 tiểu bang bị ảnh hưởng đến giao thông đường bộ bởi các công trình xây dựng nhiều nhất trong 3 năm từ 2017 tới 2019. Với 10 tiểu bang đó, liệt kê tên các con đường và phía thi công có tổng độ dài đoạn đường bị ảnh hưởng bởi các công trình lớn hơn 100 miles.

Ý nghĩa câu truy vấn:

- **Xác định 10 tiểu bang chịu ảnh hưởng nhiều nhất đến giao thông đường bộ do các công trình xây dựng.** Việc này giúp nhận biết các khu vực có hoạt động xây dựng sôi động và có thể gây ra các vấn đề liên quan đến giao thông.
- **Liệt kê tên các con đường và phía thi công có tổng độ dài đoạn đường bị ảnh hưởng lớn hơn 100 miles trong mỗi tiểu bang.** Thông tin này giúp cung cấp cái nhìn chi tiết hơn về các vị trí cụ thể và quy mô của các công trình xây dựng đang ảnh hưởng đến giao thông trong mỗi tiểu bang, từ đó có thể đưa ra các biện pháp giảm thiểu tác động và quản lý giao thông một cách hiệu quả.

Kỹ thuật thực hiện: Drill Down từ tổng quan (10 tiểu bang bị ảnh hưởng nhất) đến các con đường cụ thể trong các tiểu bang đó và Slice and Dice theo các điều kiện cụ thể như thời gian (2017-2019) và điều kiện độ dài đoạn đường bị ảnh hưởng (> 100 miles).

Thực hiện câu truy vấn:

Bước 1: Tạo NameSet [Top 10 Count] như hình bên dưới với biểu thức sau:

```
TOPCOUNT([Dim Location].[State].[State] ,10,[Measures].[Fact Count])
```

Name: [Top 10 Count]

Expression: TOPCOUNT([Dim Location].[State].[State] ,10,[Measures].[Fact Count])
No issues found Ln: 1 Ch: 69 SPC CRLF

Additional Properties:
Type: Dynamic
Display folder:

Bước 2: Tạo NameSet [Filter SumDistance] như hình bên dưới với biểu thức sau

```
FILTER([Dim Location].[Street].[Street], [Measures].[Sum Distance] > 100)
```

Name: [Filter SumDistance]

Expression: FILTER([Dim Location].[Street].[Street], [Measures].[Sum Distance] > 100)
No issues found Ln: 1 Ch: 74 SPC CRLF

Additional Properties:
Type: Dynamic
Display folder:

Bước 3: Kéo thả các thuộc tính **State, Street, Side, Sum Distance** vào cửa sổ thực thi truy vấn

The screenshot shows the 'Dimension Filter' dialog box from SQL Server Analysis Services. On the left, there's a tree view of the 'US ROAD WH' cube with nodes like 'Metadata', 'Search Model', 'Measure Group', and various dimension and fact tables. A dropdown menu 'Measure Group:' is set to '<All>'. On the right, there are two tabs: 'Dimension' and 'Hierarchy'. The 'Dimension' tab has a dropdown 'Select dimension' which is currently empty. Below it, there are four buttons: 'State', 'Street', 'Side', and 'Sum Distance'. The 'Hierarchy' tab is also empty.

Bước 4: Kéo 2 NameSet trên vào ô **Filter Expression** như hình bên dưới

Dimension	Hierarchy	Operator	Filter Expression	Param...
Dim Location	# State	In	Top 10 Count	<input type="checkbox"/> <input type="checkbox"/>
Dim Location	# Street	In	Filter SumDistance	<input type="checkbox"/> <input type="checkbox"/>
<Select dimension>				<input type="checkbox"/> <input type="checkbox"/>

Bước 5: Click vào **Click to execute the query** để thực thi câu truy vấn và xem kết quả

Ths. Đỗ Thị Minh Phụng

Dimension	Hierarchy	Operator	Filter Expression
Dim Location	# State	In	Top 10 Count
Dim Location	# Street	In	Filter SumDistance
<Select dimension>			
State	Street	Side	Sum Distance
CA	I-680 N	R	102.936012983322
CA	I-880 N	R	133.682976350188
CA	I-880 N	R	107.256878058775
CA	I-680 S	R	137.019623935223
CA	I-880 N	R	163.006126448512
CA	I-680 N	R	192.348910272121
CA	I-680 S	R	109.17274902761
CA	I-580 E	R	101.895470261574
CA	I-580 W	R	337.925602316856
CA	CA-88	R	124.068530082703
CA	CA-20	R	200.721193224192
CA	I-80 W	R	130.721097677946
CA	I-680 S	R	127.440188527107
CA	I-80 E	R	123.441954672337
CA	I-680 S	R	117.929034352303
CA	E Church Ave	R	174.011821180582
CA	I-8 W	R	196.415846049786
CA	I-8 W	R	528.45437130332

3.9.8. Câu truy vấn 8

Nội dung câu truy vấn: Với từng thành phố thuộc quận Orange của bang California, hãy cho biết độ dài đường thi công lớn nhất, tổng độ dài thi công và tỉ lệ giữa công trình thi công dài nhất so với tổng độ dài trong năm 2019. (Khó)

Ý nghĩa câu truy vấn: Mục tiêu của việc phân tích này là để hiểu rõ hơn về quy mô và phân bố của các công trình xây dựng trong một khu vực cụ thể của California. Thông tin này có thể hỗ trợ các nhà quản lý địa phương và quy hoạch xác định mức độ phát triển và nhu cầu về hạ tầng trong khu vực này.

Kỹ thuật thực hiện: Drill Down và Roll Up - Xác định và tổng hợp dữ liệu về số lượng công trình theo từng bang và từng năm.

Thực hiện câu truy vấn:

Bước 1: Tạo 1 Calculated Member [TiLeDoanDuongThiCongDaiNhatSoVoiTong] như hình dưới với biểu thức sau: [Measures].[Max Distance]/[Measures].[Sum Distance]

Ths. Đỗ Thị Minh Phụng

Name:

Parent Properties

Parent hierarchy:

Parent member:

Change

Expression

No issues found

Ln: 1 Ch: 1 SPC CRLF

Additional Properties

Format string:

Visible:

Non-empty behavior:

Associated measure group:

Display folder:

Color Expressions

Font Expressions

Bước 2: Kéo các thuộc tính City, độ đo Max Distance, Sum Distance và TiLeDoanDuongThiCongDaiNhatSoVoiTong vào cửa sổ thực thi câu truy vấn

The screenshot shows a query editor interface. On the left, there is a tree view of dimensions and measures under the heading 'KPIs'. Dimensions listed include 'Dim Location' with filters 'Fact Count > 10000', 'Filter SumDistance', and 'Top 10 Count'. Measures listed include 'Average Severity', 'Fact Count', 'Max Distance', 'Sum Distance', and 'TiLeDoanDuongThiCongDaiNhatSoVoiTong'. In the center, there are four columns: 'City', 'Max Distance', 'Sum Distance', and 'TiLeDoanDuongThiCongDaiNhatSoVoiTong'. Below these columns is a button labeled 'Click to execute the query.'

Bước 3: Chọn điều kiện là năm 2019 và quận Orange vào ô Filter Expression:

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim Time	Construction_Start	Equal	{ 2019 }	<input type="checkbox"/> <input type="checkbox"/>
Dim Location	Construction_Location	Equal	{ Orange }	<input type="checkbox"/> <input type="checkbox"/>
<Select dimension>				<input type="checkbox"/> <input type="checkbox"/>

Bước 4: Nhấn vào Click to execute the query để thực thi câu truy vấn và xem kết quả

City	Max Distance	Sum Distance	TiLeDoanDuongThiCongDaiNhatSoVoiTong
Aliso Viejo	0.518669307231903	0.929970532655716	0.557726604251362
Anaheim	15.5600004196167	70.3692457675934	0.221119329188298
Anaheim-Santa An...	1.0099999046326	2.0199998092651	0.5
Brea	9.6899995803833	87.2613822109997	0.111045680630553
Buena Park	8.39999961853027	72.2873094491661	0.116202964013169
Capistrano Beach	2	2.16181001067162	0.925150679350705
Costa Mesa	24.3199996948242	418.033287860453	0.0581771844517384
Dana Point	0.27480211853981	0.900315329432487	0.305228745480799
Foothill Ranch	0.140690609812737	0.140690609812737	1
Fountain Valley	5.44999980926514	277.520545261912	0.0196381850003994
Fullerton	4.03999996185303	55.8531887549907	0.0723324854302439
Garden Grove	3.2128221988678	10.5776039212942	0.303738183313891
Huntington Beach	6.15999984741211	67.2168032201007	0.0916437490673464
Irvine	8.80000019073486	214.734374312684	0.040980864004199
La Habra	8.06999969482422	39.4189144149423	0.204724047189011
Ladera Ranch	0.220150679349899	0.392050355672836	0.561536741809804
Laguna Beach	3.3199999332428	13.0640024095774	0.254133444648546
Laguna Hills	7.17999982833862	22.7533977627754	0.315557258884873
Laguna Niguel	2.2199999332428	10.6154116302136	0.218550161980782

3.9.9. Câu truy vấn 9

Nội dung câu truy vấn: Thống kê tổng số công trình theo từng ngày, tháng, năm.

Ý nghĩa câu truy vấn: mục đích để đếm tổng số công trình thi công được thực hiện trong mỗi ngày, mỗi tháng và mỗi năm.

Mục tiêu của việc thống kê này là để hiểu rõ hơn về xu hướng và biến động của các hoạt động xây dựng qua thời gian. Thông tin này có thể hỗ trợ các nhà quản lý xác định mức độ tải động trong quá trình thi công, phân bổ tài nguyên và lập kế hoạch công việc một cách hiệu quả.

Kỹ thuật thực hiện: Roll Up dữ liệu chi tiết về từng công trình được gộp lại để tạo thành tổng số công trình theo từng ngày, tháng, năm.

Thực hiện câu truy vấn:

Bước 1: Kéo các thuộc tính **Construction_Start** vào cửa sổ thực thi truy vấn.

The screenshot shows the 'Start' fact table in the SSAS Data Source View (DSV). The table has five columns: Start Year, Start Month, Start Day, Start Hour, and Start Minute. The 'Start Year' column is highlighted. On the left, the 'Dimensions' pane lists various dimensions like End Year, Start Day, Start Hour, etc., with 'Construction_Start' selected. Below the table, there is a link labeled 'Click to execute the query.'

Bước 2: Xóa các thuộc tính không cần thiết và thêm độ đo Fact Count vào cửa sổ thực thi câu truy vấn

The screenshot shows the 'Start' fact table in the SSAS Data Source View (DSV). The table now includes an additional column, 'Fact Count'. The 'Search Model' pane on the left shows the 'US ROAD WH' model selected. The 'Measures' section under 'Fact' contains the 'Fact Count' measure, which is highlighted. Below the table, there is a link labeled 'Click to execute the query.'

Bước 3: Nhấn vào click to execute the query để thực hiện câu truy vấn và xem kết quả

Start Year	Start Month	Start Day	Fact Count
2017	1	1	15
2017	1	2	51
2017	1	3	102
2017	1	4	157
2017	1	5	148
2017	1	6	82
2017	1	7	55
2017	1	8	46
2017	1	9	146
2017	1	10	134
2017	1	11	151
2017	1	12	176
2017	1	13	112
2017	1	14	27
2017	1	15	25
2017	1	16	79
2017	1	17	130
2017	1	18	172
---	---	---	---

3.9.10. Câu truy vấn 10

Nội dung câu truy vấn: Liệt kê số công trình thi công có các cơ sở vật chất giao thông gần đó như giao lộ (junction) và bùng binh (roundabout) theo thành phố, quận, bang trong 3 năm 2017- 2019.

Ý nghĩa câu truy vấn:

xác định được mức độ tập trung của các công trình xây dựng gây ảnh hưởng đến giao thông tại các khu vực khác nhau và thấy được sự tương quan giữa các công trình này với sự hiện diện của các cơ sở vật chất giao thông quan trọng. Điều này sẽ hỗ trợ trong việc lập kế hoạch và quản lý giao thông hiệu quả hơn, đồng thời cũng giúp đưa ra các quyết định về đầu tư và cải thiện hạ tầng giao thông tại các khu vực có nhu cầu cao.

Kỹ thuật thực hiện: Pivot xoay trục dữ liệu để hiển thị thông tin theo nhiều chiều khác nhau (thành phố, quận, bang). Slice and Dice lọc dữ liệu theo các tiêu chí cụ thể như thời

gian (2017-2019) và điều kiện có giao lộ (junction= 'Yes'), có bùng binh (roundabout='Yes').

Thực hiện câu truy vấn:

Bước 1: Tạo NameSet [Filter Junction] như hình bên dưới với biểu thức sau

```
FILTER( [Dim POI].[Junction].CHILDREN,
        [Dim POI].[Junction].CurrentMember.MemberValue <> 0)
```

Name: [Filter Junction]

Expression

```
FILTER( [Dim POI].[Junction].CHILDREN,
        [Dim POI].[Junction].CurrentMember.MemberValue <> 0)
```

No issues found

Ln: 1 Ch: 1 SPC CRLF

Tạo NameSet [Filter Roundabout] như hình bên dưới với biểu thức sau

```
FILTER( [Dim POI].[Roundabout].CHILDREN,
        [Dim POI].[Roundabout].CurrentMember.MemberValue <> 0)
```

Name: [Filter Roundabout]

Expression

```
FILTER( [Dim POI].[Roundabout].CHILDREN,
        [Dim POI].[Roundabout].CurrentMember.MemberValue <> 0)
```

No issues found

Ln: 2 Ch: 62 SPC CRLF

Bước 2: Kéo các thuộc tính State, County, City, và độ đo Fact Count vào cửa sổ thực thi câu truy vấn.

Bước 3: Kéo Dim Time và Dim POI vào cửa sổ chọn điều kiện với điều kiện Dim Time là All và Dim POI là 2 NameSet đã tạo ở bước 1.

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim POI	Junction	In	Filter Junction	<input type="checkbox"/> <input checked="" type="checkbox"/>
Dim POI	Roundabout	In	Filter Roundabout	<input checked="" type="checkbox"/> <input type="checkbox"/>
Dim Time	Construction_Start	Equal	{ All }	<input type="checkbox"/> <input type="checkbox"/>
<Select dimension>				<input type="checkbox"/> <input type="checkbox"/>
State	County	City	Fact Count	

Bước 4: Click vào **Click to execute the query** để thực hiện câu truy vấn và xem kết quả

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim POI	Junction	In	Filter Junction	<input type="checkbox"/> <input type="checkbox"/>
Dim POI	Roundabout	In	Filter Roundabout	<input checked="" type="checkbox"/> <input type="checkbox"/>
Dim Time	Construction_Start	Equal	{ All }	<input type="checkbox"/> <input type="checkbox"/>
<Select dimension>				<input type="checkbox"/> <input type="checkbox"/>
State	County	City	Fact Count	
CA	Los An...	Pala...	2	
FL	Miami-...	Hiale...	1	
GA	Clarke	Athens	1	
WA	Pierce	Tacoma	1	

3.9.11. Câu truy vấn 11

Nội dung câu truy vấn: Cho biết số công trình và mức độ nghiêm trọng trung bình của các công trình đó xảy ra tại bang Texas trong điều kiện thời tiết có liên quan tới bão (Thunderstorm) trong 3 năm 2017 – 2019.

Ý nghĩa câu truy vấn: giúp xác định mức độ ảnh hưởng của các điều kiện thời tiết xấu, cụ thể là bão, đến các công trình xây dựng và tình trạng giao thông tại bang Texas. Điều này có thể hỗ trợ các cơ quan quản lý và quy hoạch giao thông đưa ra các biện pháp phòng ngừa và ứng phó hiệu quả, giảm thiểu tác động tiêu cực của thời tiết xấu đối với giao thông và các công trình xây dựng. Ngoài ra, thông tin này còn có thể dùng để đánh giá và cải thiện các quy trình xây dựng và bảo trì công trình trong điều kiện thời tiết khắc nghiệt.

Kỹ thuật thực hiện: Slice and Dice truy vấn lọc dữ liệu để chỉ lấy các công trình tại bang Texas, xảy ra trong điều kiện thời tiết bão (Thunderstorm).

Thực hiện câu truy vấn:

Bước 1: Kéo thả Start_year, Fact_Count, Average_Severity vào cửa sổ thực thi câu truy vấn.

Start Year | Fact Count | Average Severity |

[Click to execute the query.](#)

Bước 2: Thực hiện lọc **Hierarchy** cho Weather Condition của bảng Dim_weather là {Heavy Thunderstorms and Rain, Light Thunderstorms and Rain} và lọc State của Dim Location là {TX}

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim Location	# State	Equal	{ TX }	<input type="checkbox"/> <input checked="" type="checkbox"/>
Dim Weather	# Weather Condition	Equal	{ Heavy Thunderstorms and Rain, Light Thunderstorms and Rai... }	<input type="checkbox"/> <input checked="" type="checkbox"/>
<Select dimension>				<input type="checkbox"/> <input type="checkbox"/>

Start Year | Fact Count | Average Severity |

Bước 3: Click vào **Click to execute the query** để thực thi câu truy vấn và xem kết quả

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim Location	# State	Equal	{ TX }	<input type="checkbox"/> <input checked="" type="checkbox"/>
Dim Weather	# Weather Condition	Equal	{ Heavy Thunderstorms and Rain, Light Thunderstorms and Rai... }	<input type="checkbox"/> <input checked="" type="checkbox"/>
<Select dimension>				<input type="checkbox"/> <input type="checkbox"/>

Start Year | Fact Count | Average Severity |

2017	10	2
2018	52	3.69230769230...
2019	9	2 ????????????

3.9.12. Câu truy vấn 12:

Nội dung câu truy vấn: Cho biết tiểu bang có số công trình nhiều nhất và tiểu bang có số công trình ít nhất theo từng năm.

Ý nghĩa câu truy vấn: mục đích xác định tiểu bang có số lượng công trình xây dựng gây ảnh hưởng đến giao thông đường bộ nhiều nhất và ít nhất theo từng năm.

Kỹ thuật thực hiện: Drill Down: Để xác định chi tiết số lượng công trình theo từng tiểu bang và từng năm. Roll Up: Để tổng hợp dữ liệu và sau đó so sánh để xác định tiểu bang có số công trình nhiều nhất và ít nhất trong từng năm.

Thực hiện câu truy vấn:

Bước 1: Kéo thả thuộc tính [State] từ bảng [Dim Location], [Start Year] từ bảng [Dim Time], [Fact Count] từ bảng Fact vào cửa sổ truy vấn.

Start Year | State | Fact Count

[Click to execute the query.](#)

Bước 2: Chọn vào Design Mode để có thể chỉnh sửa câu truy vấn.

```
SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS, NON EMPTY {[Dim Time].[Start Year].[Start Year].ALLMEMBERS * [Dim Location].[State].[State].ALLMEMBERS} ON ROWS
DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME
CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR,
FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

Bước 3: Chỉnh sửa câu truy vấn để có được kết quả mong muốn

```
SELECT NON EMPTY { [Measures].[Fact Count]} ON COLUMNS,
NON EMPTY {
([Dim Time].[Construction_Start].[Start Year])*(GENERATE(
Descendants([Dim Time].[Start Year].[Start Year].Members, [Dim Time].[Start Year].[Start Year].[Start Month]),
```

Ths. Đỗ Thị Minh Phụng

```
UNION (
    TopCount([Dim Location].[State].Children, 1, [Measures].[Fact Count]),
    BottomCount([Dim Location].[State].Children, 1,[Measures].[Fact Count] )
)
))

} DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS
FROM [US ROAD WH] CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR,
FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

```
SELECT NON EMPTY { [Measures].[Fact Count]} ON COLUMNS,
NON EMPTY {
    {[Dim Time].[Construction_Start].[Start Year]}*{GENERATE(
        Descendants([Dim Time].[Start Year].[Start Year].Members, [Dim Time].[Start Year].[Start Month]),
        UNION (
            TopCount([Dim Location].[State].Children, 1, [Measures].[Fact Count]),
            BottomCount([Dim Location].[State].Children, 1,[Measures].[Fact Count] )
        )
    ))
}
} DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS
FROM [US ROAD WH] CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

Start Year	State	Fact Count
2017	PA	19130
2017	AL	45
2018	PA	18495
2018	AL	268
2019	PA	28882
2019	AL	1514

3.9.13. Câu truy vấn 13

Nội dung câu truy vấn: Cho biết tốc độ hoàn thành trung bình của mỗi công trình ở gần vị trí gờ giảm tốc (bump) theo từng bang trong cả 3 năm 2017 – 2019.

Ý nghĩa câu truy vấn: mục đích tính toán tốc độ hoàn thành trung bình của các công trình xây dựng có ảnh hưởng đến giao thông đường bộ tại các vị trí gần gò giảm tốc (bump) theo từng bang, trong khoảng thời gian từ năm 2017 đến năm 2019.

Kỹ thuật thực hiện: Drill Down để chi tiết hóa dữ liệu từ tổng quát đến chi tiết về tốc độ hoàn thành trung bình của các công trình gần gò giảm tốc theo từng bang trong 3 năm.

Thực hiện câu truy vấn:

Bước 1: Tạo NameSet [**Filter Bump**] như hình bên dưới với biểu thức sau

```
FILTER( [Dim POI].[Bump].CHILDREN,  
        [Dim POI].[Bump].CurrentMember.MemberValue <> 0)
```

Name:

Expression ✖

```
FILTER( [Dim POI].[Bump].CHILDREN,  
        [Dim POI].[Bump].CurrentMember.MemberValue <> 0)
```

No issues found Ln: 2 Ch: 56 SPC CRLF

Bước 2: Tạo Calculated Member [**TocDoHoanThanhTrungBinh**] như hình bên dưới với biểu thức sau với Format string là “**Standard**”

```
IIF(ROUND(([Measures].[Sum Duration] / 86400) / [Measures].[Fact Count], 0) = 0, 1,  
    ROUND(([Measures].[Sum Duration] / 86400) / [Measures].[Fact Count], 0)
```

Ths. Đỗ Thị Minh Phụng

Name: [TocDoHoanThanhTrungBinh]

Parent Properties

Parent hierarchy: Measures Change

Expression

```
IIF(ROUND(([Measures].[Sum Duration] / 86400) / [Measures].[Fact Count], 0) = 0,  
1,  
ROUND(([Measures].[Sum Duration] / 86400) / [Measures].[Fact Count], 0))  
No issues found
```

Ln: 1 Ch: 5 SPC CRLF

Additional Properties

Format string: "Standard"

Visible: True

Non-empty behavior:

Associated measure group: (Undefined)

Display folder:

Color Expressions

Font Expressions

Bước 3: Kéo thả các thuộc tính State, [TocDoHoanThanhTrungBinh], Sum Duration, Fact Count vào trong cửa sổ thực thi câu truy vấn

State	TocDoHoanThanhTrungBinh	Sum Duration	Fact Count
-------	-------------------------	--------------	------------

[Click to execute the query.](#)

Bước 4: Kéo Dim POI vào cửa sổ chọn điều kiện với điều kiện cho cột Bump của Dim POI là NameSet Filter Bump đã tạo lúc trước

Ths. Đỗ Thị Minh Phụng

The screenshot shows a data analysis interface with the following components:

- Query Configuration:** A top bar with sections for Dimension (Dim POI), Hierarchy (Bump), Operator (In), Filter Expression (Filter Bump), and Parameters.
- Result Headers:** Below the configuration are headers for the results table: State, TocDoHoanThanhTrungBinh, Sum Duration, and Fact Count.
- Table Data:** The main area displays a table with two columns. The first column is labeled "State" and the second is "TocDoHoanThanhTrungBinh". The data rows are:
 - AL 1
 - AR 1
 - AZ 4
 - CA 1
 - CO 1
 - CT 1
 - DC 1
 - DE 1
 - FL 4
 - GA 1
 - IA 1
 - ID 1
 - IL 1
 - IN 26
 - KS 1
 - KY 1
- Action:** A link at the bottom left of the table area says "Click to execute the query."

Bước 5: Click vào **Click to execute the query** để xem kết quả truy vấn

State	TocDoHoanThanhTrungBinh
AL	1
AR	1
AZ	4
CA	1
CO	1
CT	1
DC	1
DE	1
FL	4
GA	1
IA	1
ID	1
IL	1
IN	26
KS	1
KY	1

3.9.14. Câu truy vấn 14

Nội dung câu truy vấn: Thông kê trung bình mức độ nghiêm trọng của các công trình thi công đường tại các bang ở dưới mức 3 trong năm 2017 nhưng đã lên hơn mức 3 trong năm 2018.

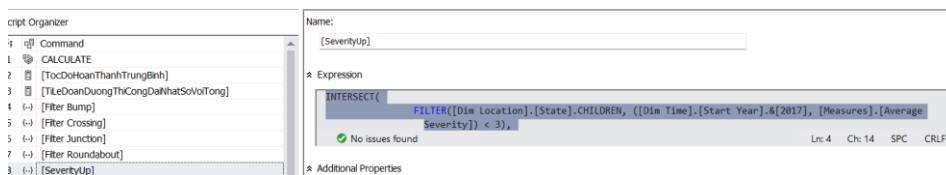
Ý nghĩa câu truy vấn: xác định và theo dõi sự gia tăng mức độ nghiêm trọng của các công trình thi công đường bộ qua hai năm liên tiếp, từ 2017 đến 2018. Kết quả của truy vấn này có thể giúp các cơ quan quản lý giao thông và các nhà lập kế hoạch phát hiện ra các xu hướng tiêu cực hoặc các vấn đề tiềm ẩn trong quá trình thi công, nhằm tìm hiểu nguyên nhân và đưa ra các biện pháp điều chỉnh, cải thiện hiệu quả và an toàn của các công trình xây dựng.

Kỹ thuật thực hiện: Slice and Dice để thực hiện việc lọc và phân tích dữ liệu dựa trên các tiêu chí cụ thể về mức độ nghiêm trọng và thời gian.

Thực hiện câu truy vấn:

Bước 1: Tạo NameSet [SeverityUp] như hình bên dưới với biểu thức sau:

```
INTERSECT( FILTER([Dim Location].[State].CHILDREN, ([Dim Time].[Start Year]&[2017], [Measures].[Average Severity]) < 3), FILTER([Dim Location].[State].CHILDREN, ([Dim Time].[Start Year]&[2018], [Measures].[Average Severity]) > 3))
```



Bước 2: Kéo thả các thuộc tính **State**, **Start Year**, **Average Severity** vào cửa sổ thực hiện câu truy vấn

Ths. Đỗ Thị Minh Phụng

State	Start Year	Average Severity
[Dim Location].[State].[State] Level: State		

[Click to execute the query.](#)

Bước 3: Kéo thả **Dim Location** và **Dim Time** vào cửa sổ lọc điều kiện cho câu truy vấn, với **Hierarchy** của 2 bảng lần lượt là **State** và **Start Year** và điều kiện lọc **Filter Expression** lần lượt là **NameSet [SeverityUp]** và **{2017, 2018}** như hình bên dưới

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim Location	# State	In	SeverityUp	<input type="checkbox"/> <input type="checkbox"/>
Dim Time	# Start Year	Equal	{ 2017, 2018 }	<input type="checkbox"/> <input type="checkbox"/>
<Select dimension>				<input type="checkbox"/> <input type="checkbox"/>

Bước 4: Click vào **Click to execute the query** để thực thi truy vấn và xem kết quả như hình bên dưới

State	Start Year	Average Severity
FL	2017	2.41686893203884
FL	2018	3.0218043063505
IL	2017	2.67758186397985
IL	2018	3.07072570725707
IN	2017	2.489443378119
IN	2018	3.02968460111317
KS	2017	2.63938618925831
KS	2018	3.28869565217391
MI	2017	2.44499178981938
MI	2018	3.05858772741289
MN	2017	2.68067226890756
MN	2018	3.3878453038674
MO	2017	2.49737670514166
MO	2018	3.48400852878465
ND	2017	2.13333333333333
ND	2018	3.6
OH	2017	2.55480868665977
OH	2018	3.75303161541793
SD	2017	2.625

3.9.15. Câu truy vấn 15

Nội dung câu truy vấn: Thông kê thời gian thi công trung bình và quãng đường thi công trung bình theo từng thành phố trong điều kiện thời tiết khắc nghiệt là có tốc độ gió trên 70mph,

Ý nghĩa câu truy vấn: mục đích thu thập và phân tích dữ liệu về thời gian thi công trung bình và quãng đường thi công trung bình của các công trình xây dựng gây ảnh hưởng đến giao thông đường bộ trong các điều kiện thời tiết khắc nghiệt, cụ thể là khi có tốc độ gió trên 70 mph, theo từng thành phố.

Ngoài ra, thông tin về thời gian thi công trung bình và quãng đường thi công trung bình trong điều kiện thời tiết khắc nghiệt sẽ hỗ trợ trong việc lập kế hoạch và dự đoán các dự án tương lai, đảm bảo rằng các công trình sẽ được hoàn thành một cách hiệu quả và an

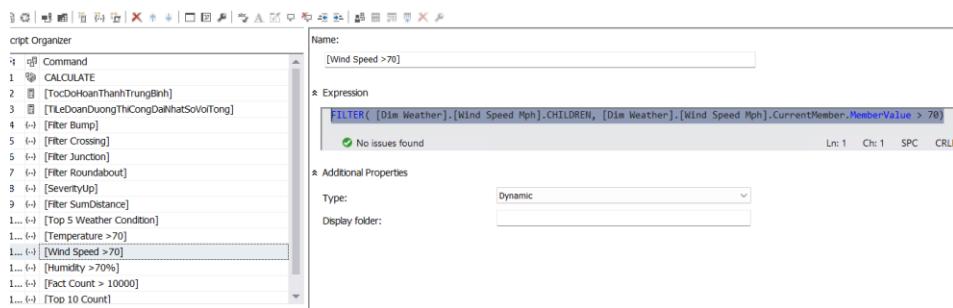
toàn, đồng thời giảm thiểu tác động tiêu cực đến giao thông trong những điều kiện thời tiết bất lợi.

Kỹ thuật thực hiện: Slice and Dice để lọc dữ liệu theo các điều kiện cụ thể về tốc độ gió và thời gian, sau đó tổng hợp dữ liệu để thống kê số lượng công trình và trung bình độ dài đoạn đường thi công.

Thực hiện câu truy vấn:

Bước 1: Tạo NameSet [Wind Speed >70] như hình bên dưới với biểu thức sau:

FILTER([Dim Weather].[Wind Speed Mph].CHILDREN, [Dim Weather].[Wind Speed Mph].CurrentMember.MemberValue > 70)



Bước 2: Kéo thả các thuộc tính City, Average Duration, Average Distance vào cửa sổ thực thi truy vấn.

City	Average Duration	Average Distance

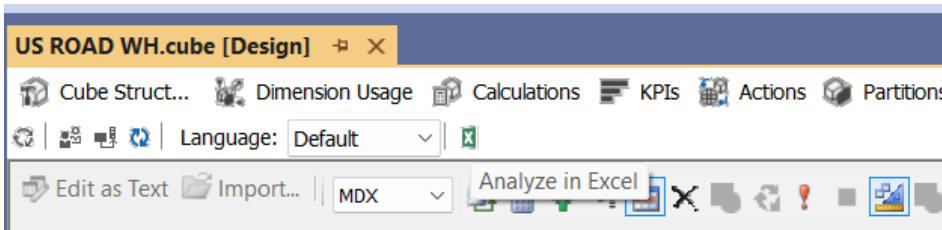
Bước 3: Kéo thả **Dim Weather** vào cửa sổ lọc điều kiện truy vấn với **Hierarchy** là **Wind Speed Mph** và **Filter Expression** là NameSet **Wind Speed >70** đã tạo lúc trước

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim Weather	# Wind Speed Mph	In	Wind Speed >70	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

Bước 4: Click vào **Click to execute the query** để thực hiện truy vấn và xem kết quả như hình bên dưới.

City	Average Duration	Average Distance
Beaver	21755	0.67901891469...
Gorham	27240	0.30713737010...
Port Matilda	21600	5.55383634567...
San Pedro	21600	0.96868723630...
Scottsdale	431992	0.02617464028...

3.10. Quá trình phân tích dữ liệu bằng công cụ Pivot Excel



3.10.1. Câu truy vấn 1:

Nội dung câu truy vấn: Thống kê số lượng công trình làm ảnh hưởng tới đường bộ xảy ra tại 4 quận Kent, Newport, Providence và Washington của bang Rhode Island (RI) trong 3 năm liên tục 2017 – 2019.

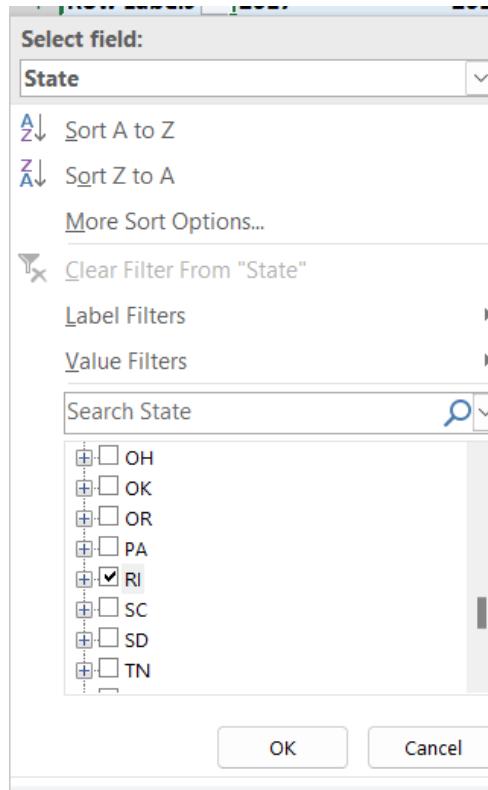
Thực hiện:

Bước 1: Tại tab PivotTable Fields:

- Kéo thả **Construction_Location** của bảng **Dim_Location** vào mục **ROWS**.
- Kéo thả measure **Fact Count** vào mục **VALUES**.
- Kéo thả **Start Year** của bảng **Dim_Time** vào mục **COLUMNS**.

		2017	2018	2019	Grand Total
Row Labels	Fact Count				
3 RI		45	268	1514	1837
4 RI-L		127	300	342	769
5 RI-K		479	1353	2218	4050
6 RI-N		5332	7326	18092	30750
7 RI-W		587	1385	6005	7977
8 RI-C		2057	1385	1982	6128
9 RI-P		230	174	1041	1441
10 RI-E		334	424	533	1291
11 RI-F		1648	3669	7222	12539
12 RI-GA		447	1426	4193	6066
13 RI-IA		1068	1751	3078	5897
14 RI-ID		309	554	4627	5490
15 RI-IL		9968	10322	11499	31899
16 RI-IN		531	1078	4295	5894
17 RI-KS		391	575	2015	2981
18 RI-KY		106	236	210	552
19 RI-LA		101	466	781	1348
20 RI-MA		1354	1453	1843	4650
21 RI-MD		5	10	10	3103
22 RI-MI		260	306	240	806
23 RI-MI		1827	3243	4990	10060
24 RI-MN		357	904	1367	2628
25 RI-MO		953	2345	1303	4601
26 RI-MS		33	186	433	652
27 RI-MT		27	50	25	102
28 RI-NC		947	1219	4638	6804
29 RI-ND		15	29	21	56

Bước 2: Chọn Filter Values để lọc lấy ra các quận của bang Rhode Island (RI).



Bước 3: Theo dõi kết quả truy vấn dạng Table

Fact Count	Column Labels	2017	2018	2019	Grand Total
Row Labels		72	221	862	1155
RI					
Kent		9	23	98	130
Newport		1	26	7	34
Providence		59	141	714	914
Washington		3	31	43	77
Grand Total		72	221	862	1155

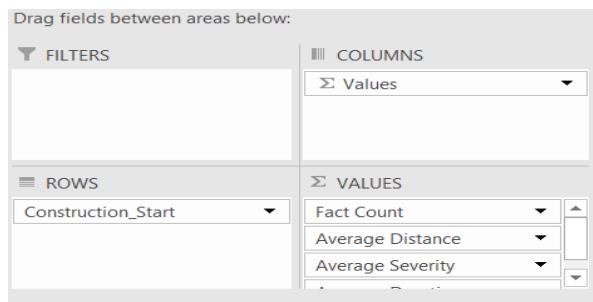
3.10.2. Câu truy vấn 2:

Nội dung câu truy vấn: Thống kê tổng số công trình, mức độ ảnh hưởng trung bình, chiều dài trung bình của đoạn đường mà có công trình thi công ảnh hưởng, và thời gian xảy ra trung bình của các công trình thi công theo các tháng trong năm 2019.

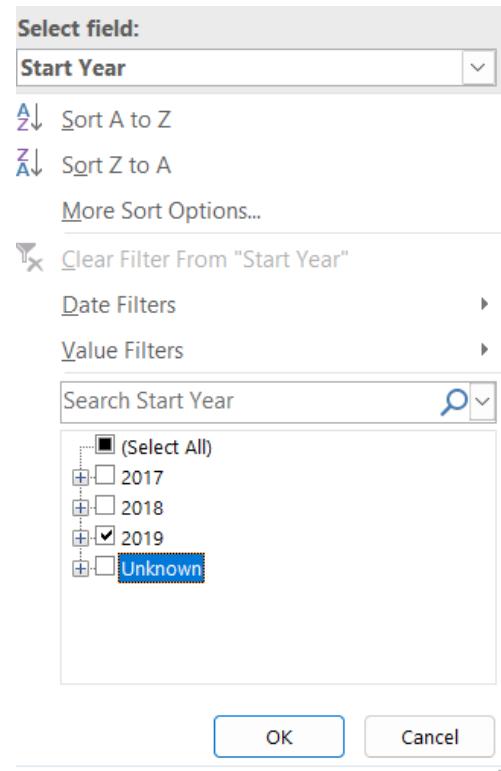
Thực hiện:

Bước 1: Tại Tab PivotTable Fields:

- Kéo thả **Construction_Start** của bảng **Dim_Time** vào mục **ROWS**.
- Kéo thả các measures **Fact Count**, **Average Duration**, **Average Severity**, **Average Distance** vào mục **VALUES**.



Bước 2: Chọn Filter Values để lọc lấy ra năm 2019.



Bước 3: Theo dõi kết quả truy vấn dạng Table:

Row Labels	Fact Count	Average Distance	Average Duration	Average Severity
2019	152421	2.717595388	557157.6391	2.811817269
+ 1	8479	2.698389172	996630.1907	2.71871683
+ 2	7333	2.87433624	959944.8992	2.704486568
+ 3	10758	2.482835824	937432.9605	2.845975088
+ 4	17061	2.758918733	707547.7226	2.838227536
+ 5	17696	2.843040741	595100.7883	2.898621157
+ 6	15147	2.575153634	628554.4095	2.8664422
+ 7	16721	2.778029395	506609.4142	2.808863106
+ 8	16563	2.46104032	372958.001	2.814526354
+ 9	14701	3.098159001	373538.7482	2.809536766
+ 10	15228	2.566955603	282845.8449	2.772130286
+ 11	7290	2.847432139	251070.1058	2.718381344
+ 12	5444	2.674107243	174070.2404	2.760102866
Grand Total	152421	2.717595388	557157.6391	2.811817269

3.10.3. Câu truy vấn 3:

Nội dung câu truy vấn: Thống kê với 5 loại điều kiện thời tiết có tổng công trình được thực hiện nhiều nhất, theo từng tháng, năm và tính tổng chiều dài các đoạn đường có công trình thi công đó.

Thực hiện:

Bước 1: Tại Tab PivotTable Fields:

- Kéo thả **Construction_Start** của bảng **Dim_Time** vào mục **ROWS**.
- Kéo thả **Set Top 5 Weather Condition** của bảng **Dim_Weather** vào mục **COLUMNS**.
- Kéo thả measure **Sum Distance** vào mục **VALUES**.

PivotTable Fields

Choose fields to add to report:

Search

Temperature F
 Weather Condition
 Weather ID
 Wind Speed Mph

Sets

Humidity > 70%
 Temperature > 70
 Top 5 Weather Condition

Drag fields between areas below:

FILTERS	COLUMNS
	Top 5 Weather Condition

ROWS	VALUES
Construction_Start	Sum Distance

Bước 2: Theo dõi kết quả truy vấn dạng Table:

Sum Distance	Column Labels					
Row Labels	Clear	Fair	Mostly Cloudy	Overcast	Partly Cloudy	
2017		75248.01644	2109.560579	22993.73751	37898.15719	13211.31025
⊕ 1		2367.94598	17.75838224	1078.7451	3265.104557	491.4279917
⊕ 2		4392.065241	38.84288005	1065.26204	2172.619091	636.4322681
⊕ 3		4272.539343	68.05514021	1078.398099	2576.317929	940.6410049
⊕ 4		5500.43488	159.9481721	1838.174384	4214.147985	656.1116887
⊕ 5		6818.92276	216.4705473	2440.085968	4331.558982	1279.594459
⊕ 6		8115.732281	346.6223468	2654.937531	2338.58061	1448.957358
⊕ 7		8839.068997	271.6544831	2338.24525	2470.066813	1628.183172
⊕ 8		9756.159645	418.6655792	3849.372221	3275.231955	1946.067129
⊕ 9		7694.593848	282.2879486	1736.439903	3002.409182	1415.61163
⊕ 10		7218.786835	104.6100959	2188.652866	3380.256973	1059.455414
⊕ 11		5402.824821	124.9824035	1654.913054	3948.623262	1041.143821
⊕ 12		4868.94181	59.6625998	1070.511091	2923.239846	667.6843111
2018		93494.84837	3784.392591	28231.98876	55872.72727	16365.08749
⊕ 1		3668.480006	243.0898245	823.014513	2103.261901	592.5245814
⊕ 2		3960.851534	56.63647645	896.7508952	3731.778963	561.2215132
⊕ 3		5778.098723	74.60076593	1366.697399	2785.343968	902.0280438
⊕ 4		6718.397903	259.1202323	2130.397667	5194.973372	1138.064553
⊕ 5		8231.328496	222.8134413	2834.384622	4028.236234	1302.881266
⊕ 6		9112.594976	134.745569	3313.431339	4107.348841	1554.151761
⊕ 7		9787.036727	615.847751	2301.080671	2180.308394	1762.940373
⊕ 8		9325.326972	294.3256035	2579.473962	2931.689451	2338.914052
⊕ 9		7480.099146	298.3924805	2672.48854	5466.360877	1311.001736
⊕ 10		10790.04104	371.1370647	3914.195021	7321.464754	2104.21079
⊕ 11		10534.42165	651.6575801	3068.767238	9456.275687	1657.350306
⊕ 12		8108.171193	562.025802	2331.306895	6565.684827	1139.798518
2019		27105.80932	176239.9081	47536.95978	18391.23206	38286.92404

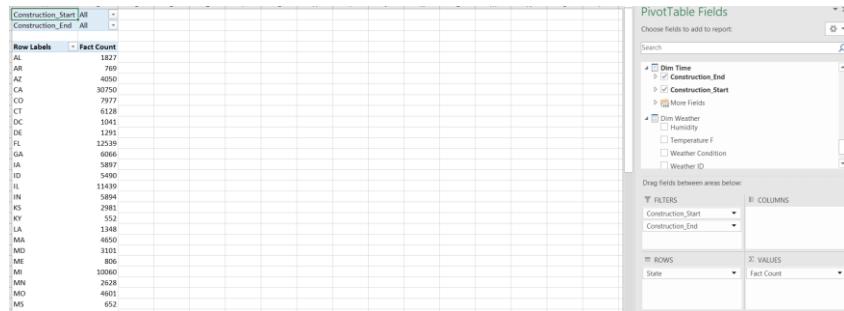
3.10.4. Câu truy vấn 4:

Nội dung câu truy vấn: Thống kê tổng số công trình (Fact Count) trên 6 tiểu bang (State): California (CA), New York (NY), Florida (FL), Ohio (OH), Texas (TX) và Washington (WA) có thời gian bắt đầu 2017 và kết thúc 2019.

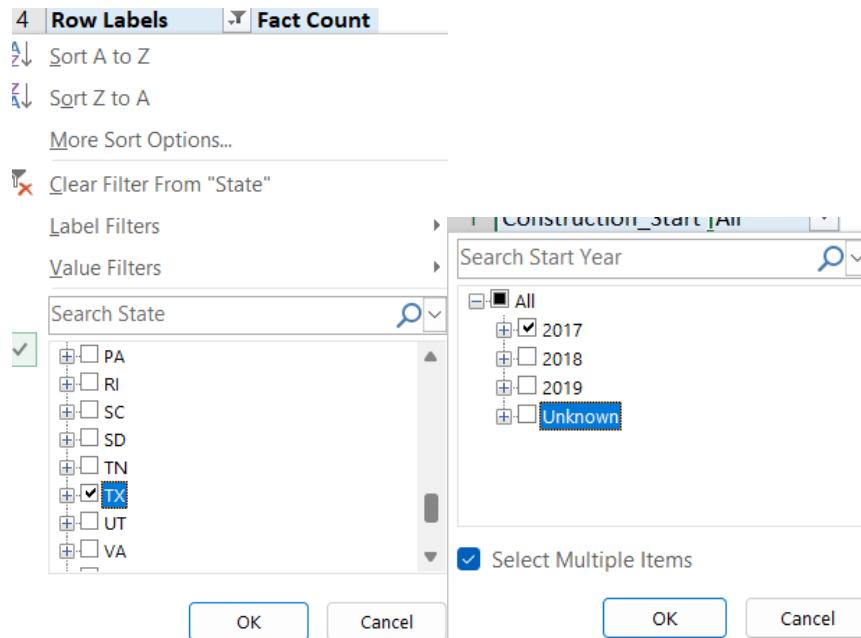
Thực hiện:

Bước 1: Tại Tab PivotTable Fields:

- Kéo thả **State** của bảng **Dim_Location** vào mục **ROWS**.
- Kéo thả measure **Fact Count** vào mục **VALUES**.
- Kéo thả **Construction_Start** và **Construction_End** của bảng **Dim_Time** vào **FILTERS**.



Bước 2: Chọn **Filter Values** để lọc lấy ra 6 tiểu bang **CA, FL, NY, OH, TX, WA** và chọn ra thời gian từ **2017** và **2019**.



Bước 3: Theo dõi kết quả truy vấn dạng Table:

Row Labels	Fact Count
CA	3
FL	1
OH	8
TX	19
WA	1
Grand Total	32

3.10.5. Câu truy vấn 5:

Nội dung câu truy vấn: Thông kê số công trình và trung bình độ dài đoạn đường thi công với nhiệt độ trên 70F và độ ẩm trên 70%.

Thực hiện:

Bước 1: Tại tab PivotTable Fields:

- Kéo thả 2 set [**Temperature > 70**] và [**Humidity > 70%**] của bảng **Dim_Weather** vào mục **ROWS**.
- Kéo thả 2 measure **Fact Count** và **Average Distance** vào mục **VALUES**.

PivotTable Fields

Choose fields to add to report:

Search

Temperature F
 Weather Condition
 Weather ID
 Wind Speed Mph

Sets

Humidity > 70%
 Temperature > 70
 Top 5 Weather Condition

Drag fields between areas below:

FILTERS	COLUMNS
	Σ Values

ROWS	VALUES
Temperature > 70	Σ Fact Count
Humidity > 70%	Σ Average Distance

Bước 2: Theo dõi kết quả truy vấn dạng Table.

Row Labels	Fact Count	Average Distance
100.400001525879		
71	1	0.00766466
70.1999969482422		
100	1	3.670000076
71	1	1.167540669
73	3	7.73789386
74	3	1.487878541
77	1	0.076100856
82	1	0.322929651
84	1	8.130000114
85	1	14.80000019
86	5	10.23600025
87	4	0.092408881
88	2	2.6648736
89	1	0.322673231
90	1	0.015374679
93	2	1.135367528
94	2	0.08726158
95	1	0.318082035
96	1	0.196100548
99	2	4.818130612
70.3000030517578		
100	3	10.67411804
74	4	0.883772664
77	1	0.07972423
79	1	1.641008377
81	7	2.434861494
84	2	0.152283145

3.10.6. Câu truy vấn 6:

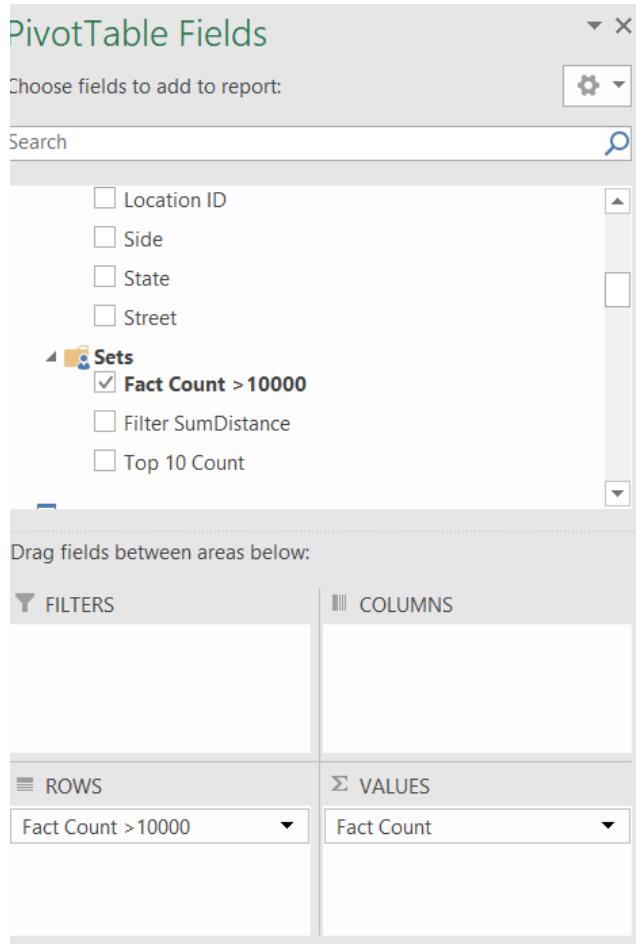
Nội dung câu truy vấn: Thống kê theo từng bang có số lượng các công trình thi công lớn 10000.

Thực hiện:

Bước 1: Tại tab PivotTable Fields:

- Kéo thả set [Fact Count > 10000] của bảng Dim_Location vào mục ROWS.

- Kéo thả measure **Fact Count** vào mục **VALUES**.



Bước 2: Theo dõi kết quả truy vấn dạng Table:

Row Labels	Fact Count
CA	30750
FL	12539
IL	11439
MI	10060
NY	21328
OH	16232
OR	10343
PA	66502
TX	14773

3.10.7. Câu truy vấn 7:

Nội dung câu truy vấn: Liệt kê 10 tiểu bang bị ảnh hưởng đến giao thông đường bộ bởi các công trình xây dựng nhiều nhất trong 3 năm từ 2017 tới 2019. Với 10 tiểu bang đó, liệt kê tên các con đường có tổng độ dài đoạn đường bị ảnh hưởng bởi các công trình lớn hơn 100 miles.

Thực hiện:

Bước 1: Tại tab PivotTable Fields:

- Kéo thả **State, County, City**, set **[Filter SumDistance], Side** của bảng **Dim_Location** vào mục **ROWS**.
- Kéo thả measure **Sum Distance** vào mục **VALUES**.

PivotTable Fields

Choose fields to add to report:

Search 

▲ Σ Fact

- Average Distance
- Average Duration
- Average Severity
- Fact Count
- Max Distance
- Min Distance
- Sum Distance
- Sum duration

Drag fields between areas below:

FILTERS	COLUMNS

ROWS	Σ VALUES
County	Sum Distance
City	
Filter SumDistance	
Side	

Bước 2: Theo dõi kết quả truy vấn dạng Table:

Row Labels		Sum Distance
CA		
CA		
Alameda		
Fremont		
R		
I-680 N		102.936013
Hayward		
R		
I-880 N		133.6829764
Oakland		
R		
I-880 N		107.2568781
Pleasanton		
R		
I-680 S		137.0196239
San Leandro		
R		
I-880 N		163.0061264
Sunol		
R		
I-680 N		192.3489103
R		
I-680 S		109.172749
Tracy		
R		
I-580 E		101.8954703
R		
I-580 W		337.9256023

3.10.8. Câu truy vấn 8:

Nội dung câu truy vấn: Với từng thành phố thuộc quận Orange của bang California, hãy cho biết độ dài đường thi công lớn nhất, tổng độ dài thi công và tỉ lệ giữa công trình thi công dài nhất so với tổng độ dài trong năm 2019.

Thực hiện:

Bước 1: Tại tab PivotTable Fields:

Commented [LQ1]: chỉnh lại, đã đổi câu tv r

Commented [PN2R1]: done

- Kéo thả **City** vào **ROWS**
- **Construction_Location, Construction_Start** vào **FILTERS**
- **Max Distance, Sum Distance, TiLeDoanDuongThiCongDaiNhatSoVoiTong** vào **VALUES**

PivotTable Fields

Choose fields to add to report:

Search

Fields list:

- Fact
- Dim Location
- Dim POI
- Dim Time
 - Construction_End
 - Construction_Start
 - More Fields
- Dim Weather

Drag fields between areas below:

FILTERS

- Construction_Location
- Construction_Start

ROWS

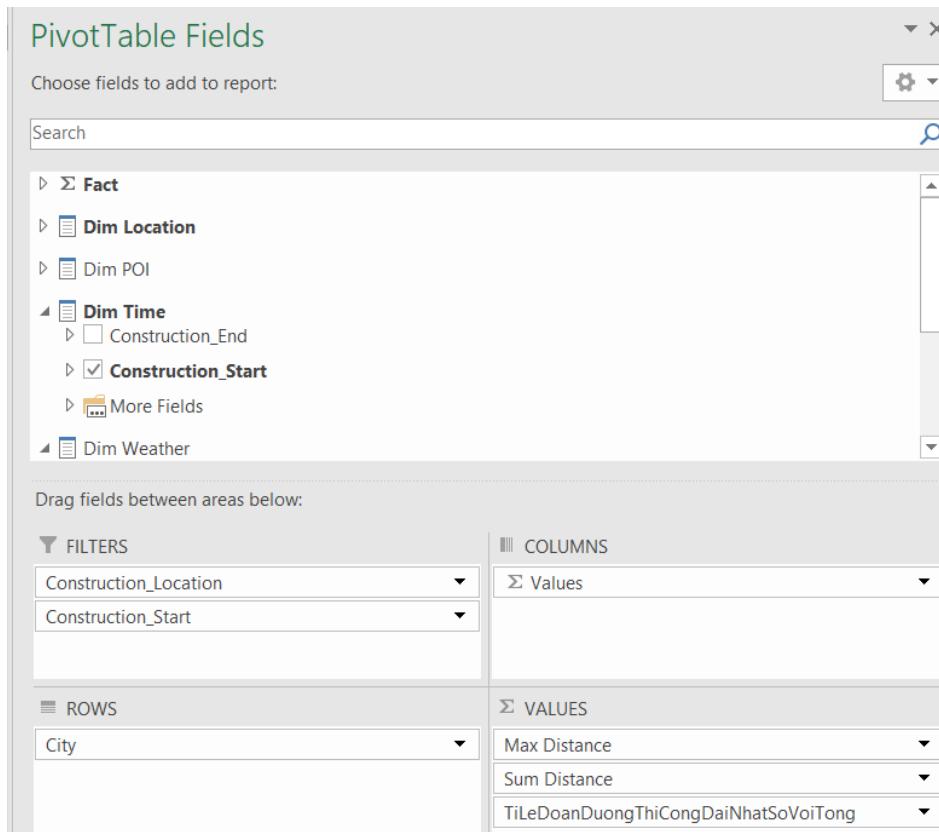
- City

COLUMNS

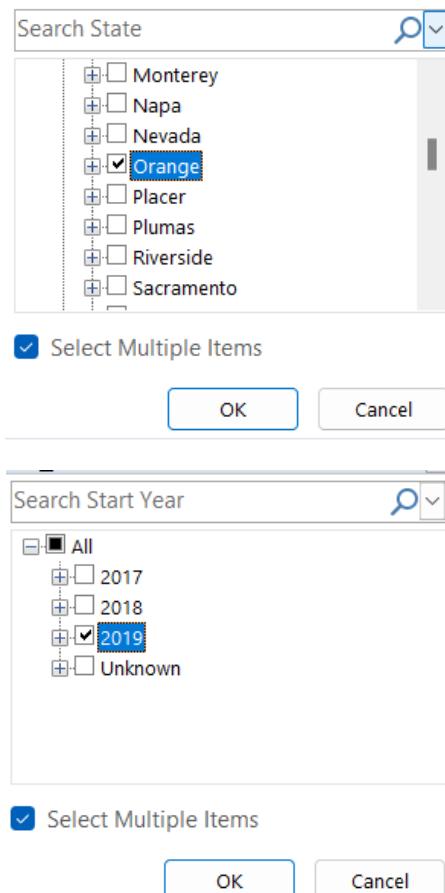
- Σ Values

VALUES

- Max Distance
- Sum Distance
- TiLeDoanDuongThiCongDaiNhatSoVoiTong



Bước 2: Chọn **Filter Values** để lọc lấy **Orange** và chọn thời gian bắt đầu (**Start_Year**) là **2019**



Bước 3: Theo dõi kết quả truy vấn dạng Table:

Construction_Location	Orange			
Construction_Start	2019			
Row Labels				
Alico Viejo	0.518669307	0.929970533		55.77%
Anaheim	15.56000042	70.36924577		22.11%
Anaheim-Santa Ana-Garden Grove	1.00999999	2.019999981		50.00%
Brea	9.68999958	87.26138221		11.10%
Buena Park	8.399999619	72.28730945		11.62%
Capistrano Beach	2	2.161810011		92.52%
Costa Mesa	24.31999969	418.0332879		5.82%
Dana Point	0.274802119	0.900315329		30.52%
Foothill Ranch	0.14069061	0.14069061		100.00%
Fountain Valley	5.449999809	277.5205453		1.96%
Fullerton	4.039999962	55.85318875		7.23%
Garden Grove	3.212822199	10.57760392		30.37%
Huntington Beach	6.159999847	67.21680322		9.16%
Irvine	8.800000191	214.7343743		4.10%
La Habra	8.069999695	39.41891441		20.47%
Ladera Ranch	0.220150679	0.392050356		56.15%
Laguna Beach	3.319999933	13.06400241		25.41%
Laguna Hills	7.179999828	22.75339776		31.56%
Laguna Niguel	2.319999933	10.61541164		21.86%
Lake Forest	0.518492162	0.518492162		100.00%
Los Alamitos	13.89999962	35.07363863		39.63%
Mission Viejo	7.269999981	65.82947646		11.04%
Newport Beach	4.920000076	72.64299986		6.77%
Newport Coast	3.369999886	7.933226585		42.48%
Orange	7.849999905	89.13317221		8.81%

3.10.9. Câu truy vấn 9:

Nội dung câu truy vấn: Thống kê tổng số công trình theo từng ngày, tháng, năm.

Thực hiện:

Bước 1: Tại tab **PivotTable Fields**:

- Kéo thả **Construction_Start** của bảng **Dim_Location** vào mục **ROWS**.
- Kéo thả **Fact Count** vào mục **VALUES**.

PivotTable Fields

Choose fields to add to report:

Search

↳ turning loop

↳ Sets

↳ **Dim Time**

↳ Construction_End

↳ **Construction_Start**

↳ More Fields

↳ End Day

↳ End Hour

↳ End Minute

Drag fields between areas below:

FILTERS	COLUMNS

ROWS	VALUES
Construction_Start	Fact Count

Bước 2: Theo dõi kết quả truy vấn dạng Table:

Row Labels	Fact Count
2017	60287
1	3497
1	15
2	51
3	102
4	157
5	148
6	82
7	55
8	46
9	146
10	134
11	151
12	176
13	112
14	27
15	25
16	79
17	130
18	172
19	242
20	102
21	44
22	44
23	127
24	151
25	221
26	189

3.10.10. Câu truy vấn 10:

Nội dung câu truy vấn: Liệt kê số công trình thi công có các cơ sở vật chất giao thông gần đó như giao lộ (junction) và bùng binh (roundabout) theo thành phố, quận, bang trong 3 năm 2017- 2019.

Ý nghĩa câu truy vấn:

Thực hiện:

Bước 1: tại tab **PivotTable Fields**:

- Kéo thả **State, County, City** của bảng **Dim_Location** vào mục **ROWS**.
- Kéo thả measure **Fact Count** vào mục **VALUES**.
- Kéo thả set **[Filter Junction]** và **[Filter Roundabout]** thuộc bảng **Dim_POI** vào mục **COLUMNS**.

PivotTable Fields

Choose fields to add to report:

Search

Sets

- Traffic Calming
- Traffic Signal
- Turning Loop
- Filter Junction
- Filter Roundabout

ROWS

- State
- County
- City

COLUMNS

- Filter Junction
- Filter Roundabout

VALUES

- Fact Count

Bước 2: Theo dõi kết quả truy vấn dạng Table:

Fact Count	Column Labels
0	
Row Labels	
CA	2
Los Angeles	2
Palmdale	2
FL	1
Miami-Dade	1
Hialeah	1
GA	1
Clarke	1
Athens	1
WA	1
Pierce	1
Tacoma	1
Grand Total	5

3.10.11. Câu truy vấn 11:

Nội dung câu truy vấn: Cho biết số công trình và mức độ nghiêm trọng trung bình của các công trình đó xảy ra tại bang Texas trong điều kiện thời tiết có liên quan tới bão (Thunderstorm) trong 3 năm 2017 – 2019.

Thực hiện:

Bước 1: tại tab PivotTable Fields:

- Kéo thả **Start Year** của bảng **Dim_Time** vào mục **ROWS**.
- Kéo thả measure **Fact Count, Average Severity** vào mục **VALUES**.
- Kéo thả **State, Weather_condition** vào mục **COLUMNS**.

PivotTable Fields

Choose fields to add to report:

Search

▽ Σ Fact

▽ Dim Location

▽ Dim POI

▽ Dim Time

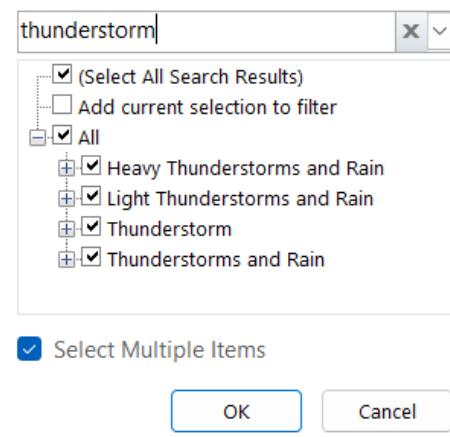
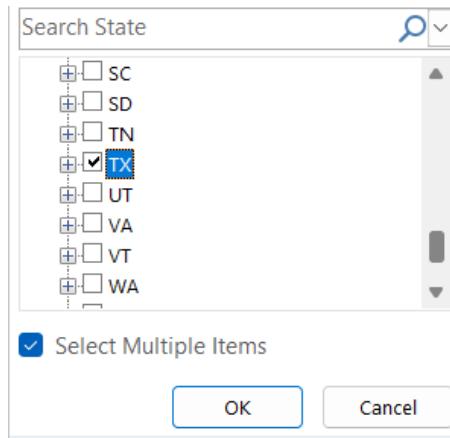
▽ Dim Weather

Drag fields between areas below:

FILTERS	COLUMNS
State	Σ Values
Weather Condition	

ROWS	VALUES
Start Year	Fact Count
	Average Severity

Bước 2: Chọn Filter Values để lọc ra tiêu bang TX và chọn all sau khi tìm kiếm Thunderstorm để lấy điều kiện thời tiết phù hợp với câu truy vấn



Bước 3: Theo dõi kết quả truy vấn dạng Table:

State	TX	
Weather Condition (Multiple Items)		
Row Labels	Fact Count	Average Severity
2017	10	2
2018	52	3.692307692
2019	3	3.33333333333
Grand Total	65	3.415384615

3.10.12. Câu truy vấn 13:

Nội dung câu truy vấn: Cho biết tốc độ hoàn thành trung bình của mỗi công trình ở gần vị trí gờ giảm tốc (bump) theo từng bang trong cả 3 năm 2017 – 2019.

Thực hiện câu truy vấn:

Bước 1: Tại Tab **PivotTable Fields**:

- Kéo thả **State** của bảng **Dim_Location** vào mục **ROWS**.
- Kéo thả **Name Set TocDoHoanThanhTrungBinh** vào mục **VALUES**.
- Kéo thả **NamSet Filter Bumps** vào **COLUMNS**

PivotTable Fields

Choose fields to add to report:

Search 🔍

Traffic Signal
 Turning Loop

↳ **Sets**

Filter Bump
 Filter Crossing
 Filter Junction
 Filter Roundabout

↳ **Dim Time**

Drag fields between areas below:

FILTERS	COLUMNS
	Filter Bump

ROWS	VALUES
State	TocDoHoanThanhTrungBinh

Bước 2: Theo dõi kết quả câu truy vấn dưới dạng Table.

TocDoHoanThanhTrungBinh	Column Labels
Row Labels	-1
AL	1
AR	1
AZ	4
CA	1
CO	1
CT	1
DC	1
DE	1
FL	4
GA	1
IA	1
ID	1
IL	1
IN	26
KS	1
KY	1
LA	1
MA	2
MD	2
ME	1
MI	1
MN	1
MO	1
MS	1
MT	1
NC	1
ND	1

3.10.13. Câu truy vấn 14:

Nội dung câu truy vấn: Thống kê trung bình mức độ nghiêm trọng của các công trình thi công đường tại các bang ở dưới mức 3 trong năm 2017 nhưng đã lên hơn mức 3 trong năm 2018.

Thực hiện câu truy vấn:

Bước 1: Tại Tab **PivotTable Fields**:

- Kéo thả **Start Year** của bảng **Dim_Time** và **NameSet SeverityUp** vào mục **ROWS**.
- Kéo thả measure **Average Severity** vào mục **VALUES**.

PivotTable Fields

Choose fields to add to report:

Search 🔍

Side
 State
 Street

📁 Sets
 Fact Count > 10000
 Filter SumDistance
 SeverityUp
 Top 10 Count

Drag fields between areas below:

FILTERS	COLUMNS

ROWS	VALUES
Start Year SeverityUp	Average Severity

Bước 2: Theo dõi kết quả truy vấn dưới dạng bảng

Row Labels	Average Severity
2017	
FL	2.416868932
IL	2.677581864
IN	2.489443378
KS	2.639386189
MI	2.44499179
MN	2.680672269
MO	2.497376705
ND	2.133333333
OH	2.554808687
SD	2.625
TX	2.426229508
UT	2.371900826
WA	2.920826162
WI	2.64882227
WY	2.791666667
2018	
FL	3.021804306
IL	3.070725707
IN	3.029684601
KS	3.288695652
MI	3.058587727
MN	3.387845304
MO	3.484008529
ND	3.6
OH	3.753031615
SD	3.473684211
TX	3.447148289

3.10.14. Câu truy vấn 15:

Nội dung câu truy vấn: Thống kê thời gian thi công trung bình và quảng đường thi công trung bình theo từng thành phố trong điều kiện thời tiết khắc nghiệt là có tốc độ gió trên 70mph,

Thực hiện câu truy vấn:

Bước 1: Tại Tab PivotTable Fields:

- Kéo thả **City** của bảng **Dim_Location** và **Name Set WindSpeed > 70** vào mục **ROWS**.
- Kéo thả measure **Average Duration** và **Average Distance** vào mục **VALUES**.

The screenshot shows the 'PivotTable Fields' pane in Power BI. At the top, it says 'Choose fields to add to report:' with a search bar and a gear icon. Below is a tree view under 'Fact':

- Average Distance
- Average Duration
- Average Severity
- Fact Count
- Max Distance
- Sum Distance
- Sum Duration
- TiLeDoanDuongThiConaDaiNhatSoVoiTong

Below the tree view, there's a section to 'Drag fields between areas below:' with three columns:

- FILTERS**: An empty area.
- COLUMNS**: A dropdown menu showing 'Σ Values'.
- ROWS**: A dropdown menu showing 'City' and 'Wind Speed > 70'.
Under 'VALUES', there are two dropdown menus:
 - 'Average Duration'
 - 'Average Distance'

Bước 2: Theo dõi kết quả truy vấn ở dạng Table

Row Labels	Average Duration	Average Distance
Beaver		
100.099998474121	21755	0.679018915
Gorham		
75	27240	0.30713737
Port Matilda		
209.399993896484	21600	5.553836346
San Pedro		
134.600006103516	21600	0.968687236
Scottsdale		
77.0999984741211	431992	0.02617464

3.11. Thực hiện 10 câu truy vấn – Quá trình phân tích dữ liệu bằng ngôn ngữ truy vấn MDX

3.11.1. Câu truy vấn 1

Nội dung câu truy vấn: Thống kê số lượng công trình làm ảnh hưởng tới đường bộ xảy ra tại 4 quận Kent, Newport, Providence và Washington của bang Rhode Island (RI) trong 3 năm liên tục 2017 – 2019.

Thực hiện câu truy vấn:

Bước 1: Kết nối SQL Server Analysis Service và viết câu lệnh MDX như hình sau trên kho dữ liệu đã tạo [US ROAD WH]

SELECT

```
NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,
NON EMPTY {
    ([Dim Time].[Start Year].[Start Year].ALLMEMBERS *
    [Dim Location].[Construction_Location].[County].ALLMEMBERS)
}
```

```
DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME  
ON ROWS  
FROM  
(SELECT  
( { [Dim Location].[Construction_Location].[State].&[RI] } )  
ON COLUMNS  
FROM [US ROAD WH])
```

Bước 2: Click vào **Execute** để thực thi câu lệnh và được kết quả như hình bên dưới

The screenshot shows a Microsoft SQL Server Management Studio (SSMS) interface. At the top, there is a code editor window containing T-SQL code. Below it is a results pane titled "Results". The results show a table with four columns: Year, County, Fact Count, and a column header "Fact Count" which is highlighted.

```
SELECT  
    NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,  
    NON EMPTY {  
        ([Dim Time].[Start Year].[Start Year].ALLMEMBERS *  
        [Dim Location].[Construction_Location].[County].ALLMEMBERS)  
    }  
    DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS  
FROM  
(SELECT  
( { [Dim Location].[Construction_Location].[State].&[RI] } )  
ON COLUMNS  
FROM [US ROAD WH])
```

		Fact Count	
2017	Kent	9	
2017	Newport	1	
2017	Providence	59	
2017	Washington	3	
2018	Kent	23	
2018	Newport	26	
2018	Providence	141	
2018	Washington	31	
2019	Kent	98	
2019	Newport	7	
2019	Providence	714	
2019	Washington	43	

3.11.2. Câu truy vấn 2

Nội dung câu truy vấn: Thông kê tổng số công trình, mức độ ảnh hưởng trung bình, chiều dài trung bình của đoạn đường mà có công trình thi công ảnh hưởng, và thời gian xảy ra trung bình của các công trình thi công theo các tháng trong năm 2019.

Thực hiện truy vấn:

Bước 1: Viết câu lệnh MDX như sau trên kho dữ liệu đã tạo [US ROAD WH]

SELECT NON EMPTY

{

[Measures].[Fact Count],

[Measures].[Average Distance],

[Measures].[Average Duration],

[Measures].[Average Severity]

} ON COLUMNS,

NON EMPTY

{

[Dim Time].[Construction_Start].[Start Month].ALLMEMBERS

} ON ROWS

FROM

[US ROAD WH]

WHERE

([Dim Time].[Start Year].&[2019])

Bước 2: Click vào **Execute** để thực thi câu lệnh và được kết quả như hình bên dưới

```
--Cau2
SELECT NON EMPTY
{
    [Measures].[Fact Count],
    [Measures].[Average Distance],
    [Measures].[Average Duration],
    [Measures].[Average Severity]
} ON COLUMNS,
NON EMPTY
{
    [Dim Time].[Construction_Start].[Start Month].ALLMEMBERS
} ON ROWS
FROM
    [US ROAD WH]
WHERE
    ( [Dim Time].[Start Year].&[2019] )
```

100 %

	FactCount	Average Distance	Average Duration	Average Severity
1	8479	2.69838917166439	996630.190706451	2.71871682981484
2	7333	2.8743362402352	959944.899222692	2.70448656757125
3	10758	2.4828358236728	937432.960494516	2.84597508830638
4	17061	2.75891873347705	707547.72258367	2.83822753648672
5	17696	2.84304074139709	595100.788313743	2.89862115732369
6	15147	2.57515363355587	628554.409454017	2.86644219977553
7	16721	2.77802939503669	506609.414209676	2.80886310627355
8	16563	2.46104032015158	372958.001026384	2.81452635392139
9	14701	3.09815900058439	373538.748180396	2.80953676620638
10	15228	2.56695560331212	282845.84489099	2.77213028631468
11	7290	2.84743213925293	251070.105761317	2.71838134430727
12	5444	2.67410724301782	174070.2404482	2.76010286554004

3.11.3. Câu truy vấn 3

Nội dung câu truy vấn: Thống kê với 5 loại điều kiện thời tiết có tổng công trình được thực hiện nhiều nhất, theo từng tháng, năm và tính tổng chiều dài các đoạn đường có công trình thi công đó

Thực hiện câu truy vấn:

Bước 1: Kết nối SQL Server Analysis Service và viết câu lệnh MDX như hình sau trên kho dữ liệu đã tạo [US ROAD WH]

```

SELECT NON EMPTY {[Dim Weather].[Weather Condition].CHILDREN * 
[Measures].[Sum Distance]} ON COLUMNS,
NON EMPTY {DrillDownLevel(DrillDownLevel([Dim Time].[Construction_Start]))}
ON ROWS
FROM (
    SELECT (TOPCOUNT([Dim Weather].[Weather Condition].CHILDREN, 5,
[Measures].[Fact Count])) ON COLUMNS
    FROM [US ROAD WH]
);
--Cau 3
SELECT NON EMPTY {[Dim Weather].[Weather Condition].CHILDREN * [Measures].[Sum Distance]} ON COLUMNS,
NON EMPTY {DrillDownLevel(DrillDownLevel([Dim Time].[Construction_Start]))} ON ROWS
FROM (
    SELECT (TOPCOUNT([Dim Weather].[Weather Condition].CHILDREN, 5, [Measures].[Fact Count])) ON COLUMNS
    FROM [US ROAD WH]
);

```

Bước 2: Click vào Execute để thực thi câu lệnh và được kết quả như hình bên dưới

	Clear	Fair	Mostly Cloudy	Overcast	Partly Cloudy
	Sum Distance				
All	195821.0088415	182130.290907334	98761.6301264637	112151.847818169	67854.5254546501
2017	75247.4974453996	2109.56057887562	22993.7375068027	37898.1571850909	13211.0590553716
1	2367.94597999845	17.7583822384477	1078.74510004765	3265.10455733849	491.176800981979
10	7218.78683511521	104.610095890937	2188.65286560365	3380.25697286889	1059.45541392982
11	5402.82482053978	124.982403473783	1654.91305391386	3948.6232617745	1041.14382073002
12	4868.42281577493	59.6625997991941	1070.51109117755	2923.23984636705	667.684311119607
2	4392.06524147249	38.8428800514666	1065.26204045713	2172.61909123123	636.432268098579
3	4272.53934276033	68.0551402061246	1078.39809921663	2576.3179291104	940.641004937977
4	5500.43487978468	159.948172128759	1838.17438358131	4214.14798452985	656.111688674806
5	6818.92276014433	216.470547292149	2440.08596838816	4331.55898205756	1279.59445924665
6	8115.7322809476	346.622346816555	2654.93753069175	2338.58060972406	1448.95735765134
7	8839.06899664096	271.654483148712	2338.2452497708	2470.06681294665	1628.18317175371
8	9756.15964469925	418.665579202701	3849.3722206707	3275.23195546983	1946.06712852199
9	7694.5938475216	282.287948626792	1736.43990328349	3002.40918167234	1415.61162972516
2018	93494.8483675563	3784.39259136764	28231.9887622843	55862.5272697744	16365.0874938549

3.11.4. Câu truy vấn 4

Nội dung câu truy vấn: Thống kê tổng số công trình (Fact Count) trên 6 tiểu bang (State): California (CA), New York (NY), Florida (FL), Ohio (OH), Texas (TX) và Washington (WA) có thời gian bắt đầu 2017 và kết thúc 2019.

Thực hiện câu truy vấn:

Bước 1: Kết nối SQL Server Analysis Service và viết câu lệnh MDX như hình sau trên kho dữ liệu đã tạo [US ROAD WH]

```
SELECT {[Measures].[Fact Count]} ON COLUMNS,
NON EMPTY {[Dim Location].[Construction_Location].[State].ALLMEMBERS} ON
ROWS
FROM (
SELECT {
[Dim Location].[State].[CA],
[Dim Location].[State].[NY],
[Dim Location].[State].[FL],
[Dim Location].[State].[OH],
[Dim Location].[State].[TX],
[Dim Location].[State].[WA]
} ON COLUMNS
FROM [US ROAD WH]
)
WHERE ([Dim Time].[Start Year].[2017], [Dim Time].[End Year].[2019])
```

```
--Câu 4
SELECT {[Measures].[Fact Count]} ON COLUMNS,
NON EMPTY {[Dim Location].[Construction_Location].[State].ALLMEMBERS} ON ROWS
FROM (
  SELECT {
    [Dim Location].[State].[CA],
    [Dim Location].[State].[NY],
    [Dim Location].[State].[FL],
    [Dim Location].[State].[OH],
    [Dim Location].[State].[TX],
    [Dim Location].[State].[WA]
  } ON COLUMNS
  FROM [US ROAD WH]
)
WHERE ([Dim Time].[Start Year].[2017], [Dim Time].[End Year].[2019])
```

Bước 2: Click vào **Execute** để thực thi câu lệnh và được kết quả như hình bên dưới:

Messages		Results
	Fact Count	
CA	3	
FL	1	
OH	8	
TX	19	
WA	1	

3.11.5. Câu truy vấn 5

Nội dung câu truy vấn: Thống kê số công trình và trung bình độ dài đoạn đường thi công với nhiệt độ trên 70F và độ ẩm trên 70%.

Thực hiện câu truy vấn:

Bước 1: Kết nối SQL Server Analysis Service và viết câu lệnh MDX như hình sau trên kho dữ liệu đã tạo **[US ROAD WH]**

```
SELECT NON EMPTY {[Measures].[Fact Count],[Measures].[Average Distance]} ON COLUMNS,
```

```
NON EMPTY CROSSJOIN(
```

FILTER([Dim Weather].[Temperature F].CHILDREN, [Dim Weather].[Temperature F].CurrentMember.MemberValue > 70),
 FILTER([Dim Weather].[Humidity].CHILDREN, [Dim Weather].[Humidity].CurrentMember.MemberValue > 70)
) ON ROWS
 FROM [US ROAD WH]

--Câu 5

```

SELECT
  NON EMPTY {[Measures].[Fact Count],[Measures].[Average Distance]} ON COLUMNS,
  NON EMPTY CROSSJOIN(
    FILTER([Dim Weather].[Temperature F].CHILDREN, [Dim Weather].[Temperature F].CurrentMember.MemberValue > 70),
    FILTER([Dim Weather].[Humidity].CHILDREN, [Dim Weather].[Humidity].CurrentMember.MemberValue > 70)
  ) ON ROWS
FROM [US ROAD WH]
  
```

Bước 2: Click vào **Execute** để thực thi câu truy vấn và được kết quả như hình bên dưới

		Fact Count	Average Distance
100.400001525879	71	1	7.66466045752168E-03
70.1999969482422	100	1	3.67000007629395
70.1999969482422	71	1	1.16754066944122
70.1999969482422	73	3	7.73789385954539
70.1999969482422	74	3	1.48787854115168
70.1999969482422	77	1	7.61008560657501E-02
70.1999969482422	82	1	0.32292965054512
70.1999969482422	84	1	8.13000011444092
70.1999969482422	85	1	14.8000001907349

3.11.6. Câu truy vấn 6

Nội dung câu truy vấn: Thông kê theo từng bang có số lượng các công trình thi công lớn 10000.

Thực hiện câu truy vấn:

Bước 1: Viết câu lệnh **MDX** như hình bên dưới

SELECT

```
NON EMPTY {[Measures].[Fact Count]} ON  
COLUMNS,  
FILTER({[Dim Location].[Construction_Location].[State].ALLMEMBERS},  
[Measures].[Fact Count] > 10000) ON ROWS  
FROM [US ROAD WH];
```

```
--Cau 6  
SELECT  
NON EMPTY {[Measures].[Fact Count]} ON  
COLUMNS,  
FILTER({[Dim Location].[Construction_Location].[State].ALLMEMBERS},  
[Measures].[Fact Count] > 10000) ON ROWS  
FROM [US ROAD WH];
```

Bước 2: Click vào **Execute** để thực thi câu truy vấn và hiện kết quả như hình bên dưới

	Fact Count
CA	30750
FL	12539
IL	11439
MI	10060
NY	21328
OH	16232
OR	10343
PA	66502
TX	14773

3.11.7. Câu truy vấn 7

Nội dung câu truy vấn: Liệt kê 10 tiểu bang bị ảnh hưởng đến giao thông đường bộ bởi các công trình xây dựng nhiều nhất trong 3 năm từ 2017 tới 2019. Với 10 tiểu bang đó, liệt kê tên các con đường có tổng độ dài đoạn đường bị ảnh hưởng bởi các công trình lớn hơn 100 miles.

Thực hiện câu truy vấn:

Bước 1: Viết câu lệnh MDX như hình bên dưới

```
SELECT NON EMPTY { [Measures].[Sum Distance]} ON COLUMNS,
```

```
NON EMPTY {  
    (  
        [Dim Location].[State].CHILDREN *  
        [Dim Location].[County].CHILDREN *  
        [Dim Location].[City].CHILDREN *  
        FILTER([Dim Location].[Street].CHILDREN, [Measures].[Sum Distance] > 100) *  
        [Dim Location].[Side]. CHILDREN  
    )  
} ON ROWS  
FROM (  
    SELECT (  
        TOPCOUNT([Dim Location].[State].CHILDREN, 10, [Measures].[Fact Count])  
    ) ON COLUMNS  
FROM [US ROAD WH] )  
  
--Cau 7  
SELECT NON EMPTY { [Measures].[Sum Distance] } ON COLUMNS,  
NON EMPTY {  
    (  
        [Dim Location].[State].CHILDREN *  
        [Dim Location].[County].CHILDREN *  
        [Dim Location].[City].CHILDREN *  
        FILTER([Dim Location].[Street].CHILDREN, [Measures].[Sum Distance] > 100) *  
        [Dim Location].[Side]. CHILDREN  
    )  
} ON ROWS  
FROM (  
    SELECT (  
        TOPCOUNT([Dim Location].[State].CHILDREN, 10, [Measures].[Fact Count])  
    ) ON COLUMNS  
FROM [US ROAD WH]  
)|
```

Bước 2: Click vào **Execute** để thực thi câu truy vấn và được kết quả như hình bên dưới

			Sum Distance
CA	Alameda	Fremont	I-680 N
CA	Alameda	Hayward	I-880 N
CA	Alameda	Oakland	I-880 N
CA	Alameda	Pleasanton	I-680 S
CA	Alameda	San Leandro	I-880 N
CA	Alameda	Sunol	I-680 N
CA	Alameda	Sunol	I-680 S
CA	Alameda	Tracy	I-580 E
CA	Alameda	Tracy	I-580 W

3.11.8. Câu truy vấn 8

Nội dung câu truy vấn: Với từng thành phố thuộc quận Orange của bang California, hãy cho biết độ dài đường thi công lớn nhất, tổng độ dài thi công và tỉ lệ giữa công trình thi công dài nhất so với tổng độ dài trong năm 2019.

Thực hiện câu truy vấn:

Viết câu lệnh MDX như hình bên dưới:

SELECT

```
{
    [Measures].[Max Distance],
    [Measures].[Sum Distance],
    [Measures].[TiLeDoanDuongThiCongDaiNhatSoVoiTong]
```

} ON COLUMNS,

NON EMPTY

```
{
    [Dim Location].[City].[City].ALLMEMBERS
```

} ON ROWS

FROM [US ROAD WH]

WHERE

```
CrossJoin(
    {[Dim Location].[Construction_Location].[County].&[Orange]&[CA]},
    {[Dim Time].[Construction_Start].[Start Year].&[2019]}

);
```

```
--Cau8
SELECT
{
    [Measures].[Max Distance],
    [Measures].[Sum Distance],
    [Measures].[TiLeDoanDuongThiCongDaiNhatSoVoiTong]
} ON COLUMNS,
NON EMPTY
{
    [Dim Location].[City].[City].ALLMEMBERS
} ON ROWS
FROM [US ROAD WH]
WHERE
CrossJoin(
    {[Dim Location].[Construction_Location].[County].&[Orange]&[CA]},
    {[Dim Time].[Construction_Start].[Start Year].&[2019]}

);
```

100 % ▾

	Max Distance	Sum Distance	TiLeDoanDuongThiCongDaiNhatSoVoiTong
Aliso Viejo	0.518669307231903	0.929970532655716	55.77%
Anaheim	15.5600004196167	70.3692457675934	22.11%
Anaheim-Santa Ana-Garden Grove	1.0099999046326	2.0199998092651	50.00%
Brea	9.689995803833	87.2613822109997	11.10%
Buena Park	8.39999961853027	72.2873094491661	11.62%
Capistrano Beach	2	2.16181001067162	92.52%
Costa Mesa	24.3199996948242	418.03287860453	5.82%
Dana Point	0.27480211853981	0.900315329432487	30.52%
Foothill Ranch	0.140690609812737	0.140690609812737	100.00%
Fountain Valley	5.44999980926514	277.520545261912	1.96%
Fullerton	4.03999996185303	55.8531887549907	7.23%
Garden Grove	3.2128221988678	10.5776039212942	30.37%
Huntington Beach	6.15999984741211	67.2168032201007	9.16%

3.11.9. Câu truy vấn 9

Nội dung câu truy vấn: Thông kê tổng số công trình theo từng ngày, tháng, năm.

Thực hiện câu truy vấn:

Bước 1: Viết câu lệnh MDX nhu hình bên dưới

```
SELECT  
NON EMPTY  
{  
    [Dim Time].[Start Year].CHILDREN*  
    [Dim Time].[Start Month].CHILDREN*  
    [Dim Time].[Start Day].CHILDREN } ON ROWS,  
NON EMPTY  
{  
    [Measures].[Fact Count]  
} ON COLUMNS  
FROM [US ROAD WH];
```

Bước 2: Click vào **Execute** để thực thi câu lệnh và được kết quả nhu hình bên dưới

```
--Cau9
SELECT
    NON EMPTY
    {
        [Dim Time].[Start Year].CHILDREN*
        [Dim Time].[Start Month].CHILDREN*
        [Dim Time].[Start Day].CHILDREN
    } ON ROWS,
    NON EMPTY
    {
        [Measures].[Fact Count]
    } ON COLUMNS
FROM [US ROAD WH];
```

100 % ▶

			Fact Count
2017	1	1	15
2017	1	2	51
2017	1	3	102
2017	1	4	157
2017	1	5	148
2017	1	6	82
2017	1	7	55
2017	1	8	46
2017	1	9	146
2017	1	10	134
2017	1	11	151
2017	1	12	176
2017	1	13	112

3.11.10. Câu truy vấn 10

Nội dung câu truy vấn: Liệt kê số công trình thi công có các cơ sở vật chất giao thông gần đó như giao lộ (junction) và bùng binh (roundabout) theo thành phố, quận, bang trong 3 năm 2017- 2019.

Thực hiện câu truy vấn:

Bước 1: Viết câu lệnh MDX như hình bên dưới

```
SELECT  
NON EMPTY {[Measures].[Fact Count]} ON COLUMNS,  
NON EMPTY {  
    [Dim Location].[State].[State].MEMBERS *  
    [Dim Location].[County].[County].MEMBERS *  
    [Dim Location].[City].[City].MEMBERS  
} ON ROWS  
FROM [US ROAD WH]  
WHERE  
Crossjoin(  
[Dim POI].[Junction].[Junction].&[-1],  
[Dim POI].[Roundabout].[Roundabout].&[-1])
```

Bước 2: Click vào **Execute** để thực thi câu lệnh và được kết quả như hình bên dưới

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) environment. On the left, there's a tree view of database objects. In the center, a query editor window displays the following MDX code:

```
--Cau10
SELECT
    NON EMPTY {[Measures].[Fact Count]} ON COLUMNS,
    NON EMPTY {
        [Dim Location].[State].[State].MEMBERS *
        [Dim Location].[County].[County].MEMBERS *
        [Dim Location].[City].[City].MEMBERS
    } ON ROWS
FROM [US ROAD WH]
WHERE
Crossjoin(
    [Dim POI].[Junction].[Junction].&[-1],
    [Dim POI].[Roundabout].[Roundabout].&[-1])
```

Below the query editor is a results grid titled "Messages" showing the following data:

			Fact Count
CA	Los Angeles	Palmdale	2
FL	Miami-Dade	Hialeah	1
GA	Clarke	Athens	1
WA	Pierce	Tacoma	1

3.11.11. Câu truy vấn 11

Nội dung câu truy vấn: Cho biết số công trình và mức độ nghiêm trọng trung bình của các công trình đó xảy ra tại bang Texas trong điều kiện thời tiết có liên quan tới bão (Thunderstorm) trong 3 năm 2017 – 2019.

Thực hiện:

Bước 1: Viết câu lệnh **MDX** như sau trên kho dữ liệu đã tạo **[US ROAD WH]**

SELECT

```
{  
    [Measures].[Fact Count],  
    [Measures].[Average Severity]  
} ON COLUMNS,  
NON EMPTY  
{  
    [Dim Time].[Construction_Start].[Start Year].ALLMEMBERS  
} ON ROWS  
FROM [US ROAD WH]  
WHERE  
(  
{  
    [Dim Weather].[Weather Condition].&[Heavy Thunderstorms and Rain],  
    [Dim Weather].[Weather Condition].&[Light Thunderstorms and Rain],  
    [Dim Weather].[Weather Condition].&[Thunderstorm],  
    [Dim Weather].[Weather Condition].&[Thunderstorms and Rain]  
}  
,  
[Dim Location].[State].&[TX]  
);
```

Bước 2: Click vào **Execute** để thực thi câu lệnh và được kết quả như hình bên dưới

```
--Cau11
SELECT
{
    [Measures].[Fact Count],
    [Measures].[Average Severity]
} ON COLUMNS,
NON EMPTY
{
    [Dim Time].[Construction_Start].[Start Year].ALLMEMBERS
} ON ROWS
FROM [US ROAD WH]
WHERE
(
{
    [Dim Weather].[Weather Condition].&[Heavy Thunderstorms and Rain],
    [Dim Weather].[Weather Condition].&[Light Thunderstorms and Rain],
    [Dim Weather].[Weather Condition].&[Thunderstorm],
    [Dim Weather].[Weather Condition].&[Thunderstorms and Rain]
}
,
[Dim Location].[State].&[TX]
);

```

100 %

	Fact Count	Average Severity
2017	10	2
2018	52	3.69230769230769
2019	3	3.33333333333333

3.11.12. Câu truy vấn 12

Nội dung câu truy vấn: Cho biết tiểu bang có số công trình nhiều nhất và tiểu bang có số công trình ít nhất theo từng năm.

Thực hiện câu truy vấn:

Bước 1: Viết câu lệnh MDX như sau trên kho dữ liệu đã tạo **[US ROAD WH]**

SELECT

NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,

NON EMPTY {

[Dim Time].[Construction_Start].[Start Year].ALLMEMBERS *

```

GENERATE(
    Descendants([Dim Time].[Start Year].[Start Year].MEMBERS, [Dim Time].[Start
Year].[Start Month]),
UNION (
    TopCount([Dim Location].[State].CHILDREN, 1, [Measures].[Fact Count]),
    BottomCount([Dim Location].[State].CHILDREN, 1, [Measures].[Fact Count])
)
)
} ON ROWS
FROM [US ROAD WH];

```

Bước 2: Click vào **Execute** để thực thi câu lệnh và được kết quả như hình bên dưới

The screenshot shows the SQL Server Management Studio interface. On the left, there's a tree view labeled 'dation'. In the center, a query window titled 'Cau12' contains the following T-SQL code:

```

-- Cau12
SELECT
    NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,
    NON EMPTY [
        [Dim Time].[Construction_Start].[Start Year].ALLMEMBERS *
        GENERATE(
            Descendants([Dim Time].[Start Year].[Start Year].MEMBERS, [Dim Time].[Start Year].[Start Month]),
            UNION (
                TopCount([Dim Location].[State].CHILDREN, 1, [Measures].[Fact Count]),
                BottomCount([Dim Location].[State].CHILDREN, 1, [Measures].[Fact Count])
            )
        )
    } ON ROWS
FROM [US ROAD WH];

```

Below the query window, the results pane shows a table titled 'Fact Count' with the following data:

		Fact Count
2017	PA	19130
2017	AL	45
2018	PA	18495
2018	AL	268
2019	PA	28882
2019	AL	1514

3.11.13. Câu truy vấn 13

Nội dung câu truy vấn: Cho biết tốc độ hoàn thành trung bình của mỗi công trình ở gần vị trí gờ giảm tốc (bump) theo từng bang trong cả 3 năm 2017 – 2019.

Thực hiện câu truy vấn:

Bước 1: Viết câu lệnh MDX như sau trên kho dữ liệu đã tạo [US ROAD WH]

WITH MEMBER [Measures].[Custom Calculation] AS

IIF(

ROUND(

SUM([Measures].[Sum Duration]) / 86400 / COUNT([Measures].[Fact Count]), 0

) = 0,

1,

ROUND(

SUM([Measures].[Sum Duration]) / 86400 / COUNT([Measures].[Fact Count]), 0

)

)

SELECT

[Measures].[Custom Calculation] ON COLUMNS,

[Dim Location].[State].[State].MEMBERS ON ROWS

FROM [US ROAD WH]

WHERE [Dim POI].[Bump].[Bump].&[-1]

Bước 2: Click vào **Execute** để thực thi câu lệnh và được kết quả như hình bên dưới

```
--Cau13
WITH MEMBER [Measures].[Custom Calculation] AS
  IIF(
    ROUND(
      SUM([Measures].[Sum Duration]) / 86400 / COUNT([Measures].[Fact Count]), 0
    ) = 0,
    1,
    ROUND(
      SUM([Measures].[Sum Duration]) / 86400 / COUNT([Measures].[Fact Count]), 0
    )
  )
SELECT
  [Measures].[Custom Calculation] ON COLUMNS,
  [Dim Location].[State].[State].MEMBERS ON ROWS
FROM [US ROAD WH]
WHERE [Dim POI].[Bump].[Bump].&[-1]
```

100 %

	Custom Calculation
AL	1
AR	1
AZ	12
CA	26
CO	4
CT	1
DC	1
DE	1
FL	4
GA	1
IA	1

3.11.14. Câu truy vấn 14

Nội dung câu truy vấn: Thông kê trung bình mức độ nghiêm trọng của các công trình thi công đường tại các bang ở dưới mức 3 trong năm 2017 nhưng đã lên hơn mức 3 trong năm 2018.

Thực hiện câu truy vấn:

Bước 1: Viết câu lệnh MDX như sau trên kho dữ liệu đã tạo [US ROAD WH]

SELECT

Ths. Đỗ Thị Minh Phụng

NON EMPTY { [Measures].[Average Severity] } ON COLUMNS,

NON EMPTY

(

[Dim Location].[State].[State].ALLMEMBERS

*

[Dim Time].[Start Year].[Start Year].ALLMEMBERS

)

ON ROWS

FROM

(

SELECT

(

[SeverityUp]

)

ON COLUMNS

FROM

(

SELECT

{

[Dim Time].[Start Year]&[2017],

[Dim Time].[Start Year]&[2018]

}

ON COLUMNS

```

FROM
[US ROAD WH]
)
)

```

Bước 2: Click vào **Execute** để thực thi câu lệnh và được kết quả như hình bên dưới

```

--Cau14
SELECT
    NON EMPTY { [Measures].[Average Severity] } ON COLUMNS,
    NON EMPTY (
        (
            [Dim Location].[State].[State].ALLMEMBERS
            *
            [Dim Time].[Start Year].[Start Year].ALLMEMBERS
        )
    ) ON ROWS
FROM
(
    SELECT
        (
            [SeverityUp]
        )
    ON COLUMNS
    FROM
    (
        SELECT
            {
                [Dim Time].[Start Year].&[2017],
                [Dim Time].[Start Year].&[2018]
            }
        ON COLUMNS
        FROM
        [US ROAD WH]
    )
)

```

	Average Severity
FL 2017	2.41686893203884
FL 2018	3.0218043063505
IL 2017	2.67758186397985
IL 2018	3.07072570725707
IN 2017	2.489443378119
IN 2018	3.02968460111317
KS 2017	2.63938618925831

3.11.15. Câu truy vấn 15

Nội dung câu truy vấn: Thông kê thời gian thi công trung bình và quãng đường thi công trung bình theo từng thành phố trong điều kiện thời tiết khắc nghiệt là có tốc độ gió trên 70mph,

Thực hiện câu truy vấn:

Bước 1: Viết câu lệnh MDX như sau trên kho dữ liệu đã tạo [US ROAD WH]

```
SELECT NON EMPTY { [Measures].[Average Duration], [Measures].[Average Distance] } ON COLUMNS,  
NON EMPTY { ([Dim Location].[City].[City].ALLMEMBERS ) } ON ROWS  
FROM  
( SELECT ( FILTER( [Dim Weather].[Wind Speed Mph].CHILDREN, [Dim  
Weather].[Wind Speed Mph].CurrentMember.MemberValue > 70 ) ) ON COLUMNS  
FROM [US ROAD WH]  
)
```

Bước 2: Click vào **Execute** để thực thi câu lệnh và được kết quả như hình bên dưới

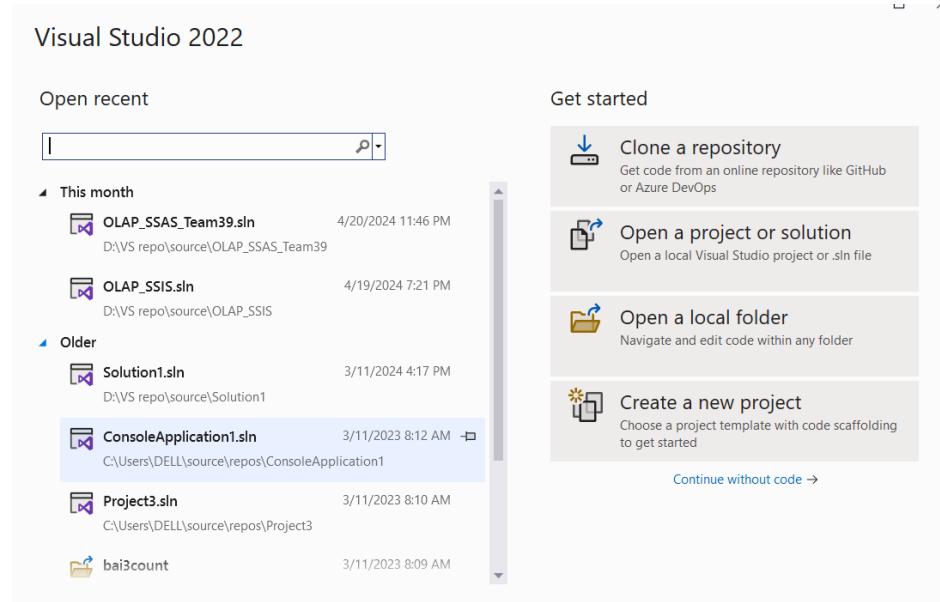
```
--Cau15  
SELECT  
NON EMPTY { [Measures].[Average Duration], [Measures].[Average Distance] } ON COLUMNS,  
NON EMPTY { ([Dim Location].[City].[City].ALLMEMBERS ) } ON ROWS  
FROM  
( SELECT ( FILTER( [Dim Weather].[Wind Speed Mph].CHILDREN, [Dim Weather].[Wind Speed Mph].CurrentMember.MemberValue > 70 ) ) ON COLUMNS  
FROM [US ROAD WH]  
)
```

	Average Duration	Average Distance
Beaver	21755	0.679018914699554
Gorham	27240	0.307137370109558
Port Matilda	21600	5.55383634567261
San Pedro	21600	0.968867236309052
Scottsdale	431992	2.61746402829885E-02

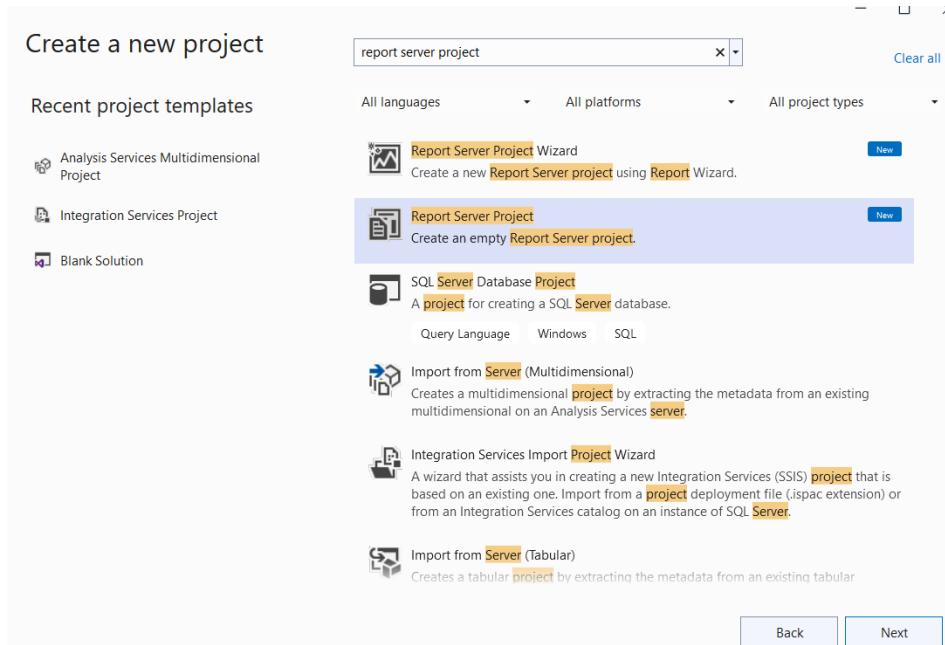
CHƯƠNG 4. QUÁ TRÌNH XÂY DỰNG BÁO CÁO VÀ SỬ DỤNG CÔNG CỤ POWER BI

4.1. Tạo project SSRS và thêm các Datasource và Dataset

Bước 1: Tạo project trong Visual Studio



Bước 2: Nhập tìm kiếm và chọn Report Server Project



Bước 3: Đặt tên project và nhấp **Create**

Ths. Đỗ Thị Minh Phụng

Configure your new project

Report Server Project

Project name

OLAP_SSRS_Team39

Location

D:\VS repo\source



Solution name ⓘ

OLAP_SSRS_Team39

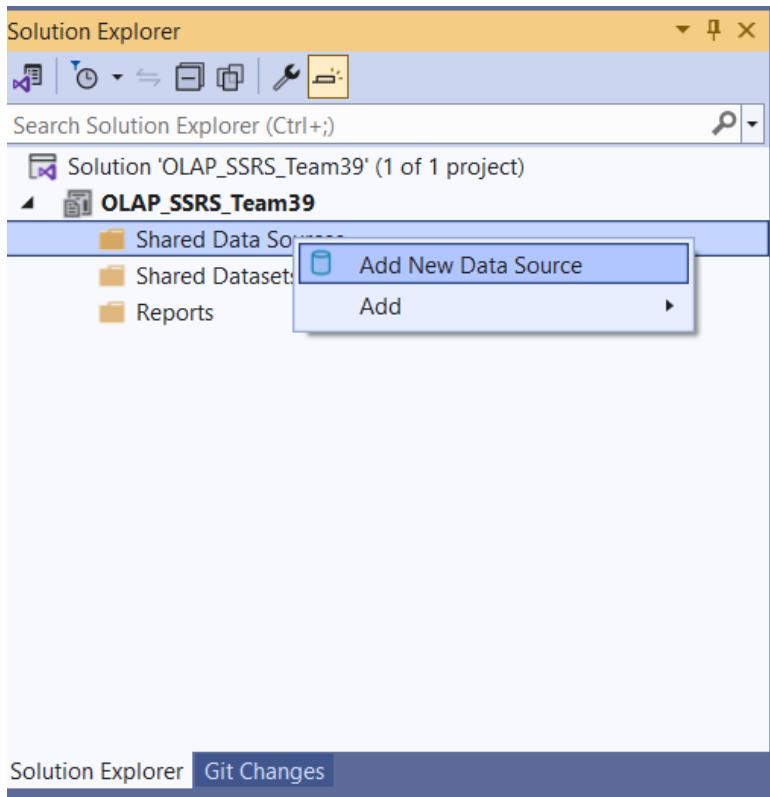
Place solution and project in the same directory

Project will be created in "D:\VS repo\source\OLAP_SSRS_Team39\OLAP_SSRS_Team39"

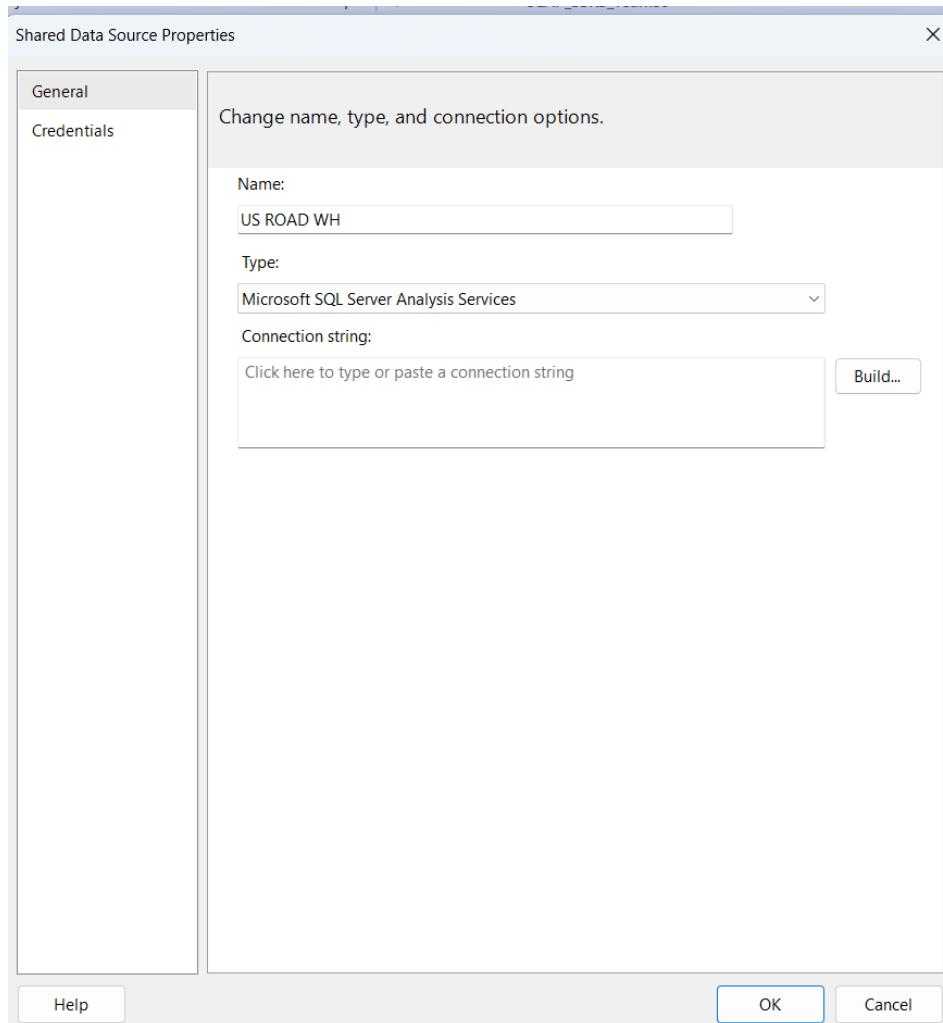
Back

Create

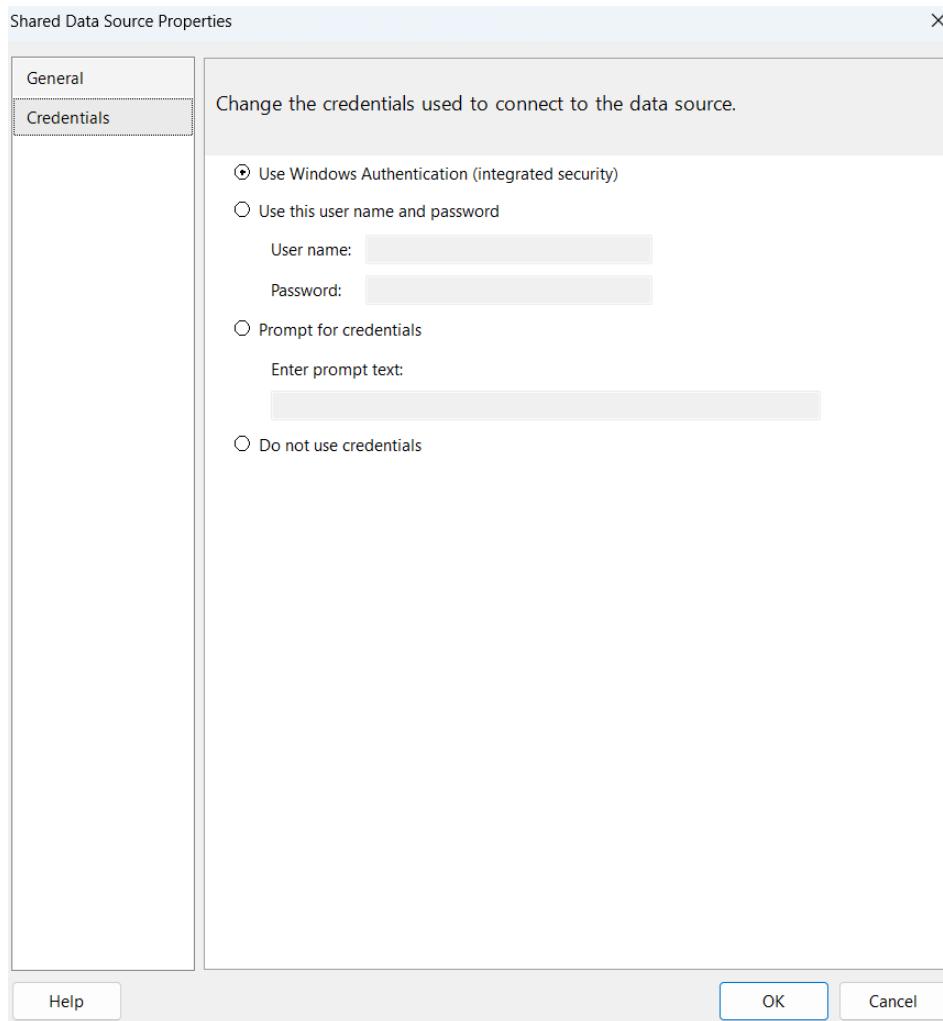
Bước 4: Click chuột phải lên **Shared Data Sources** trong cửa sổ **Solution** và chọn **Add New Data Source** để thực hiện kết nối với **Data Sources**.



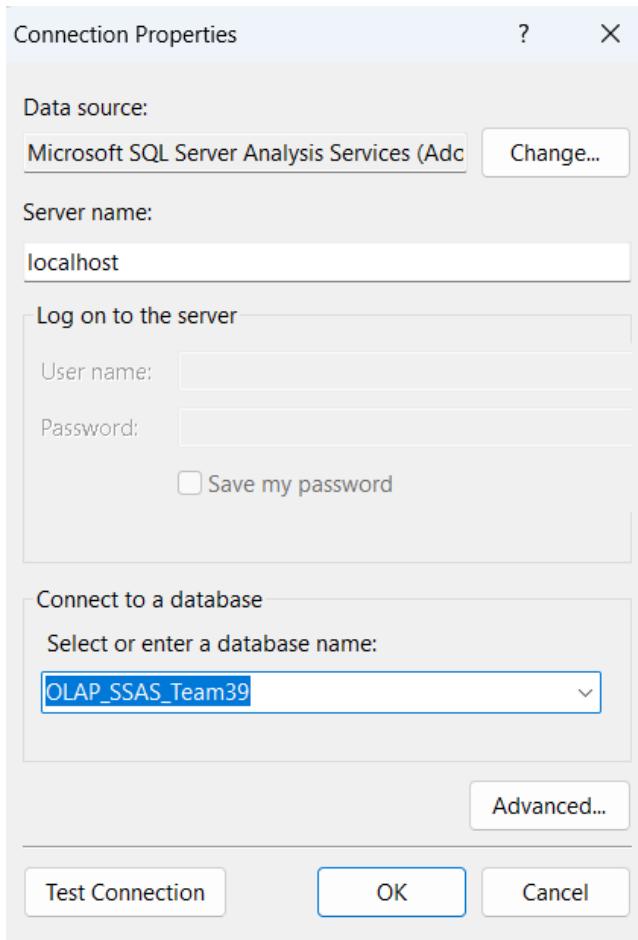
Bước 5: Đặt tên cho **Data source** đồng thời chọn **Type** là **Microsoft SQL Server Analysis Services** để kết nối với database có các độ đo từ quá trình SSAS



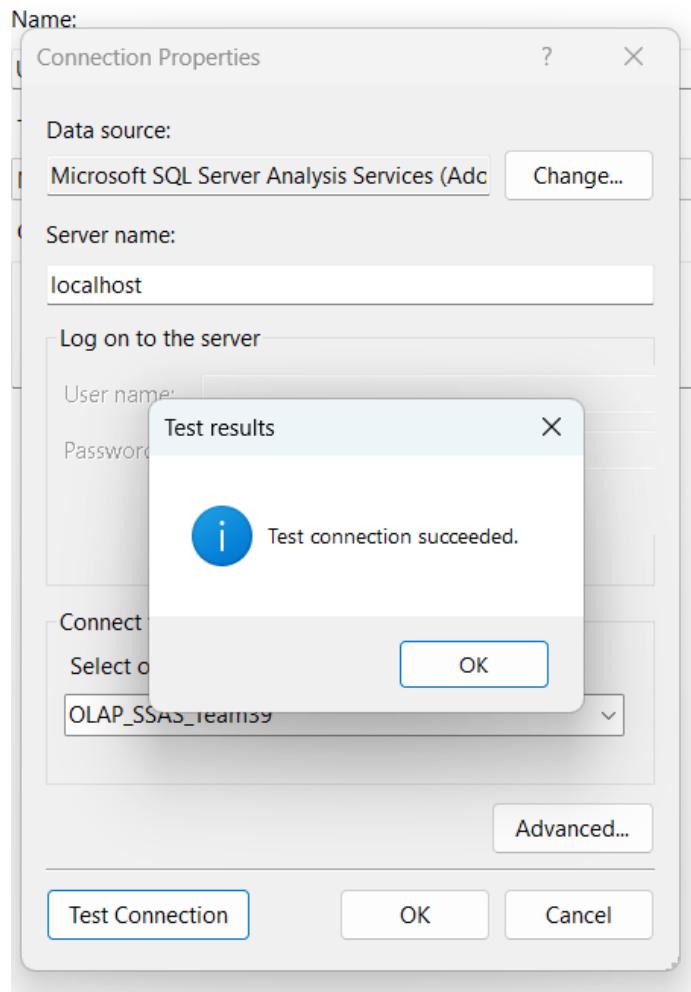
Bước 6: Chuyển sang tab **Credentials** và tick vào **Use Windows Authentication (integrated security)** và nhấn **Ok**



Bước 7: Click vào **Build**, nhập **server name** là **localhost** vào cửa sổ vừa hiện lên. Nhập **database name** là tên của database sau khi deploy từ quá trình SSAS và nhấn **Test Connection** để kiểm tra kết nối



Bước 8: Kết nối thành công và nhấn **OK**



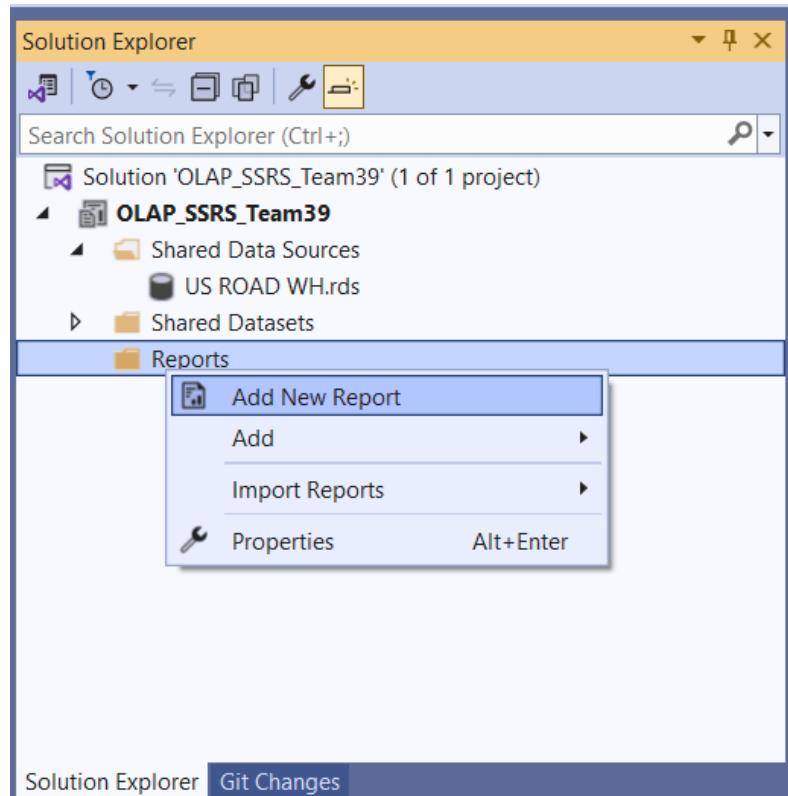
4.2. Tạo các báo cáo từ project SSRS

4.2.1. Báo cáo 1

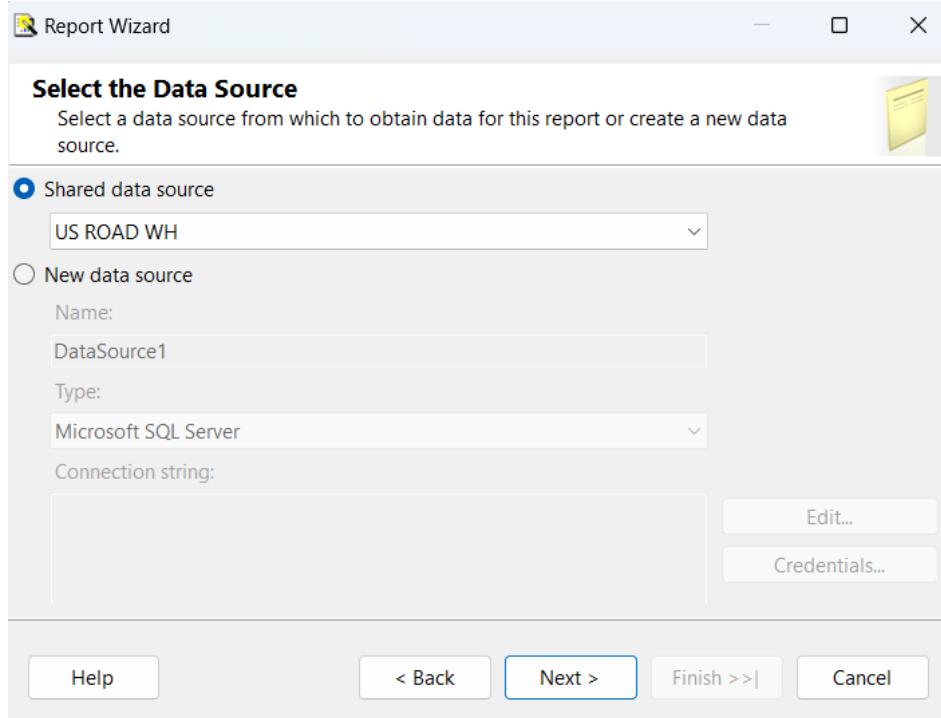
Nội dung câu truy vấn: Thống kê số lượng công trình làm ảnh hưởng tới đường bộ xảy ra tại 5 quận Los Angeles, San Diego, San Francisco, Orange và Alameda của bang California (CA) trong 3 năm liên tục 2017 – 2019.

4.2.1.1. Tạo báo cáo

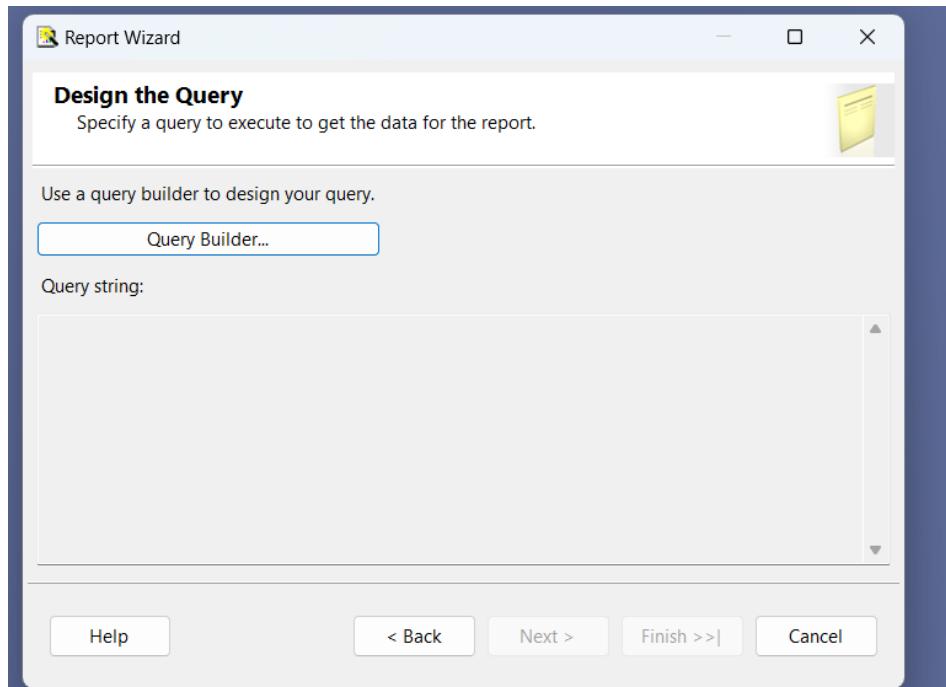
Bước 1: Click chuột phải lên thư mục **Reports** chọn **Add New Report**



Bước 2: tick chọn **Shared data source** và chọn vào **US ROAD WH** và nhấn **Next**



Bước 3: Sử dụng **Query Builder** để tạo query data dùng cho report. Click chọn **Query Builder**



Query Designer

Edit as Text Import... MDX

Dimension Hierarchy Operator Filter Expression Param...

Dim Location Hierarcy Equal [Alameda, Los Angeles, Orange, San Di...]

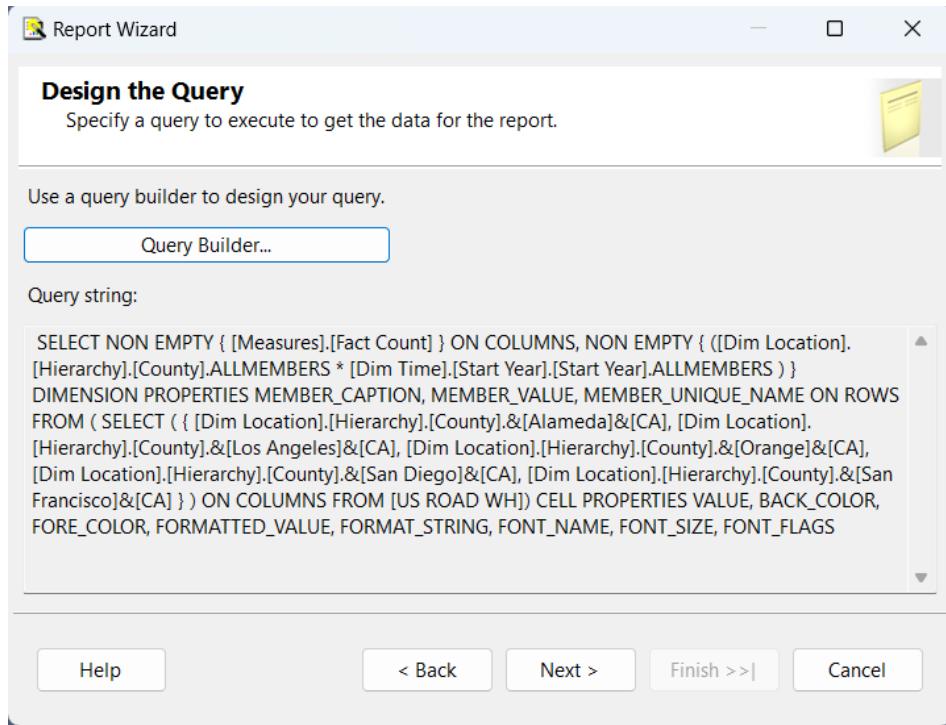
<Select dimension>

State	County	Start Year	Fact Count
CA	Alame...	2017	211
CA	Alame...	2018	644
CA	Alame...	2019	859
CA	Los An...	2017	1774
CA	Los An...	2018	2471
CA	Los An...	2019	6324
CA	Orange	2017	262
CA	Orange	2018	356
CA	Orange	2019	1925
CA	San Di...	2017	392
CA	San Di...	2018	338
CA	San Di...	2019	2495
CA	San Fr...	2017	143
CA	San Fr...	2018	250
CA	San Fr...	2019	509

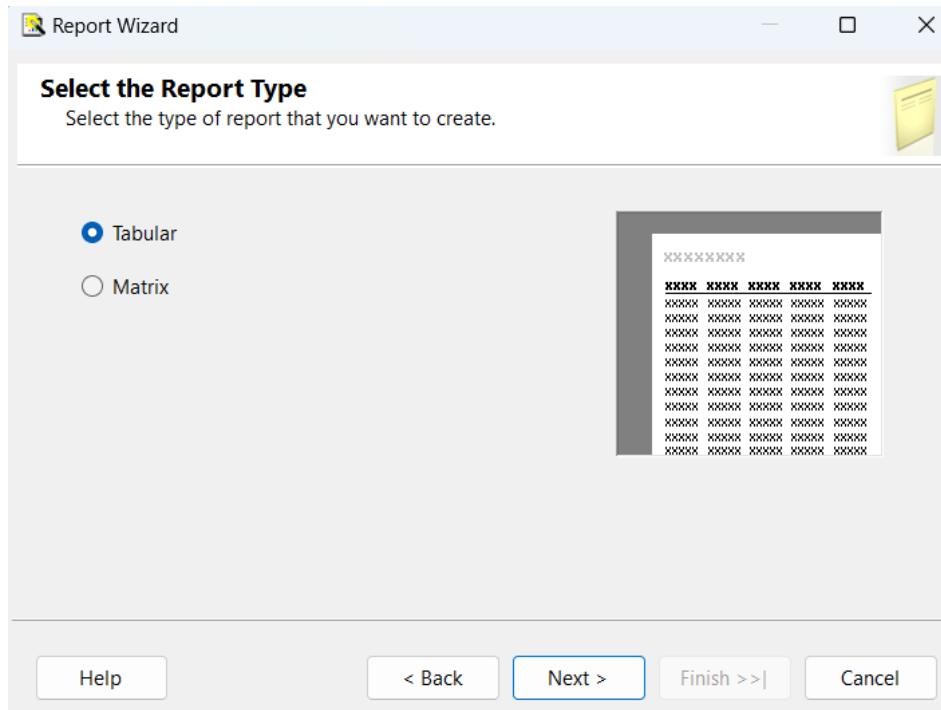
Calculated Members

Help OK Cancel

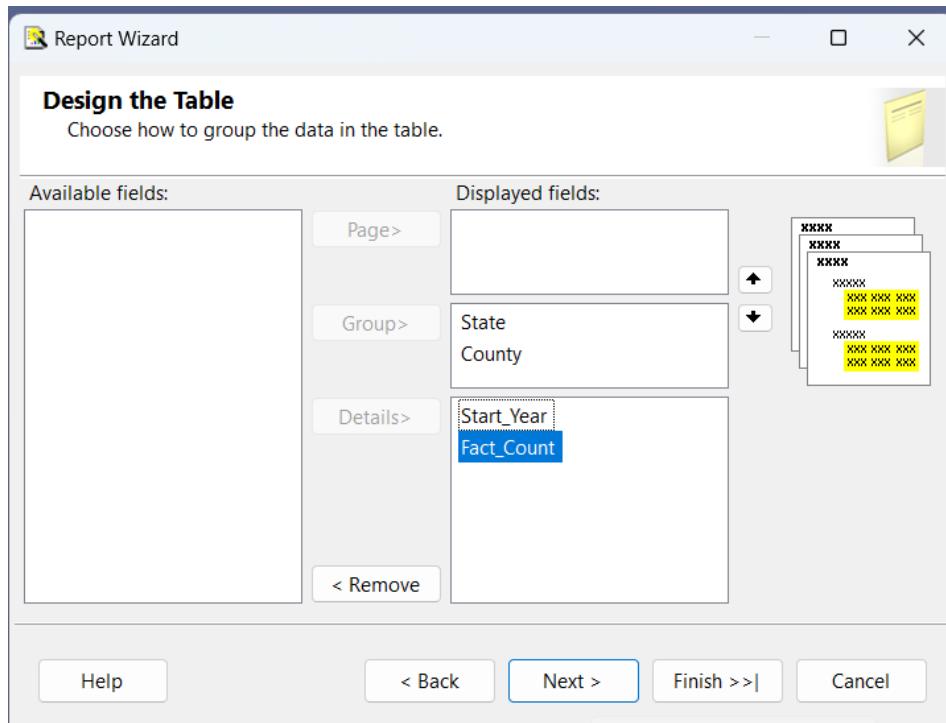
Bước 4: Chuỗi câu lệnh truy vấn được hiện ra như sau và nhấn Next



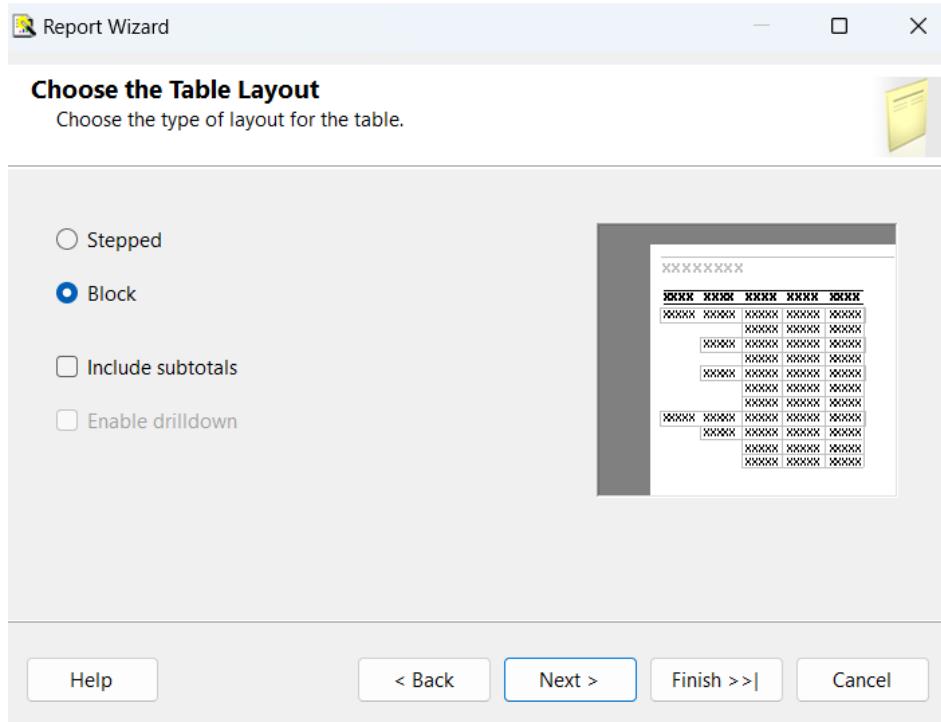
Bước 5: Chọn kiểu báo cáo (dạng bảng hoặc ma trận) sau đó nhấn Next



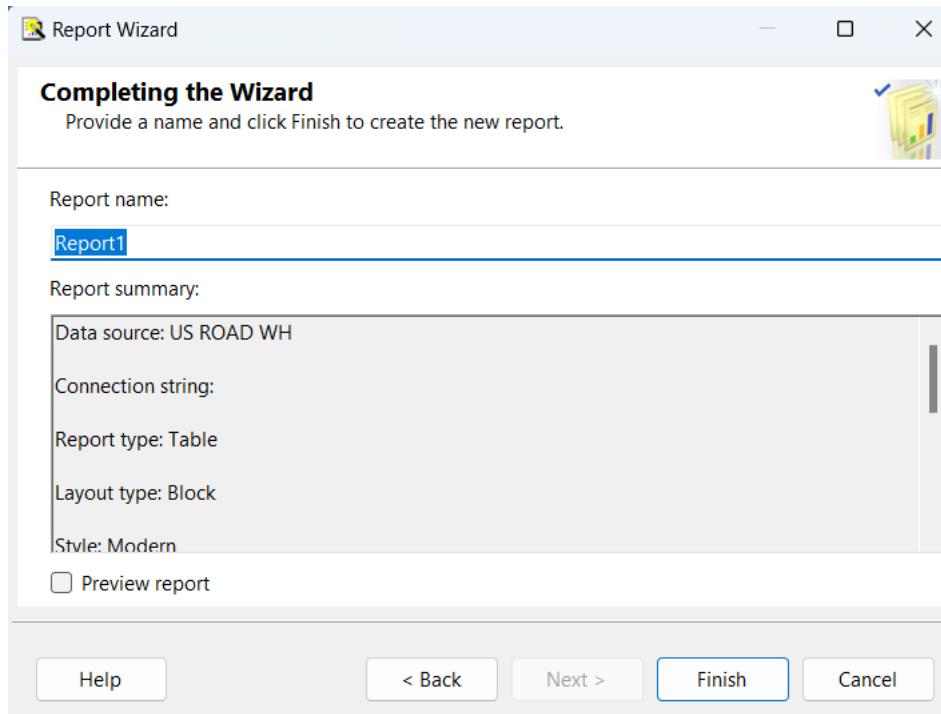
Bước 6: Đưa thuộc tính State, County vào Group, độ đo Fact_Count và thuộc tính Start_Year vào Details sau đó nhấn Next



Bước 7: Tick chọn **Block** để chọn layout cho báo cáo, sau đó nhấn **Next**



Bước 8: đặt tên cho Report và kiểm tra **Report summary** sau đó nhấn **Finish**



Bước 9: Sau khi nhấn **Finish**, sẽ được màn hình bên dưới, chuyển sang tab **Preview** để xem trước báo cáo

Ths. Đỗ Thị Minh Phụng

The screenshot shows the Report Designer interface for a report named "Report1.rdl". The main area displays a table with four columns: "State", "County", "Start Year", and "Fact Count". The table has three row groups: "table1_State", "table1_County", and "table1_Details_Group". The "Output" pane at the bottom shows log messages:

```
Output
Show output from: Windows Forms
[09:00:50.434504] warn: Failed to retrieve the logging service, no log files will be created.
[09:00:53.768923] warn: Failed to retrieve the logging service, no log files will be created.
```

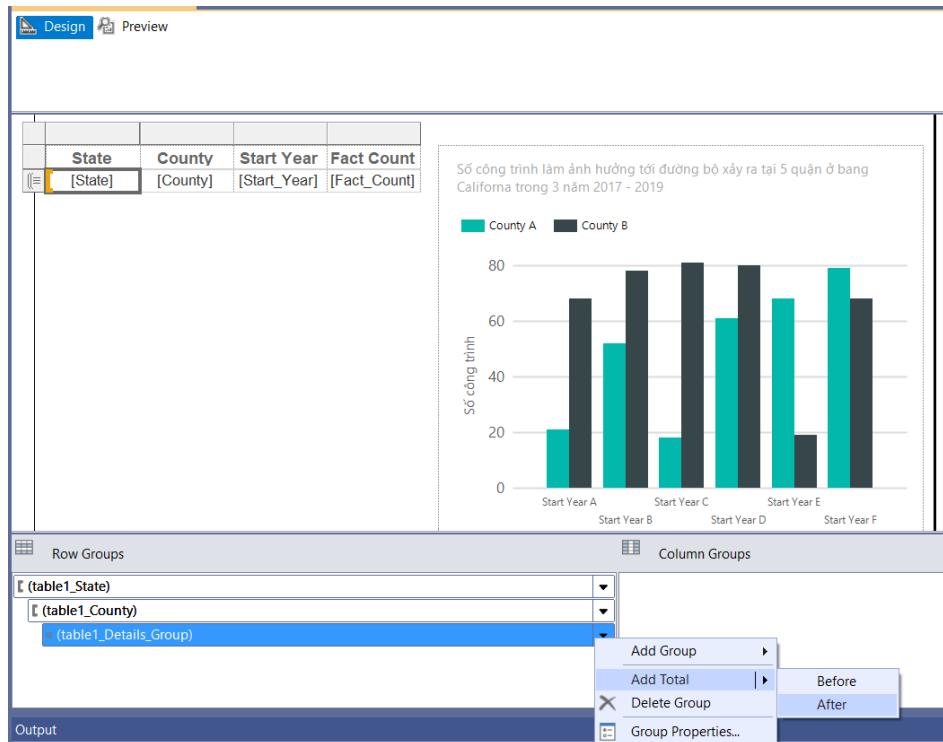
The screenshot shows the Report Designer interface for Report1.rdl [Design]. The main area displays a table titled "Report1" with the following data:

State	County	Start Year	Fact Count	
CA	Alameda	2017	211	
		2018	644	
		2019	859	
Los Angeles	2017	1774		
		2018	2471	
		2019	6324	
Orange	2017	262		
		2018	356	
		2019	1925	
San Diego	2017	392		
		2018	338	
		2019	2495	
San Francisco	2017	143		
		2018	250	
		2019	509	

The bottom section shows the Output window with the following log:

```
Show output from: Build
Skipping 'Report1.rdl'. Item is up to date.
Build complete -- 0 errors, 0 warnings
=====
===== Build: 1 succeeded on up-to-date, 0 failed, 0 skipped =====
===== Build completed at 9:02 AM and took 00.253 seconds =====
```

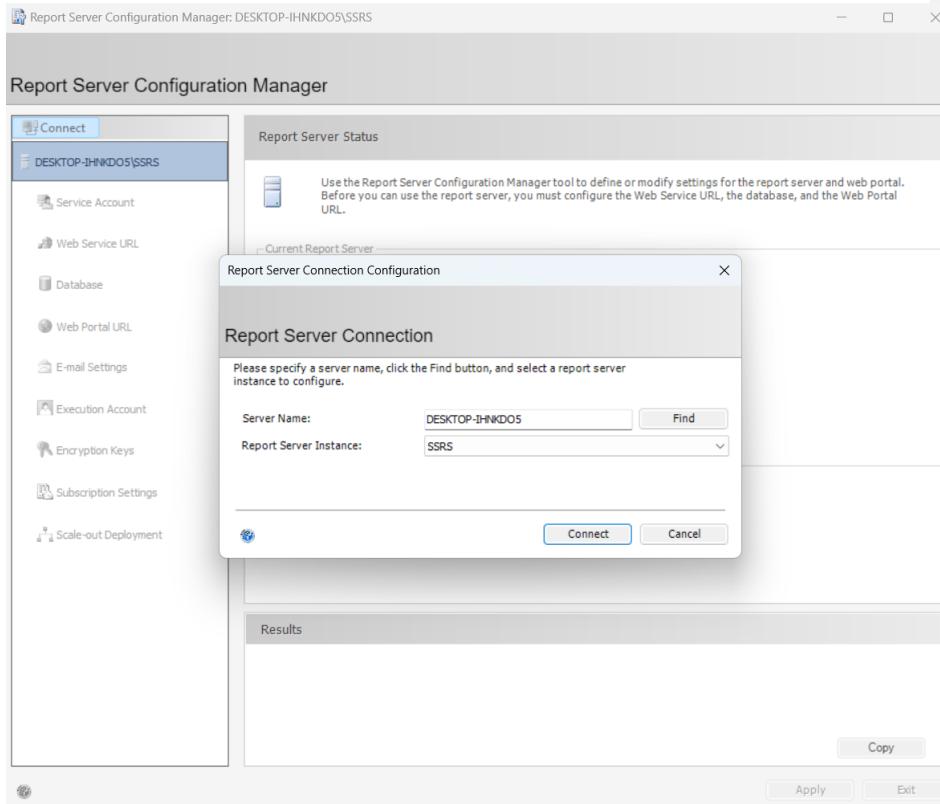
Bước 10: Click chuột phải vào (**table1_Details_Group**) trong nhóm **Row Groups**, chọn **Add Total** và chọn **After** để thêm tổng cho cột **Fact_Count**



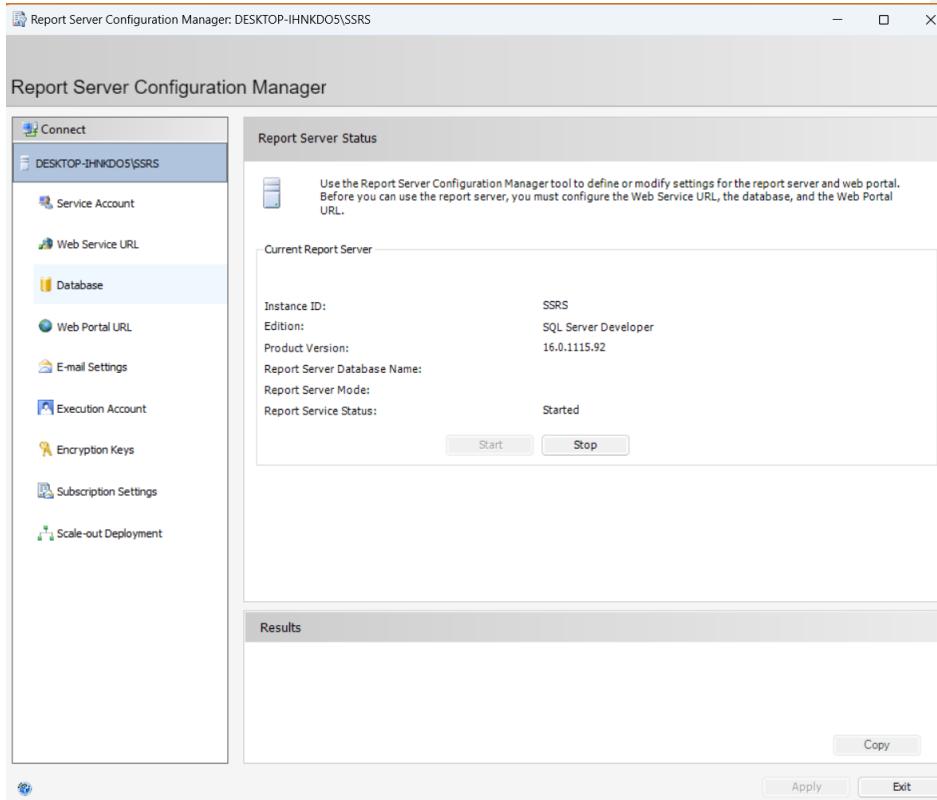
4.2.1.2. Xuất báo cáo

- Cấu hình Report Server Configuration Manager

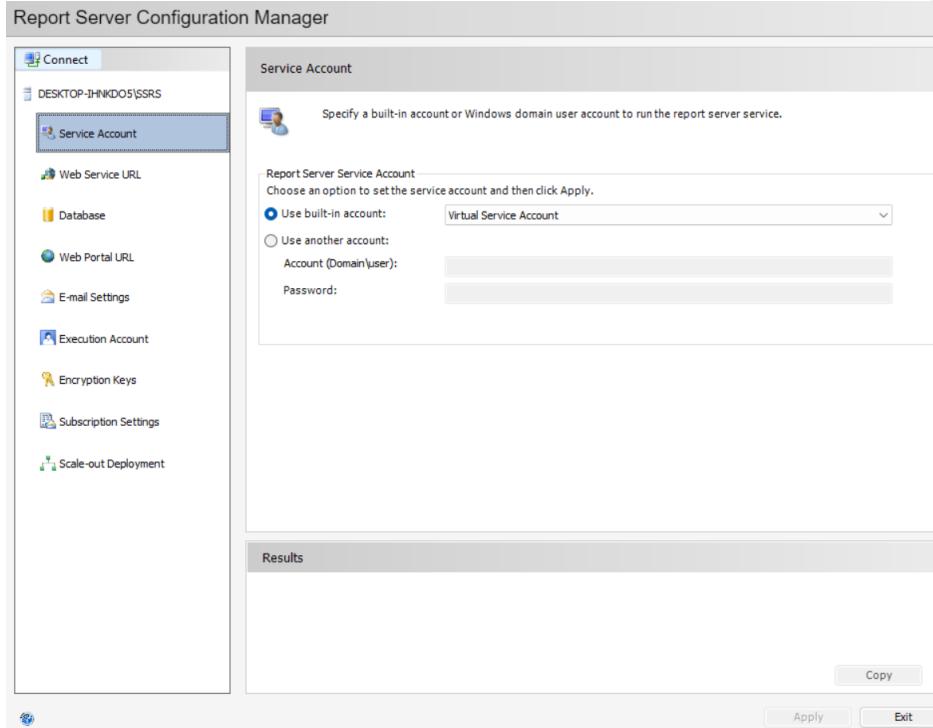
Bước 1: Mở Report Server Configuration Manager bằng công cụ Report Server Configuration Connection trong thanh tìm kiếm hoặc chọn Configure report server



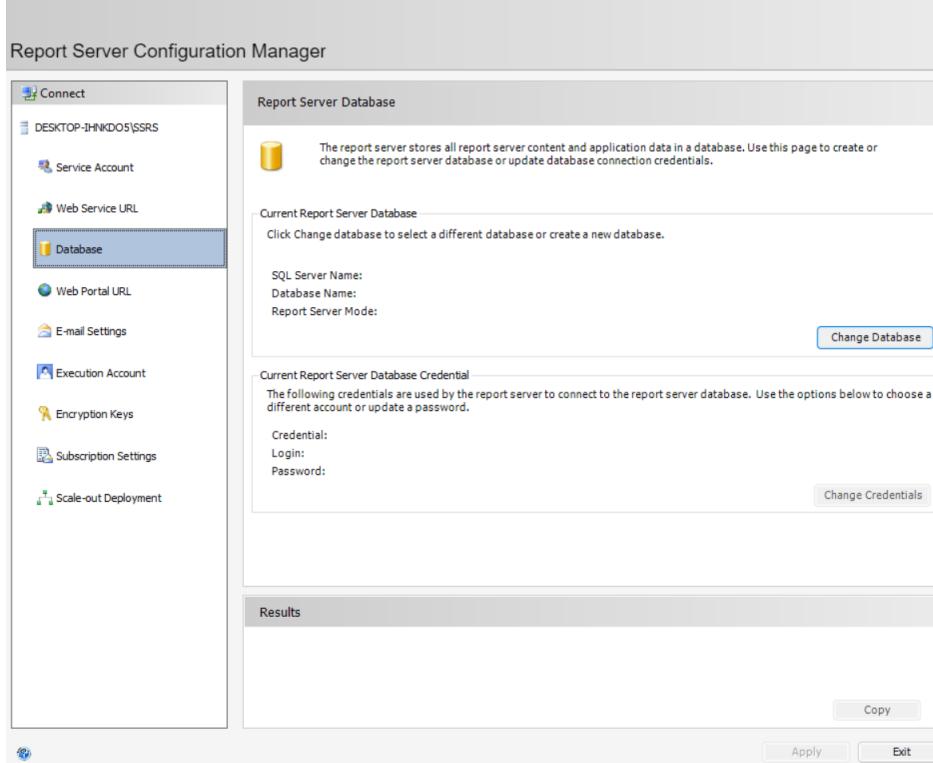
Bước 2: Giao diện tổng thể của công cụ **Report Server Configuration Manager**.



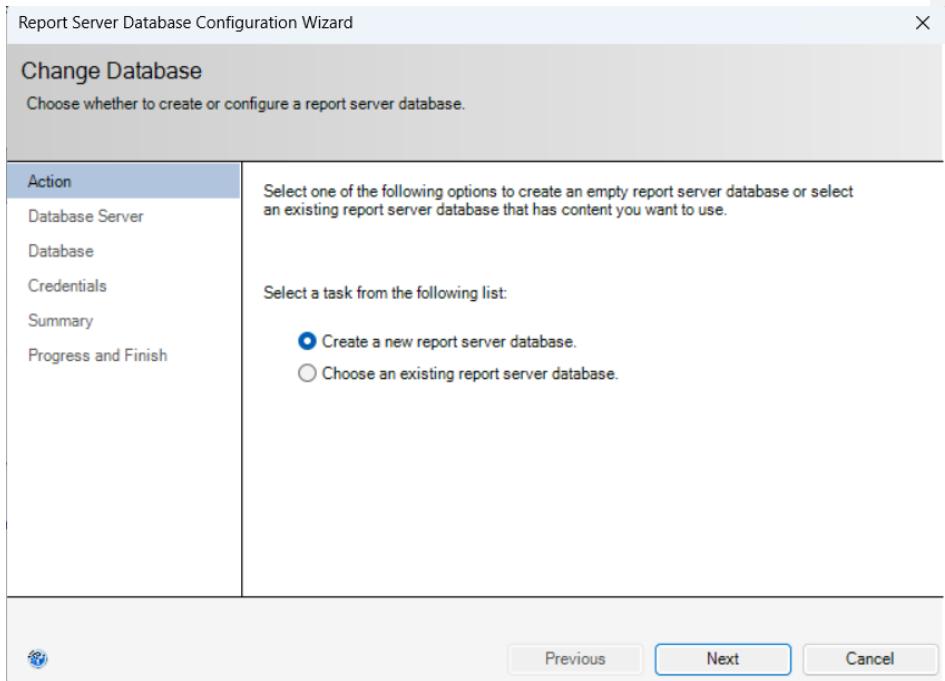
Bước 3: Chọn vào tab **Service Account**, tick chọn **Use built – in account** và chọn **Network Service** và nhấn **Apply**



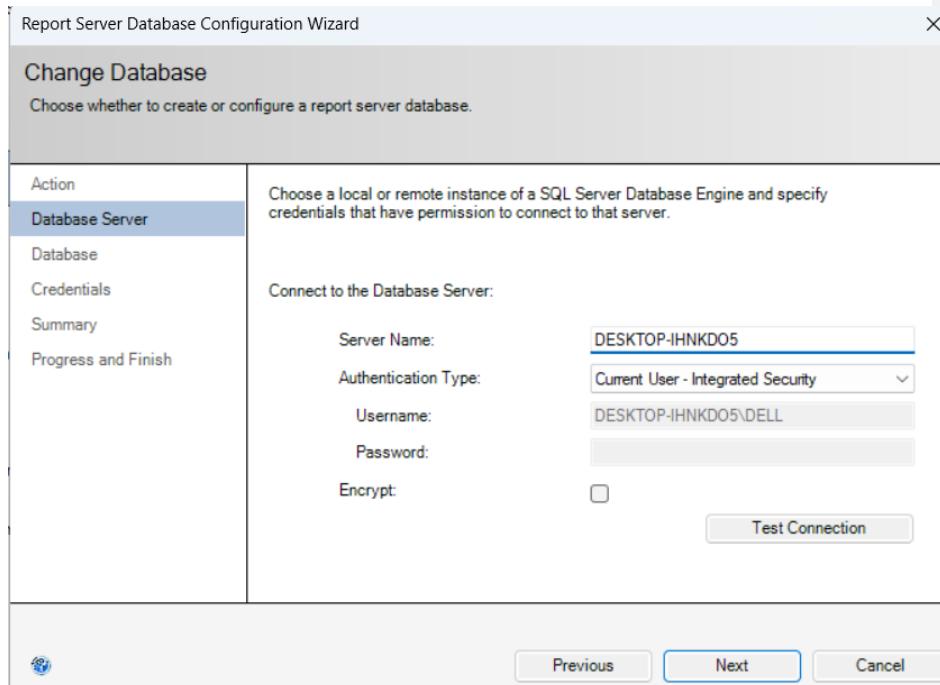
Bước 4: Chuyển đến tab **Database**, chọn **Change Database** để tạo mới 1 database



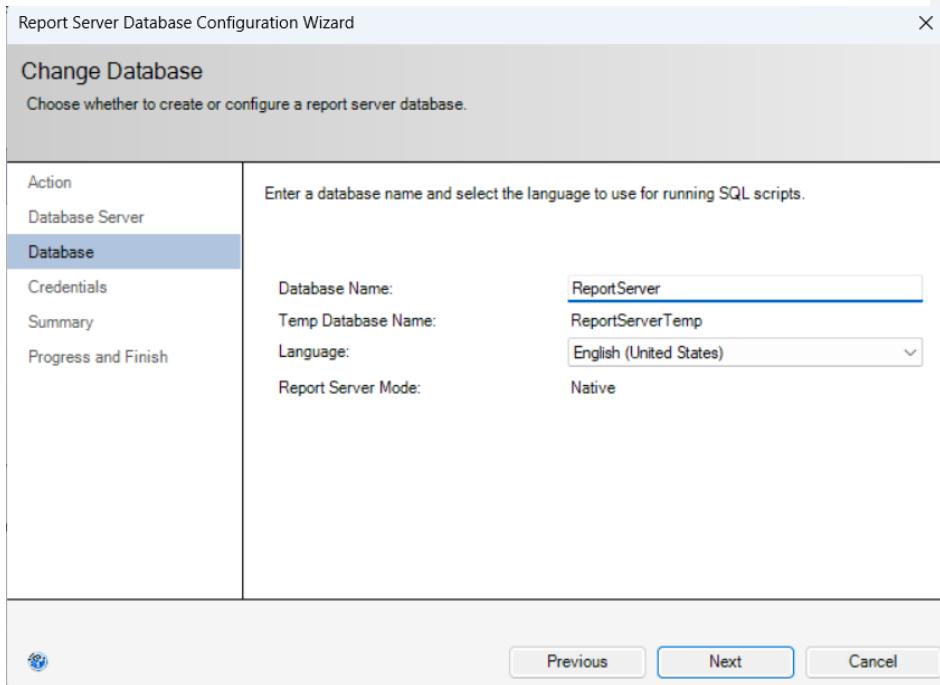
Bước 5: Chọn **Create a new report server database** và nhấn **Next**



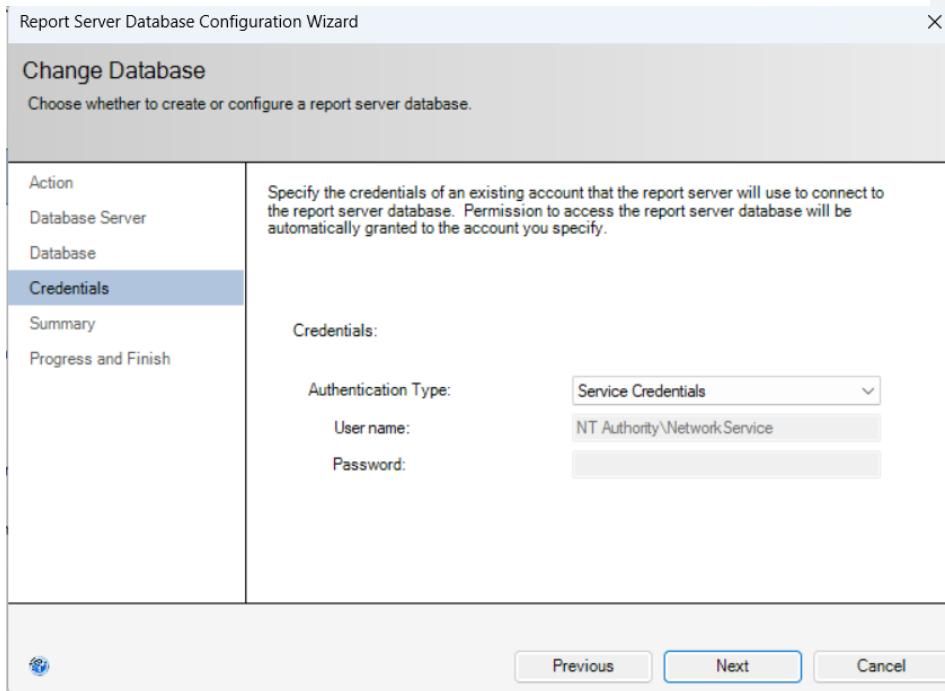
Bước 6: Điền tên server vào **Server Name** và nhấn **Next**



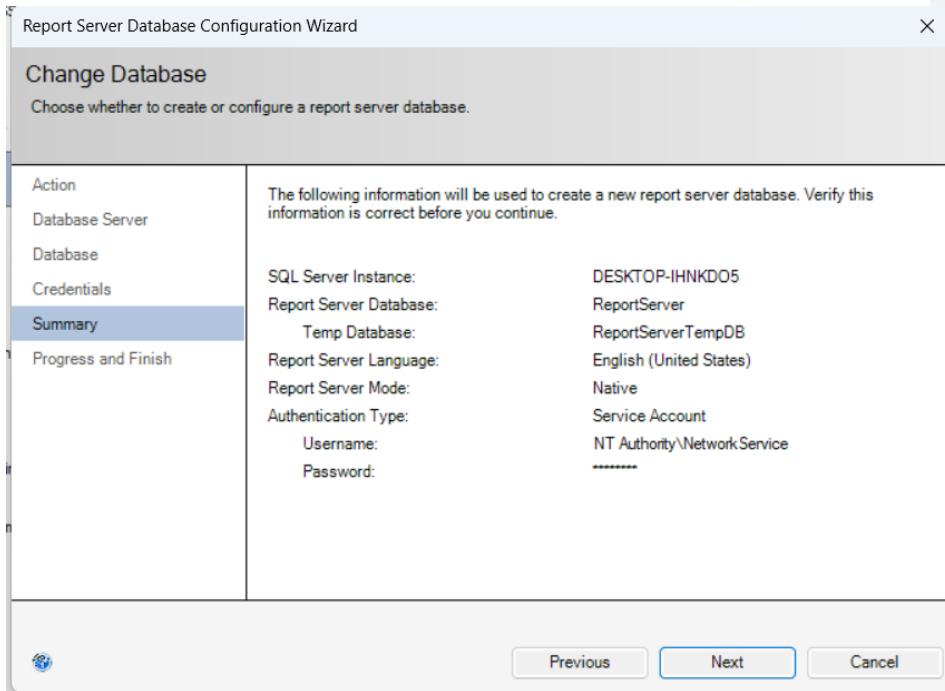
Bước 7: Đặt tên cho Database và nhấn Next



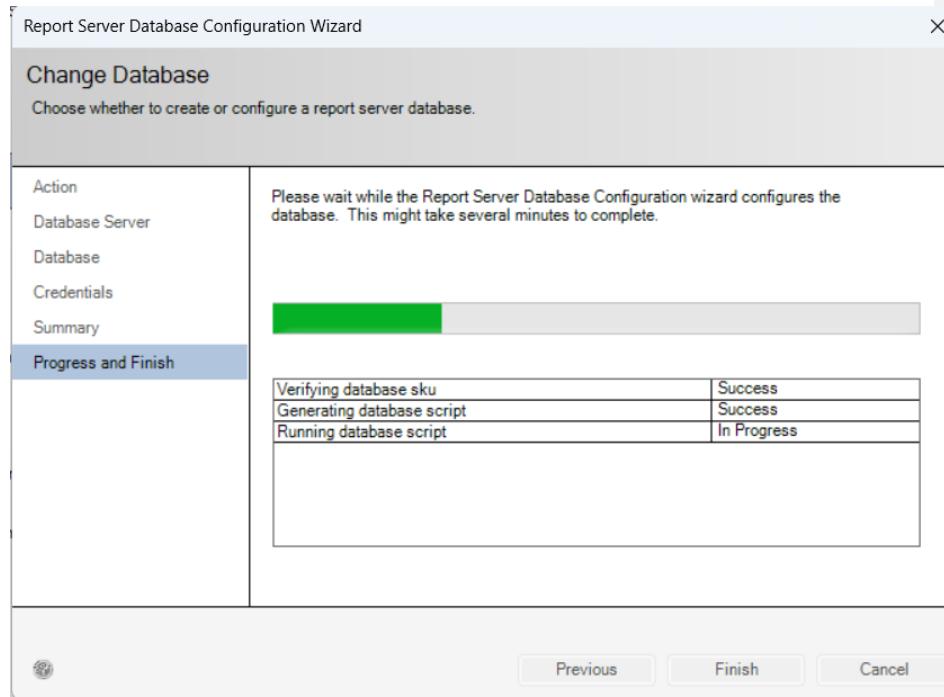
Bước 8: Chuyển sang tab **Credentials** và nhấn **Next**



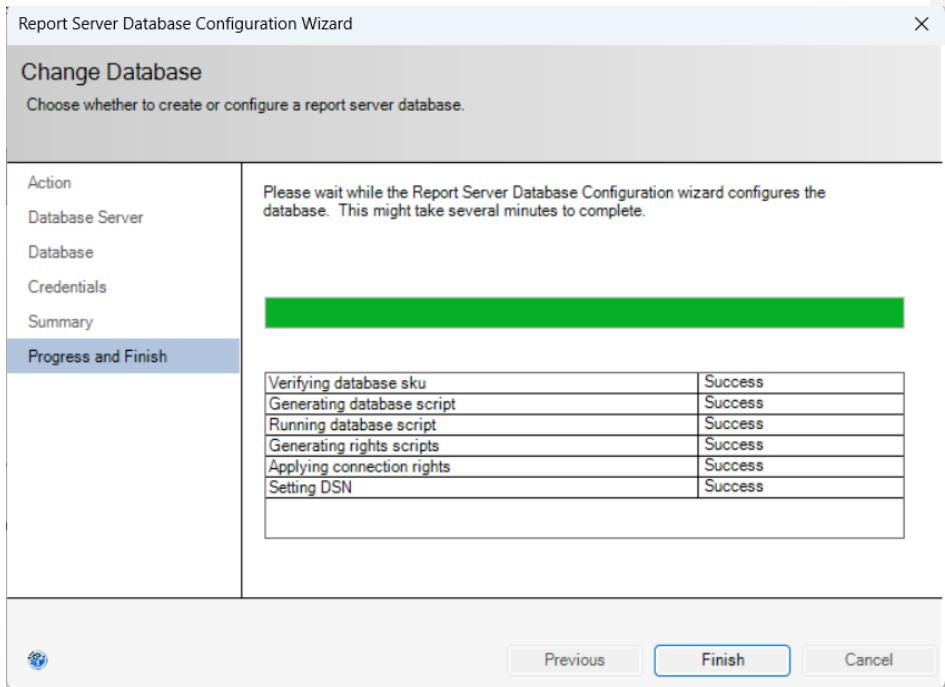
Bước 9: Chuyển sang tab **Summary** và nhấp **Next**



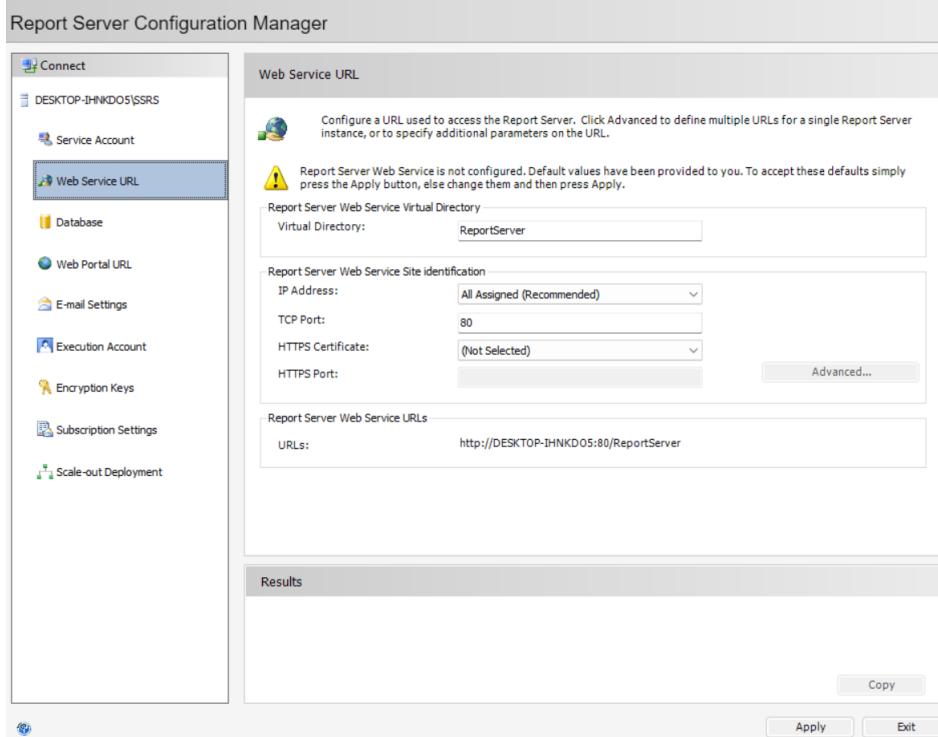
Bước 10: Chờ hệ thống tạo database

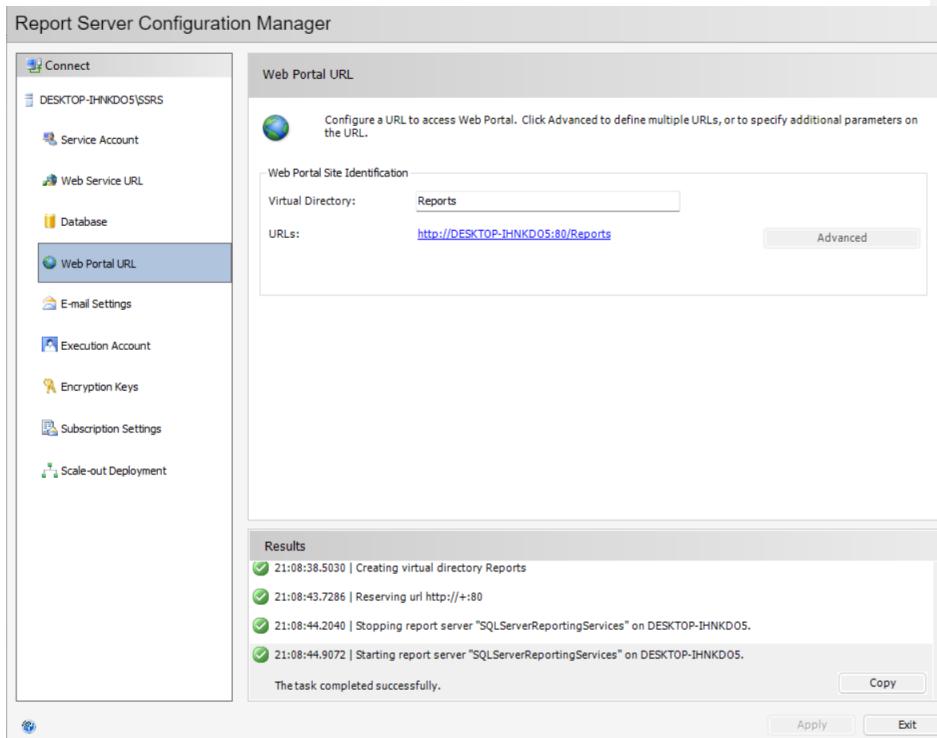


Bước 11: Tạo database thành công



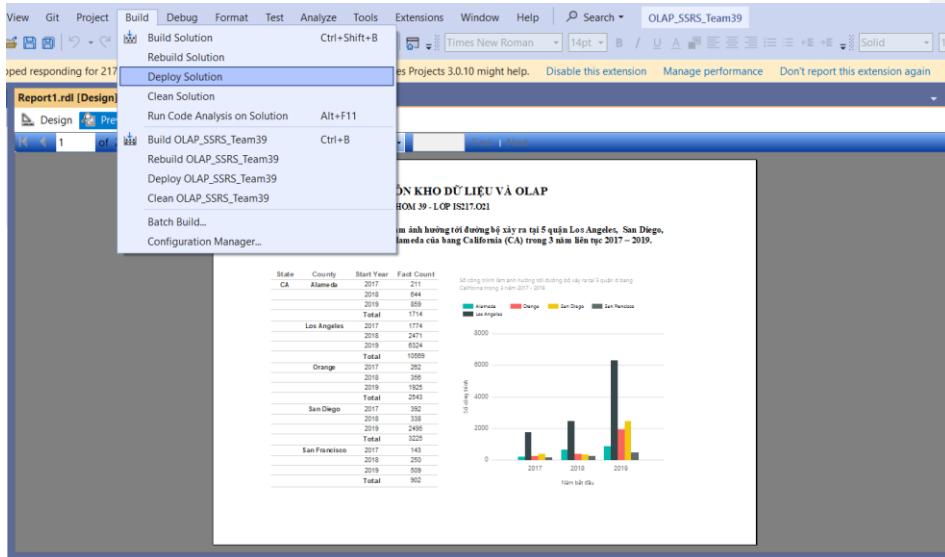
Bước 12: Chuyển sang tab **Web Service URL** và **Web Portal URL** xem các thông tin cấu hình, kiểm tra ô **Virtual Directory là Report** và nhấn **Apply**



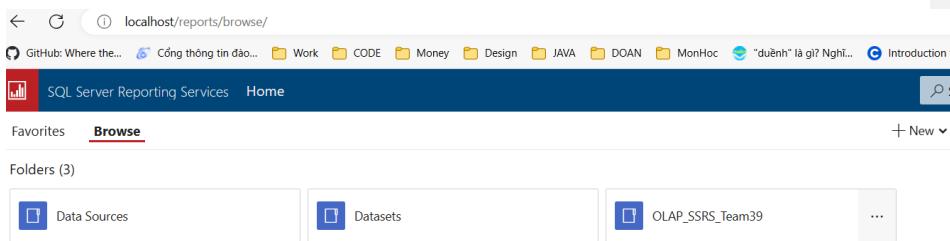


Bước 13: Chọn **Build** trên thanh menu của project SSRS -> Deploy Solution

Ths. Đỗ Thị Minh Phụng



Bước 14: Truy cập vào địa chỉ **localhost/reports** bằng trình duyệt web



Bước 15: Mở thư mục **OLAP_SSRS_Team39** và mở Report **Report1**

The screenshot shows the 'Browse' section of the SQL Server Reporting Services interface. At the top, there's a red header bar with the SSRS logo and the text 'SQL Server Reporting Services'. Below it, the URL 'Home > OLAP_SSRS_Team39' is visible. The main area has two tabs: 'Favorites' and 'Browse', with 'Browse' being the active tab. Underneath, a heading says 'Paginated Reports (1)'. A single report item is listed: 'Report1' with a blue icon. To the right of the report name are three small icons: an ellipsis, a star, and a printer.

Bước 16: Nhấn nút in để xuất ra file pdf báo cáo **report1**.

This screenshot displays the content of the 'Report1' report. At the top, the title 'ĐỒ ÁN MÔN KHO DỮ LIỆU VÀ OLAP' and subtitle 'NHÓM 39 - LỚP IS217.O21' are shown. Below the title, a descriptive text block states: 'Thông kê số lượng công trình làm ảnh hưởng tới đường bộ xảy ra tại 5 quận Los Angeles, San Diego, San Francisco, Orange và Alameda của bang California (CA) trong 3 năm liên tục 2017 – 2019.' The report interface includes standard navigation buttons (back, forward, search), a zoom control, and a large 'Print' button highlighted with a dashed border.

Bước 17: Chọn kiểu cấu hình phù hợp để in hợp lý và nhấn **Print**.

Print



We'll create a printer-friendly PDF version of your report.

Page size:

Tabloid (11" x 17")



Page orientation:

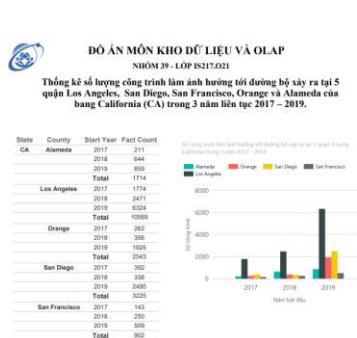
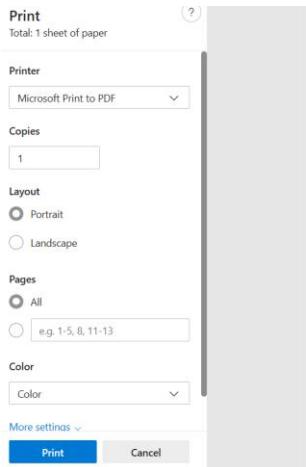
Portrait



Print

Cancel

Bước 18: Xem lại trước khi in



4.2.1.3. Kết quả báo cáo



ĐỒ ÁN MÔN KHO DỮ LIỆU VÀ OLAP

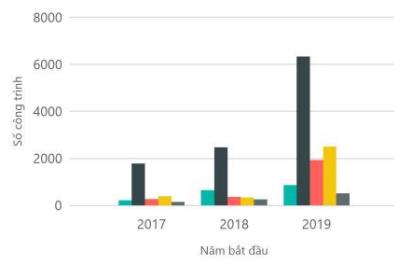
NHÓM 39 - LỚP IS217.O21

Thống kê số lượng công trình làm ảnh hưởng tới đường bộ xảy ra tại 5 quận Los Angeles, San Diego, San Francisco, Orange và Alameda của bang California (CA) trong 3 năm liên tục 2017 – 2019.

State	County	Start Year	Fact Count
CA	Alameda	2017	211
		2018	644
		2019	859
		Total	1714
Los Angeles	Los Angeles	2017	1774
		2018	2471
		2019	6324
		Total	10569
Orange	Orange	2017	262
		2018	356
		2019	1925
		Total	2543
San Diego	San Diego	2017	392
		2018	338
		2019	2495
		Total	3225
San Francisco	San Francisco	2017	143
		2018	250
		2019	509
		Total	902

Số công trình làm ảnh hưởng tới đường bộ xảy ra tại 5 quận ở bang California trong 3 năm 2017 - 2019

Alameda Orange San Diego San Francisco
Los Angeles

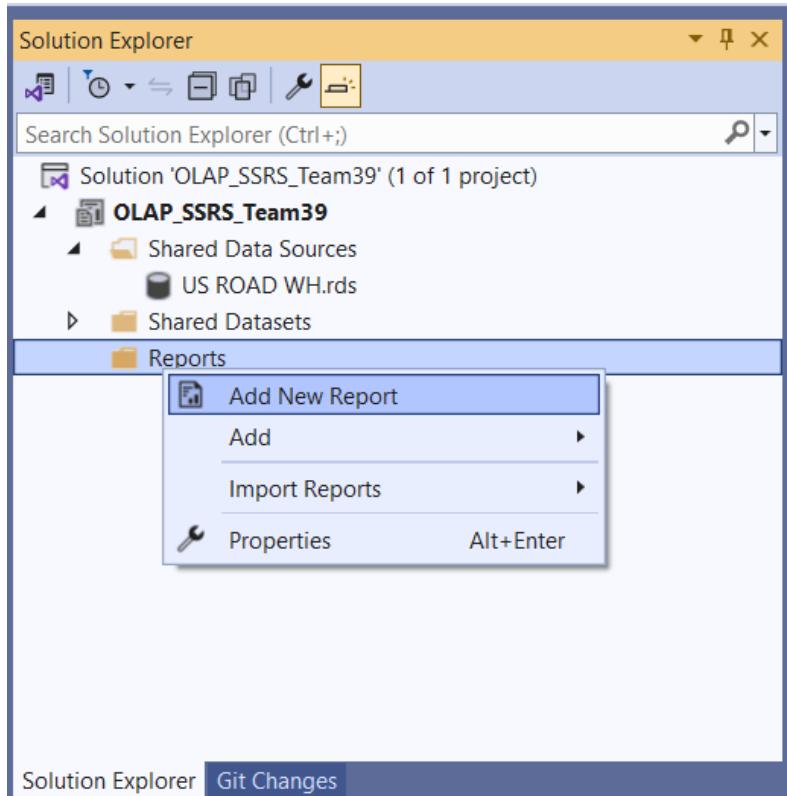


4.2.2. Báo cáo 2

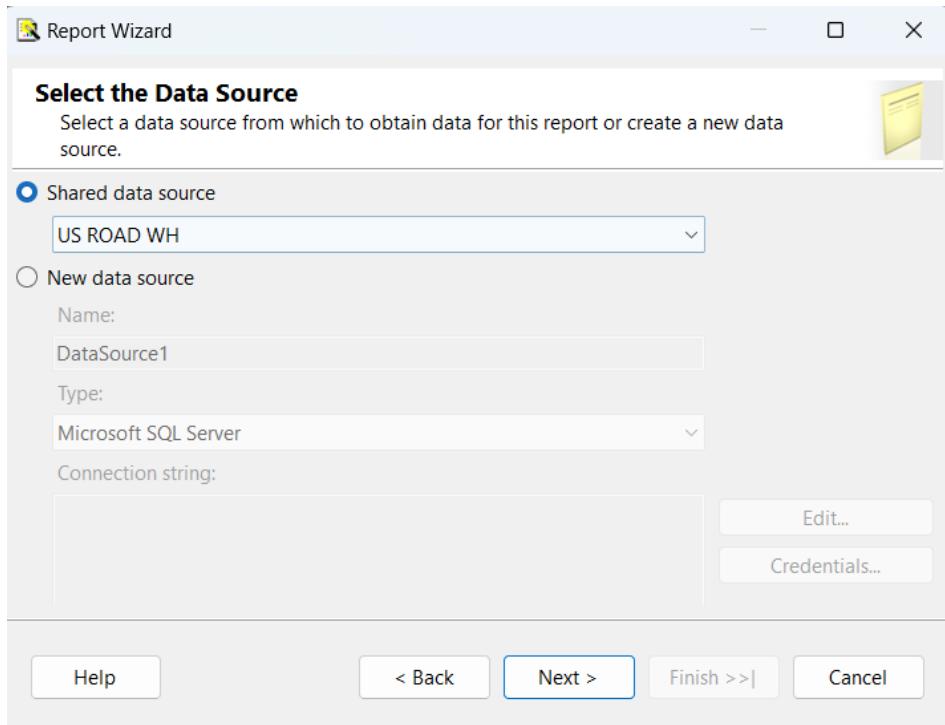
Nội dung câu truy vấn: Thống kê theo từng bang có số lượng các công trình thi công lớn 10000.

4.2.2.1. Tạo báo cáo

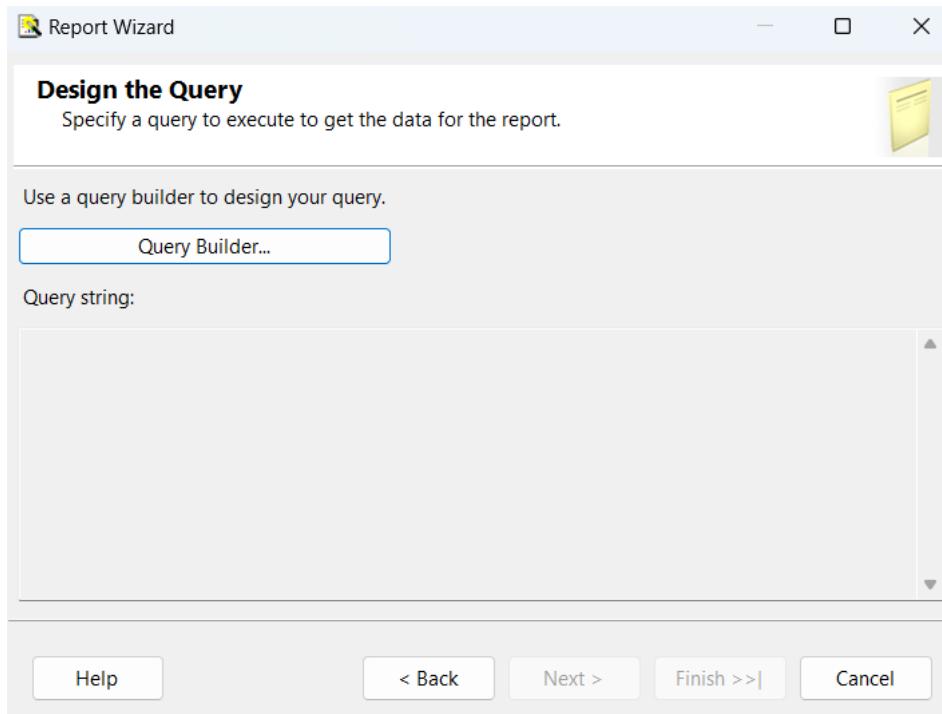
Bước 1: Quay trở lại project SSRS click chuột phải lên thư mục **Report** ở cửa sổ **Solution** chọn **Add New Report**



Bước 2: Cửa sổ **Report Wizard** hiện ra, tick chọn **Shared data source** và chọn **US ROAD** và nhấp **Next**



Bước 3: Sử dụng **Query Builder** để tạo query data dùng cho report. Click chọn **Query Builder**



Query Designer

Edit as Text Import... MD

Dimension Hierarchy Operator Filter Expression Param...

Dim Location Hierarchy In Fact Count > 10000

<Select dimension>

State	Fact Count
CA	30750
FL	12539
IL	11439
MI	10060
NY	21328
OH	16232
OR	10343
PA	66502
TX	14773

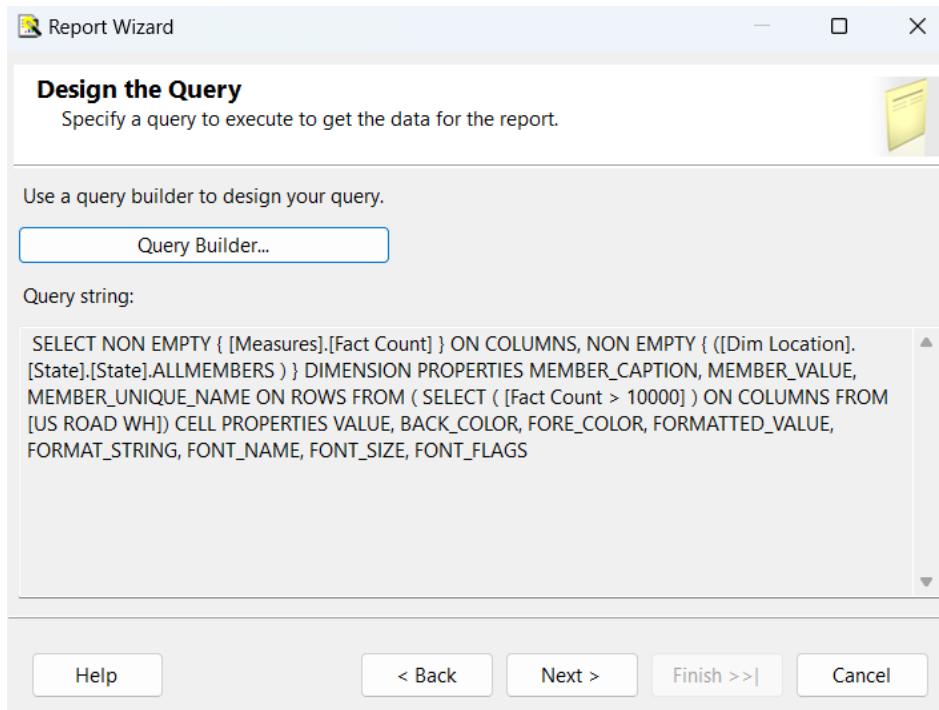
Calculated Members

Help OK Cancel

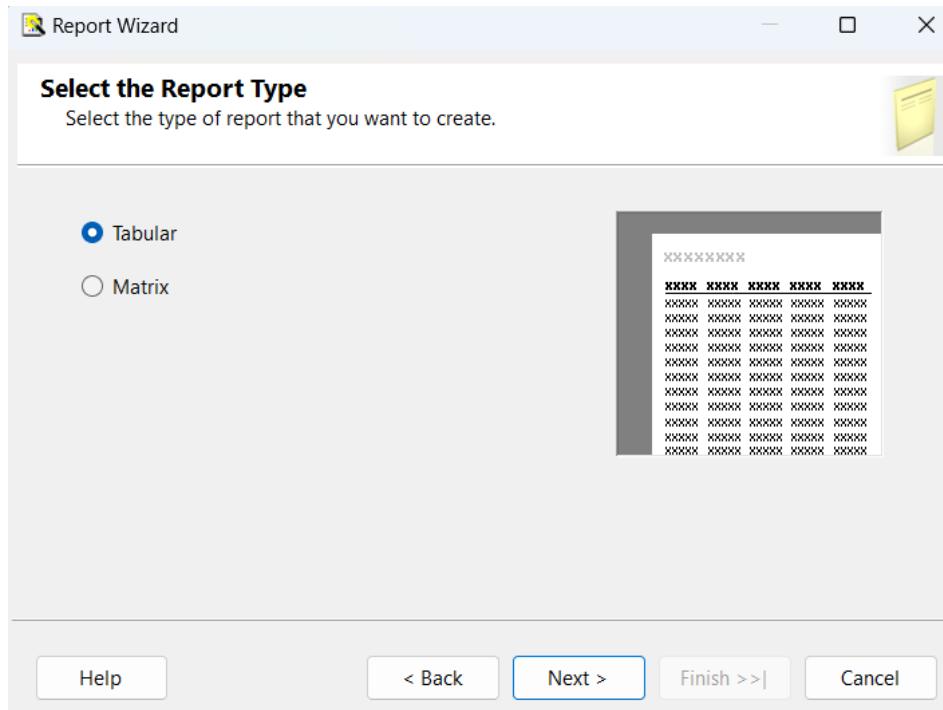
The screenshot shows the Microsoft Analysis Services Query Designer. On the left, there's a navigation pane with 'US ROAD WH' selected, showing 'Metadata', 'Search Model', 'Measure Group', and 'Fact' (with sub-options like Average Distance, Average Duration, etc.). Below that is 'KPIs' and 'Dim Location' with a specific filter applied: 'Fact Count > 10000'. Under 'Calculated Members', there is no content. The main area displays a table with 'State' and 'Fact Count' columns. The data is as follows:

State	Fact Count
CA	30750
FL	12539
IL	11439
MI	10060
NY	21328
OH	16232
OR	10343
PA	66502
TX	14773

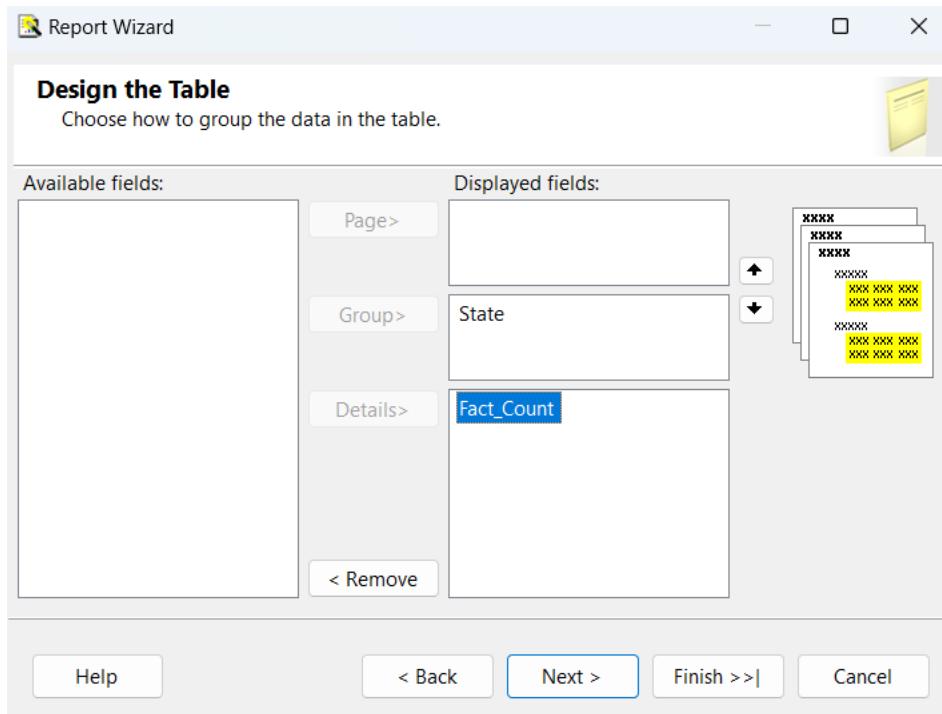
Bước 4: Chuỗi câu lệnh truy vấn được hiện ra như sau sau khi nhấn ok ở cửa sổ **Query Design** và nhấn Next



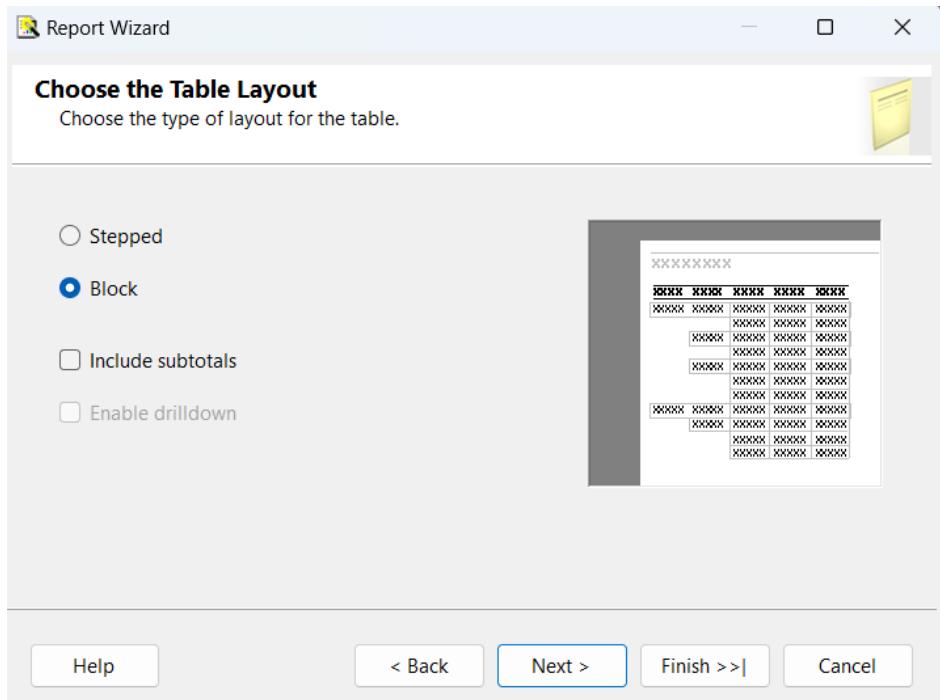
Bước 5: Chọn kiểu báo cáo (dạng bảng hoặc ma trận) sau đó nhấn **Next**



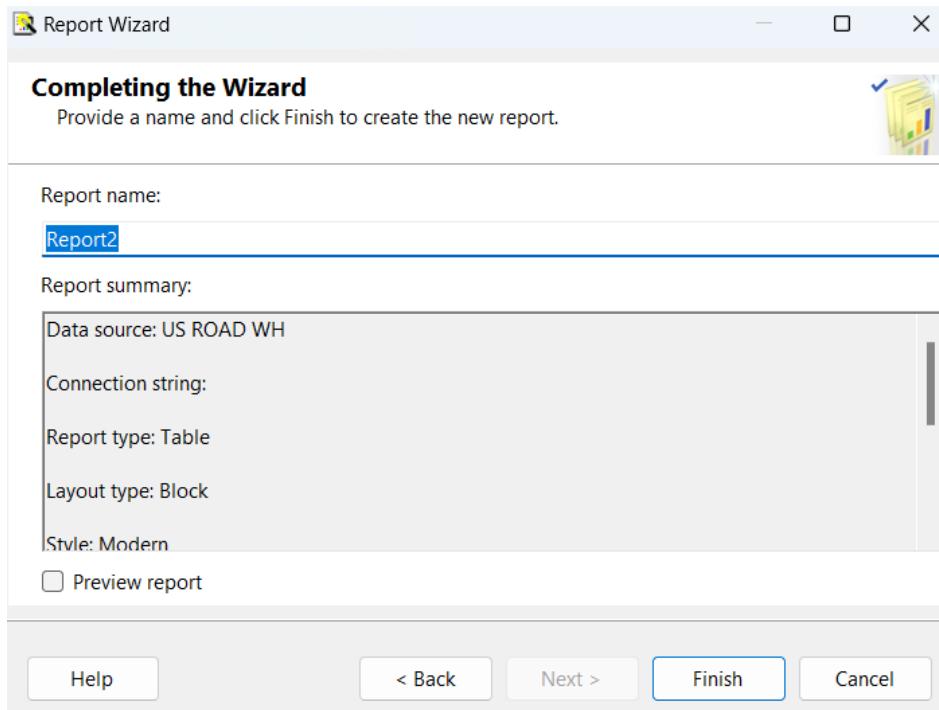
Bước 6: Đưa thuộc tính **State** vào **Group**, độ đo **Fact_Count** vào **Details** sau đó nhấn **Next**



Bước 7: Tick chọn **Block** để chọn layout cho báo cáo, sau đó nhấn **Next**



Bước 8: đặt tên cho **Report** và kiểm tra **Report summary** sau đó nhấn **Finish**



Bước 9: Sau khi nhấn **Finish**, sẽ được màn hình bên dưới, chuyển sang tab **Preview** để xem trước báo cáo

The screenshot shows the 'Report2.rdl [Design]' view. The top navigation bar has tabs for 'Design' (selected) and 'Preview'. The main area contains a table with the following data:

State	Fact Count
[State]	[Fact_Count]

Below the table, a text box displays the message: 'Thống kê theo từng bang có số lượng các công trình thi công lớn 10000.' A small chart icon is shown next to the text.

Bước 10: Click chuột phải vào **(table1_Details_Group)** trong nhóm **Row Groups**, chọn **Add Total** và chọn **After** để thêm tổng cho cột **Fact_Count**

The screenshot shows the Microsoft Power BI Report Designer interface. At the top, there's a title card with a logo, the text 'ĐỒ ÁN MÔN KHO DỮ LIỆU VÀ OLAP', and 'NHÓM 39 - LỚP IS217.O21'. Below this is a subtitle: 'Thống kê theo từng bang có số lượng các công trình thi công lớn 10000.' A table is displayed with two columns: 'State' and 'Fact Count'. The 'Fact Count' column contains the formula 'Sum(Fact_Count)'. In the bottom left, the 'Row Groups' pane is open, showing '(table1_State)' and '(table1_Details_Group)'. A context menu is open over the '(table1_Details_Group)' entry, with 'Add Group' expanded. Under 'Add Group', 'Add Total' is highlighted, and its submenu shows 'Before' and 'After' options, with 'After' being the selected choice.

Bước 11: chuyển sang tab **Preview** để xem trước báo cáo

Design Preview

1 of 1 | 100% Find | Next

ĐỒ ÁN MÔN KHO DỮ LIỆU VÀ OLAP

NHÓM 39 - LỚP IS217.O21

Thống kê theo từng bang có số lượng các công trình thi công lớn 10000.

State	Fact Count
CA	30750
FL	12539
IL	11439
MI	10060
NY	21328
OH	16232
OR	10343
PA	66502
TX	14773
Total	193966

Thống kê theo từng bang có số lượng các công trình thi công lớn 10000.

Fact Count

Tiểu bang	Số công trình
CA	30750
FL	12539
IL	11439
MI	10060
NY	21328
OH	16232
OR	10343
PA	66502
TX	14773

4.2.2.2. Xuất báo cáo

Bước 1: Chọn Build trên thanh menu của project SSRS -> Deploy Solution

The screenshot shows the Microsoft Visual Studio interface. The 'Build' menu is open, with 'Deploy Solution' highlighted. The background displays a report titled 'VÀ OLAP 021' containing a chart and a table.

Table:

State	Fact Count
CA	30750
FL	12539
IL	11439
MI	10060
NY	21328
OH	16232
OR	10343
PA	66502
TX	14773
Total	193966

Chart:

Thống kê theo từng bang có số lượng các công trình thi công lớn 10000.

The chart displays the 'Fact Count' (Số công trình) for each state. The Y-axis ranges from 0 to 80,000. The X-axis lists the states: CA, FL, IL, MI, NY, OH, OR, PA, TX. The data shows a general decline from CA to OR, followed by a sharp peak for PA (approx. 66,502).

Bước 2: Truy cập vào địa chỉ localhost/reports bằng trình duyệt web

The screenshot shows a web browser displaying the SQL Server Reporting Services Home page. The URL is 'localhost/reports/browse/'. The page includes a navigation bar with links like GitHub, Cổng thông tin đào..., Work, CODE, Money, Design, JAVA, DOAN, MonHoc, and Introduction. Below the navigation bar, there are sections for Favorites (Browse), Folders (3), Data Sources, Datasets, and OLAP_SSRS_Team39.

Ths. Đỗ Thị Minh Phụng

Bước 3: Mở thư mục **OLAP_SSRS_Team39** và mở Report **Report2**

The screenshot shows the 'Browse' section of the SQL Server Reporting Services interface. At the top, it says 'SQL Server Reporting Services Home > OLAP_SSRS_Team39'. Below that, under 'Favorites', is a link to 'Report2'. Under 'Paginated Reports (2)', there are two items: 'Report1' and 'Report2', each with a small icon.

Bước 4: Nhấn nút in để xuất ra file pdf báo cáo **report2**.

The screenshot shows the 'Report2' page. At the top, it says 'SQL Server Reporting Services Home > OLAP_SSRS_Team39 > Report2'. Below that is a navigation bar with various icons. The main content area displays a chart titled 'ĐỒ ÁN MÔN KHO DỮ LIỆU VÀ OLAP' and 'NHÓM 39 - LỚP IS217.O21'. The chart is a line graph titled 'Thống kê theo từng bang có số lượng các công trình thi công lớn 10000.' It shows 'Fact Count' on the Y-axis (0 to 80,000) and states on the X-axis: CA, FL, IL, MI, NY, OH, OR, PA. The data table below the chart is:

State	Fact Count
CA	30750
FL	12539
IL	11439
MI	10060
NY	21328
OH	16232
OR	10343
PA	66502
TX	14773
Total	193966

A 'Print' dialog box is overlaid on the right side of the chart. It contains the following text: 'We'll create a printer-friendly PDF version of your report.', 'Page size: Letter (8.5" x 11")', 'Page orientation: Portrait', and two buttons: 'Print' (red) and 'Cancel'.

4.2.2.3. Kết quả báo cáo



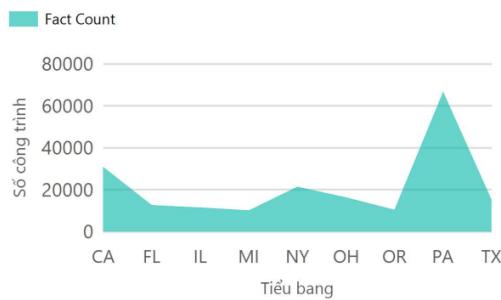
ĐỒ ÁN MÔN KHO DỮ LIỆU VÀ OLAP

NHÓM 39 - LỚP IS217.O21

**Thống kê theo từng bang có số lượng các công trình thi công
lớn 10000.**

State	Fact Count
CA	30750
FL	12539
IL	11439
MI	10060
NY	21328
OH	16232
OR	10343
PA	66502
TX	14773
Total	193966

Thống kê theo từng bang có số lượng các công trình thi công
lớn 10000.

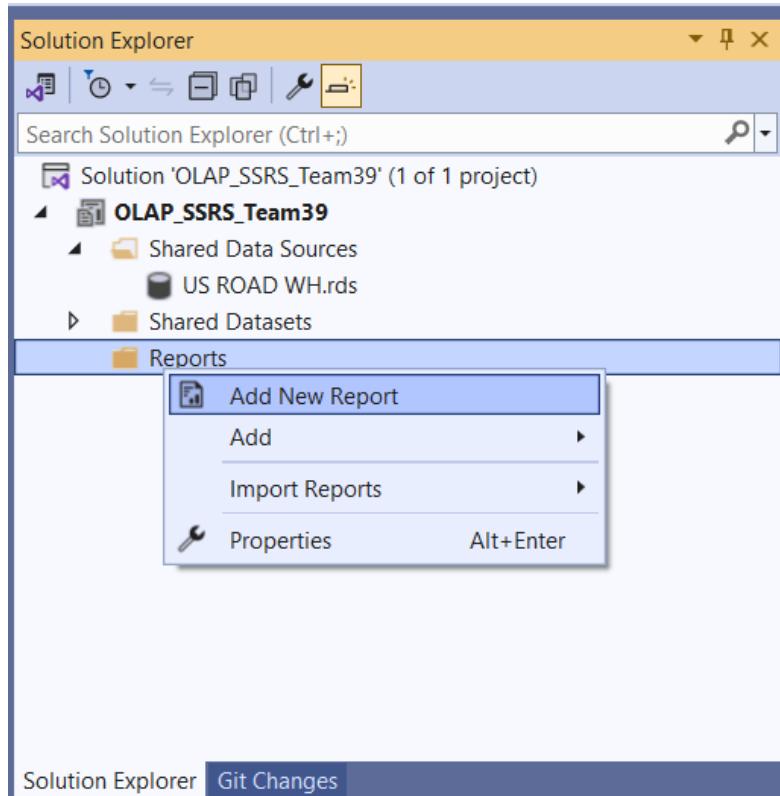


4.2.3. Báo cáo 3

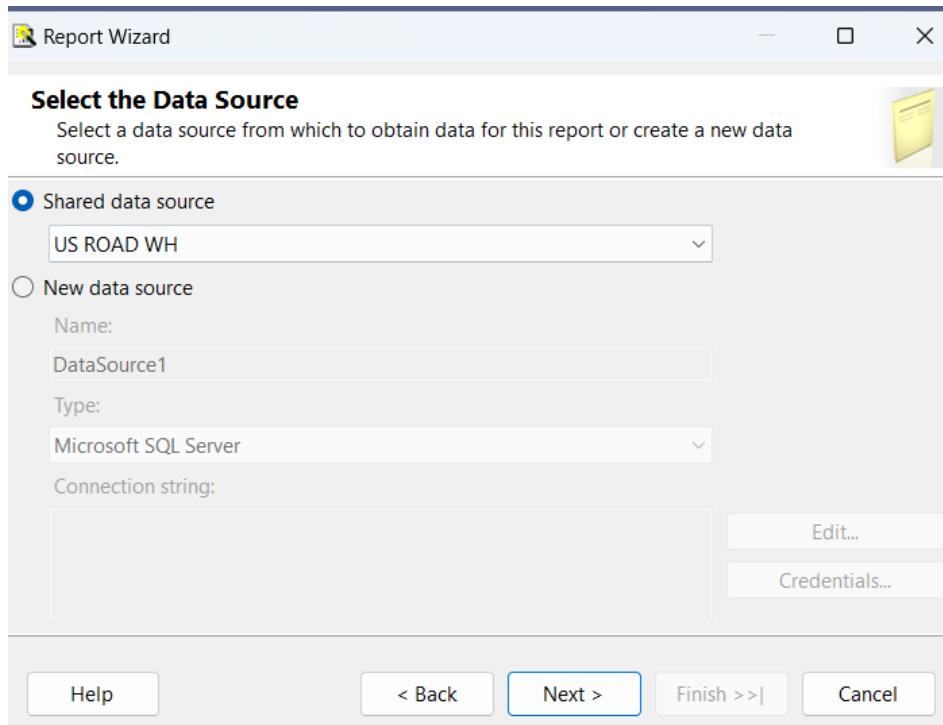
Nội dung câu truy vấn: Thống kê tổng số công trình (**Fact Count**) trên 6 tiểu bang (**State**): California (CA), New York (NY), Florida (FL), Ohio (OH), Texas (TX) và Washington (WA) có thời gian bắt đầu 2017 và kết thúc 2019.

4.2.3.1. Tạo báo cáo

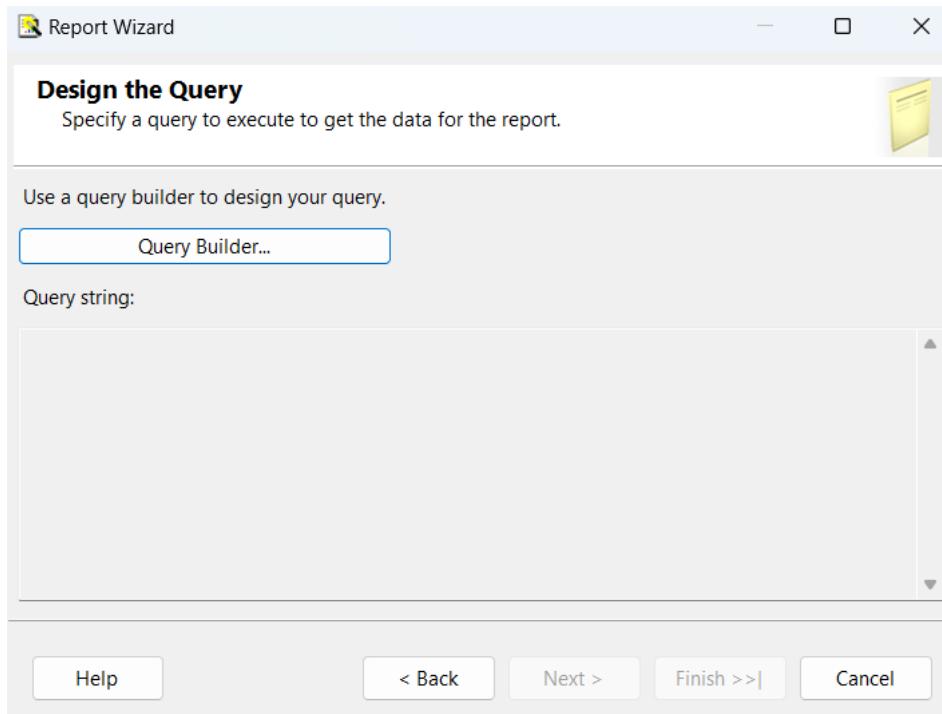
Bước 1: Quay trở lại dự án SSRS, Click chuột phải lên thư mục **Reports** chọn **Add New Report**

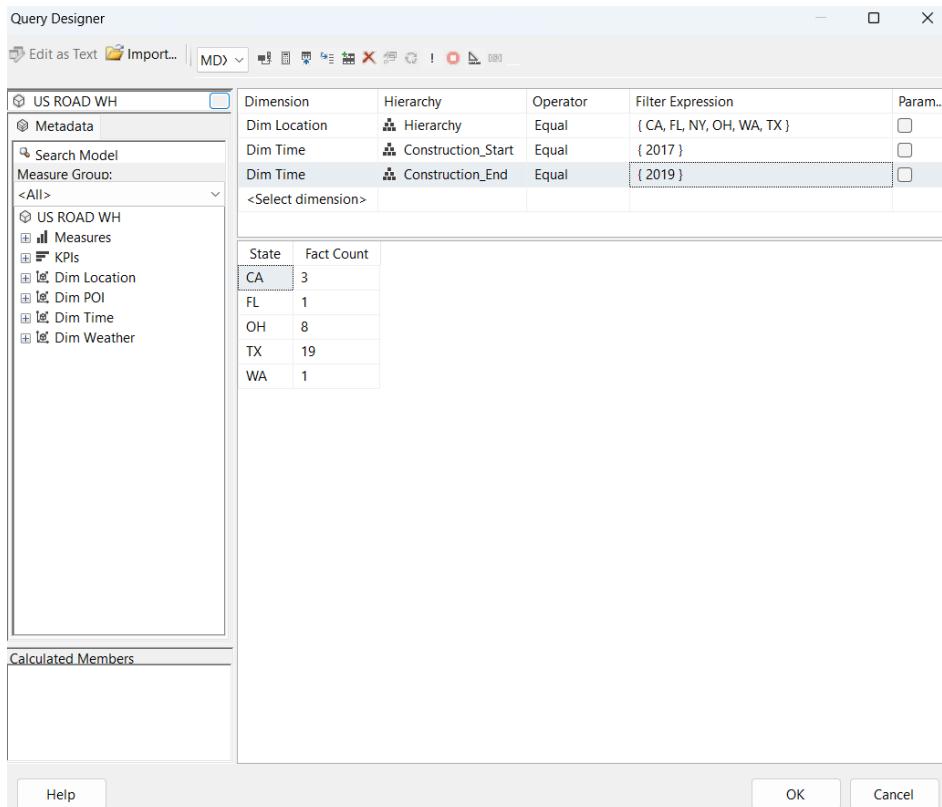


Bước 2: tick chọn **Shared data source** và chọn vào **US ROAD WH** và nhấn **Next**

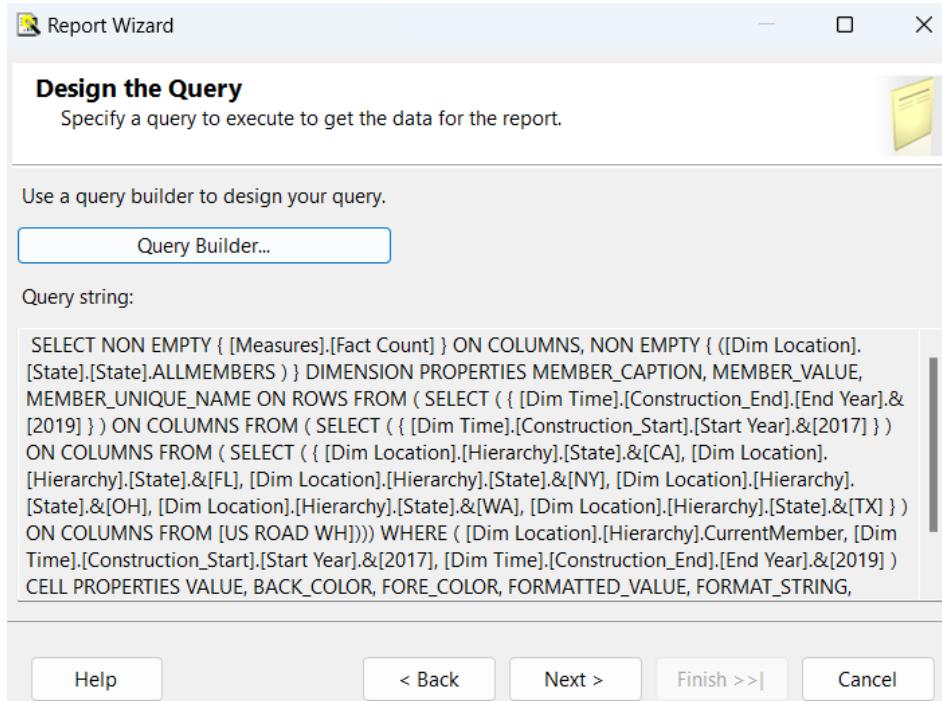


Bước 3: Sử dụng **Query Builder** để tạo query data dùng cho report. Click chọn **Query Builder**

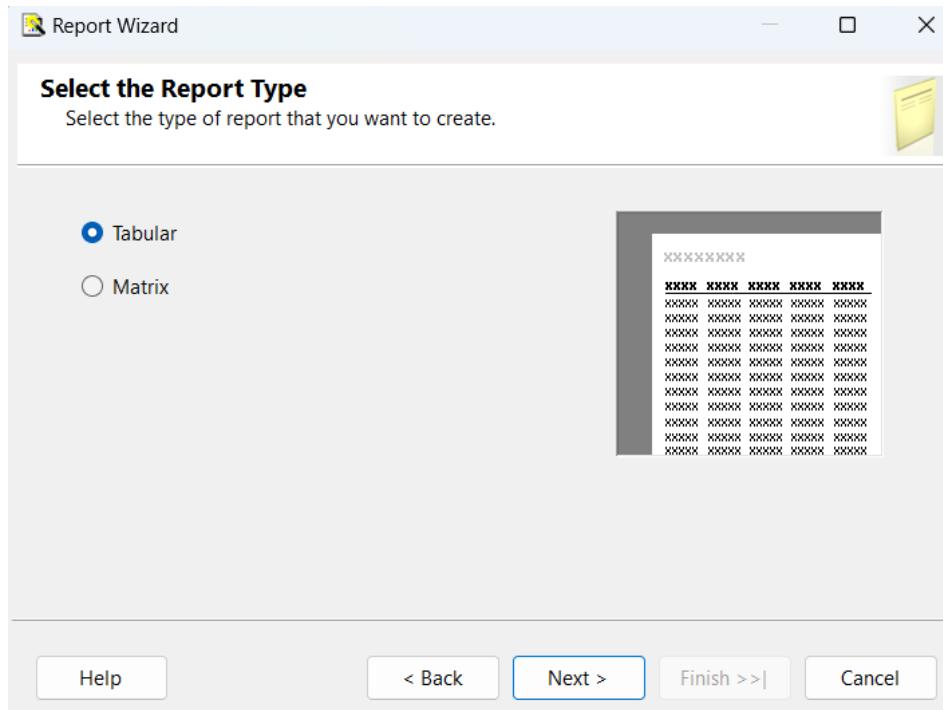




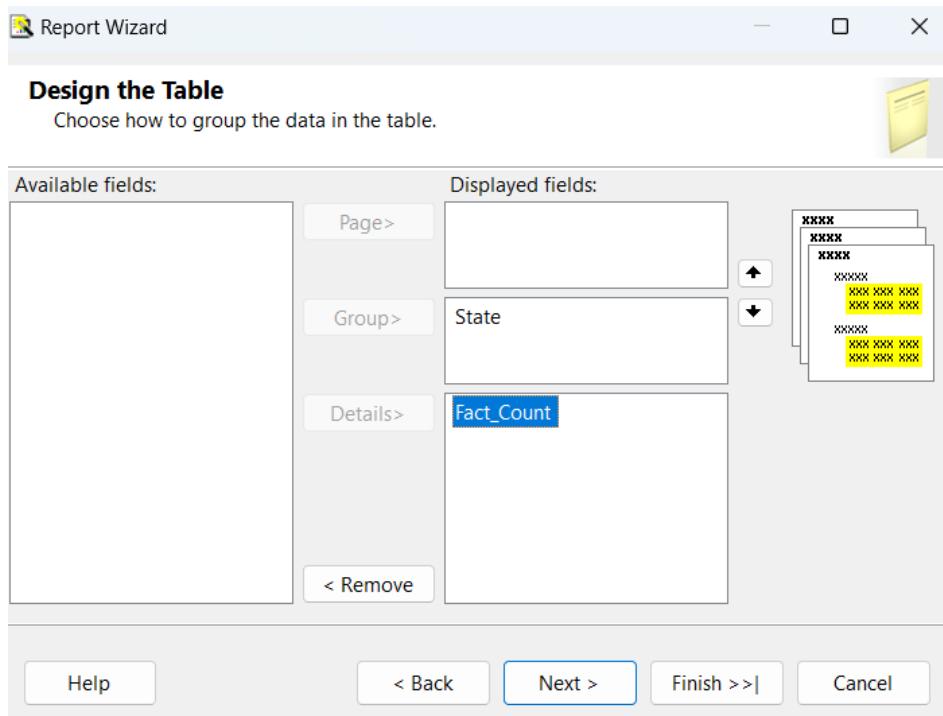
Bước 4: Chuỗi câu lệnh truy vấn được hiện ra như sau sau khi nhấn **OK** ở cửa sổ **Query Design** và nhấn **Next**



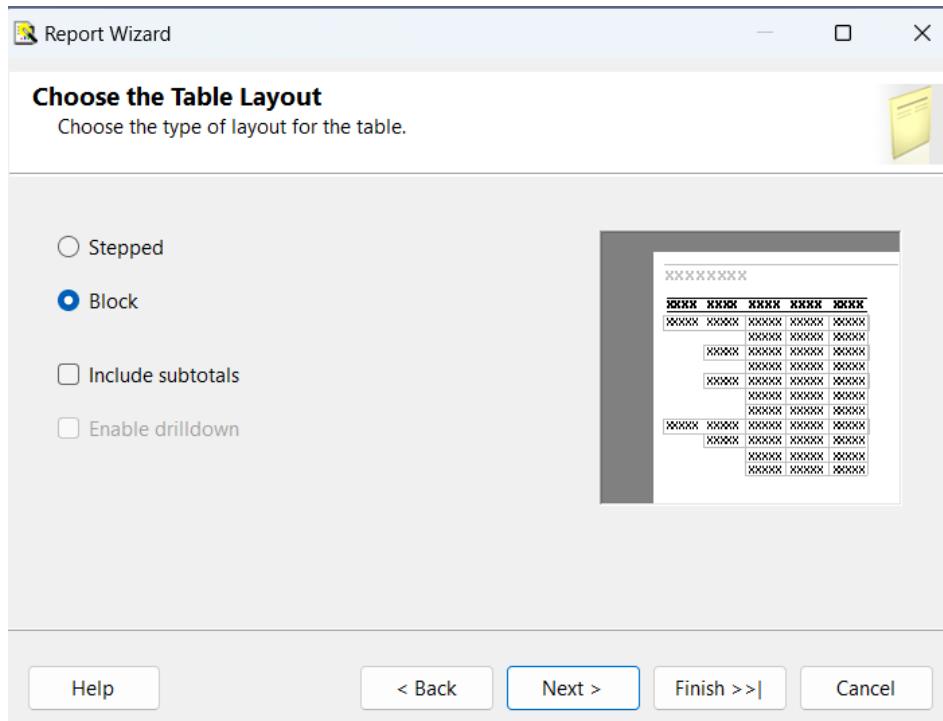
Bước 5: Chọn kiểu báo cáo (dạng bảng hoặc ma trận) sau đó nhấn Next



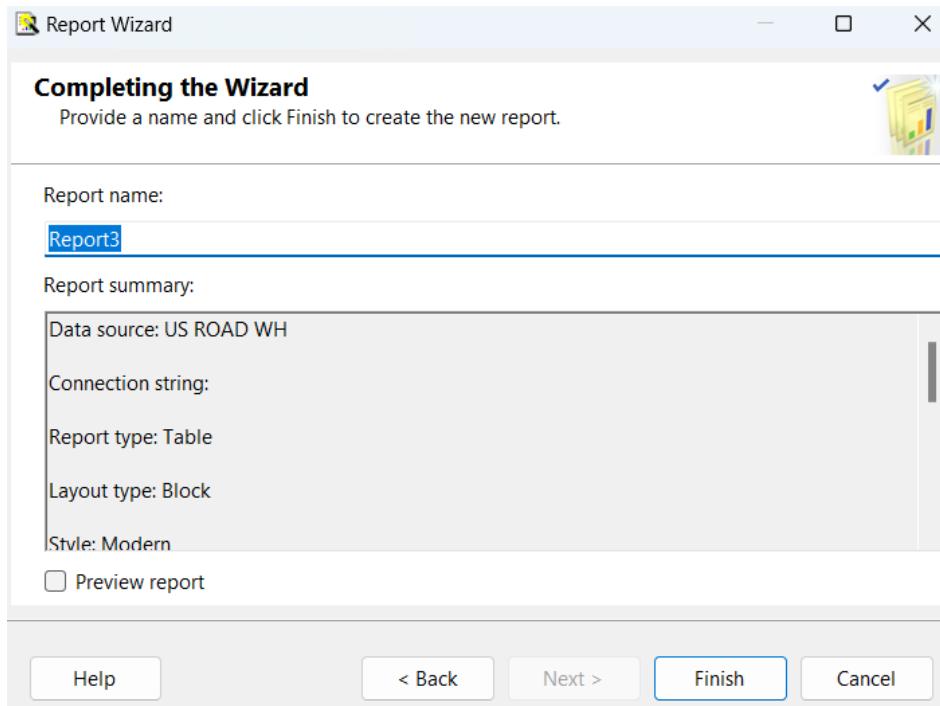
Bước 6: Đưa thuộc tính **State** vào **Group**, độ đo **Fact_Count** vào **Details** sau đó nhấn **Next**



Bước 7: Tick chọn Block để chọn layout cho báo cáo, sau đó nhấn Next



Bước 8: đặt tên cho **Report** và kiểm tra **Report summary** sau đó nhấn **Finish**



Bước 9: Sau khi nhấn **Finish**, sẽ được màn hình bên dưới, chuyển sang tab **Preview** để xem trước báo cáo

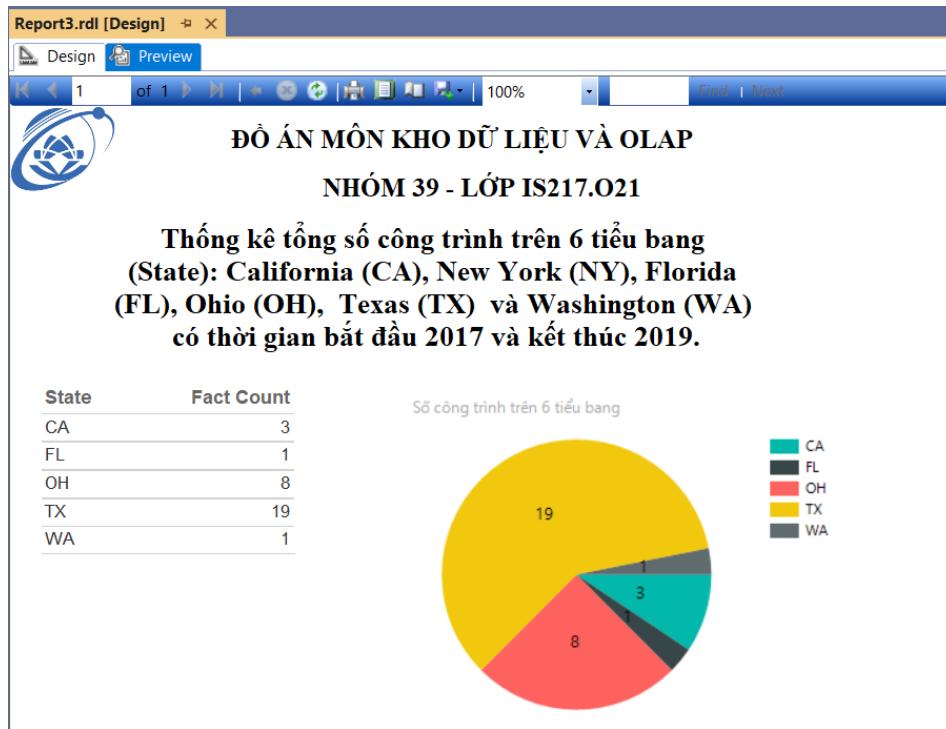
Ths. Đỗ Thị Minh Phụng

Report3.rdl [Design] * X

Design Preview

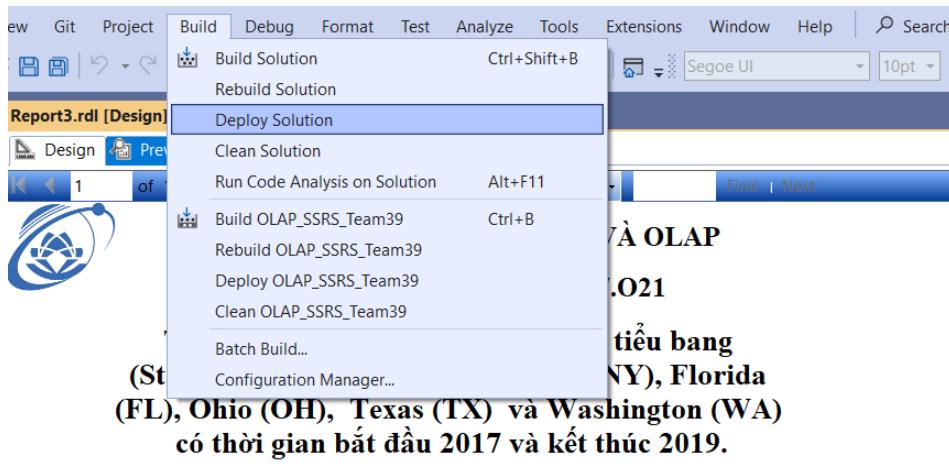
Thống kê tổng số công trình trên 6 tiêu bang

State	Fact Count
[State]	[Fact_Count]

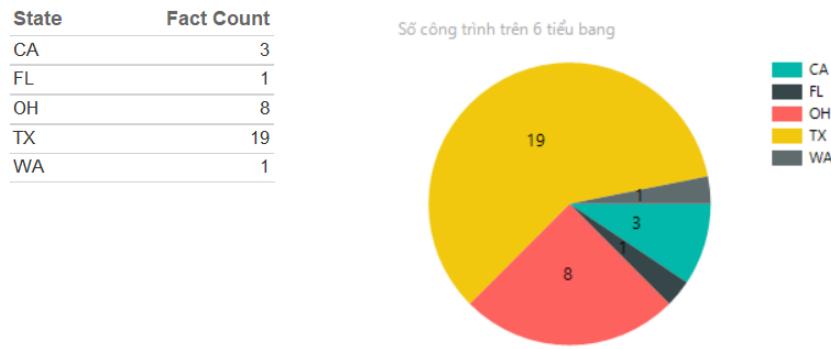


4.2.3.2. Xuất báo cáo

Bước 1: Quay trở lại Project SSRS Chọn **Build** trên thanh menu của project **SSRS** -> **Deploy Solution**



(St
(FL), Ohio (OH), Texas (TX) và Washington (WA)
có thời gian bắt đầu 2017 và kết thúc 2019.



Bước 2: Truy cập vào địa chỉ `localhost/reports` bằng trình duyệt web

Bước 3: Mở thư mục **OLAP_SSRS_Team39** và mở Report **Report3**

Ths. Đỗ Thị Minh Phụng

The screenshot shows the SQL Server Reporting Services interface. At the top, it says "SQL Server Reporting Services Home > OLAP_SSRS_Team39". Below that, there are tabs for "Favorites" and "Browse", with "Browse" being the active tab. Under "Paginated Reports (3)", there are three items: "Report1", "Report2", and "Report3", each with a preview icon.

Bước 4: Nhấn nút in để xuất ra file pdf báo cáo **report3**.

The screenshot shows the details of "Report3". The title is "ĐỒ ÁN MÔN KHO DỮ LIỆU VÀ OLAP" and subtitle is "NHÓM 39 - LỚP IS217.O21". It displays a table of "Thống kê tổng số công trình trên 6 tiểu bang (State)" and a pie chart titled "Số công trình trên 6 tiểu bang". The table data is as follows:

State	Fact Count
CA	3
FL	1
OH	8
TX	19
WA	1

The pie chart shows the distribution of projects across six states: Texas (19), California (3), Florida (1), Ohio (8), Washington (1), and another state (3). A print dialog box is overlaid on the right side of the screen, prompting the user to create a printer-friendly PDF version of the report. It includes fields for "Page size" (set to "Letter (8.5" x 11")") and "Page orientation" (set to "Portrait"). There are "Print" and "Cancel" buttons at the bottom of the dialog.

Bước 5: Chọn kiểu cấu hình phù hợp để in hợp lý và nhấn **Print**.

Ths. Đỗ Thị Minh Phụng

Print

Total: 1 sheet of paper

Printer

Microsoft Print to PDF

Copies

1

Layout

Portrait

Landscape

Pages

All

e.g. 1-5, 8, 11-13

Color

Color

More settings ▼

Print **Cancel**

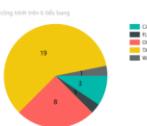
BỘ ÁN MÔN KHO DỮ LIỆU VÀ OLAP

NHÓM 39 - LỚP IS217.O21

Thống kê tổng số công trình trên 6 tiểu bang (State): California (CA), New York (NY), Florida (FL), Ohio (OH), Texas (TX) và Washington (WA) có thời gian bắt đầu 2017 và kết thúc 2019.

State	Fact Count
CA	3
FL	1
OH	8
TX	19
WA	1

Tổng cộng có 34 công trình.



4.2.3.3. Kết quả báo cáo



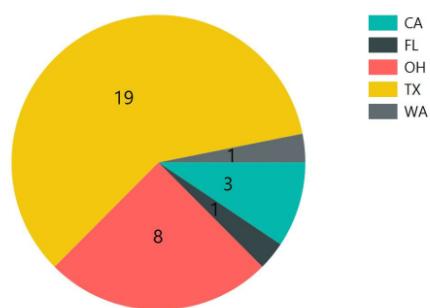
ĐỒ ÁN MÔN KHO DỮ LIỆU VÀ OLAP

NHÓM 39 - LỚP IS217.O21

Thông kê tổng số công trình trên 6 tiểu bang (State): California (CA), New York (NY), Florida (FL), Ohio (OH), Texas (TX) và Washington (WA) có thời gian bắt đầu 2017 và kết thúc 2019.

State	Fact Count
CA	3
FL	1
OH	8
TX	19
WA	1

Số công trình trên 6 tiểu bang

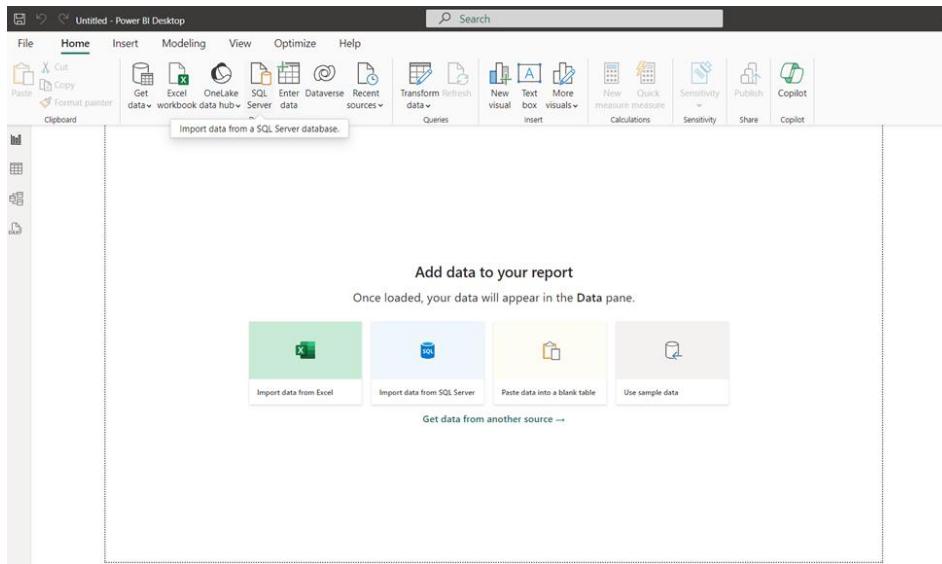


4.3. Tạo báo cáo bằng POWER BI

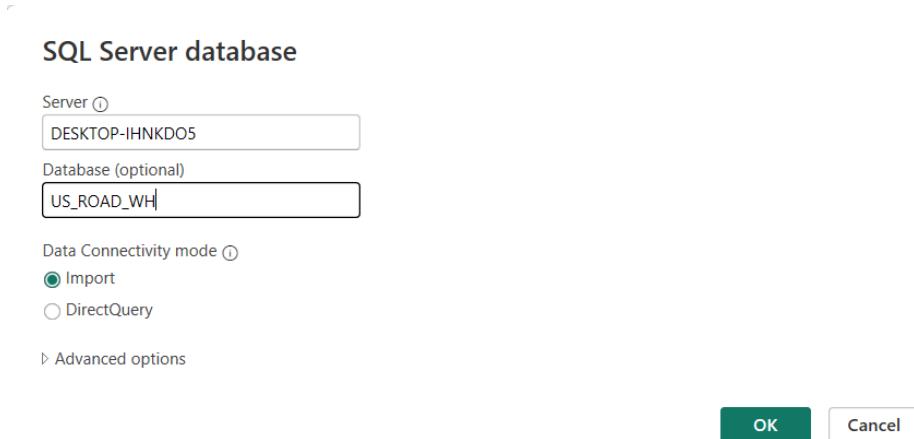
4.3.1. Kết nối Power BI với SQL Server

Bước 1: Trong Power BI. Tại thẻ **Home** trong nhóm **Data**, chọn **SQL Server**.

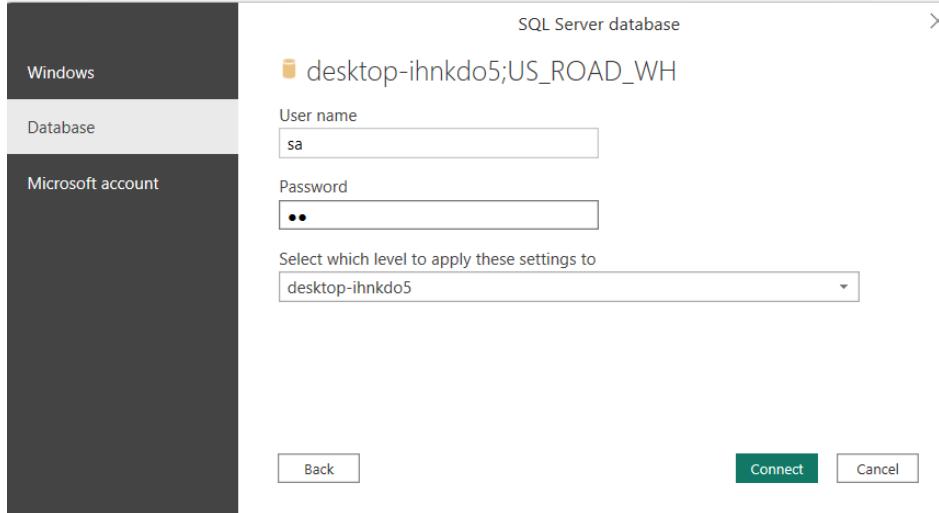
Ths. Đỗ Thị Minh Phụng



Bước 2: Nhập tên Server và tên Database và nhấn OK



Bước 3: Chọn tab Database ở cửa sổ tiếp thiêu và nhấn Connect



Bước 4: Chọn bảng **Fact** và các bảng **Dim** và chờ **Power BI** load dữ liệu

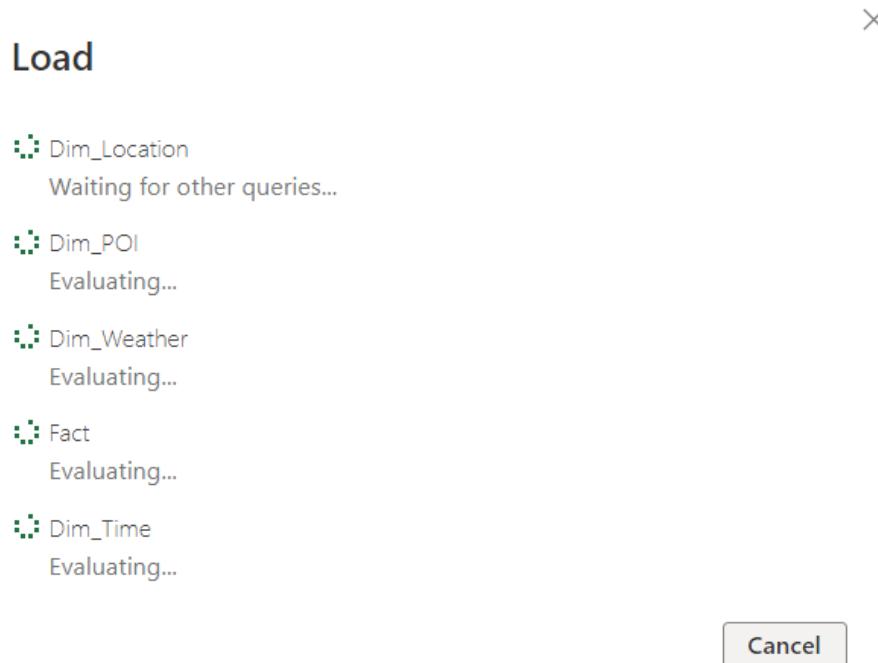
Navigator

The screenshot shows the Navigator interface with the following details:

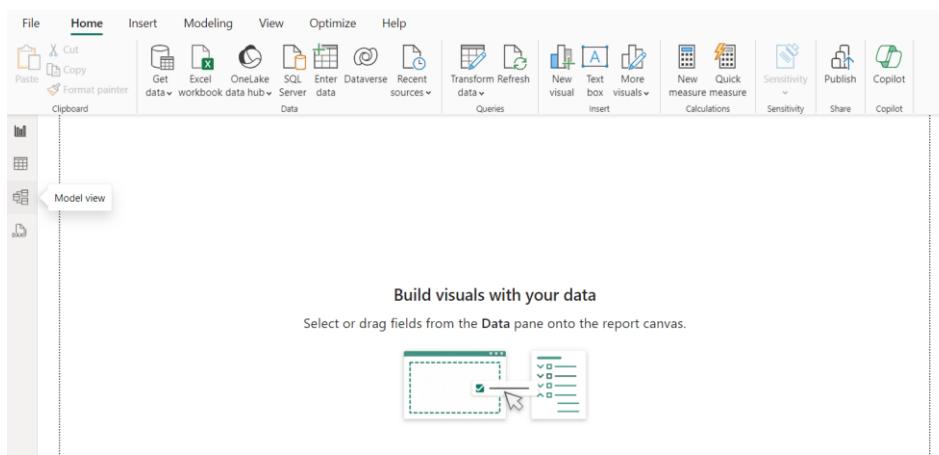
- Display Options** dropdown menu is open, showing the current connection: DESKTOP-IHNKDO5: US_ROAD_WH [8].
- Dim_Time** table is selected and highlighted.
- Table Structure:**

Time_ID	Start_day	Start_month	Start_year	Start_hour
1/1/2017 1:01:04 AM	1	1	2017	
1/1/2017 2:01:42 AM	1	1	2017	
1/1/2017 3:01:00 AM	1	1	2017	
1/1/2017 6:00:00 AM	1	1	2017	
1/1/2017 6:00:33 AM	1	1	2017	
1/1/2017 6:42:30 AM	1	1	2017	
1/1/2017 8:42:28 AM	1	1	2017	
1/1/2017 8:42:29 AM	1	1	2017	
1/1/2017 11:45:39 AM	1	1	2017	
1/1/2017 11:45:48 AM	1	1	2017	
1/1/2017 2:00:29 PM	1	1	2017	
1/1/2017 7:15:28 PM	1	1	2017	
1/1/2017 7:51:37 PM	1	1	2017	
1/1/2017 8:00:00 PM	1	1	2017	
1/1/2017 11:46:25 PM	1	1	2017	
1/2/2017 12:02:36 AM	2	1	2017	
1/2/2017 1:52:13 AM	2	1	2017	
1/2/2017 2:10:58 AM	2	1	2017	
1/2/2017 4:14:19 AM	2	1	2017	
1/2/2017 5:06:28 AM	2	1	2017	
1/2/2017 5:08:53 AM	2	1	2017	
1/2/2017 5:55:29 AM	2	1	2017	
1/2/2017 6:00:14 AM	2	1	2017	
- Buttons:** Select Related Tables, Load, Transform Data, Cancel.

Ths. Đỗ Thị Minh Phụng



Bước 5: Sau khi quá trình load dữ liệu hoàn tất, chọn tab **Model View** ở thanh bên trái.
Quan hệ giữa các bảng sẽ được hiện.





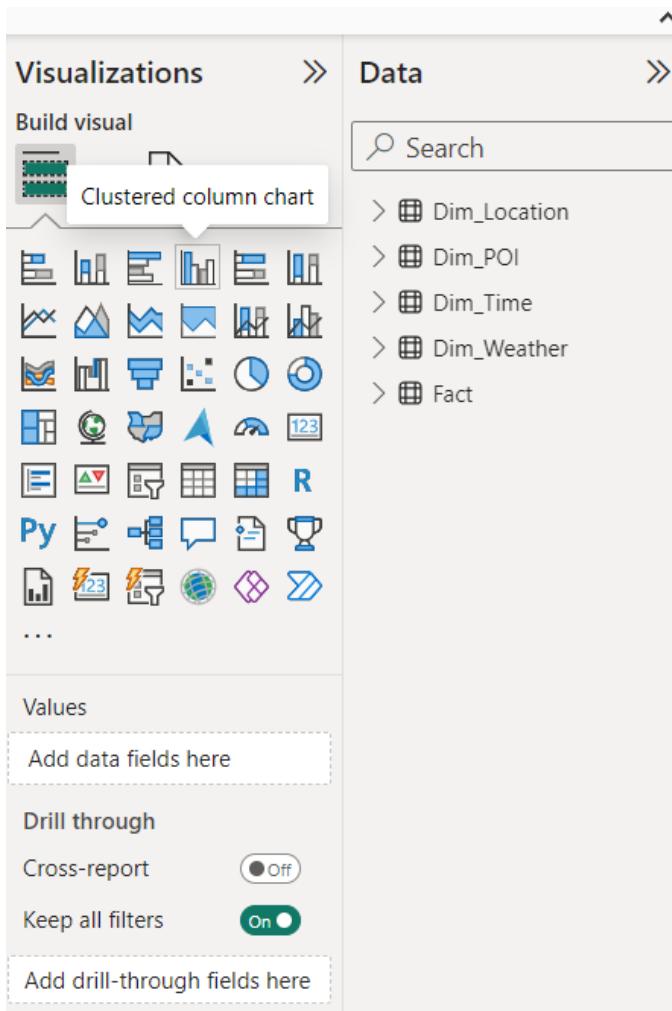
4.3.2. Thực hiện Report bằng Power BI

4.3.2.1. Báo cáo 1

Nội dung câu truy vấn: Thống kê số lượng công trình làm ảnh hưởng tới đường bộ xảy ra tại 5 quận Los Angeles, San Diego, San Francisco, Orange và Alameda của bang California (CA) trong 3 năm liên tục 2017 – 2019.

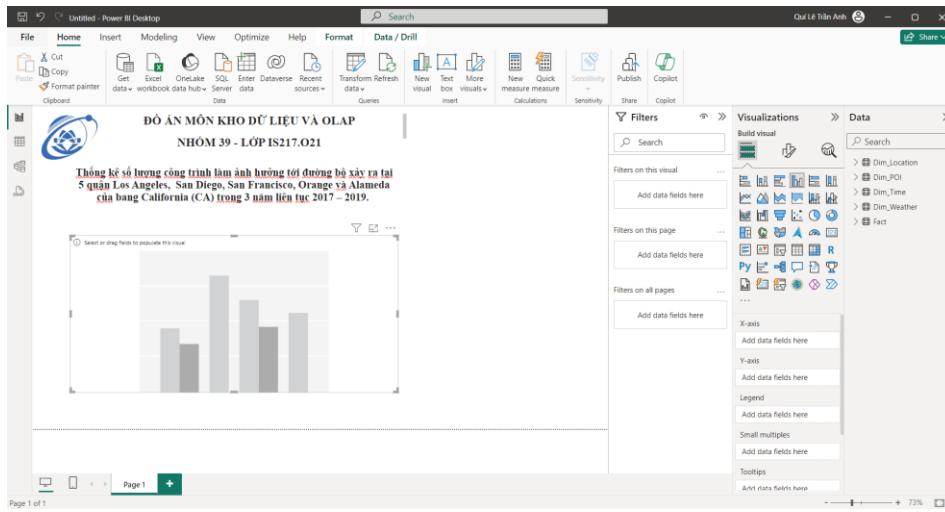
Thực hiện report:

Bước 1: Hiển thị kết quả dạng **Chart**, chọn dạng **Clustered column chart** Trong tab **Visualizations**



Bước 2: Tạo mẫu report

Ths. Đỗ Thị Minh Phụng



Bước 3: Điều chỉnh X-axis, Y-axis, Legend như hình bên dưới

Filters

Search:

Filters on this visual:

- Count of Fact_ID is (All)
- County**: is San Diego, Los Ang...
- Filter type: Basic filtering

 - Select all
 - Acadia (9)
 - Accomack (1)
 - Ada (44)
 - Adair (19)
 - Adams (202)
 - Addison (6)

- Require single selection

Start_year is (All)

Add data fields here

Visualizations

Build visual:

X-axis: Start_year

Y-axis: Count of Fact_ID

Legend: County

Small multiples: Add data fields here

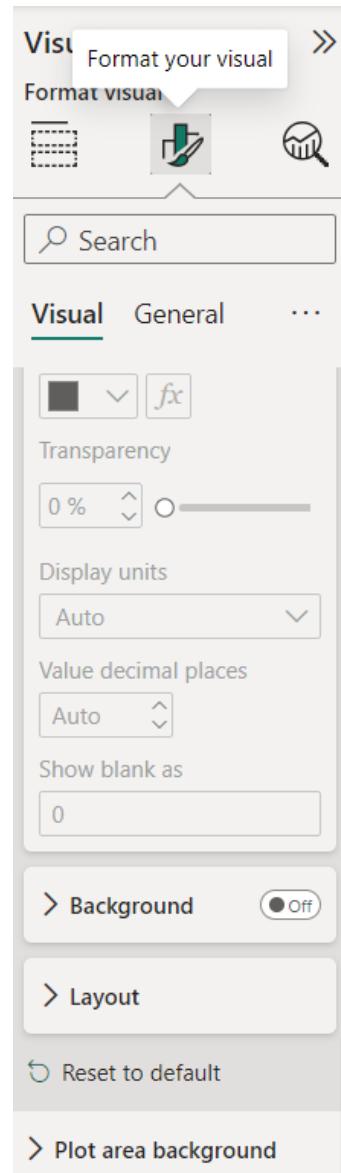
Toolips: Add data fields here

Data

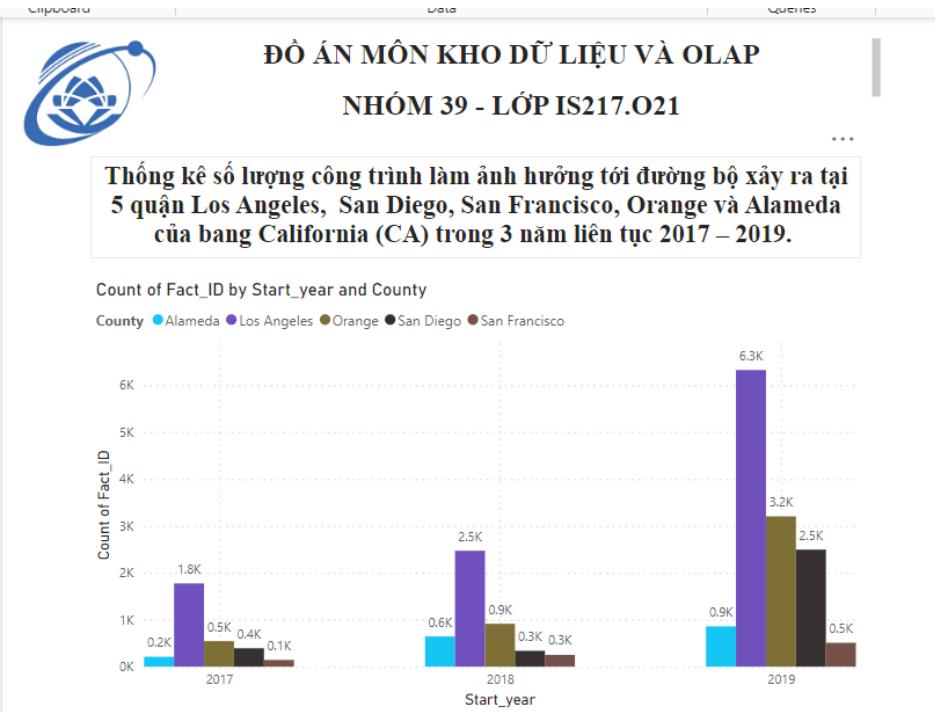
Search:

- Dim_Location** (selected):
 - City
 - County (checked)
 - Location_ID
 - Side
 - State
 - Street
- Dim_POI**
- Dim_Time** (selected):
 - Σ End_day
 - Σ End_hour
 - Σ End_minute
 - Σ End_month
 - Σ End_year
 - Σ Start_day
 - Σ Start_hour
 - Σ Start_minute
 - Σ Start_month
 - Σ Start_year (checked)
- Time_ID**
- Dim_Weather**
- Fact** (selected):
 - Σ Distance mi

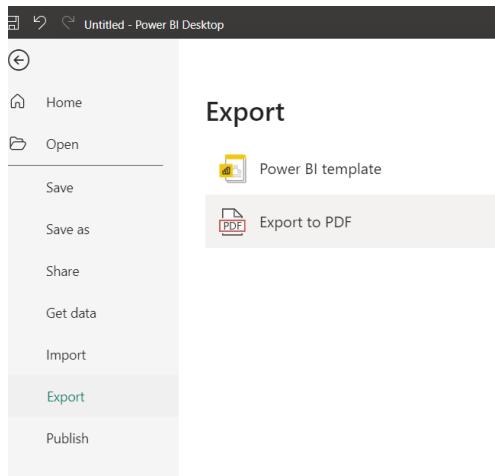
Bước 4: Chuyển sang tab **Format your visual**



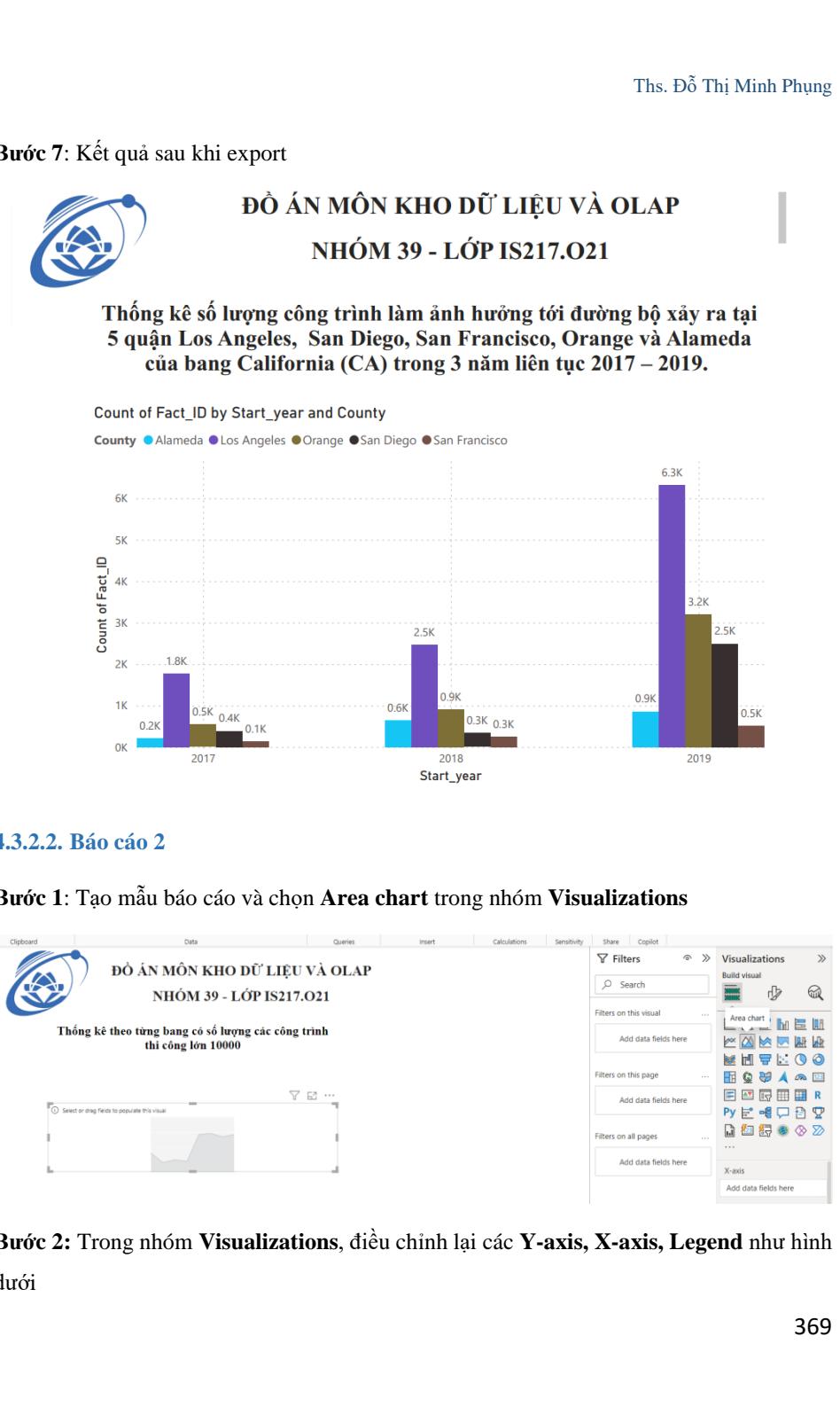
Bước 5: Tổng hợp kết quả report



Bước 6: Export report



Bước 7: Kết quả sau khi export



4.3.2.2. Báo cáo 2

Bước 1: Tạo mẫu báo cáo và chọn Area chart trong nhóm Visualizations



Bước 2: Trong nhóm Visualizations, điều chỉnh lại các Y-axis, X-axis, Legend như hình dưới

The screenshot shows the Power BI 'Visualizations' pane on the left and the 'Data' pane on the right. In the 'Visualizations' pane, a 'Scatter chart' icon is highlighted. The 'Data' pane displays a search bar and two main sections: 'Dim_Location' and 'Fact'. Under 'Dim_Location', 'State' is selected. Under 'Fact', 'Fact_ID' is selected. The 'X-axis' section shows 'State' assigned, and the 'Y-axis' section shows 'Count of Fact_ID' assigned.

Visualizations > Data

Build visual

Scatter chart

X-axis: State

Y-axis: Count of Fact_ID

Dim_Location

- City
- County
- Location_ID
- Side
- State
- Street

Dim_POI

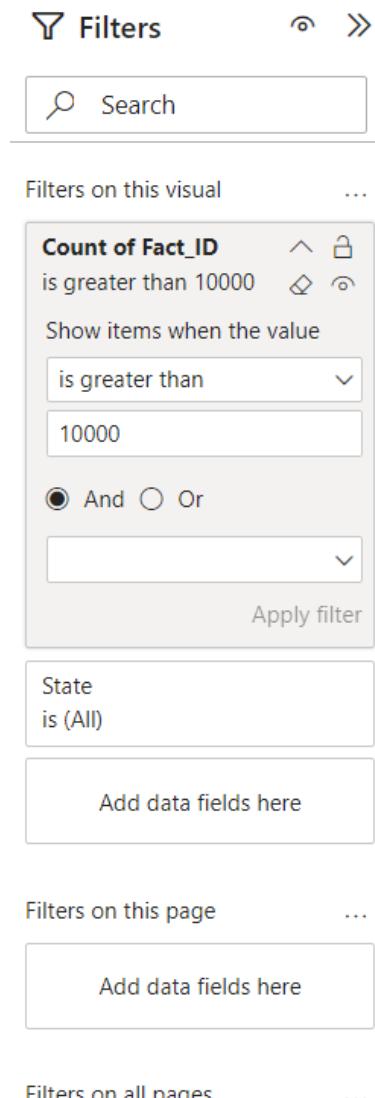
Dim_Time

Dim_Weather

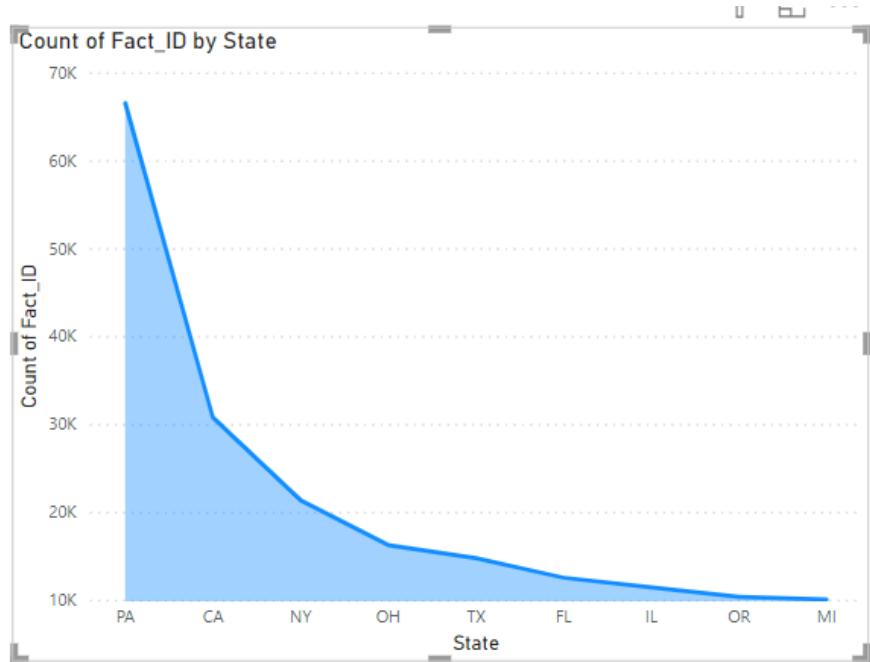
Fact

- \sum Distance_mi
- \sum Duration
- Fact_ID
- Location_ID
- POI_ID
- \sum Severity
- Time_ID
- Weather_ID

Bước 3: Trong nhóm Filters, thực hiện việc lọc như hình bên dưới và nhấn **Apply**

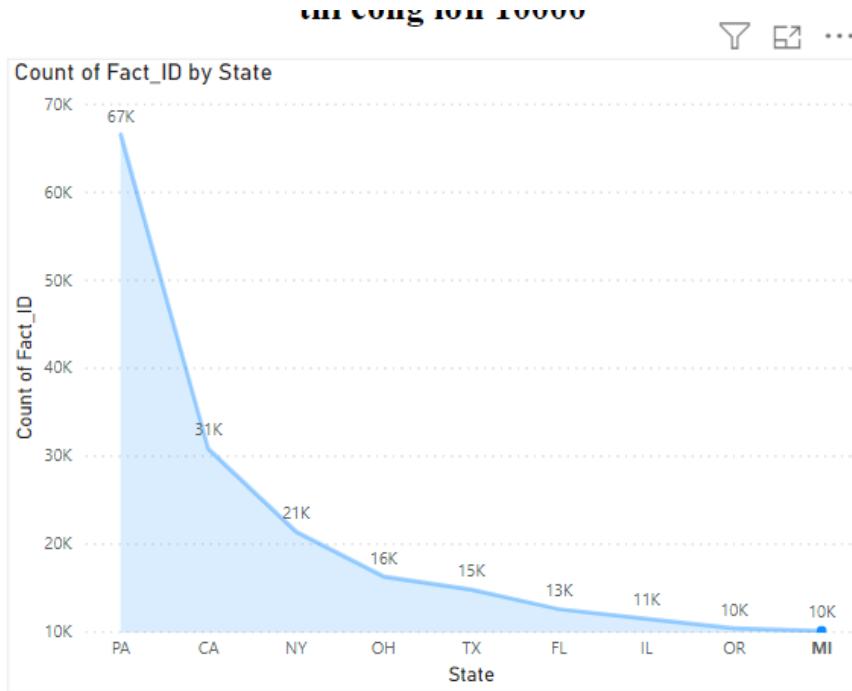


Bước 4: Xem kết quả dạng Area

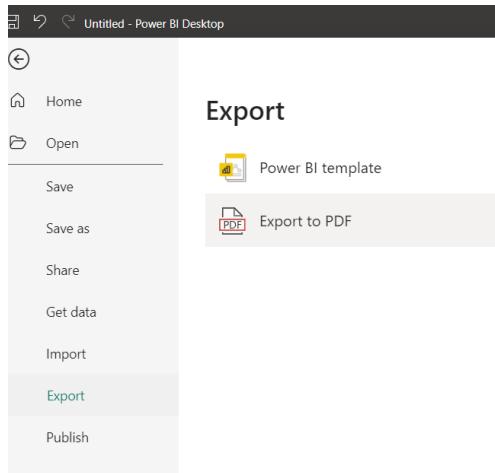


Bước 5: Bấm vào Format your visual:

Ths. Đỗ Thị Minh Phụng



Bước 6: Xuất báo cáo



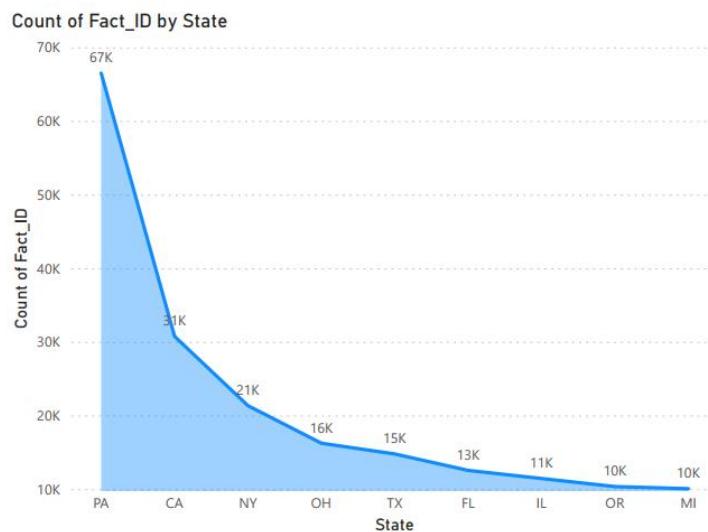
Bước 7: Kết quả sau khi xuất báo cáo



ĐỒ ÁN MÔN KHO DỮ LIỆU VÀ OLAP

NHÓM 39 - LỚP IS217.O21

Thống kê theo từng bang có số lượng các công trình thi công lớn 10000



4.3.2.3. Báo cáo 3

Bước 1: Tạo mẫu báo cáo và chọn **Table** trong nhóm **Visualization**

The screenshot shows the Power BI desktop interface. On the left, there is a visual placeholder with the text "Select or drag fields to populate this visual". On the right, the "Visualizations" pane is open, displaying various chart and table options. The "Filters" pane is also visible, showing filter settings for the current visual.

Bước 2: từ tab Data, kéo State của bảng Dim_Location vào Columns trong tab Visualization. Kéo Fact_ID từ bảng Fact vào Columns

The screenshot shows the Power BI Data view interface. On the left, there's a sidebar with 'Visualizations' and 'Build visual' sections, followed by a grid of visualization icons. Below that are sections for 'Columns' (with 'State' and 'Count of Fact_ID' selected), 'Drill through' (with 'Cross-report off' and 'Keep all filters On'), and a button 'Add drill-through fields here'. On the right, under 'Data', there's a search bar and a tree view of data models. The 'Dim_Location' node is expanded, showing fields: City, County, Location_ID, Side, State (which is checked), and Street. The 'Fact' node is also expanded, showing fields: \sum Distance_mi, \sum Duration, Fact_ID (which is checked), Location_ID, POI_ID, \sum Severity, Time_ID, and Weather_ID.

Bước 3: Trong nhóm Filter, chọn điều kiện lọc như hình bên dưới

The screenshot shows the Power BI interface with the 'Visualizations' pane on the left and the 'Data' pane on the right.

Visualizations pane:

- Build visual: Includes icons for various chart types like bar, line, pie, etc.
- Columns section: Shows 'State' and 'Count of Fact_ID' selected.
- Drill through section: Shows 'Cross-report' off and 'Keep all filters' on.
- Add drill-through fields here: A button.

Data pane:

- Search bar: 'Search'.
- Dim_Location: Contains fields: City, County, Location_ID, Side, State (selected), Street.
- Dim_POI: A folder icon.
- Dim_Time: A folder icon.
- Dim_Weather: A folder icon.
- Fact: Contains fields: \sum Distance_mi, \sum Duration, Fact_ID (selected), Location_ID, POI_ID, \sum Severity, Time_ID, Weather_ID.

Bước 4: Xem kết quả ở dạng Tables

State	Count of Fact_ID
CA	3
FL	1
OH	8
TX	19
WA	1
Total	32

Bước 5: Từ nhóm **Visualization**, chọn **Pie chart**

The screenshot shows the Power BI interface with the 'Visualizations' pane on the left and the 'Data' pane on the right.

Visualizations pane:

- Build visual section: Includes icons for Bar chart, Line chart, Map, Pie chart, Stacked bar chart, Stacked line chart, Gantt chart, Timeline, Heatmap, Scatter plot, Radar chart, Treemap, Pyramid chart, and Gauge.
- Values section: A dashed box labeled 'Add data fields here'.
- Drill through section: Includes 'Cross-report' (Off) and 'Keep all filters' (On).
- Add drill-through fields here: A dashed box.

Data pane:

- Search bar: 'Search'
- Dim_Location:
 - City
 - County
 - Location_ID
 - Side
 - State
 - Street
- > Dim_POI
- > Dim_Time:
 - \sum End_day
 - \sum End_hour
 - \sum End_minute
 - \sum End_month
 - \sum End_year
 - \sum Start_day
 - \sum Start_hour
 - \sum Start_minute
 - \sum Start_month
 - \sum Start_year
- > Time_ID
- > Dim_Weather
- > Fact

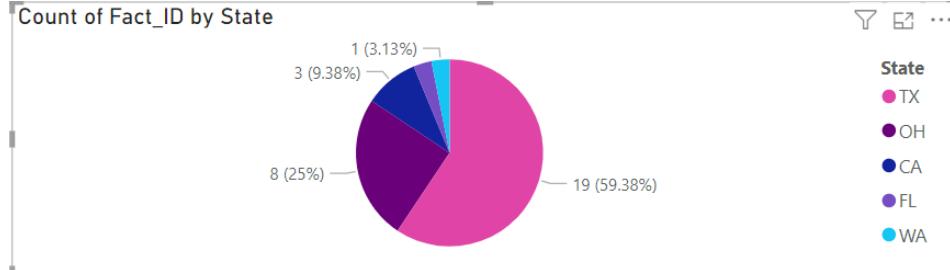
Bước 6: kéo cột **State** vào **Legend**, **Fact_ID** vào **Values** và điều chỉnh điều kiện lọc trong nhóm **Filter** như hình bên dưới

The screenshot shows the Power BI desktop interface with three main panes:

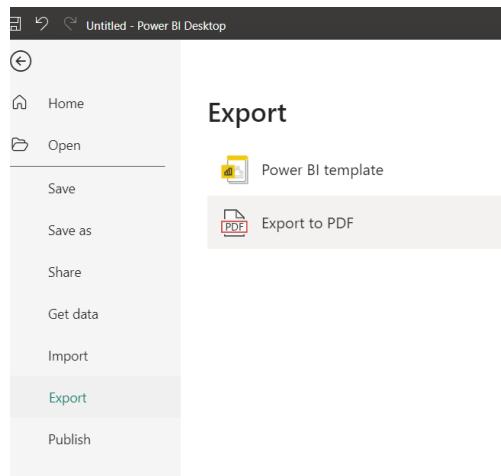
- Filters pane (left):**
 - Search bar: Search
 - Section: Filters on this visual
 - Count of Fact_ID is (All)
 - End_year is 2019
 - Start_year is 2017
 - State is CA, NY, FL, OH, TX, ...
 - Add data fields here
 - Section: Filters on this page
 - Add data fields here
 - Section: Filters on all pages
 - Add data fields here
- Visualizations pane (center):**
 - Build visual section with icons for various chart types.
 - Legend section: State
 - Values section: Count of Fact_ID
 - Details section: Add data fields here
 - Toolips section: Add data fields here
 - Drill through section: Drill through
- Data pane (right):**
 - Search bar: Search
 - Dim_Location section: City, County, Location_ID, Side, State (selected), Street
 - Dim_POI section: ...
 - Dim_Time section: End_day, End_hour, End_minute, End_month, End_year, Start_day, Start_hour, Start_minute, Start_month, Start_year
 - Time_ID section: ...
 - Dim_Weather section: ...
 - Fact section: ...

Bước 7: Xem kết quả ở dạng Pie chart

Ths. Đỗ Thị Minh Phụng



Bước 8: Xuất báo cáo



Bước 9: Kết quả sau khi xuất báo cáo



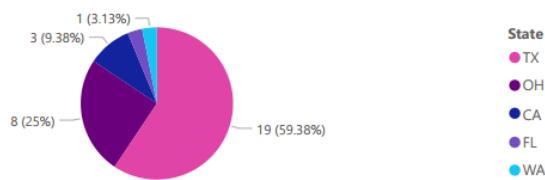
ĐỒ ÁN MÔN KHO DỮ LIỆU VÀ OLAP

NHÓM 39 - LỚP IS217.O21

**Thống kê tổng số công trình (Fact Count) trên 6 tiểu bang
(State): California (CA), New York (NY), Florida (FL),
Ohio (OH), Texas (TX) và Washington (WA) có thời gian
bắt đầu 2017 và kết thúc 2019**

State	Count of Fact_ID
CA	3
FL	1
OH	8
TX	19
WA	1
Total	32

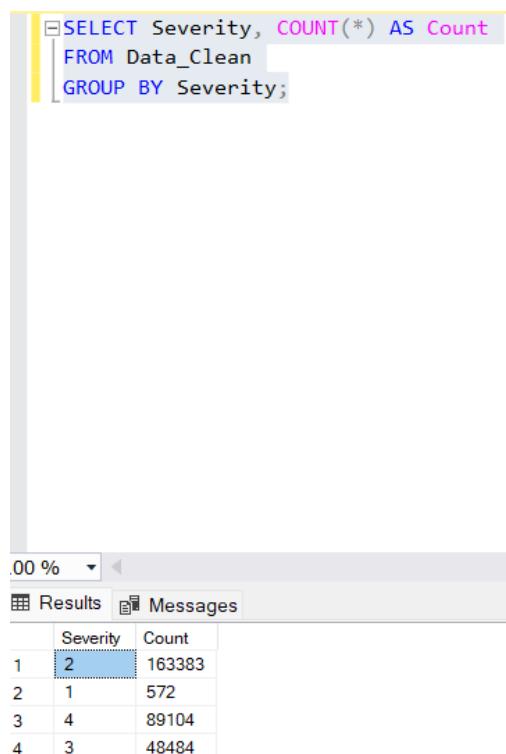
Count of Fact_ID by State



CHƯƠNG 5. KHAI THÁC DỮ LIỆU

5.1. Chuẩn bị dữ liệu:

Lúc đầu, nhóm định sử dụng dữ liệu đã có từ quá trình SSIS trước đó để thực hiện dự đoán mức độ nghiêm trọng của các công trình thi công đường bộ ở Mỹ. Tuy nhiên khi kiểm tra sự phân bố của các dòng Severity thì nhóm nhận thấy có sự chênh lệch lớn như sau:



Do đó nhóm quyết định sẽ tái xử lý dữ liệu từ dataset đã chọn ban đầu. Ở phần data mining này, nhóm sẽ chọn dữ liệu trong 2 năm 2019 và 2020.

```

df_final = df.loc[(df['Start_Time'].dt.year >= 2019) & (df['Start_Time'].dt.year <= 2020) &
                   (df['End_Time'].dt.year >= 2019) & (df['End_Time'].dt.year <= 2020)]

```

Nhóm sẽ lấy các thuộc tính: **ID, Severity, Start_Time, Duration, Distance(mi), State, Temperature(F), Wind_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Direction, Wind_Speed(mph), Precipitation(in), Weather_Condition, Amenity, Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, Sunrise_Sunset**. Bỏ các thuộc tính còn lại do không cần thiết và bị trùng lặp.

	ID	Severity	Start_Time	Duration	Distance(mi)	State	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	...	Give_Way	Junction	No_Exit	Rai
0	C-1	4	2019-04-05 16:00:00	46900437	1.103497	LA	75.0	75.0	58.0	29.72	...	False	False	False	
4	C-5	2	2020-09-24 15:58:00	104814	0.996057	PA	73.0	73.0	NaN	28.78	...	False	False	False	
15	C-16	2	2019-11-06 09:04:01	345576	0.118220	MI	33.0	33.0	78.0	29.73	...	False	False	False	
28	C-29	4	2019-05-14 16:35:59	17990	0.235121	ID	81.0	81.0	20.0	27.19	...	False	False	False	
30	C-31	2	2020-09-22 19:00:00	39600	25.980000	TN	68.0	68.0	55.0	29.19	...	False	False	False	
...
6170615	C-6293065	2	2020-12-18 04:00:00	35100	0.703915	CA	47.0	45.0	51.0	30.15	...	False	False	False	
6170616	C-6293066	2	2020-12-21 17:01:00	31500	1.434852	CA	55.0	55.0	74.0	29.78	...	False	True	False	
6170620	C-6293070	2	2020-07-15 16:00:00	226799	0.100457	NY	81.0	81.0	56.0	29.78	...	False	False	False	
6170621	C-6293071	2	2020-08-26 14:10:00	41608	0.097580	IL	92.0	92.0	44.0	29.23	...	False	False	False	
6170623	C-6293073	2	2020-01-29 21:00:00	28800	3.040000	CA	54.0	54.0	77.0	30.20	...	False	False	False	

2219091 rows × 28 columns

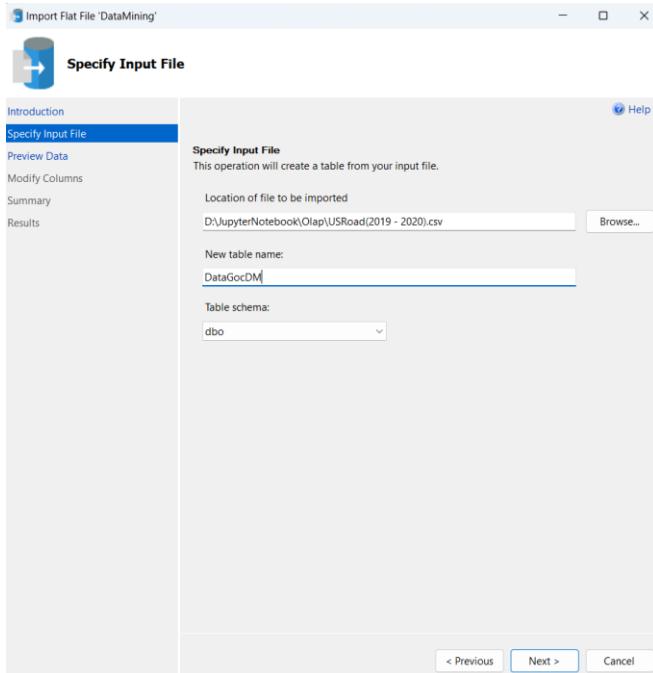
Lưu dữ liệu ra thành 1 file csv : USRoad(2019 – 2020).csv

```
: df_filter1.to_csv('USRoad(2019 - 2020).csv', index=False)
```

5.2. Đồ dữ liệu cho Data Mining bằng công cụ SSIS

Bước 1: Trong SQL Server 2019, tạo database **DataMining** và import dữ liệu file **USRoad(2019 – 2020).csv** vào bảng **DataGocDM**

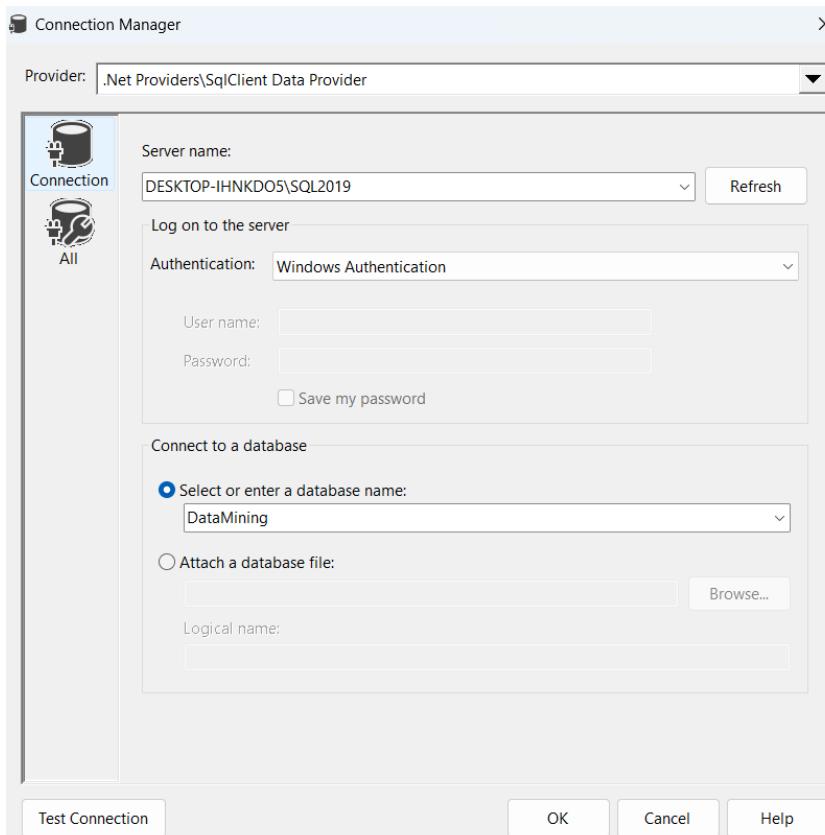
Ths. Đỗ Thị Minh Phụng



Import dữ liệu thành công:

SQL Query Log - D:\HMKD01\DELL (64) - x - SQLQueryLog - L:\HMKD01\SQL (55)													
SELECT TOP (1000) [ID] ,[Severity] ,[Start_Time] ,[Duration] ,[Distance_mi] ,[State] ,[Temperature_F] ,[Wind_Chill_F] ,[Humidity] ,[Pressure_in] ,[Visibility_mi] ,[Wind_Direction] ,[Wind_Speed_mph] ,[Precipitation_in] ,[Weather_Condition] ,[Amenity] ,[Bump] ,[Crossing]													
	ID	Severity	Start_Time	Duration	Distance_mi	Date	Temperature_F	Wind_Chill_F	Humidity	Precipitation	Visibility_mi	Wind_Direction	Wind_Speed_mph
1	1	4	2019-01-05 16:00:00.000	00:00:037	1.9980437	2019-01-05 16:00:00.000	50	30.199999132046	10	0	10	SW	NULL
2	C-1000000	2	2019-01-02 02:27:00.000	215651	0.1032512194193	OR	52	NULL	100	30.199999642242	0.20000002880232	Calm	NULL
3	C-1000000	2	2020-05-01 00:34:15.000	17960	0.10498535942313	NJ	49	47	56	30.420000782939	10	E	5
4	C-1000000	2	2020-08-14 17:00:00.000	89524	0.5032853996796	PA	87	87	49	29.839993536450	10	NE	8
5	C-1000000	2	2020-09-04 17:05:22.000	16748	54.5761326293945	CT	84	84	40	29.760002288118	10	NW	7
6	C-1000000	2	2020-09-04 17:05:22.000	17616	54.5761326293945	NY	86	86	52	29.89999474121	10	W	7
7	C-1000000	2	2019-01-20 10:34:15.000	12659	2.5393999551074	NY	48	30.199994594244	10	0	10	SW	6.900000936743
8	C-1000000	2	2019-01-20 13:49:07.000	19566874	0.92493043493024	OH	35	27	54	29.5	10	NW	10
9	C-1000000	2	2019-03-04 02:08:00.000	345572	0.041558265723448	NY	33.769999237065	24.299999237065	93	29.790000159273	1.5	NE	13.800000190734
10	C-1000000	2	2019-07-10 09:00:00.000	25200	0.430000000000000	NY	73	73	77	29.190000230376	10	W	10
11	C-1000000	2	2020-01-01 00:00:00.000	40000	0.000000000000000	NY	73	73	46	29.559994524040	10	CALM	0
12	C-1000000	2	2020-01-01 00:00:00.000	43200	0.000734049038623	CO	19.8999996185303	14.8999996185303	85	29.559994524040	10	ESE	0
13	C-1000000	2	2020-12-07 08:00:00.000	43199	0.56617055443973	MD	30	30	88	29.559999458242	10	CALM	0
14	C-1000000	2	2020-10-10 14:00:00.000	40742	1.09994596264626	NY	70	70	76	29.44999158303	10	N	8
15	C-1000000	2	2020-02-02 02:23:00.000	34460	0.672841096575	WA	57	57	10	29.799991677663	10	CALM	0

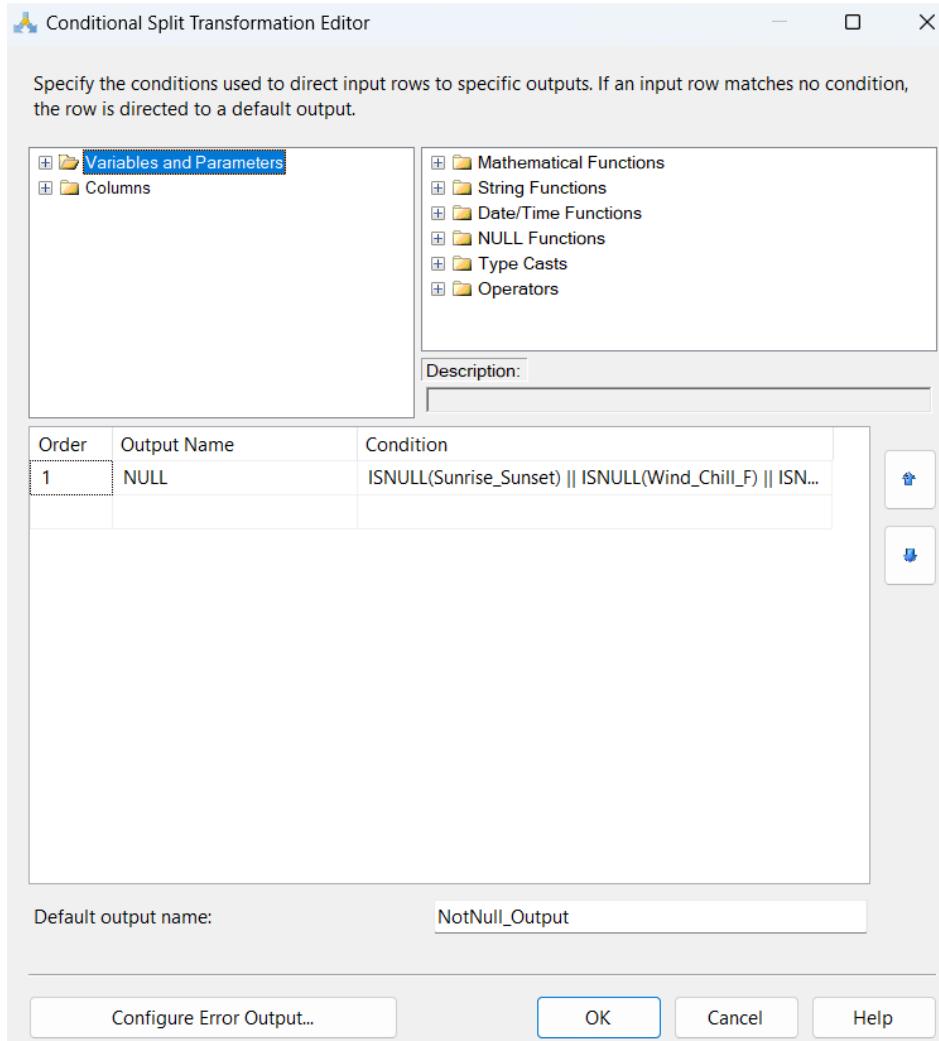
Bước 2: Tạo 1 project SSIS mới và thực hiện tạo 1 connection mới:



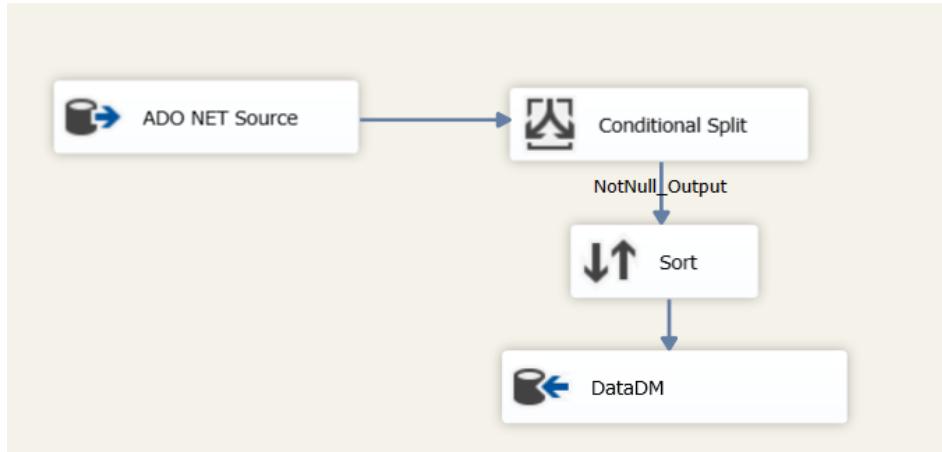
Bước 3: Thực hiện tương tự **Quá trình làm sạch dữ liệu** đã thực hiện trong chương 2.

Trong đó, tạo 1 **Conditional Split** với đoạn code sau:

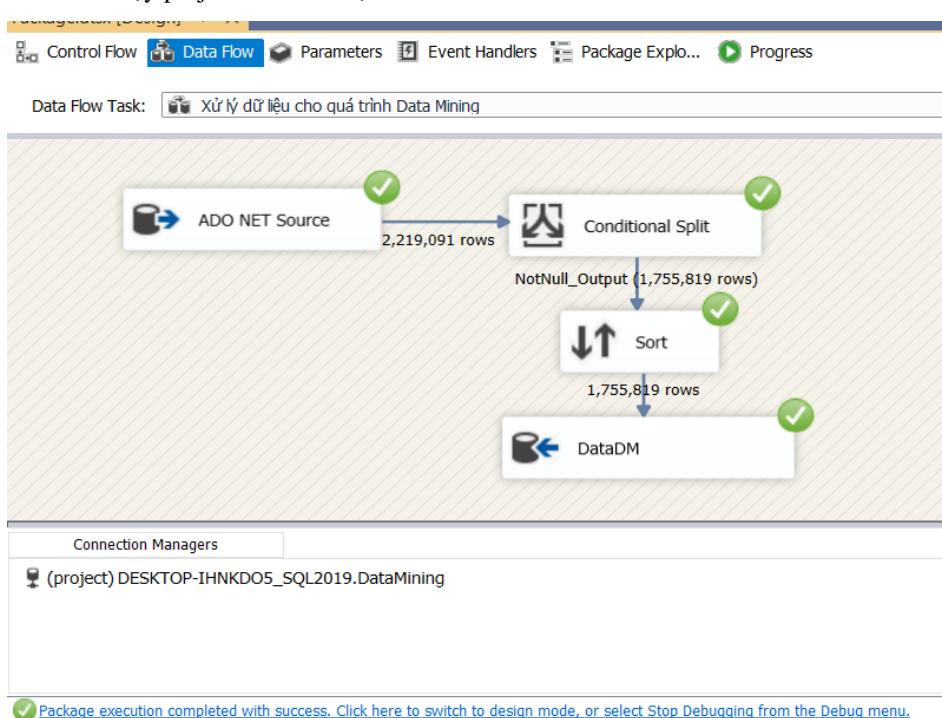
```
ISNULL( Sunrise_Sunset ) || ISNULL( Wind_Chill_F ) || ISNULL( Pressure_in ) ||
ISNULL(Humidity) || ISNULL( Visibility_mi ) || ISNULL( Wind_Direction ) ||
ISNULL(Wind_Speed_mph) || ISNULL( Precipitation_in ) ||
ISNULL(Weather_Condition) || REPLACENULL(Distance_mi,0) == 0 ||
ISNULL(Temperature_F)
```



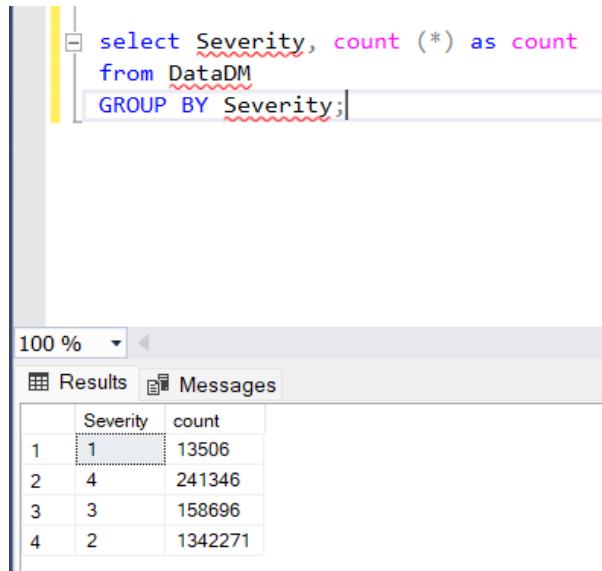
Thực hiện tương tự với các bước ở Chương 2, ta được 1 data flow như sau:



Bước 4: Chạy project SSIS vừa tạo.



Bước 5: Trong SQL Server, thực hiện kiểm tra sự phân bố của các dòng Severity:



The screenshot shows a SQL query window with the following code:

```
select Severity, count (*) as count
from DataDM
GROUP BY Severity;
```

Below the code is a results grid:

	Severity	count
1	1	13506
2	4	241346
3	3	158696
4	2	1342271

Bước 6: Thực hiện tạo ra 1 bảng dữ liệu **Data_UseDm** để phục vụ cho quá trình **Data Mining** với đoạn code sau:

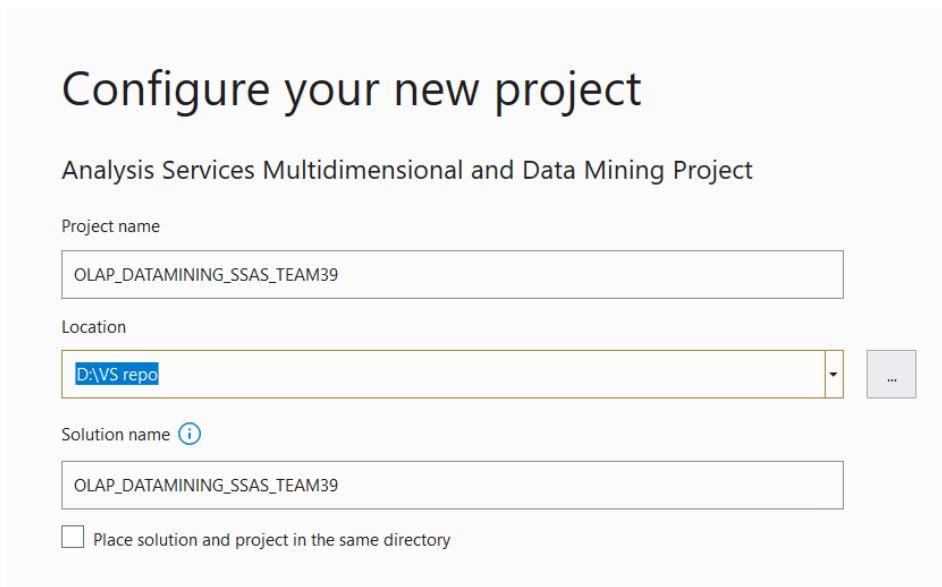
```
SELECT * INTO Data_UseDM
FROM DataDM
WHERE ID IN (
    (SELECT TOP 10000 ID FROM DataDM WHERE Severity = 1)
    UNION
    (SELECT TOP 10000 ID FROM DataDM WHERE Severity = 2)
    UNION
    (SELECT TOP 10000 ID FROM DataDM WHERE Severity = 3)
    UNION
    (SELECT TOP 10000 ID FROM DataDM WHERE Severity = 4)
)
```

```
SELECT * INTO Data_UsedM
FROM DataDM
WHERE ID IN (
    (SELECT TOP 10000 ID FROM DataDM WHERE Severity = 1)
UNION
    (SELECT TOP 10000 ID FROM DataDM WHERE Severity = 2)
UNION
    (SELECT TOP 10000 ID FROM DataDM WHERE Severity = 3)
UNION
    (SELECT TOP 10000 ID FROM DataDM WHERE Severity = 4)
)
```

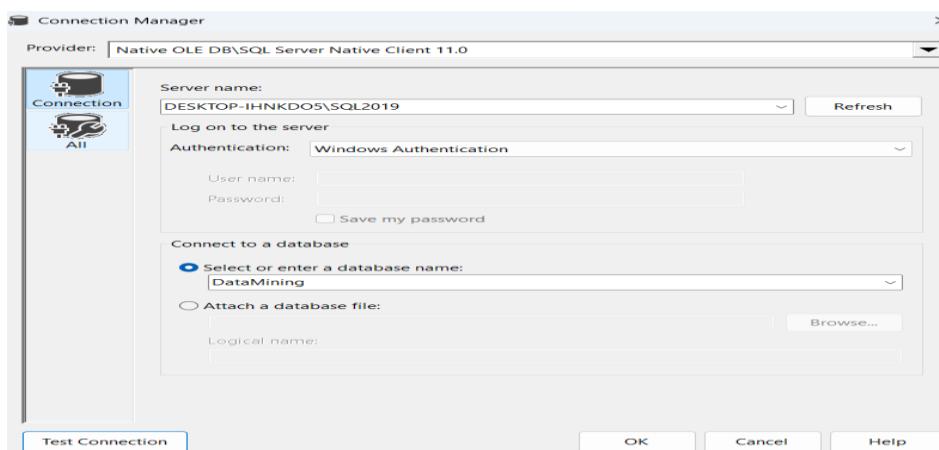
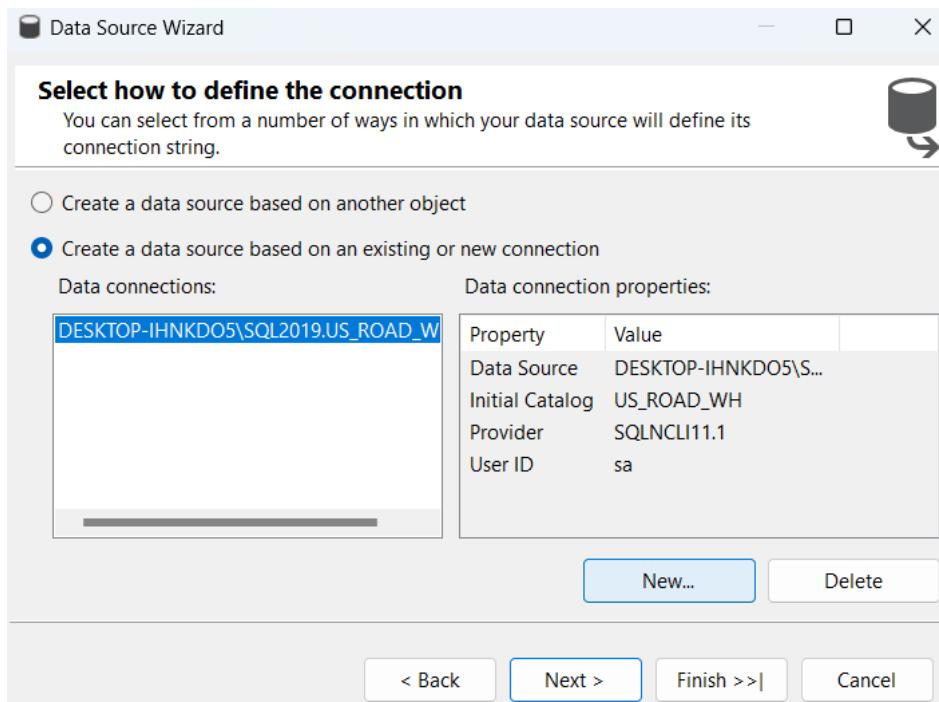
00 % ▶ Messages (40000 rows affected)
Completion time: 2024-05-20T01:32:02.4716298+07:00

5.3. Thực hiện Data Mining bằng SSAS:

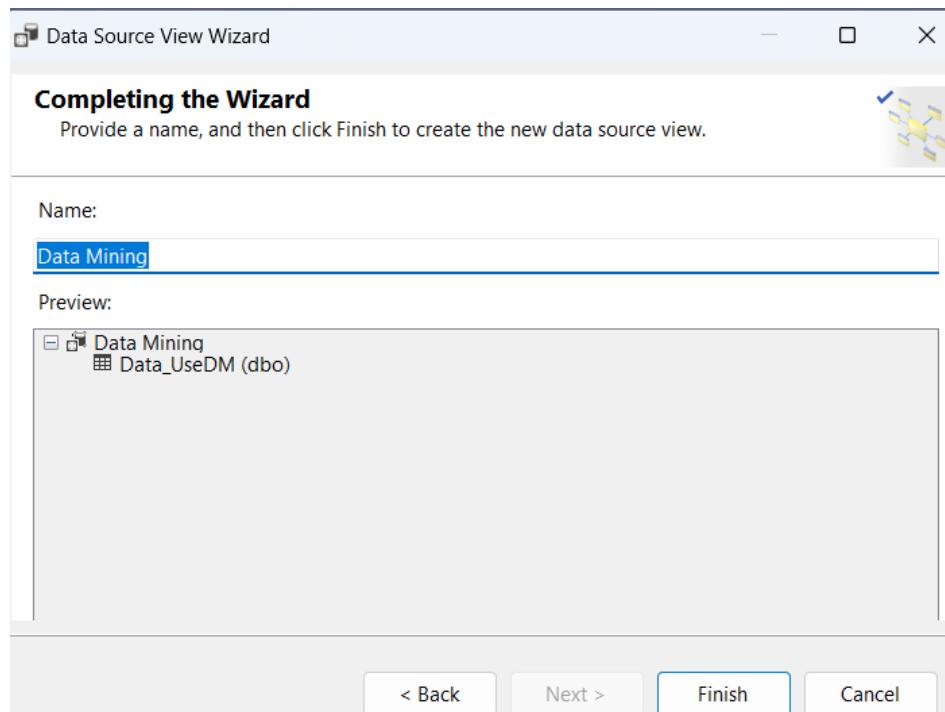
Bước 1: Tạo 1 project SSAS mới.



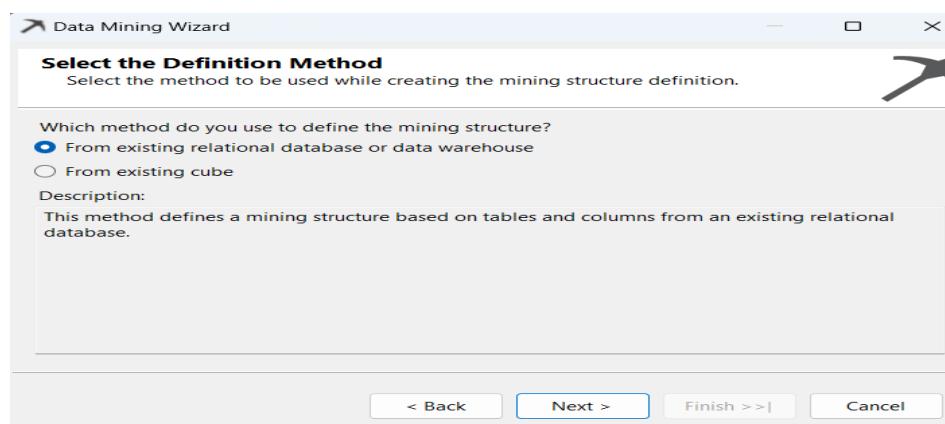
Bước 2: Nhấn tạo 1 data source mới dựa vào 1 connection mới được tạo.



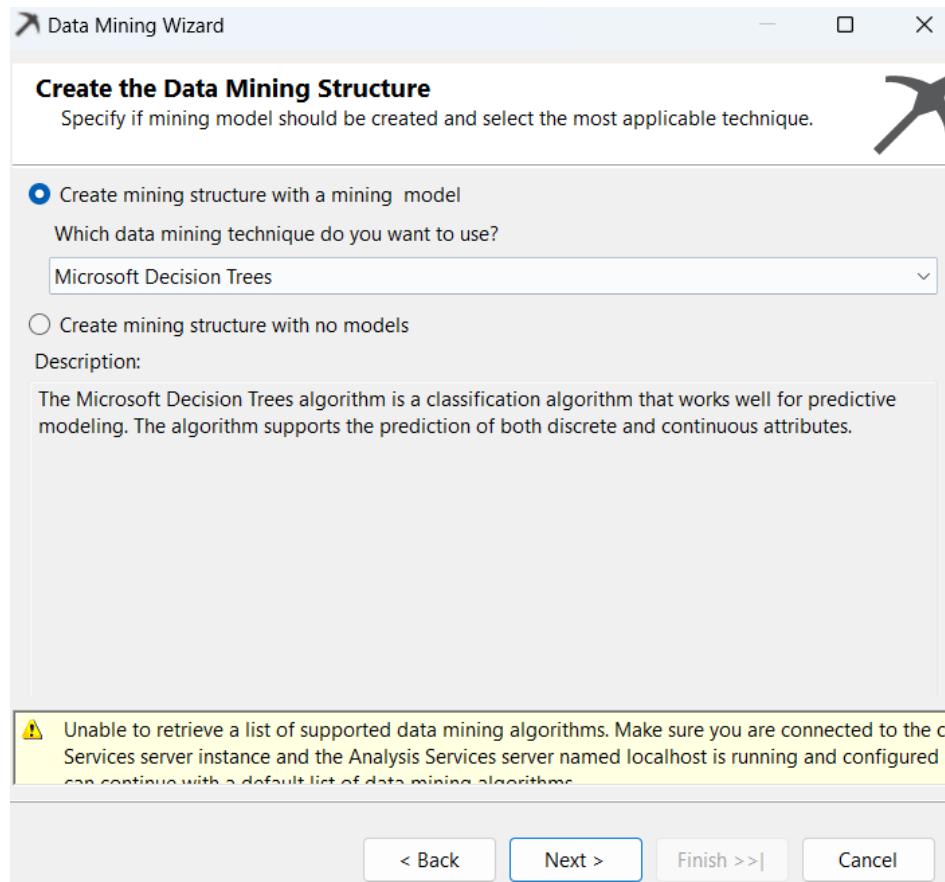
Bước 3: Tạo Data Source View mới sử dụng bảng Data_UseDM:



Bước 4: Tạo Mining Structures

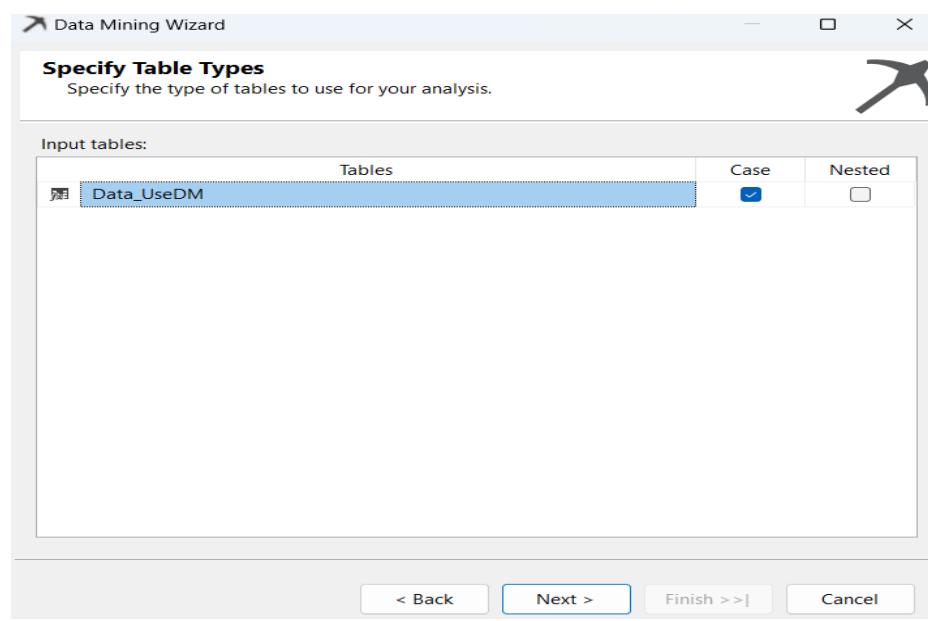
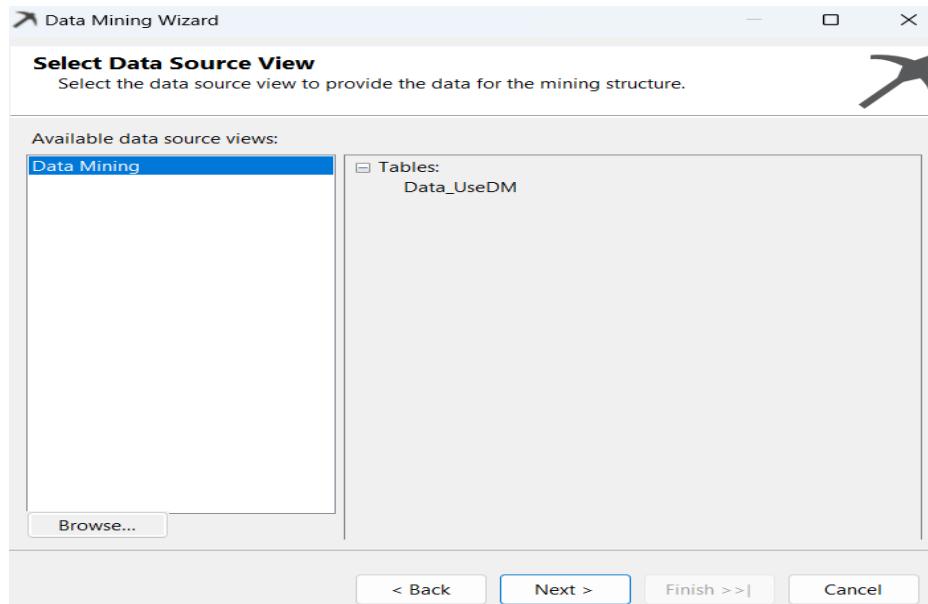


Bước 5: Lựa chọn thuật toán đầu tiên là **Microsoft Decision Trees** (Cây quyết định)



Bước 6: Lựa chọn **Data Source View** cho thuật toán là **Data Mining** và bảng **Data_UseDM**

Ths. Đỗ Thị Minh Phụng



Bước 7: Lựa chọn các cột để làm **Key**, **Input** và **Prediction**. Ở đây sẽ chọn cột **ID** làm **Key**, cột **Severity** để làm **Prediction** và lựa chọn các cột còn lại (trừ **Start_Time**) để làm **Input**

Tables/Columns	Key	Input	Predict...
Give_Way	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Humidity	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
ID	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Junction	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
No_Exit	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Precipitation_in	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Pressure_in	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Railway	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Roundabout	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Severity	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Start_Time	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
State	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Recommend inputs for currently selected predictable:

Bước 8: Detect các cột dữ liệu

Data Mining Wizard

Specify Columns' Content and Data Type

Specify mining structure columns' content and data type.

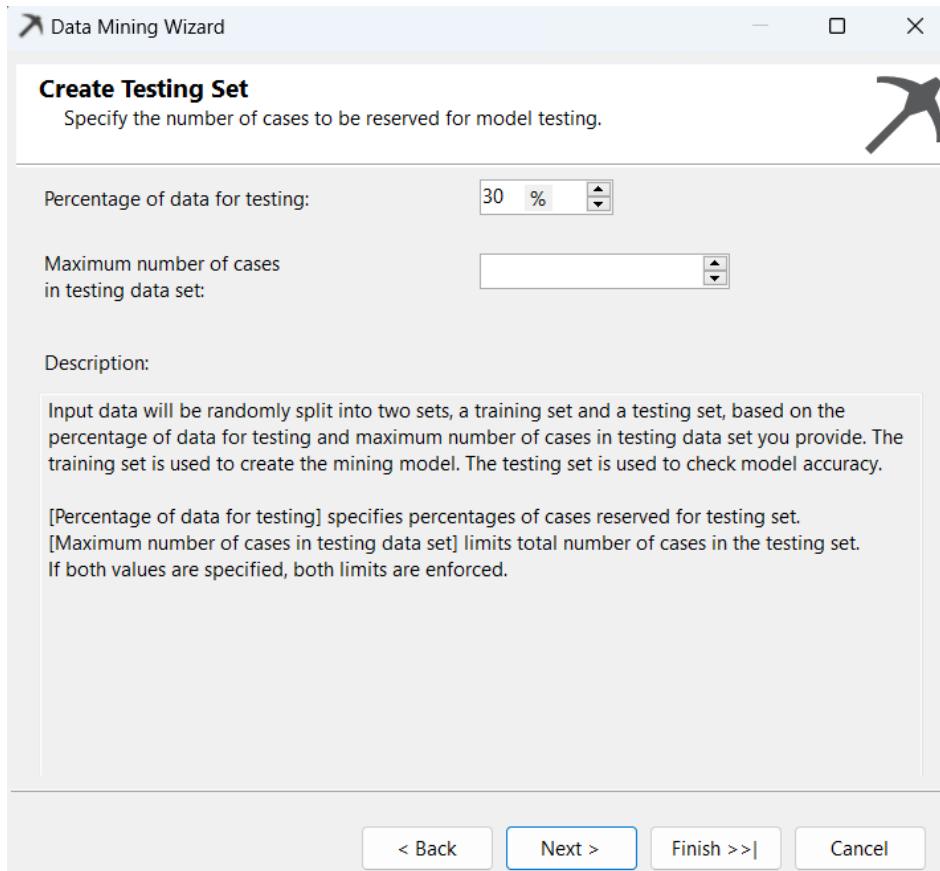
Mining model structure:

Columns	Content Type	Data Type
Humidity	Continuous	Double
ID	Key	Text
Junction	Discrete	Boolean
No Exit	Discrete	Boolean
Precipitation In	Continuous	Double
Pressure In	Continuous	Double
Railway	Discrete	Boolean
Roundabout	Discrete	Boolean
Severity	Discrete	Long
State	Continuous	Text
Station	Cyclical	Boolean
Stop	Discrete	Boolean
	Discretized	Boolean
	Ordered	Boolean

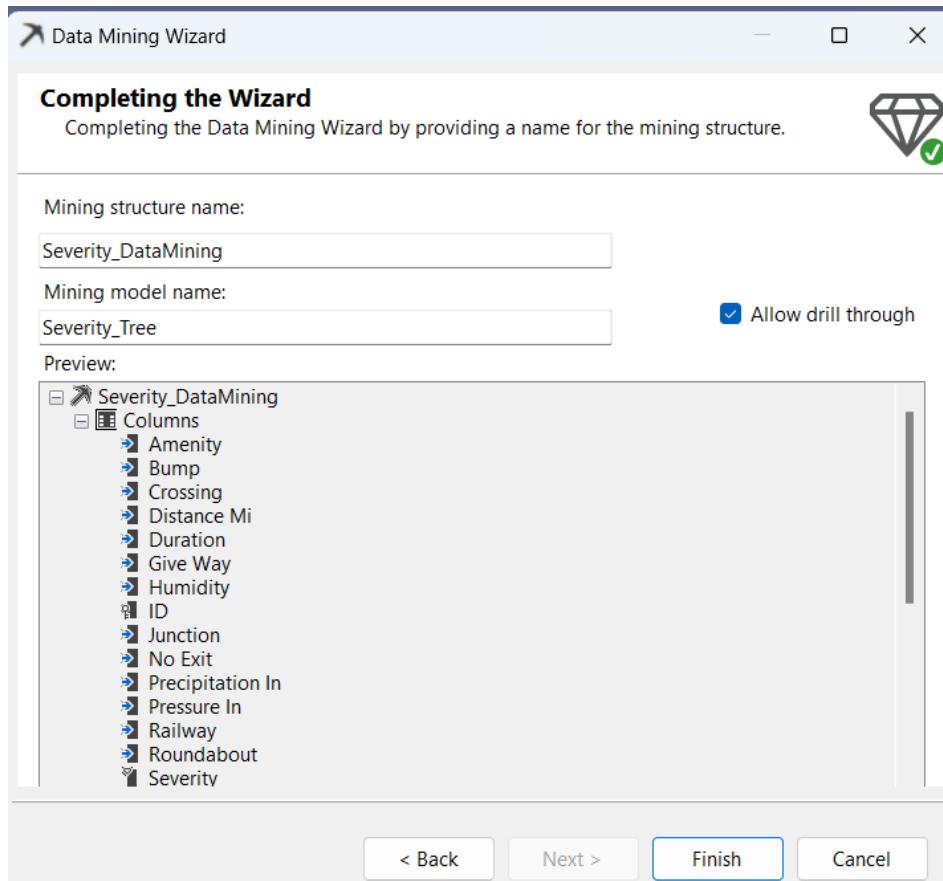
Detect continuous or discrete for numeric columns:

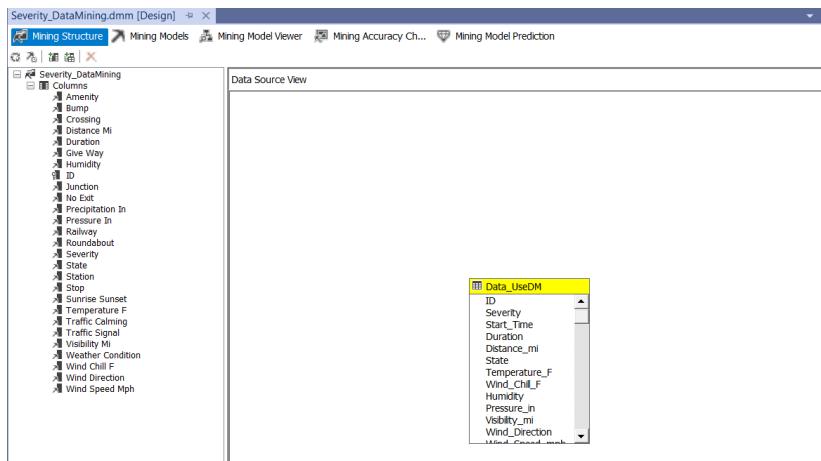
< Back Finish >> Cancel

Bước 9: Chia dữ liệu thành các tập Train và Test với tỉ lệ 70:30.

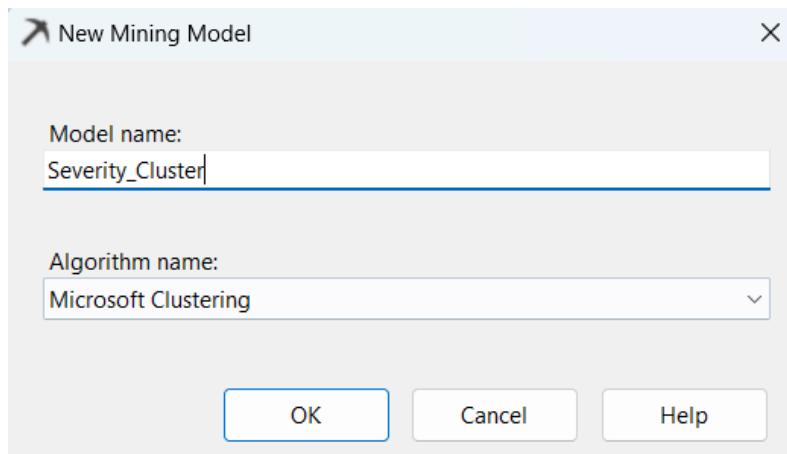


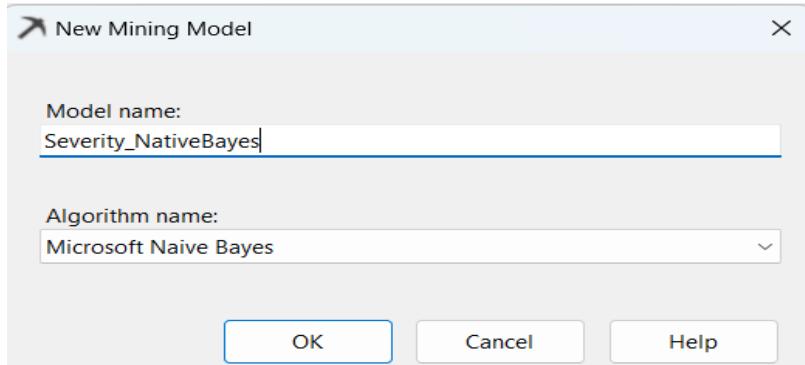
Bước 10: Đặt lại tên cho model là **Severity_DataMining** và nhấn **Finish**.



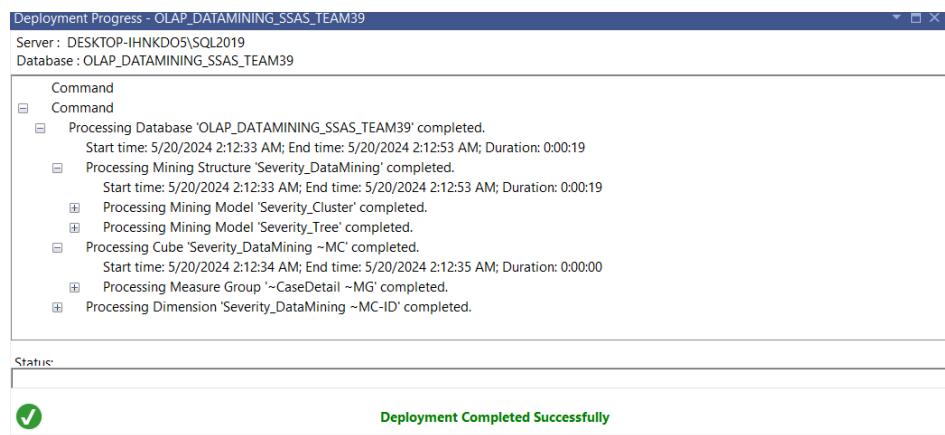


Bước 11: Tiếp tục thêm 2 model cho 2 thuật toán **Clustering** (Gom cụm) và **Naïve Bayes**





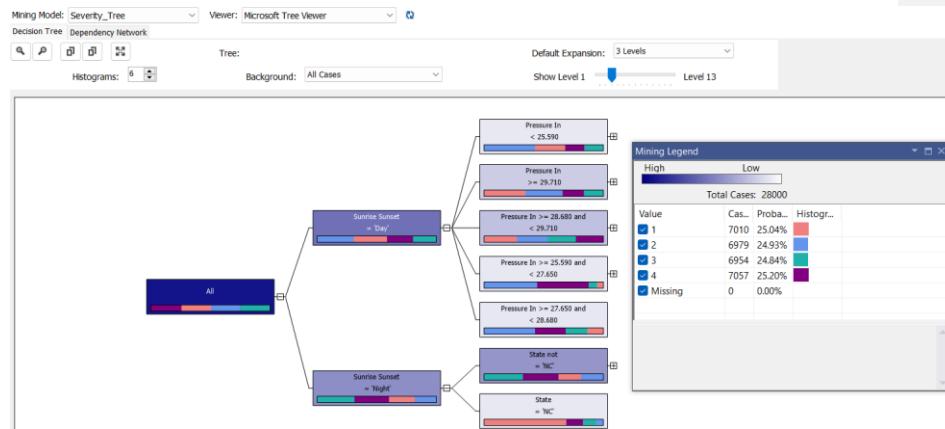
Bước 12: Sau khi tạo xong các mô hình, click chuột phải vào tên project , sau đó chọn Deploy.



5.4. Phân tích các mô hình khai phá dữ liệu:

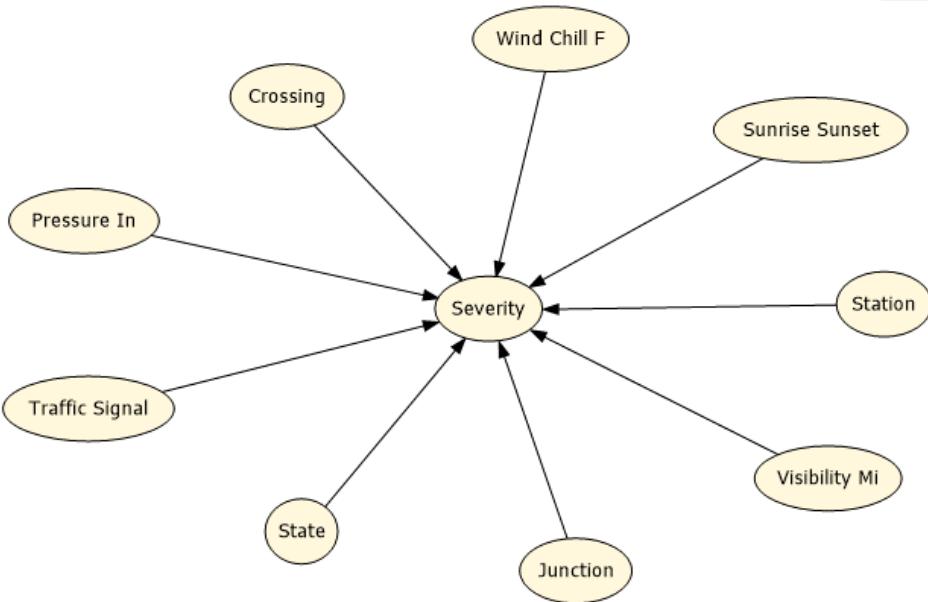
5.4.1. Mô hình cây quyết định:

Cây được vẽ thành công thì sẽ hiển thị ở **cấp độ 3** và **Background là “All”**.



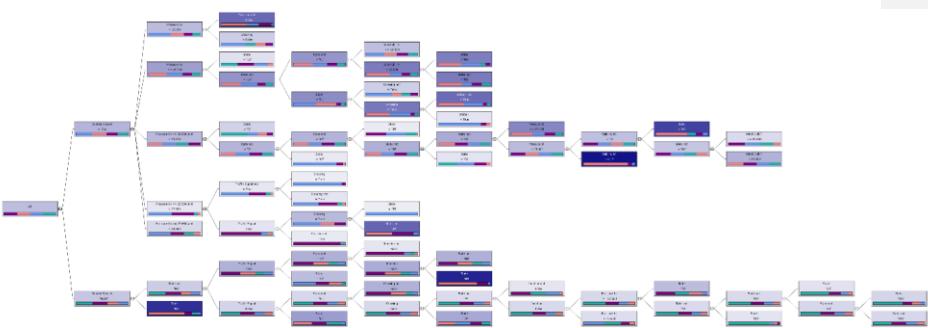
Xét trên thuộc tính dự đoán Severity:

- Tỉ lệ ở mức 1 là 25,04%
 - Tỉ lệ ở mức 2 là 24,93%
 - Tỉ lệ ở mức 3 là 24,84%
 - Tỉ lệ ở mức 4 là 25,2%
- Nhấn chọn mục **Dependency Network** để hiển thị quan hệ của các thuộc tính có thể trực tiếp kết luận đến cột dự đoán.

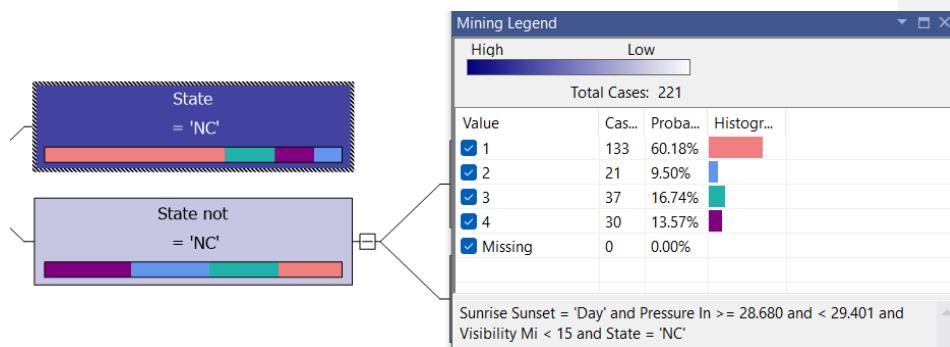


Ở hình trên, ta có thể thấy các thuộc tính ảnh hưởng nhiều tới thuộc tính dự đoán **Severity**.
Những thuộc tính đó sắp xếp giảm dần theo mức độ ảnh hưởng như sau: **Sunrise Sunset > State > Pressure In > Traffic Signal > Crossing > Junction > Wind Chill F > Visibility Mi > Station.**

Xét trường hợp **Severity = 1**: Ta chọn Background là 1 và chọn Level 13 (**level max**)



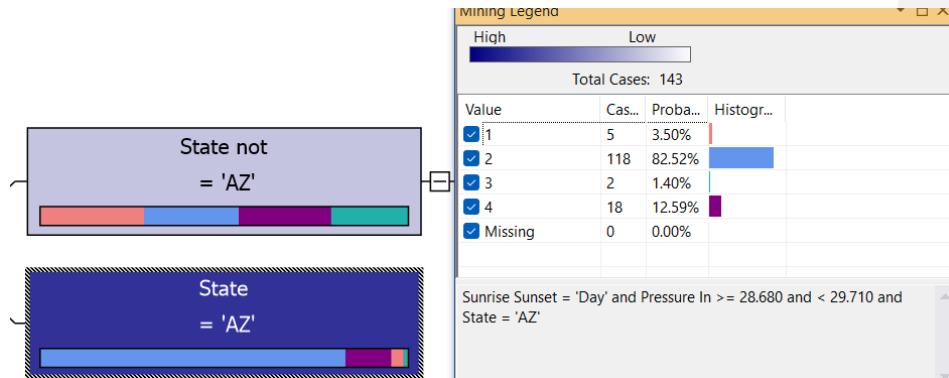
Với các công trình xây dựng đường bộ với mức nghiêm trọng 1 có độ tin tưởng và tỷ lệ cao là **60,18%** là những công trình được thi công vào ban ngày, áp suất không khí nằm trong khoảng từ **28,68in** đến dưới **29,401in**, tầm nhìn xa từ 15 dặm trở lên và không thuộc 3 bang **Arizona, Indiana, California**.



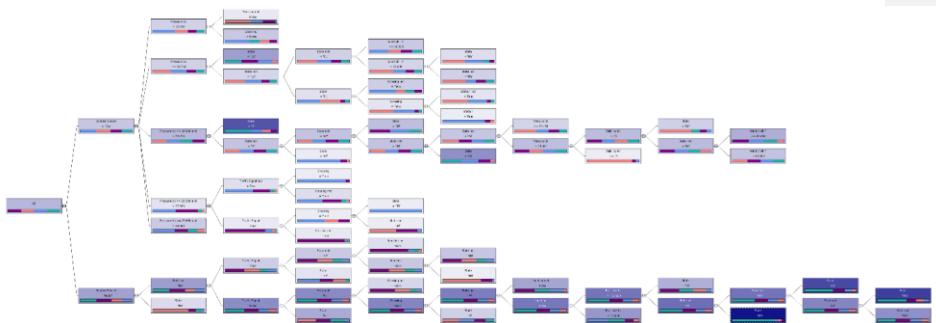
Xét trường hợp Severity = 2: Ta chọn Background là 2 và chọn Level 13 (level max)



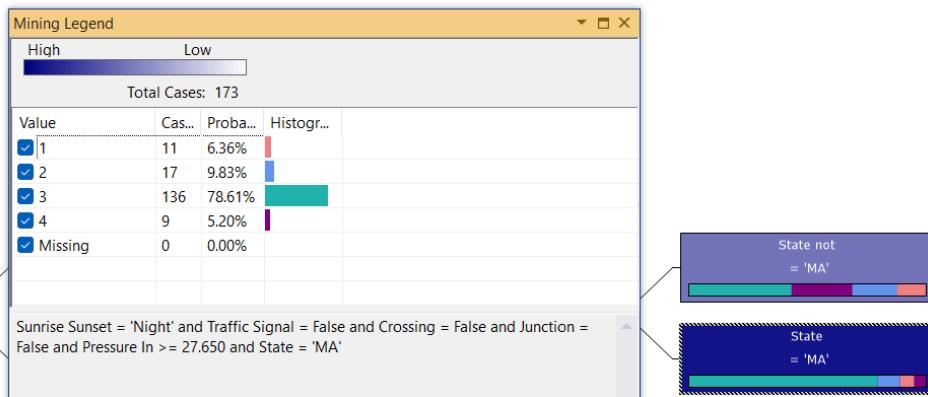
Với các công trình xây dựng đường bộ với mức nghiêm trọng 2 có độ tin tưởng và tỷ lệ cao **82,52%** là những công trình được thi công vào ban ngày, áp suất không khí nằm trong khoảng từ **bằng 28,68 in** đến dưới **29,71 in** và thuộc tiểu bang **Arizona**.



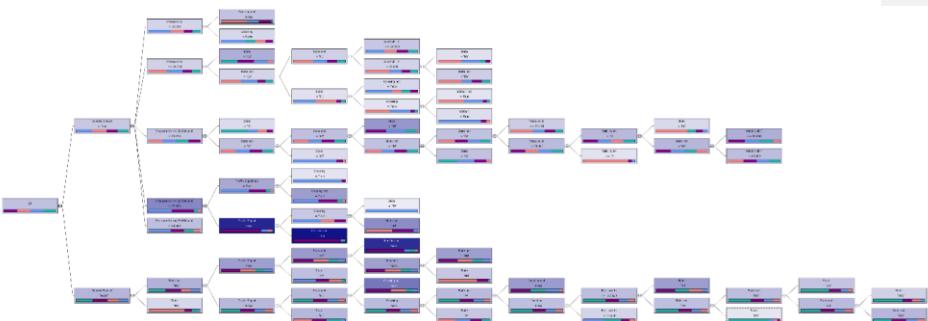
Xét trường hợp Severity = 3: Ta chọn Background là 3 và chọn Level 13 (level max)



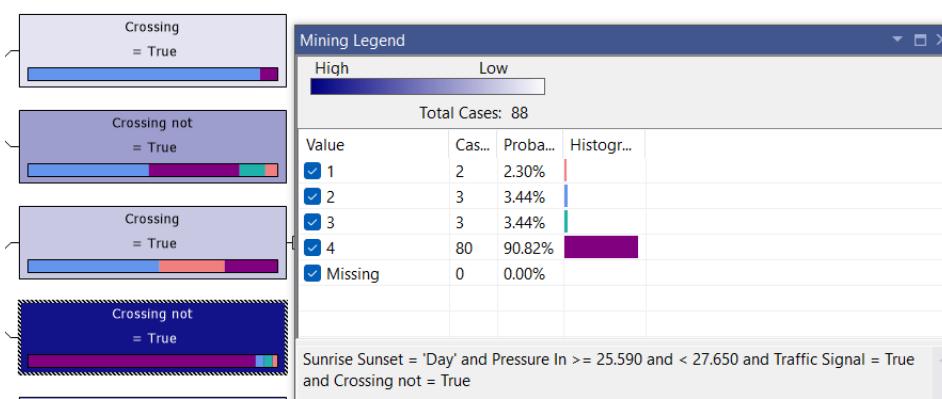
Với các công trình xây dựng đường bộ với mức nghiêm trọng 3 có độ tin tưởng và tỷ lệ cao **78,61%** là những công trình được thi công vào ban đêm, thuộc tiểu bang **Massachusetts**, áp suất không khí từ **27,65 in** trở lên và không có đèn tín hiệu giao thông, ngã tư hay giao lộ gần chỗ thi công.



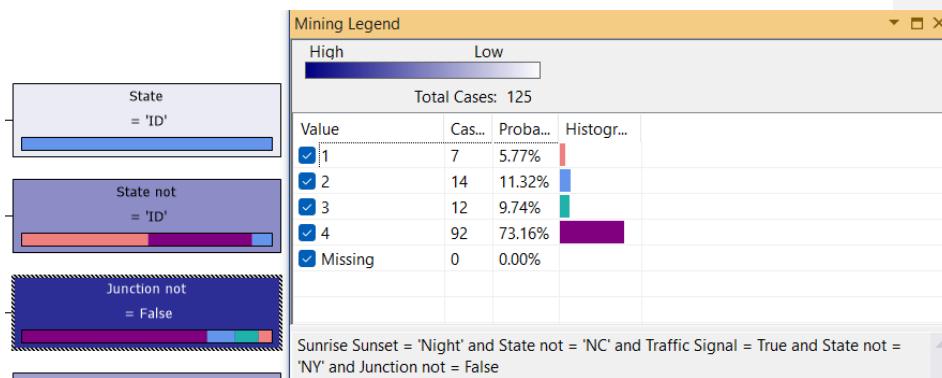
Xét trường hợp Severity = 4: Ta chọn Background là 4 và chọn Level 13 (level max)



Với các công trình xây dựng đường bộ với mức nghiêm trọng 4 có độ tin tưởng cao và tỷ lệ cao **90,82%** là những công trình được thi công vào ban ngày, áp suất không khí nằm trong khoảng từ **25,59 in** đến **27,65 in** và có đèn tín hiệu giao thông, ngã tư nằm gần công trình thi công.



Với các công trình xây dựng đường bộ với mức nghiêm trọng 4 có độ tin tưởng cao và tỷ lệ cao **73,16%** là những công trình được thi công vào ban ngày, không thuộc bang **New York, North Carolina** và có đèn tín hiệu giao thông, giao lộ nằm gần công trình thi công.



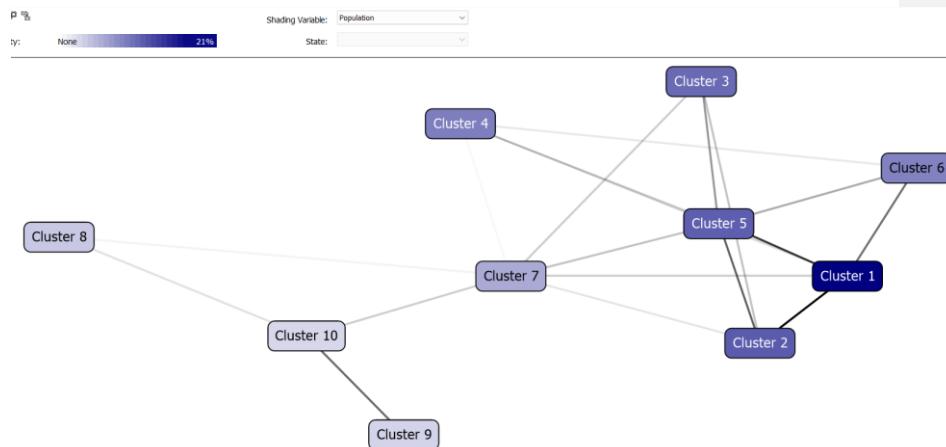
➔ Rút ra được tập luật:

- Công trình thi công đường bộ Hoa Kỳ sẽ có **mức độ nghiêm trọng là 1** nếu được thi công vào ban ngày, áp suất không khí nằm trong khoảng từ **28,68in** đến dưới **29,401 in**, tầm nhìn xa từ **15 dặm** trở lên và không thuộc 3 bang **Arizona, Indiana, California**.

- Công trình thi công đường bộ Hoa Kỳ sẽ có **mức độ nghiêm trọng là 2** nếu được thi công vào ban ngày, áp suất không khí nằm trong khoảng từ bằng 28,68 in đến dưới 29,71 in và thuộc tiểu bang Arizona.
- Công trình thi công đường bộ Hoa Kỳ sẽ có **mức độ nghiêm trọng là 3** nếu được thi công vào ban đêm, thuộc tiểu bang Massachusetts, áp suất không khí từ 27,65 in trở lên và không có đèn tín hiệu giao thông, ngã tư hay giao lộ gần chỗ thi công.
- Công trình thi công đường bộ Hoa Kỳ sẽ có **mức độ nghiêm trọng là 4** nếu được thi công vào ban ngày, áp suất không khí nằm trong khoảng từ 25,59 in đến 27,65 in và có đèn tín hiệu giao thông, ngã tư nằm gần công trình thi công. Hoặc nếu được thi công vào ban ngày, không thuộc bang New York, North Carolina và có đèn tín hiệu giao thông, giao lộ nằm gần công trình thi công.

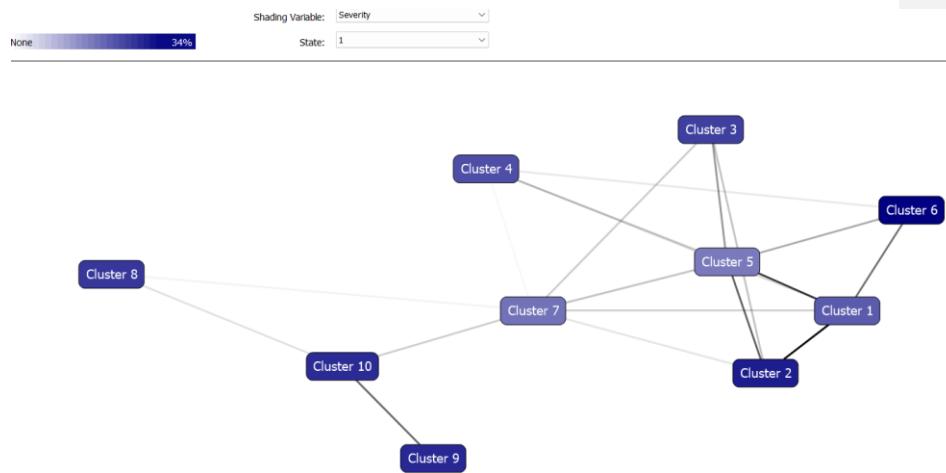
5.4.2. Mô hình Gom cụm:

Mặc định là ta sẽ được thuộc tính **Population** (hệ thống tự tạo) làm thuộc tính xét chung toàn bộ. Ta có thể kéo trượt thanh lượng liên kết bên trái để hiển thị các liên kết theo nhu cầu. Màu cụm càng đậm thì càng tập trung nhiều dữ liệu. Ở đây ta thấy cụm 1 màu đậm nhất.



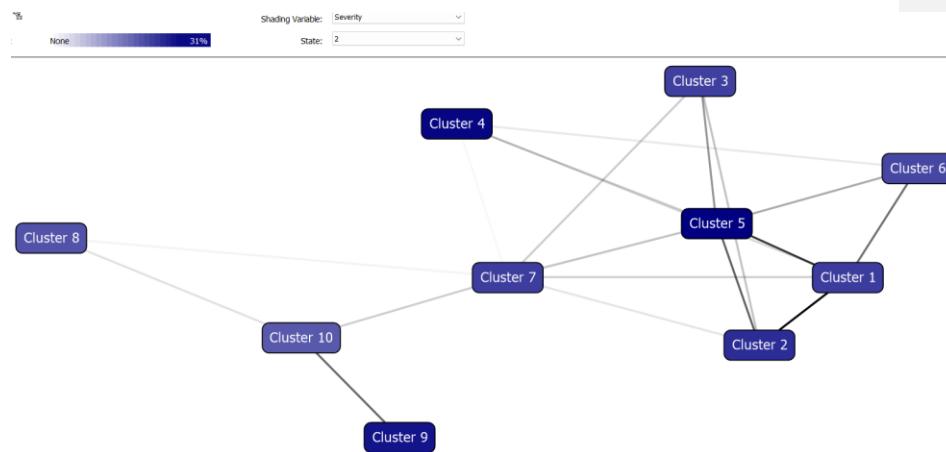
Ths. Đỗ Thị Minh Phụng

Xét **Clustering Diagram** của thuật toán là **Shading Variable** là **Severity** với value là 1 (mức nghiêm trọng 1):



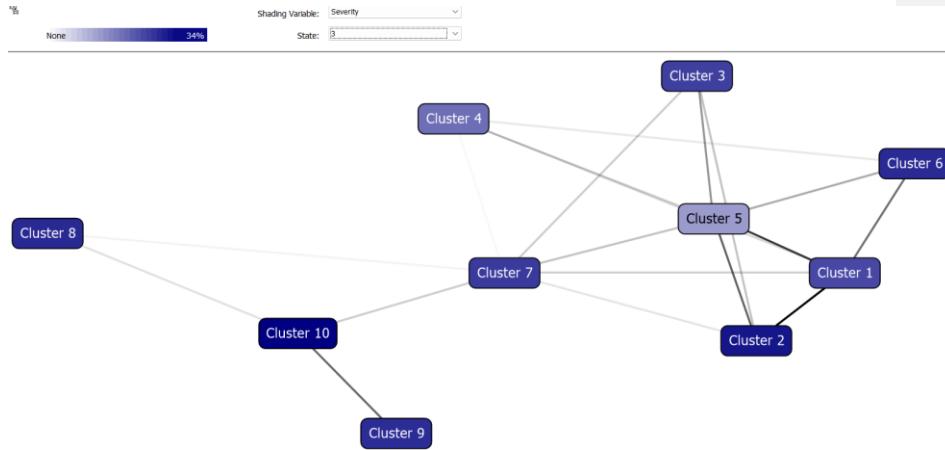
Ta thấy cụm 6 màu đậm nhất nên có tỉ lệ mức 1 cao nhất (34%)

Xét **Clustering Diagram** của thuật toán là **Shading Variable** là **Severity** với value là 2:



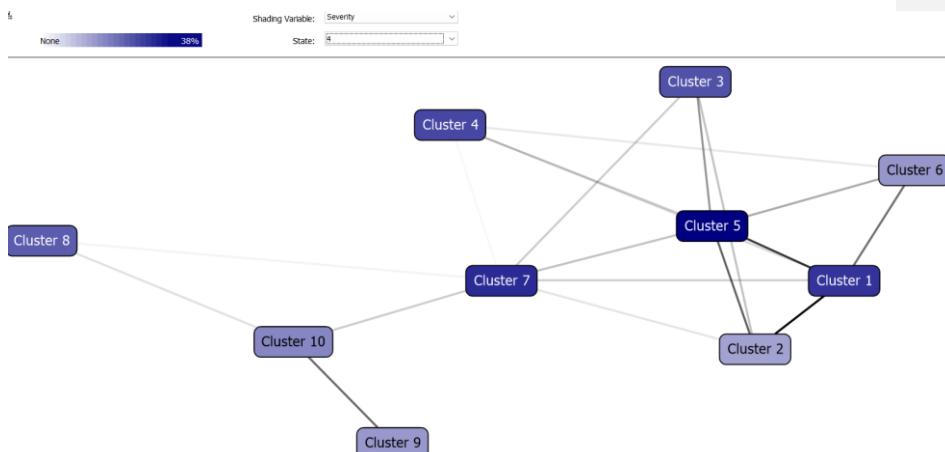
Ta thấy cụm 5 màu đậm nhất nên có tỉ lệ mức 2 cao nhất (31%)

Xét **Clustering Diagram** của thuật toán là **Shading Variable** là **Severity** với value là 3:



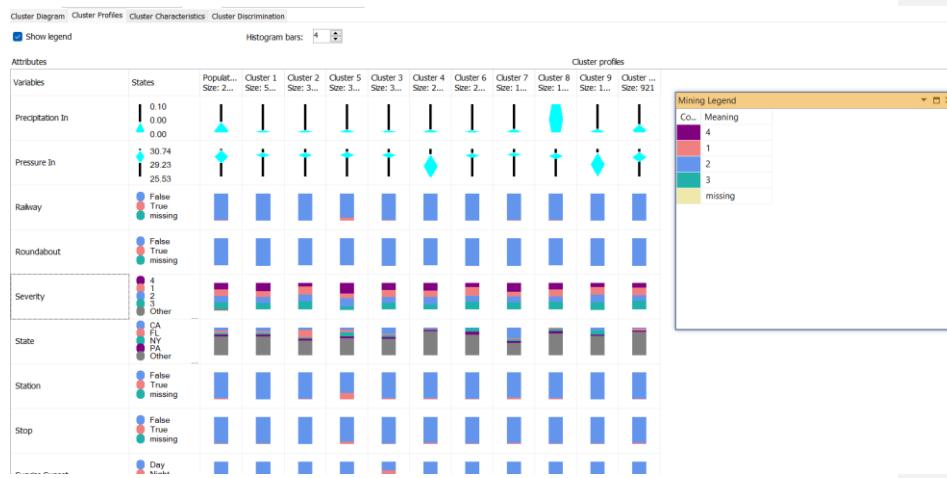
Ta thấy cụm 10 màu đậm nhất nên có tỉ lệ mức 3 cao nhất (34%)

Xét Clustering Diagram của thuật toán là Shading Variable là Severity với value là 4:



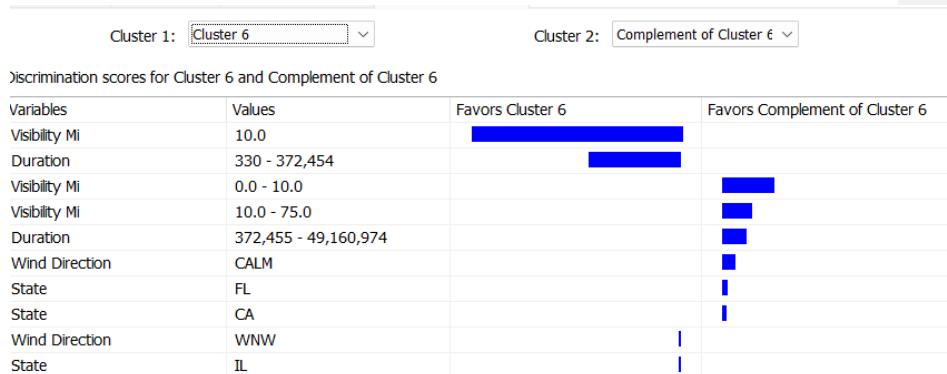
Ta thấy cụm 5 màu đậm nhất nên có tỉ lệ mức 4 cao nhất (38%)

Trong mục Cluster Profile, ta có thể quan sát kỹ hơn thông số tỉ lệ và số lượng trường hợp của mỗi cụm với từng thuộc tính.

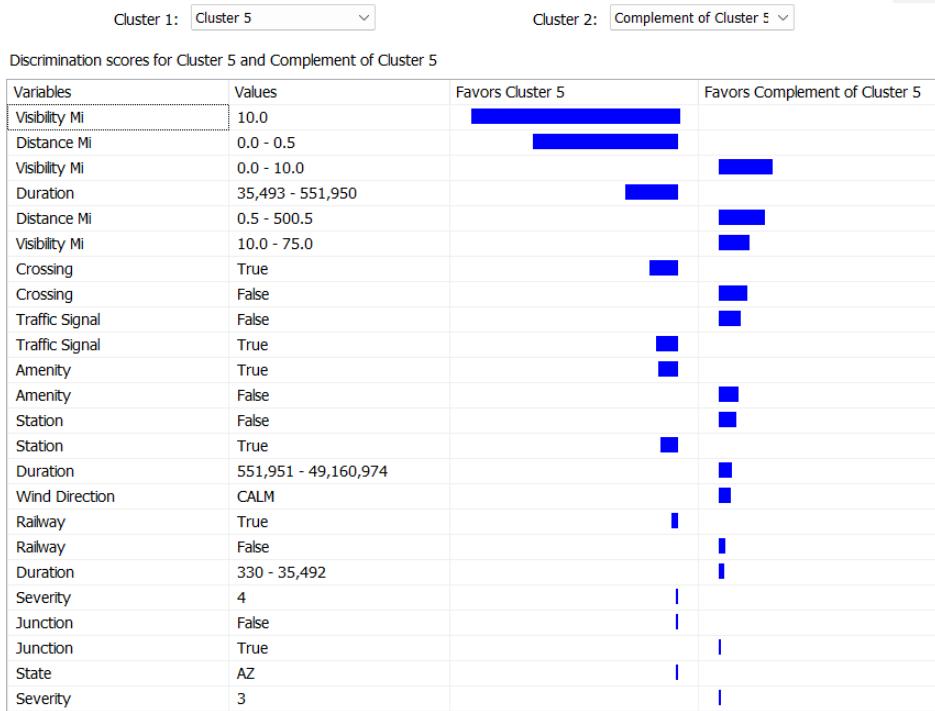


Trong mục **Cluster Discrimination**, ta có thể phân tích các thuộc tính có chiếm tỉ lệ áp đảo về một giá trị của một cụm.

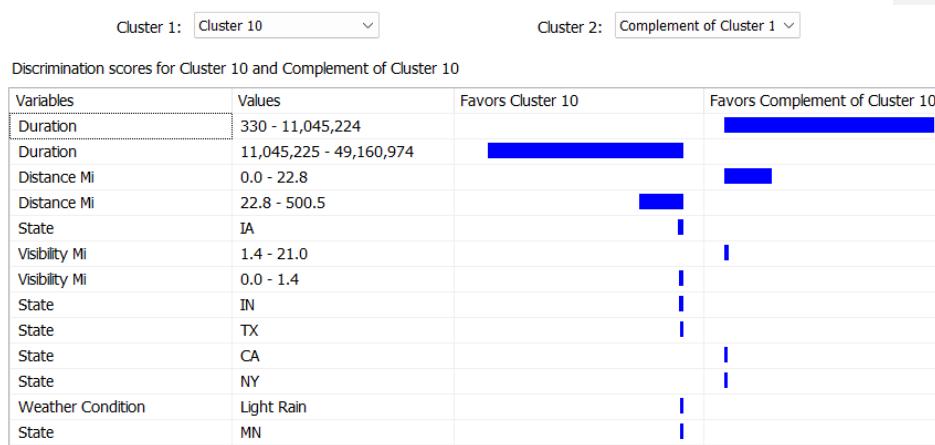
Xét với cụm 6 với tỉ lệ **severity = 1** cao nhất: ta thấy nếu tầm nhìn xa bằng 10 dặm thì khả năng cao **severity = 1**.



Xét với cụm 5 với tỉ lệ **severity = 2,4** cao nhất: ta thấy nếu tầm nhìn xa bằng 10 dặm và độ dài công trình thi công từ 0 – 0,5 dặm thì khả năng cao sẽ có mức nghiêm trọng 2 hoặc 4.

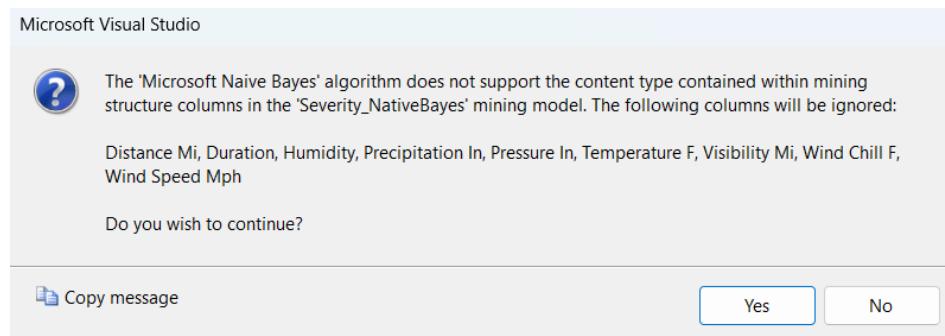


Xét với cụm 10 với tỉ lệ **severity = 3** cao nhất: ta thấy nếu thời gian thi công nằm trong khoảng 11045225 giây tới 49160974 giây thì khả năng mức nghiêm trọng 3 cao.



Mô hình Naive Bayes

Tuy nhiên **Naive Bayes** không thể xử lý kiểu dữ liệu số thực (kiểu biến liên tục) nên các cột thuộc tính có kiểu số thực như độ dài, thời gian, độ ẩm,... không thể tham gia vào thuật toán **Naive Bayes**.



Vì các thuộc tính trên có tính ảnh hưởng lớn đến thuộc tính severity theo 2 mô hình trên nên nhóm chúng em sẽ không nghiên cứu về thuật toán này.

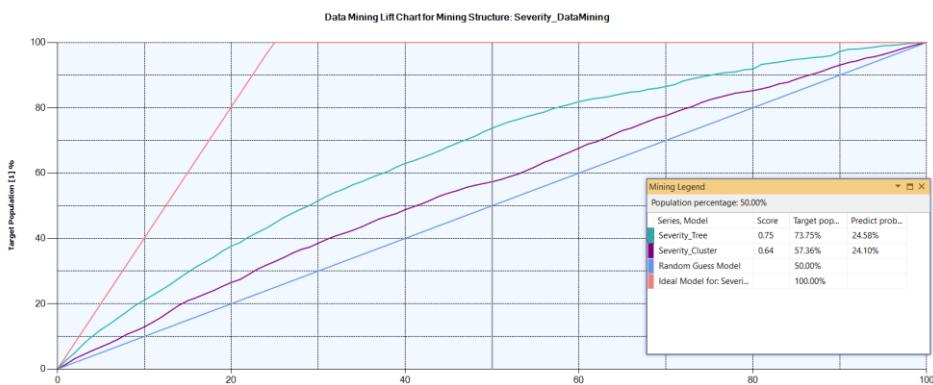
5.5. Đánh giá độ chính xác của các mô hình thuật toán bằng Mining Accuracy Chart:

- Chuyển sang tab **Mining Accuracy Chart** sau đó chọn **Predict Value** là 1 để đánh giá độ chính xác trong trường hợp mức nghiêm trọng là 1

The screenshot shows the "Mining Accuracy Chart" tab selected in the Microsoft SQL Server Data Mining interface. It includes tabs for Mining Structure, Mining Models, Mining Model Viewer, Mining Accuracy Chart, and Mining Model Prediction. Below the tabs, there are buttons for Input Selection, Lift Chart, Classification Matrix, and Cross Validation. A section titled "Select predictable mining model columns to show in the lift chart:" contains a checked checkbox for "Synchronize Prediction Columns and Values". A table lists "Show" columns (Severity_Tree, Severity_Cluster), "Predictable Column Name" (Severity), and "Predict Value" (1). Both rows have checked checkboxes in the "Show" column.

Show	Mining Model	Predictable Column Name	Predict Value
<input checked="" type="checkbox"/>	Severity_Tree	Severity	1
<input checked="" type="checkbox"/>	Severity_Cluster	Severity	1

Ta nhấp vào **Lift Chart** để so sánh 2 thuật toán cây quyết định và gom cụm.



Nhận xét:

- ✓ Thuật toán Cây quyết định có số điểm là 0.75 với hiệu suất chính xác là 73,75% và dự đoán cụ thể là 24,58%.
- ✓ Thuật toán Gom cụm có số điểm là 0.64 với hiệu suất chính xác là 57,36% và dự đoán cụ thể là 24,1%.
- Chọn **Predict Value** là 2 để đánh giá độ chính xác trong trường hợp mức nghiêm trọng là 2

Mining Structure Mining Models Mining Model Viewer Mining Accuracy Ch... Mining Model Prediction

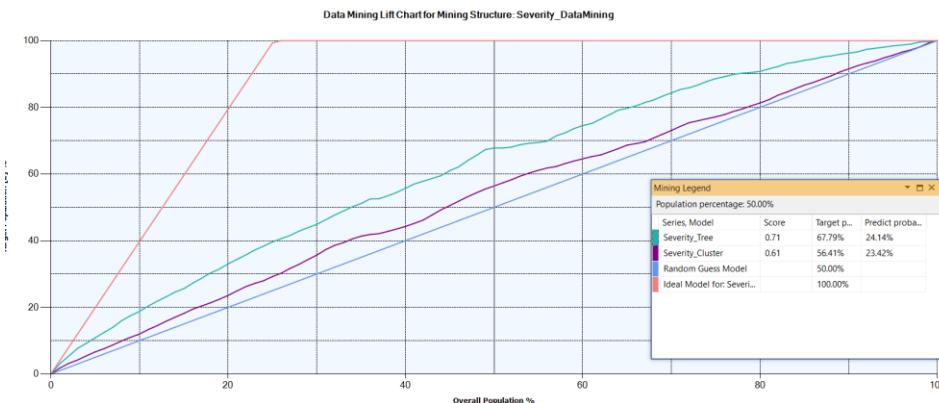
Input Selection Lift Chart Classification Matrix Cross Validation

Select predictable mining model columns to show in the lift chart:

Synchronize Prediction Columns and Values

Show	Mining Model	Predictable Column Name	Predict Value
<input checked="" type="checkbox"/>	Severity_Tree	Severity	2
<input checked="" type="checkbox"/>	Severity_Cluster	Severity	2

Ta nháp vào **Lift Chart** để so sánh 2 thuật toán cây quyết định và gom cụm.



Nhận xét:

- ✓ Thuật toán Cây quyết định có số điểm là 0.71 với hiệu suất chính xác là 67,79% và dự đoán cụ thể là 24,14%.
- ✓ Thuật toán Gom cụm có số điểm là 0.61 với hiệu suất chính xác là 56,41% và dự đoán cụ thể là 23,42%.
- Chọn Predict Value là 3 để đánh giá độ chính xác trong trường hợp mức nghiêm trọng là 3

Mining Structure Mining Models Mining Model Viewer Mining Accuracy Ch...

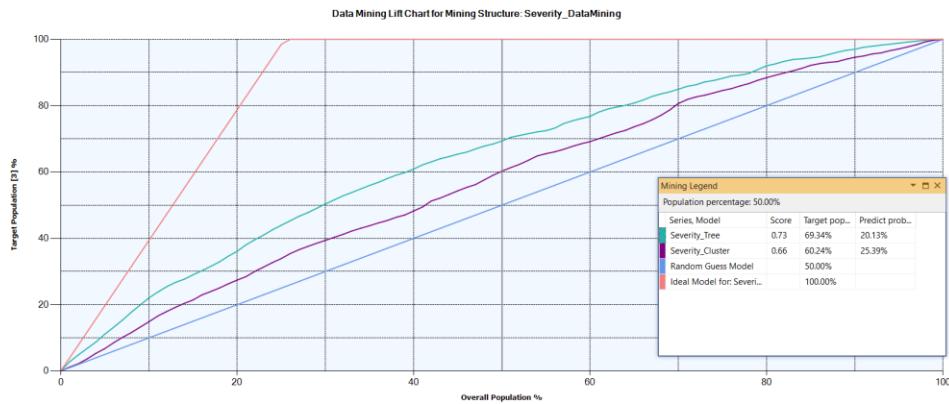
Input Selection Lift Chart Classification Matrix Cross Validation

Select predictable mining model columns to show in the lift chart:

Synchronize Prediction Columns and Values

Show	Mining Model	Predictable Column Name	Predict Value
<input checked="" type="checkbox"/>	Severity_Tree	Severity	3
<input checked="" type="checkbox"/>	Severity_Cluster	Severity	3

Ta nháp vào **Lift Chart** để so sánh 2 thuật toán cây quyết định và gom cụm.



Nhận xét:

- ✓ Thuật toán Cây quyết định có số điểm là 0.73 với hiệu suất chính xác là 69,34% và dự đoán cụ thể là 20,13%.
- ✓ Thuật toán Gom cụm có số điểm là 0.66 với hiệu suất chính xác là 60,24% và dự đoán cụ thể là 25,39%.
- Chọn **Predict Value** là 4 để đánh giá độ chính xác trong trường hợp mức nghiêm trọng là 4

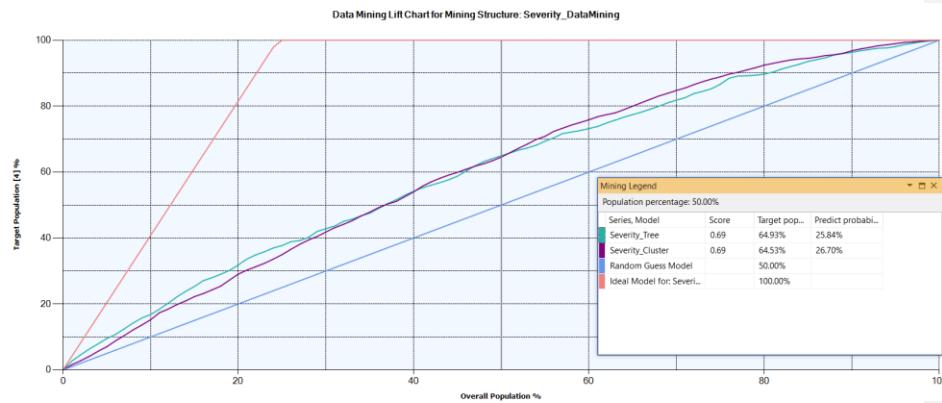
Input Selection Lift Chart Classification Matrix Cross Validation

Select predictable mining model columns to show in the lift chart:

Synchronize Prediction Columns and Values

Show	Mining Model	Predictable Column Name	Predict Value
<input checked="" type="checkbox"/>	Severity_Tree	Severity	4
<input checked="" type="checkbox"/>	Severity_Cluster	Severity	4

Ta nhấp vào **Lift Chart** để so sánh 2 thuật toán cây quyết định và gom cụm.



Nhận xét:

- ✓ Thuật toán Cây quyết định có số điểm là 0.69 với hiệu suất chính xác là 64,93% và dự đoán cụ thể là 25,84%.
- ✓ Thuật toán Gom cụm có số điểm là 0.69 với hiệu suất chính xác là 64,53 % và dự đoán cụ thể là 26,7%.

==> Tóm lại, để dự đoán mức nghiêm trọng của các công trình thi công đường bộ ở Mỹ thì dùng thuật toán cây quyết định (Decision Trees) để đạt được độ chính xác cao nhất.

Ths. Đỗ Thị Minh Phụng

TÀI LIỆU THAM KHẢO

1. Báo cáo đồ án tham khảo Kho dữ liệu và OLAP, tác giả: Đỗ Hữu Lương – Lê Minh Quốc.
2. Báo cáo đồ án tham khảo Kho dữ liệu và OLAP, tác giả: Nguyễn Trần Hà Đức – Nguyễn Tuấn Anh.
3. Báo cáo đồ án tham khảo Kho dữ liệu và OLAP, tác giả: Nguyễn Thị Mỹ Trân.
4. Báo cáo đồ án tham khảo Kho dữ liệu và OLAP, tác giả: Nguyễn Minh Nhựt – Trần Quang Phúc.

LINK DRIVE đồ án nhóm: <https://drive.google.com/drive/folders/1KLjWFxYSNyMP7EJ-CcImk0vi9ilG6oz8?usp=sharing>