

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



**ĐÒ ÁN MÔN HỌC**  
**KHO DỮ LIỆU OLAP**  
**ĐỀ TÀI:**  
**PHÂN TÍCH BỘ DỮ LIỆU TAI NẠN GIAO THÔNG**  
**Giảng viên hướng dẫn: ThS. Nguyễn Thị Kim Phụng**

**Lớp: IS217.N22.HTCL**

**Danh sách thành viên**

STT	Họ và tên	MSSV
1	Lê Thị Ngọc Hảo	20521292
2	Hồ Thị Hằng	20521285

**TP. Hồ Chí Minh, tháng 5 năm 2023**

## LỜI CẢM ƠN

Đầu tiên, nhóm tác giả xin chân thành cảm ơn quý thầy cô Trường Đại học Công nghệ Thông tin – Đại học Quốc gia TP.HCM và quý thầy cô khoa Hệ thống thông tin đã giúp cho nhóm có những kiến thức nền tảng quan trọng để thực hiện đề tài này.

Nhóm tác giả xin gửi lời cảm ơn và lòng biết ơn sâu sắc tới ThS. Nguyễn Thị Kim Phụng - giảng viên môn Kho dữ liệu và OLAP. Cô đã tận tình giảng dạy, hướng dẫn, giúp đỡ, cung cấp tài liệu và tạo điều kiện cho nhóm tác giả hoàn thành đồ án này.

Trong thời gian một học kỳ thực hiện đề tài, nhóm tác giả đã cố gắng vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới để ứng dụng vào thực hiện đề tài này. Mặc dù đã hết sức cố gắng, nhưng chắc chắn đồ án vẫn còn những thiếu sót. Chính vì vậy, nhóm rất mong nhận được những sự góp ý quý báu từ Cô nhằm hoàn thiện những kiến thức mà nhóm tác giả đã học tập và là hành trang để nhóm tác giả thực hiện tiếp các đề tài khác trong tương lai.

Xin chân thành cảm ơn Cô!

Nhóm sinh viên thực hiện

## NHẬN XÉT CỦA GIẢNG VIÊN

## I. Phân tích dữ liệu đồ án môn học

**Đề tài:** Phân tích dữ liệu tai nạn tại Hoa Kỳ

### Chương 1: Giới thiệu và tổng quan

#### 1.1 Lý do chọn đề tài:

Tai nạn giao thông là một trong những vấn đề nóng bỏng nhất hiện nay, nó được xem là một trong những thảm họa lớn nhất đe dọa đến sinh mạng và sức khỏe của con người. Hậu quả của tai nạn giao thông rất nặng nề, không chỉ ảnh hưởng về mặt tinh thần mà còn dễ dẫn đến nghèo đói, lạc hậu, bệnh tật bởi có tới 70% số vụ, số người tử vong là đối tượng thanh niên, trụ cột trong gia đình. Hậu quả của tai nạn giao thông là không kể hết khi nó tác động và gây tổn thương đến toàn xã hội và gia đình người bị nạn. Song, tai nạn giao thông vẫn đang ngày càng gia tăng và ngày càng trở nên nghiêm trọng trong những năm gần đây.

**Số người chết tăng nhiều nhất giai đoạn 2020 - 2021 đến từ các vụ va chạm giữa nhiều ô tô, va chạm trên đường đô thị và các vụ tai nạn liên quan đến người lái xe trên 65 tuổi.**

Cơ quan Quản lý An toàn Giao thông Đường cao tốc Quốc gia thông báo, vào năm ngoái, tỷ lệ tử vong do tai nạn giao thông ở Mỹ tăng lên mức cao nhất trong vòng 16 năm. Uớc tính có hơn 42.900 người chết vì tai nạn ô tô, tăng 10,5% so với năm 2020 và là mức cao nhất kể từ năm 2005. Số người chết tăng nhiều nhất giai đoạn 2020 - 2021 đến từ các vụ va chạm giữa nhiều ô tô, va chạm trên đường đô thị và các vụ tai nạn liên quan đến người lái xe trên 65 tuổi.

Việc phân tích nguồn dữ liệu về tai nạn giao thông sẽ giúp xác định sự ảnh hưởng xấu của tai nạn đến cộng đồng, cũng như là những địa điểm thường xuyên xảy ra sự cố, ở các ảnh hưởng bên ngoài như thời tiết,.. Từ đó xác định và ngăn chặn, phát hiện sớm và ngăn chặn sự cố xảy ra trong tương lai

Vì vậy, nhóm chúng em quyết định tiến hành nghiên cứu, phân tích và thiết kế kho dữ liệu nhằm: nghiên cứu vị trí điểm nóng về tai nạn, phân tích và rút ra các quy tắc, nguyên nhân và kết quả để dự đoán tai nạn, hoặc nghiên cứu tác động của lượng mưa hoặc các kích thích môi trường khác đối với sự cố xảy ra tai

#### 1.2 Mô tả bộ dữ liệu

##### 1.2.1 Giới thiệu nguồn dữ liệu

Tên dataset: US Accidents (Dữ liệu ứng dụng tai nạn tại Hoa Kỳ).

Ngày cập nhật gần nhất: 12/03/2022.

Đơn vị cung cấp: Sobhan Moosavi - một nhà khoa học Hoa Kỳ

Đường dẫn: <https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>

Bộ dữ liệu gồm 2,845,342 dòng dữ liệu và 47 cột thuộc tính.

Sau khi lọc dữ liệu ta được 1,394,113 dòng dữ liệu và 25 thuộc tính để sử dụng cho việc phân tích đề tài.

Bộ dữ liệu được thu thập từ tháng 2 năm 2016 đến tháng 12 năm 2021.

### 1.3 Thiết kế kho dữ liệu

STT	Tên thuộc tính	Kiểu DL	Mô tả
1	ID	String	Đây là số nhận dạng duy nhất của hồ sơ vụ tai nạn.
2	Severity	Int	Cho biết mức độ nghiêm trọng của vụ tai nạn. Có giá trị từ 1 đến 4, trong đó 1 cho biết tác động ít nhất đến giao thông (short delay).
3	Start_Time	DateTime	Hiển thị thời gian bắt đầu vụ tai nạn theo múi giờ địa phương (time_zone). Định dạng ngày: MM / DD / YYYY hh: mm.
4	Distance(mi)	Float	Chiều dài của đoạn đường bị ảnh hưởng bởi vụ tai nạn. (Đo bằng mile, trong đó 1 mile = 1,609344 km = 1.609,344 m)
5	Street	String	Hiển thị tên đường trong bản ghi địa chỉ.
6	Side	String	Hiển thị phía tương đối của đường (Phải / Trái) trong bản ghi địa chỉ.
7	City	String	Hiển thị thành phố trong bản ghi địa chỉ

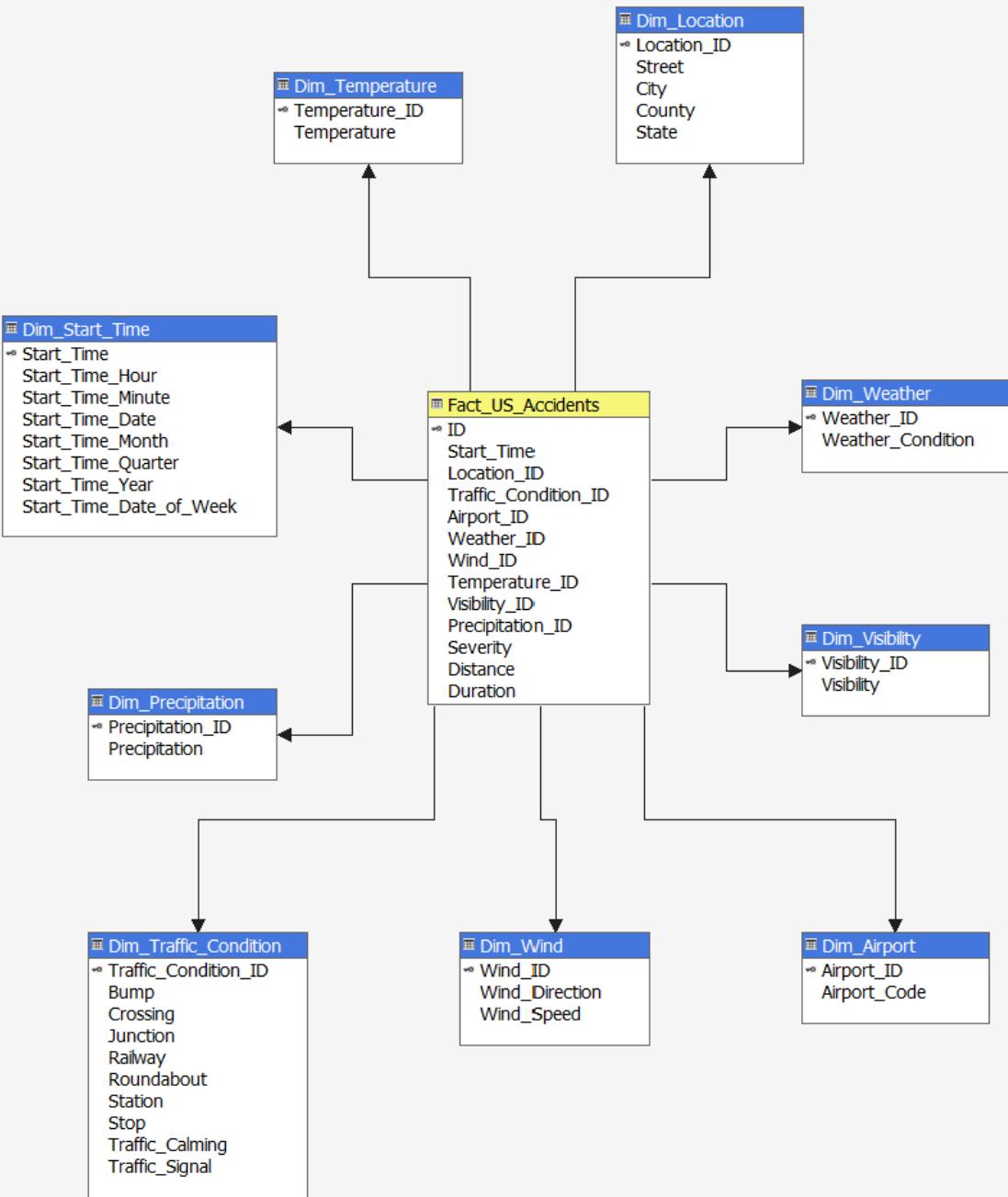
8	County	String	Hiển thị quận trong hồ sơ địa chỉ.
9	State	String	Hiển thị tiểu bang trong bản ghi địa chỉ.
10	Airport_Code	String	Biểu thị một trạm thời tiết ở sân bay và là trạm gần nhất với vị trí xảy ra tai nạn.
11	Temperature(F)	Float	Hiển thị nhiệt độ (tính bằng °F).
12	Visibility(mi)	Float	Cho biết khả năng nhìn xa (tính bằng dặm)
13	Wind_Direction	String	Hiển thị hướng gió. Các giá trị có thể có là: – Calm: gió lặng – E, East: có gió hướng đông – ENE, NNE, NNW, NW....
14	Wind_Speed(mph)	Float	Hiển thị tốc độ gió (tính bằng dặm / giờ).
15	Precipitation(in)	Float	Hiển thị lượng mưa (tính bằng inch), nếu có.
16	Weather_Condition	String	Hiển thị tình trạng thời tiết (mưa, tuyết, giông, sương mù, v.v.)
17	Bump	Boolean	Cho biết sự hiện diện của gờ giảm tốc ở một vị trí gần đó.
18	Crossing	Boolean	Cho biết sự hiện diện của lối qua đường cho người đi bộ ở một địa điểm gần đó.
19	Junction	Boolean	Cho biết sự hiện diện của đường giao nhau ở một vị trí gần đó
20	Railway	Boolean	Cho biết sự hiện diện của một đường sắt ở gần đó

21	Roundabout	Boolean	Cho biết sự hiện diện của một bùng binh ở gần đó
22	Station	Boolean	Cho biết sự hiện diện của một trạm ga điện ở gần đó.
23	Stop	Boolean	Cho biết sự hiện diện của một điểm dừng ở gần đó
24	Traffic_Calming	Boolean	Cho biết sự hiện diện của điều tiết giao thông đang ở gần đó.
25	Traffic_Signal	Boolean	Cho biết sự hiện diện của đèn tín hiệu giao thông ở vị trí gần đó

Bảng 1 Mô tả chi tiết các thuộc tính

#### 1.4 Thiết kế kho dữ liệu

##### 1.4.1 Lược đồ hình sao



#### 1.4.2 Mô tả chi tiết bảng Fact

STT	Tên thuộc tính	Kiểu DL	Ràng buộc	Mô tả
1	Fact_ID	String	PK	Số hồ sơ vụ án

2	Start_Time	DateTime	FK	Thời gian bắt đầu vụ án
3	Location_ID	Int	FK	Mã địa điểm xảy ra vụ án
4	Airport_ID	Int	FK	Mã bản ghi thời tiết do trạm thời tiết gần nhất với vị trí xảy ra vụ án ghi lại
5	Traffic_Condition_ID	Int	FK	Mã tình trạng giao thông nơi xảy ra vụ án
6	Weather_ID	Int	FK	Mã bảng ghi thông tin về điều kiện thời tiết
7	Wind_ID	Int	FK	Mã bảng ghi thông tin về gió
8	Temperature_ID	Int	FK	Mã bảng ghi thông tin về nhiệt độ
9	Visibility_ID	Int	FK	Mã bảng ghi thông tin về tầm nhìn xa
10	Precipitation_ID	Int	FK	Mã bảng ghi thông tin về lượng mưa
11	Severity	Int		Mức độ nghiêm trọng của vụ án
12	Distance(m)	Int		Chiều dài của đoạn đường bị ảnh hưởng bởi vụ tai nạn
13	Duration*	Int		Khoảng thời gian vụ tai nạn kéo dài, được tính bằng giây

Bảng 2 Mô tả chi tiết bảng Fact

\*Duration: được tính toán bằng Python trước khi bắt đầu quá trình SSIS. Duration được tính bằng cách sử dụng bộ dữ liệu gốc, lấy thời gian kết thúc vụ tai nạn (End\_Time) trừ cho thời gian bắt đầu vụ tai nạn (Start\_Time). Sau đó, dùng hàm total\_seconds() để lấy ra khoảng thời gian đó (tính bằng giây).

### 1.4.3 Mô tả chi tiết bảng Dimension

#### a. Bảng Dim\_Time

STT	Tên thuộc tính	Kiểu DL	Ràng buộc	Mô tả
1	Start_Time	DateTime	PK	Hiển thị thời gian bắt đầu vụ tai nạn theo múi giờ địa phương (time_zone). Định dạng ngày: MM / DD / YYYY hh:mm.
2	Start_Time_Hour	Int		Lấy ra giờ trong thời gian bắt đầu vụ tai nạn.
3	Start_Time_Minute	Int		Lấy ra phút trong thời gian bắt đầu vụ tai nạn.
4	Start_Time_Day	Int		Lấy ra ngày trong thời gian bắt đầu vụ tai nạn.
5	Start_Time_Month	Int		Lấy ra tháng trong thời gian bắt đầu vụ tai nạn.
6	Start_Time_Year	Int		Lấy ra năm trong thời gian bắt đầu vụ tai nạn.
7	Start_Time_Quarter	Int		Lấy ra quý trong thời gian bắt đầu vụ tai nạn.
8	Start_Time_Day_of_Week	Int		Lấy ra ngày trong tuần trong thời gian bắt đầu vụ tai nạn.

Bảng 3 Mô tả chi tiết bảng DimTime

#### b. Bảng Dim\_Location

<b>STT</b>	<b>Tên thuộc tính</b>	<b>Kiểu DL</b>	<b>Ràng buộc</b>	<b>Mô tả</b>
1	Location_ID	Int	<b>PK</b>	Mã địa điểm xảy ra tai nạn
2	Street	String		Hiển thị tên đường trong bản ghi địa chỉ
3	City	String		Hiển thị thành phố trong bản ghi địa chỉ
4	Country	String		Hiển thị quốc trong bản ghi địa chỉ
5	State	String		Hiển thị tiểu bang trong bản ghi địa chỉ

*Bảng 4 Mô tả chi tiết bảng Dim Location*

c. *Bảng Dim\_Traffic\_Condition*

<b>STT</b>	<b>Tên thuộc tính</b>	<b>Kiểu DL</b>	<b>Ràng buộc</b>	<b>Mô tả</b>
1	Traffic_Condition_ID	Int	<b>PK</b>	Mã tình trạng giao thông nơi xảy ra tai nạn.
2	Bump	Boolean		Cho biết sự hiện diện của gờ giảm tốc ở một vị trí gần đó.
3	Crossing	Boolean		Cho biết sự hiện diện của lối qua đường cho người đi bộ ở một địa điểm gần đó.
4	Junction	Boolean		Cho biết sự hiện diện của đường giao nhau ở một vị trí gần đó.
5	Railway	Boolean		Cho biết sự hiện diện của một đường sắt ở gần đó

6	Roundabout	Boolean		Cho biết sự hiện diện của một bùng binh ở gần đó
7	Station	Boolean		Cho biết sự hiện diện của một trạm ga điện ở gần đó.
8	Stop	Boolean		Cho biết sự hiện diện của một điểm dừng ở gần đó.
9	Tracffic_Calming	Boolean		Cho biết sự hiện diện của điều tiết giao thông đang ở gần đó.
10	Trafic_Signal	Boolean		Cho biết sự hiện diện của đèn tín hiệu giao thông ở vị trí gần đó.

Bảng 5 Mô tả chi tiết bảng Dim Traffic Condition

d. Bảng Dim\_Airport

STT	Tên thuộc tính	Kiểu DL	Ràng buộc	Mô tả
1	Airport_Record_ID	Int	PK	Mã bản ghi thời tiết do trạm thời tiết gần nhất với vị trí xảy ra tai nạn ghi lại.
2	Airport_Code	String		Biểu thị một trạm thời tiết ở sân bay và là trạm gần nhất với vị trí xảy ra tai nạn.

Bảng 6 Mô tả chi tiết bảng Dim Airport

e. Bảng Dim\_Wind

STT	Tên thuộc tính	Kiểu DL	Ràng buộc	Mô tả
1	Wind_ID	Int	<b>PK</b>	Mã bản ghi thông tin về gió
2	Wind_Direction	String		Hiển thị hướng gió. Các giá trị có thể có là: – Calm: gió lặng – E, East: có gió hướng đông – ENE, NNE, NNW, NW....
3	Wind_Speed(mph)	Float		Hiển thị tốc độ gió (tính bằng dặm / giờ).

Bảng 7 Mô tả chi tiết bảng Dim Wind

f. Bảng Dim\_Temperature

STT	Tên thuộc tính	Kiểu DL	Ràng buộc	Mô tả
1	Temperature_ID	Int	<b>PK</b>	Mã bản ghi thông tin nhiệt độ
2	Temperature(F)	Float		Hiển thị nhiệt độ (tính bằng °F)

Bảng 8 Mô tả chi tiết bảng Dim Temperature

g. Bảng Dim\_Visibility

STT	Tên thuộc tính	Kiểu DL	Ràng buộc	Mô tả
1	Visibility_ID	Int	<b>PK</b>	Mã bản ghi thông tin tầm nhìn xa
2	Visibility	Float		Cho biết khả năng nhìn xa (tính bằng dặm)

Bảng 9 Mô tả chi tiết bảng Dim Visibility

h. Bảng Dim\_Weather

STT	Tên thuộc tính	Kiểu DL	Ràng buộc	Mô tả
1	Weather_ID	Int	<b>PK</b>	Mã bản ghi thông tin điều kiện thời tiết
2	Weather_Condition	String		Hiển thị tình trạng thời tiết (mưa, tuyết, giông, sương mù, v.v.)

Bảng 10 Mô tả chi tiết bảng Dim Weather

### 1.5 Các câu truy vấn

**Câu 1:** Liệt kê 10 quận xảy ra nhiều vụ tai nạn nhất và sắp xếp theo thứ tự tăng dần.

**Câu 2:** Thống kê số vụ tai nạn xảy ra tại đường ray xe lửa ở các bang từ năm 2020 đến 2021

**Câu 3:** Cho biết số lượng các vụ tai nạn xảy ra của bốn hướng gió: Đông, Tây, Nam, Bắc thuộc các thành phố tại các tiểu bang với tốc độ gió lớn hơn 10 dặm/ giờ.

**Câu 4:** Thống kê theo năm, top 5 loại thời tiết có nhiều vụ tai nạn xảy ra nhất trong năm 2020 và năm 2021

**Câu 5:** Thống kê tổng chiều dài của đoạn đường bị ảnh hưởng bởi các vụ tai nạn xảy ra tại nơi vừa có tín hiệu giao thông (Traffic Signal), vừa có trạm xe (Station) theo năm, quý, tiểu bang, quận, thành phố trong năm 2020, 2021.

**Câu 6:** Thống kê theo năm, tháng tổng số vụ tai nạn, chỉ số nghiêm trọng mà bị ảnh hưởng tại những con đường có tai nạn gờ giảm tốc khi nhiệt độ trong ngày trong khoảng -10 F đến 100 F.

**Câu 7:** Liệt kê tổng số vụ tai nạn, mức độ nghiêm trọng trung bình, và thời gian trung bình của các vụ tai nạn giao thông theo quý trong năm.

**Câu 8:** Cho biết tên con đường có số vụ tai nạn giao thông cao nhất và tên con đường có chỉ số nghiêm trọng tai nạn giao thông cao nhất.

**Câu 9:** Lấy ra 3 tiêu bang xảy ra nhiều vụ tai nạn nhất với mỗi quý trong năm 2020.

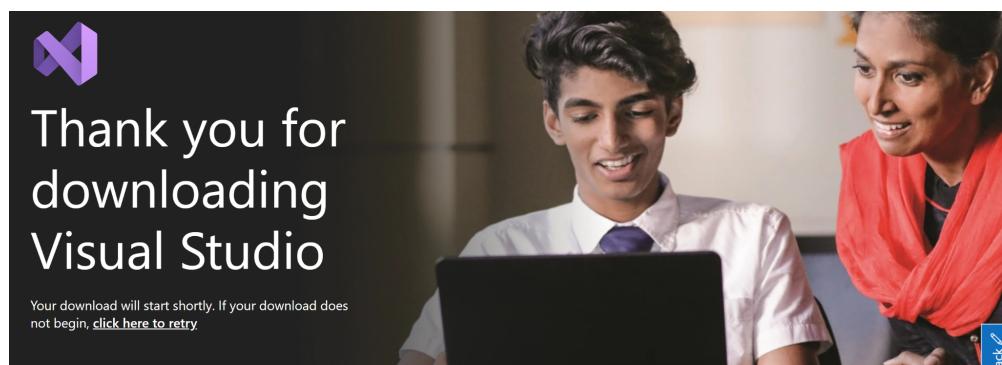
**Câu 10:** Liệt kê chiều dài lớn nhất của đoạn đường và mức độ nghiêm trọng bị ảnh hưởng bởi các vụ tai nạn giao thông có tầm nhìn xa trên 2 dặm tại tiểu bang California (CA) và Washington (WA)

## **CHƯƠNG 2: QUÁ TRÌNH TRÍCH XUẤT DỮ LIỆU, BIẾN ĐỔI VÀ NẠP DỮ LIỆU VÀO KHO DỮ LIỆU (QUÁ TRÌNH SSIS)**

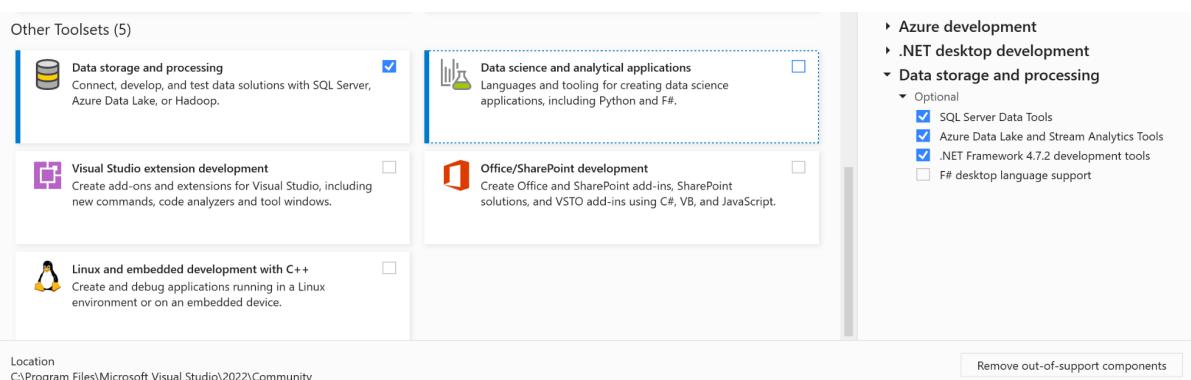
### **2.1 Chuẩn bị công cụ và Data Warehouse :**

#### **2.1.1 Công cụ SQL Server Data Tools:**

**Bước 1:** Truy cập vào Visual Studio, sau đó download



**Bước 2:** Sau khi đã tải xong, kéo xuống mục **Other Toolset (5)**, chọn **Data storage and processing** và chọn **Optional** là **SQL Server Data Tools**



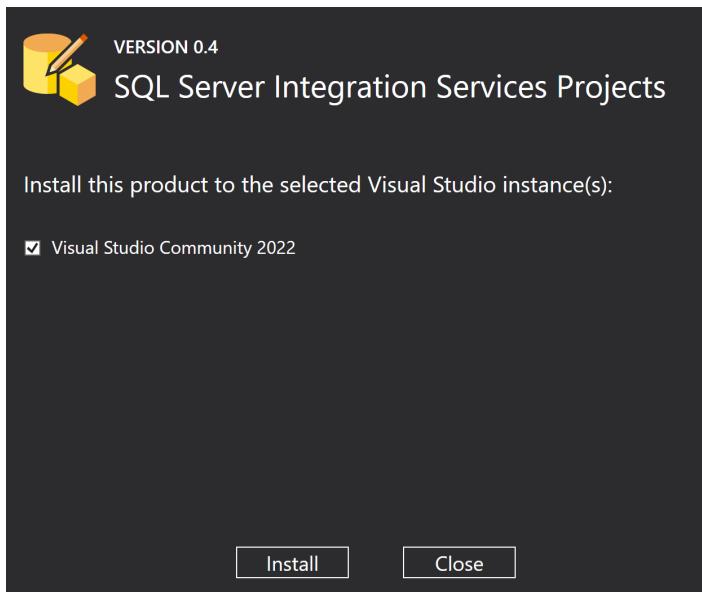
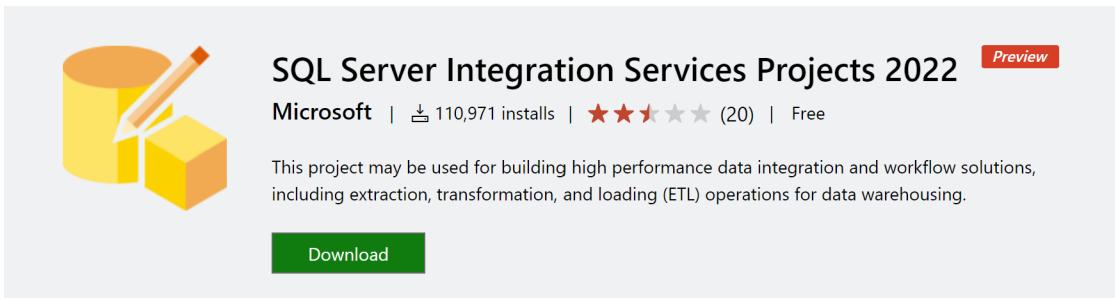
### **2.2 Tạo project và thiết lập kết nối**

## 2.2.1 Tạo project “Integration Service Project” mới

**Bước 1:** Tải công cụ SQL Server Integration Service Projects 2022

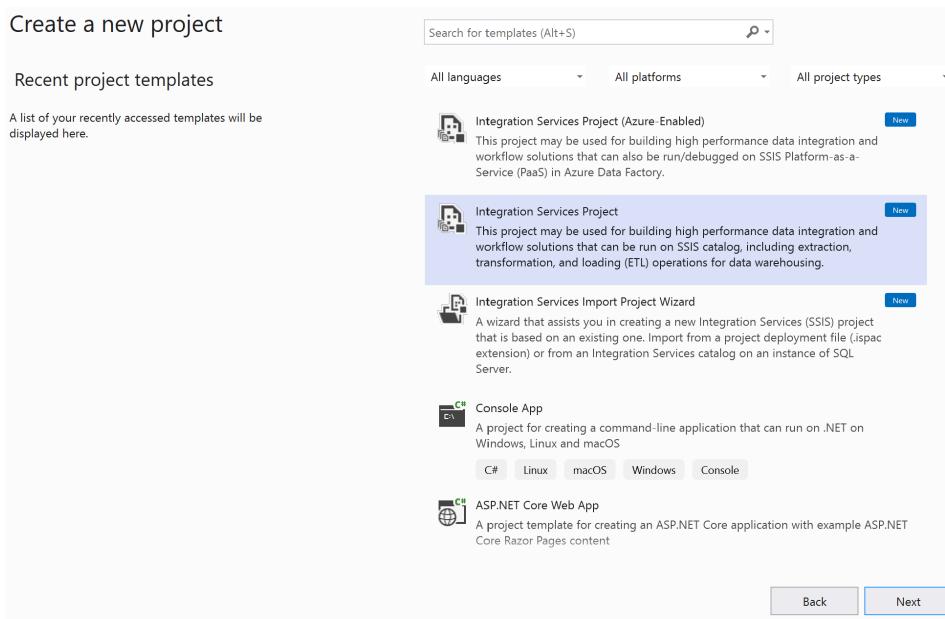
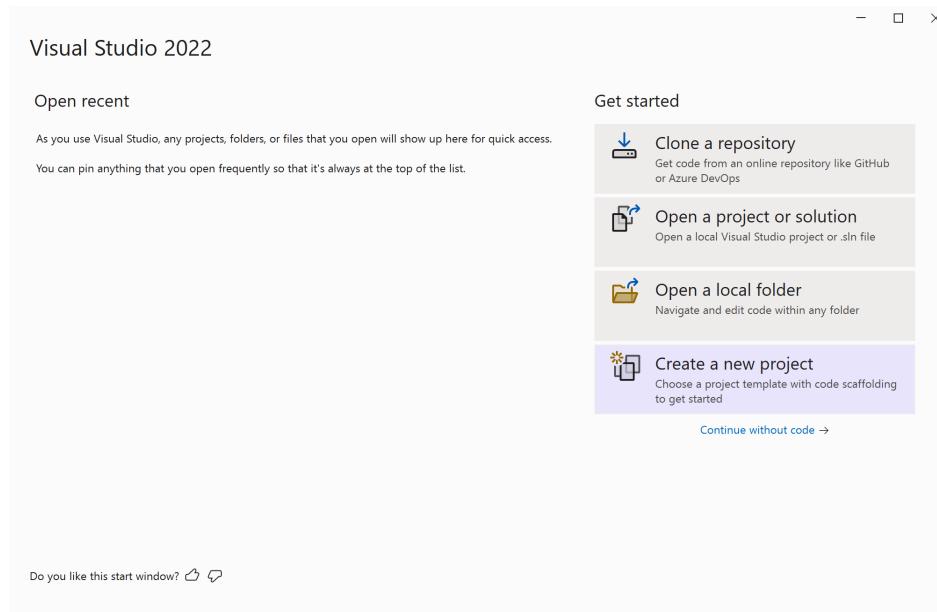
**Đường link:**

<https://marketplace.visualstudio.com/items?itemName=SSIS.MicrosoftDataToolsIntegrationServices>

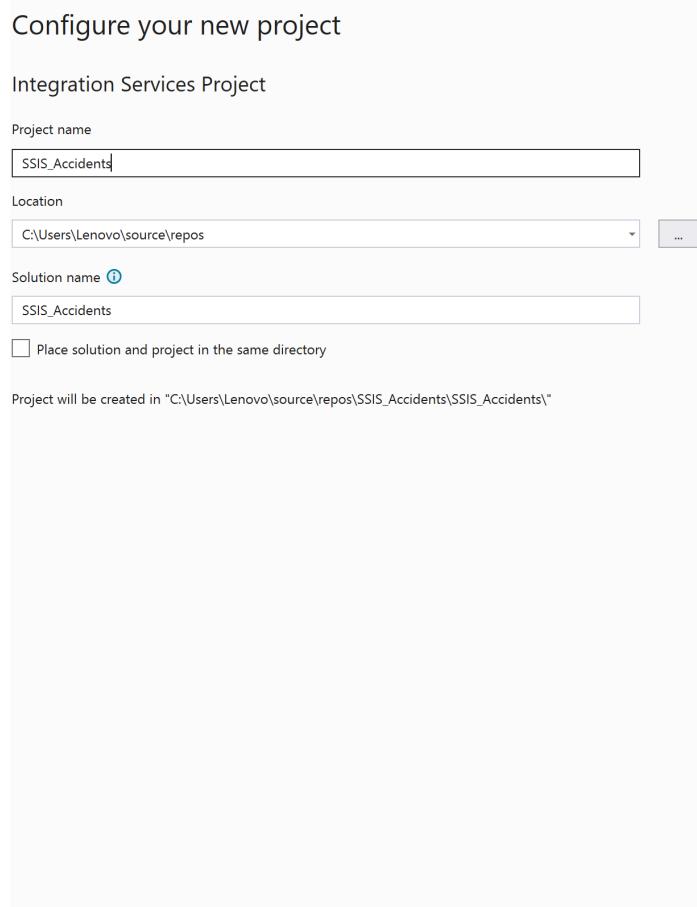


Sau khi cài đặt thành công, thông báo sẽ hiện trên màn hình và chọn OK

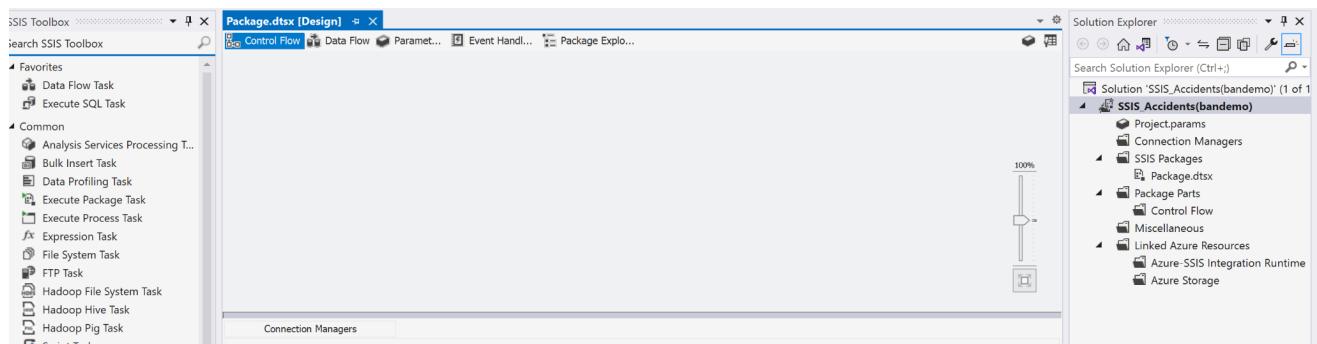
**Bước 2:** Mở Visual Studio. Chọn **create a new project**, sau đó chọn **Integration Services Project**



### Bước 3: Tạo một dự án mới

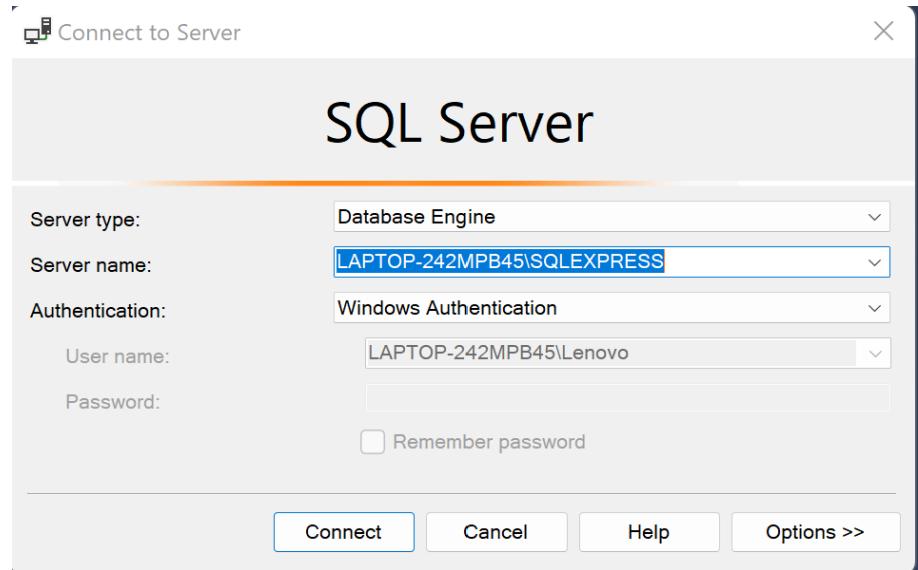


Sau khi tạo thành công, giao diện hiện ra như sau:

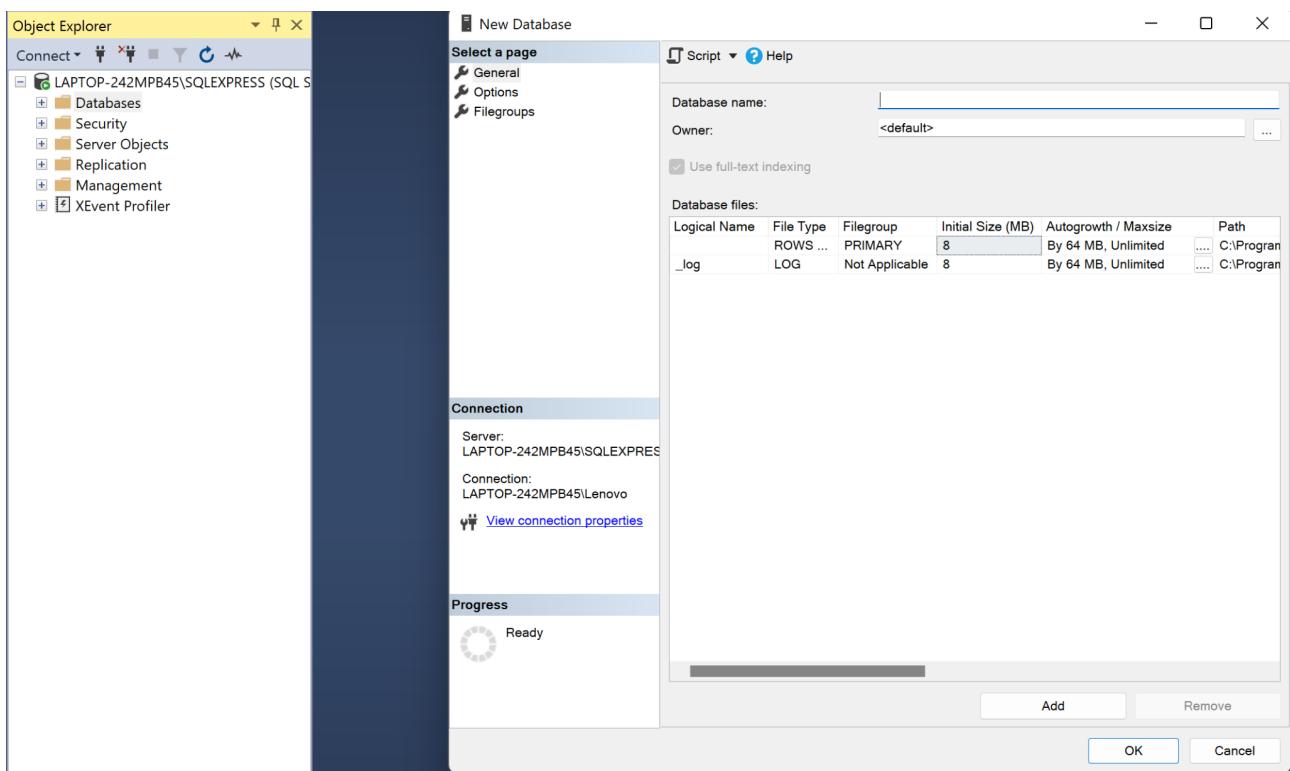


## 2.1.2 Tạo kho dữ liệu OLAP

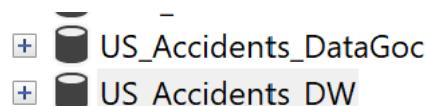
**Bước 1:** Kết nối SQL Server



**Bước 2:** Click chuột trái vào Databases ở cột bên trái để tạo database mới. Sau đó tạo Database US\_Accidents\_DataGoc và database US\_Accidents\_DW



+ Đã tạo xong 2 database cho đê tài



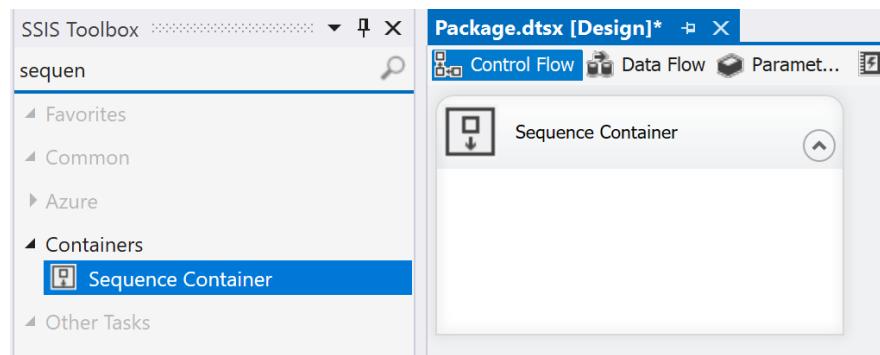
Trong đó:

**US\_Accidents\_DataGoc:** Nơi lưu dữ liệu gốc đã được làm sạch. Dữ liệu này được đổ từ file .csv tải xuống tại Kaggle

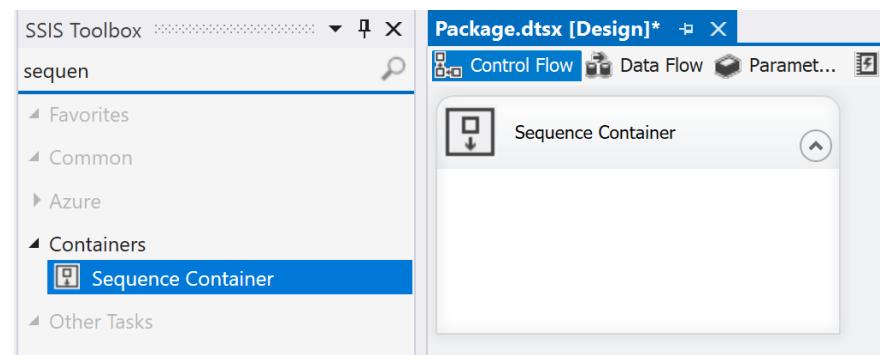
**US\_Accidents\_DW:** Nơi lưu các bảng Dim và bảng Fact cùng dữ liệu của các bảng đó. Dữ liệu này được đổ từ US\_Accidents\_DataGoc vào.

### 2.1.3 Các bước tiến hành

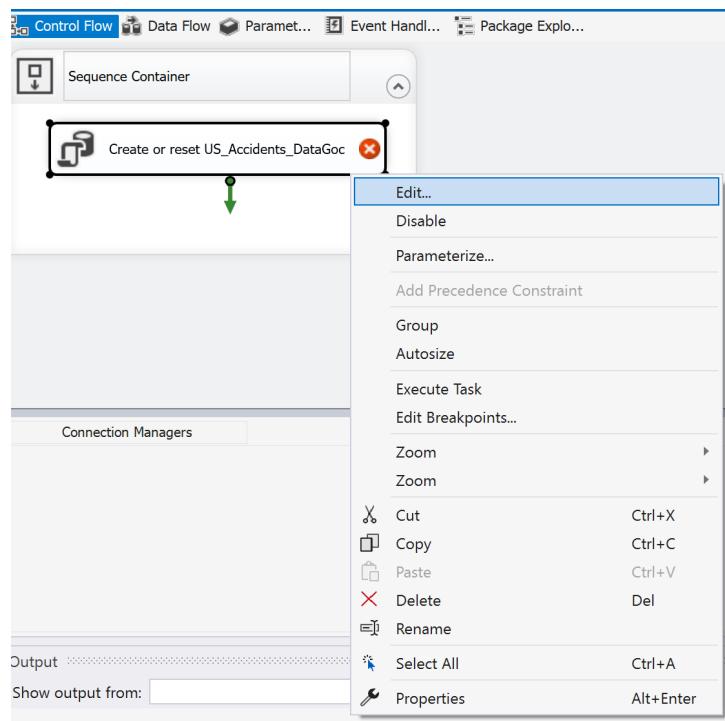
**Bước 1:** Tạo ra một Sequence Container và đặt tên là Setup database



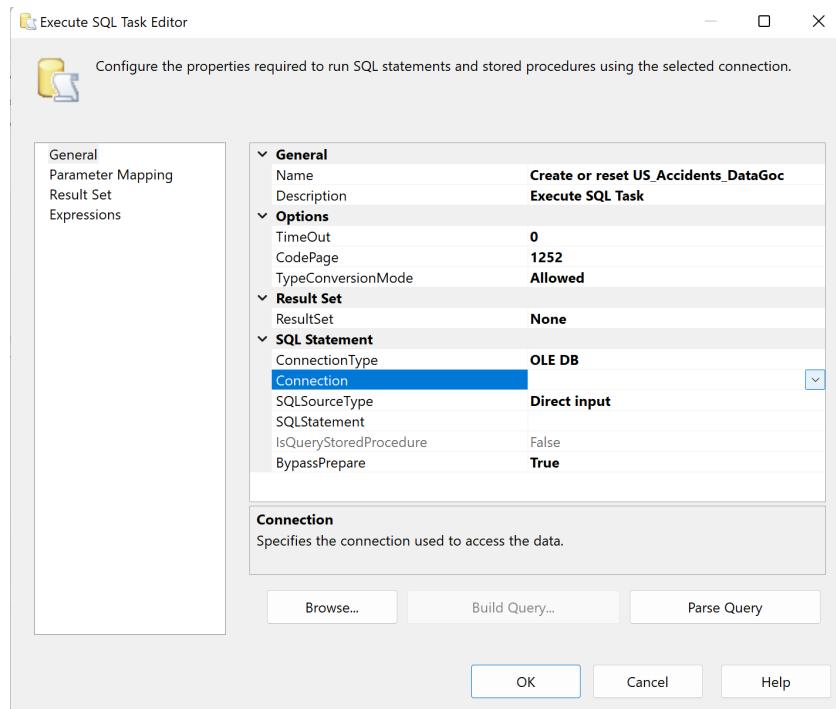
**Bước 2:** Trong Sequence Container vừa tạo, tạo ra một Execute SQL Task và đặt tên là “Create or reset US\_Accidents\_DataGoc”. Execute SQL Task này có chức năng thực hiện các câu lệnh SQL trên SQL Server để khởi tạo hoặc reset cơ sở dữ liệu lưu dữ liệu gốc.



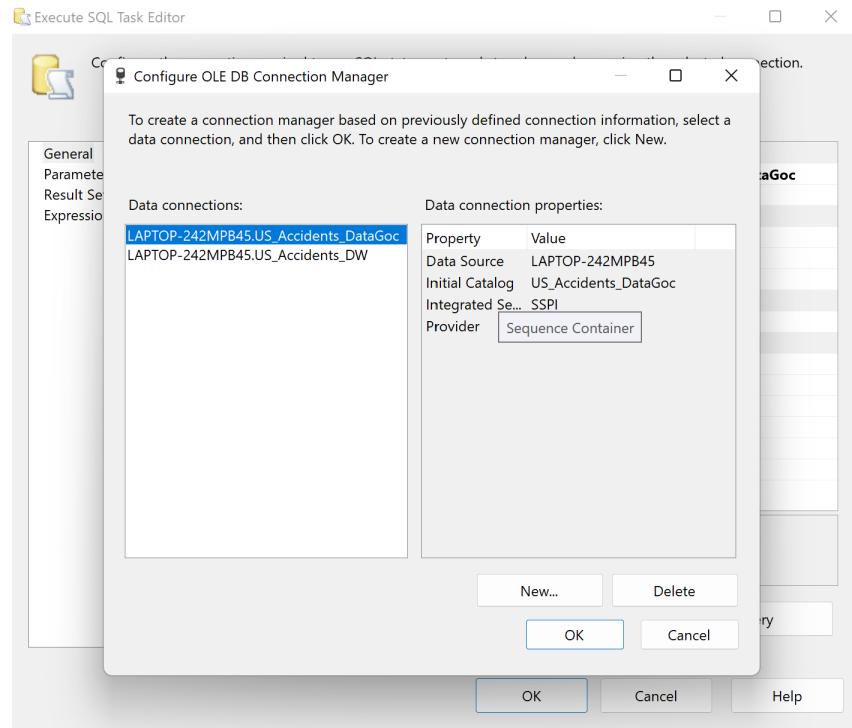
**Bước 3:** Nhấn chuột phải vào Execute SQL Task vừa tạo, chọn Edit.



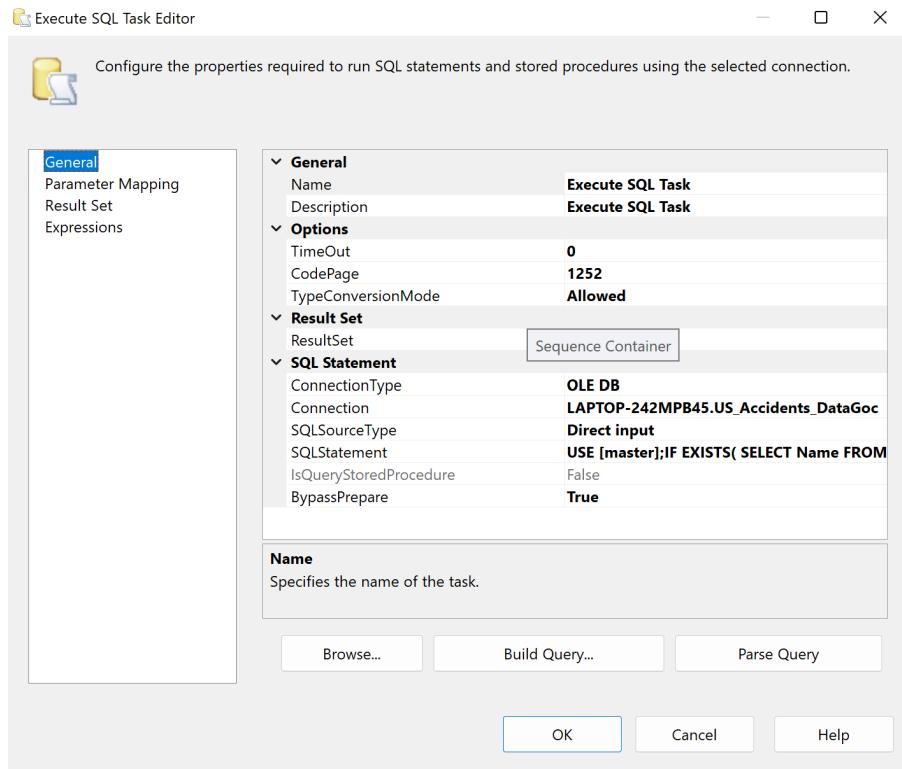
**Bước 4:** Tại hộp thoại Execute SQL Task, ta chọn mục connection  New connection để tạo mới một kết nối đến cơ sở dữ liệu master trên SQL Server



**Bước 5:** Tiếp tục chọn New... để tạo data connection.



**Bước 6:** Tại hộp thoại Connection Manager, nhập vào Server name và chọn tên cơ sở dữ liệu master. Sau đó, chọn Test connection  OK để hoàn tất tạo connection



**Bước 7:** Ta tiếp tục chọn mục SQL Statement trên hộp thoại Execute SQL Task để thêm các câu lệnh SQL vào.

```

USE [master];
IF EXISTS( SELECT Name FROM SysDatabases WHERE Name = 'US_Accidents_DataGoc' )
BEGIN
    ALTER DATABASE US_Accidents_DataGoc SET SINGLE_USER WITH
    ROLLBACK IMMEDIATE;
    DROP DATABASE US_Accidents_DataGoc;
END
GO
CREATE DATABASE US_Accidents_DataGoc;
GO

USE US_Accidents_DataGoc;
GO

IF(object_id('Data_Clean') IS NOT NULL)
    DROP TABLE Data_Clean;
GO
CREATE TABLE [Data_Clean] (
    [ID] nvarchar(100),
    [Severity] tinyint,
    [Start_Time] datetime2(7),
    [End_Time] datetime2(7),
    [Distance] numeric(6,3),
    [Description] nvarchar(max),
    [Street] nvarchar(150),
    [Side] nvarchar(150)
)

```

Nội dung các câu lệnh SQL trong Execute SQL Task “Create or reset US\_Accidents\_DataGoc” như sau:

```
USE [master];

IF EXISTS( SELECT Name FROM SysDatabases WHERE Name =
'US_Accidents_DataGoc')
BEGIN
    ALTER DATABASE US_Accidents_DataGoc SET SINGLE_USER WITH ROLLBACK
IMMEDIATE;
    DROP DATABASE US_Accidents_DataGoc;
END
GO
CREATE DATABASE US_Accidents_DataGoc;
GO
```

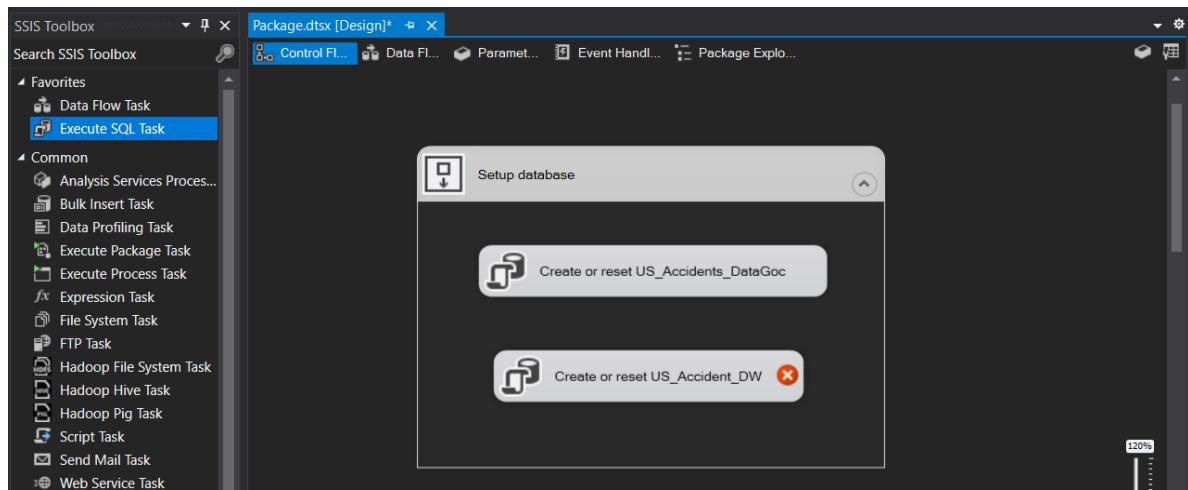
```
USE US_Accidents_DataGoc;
GO
```

```
IF(object_id('Data_Clean') IS NOT NULL)
DROP TABLE Data_Clean;
GO
CREATE TABLE [Data_Clean] (
    [ID] nvarchar(100),
    [Severity] tinyint,
    [Start_Time] datetime2(7),
    [Distance] numeric(6,3),
    [Duration] int,
    [Street] nvarchar(150),
    [City] nvarchar(150),
    [County] nvarchar(150),
    [State] nvarchar(150),
```

```
[Airport_Code] nvarchar(150),  
[Temperature] int,  
[Visibility] numeric(6,3),  
[Wind_Direction] nvarchar(150),  
[Wind_Speed] int,  
[Precipitation] numeric(6,3),  
[Weather_Condition] nvarchar(150),  
[Bump] nvarchar(50),  
[Crossing] nvarchar(50),  
[Junction] nvarchar(50),  
[Railway] nvarchar(50),  
[Roundabout] nvarchar(50),  
[Station] nvarchar(50),  
[Stop] nvarchar(50),  
[Traffic_Calming] nvarchar(50),  
[Traffic_Signal] nvarchar(50)  
)
```

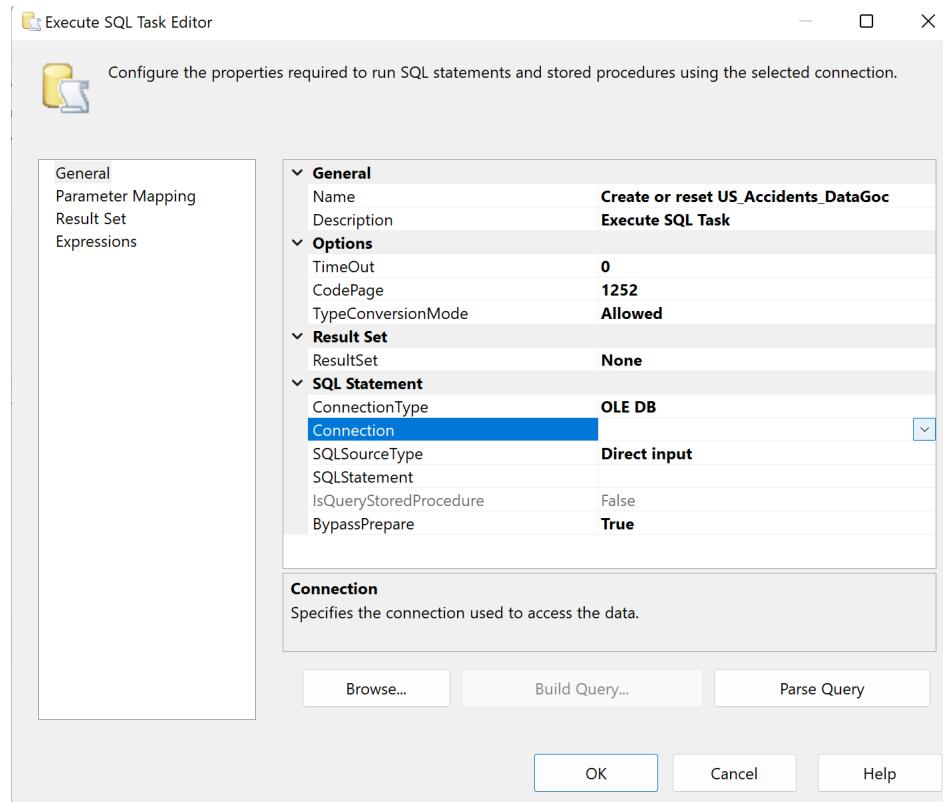
Sau khi đã thêm các câu lệnh SQL vào, ta chọn OK để hoàn tất thiết lập Execute SQL Task.

**Bước 8:** Tương tự như vậy, ta cũng thiết lập một Execute SQL Task để chứa các bảng Dim, bảng Fact và đặt tên là “Create or reset US\_Accident\_DW”.



**Bước 9:** Double click vào Execute SQL Task để thiết lập connection và SQL Statement cho Execute SQL Task vừa tạo.

Chọn connection là connection liên kết đơn cơ sở dữ liệu master vừa được khởi tạo ở các bước trước.



**Bước 10:** Ta chọn mục SQL Statement để thêm các câu lệnh SQL vào.

```
USE [master];
IF EXISTS( SELECT Name FROM SysDatabases WHERE Name = 'US_Accidents_DW')
BEGIN
ALTER DATABASE US_Accidents_DW SET SINGLE_USER WITH ROLLBACK IMMEDIATE;
DROP DATABASE US_Accidents_DW;
END
GO

CREATE DATABASE US_Accidents_DW;
GO

USE US_Accidents_DW;
GO

IF(object_id('Dim_Start_Time') IS NOT NULL)
DROP TABLE Dim_Start_Time;
GO
CREATE TABLE [Dim_Start_Time] (
[Start_Time] datetime2(7) NOT NULL PRIMARY KEY,
[Start_Time_Date] int,
[Start_Time_Month] int,
[Start_Time_Quarter] int,
```

Nội dung các câu lệnh SQL trong Execute SQL Task “Create or reset US\_Accidents\_DW” như sau:

USE [master];

```
IF EXISTS( SELECT Name FROM SysDatabases WHERE Name = 'US_Accidents_DW')
BEGIN
    ALTER DATABASE US_Accidents_DW SET SINGLE_USER WITH ROLLBACK
    IMMEDIATE;
    DROP DATABASE US_Accidents_DW;
END
GO
```

```
CREATE DATABASE US_Accidents_DW;
GO
```

```
USE US_Accidents_DW;
GO
```

```
IF(object_id('Dim_Start_Time') IS NOT NULL)
DROP TABLE Dim_Start_Time;
GO
CREATE TABLE [Dim_Start_Time] (
    [Start_Time] datetime2(7) NOT NULL PRIMARY KEY,
    [Start_Time_Hour] int,
    [Start_Time_Minute] int,
    [Start_Time_Date] int,
    [Start_Time_Month] int,
    [Start_Time_Quarter] int,
    [Start_Time_Year] int,
    [Start_Time_Date_of_Week] int
)
```

```
IF(object_id('Dim_Location') IS NOT NULL)
DROP TABLE Dim_Location;
GO
CREATE TABLE [Dim_Location] (
[Location_ID] int identity (1,1) not null primary key,
[Street] nvarchar(150),
[City] nvarchar(150),
[County] nvarchar(150),
[State] nvarchar(150)
)
```

```
IF(object_id('Dim_Traffic_Condition') IS NOT NULL)
DROP TABLE Dim_Traffic_Condition;
GO
CREATE TABLE [Dim_Traffic_Condition] (
[Traffic_Condition_ID] int identity (1,1) not null primary key,
[Bump] nvarchar(50),
[Crossing] nvarchar(50),
[Junction] nvarchar(50),
[Railway] nvarchar(50),
[Roundabout] nvarchar(50),
[Station] nvarchar(50),
[Stop] nvarchar(50),
[Traffic_Calming] nvarchar(50),
[Traffic_Signal] nvarchar(50)
)
```

```
IF(object_id('Dim_Airport') IS NOT NULL)
DROP TABLE Dim_Airport;
```

GO

```
CREATE TABLE [Dim_Airport] (
    [Airport_ID] int identity (1,1) not null primary key,
    [Airport_Code] nvarchar(150)
)
```

IF(object\_id('Dim\_Weather') IS NOT NULL)

```
DROP TABLE Dim_Weather;
```

GO

```
CREATE TABLE [Dim_Weather] (
    [Weather_ID] int identity (1,1) not null primary key,
    [Weather_Condition] nvarchar(150)
)
```

IF(object\_id('Dim\_Temperature') IS NOT NULL)

```
DROP TABLE Dim_Temperature;
```

GO

```
CREATE TABLE [Dim_Temperature] (
    [Temperature_ID] int identity (1,1) not null primary key,
    [Temperature] int
)
```

IF(object\_id('Dim\_Visibility') IS NOT NULL)

```
DROP TABLE Dim_Visibility;
```

GO

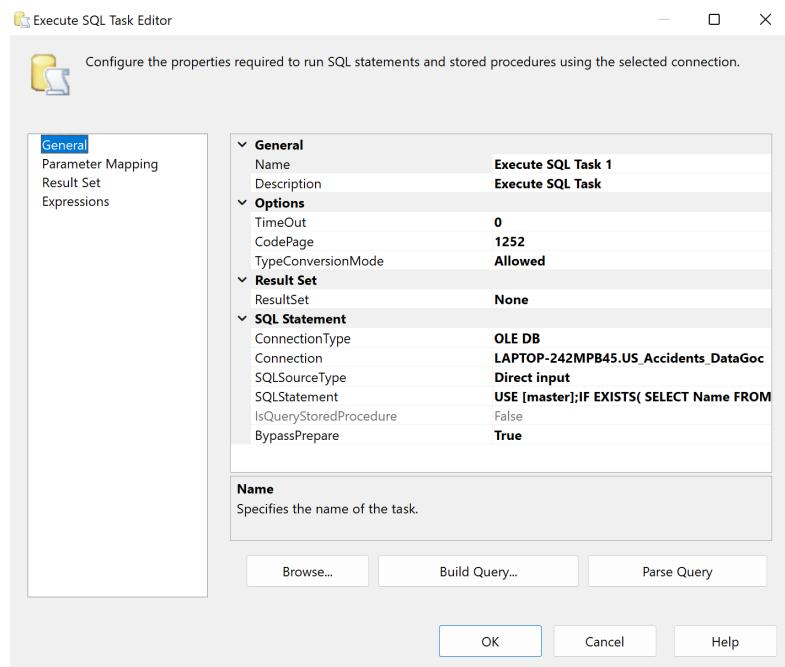
```
CREATE TABLE [Dim_Visibility] (
    [Visibility_ID] int identity (1,1) not null primary key,
    [Visibility] numeric(6,3)
)
```

```
IF(object_id('Dim_Precipitation') IS NOT NULL)
DROP TABLE Dim_Precipitation;
GO
CREATE TABLE [Dim_Precipitation] (
    [Precipitation_ID] int identity (1,1) not null primary key,
    [Precipitation] numeric(6,3),
)
```

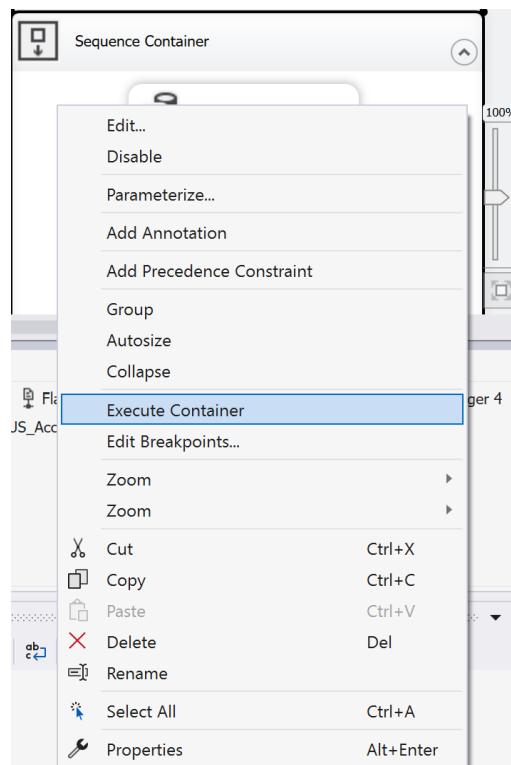
```
IF(object_id('Dim_Wind') IS NOT NULL)
DROP TABLE Dim_Wind;
GO
CREATE TABLE [Dim_Wind] (
    [Wind_ID] int identity (1,1) not null primary key,
    [Wind_Direction] nvarchar(150),
    [Wind_Speed] int
)
```

```
IF(object_id('Fact_US_Accidents') IS NOT NULL)
DROP TABLE Fact_US_Accidents;
GO
CREATE TABLE [Fact_US_Accidents] (
    [ID] nvarchar(100) NOT NULL PRIMARY KEY,
    [Start_Time] datetime2(7),
    [Location_ID] int,
    [Traffic_Condition_ID] int,
    [Airport_ID] int,
    [Weather_ID] int,
    [Wind_ID] int,
    [Temperature_ID] int,
    [Visibility_ID] int,
    [Precipitation_ID] int,
    [Severity] tinyint,
    [Distance] numeric(6,3),
    [Duration] int
)
```

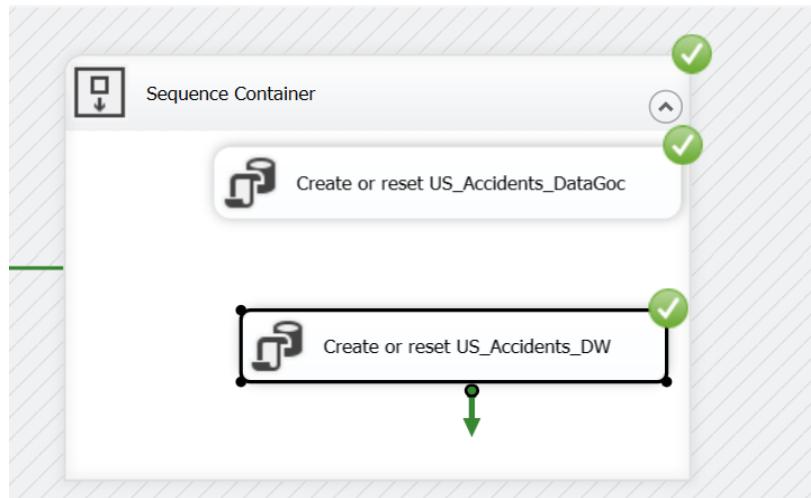
Sau khi đã thêm các câu lệnh SQL vào, ta chọn OK để hoàn tất thiết lập Execute SQL Task.



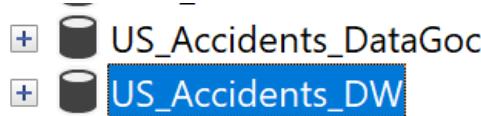
**Bước 11:** Tại Sequence Container, ta chọn chuột phải và chọn Execute Container để chạy các câu lệnh SQL trong các Execute SQL Task vừa được tạo.



Như vậy ta đã hoàn tất quá trình chuẩn bị cơ sở dữ liệu cho quá trình SSIS.

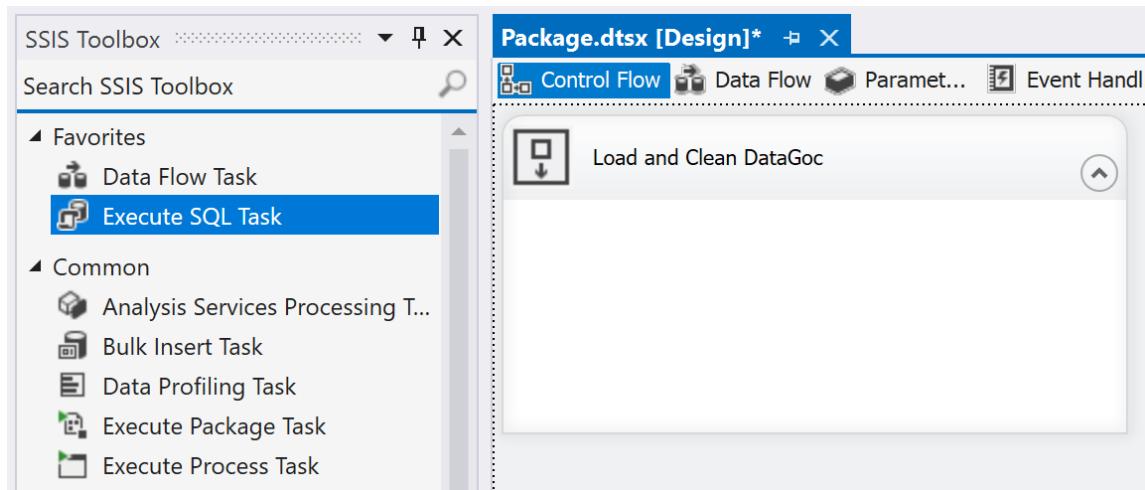


Ta kiểm tra lại cơ sở dữ liệu vừa được tạo trong SQL Server:



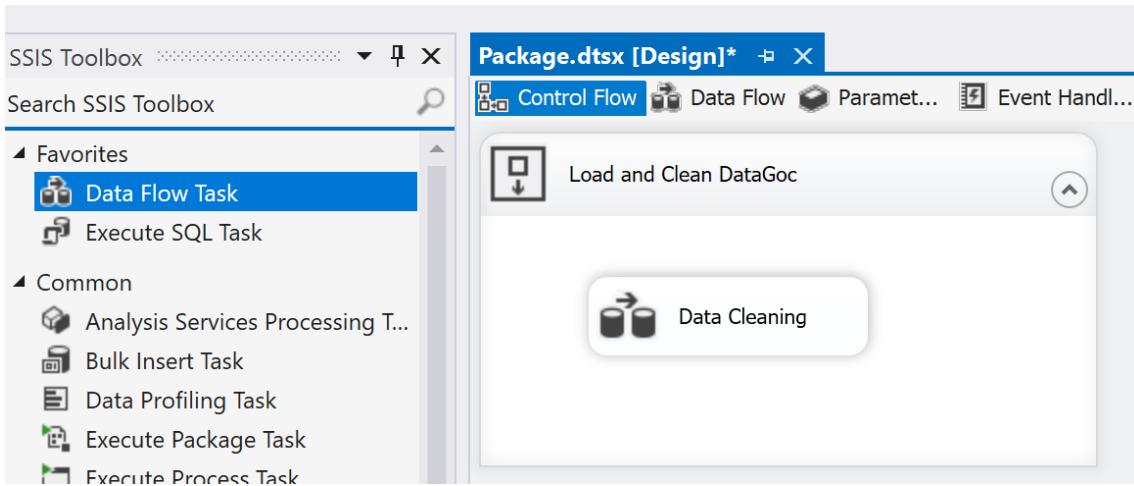
## 2.4. Quá trình làm sạch dữ liệu và đổ dữ liệu gốc vào cơ sở dữ liệu

**Bước 1:** Tạo ra một Sequence Container và đặt tên là Load and clean DataGoc.

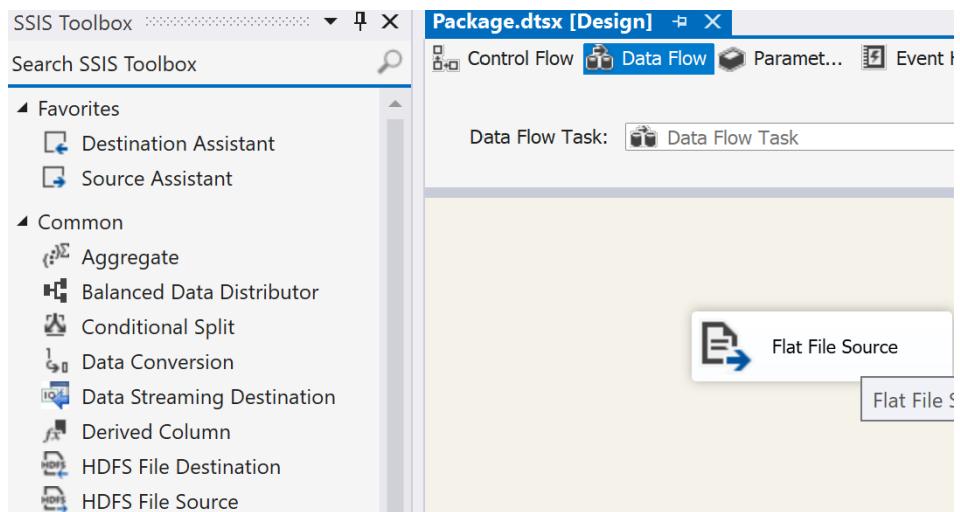


**Bước 2:** Trong Sequence Container vừa khởi tạo, tạo mới một Data Flow Task và đặt tên là Data cleaning.

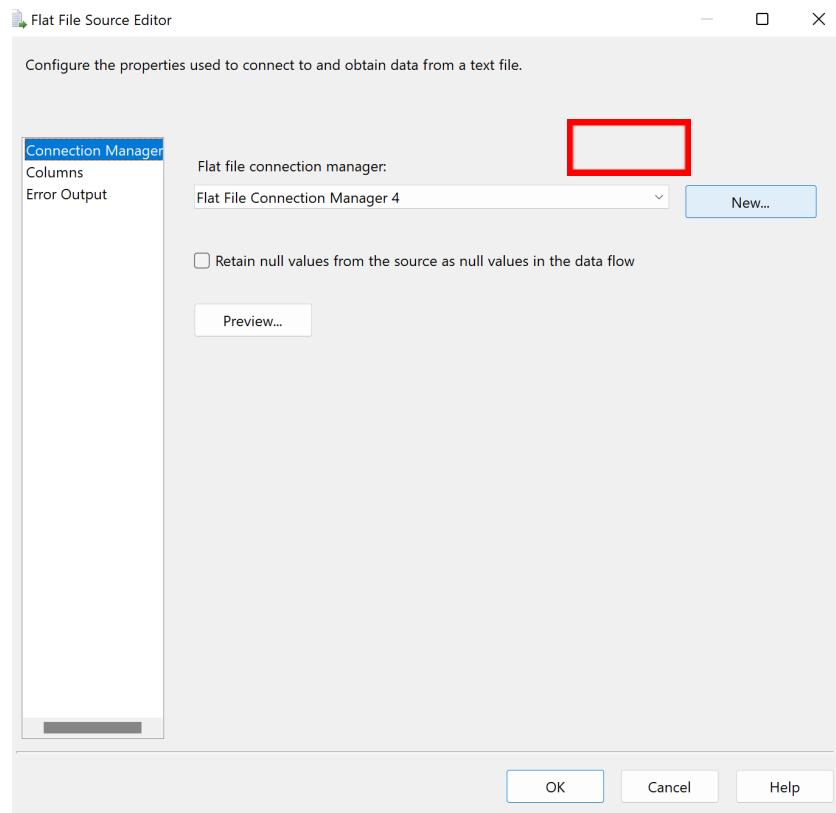
Ta sẽ thực hiện quá trình làm sạch dữ liệu và đổ dữ liệu gốc vào database tại Data Flow Task này.



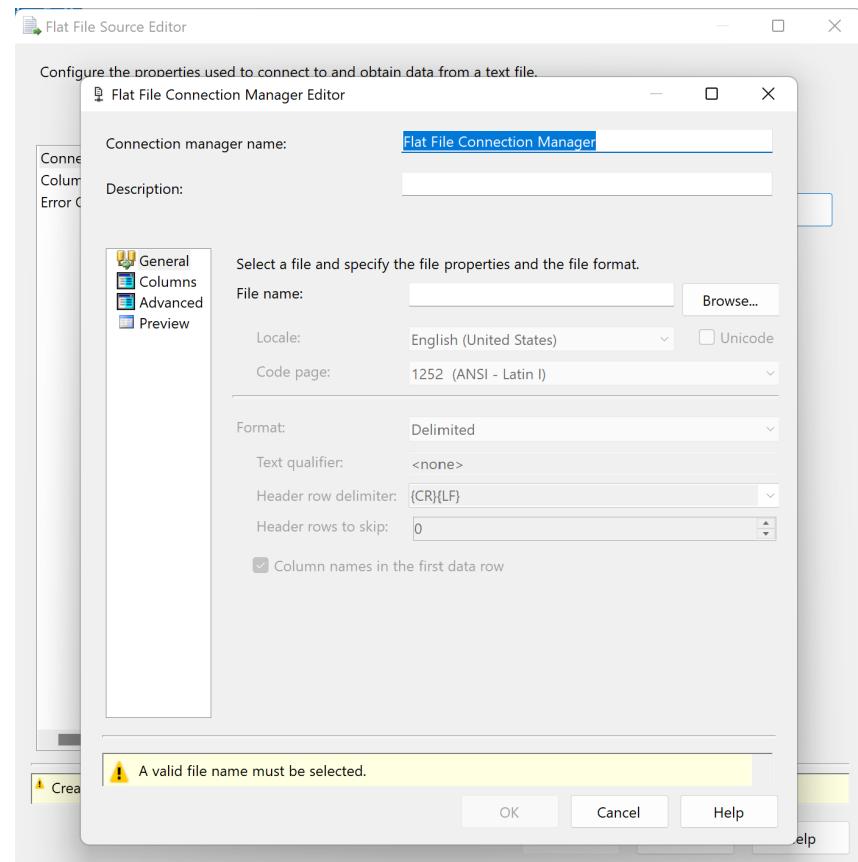
**Bước 3:** Double click vào Data Flow Task vừa tạo, tạo mới một Flat File Source. Flat File Source này sẽ giúp ta lấy dữ liệu từ file csv vào project.



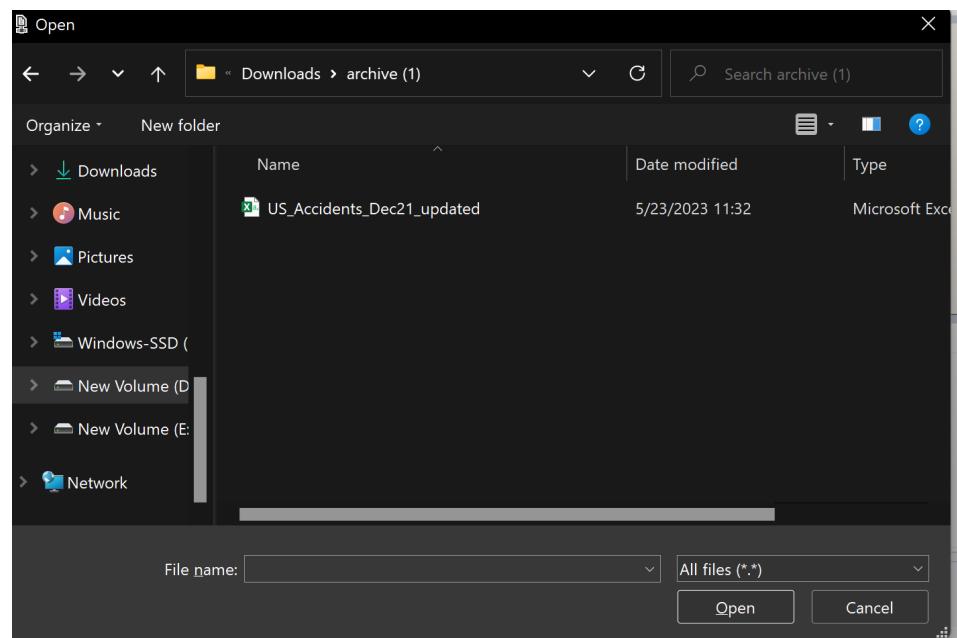
**Bước 4:** Double click vào Flat File Source và chọn New để tạo mới connection đến file csv chứa dữ liệu gốc.



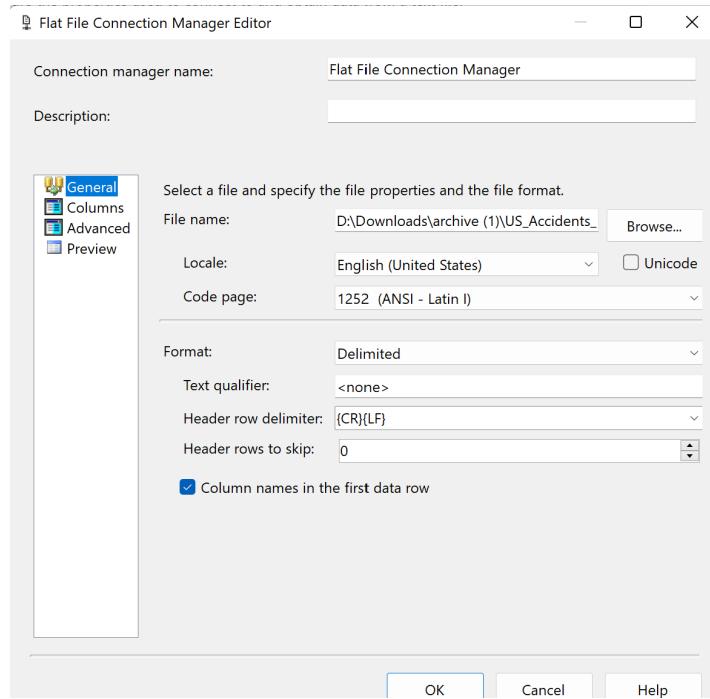
**Bước 5:** Tại mục File name, ta chọn Browse...



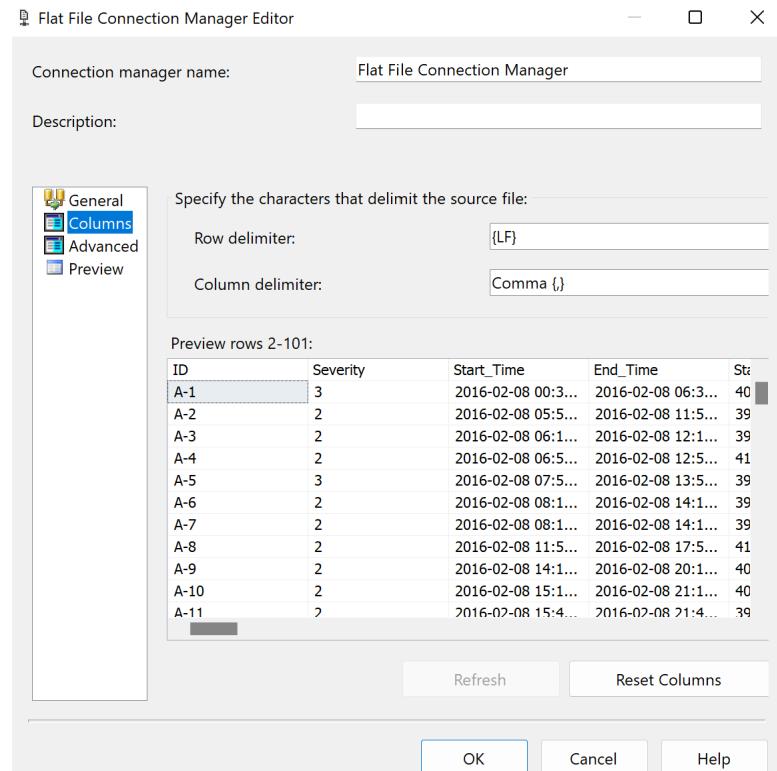
Sau đó mở file .csv chứa dữ liệu gốc vào. Chú ý: Ta cần chọn đúng loại file là .csv thì file mới xuất hiện.



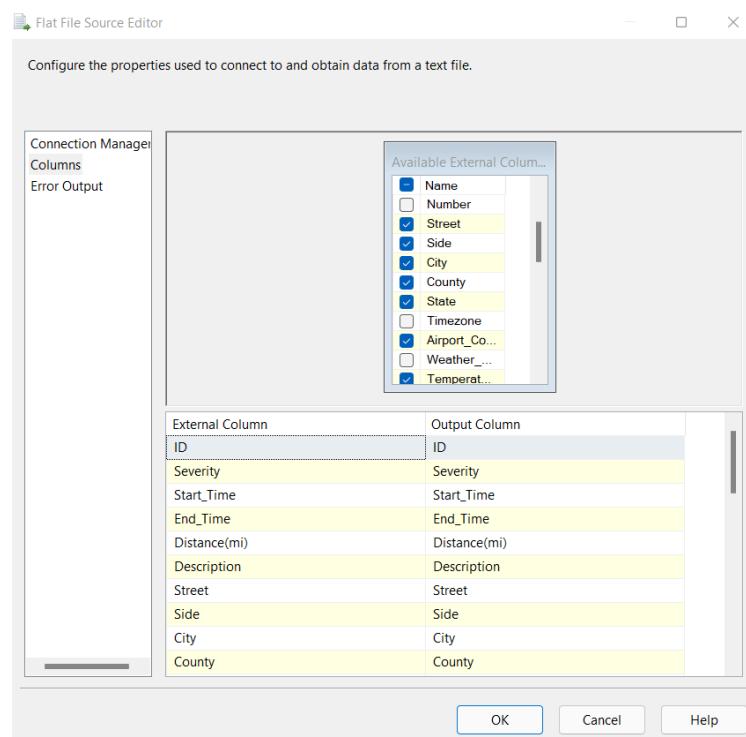
**Bước 6:** Tại hộp thoại, ta chọn mục Columns để xác định các trường dữ liệu đầu vào.



**Bước 7:** Sau khi đã xem dữ liệu xong, ta chọn OK để hoàn thành việc thiết lập Flat File Source.

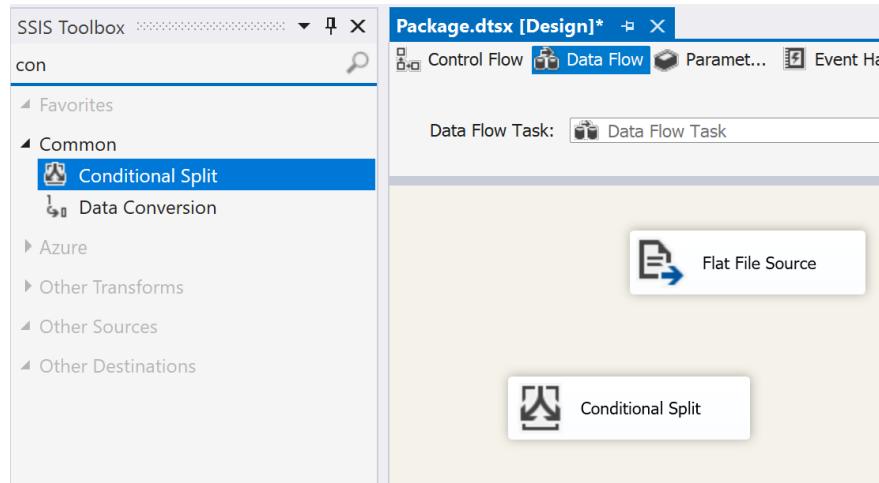


**Bước 8:** Tại hộp thoại Flat File Source Editor, ta tiếp tục chọn mục columns và bỏ chọn những cột không cần dùng.



**Bước 9:** Tiếp theo, ta thêm một Conditional Split và liên kết Flat File Source với Conditional Split vừa tạo.

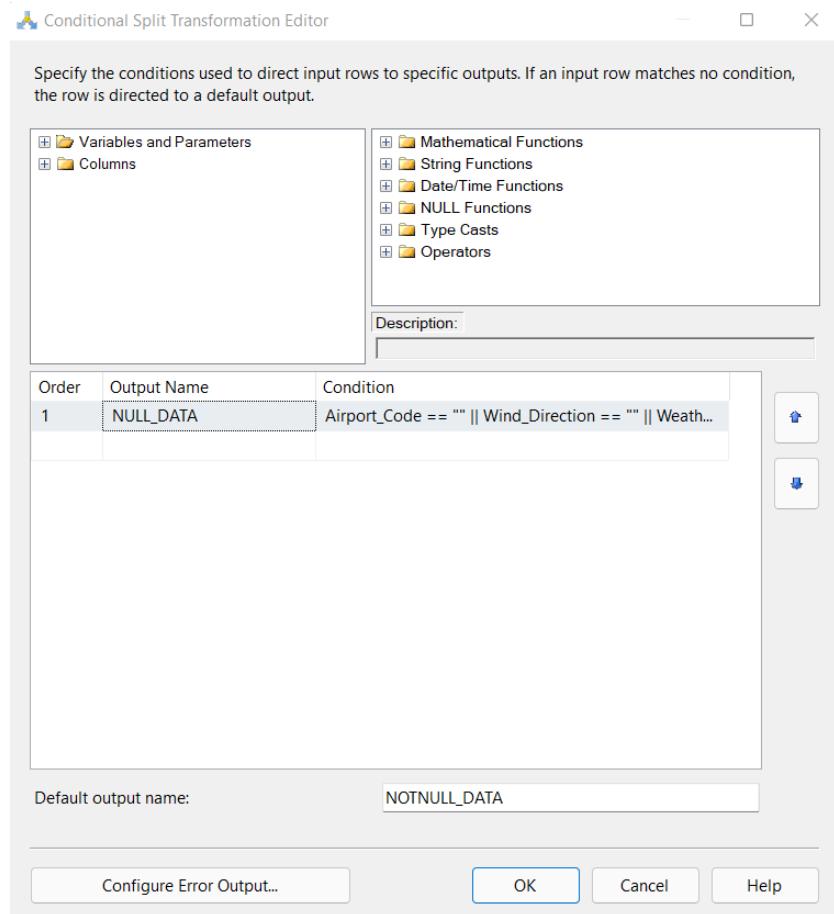
Conditional Split này sẽ phân tách dữ liệu gốc thành 2 phần dữ liệu rỗng và giữ liệu không rỗng.



**Bước 10:** Double click vào Conditional Split vừa tạo. Sau đó, ta thêm các điều kiện để lọc ra các trường dữ liệu rỗng như sau:

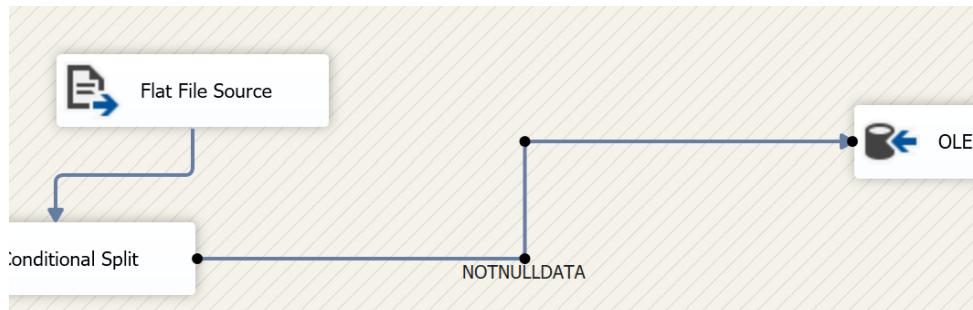
```
Airport_Code == "" || Wind_Direction == "" || Weather_Condition == "" || ....
```

Sau đó, ta đổi tên Output Name trong trường hợp này là NULL\_DATA và đổi tên Default output name (các dữ liệu còn lại) là NOTNULL\_DATA. Cuối cùng, ta chọn OK để hoàn tất thiết lập.



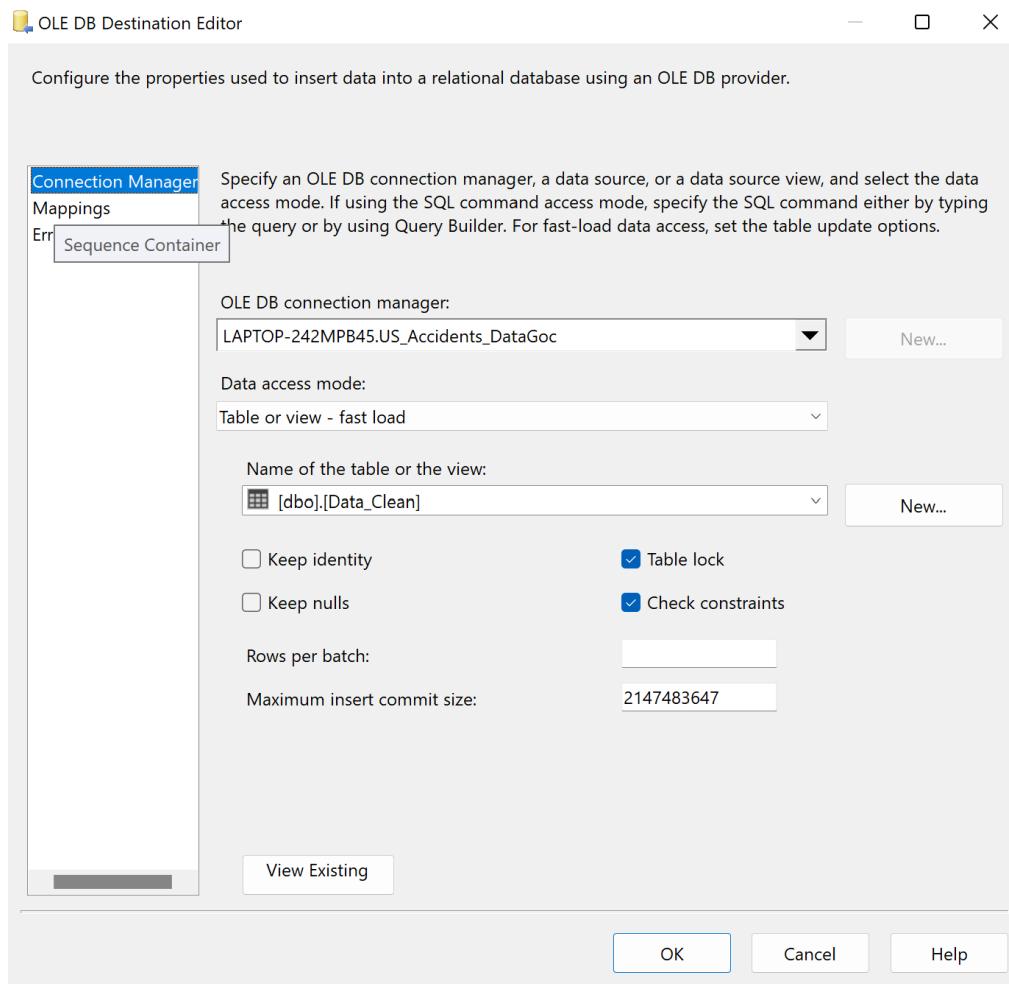
**Bước 11:** Ta tạo ra một OLE DB Destination, đồng thời liên kết Conditional Split với OLE DB Destination này.

Khi hộp thoại Input Output Selection xuất hiện, ta chọn Output là NOTNULL\_DATA để đồ dữ liệu đã được làm sạch vào cơ sở dữ liệu. Sau đó chọn OK.



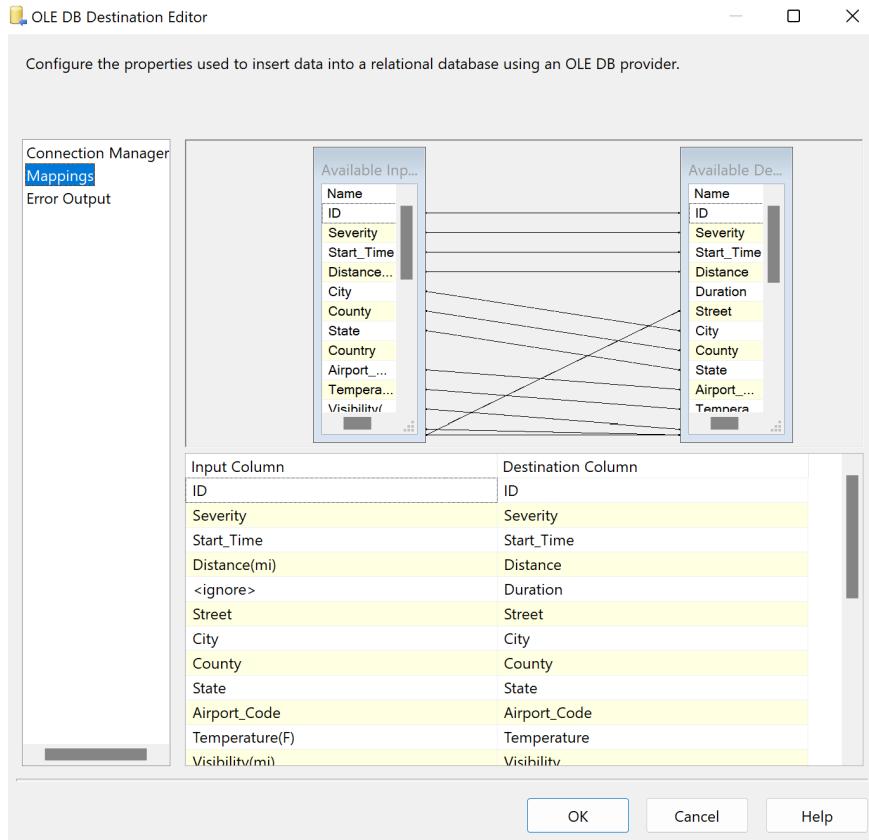
**Bước 12:** Double click vào OLE DB Destination. Tại OLE DB connection manager, ta chọn New... để tạo kết nối mới đến cơ sở dữ liệu US\_Accidents\_DataGoc vừa được khởi tạo ở các bước trên.

Tại connection mới tạo, chọn bảng Data\_Clean là bảng đã được tạo ở các bước trước và cũng là nơi lưu dữ liệu gốc đã được làm sạch.

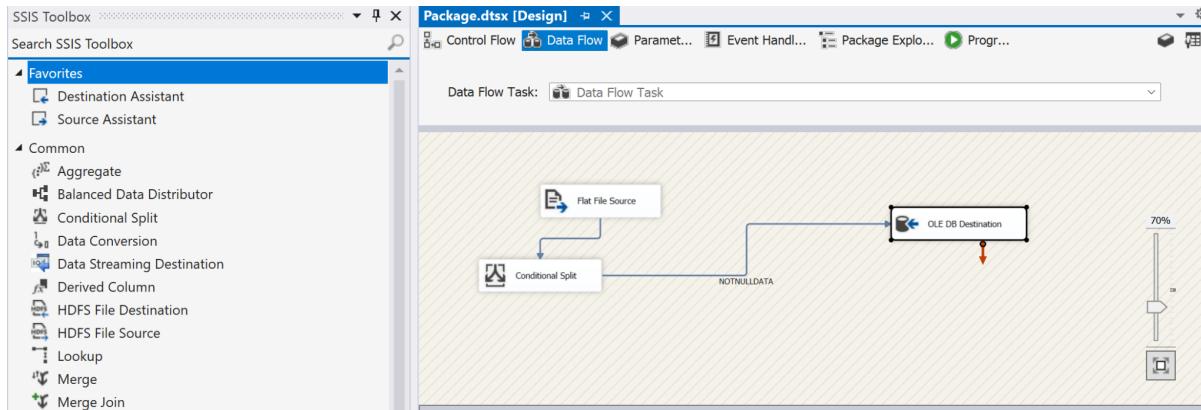


**Bước 13:** Trước khi hoàn tất thiết lập này, ta chọn mục Mappings để kiểm tra sự ánh xạ giữa các cột dữ liệu.

Nếu có các cột chưa được ánh xạ, ta tiến hành ánh xạ thủ công chúng và chọn OK để hoàn tất thiết lập.

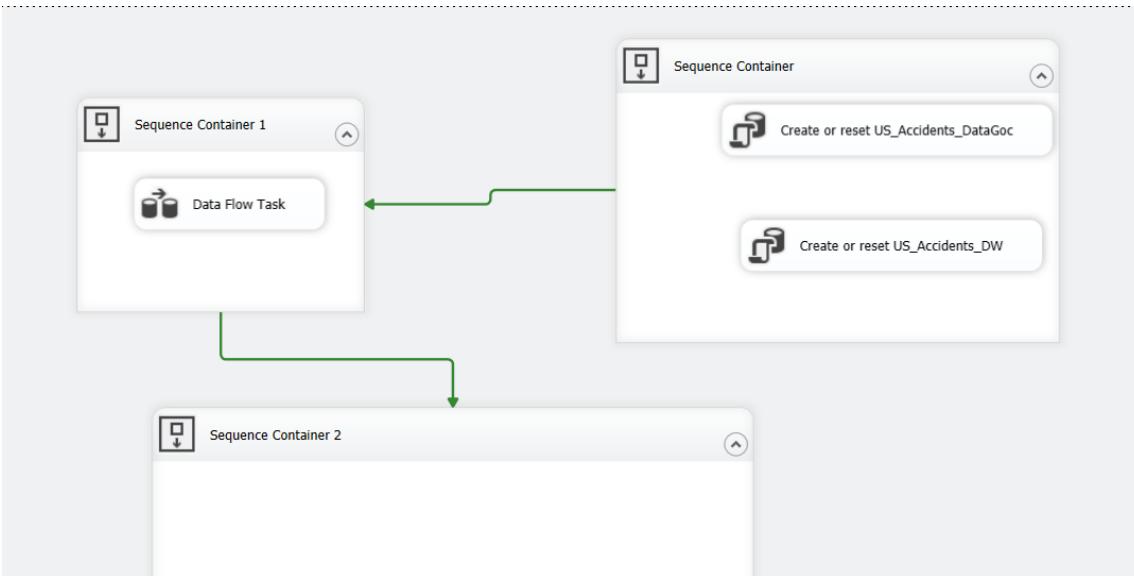


Như vậy, chúng ta đã hoàn tất quá trình làm sạch dữ liệu và đổ dữ liệu gốc vào cơ sở dữ liệu.



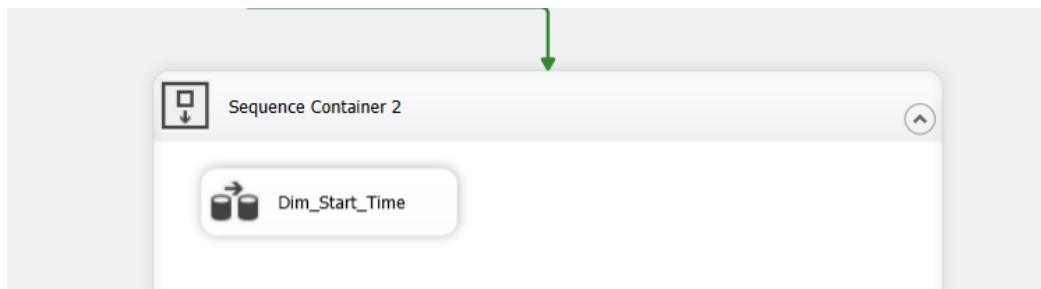
## 2.5. Đổ dữ liệu từ DataGoc vào các bảng Dimension

Đầu tiên ta tạo một Sequence Container mới và đặt tên là Load Dimension Tables để có thể thực hiện việc đổ dữ liệu vào các bảng Dimension cùng một lúc.

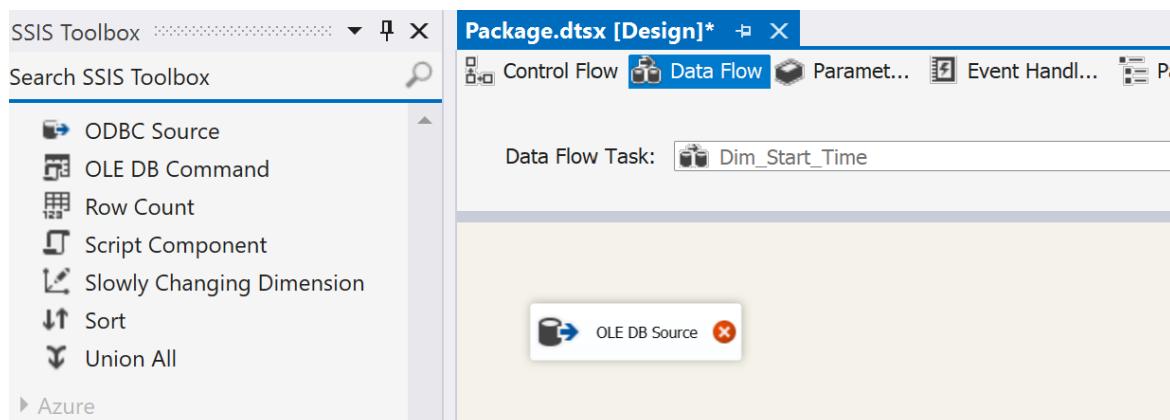


### 2.5.1. Bảng Dim\_Start\_Time

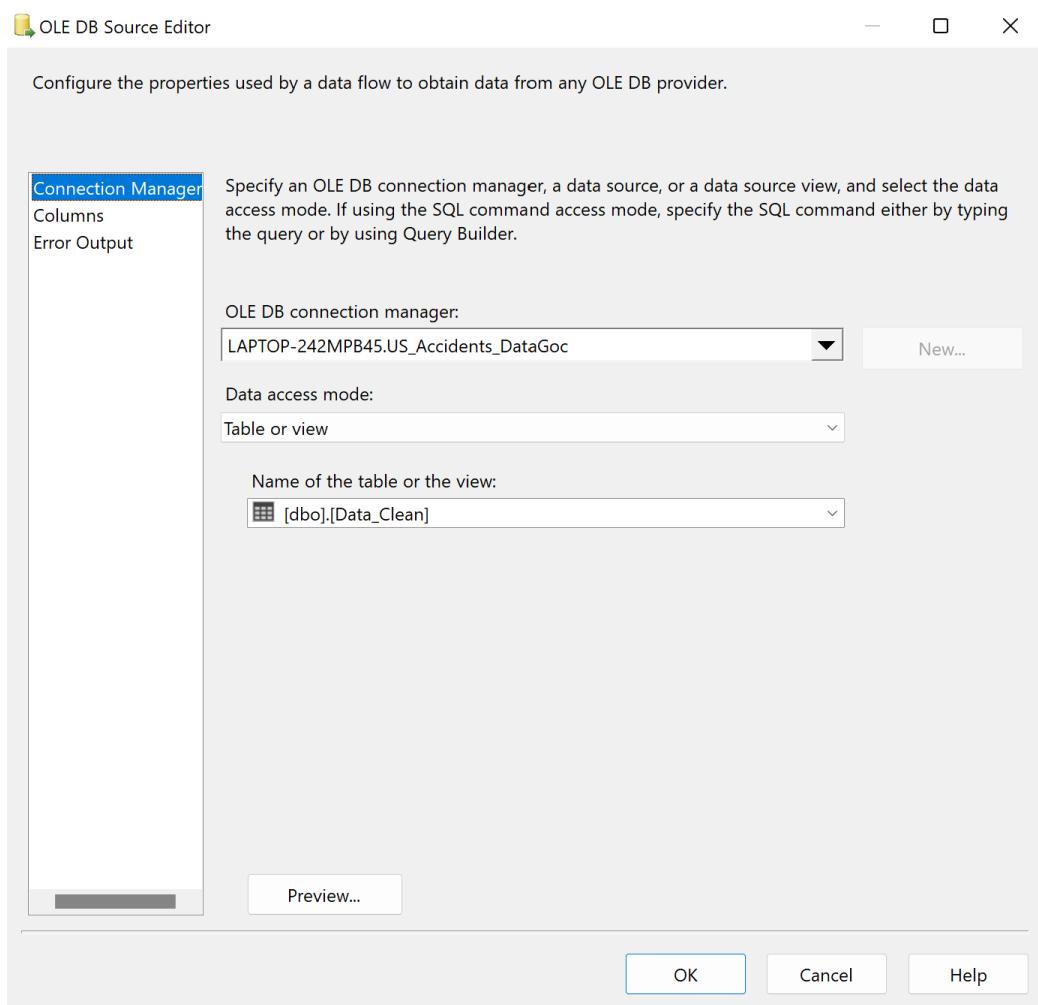
**Bước 1:** Tại Sequence Container mới tạo, tạo một Data Flow Task mới và đặt tên là Load Dim\_Start\_Time.



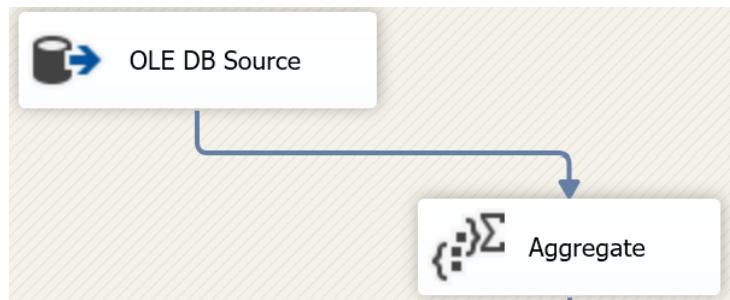
**Bước 2:** Double click vào Data Flow Task mới tạo và tạo mới một OLE DB Source.



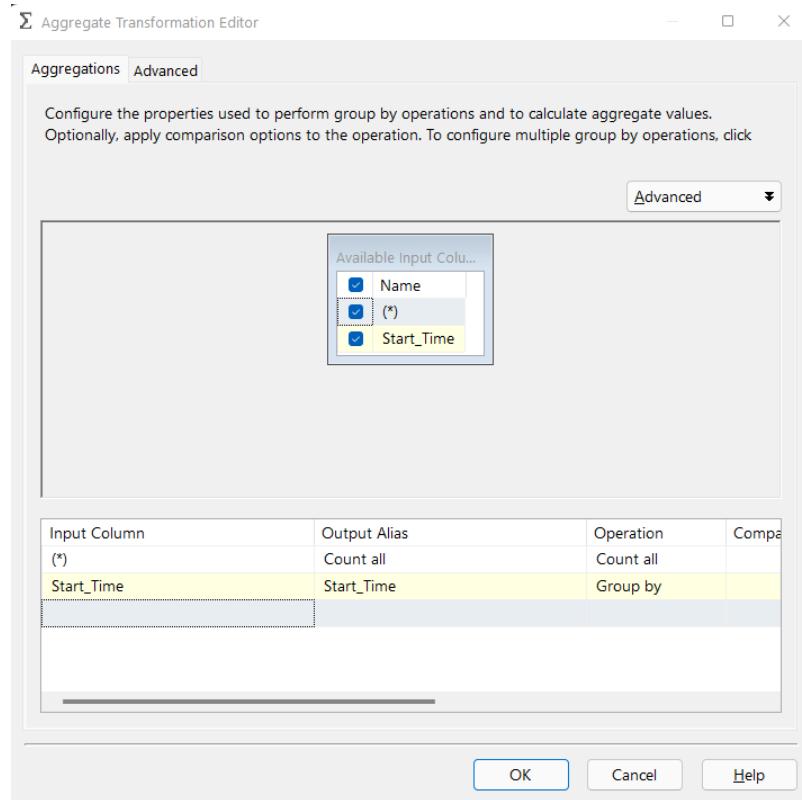
**Bước 3:** Double click vào OLE DB Source này, ta chọn connection manager là cơ sở dữ US\_Accident\_DataGoc và chọn bảng Data\_Clean. Sau đó chọn OK để hoàn tất.



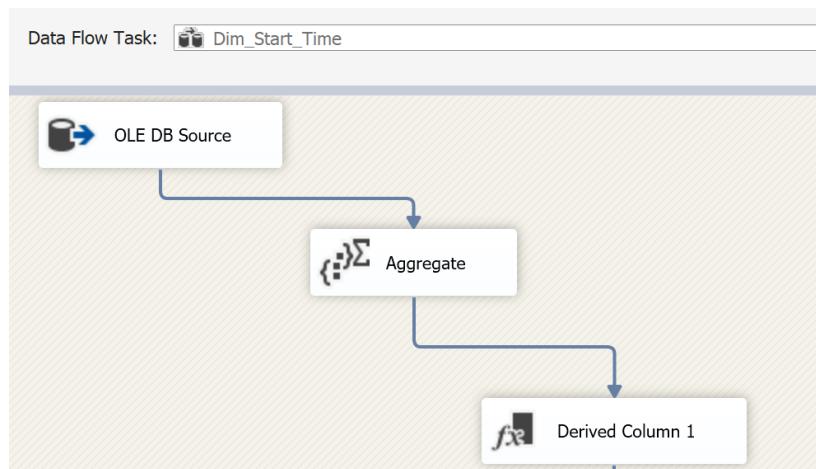
**Bước 4:** Tạo mới một Aggregate và tạo liên kết từ OLE DB Source đến Aggregate vừa tạo



**Bước 5:** Double click vào Aggregate, chọn 2 cột dữ liệu liên quan đến bảng Dim\_Start\_Time với Operation lần lượt là Count all và groupby. Sau đó chọn OK.



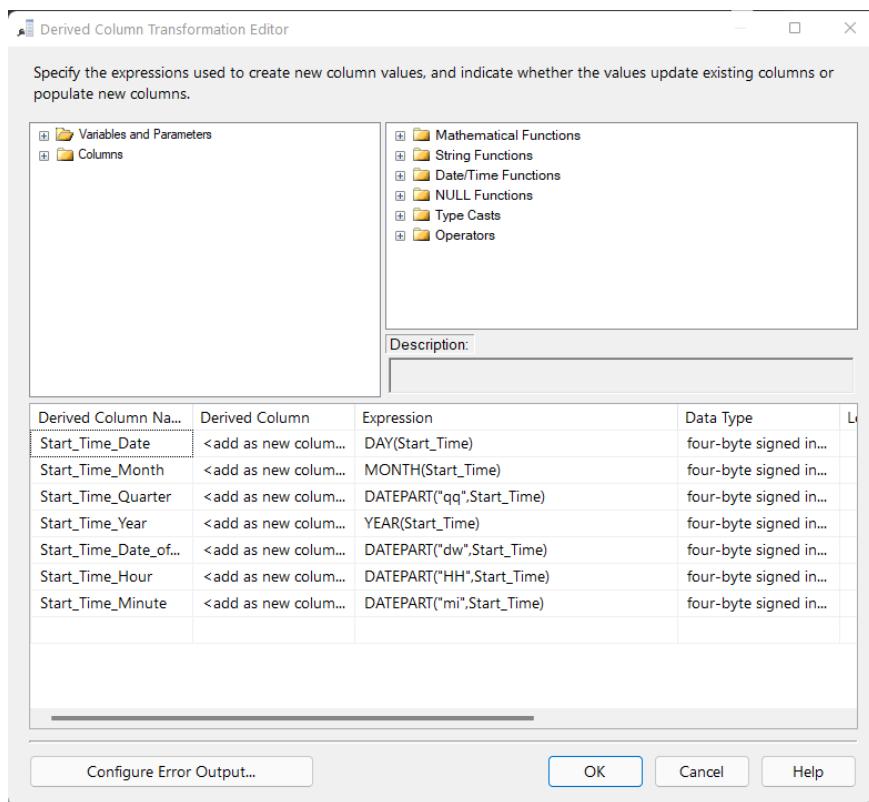
**Bước 6:** Thêm mới một Derived Column và tạo liên kết Derived Column này với Aggregate.



**Bước 7:** Double click vào Derived Column. Sau đó sử dụng các hàm để tạo ra các trường dữ liệu cần thiết cho câu truy vấn dựa trên cột Start\_Time đã có. Dữ liệu này bao gồm:

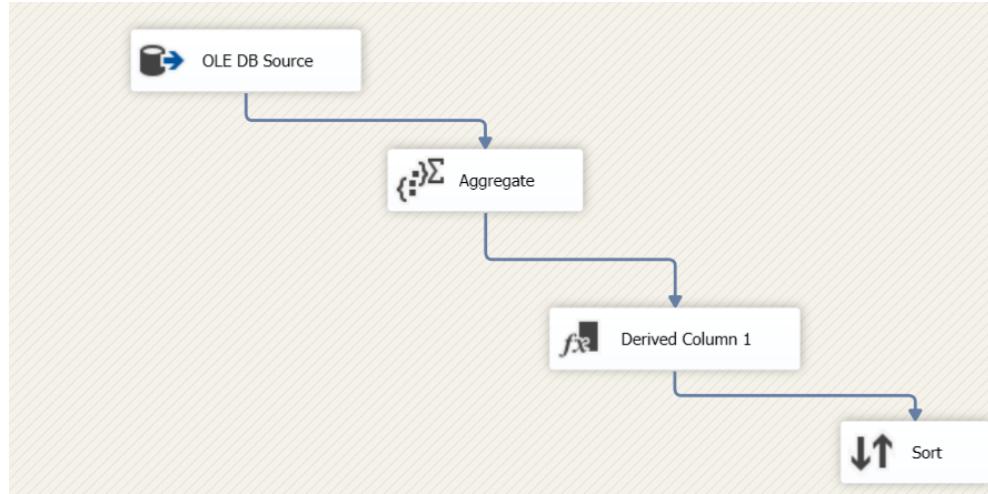
- + Start\_Time\_Date: lấy ra ngày từ Start\_Time.
- + Start\_Time\_Month: lấy ra tháng từ Start\_Time.

- + Start\_Time\_Quarter: lấy ra quý từ Start\_Time.
- + Start\_Time\_Year: lấy ra năm từ Start\_Time.
- + Start\_Time\_Date\_of\_week: lấy ra ngày trong tuần từ Start\_Time.
- + Start\_Time\_Hour: lấy ra giờ từ Start\_Time
- + Start\_Time\_Minute: lấy ra phút trong Start\_Time



Sau đó chọn OK để hoàn tất thiết lập.

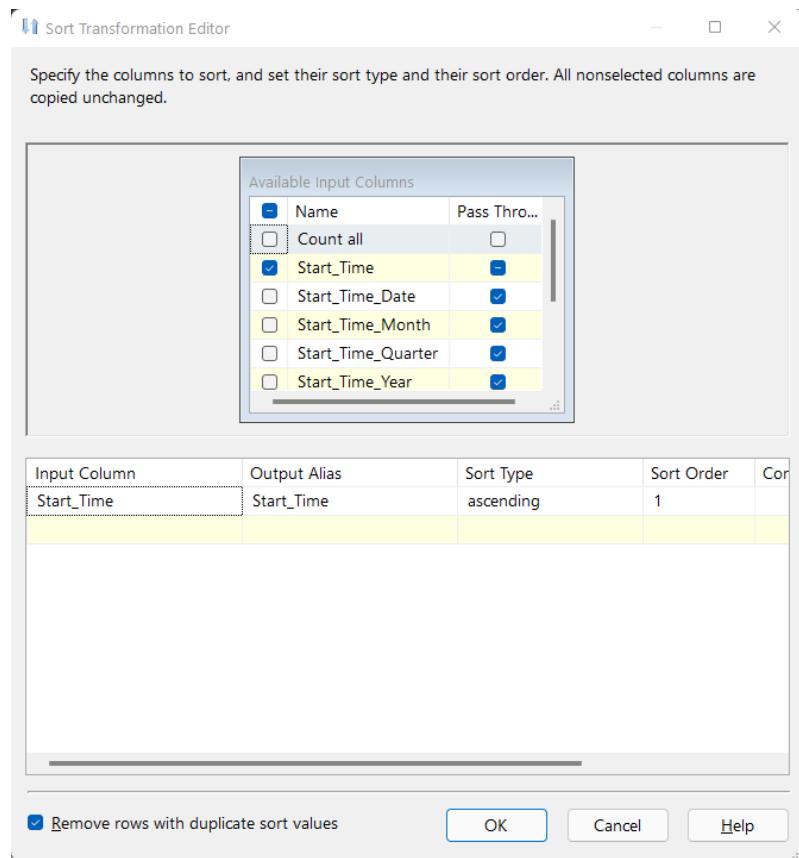
**Bước 8:** Tiếp theo, ta tạo ra một khôi Sort để sắp xếp dữ liệu



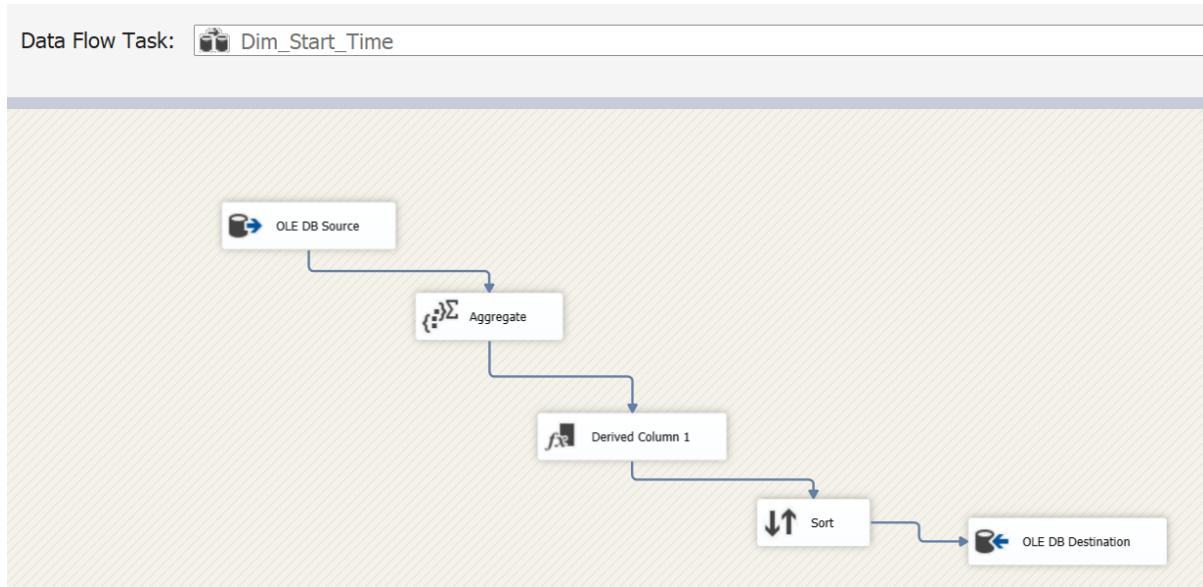
**Bước 9:** Double click vào Sort. Sau đó, chọn trường dữ liệu Start\_Time và sắp xếp theo thứ tự tăng dần.

Đồng thời, ta tick chọn checkbox “Remove rows with duplicate sort values” để loại bỏ các hàng có giá trị sắp xếp trùng lặp.

Cuối cùng, chọn OK để kết thúc thiết lập.



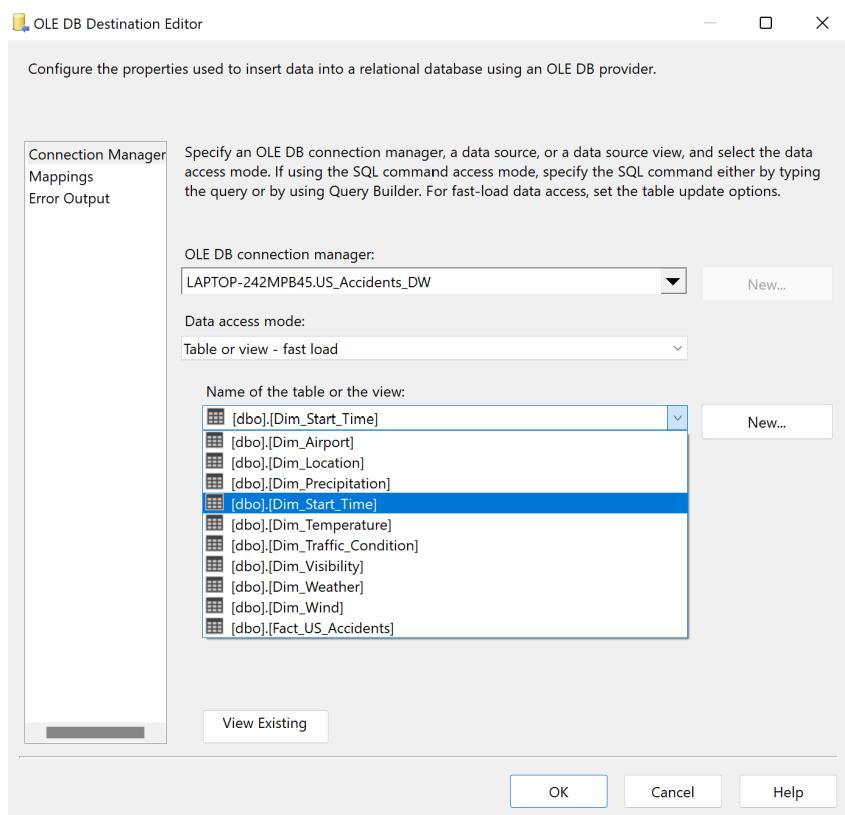
**Bước 10:** Tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong kho dữ liệu US\_Accidents\_DW.

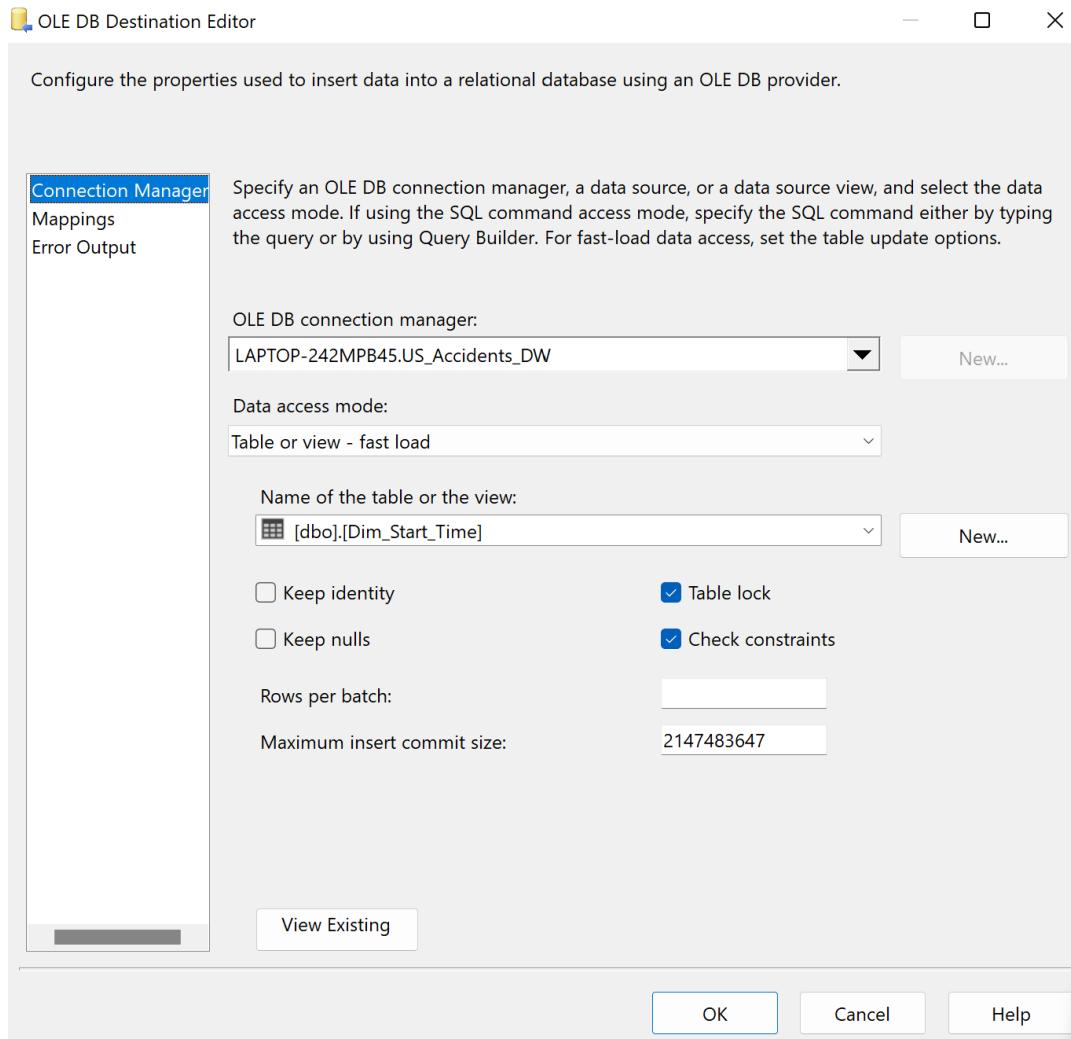


**Bước 11:** Double click vào **OLE DB Destination** vừa tạo, ta chọn New... để tạo một kết nối đến kho dữ liệu US\_Accidents\_DW.

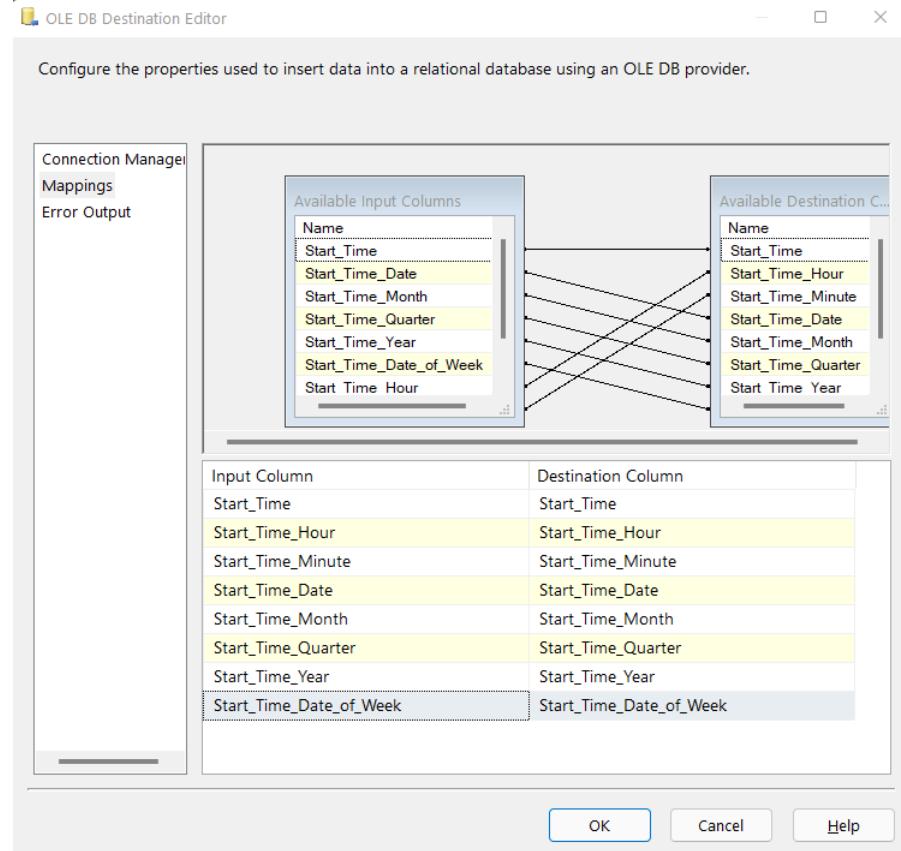
Tại hộp thoại Connection Manager, chọn Server Name và chọn tên cơ sở dữ liệu là US\_Accidents\_DW. Sau đó chọn Test Connection  OK để hoàn tất tạo kết nối.

Tiếp theo ta chọn địa điểm đổ dữ liệu là bảng Dim\_Start\_Time trong kho dữ liệu US\_Accidents\_DW.





Tiếp đến ta cần chọn mục Mappings để xem xét việc ánh xạ các cột dữ liệu

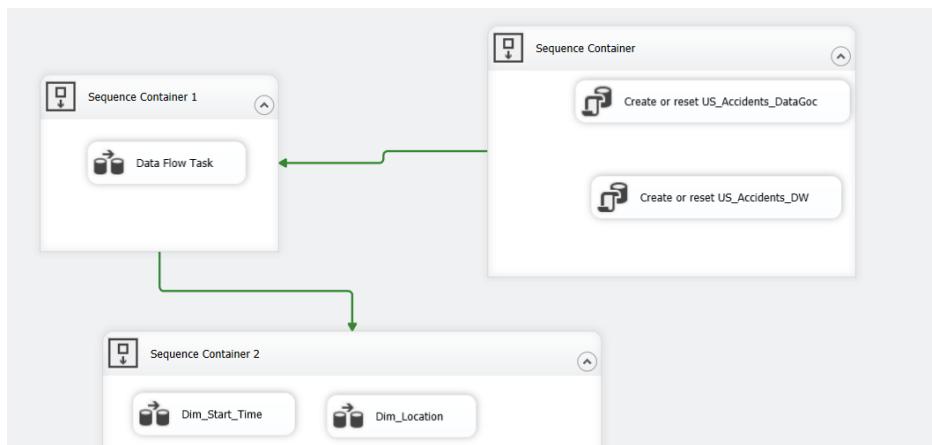


Cuối cùng, chọn OK để hoàn tất thiết lập.

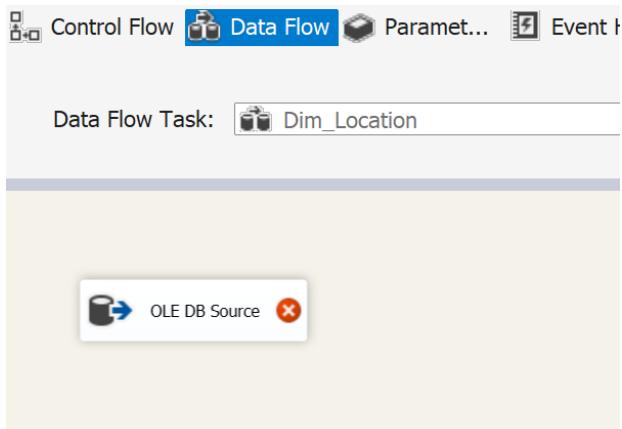
Như vậy, chúng ta đã hoàn tất quá trình đổ dữ liệu vào bảng Dim\_Start\_Time.

### **2.5.2. Bảng Dim\_Location**

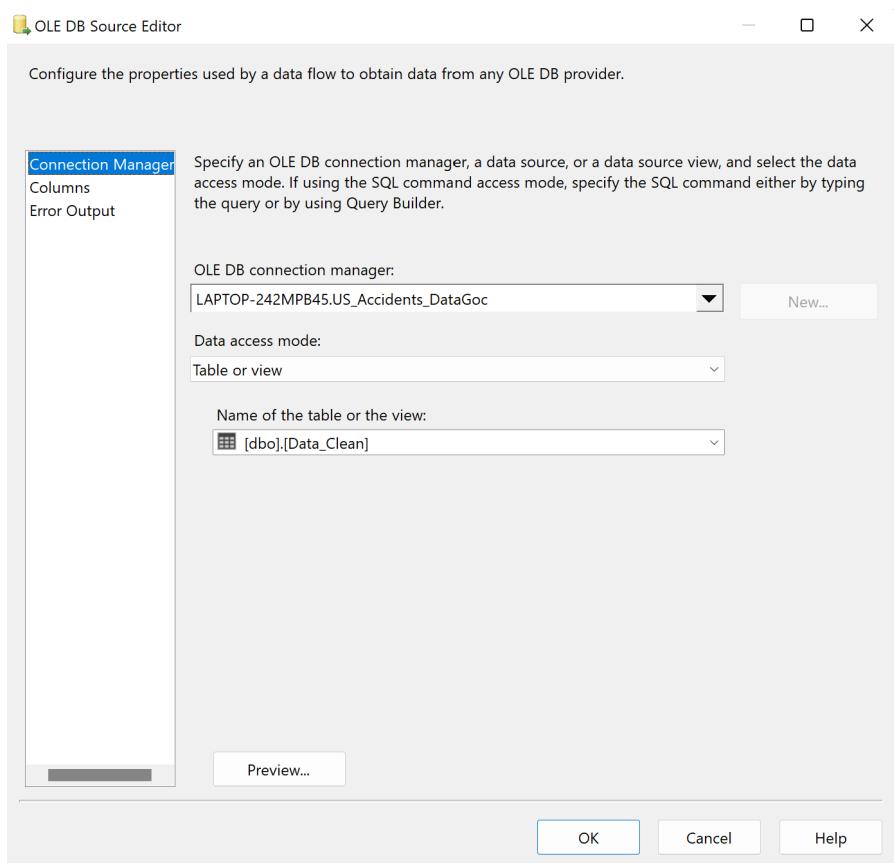
**Bước 1:** Tại Sequence Container Load Dimension Tables, tạo một Data Flow Task mới và đặt tên là Load Dim\_Location



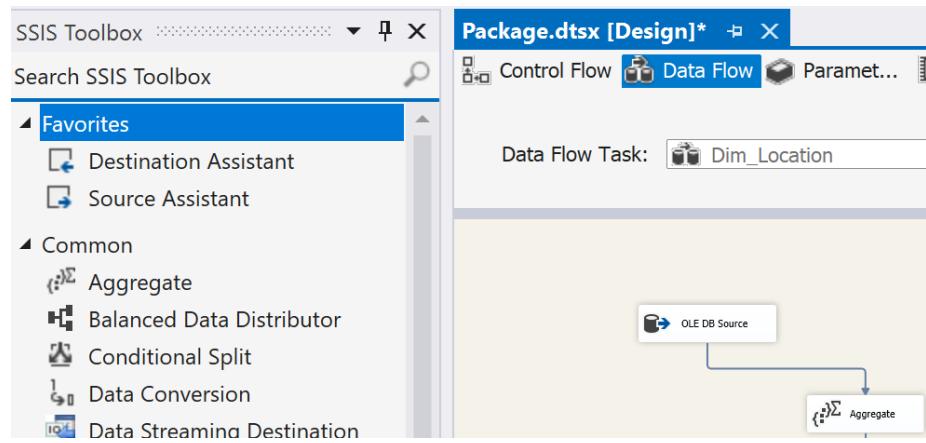
**Bước 2:** Double click vào Data Flow Task mới tạo và tạo mới một OLE DB Source.



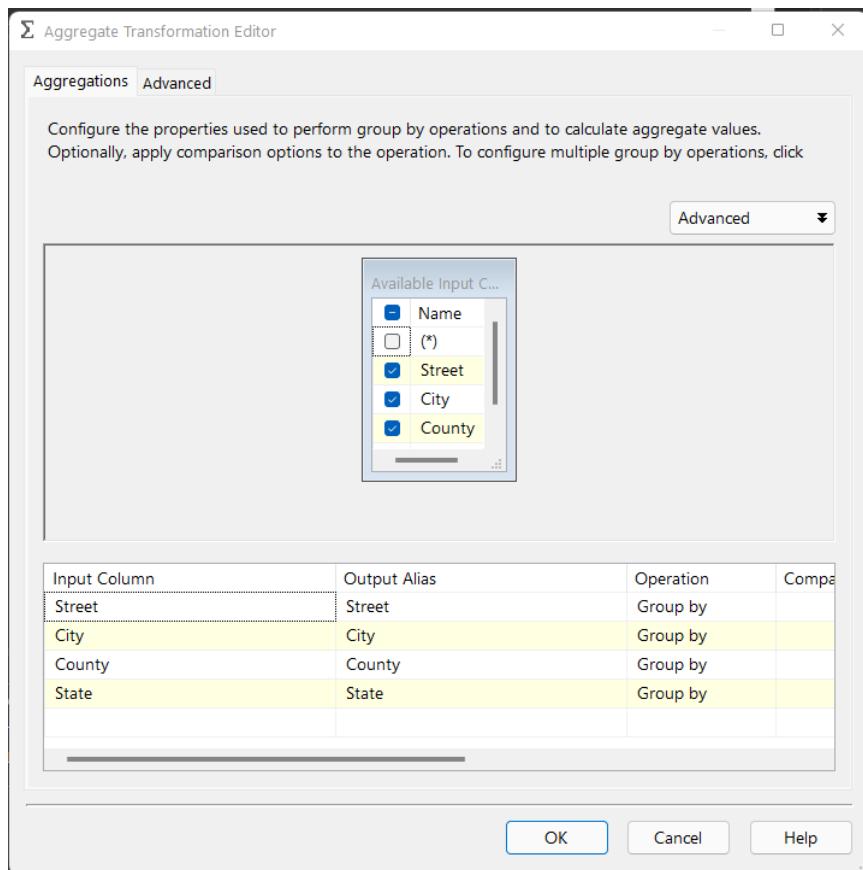
**Bước 3:** Double click vào OLE DB Source này, ta chọn connection manager là cơ sở dữ US\_Accident\_DataGoc và chọn bảng Data\_Clean. Sau đó chọn OK để hoàn tất.



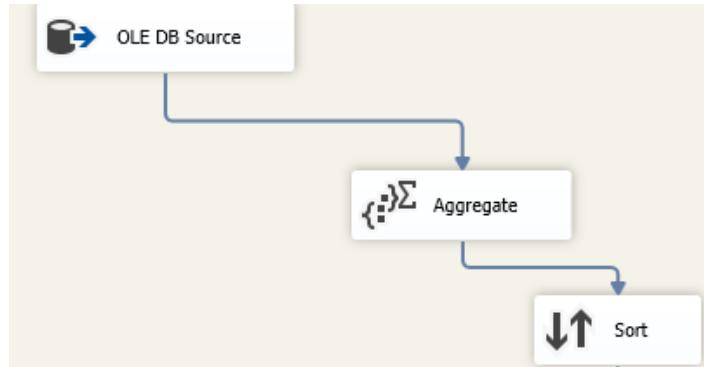
**Bước 4:** Tạo mới một Aggregate và tạo liên kết từ OLE DB Source đến Aggregate vừa tạo



**Bước 5:** Double click vào Aggregate, chọn cột dữ liệu liên quan đến bảng Dim\_Location với Operation là groupby để gom nhóm dữ liệu. Sau đó chọn OK.



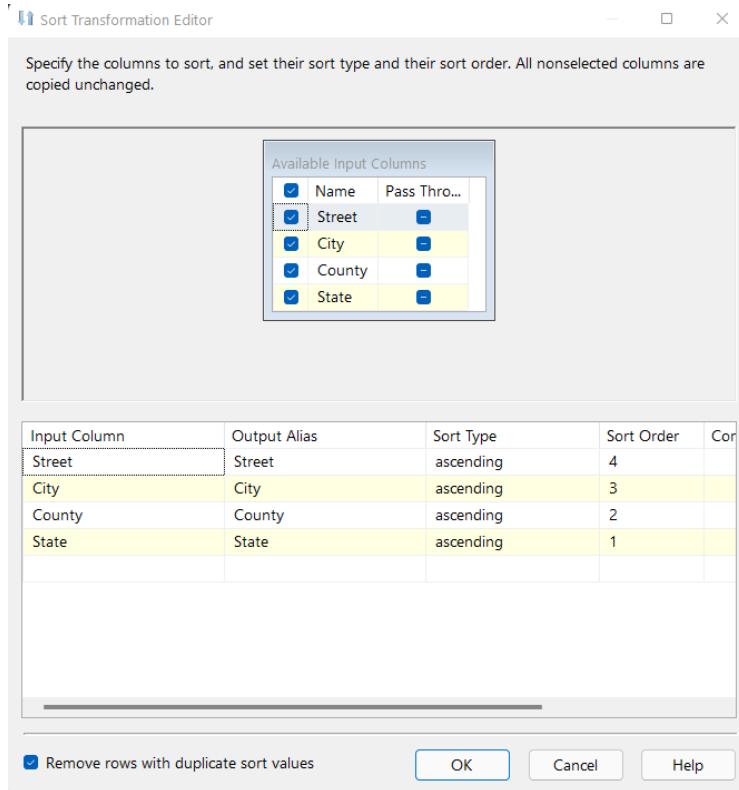
**Bước 6:** Tiếp theo, ta tạo ra một khối Sort để sắp xếp dữ liệu



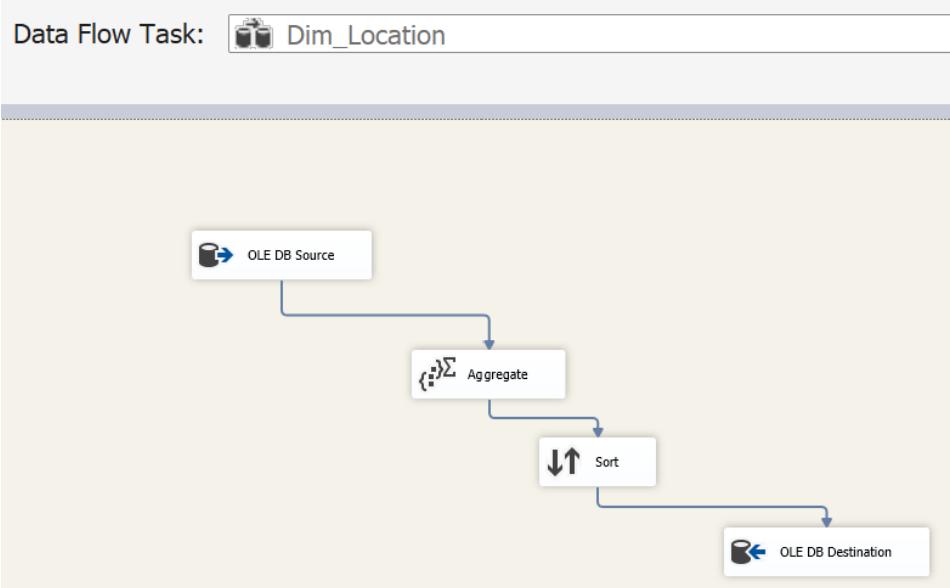
**Bước 7:** Double click vào Sort. Sau đó, chọn tất cả các trường dữ liệu đang có và sắp xếp theo thứ tự tăng dần.

Đồng thời, ta tick chọn checkbox “Remove rows with duplicate sort values” để loại bỏ các hàng có giá trị sắp xếp trùng lặp.

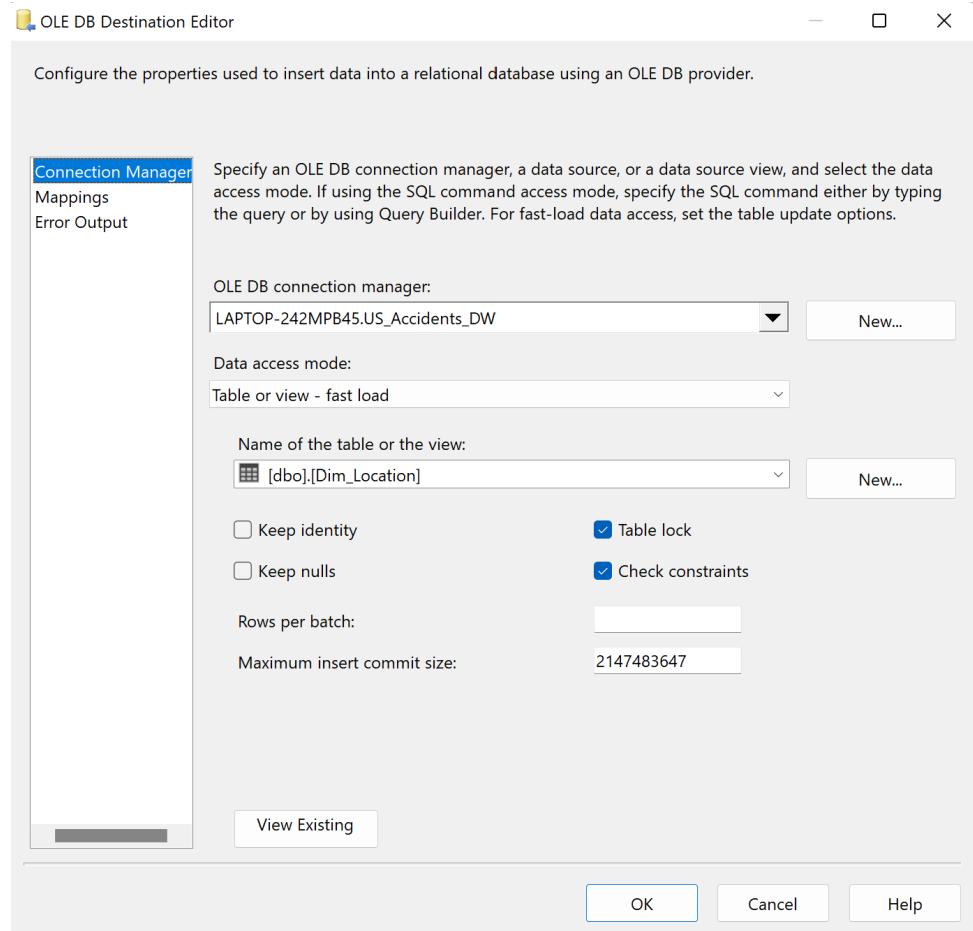
Cuối cùng, chọn OK để kết thúc thiết lập.



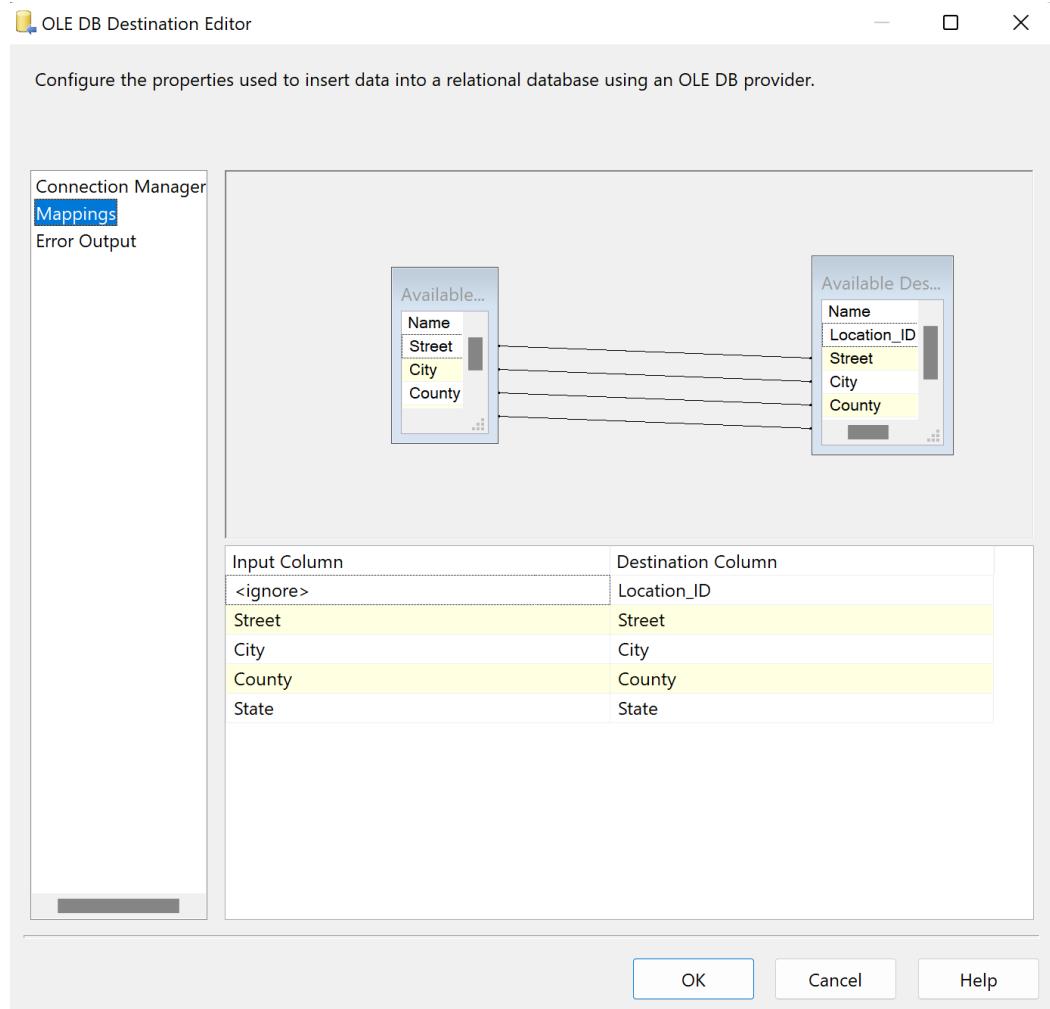
**Bước 8:** Tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong kho dữ liệu US\_Accidents\_DW.



**Bước 9:** Tiếp theo ta double click vào **OLE DB Destination** và chọn địa điểm để dữ liệu là bảng Dim\_Location trong kho dữ liệu US\_Accidents\_DW.



Tiếp đến ta cần chọn mục Mappings để xem xét việc ánh xạ các cột dữ liệu

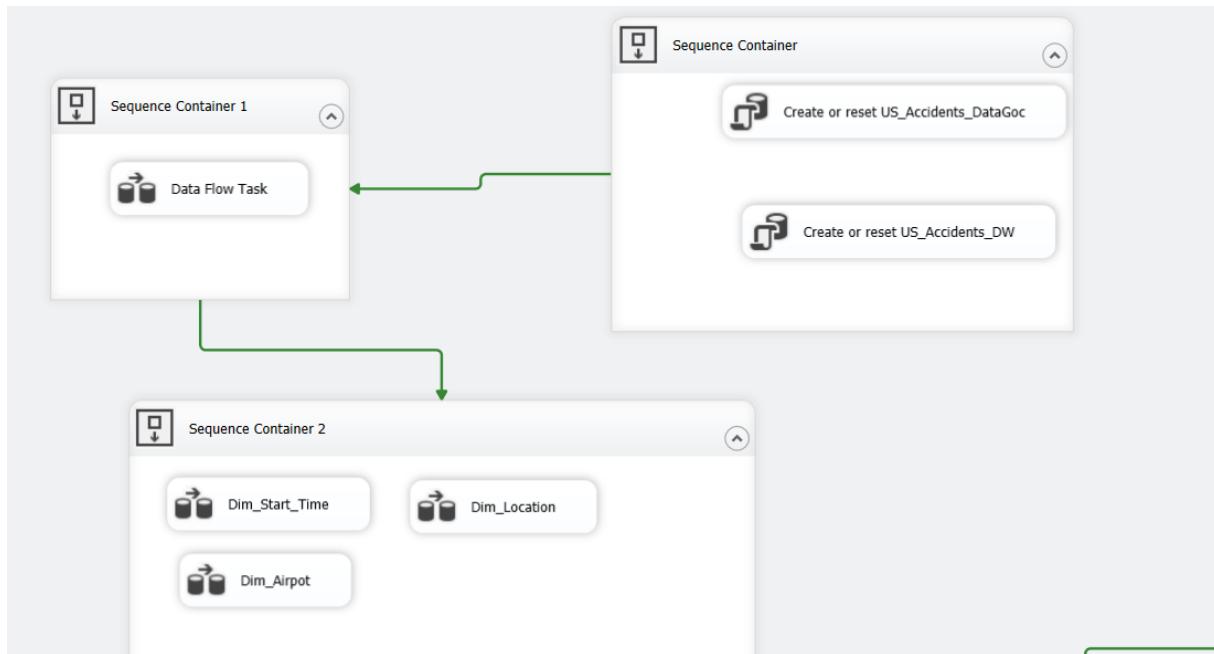


Cuối cùng, chọn OK để hoàn tất thiết lập.

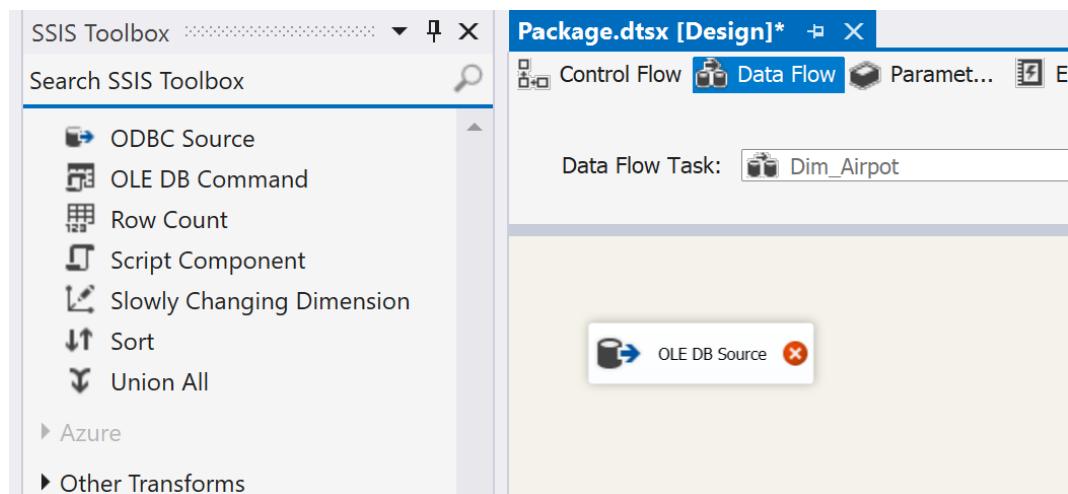
Như vậy, chúng ta đã hoàn tất quá trình đổ dữ liệu vào bảng Dim\_Location.

### 2.5.3. Bảng Dim\_Airport

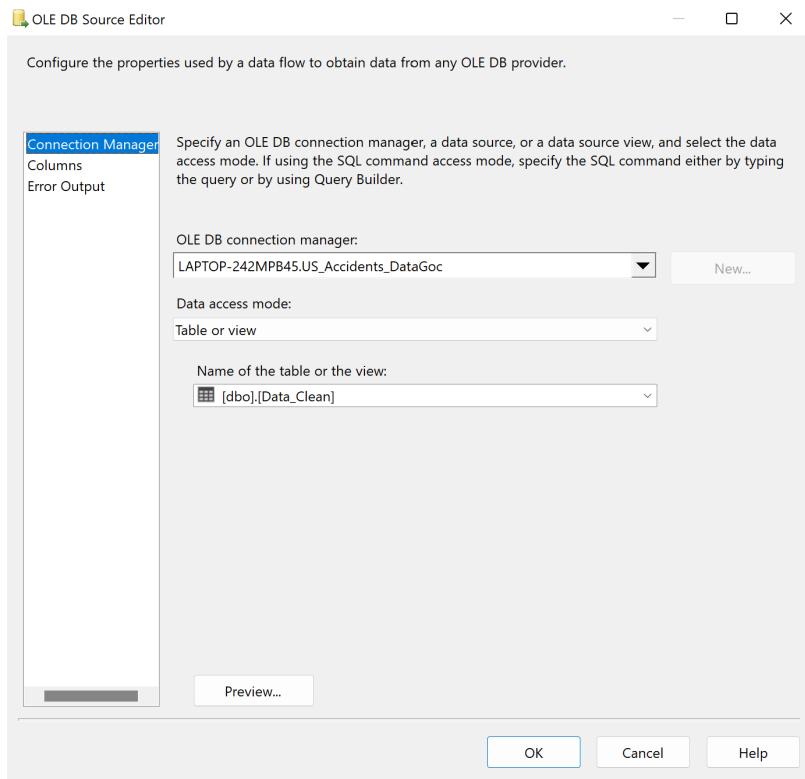
**Bước 1:** Tại Sequence Container Load Dimension Tables, tạo một Data Flow Task mới và đặt tên là Load Dim\_Airport.



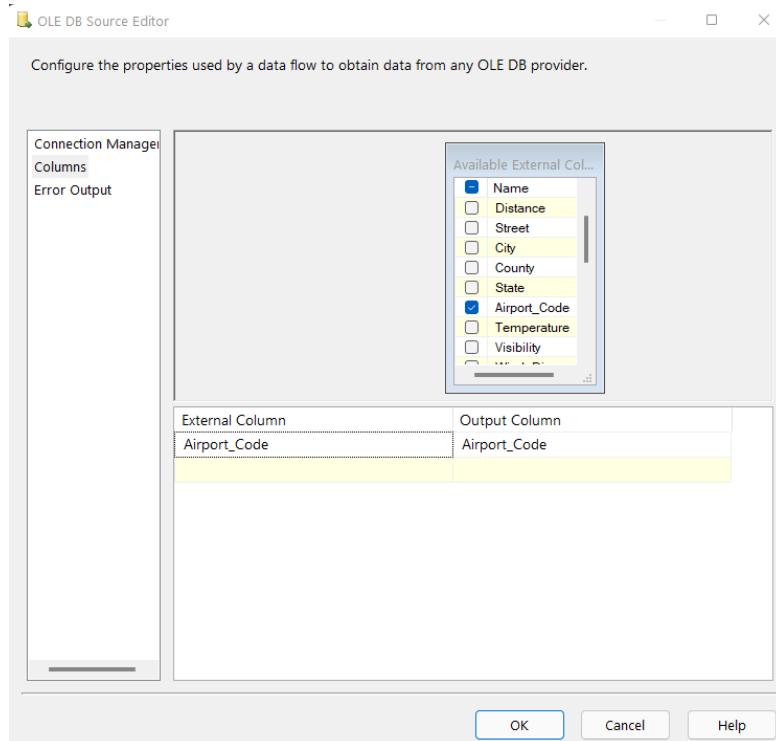
**Bước 2:** Double click vào Data Flow Task mới tạo và tạo mới một OLE DB Source.



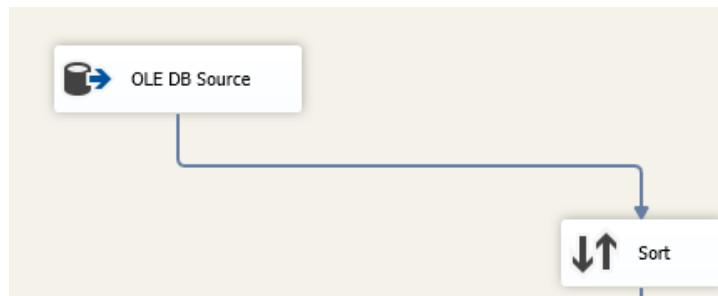
**Bước 3:** Double click vào OLE DB Source này, ta chọn connection manager là cơ sở dữ US\_Accident\_DataGoc và chọn bảng Data\_Clean.



**Bước 4:** Tại mục Columns, ta chọn các cột dữ liệu cần thiết cho bảng Dim\_Airport và lược bỏ các cột không cần thiết.

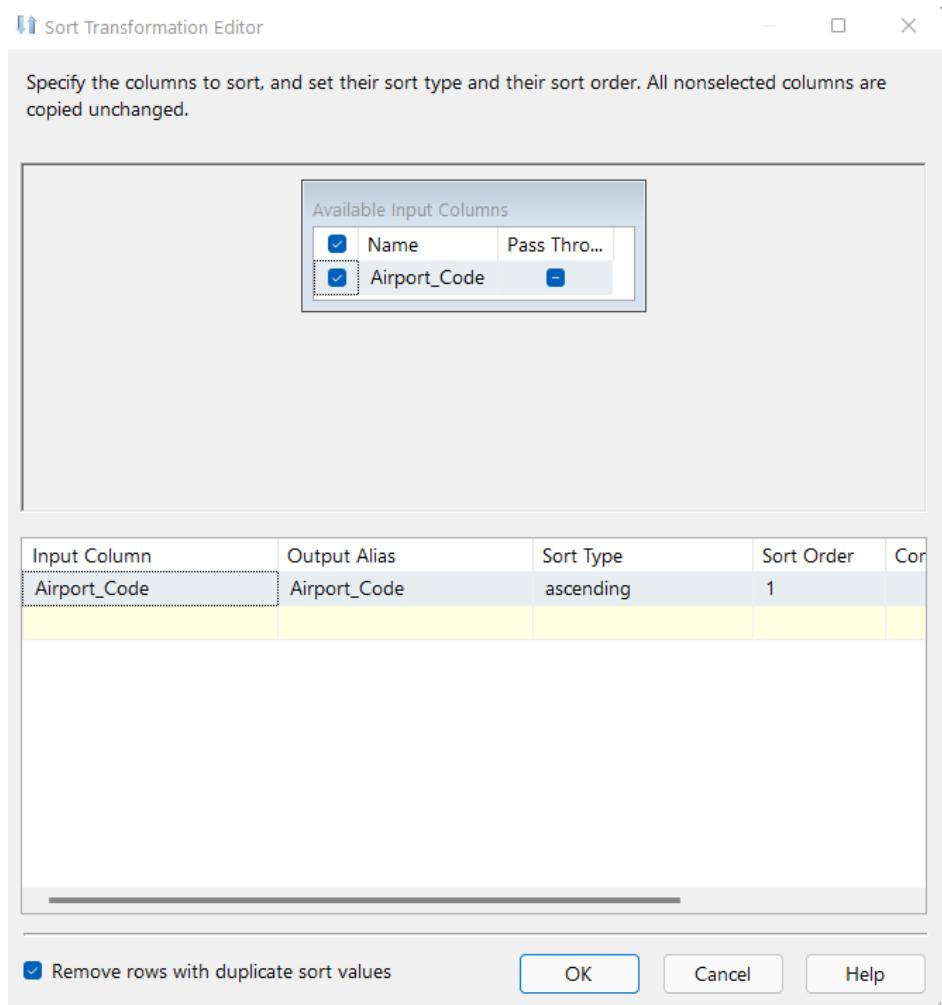


**Bước 5:** Tiếp theo, ta tạo ra một khối Sort để sắp xếp dữ liệu.

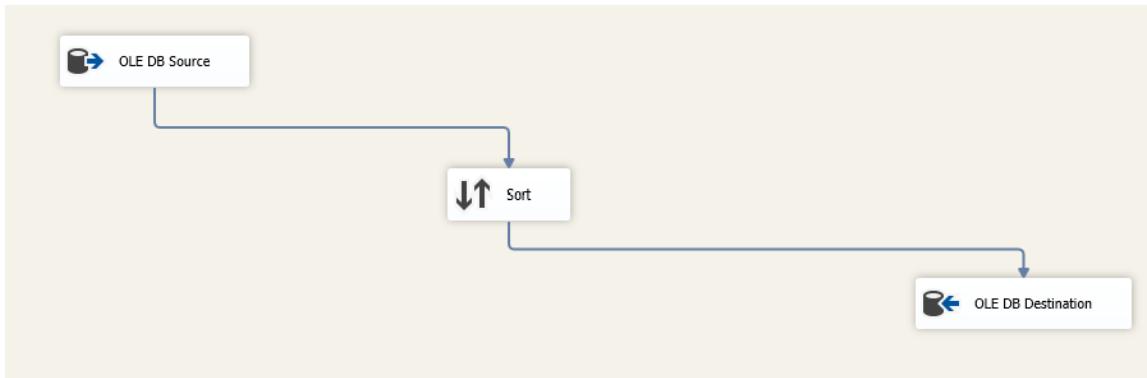


**Bước 6:** Double click vào Sort. Sau đó, chọn trường dữ liệu đang có và sắp xếp theo thứ tự tăng dần.

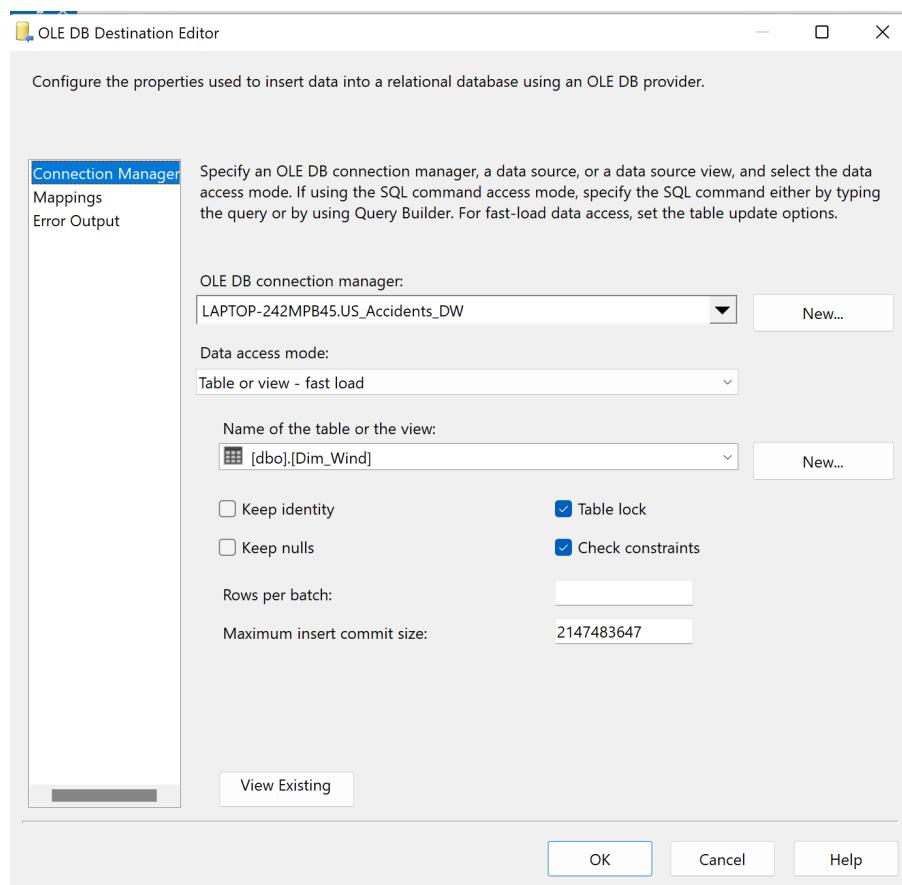
Cuối cùng, chọn OK để kết thúc thiết lập.



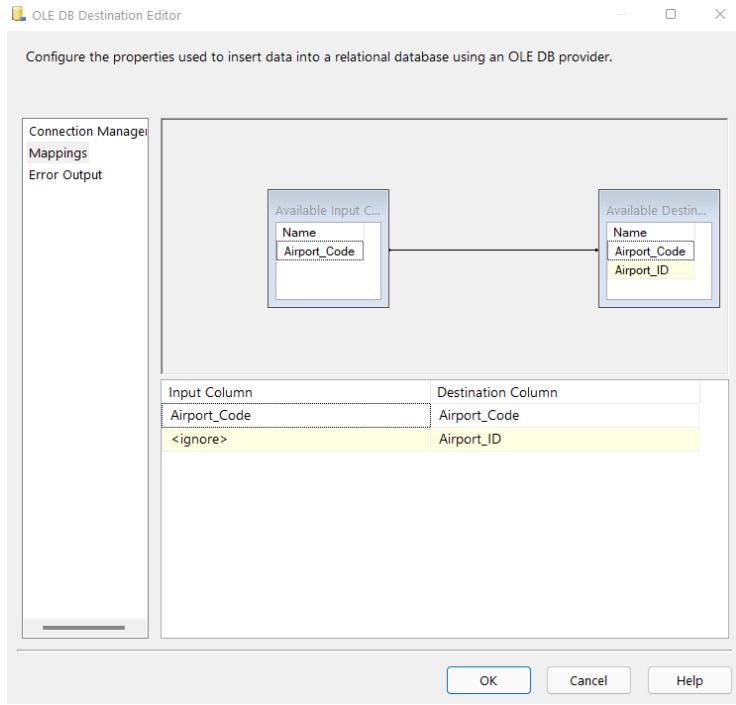
**Bước 7:** Tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong kho dữ liệu US\_Accidents\_DW.



**Bước 8:** Tiếp theo ta double click vào **OLE DB Destination** và chọn địa điểm đổ dữ liệu là bảng Dim\_Airport trong kho dữ liệu US\_Accidents\_DW.



Tiếp đến ta cần chọn mục Mappings để xem xét việc ánh xạ các cột dữ liệu

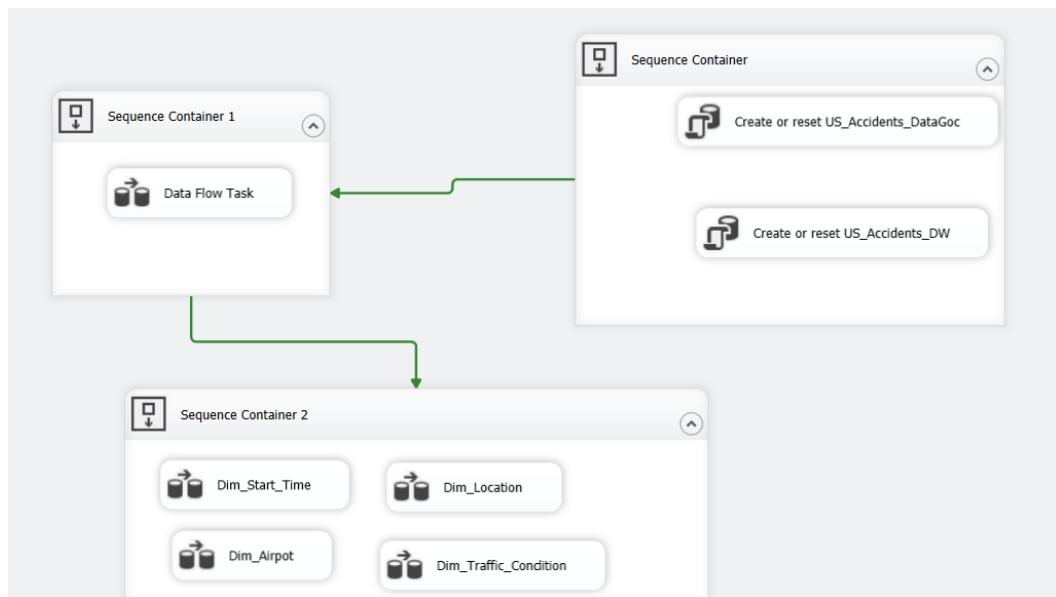


Cuối cùng, chọn OK để hoàn tất thiết lập.

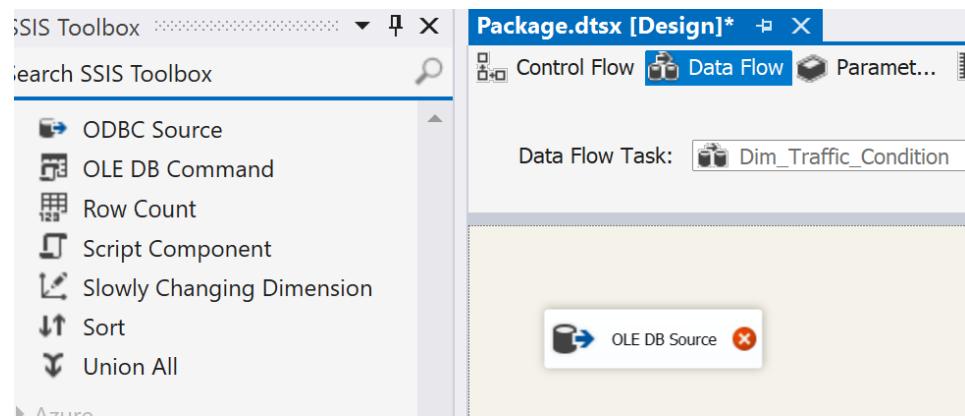
Như vậy, chúng ta đã hoàn tất quá trình đổ dữ liệu vào bảng Dim\_Airport.

#### 2.5.4. Bảng Dim\_Traffic\_Condition

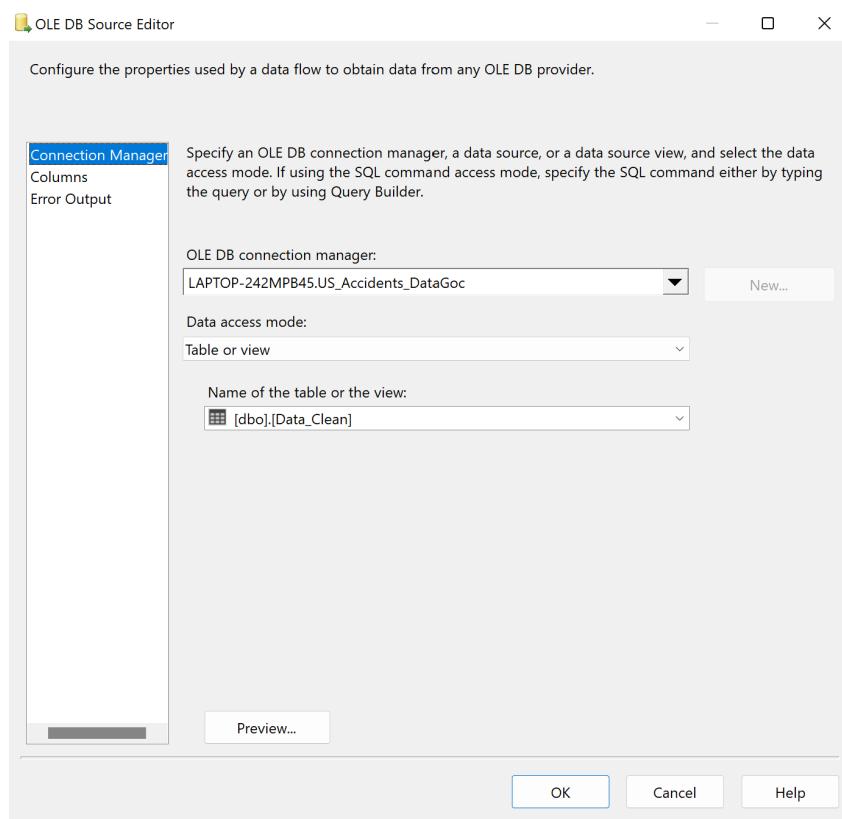
**Bước 1:** Tại Sequence Container Load Dimension Tables, tạo một Data Flow Task mới và đặt tên là Load Dim\_Traffic\_Condition.



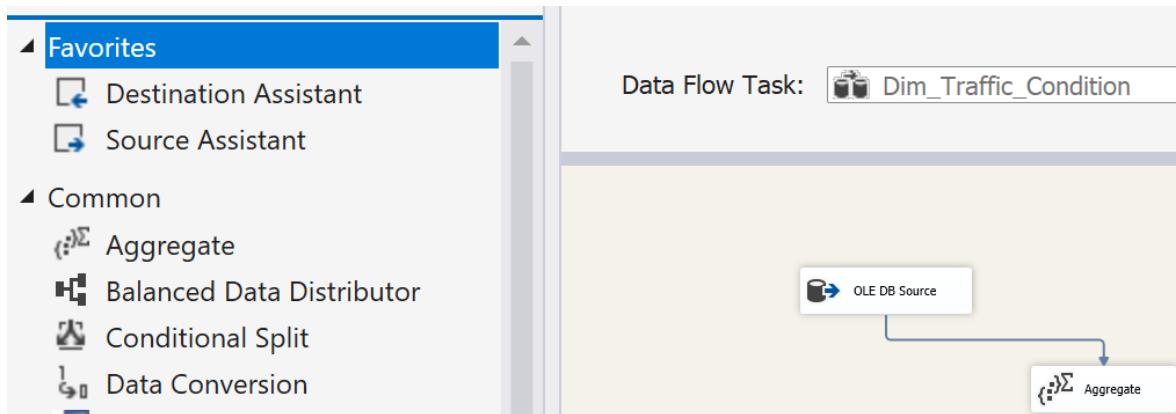
**Bước 2:** Double click vào Data Flow Task mới tạo và tạo mới một OLE DB Source.



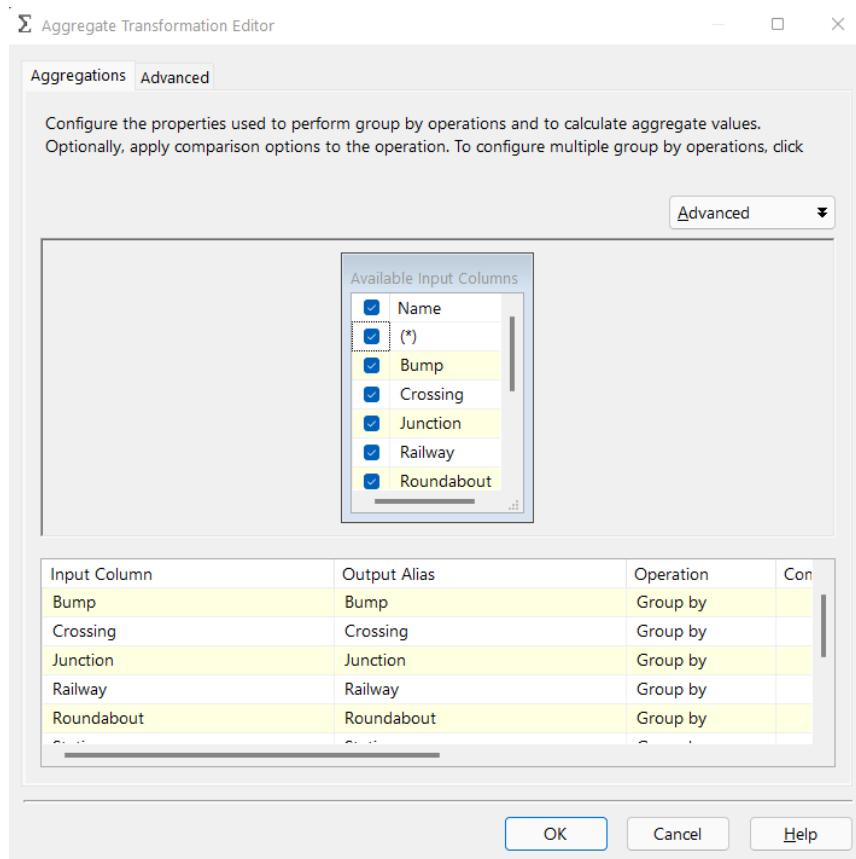
**Bước 3:** Double click vào OLE DB Source này, ta chọn connection manager là cơ sở dữ US\_Accident\_DataGoc và chọn bảng Data\_Clean. Sau đó chọn OK để hoàn tất.



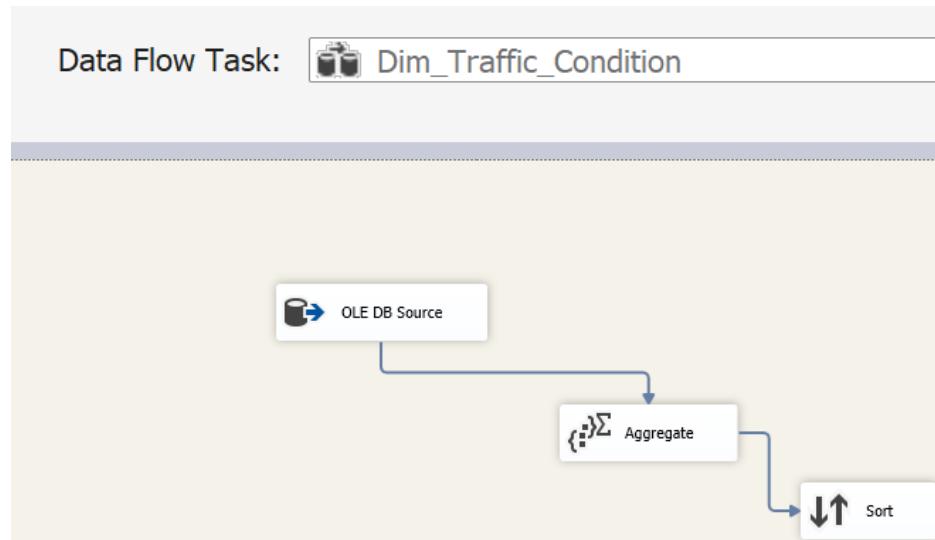
**Bước 4:** Tạo mới một Aggregate và tạo liên kết từ OLE DB Source đến Aggregate vừa tạo



**Bước 5:** Double click vào Aggregate, chọn cột dữ liệu liên quan đến bảng Dim\_Traffic\_Condition với Operation là groupby để gom nhóm dữ liệu. Sau đó chọn OK.



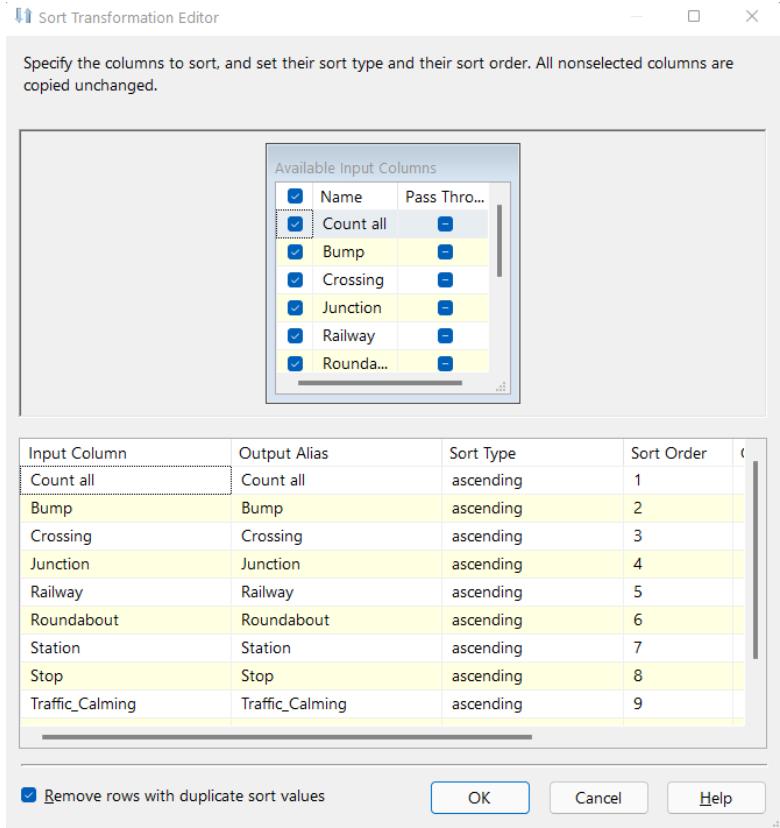
**Bước 6:** Tiếp theo, ta tạo ra một khối Sort để sắp xếp dữ liệu



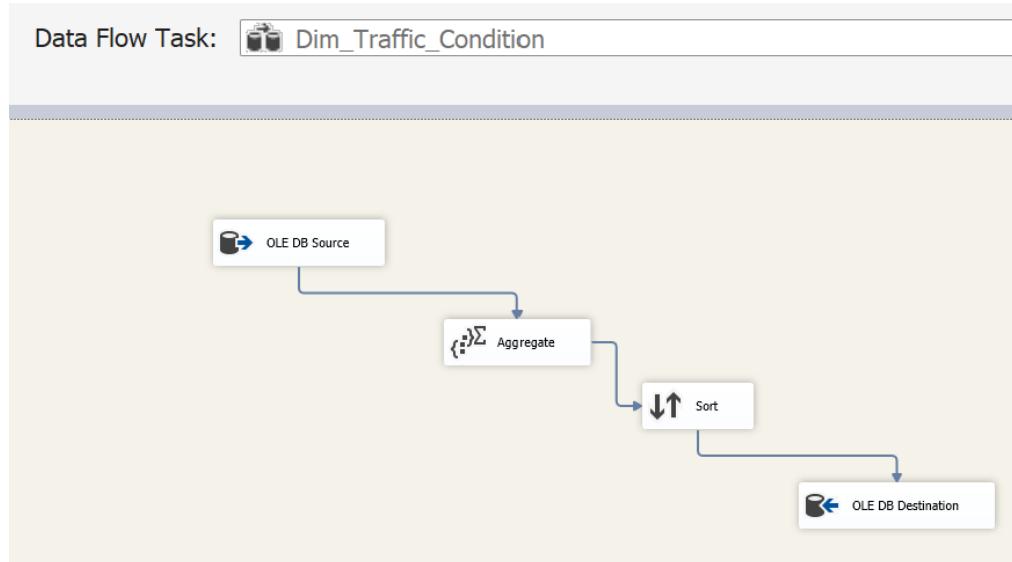
**Bước 7:** Double click vào Sort. Sau đó, chọn tất cả các trường dữ liệu đang có và sắp xếp theo thứ tự tăng dần.

Đồng thời, ta tick chọn checkbox “Remove rows with duplicate sort values” để loại bỏ các hàng có giá trị sắp xếp trùng lặp.

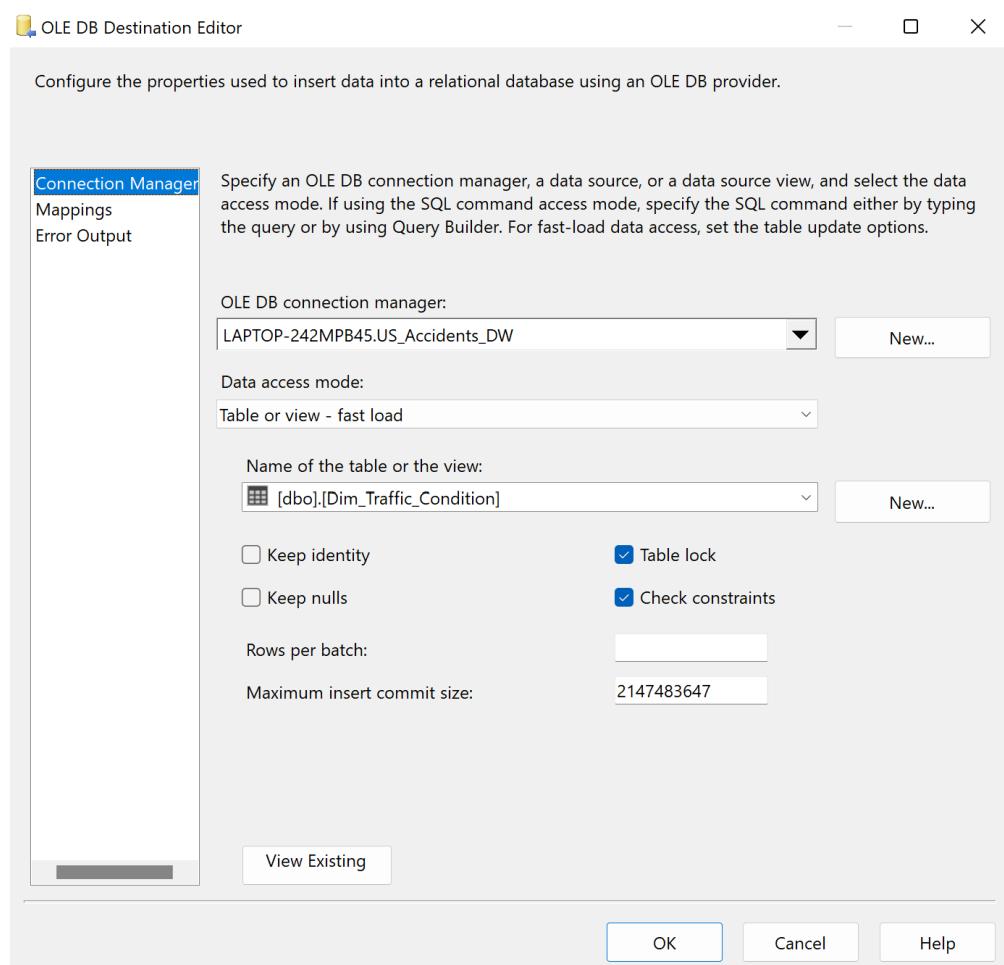
Cuối cùng, chọn OK để kết thúc thiết lập.



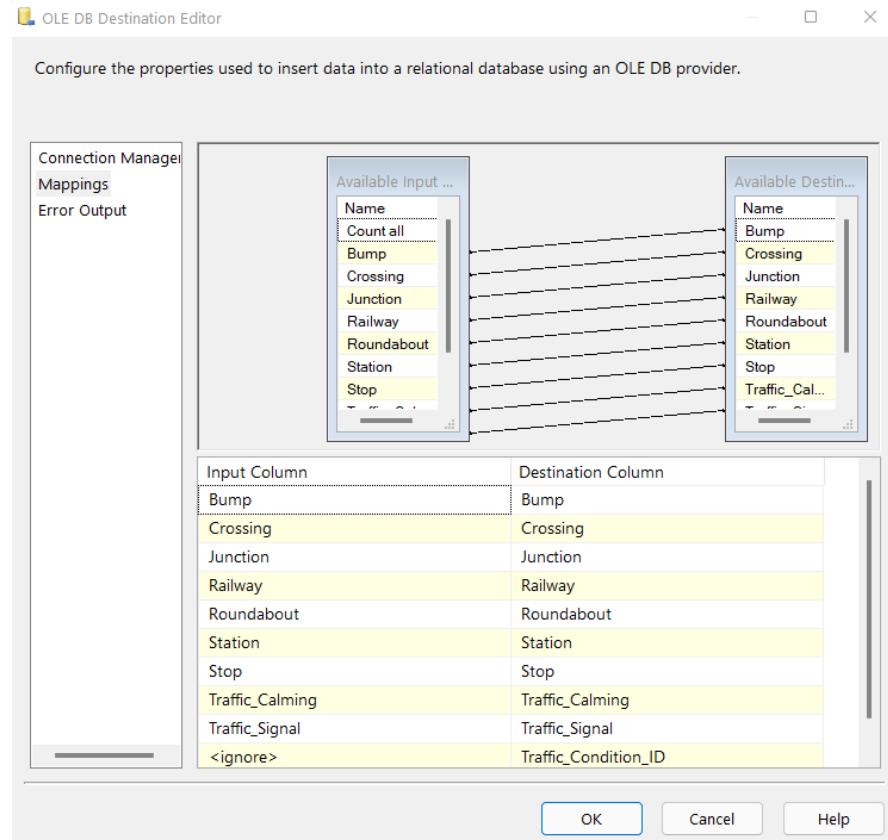
**Bước 9:** Tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong kho dữ liệu US\_Accidents\_DW.



**Bước 11:** Tiếp theo ta double click vào **OLE DB Destination** và chọn địa điểm để dữ liệu là bảng Dim\_Traffic\_Condition trong kho dữ liệu US\_Accidents\_DW.



Tiếp đến ta cần chọn mục Mappings để xem xét việc ánh xạ các cột dữ liệu

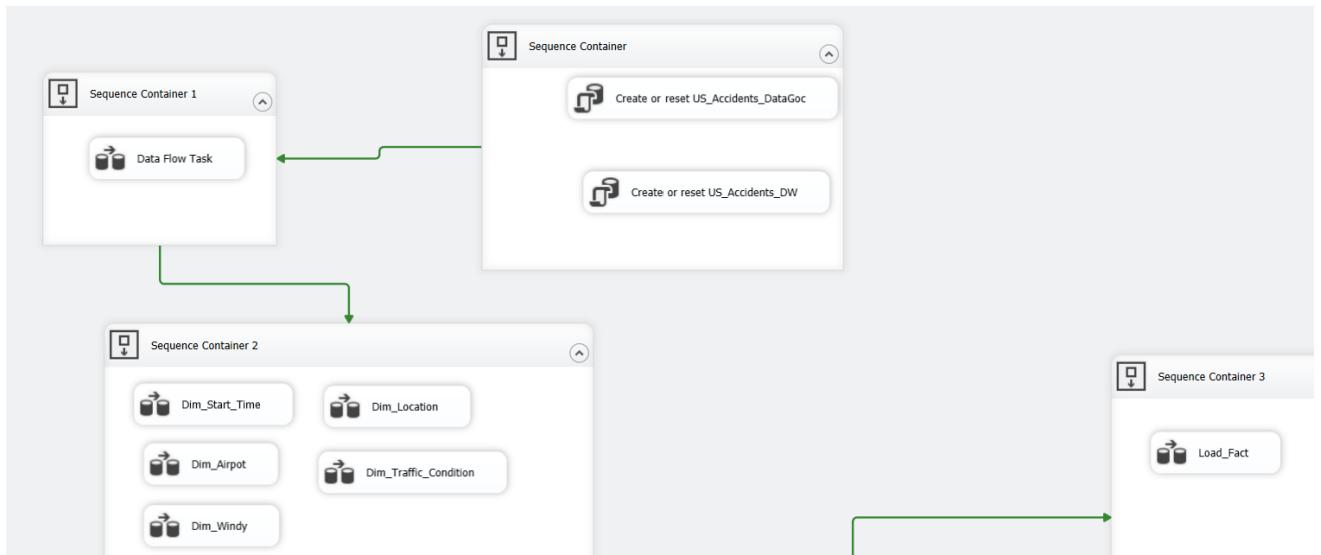


Cuối cùng, chọn OK để hoàn tất thiết lập.

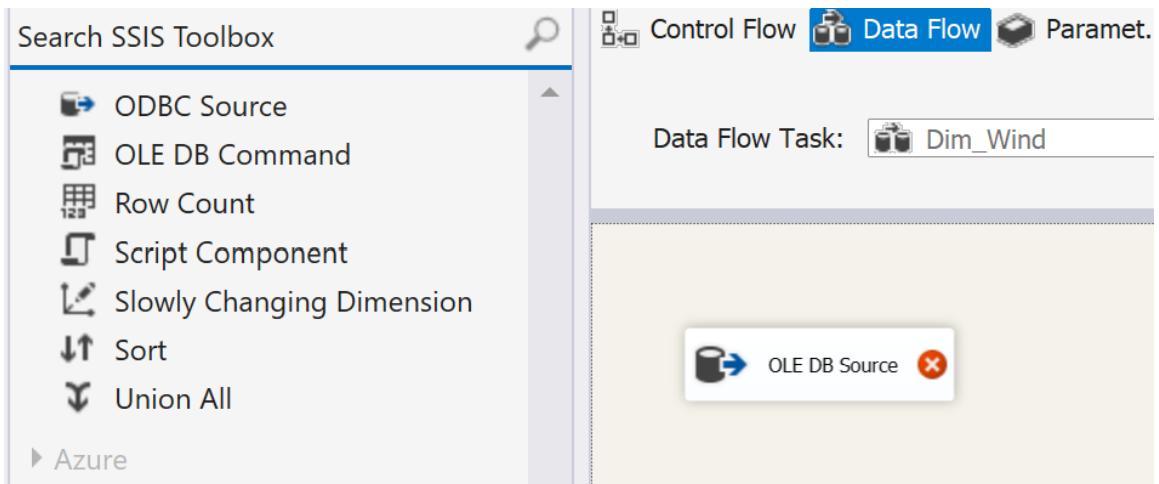
Như vậy, chúng ta đã hoàn tất quá trình đổ dữ liệu vào bảng Dim\_Traffic\_Condition.

### 2.5.5. Bảng Dim\_Wind

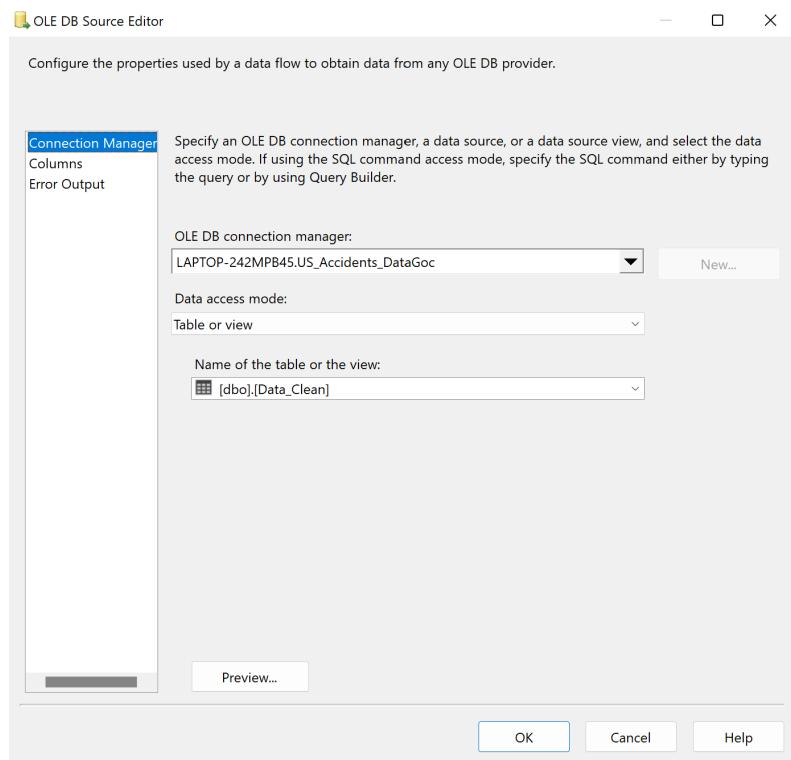
**Bước 1:** Tại Sequence Container Load Dimension Tables, tạo một Data Flow Task mới và đặt tên là Load Dim\_Wind.



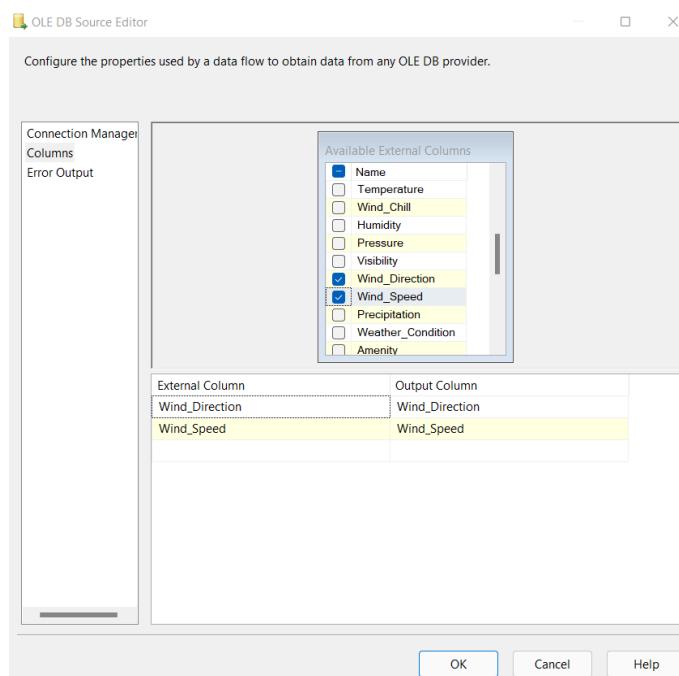
**Bước 2:** Double click vào Data Flow Task mới tạo và tạo mới một OLE DB Source.



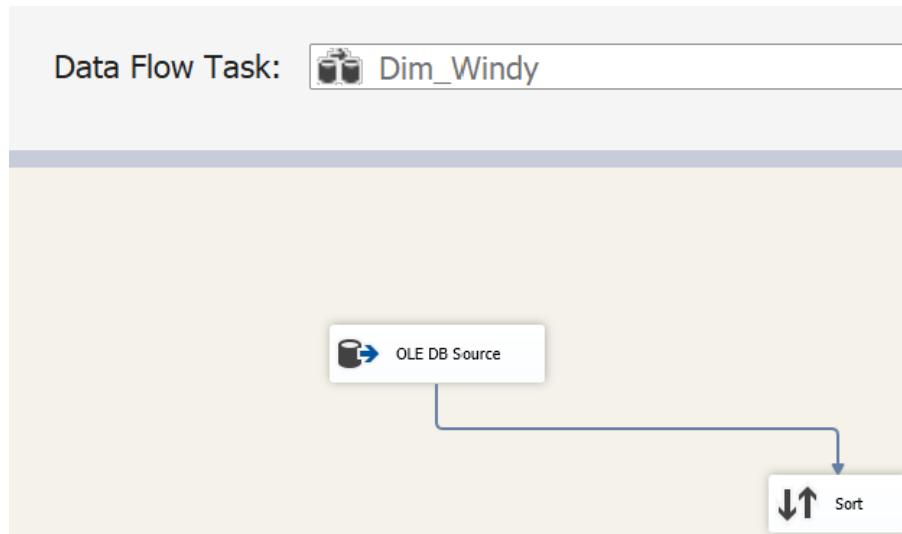
**Bước 3:** Double click vào OLE DB Source này, ta chọn connection manager là cơ sở dữ US\_Accident\_DataGoc và chọn bảng Data\_Clean.



**Bước 4:** Tại mục Columns, ta chọn các cột dữ liệu cần thiết cho bảng Dim\_Wind và lược bỏ các cột không cần thiết.



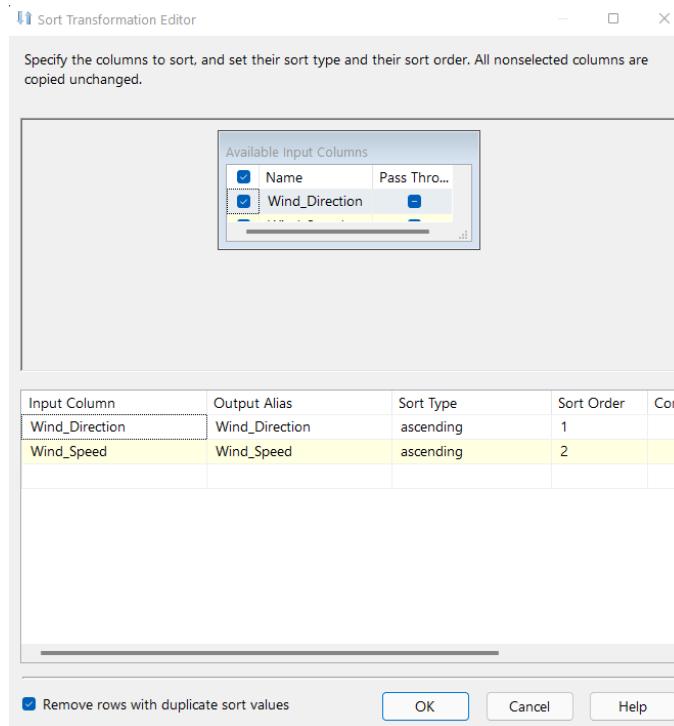
**Bước 5:** Tiếp theo, ta tạo ra một khối Sort để sắp xếp dữ liệu.



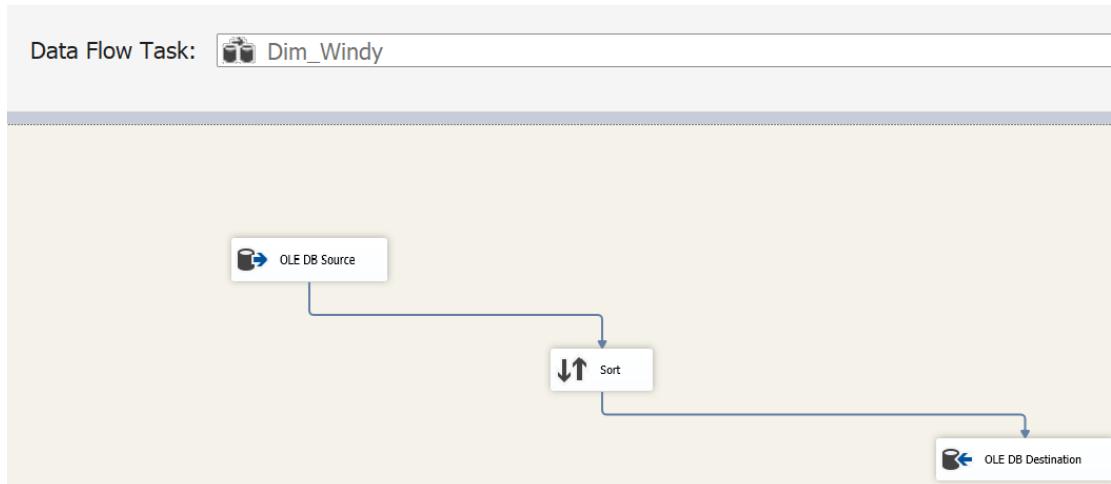
**Bước 6:** Double click vào Sort. Sau đó, chọn tất cả các trường dữ liệu đang có và sắp xếp theo thứ tự tăng dần.

Đồng thời, ta tick chọn checkbox “Remove rows with duplicate sort values” để loại bỏ các hàng có giá trị sắp xếp trùng lặp.

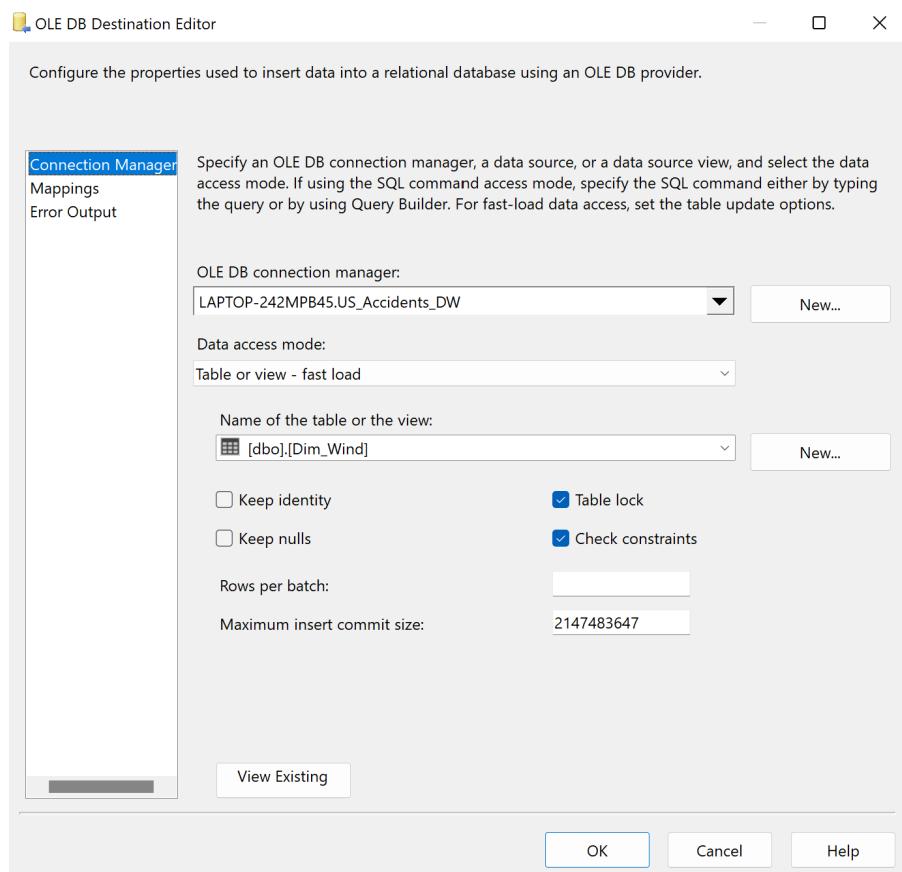
Cuối cùng, chọn OK để kết thúc thiết lập.



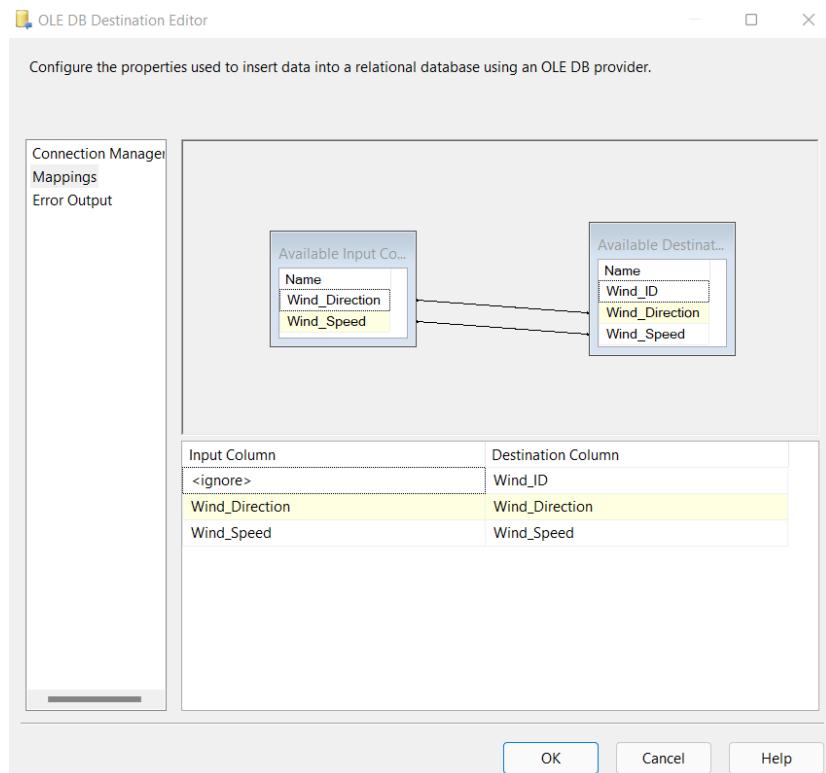
**Bước 7:** Tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong kho dữ liệu US\_Accidents\_DW.



**Bước 8:** Tiếp theo ta double click vào **OLE DB Destination** và chọn địa điểm đổ dữ liệu là bảng Dim\_Wind trong kho dữ liệu US\_Accidents\_DW.



Tiếp đến ta cần chọn mục Mappings để xem xét việc ánh xạ các cột dữ liệu

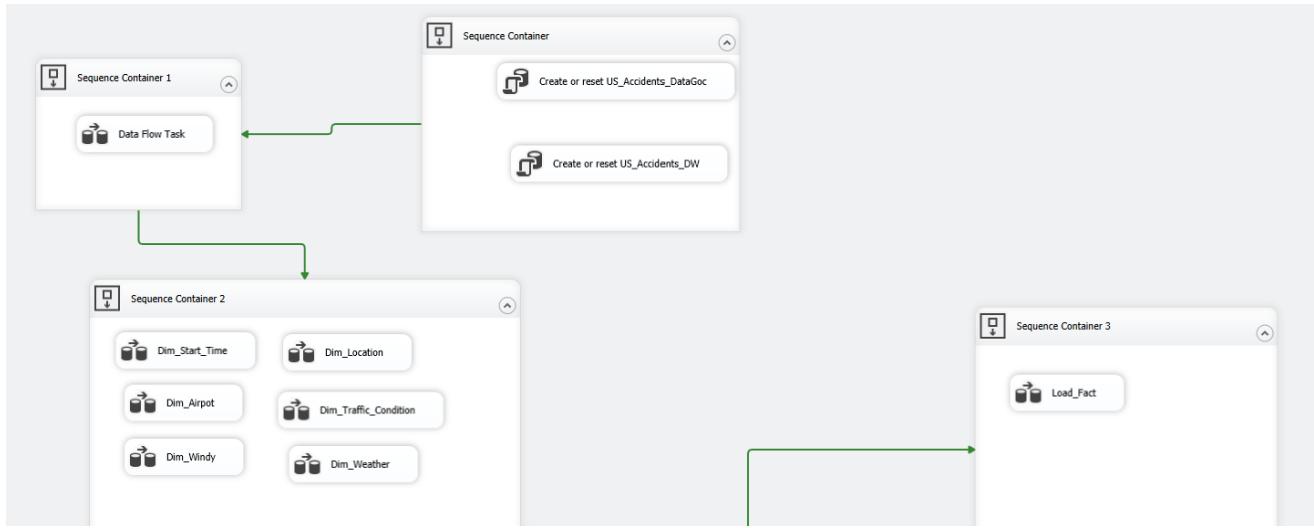


Cuối cùng, chọn OK để hoàn tất thiết lập.

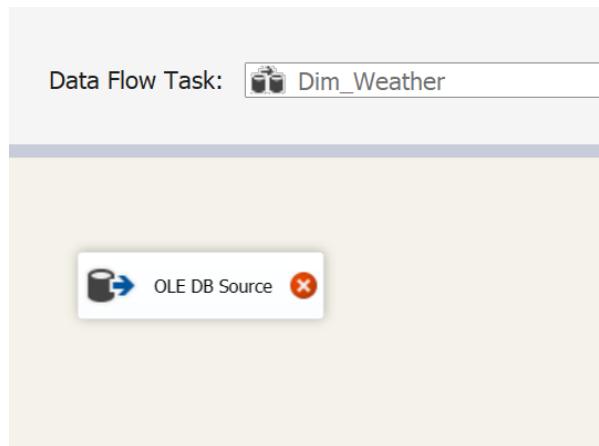
Như vậy, chúng ta đã hoàn tất quá trình đổ dữ liệu vào bảng Dim\_Wind.

### 2.5.6. Bảng Dim\_Weather

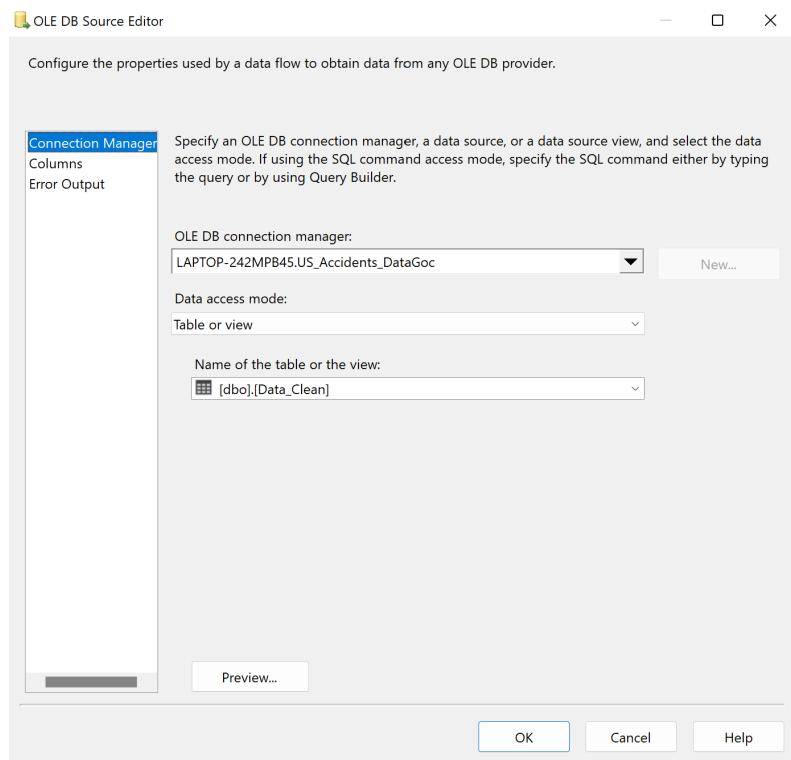
**Bước 1:** Tại Sequence Container Load Dimension Tables, tạo một Data Flow Task mới và đặt tên là Load Dim\_Weather.



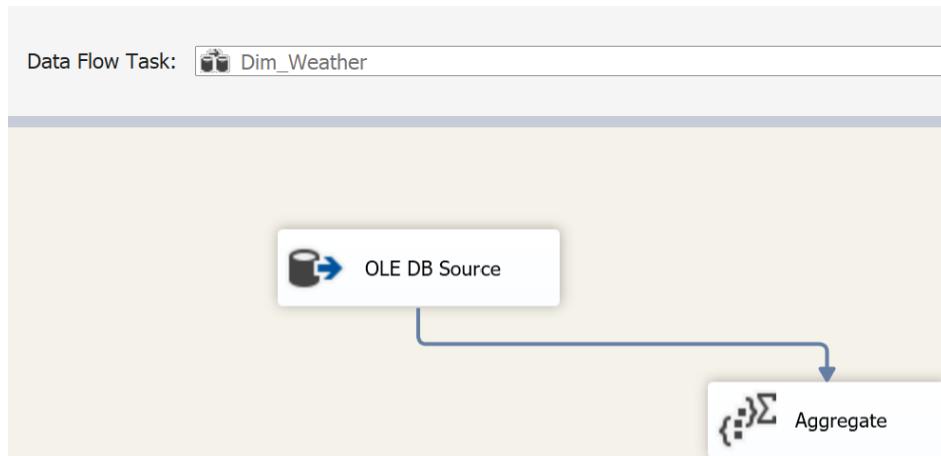
**Bước 2:** Double click vào Data Flow Task mới tạo và tạo mới một OLE DB Source.



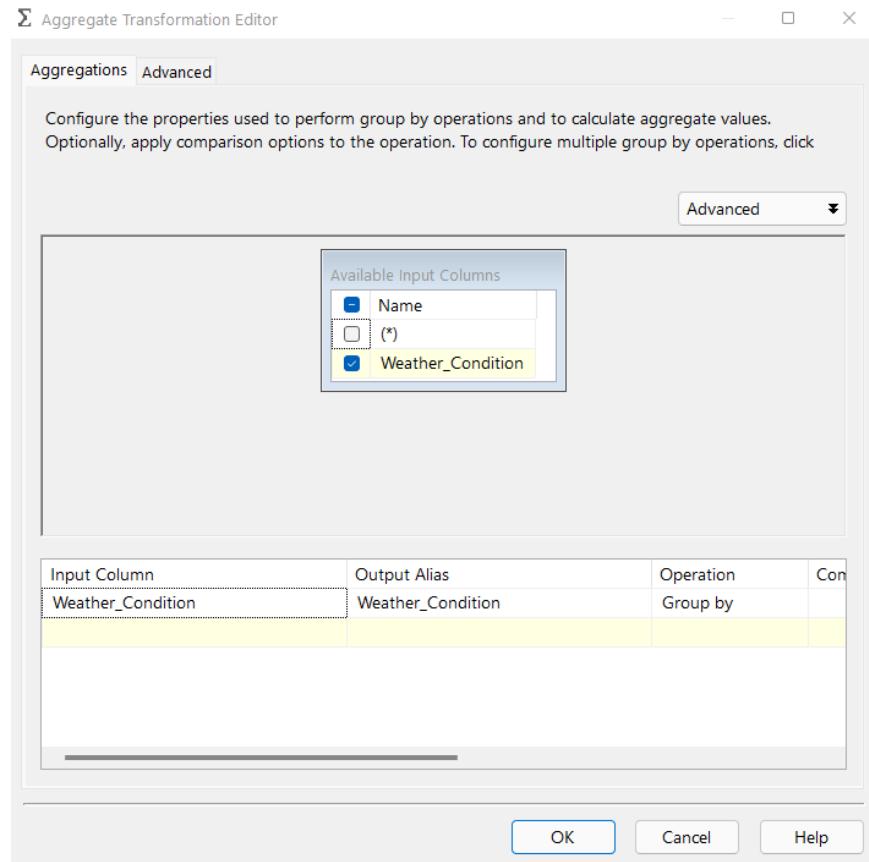
**Bước 3:** Double click vào OLE DB Source này, ta chọn connection manager là cơ sở dữ US\_Accident\_DataGoc và chọn bảng Data\_Clean. Sau đó chọn OK để hoàn tất.



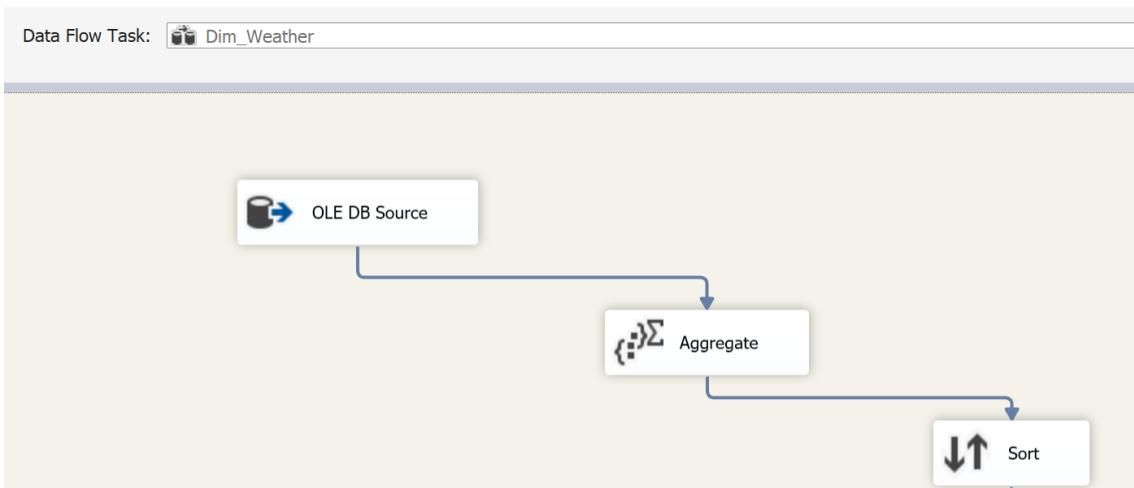
**Bước 4:** Tạo mới một Aggregate và tạo liên kết từ OLE DB Source đến Aggregate vừa tạo



**Bước 5:** Double click vào Aggregate, chọn cột dữ liệu liên quan đến bảng Dim\_Weather với Operation là groupby để gom nhóm dữ liệu. Sau đó chọn OK.



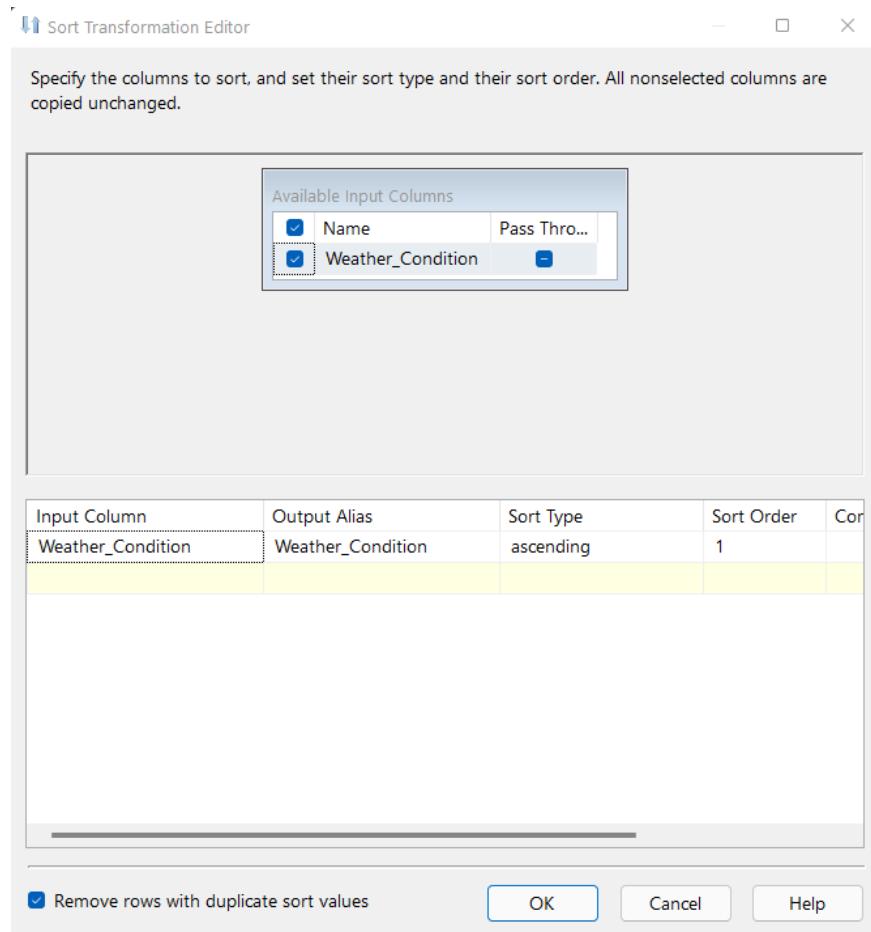
**Bước 6:** Tiếp theo, ta tạo ra một khối Sort để sắp xếp dữ liệu



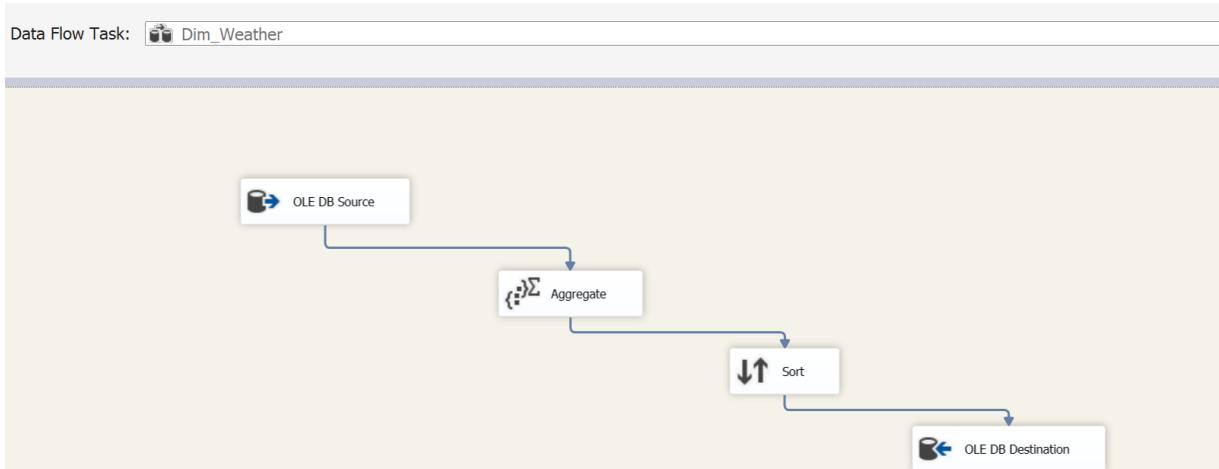
**Bước 7:** Double click vào Sort. Sau đó, chọn tất cả các trường dữ liệu đang có và sắp xếp theo thứ tự tăng dần.

Đồng thời, ta tick chọn checkbox “Remove rows with duplicate sort values” để loại bỏ các hàng có giá trị sắp xếp trùng lặp.

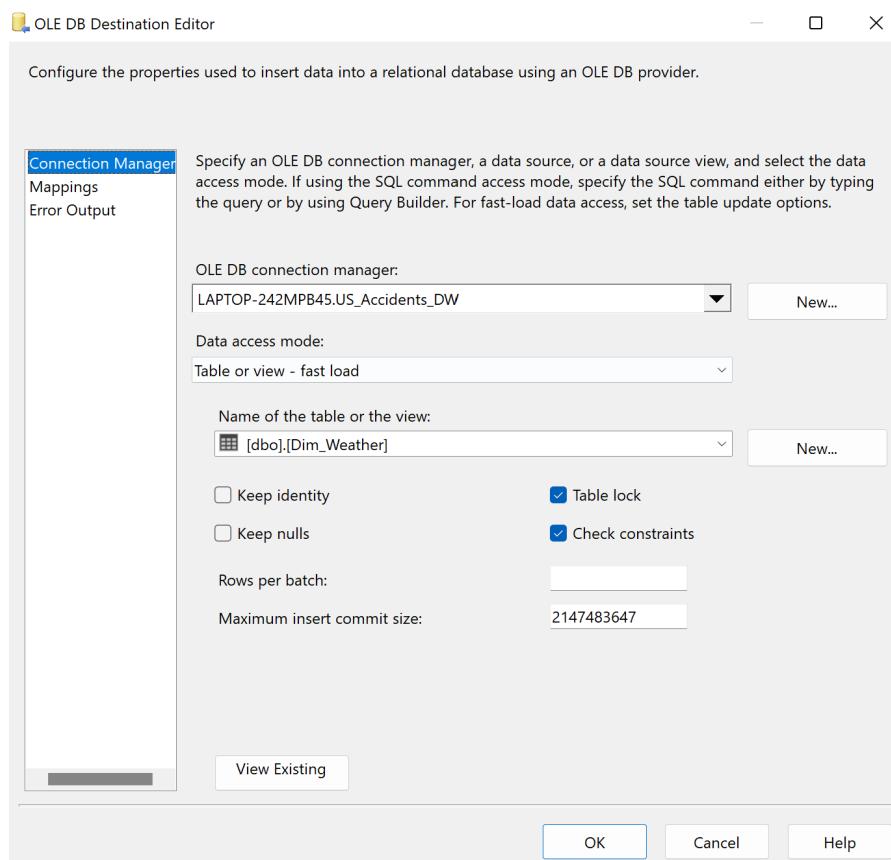
Cuối cùng, chọn OK để kết thúc thiết lập.



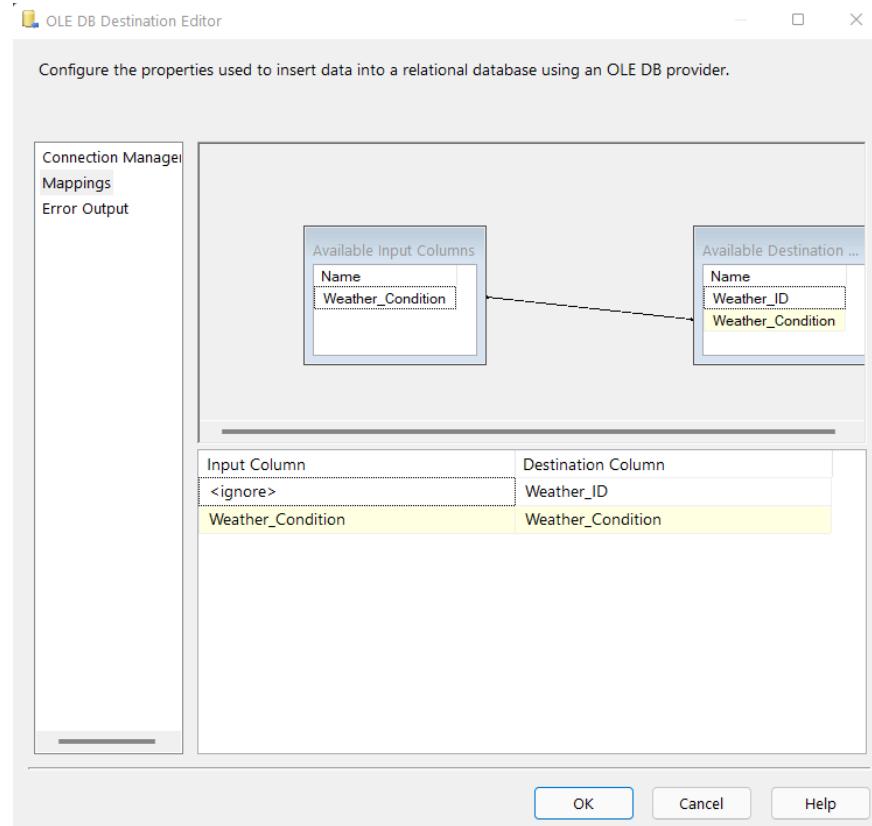
**Bước 9:** Tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong kho dữ liệu US\_Accidents\_DW.



**Bước 11:** Tiếp theo ta double click vào **OLE DB Destination** và chọn địa điểm đổ dữ liệu là bảng Dim\_Weather trong kho dữ liệu US\_Accidents\_DW.



Tiếp đến ta cần chọn mục Mappings để xem xét việc ánh xạ các cột dữ liệu

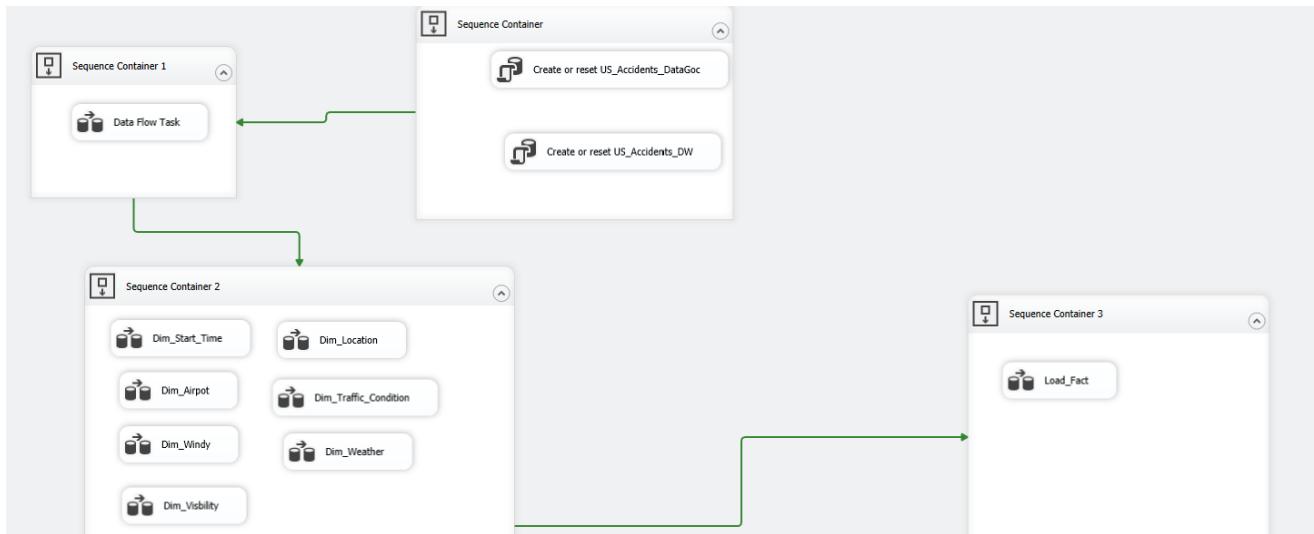


Cuối cùng, chọn OK để hoàn tất thiết lập.

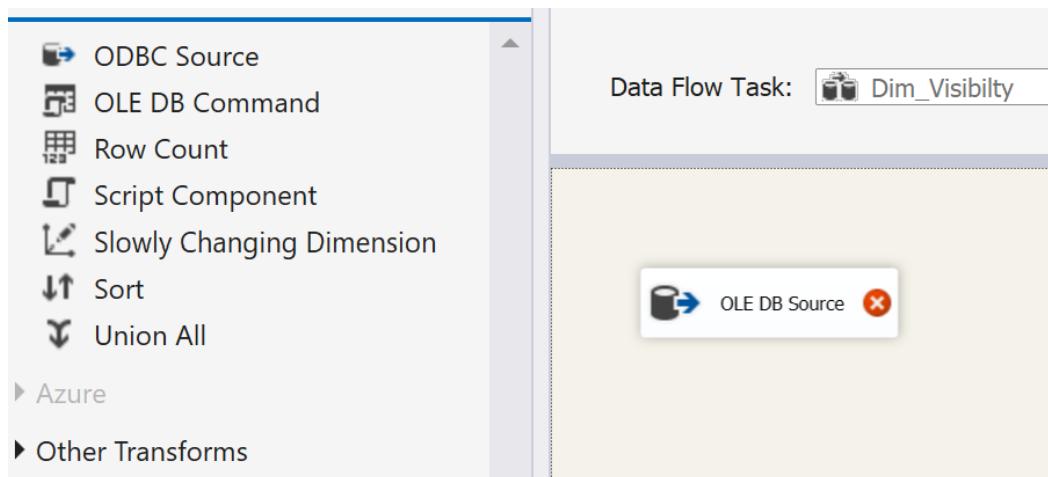
Như vậy, chúng ta đã hoàn tất quá trình đổ dữ liệu vào bảng Dim\_Weather.

### 2.5.7. Bảng Dim\_Visibility

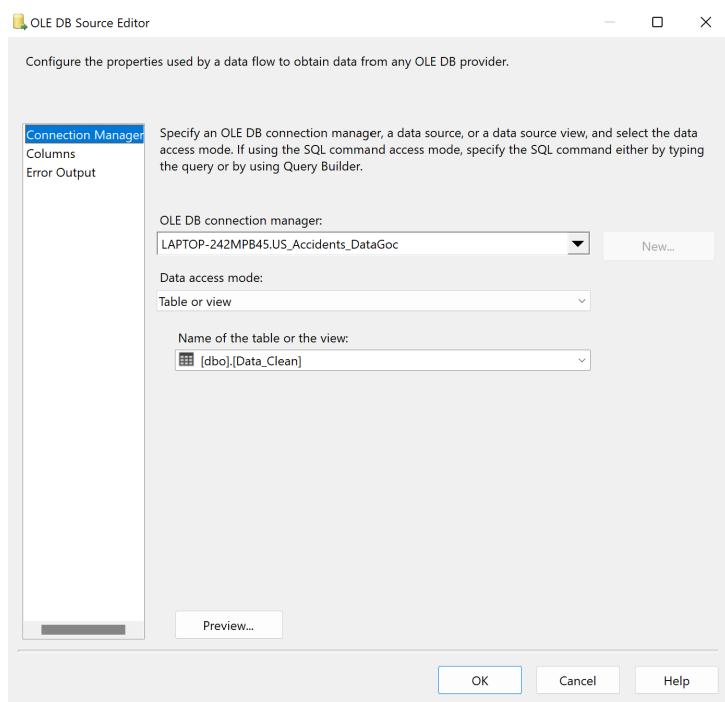
**Bước 1:** Tại Sequence Container Load Dimension Tables, tạo một Data Flow Task mới và đặt tên là Load Dim\_Visibility.



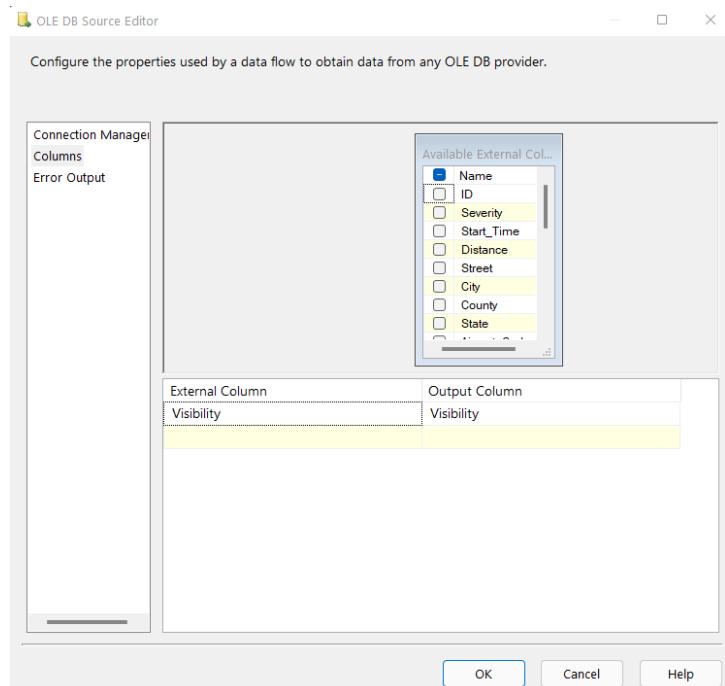
**Bước 2:** Double click vào Data Flow Task mới tạo và tạo mới một OLE DB Source.



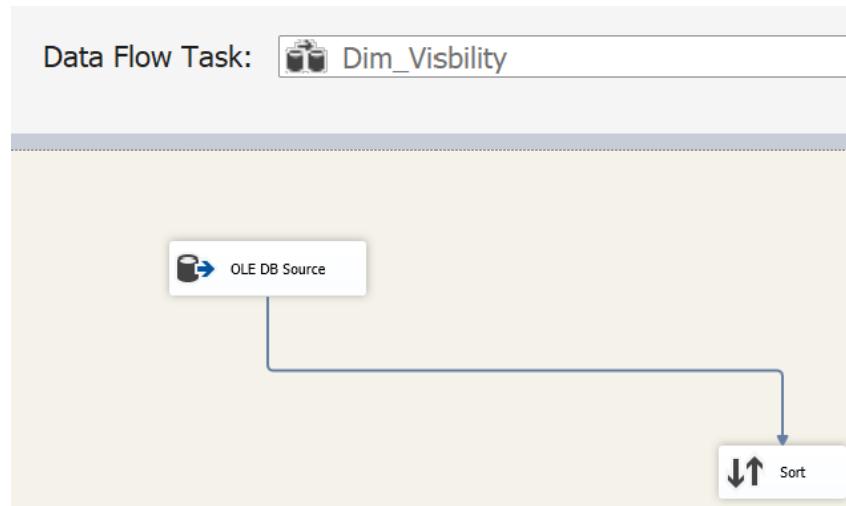
**Bước 3:** Double click vào OLE DB Source này, ta chọn connection manager là cơ sở dữ US\_Accident\_DataGoc và chọn bảng Data\_Clean.



**Bước 4:** Tại mục Columns, ta chọn các cột dữ liệu cần thiết cho bảng Dim\_Visibility và lược bỏ các cột không cần thiết.



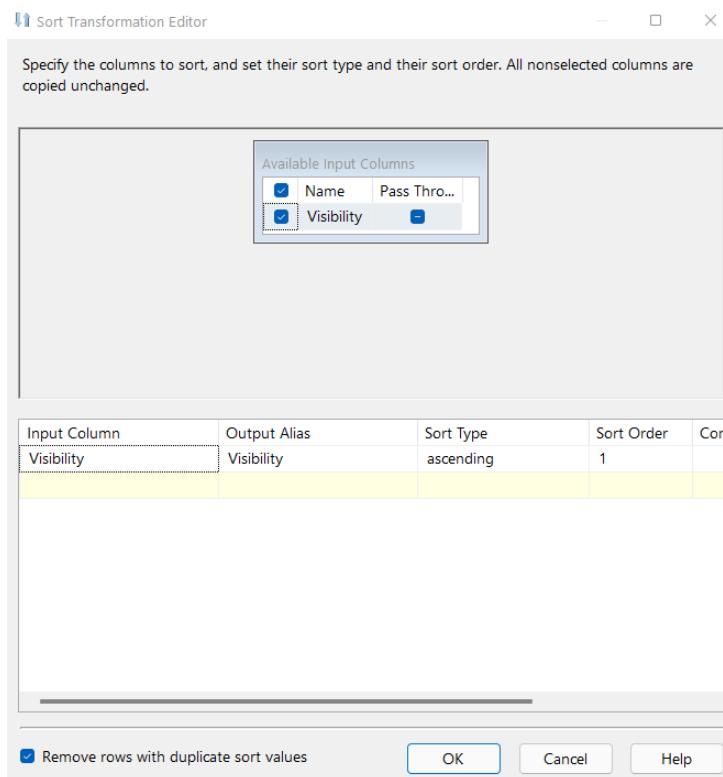
**Bước 5:** Tiếp theo, ta tạo ra một khối Sort để sắp xếp dữ liệu.



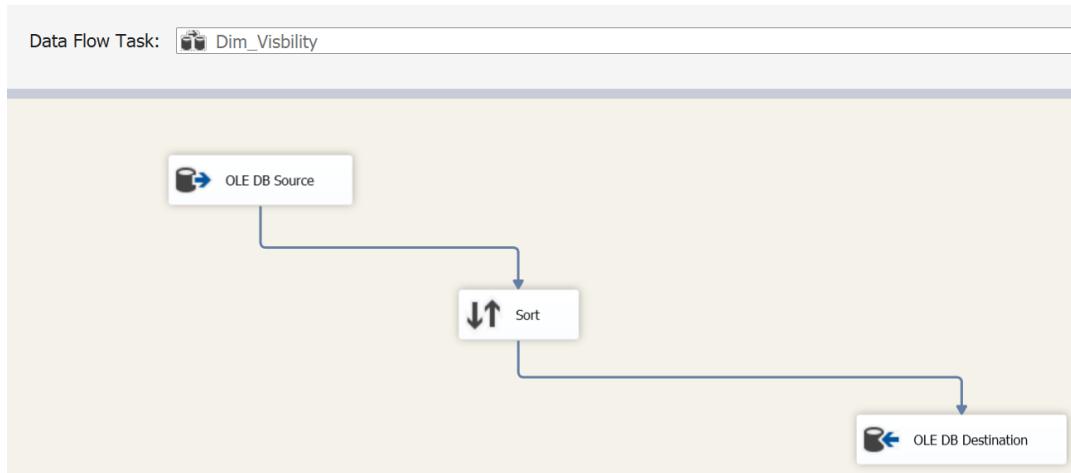
**Bước 6:** Double click vào Sort. Sau đó, chọn tất cả các trường dữ liệu đang có và sắp xếp theo thứ tự tăng dần.

Đồng thời, ta tick chọn checkbox “Remove rows with duplicate sort values” để loại bỏ các hàng có giá trị sắp xếp trùng lặp.

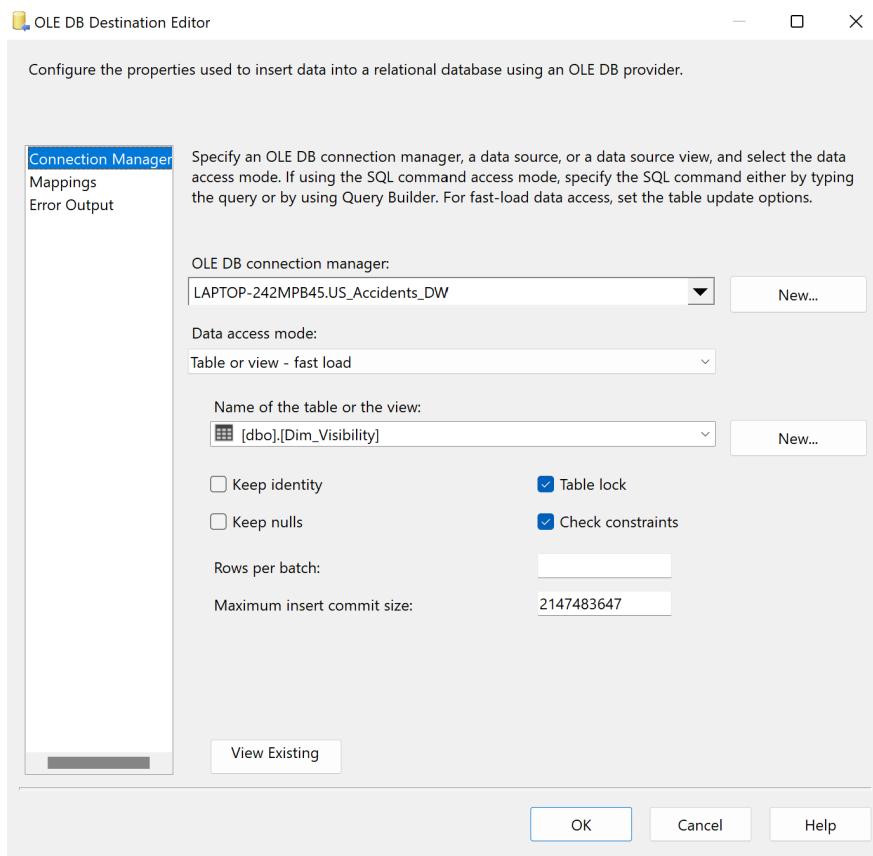
Cuối cùng, chọn OK để kết thúc thiết lập.



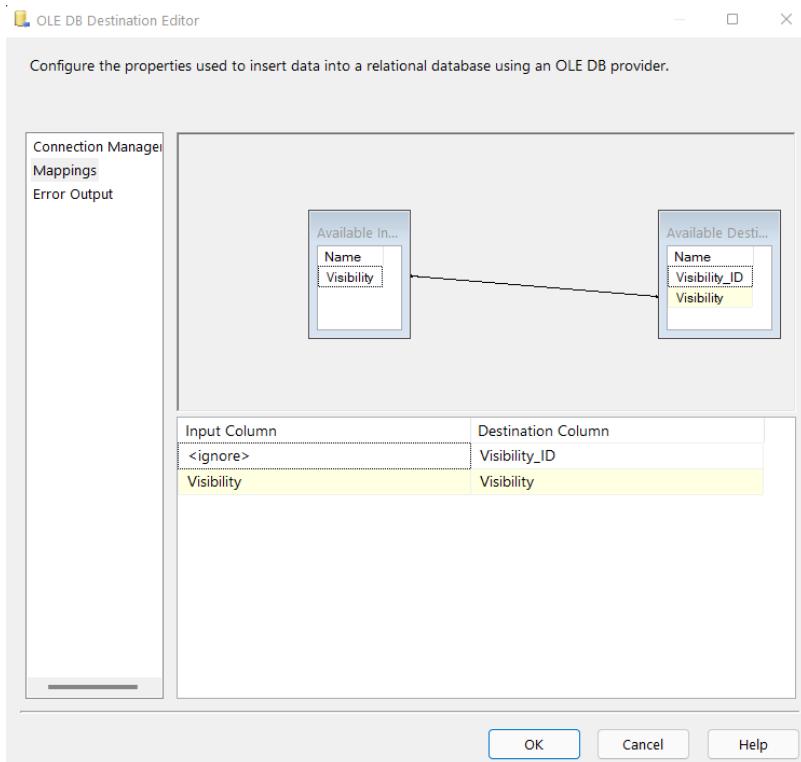
**Bước 7:** Tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong kho dữ liệu US\_Accidents\_DW.



**Bước 8:** Tiếp theo ta double click vào **OLE DB Destination** và chọn địa điểm đổ dữ liệu là bảng Dim\_Wind trong kho dữ liệu US\_Accidents\_DW.



Tiếp đến ta cần chọn mục Mappings để xem xét việc ánh xạ các cột dữ liệu

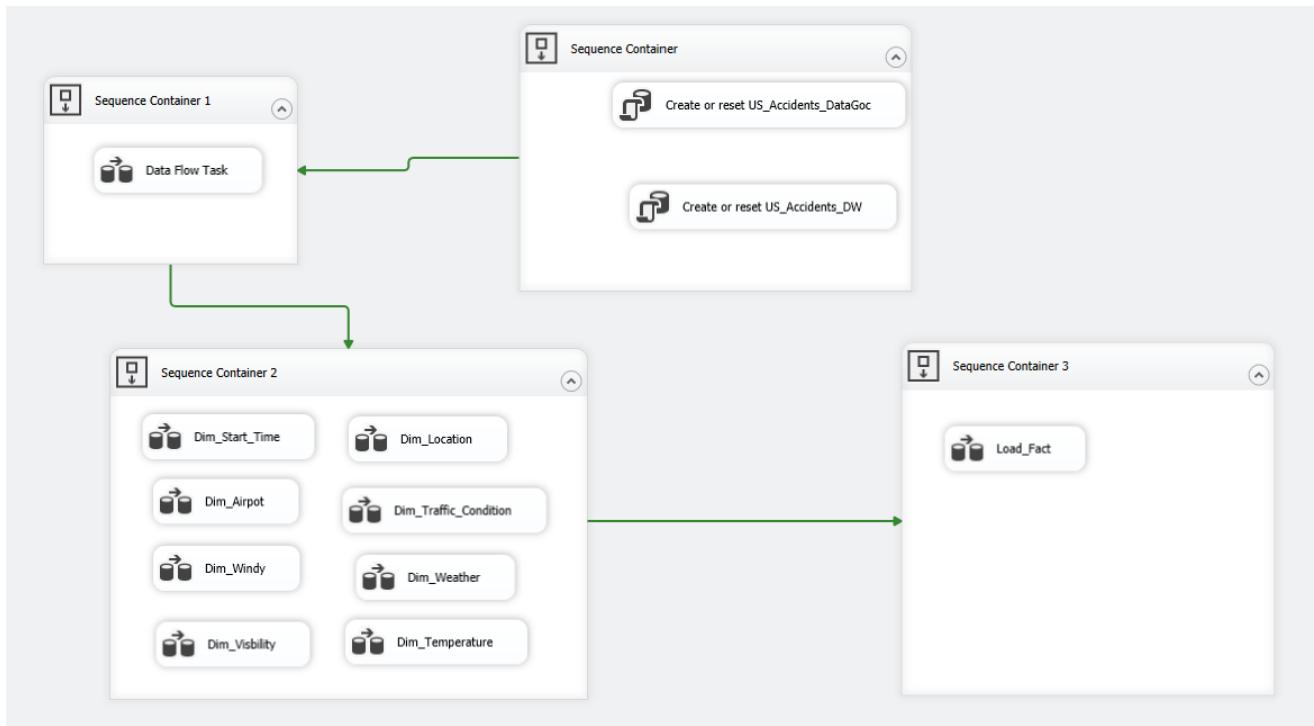


Cuối cùng, chọn OK để hoàn tất thiết lập.

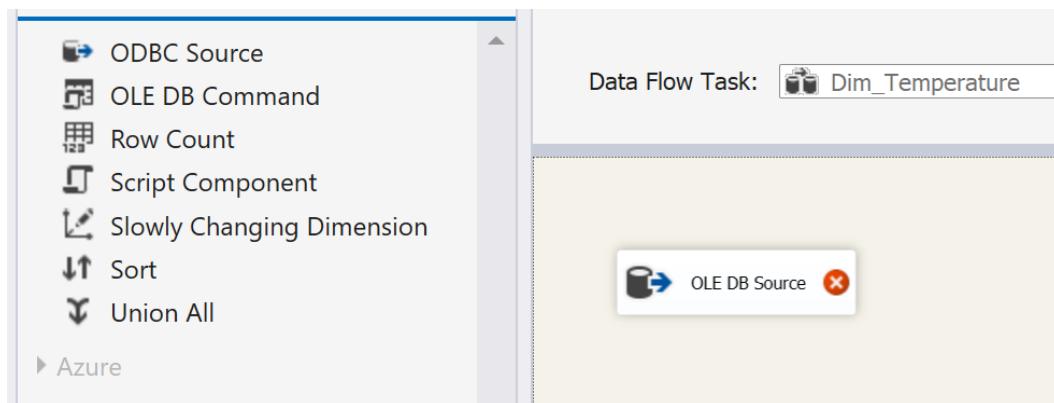
Như vậy, chúng ta đã hoàn tất quá trình đổ dữ liệu vào bảng Dim\_Visibility.

### 2.5.8. Bảng Dim\_Temperature

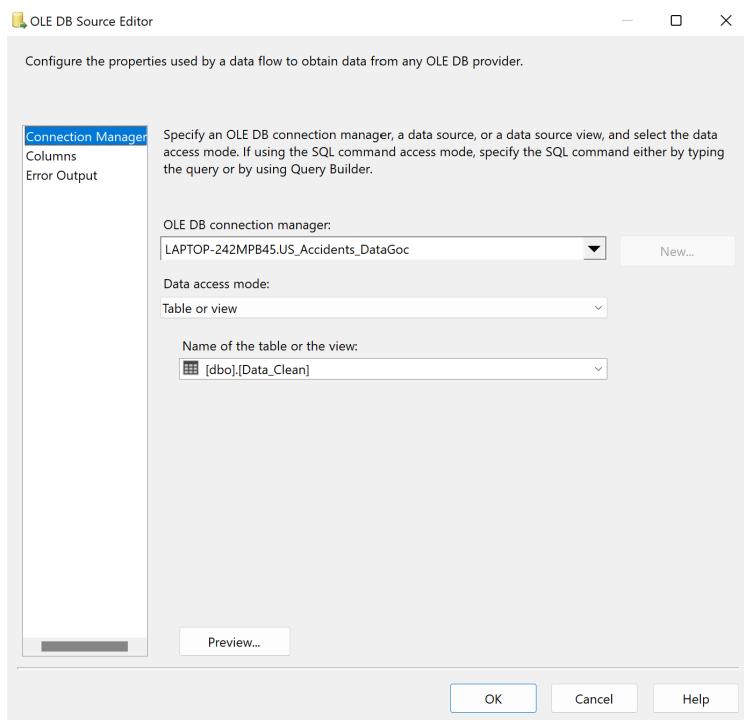
**Bước 1:** Tại Sequence Container Load Dimension Tables, tạo một Data Flow Task mới và đặt tên là Load Dim\_Temperature.



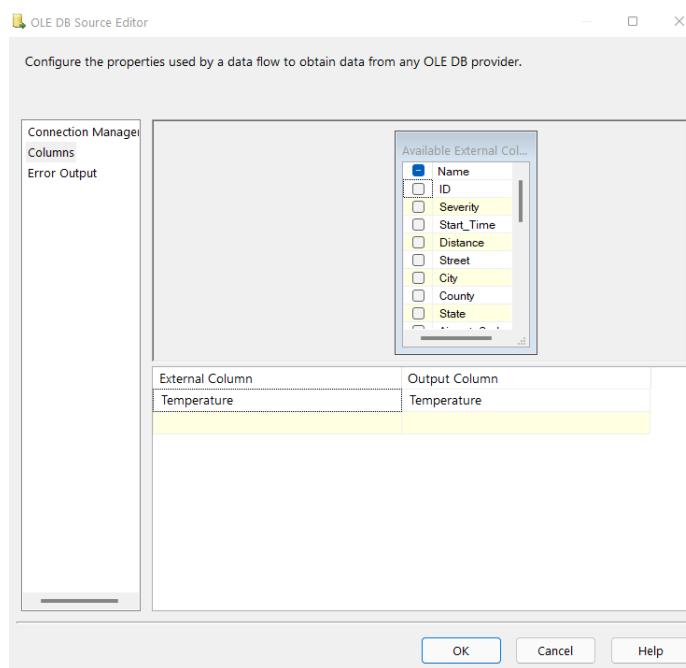
**Bước 2:** Double click vào Data Flow Task mới tạo và tạo mới một OLE DB Source.



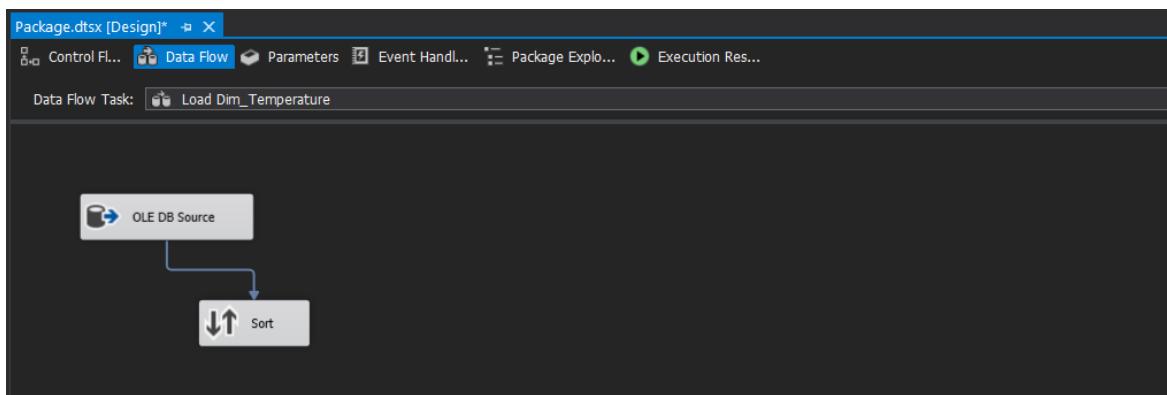
**Bước 3:** Double click vào OLE DB Source này, ta chọn connection manager là cơ sở dữ US\_Accident\_DataGoc và chọn bảng Data\_Clean.



**Bước 4:** Tại mục Columns, ta chọn các cột dữ liệu cần thiết cho bảng Dim\_Precipitation và lược bỏ các cột không cần thiết.



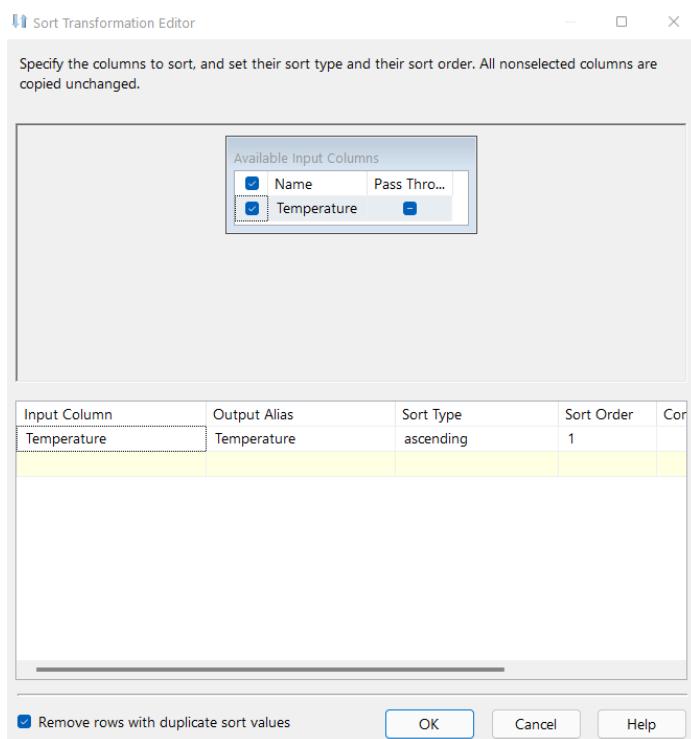
**Bước 5:** Tiếp theo, ta tạo ra một khối Sort để sắp xếp dữ liệu.



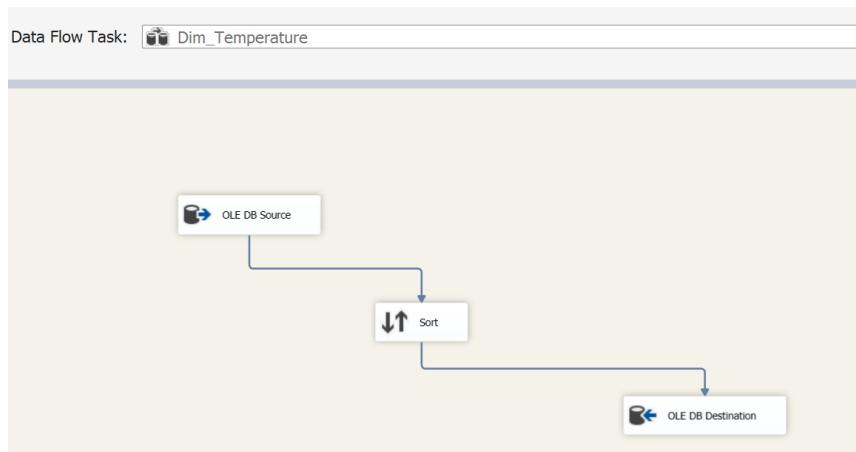
**Bước 6:** Double click vào Sort. Sau đó, chọn tất cả các trường dữ liệu đang có và sắp xếp theo thứ tự tăng dần.

Đồng thời, ta tick chọn checkbox “Remove rows with duplicate sort values” để loại bỏ các hàng có giá trị sắp xếp trùng lặp.

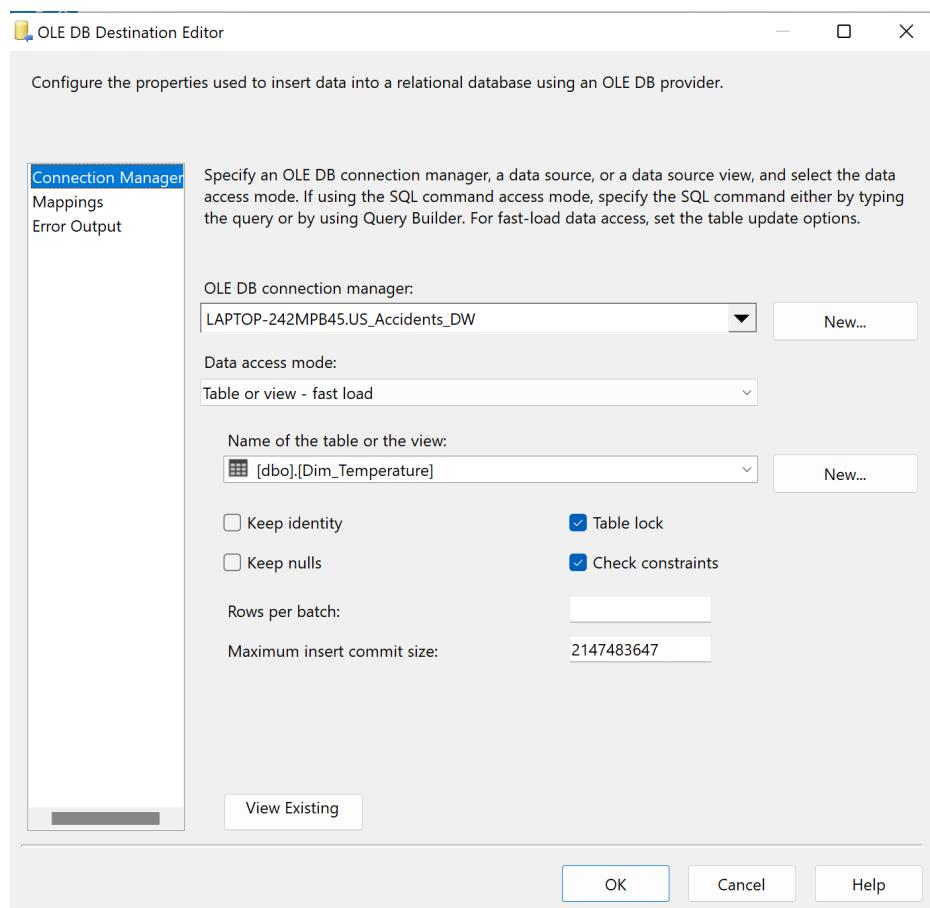
Cuối cùng, chọn OK để kết thúc thiết lập.



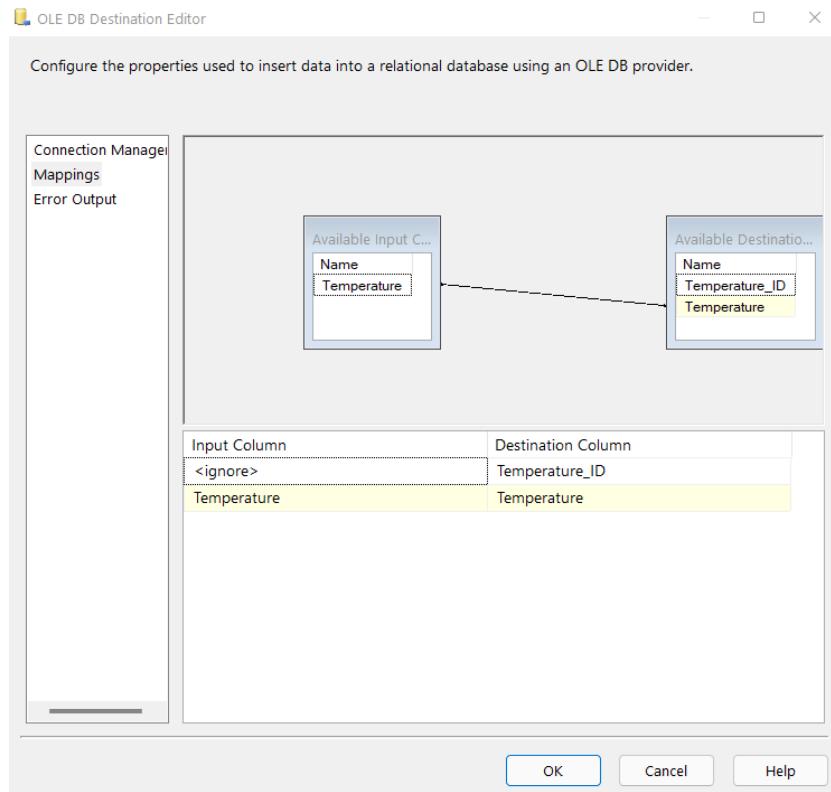
**Bước 7:** Tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong kho dữ liệu US\_Accidents\_DW.



**Bước 8:** Tiếp theo ta double click vào **OLE DB Destination** và chọn địa điểm đổ dữ liệu là bảng Dim\_Temperature trong kho dữ liệu US\_Accidents\_DW.



Tiếp đến ta cần chọn mục Mappings để xem xét việc ánh xạ các cột dữ liệu



Cuối cùng, chọn OK để hoàn tất thiết lập.

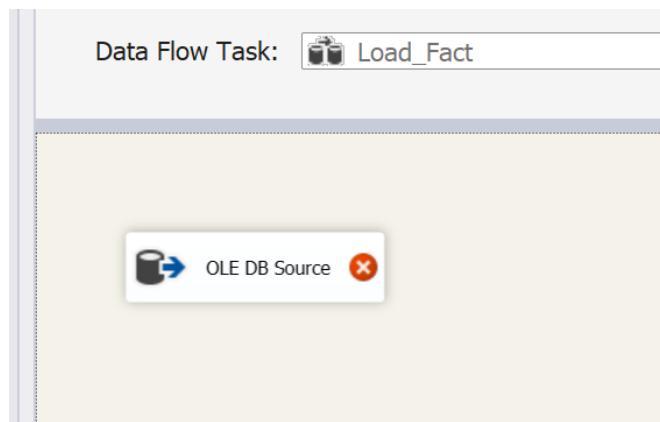
Như vậy, chúng ta đã hoàn tất quá trình đổ dữ liệu vào bảng Dim\_Temperature.

## 2.6. Đổ dữ liệu từ DataGoc vào bảng Fact

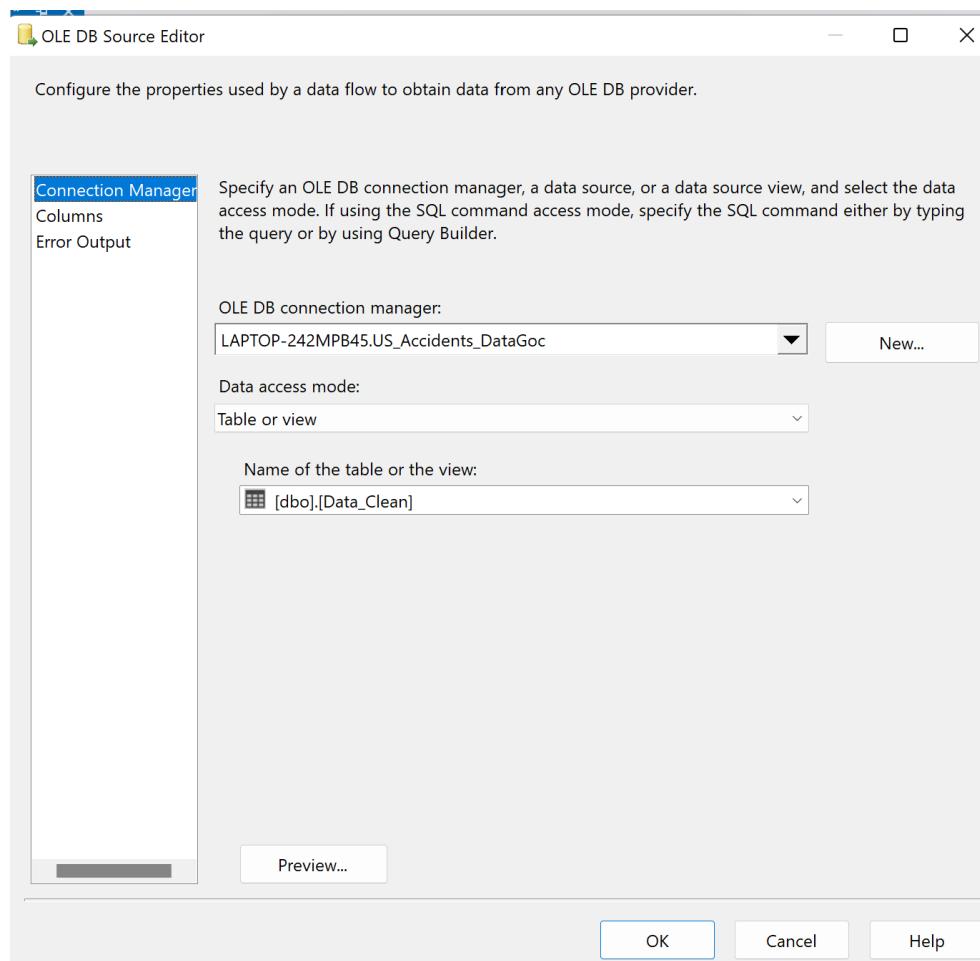
**Bước 1:** Tạo mới một Sequence Container và đặt tên là Load Fact Tables. Trong container này, tạo một Data Flow Task và đặt tên là Load Fact\_US\_Accidents.



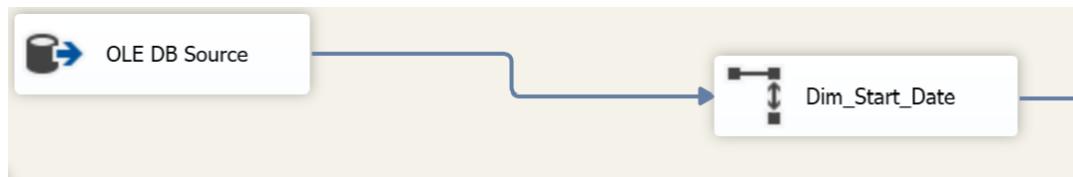
**Bước 2:** Double click vào Data Flow Task mới tạo và tạo mới một OLE DB Source.



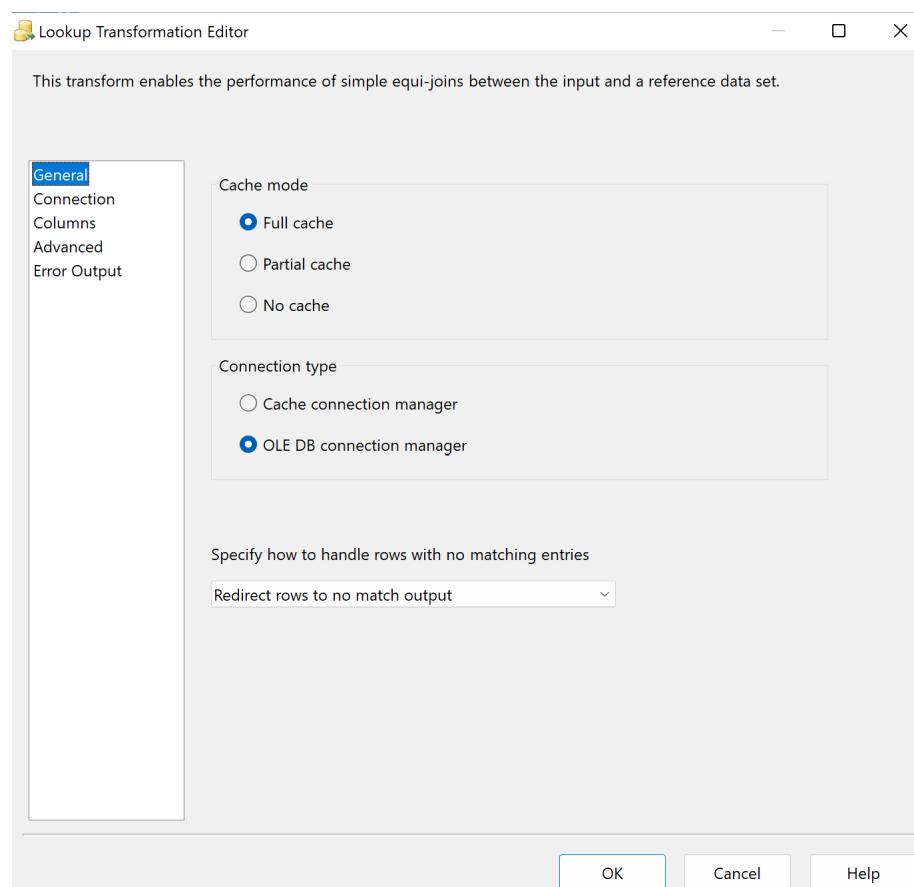
**Bước 3:** Double click vào OLE DB Source này, ta chọn connection manager là cơ sở dữ US\_Accident\_DataGoc và chọn bảng Data\_Clean.



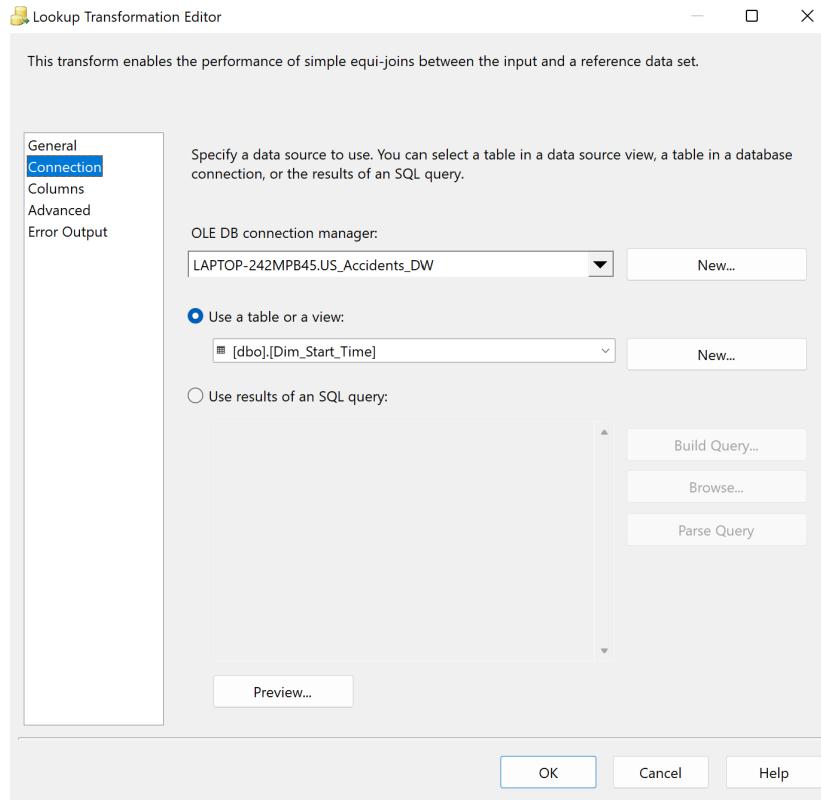
**Bước 4:** Tiếp theo, ta tạo ra một khối Lookup ánh xạ dữ liệu với bảng Dim\_Start\_Time và đặt tên là Lookup Dim\_Start\_Time.



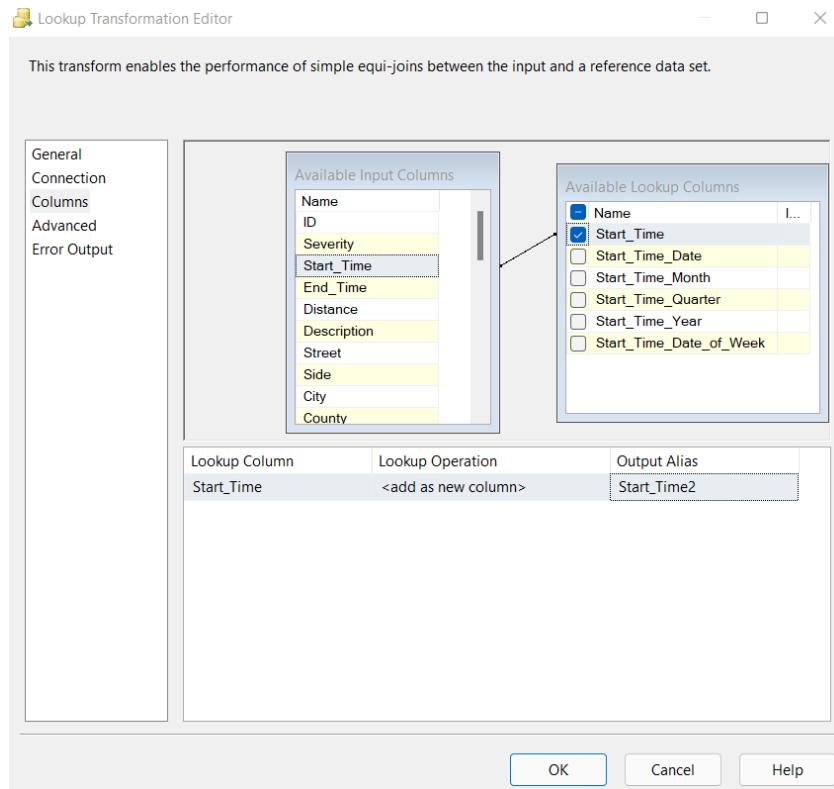
**Bước 5:** Double click vào Lookup tại mục “Specify how to handle rows with no matching entries” ta chọn “Redirect rows to no match output”.



**Bước 6:** Sau đó ta tạo connection cho Lookup này với connection manager là cơ sở dữ liệu US\_Accidents\_DW và ta chọn bảng Dim\_Start\_Time.

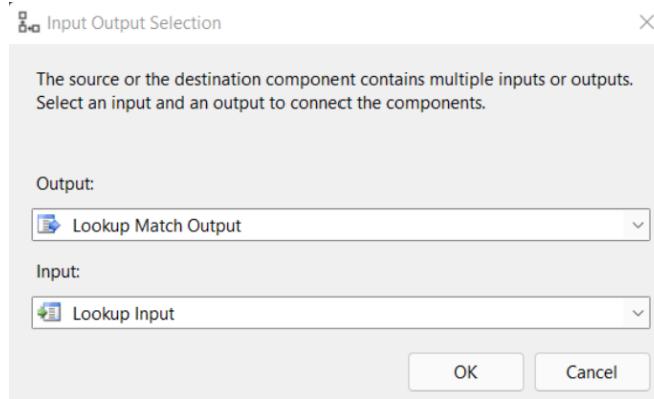


**Bước 7:** Tại mục columns, ta ánh xạ cột Start\_Time có trong OLE DB Source với cột dữ liệu Start\_Time trong bảng Dim\_Start\_Time. Tiếp theo, ta tick chọn cột Start\_Time để thêm cột này vào bảng Fact. Đồng thời, ta cũng đặt tên cho Output Alias là Start\_Time2 – là cột vừa được thêm vào, để phân biệt với cột Start\_Time đã tồn tại trước đó.



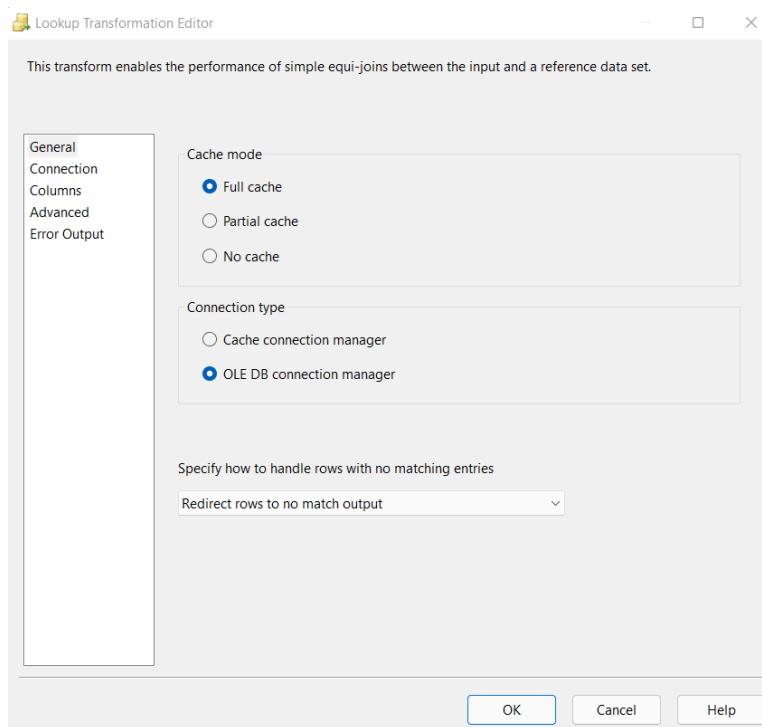
Sau đó, ta chọn OK để hoàn tất thiết lập.

**Bước 8:** Tiếp theo, ta tạo ra một khối Lookup ánh xạ dữ liệu với bảng Dim\_Location và đặt tên là Lookup Dim\_Location.

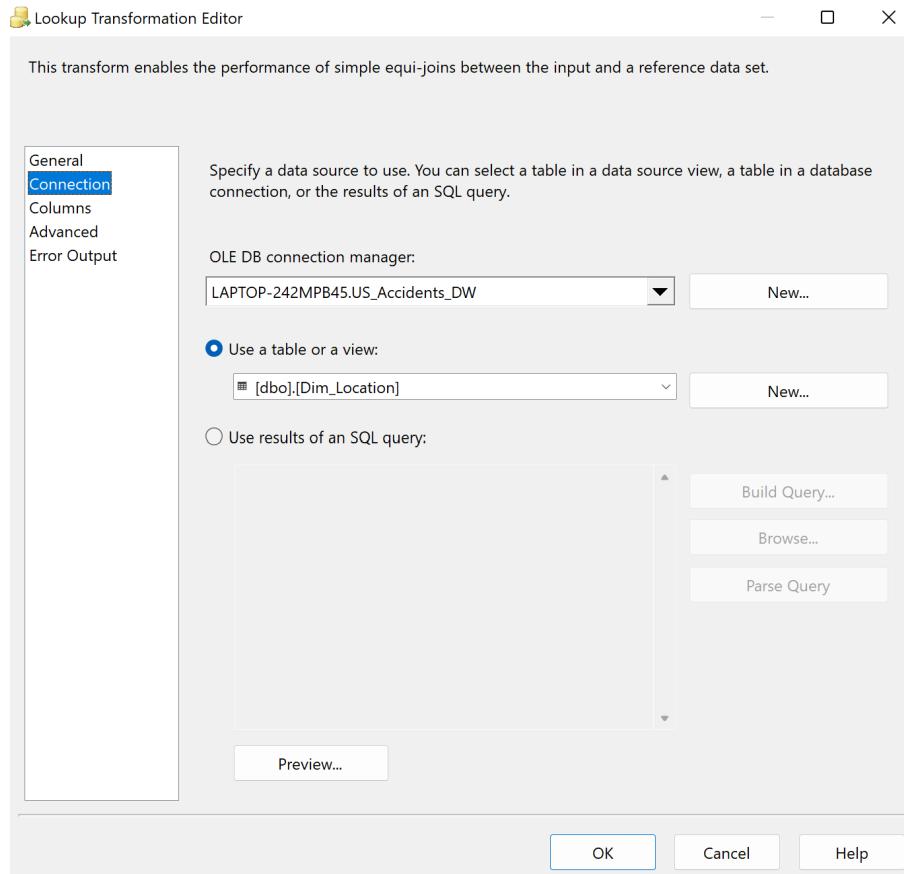


Tại hộp thoại Input Output Selection, ta chọn mục Output là “Lookup Match Output”, Input là “Lookup Input”. Sau đó chọn OK.

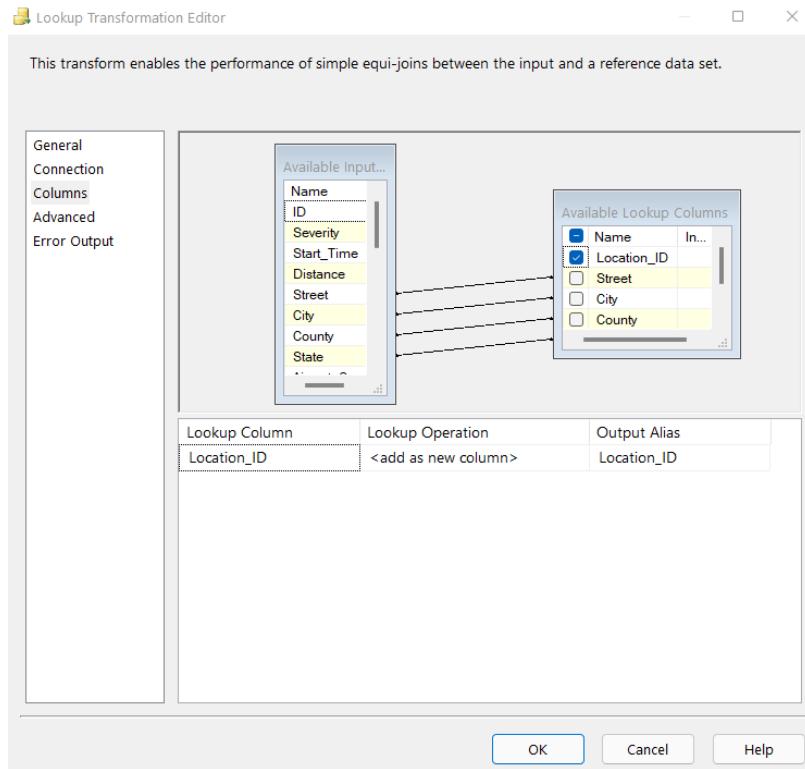
**Bước 9:** Double click vào Lookup tại mục “Specify how to handle rows with no matching entries” ta chọn “Redirect rows to no match output”.



**Bước 10:** Sau đó ta tạo connection cho Lookup này với connection manager là cơ sở dữ liệu US\_Accidents\_DW và ta chọn bảng Dim\_Location.

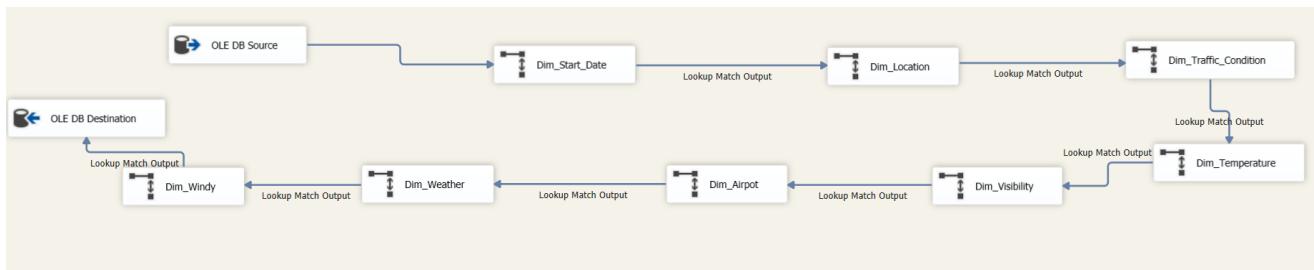


**Bước 11:** Tại mục columns, ta ánh xạ các cột dữ liệu có trong OLE DB Source với các cột dữ liệu trong bảng Dim\_Location. Tiếp theo, ta tick chọn cột Location\_ID để thêm cột này vào bảng Fact.



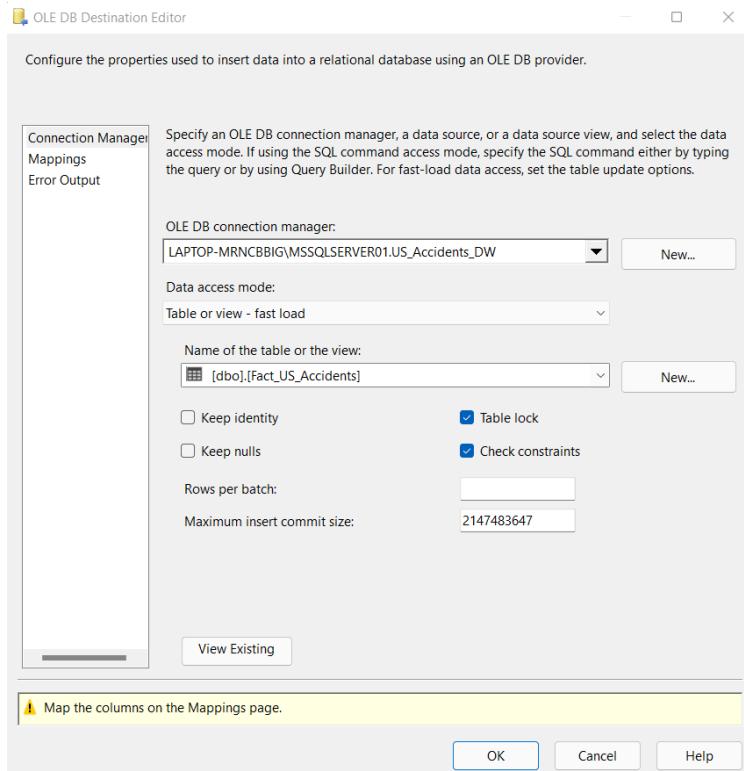
Sau đó, ta chọn OK để hoàn tất thiết lập.

**Bước 12:** Tiếp tục sử dụng Lookup cho các bảng Dim còn lại, ta được kết quả như sau:

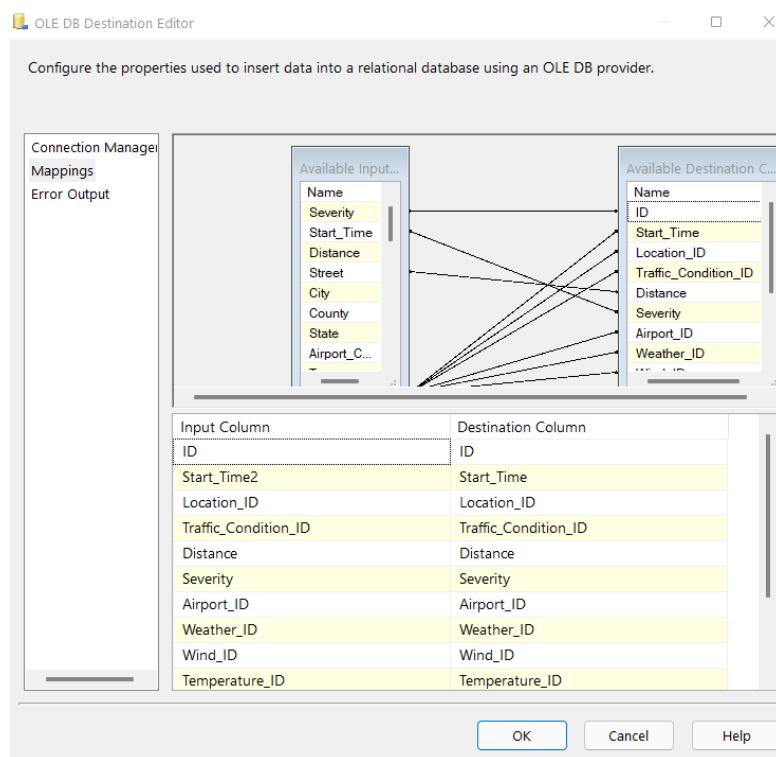


Sau đó, tạo mới một **OLE DB Destination** để đổ dữ liệu gốc sau khi đã được xử lý vào trong kho dữ liệu US\_Accidents\_DW.

**Bước 18:** Tiếp theo ta double click vào **OLE DB Destination** và chọn địa điểm đổ dữ liệu là bảng Fact\_US\_Accidents trong kho dữ liệu US\_Accidents\_DW.



Tiếp đến ta cần chọn mục **Mappings** để xem xét việc ánh xạ các cột dữ liệu



Cuối cùng, chọn OK để hoàn tất thiết lập.

Như vậy, chúng ta đã hoàn tất quá trình đỗ dữ liệu vào bảng Fact\_US\_Accidents.

Nội dung của các câu lệnh SQL như sau:

```
ALTER TABLE Fact_US_Accidents  
ADD CONSTRAINT FK_FACT_US_ACCIDENTS_START_TIME  
FOREIGN KEY(Start_Time) REFERENCES Dim_Start_Time(Start_Time)
```

```
ALTER TABLE Fact_US_Accidents  
ADD CONSTRAINT FK_FACT_US_ACCIDENTS_LOCATION_ID  
FOREIGN KEY(Location_ID) REFERENCES Dim_Location(Location_ID)
```

```
ALTER TABLE Fact_US_Accidents  
ADD CONSTRAINT FK_FACT_US_ACCIDENTS_TRAFFIC_CONDITION_ID  
FOREIGN KEY(Traffic_Condition_ID) REFERENCES  
Dim_Traffic_Condition(Traffic_Condition_ID)
```

```
ALTER TABLE Fact_US_Accidents  
ADD CONSTRAINT FK_FACT_US_ACCIDENTS_AIRPORT_ID  
FOREIGN KEY(Airport_ID) REFERENCES Dim_Airport(Airport_ID)
```

```
ALTER TABLE Fact_US_Accidents  
ADD CONSTRAINT FK_FACT_US_ACCIDENTS_WIND_ID  
FOREIGN KEY(Wind_ID) REFERENCES Dim_Wind(Wind_ID)
```

```
ALTER TABLE Fact_US_Accidents  
ADD CONSTRAINT FK_FACT_US_ACCIDENTS_TEMPERATURE_ID  
FOREIGN KEY(Temperature_ID) REFERENCES Dim_Temperature(Temperature_ID)
```

```
ALTER TABLE Fact_US_Accidents  
ADD CONSTRAINT FK_FACT_US_ACCIDENTS_VISIBILITY_ID  
FOREIGN KEY(Visibility_ID) REFERENCES Dim_Visibility(Visibility_ID)
```

```
ALTER TABLE Fact_US_Accidents  
ADD CONSTRAINT FK_FACT_US_ACCIDENTS_PRECIPITATION_ID  
FOREIGN KEY(Precipitation_ID) REFERENCES Dim_Precipitation(Precipitation_ID)
```

```
ALTER TABLE Fact_US_Accidents  
ADD CONSTRAINT FK_FACT_US_ACCIDENTS_WEATHER_ID  
FOREIGN KEY(Weather_ID) REFERENCES Dim_Weather(Weather_ID)
```

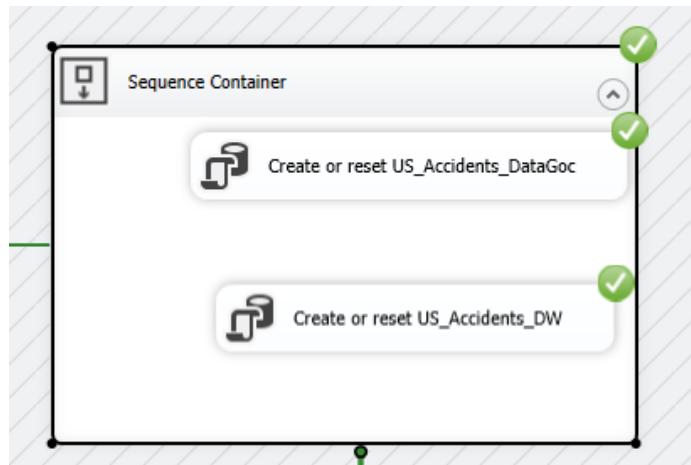
## 2.7. Thực thi project và kết quả SSIS

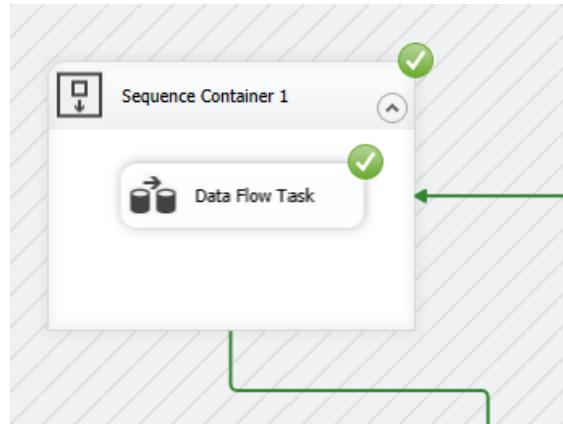
### 2.7.1. Kết quả chạy chương trình SSIS

Cuối cùng ta chọn nút Start trên thanh công cụ để chạy cả project.

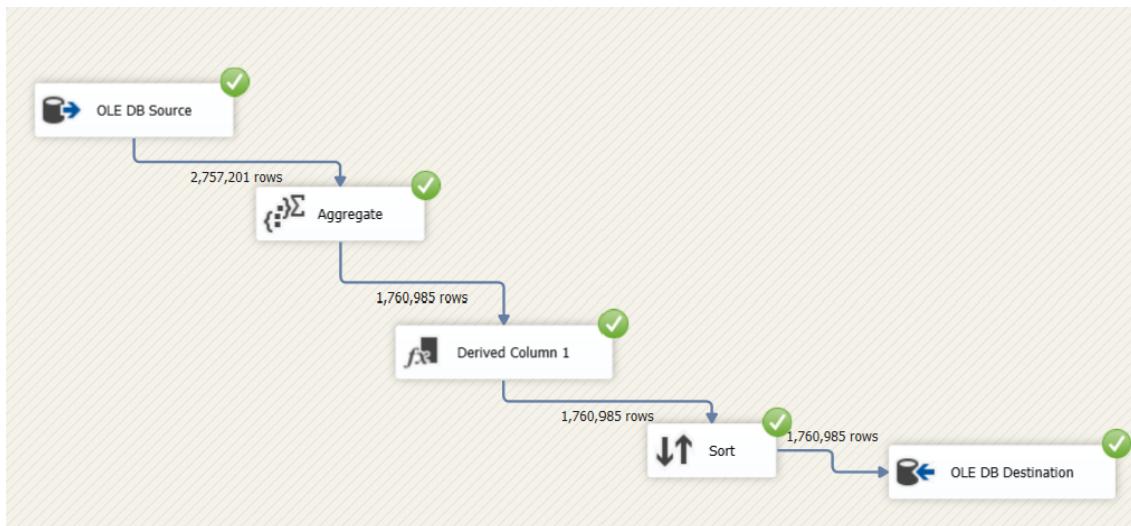


Kết quả Data cleaning:

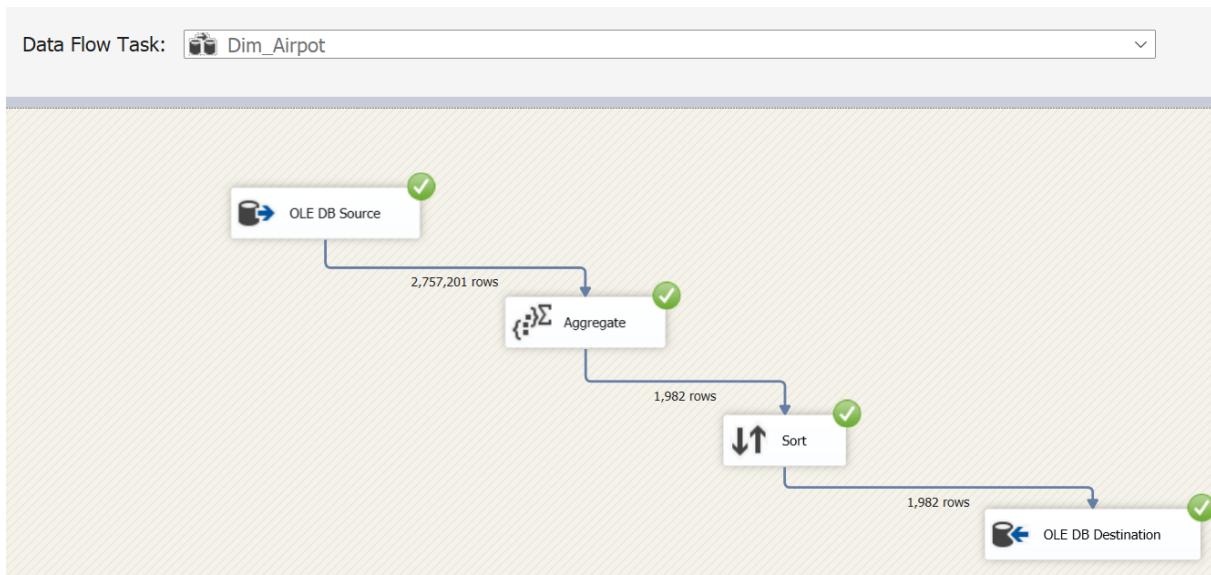




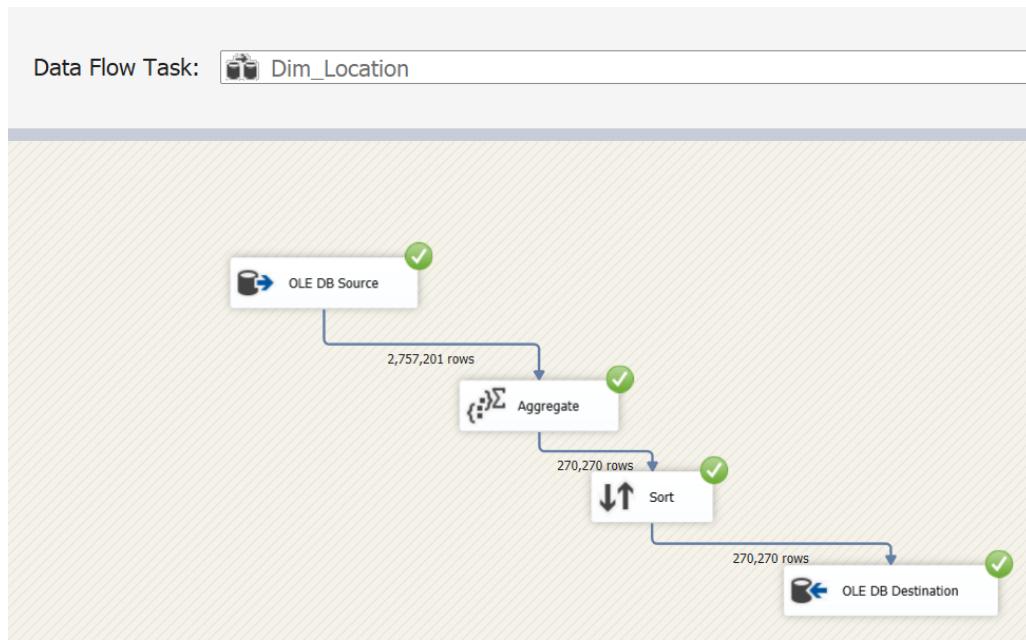
Kết quả bảng Dim\_Start\_Time:



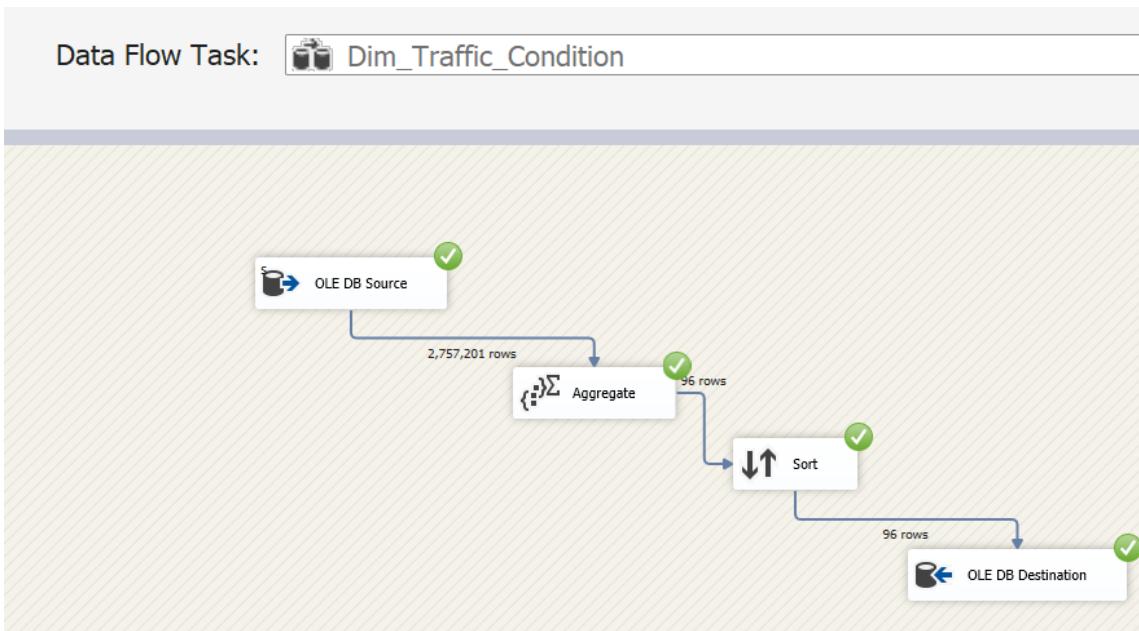
Kết quả bảng Dim\_Airport:



Kết quả bảng Dim\_Location:



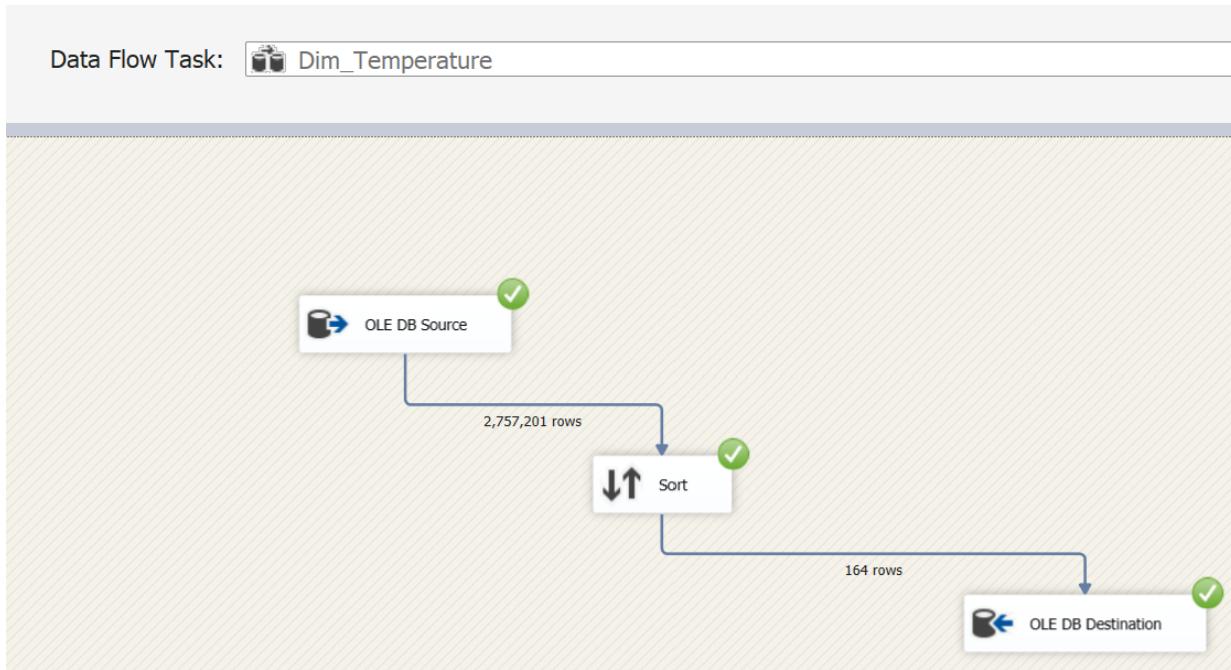
Kết quả bảng Dim\_Traffic\_Condition:



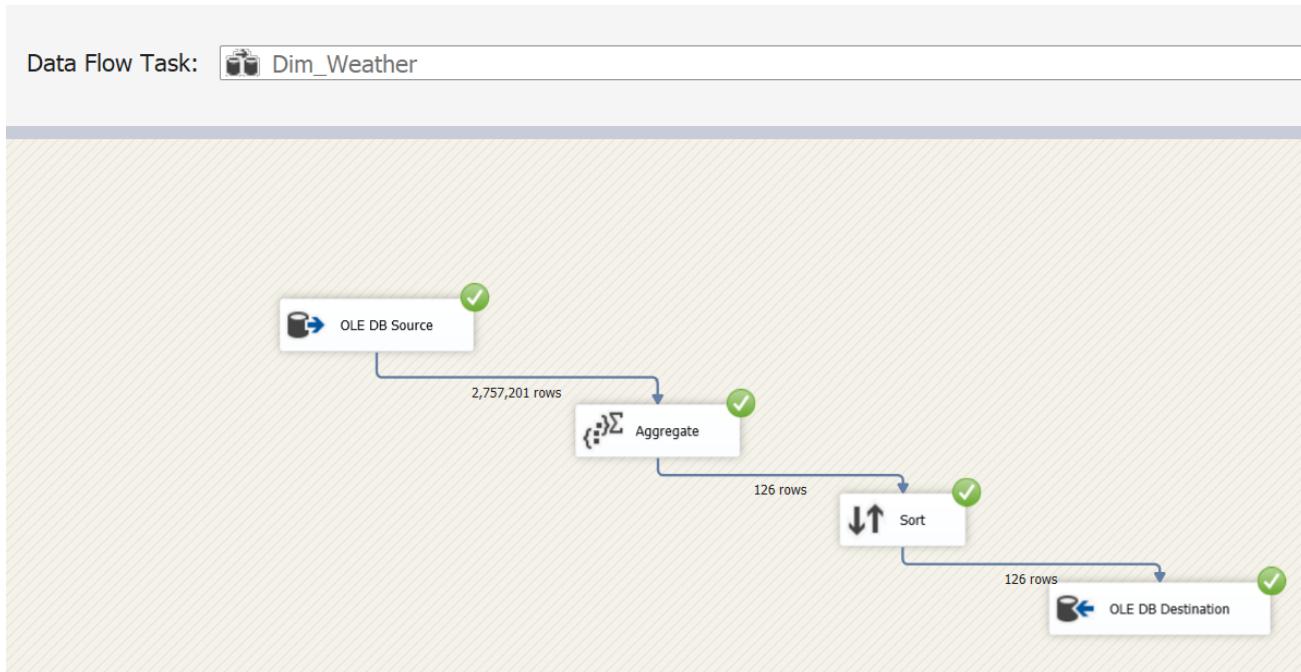
Kết quả bảng Dim\_Wind:



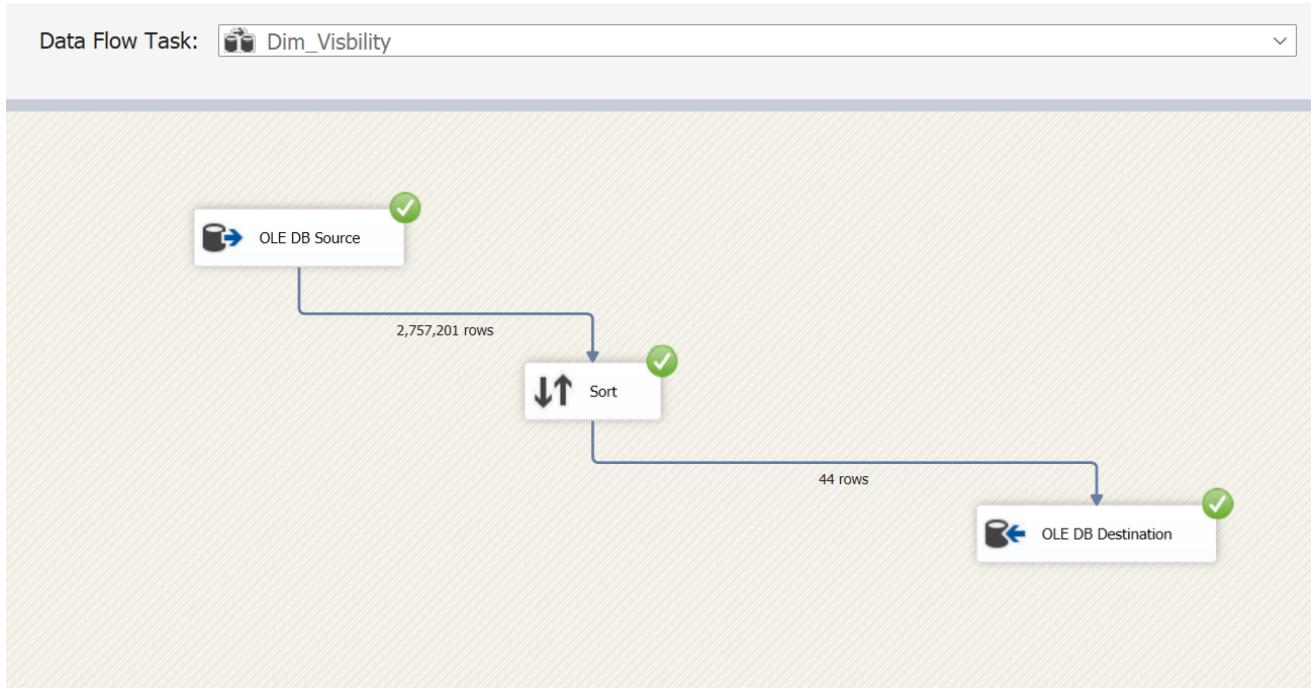
Kết quả bảng Dim\_Temperature:



Kết quả bảng Dim\_Weather:



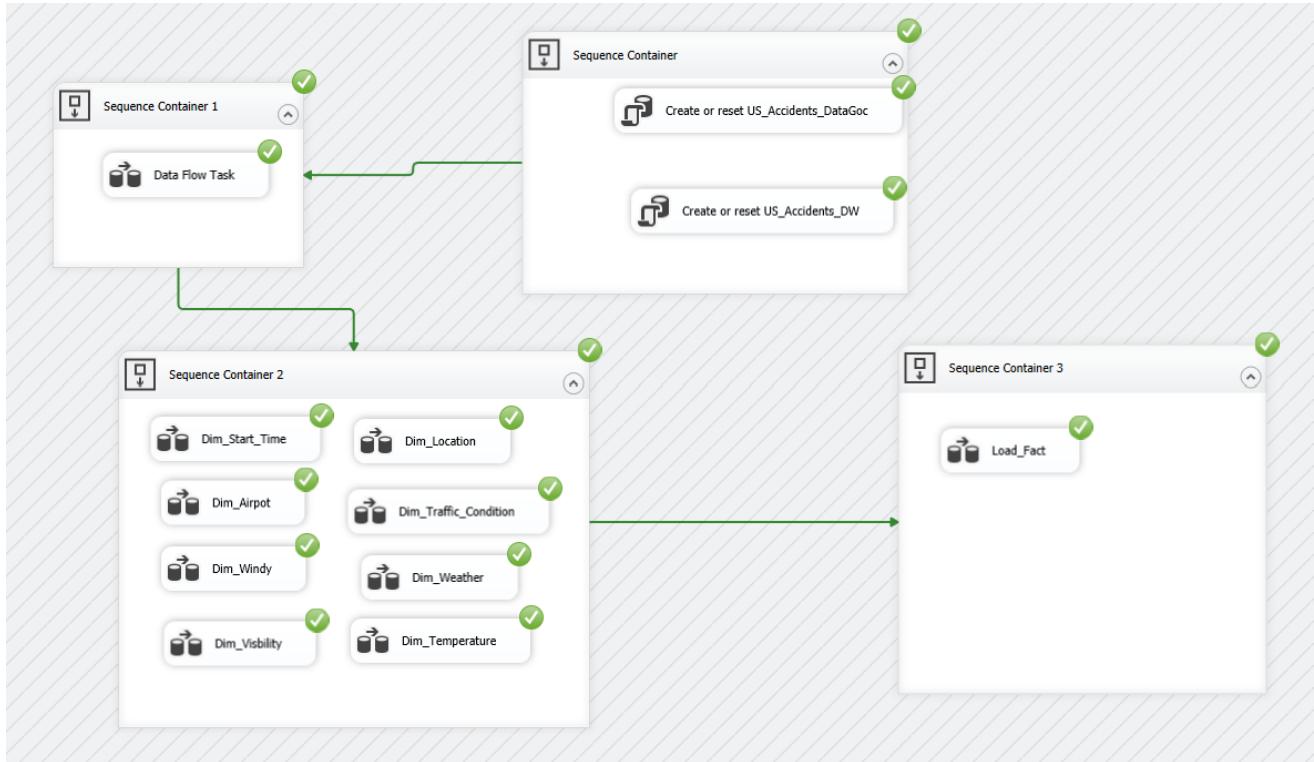
Kết quả bảng Dim\_Visibility



Kết quả bảng Fact\_US\_Accidents:



Kết quả cả quá trình SSIS:



## 2.7.2. Kiểm tra dữ liệu các bảng

- Bảng Dim\_Airpot

	Airport_ID	Airport_Code
1	1	K01M
2	2	K04V
3	3	K04W
4	4	K06D
5	5	K08D
6	6	K0A9
7	7	K0CO
8	8	K0E0
9	9	K0F2
10	10	K0J4
11	11	K0VG
12	12	K11R
13	13	K12N
14	14	K14Y
15	15	K1A5
16	16	K1A6
17	17	K1F0
18	18	K1H2
19	19	K1I1
20	20	K1J0
21	21	K1M4
22	22	K1P1
23	23	K1R7
24	24	K1V6
25	25	K1YT
26	26	K20U
27	27	K20V
28	28	K21D
29	29	K27A
30	30	K2C8
31	31	K2D5
32	32	K2DP
33	33	K2G4
34	34	K2I0
35	35	K2J9
36	36	K2V5
37	37	K2W6
..	..	..

- Bảng Dim\_Start\_Time

	Start_Time	Start_Time_Hour	Start_Time_Minute	Start_Time_Date	Start_Time_Month	Start_Time_Quarter	Start_Time_Year	Start_Time_Date_of_Week
1	2016-01-14 20:18:33.0000000	20	18	14	1	1	2016	5
2	2016-02-08 00:37:08.0000000	0	37	8	2	1	2016	2
3	2016-02-08 05:56:20.0000000	5	56	8	2	1	2016	2
4	2016-02-08 06:15:39.0000000	6	15	8	2	1	2016	2
5	2016-02-08 06:51:45.0000000	6	51	8	2	1	2016	2
6	2016-02-08 07:53:43.0000000	7	53	8	2	1	2016	2
7	2016-02-08 08:15:41.0000000	8	15	8	2	1	2016	2
8	2016-02-08 08:16:57.0000000	8	16	8	2	1	2016	2
9	2016-02-08 11:51:46.0000000	11	51	8	2	1	2016	2
10	2016-02-08 14:19:57.0000000	14	19	8	2	1	2016	2
11	2016-02-08 15:16:43.0000000	15	16	8	2	1	2016	2
12	2016-02-08 15:43:50.0000000	15	43	8	2	1	2016	2
13	2016-02-08 16:50:57.0000000	16	50	8	2	1	2016	2
14	2016-02-08 17:27:39.0000000	17	27	8	2	1	2016	2
15	2016-02-08 17:30:18.0000000	17	30	8	2	1	2016	2
16	2016-02-08 18:11:11.0000000	18	11	8	2	1	2016	2
17	2016-02-08 19:47:42.0000000	19	47	8	2	1	2016	2
18	2016-02-08 20:13:22.0000000	20	13	8	2	1	2016	2
19	2016-02-08 21:00:17.0000000	21	0	8	2	1	2016	2
20	2016-02-08 21:10:10.0000000	21	10	8	2	1	2016	2
21	2016-02-08 21:30:31.0000000	21	30	8	2	1	2016	2
22	2016-02-09 05:54:01.0000000	5	54	9	2	1	2016	3
23	2016-02-09 06:10:59.0000000	6	10	9	2	1	2016	3
24	2016-02-09 06:46:32.0000000	6	46	9	2	1	2016	3
25	2016-02-09 07:18:10.0000000	7	18	9	2	1	2016	3
26	2016-02-09 07:25:40.0000000	7	25	9	2	1	2016	3

- Bảng Dim\_Temperature

Results Messages

	Temperature_ID	Temperatu
1	1	-89
2	2	-58
3	3	-50
4	4	-40
5	5	-33
6	6	-30
7	7	-29
8	8	-28
9	9	-27
10	10	-26
11	11	-25
12	12	-24
13	13	-23
14	14	-22
15	15	-21
16	16	-20
17	17	-19
18	18	-18
19	19	-17
20	20	-16
21	21	-15
22	22	-14
23	23	-13
24	24	-12
25	25	-11
26	26	-10
27	27	-9
28	28	-8
29	29	-7

- Bảng Dim\_Traffic\_Condition

	Traffic_Condition_ID	Bump	Crossing	Junction	Railway	Roundabout	Station	Stop	Traffic_Calming	Traffic_Signal
1	1	False	False	False	False	False	True	False	True	True
2	2	False	False	False	False	False	True	True	True	False
3	3	False	False	True	False	True	False	True	True	False
4	4	False	True	False	False	False	True	True	True	True
5	5	True	False	False	False	False	False	True	True	True
6	6	True	False	False	False	False	True	False	True	True
7	7	True	True	False	False	False	False	True	True	True
8	8	False	False	True	False	False	False	True	False	True
9	9	True	False	False	False	False	True	True	True	False
10	10	True	True	False	True	False	True	False	True	False
11	11	False	False	True	False	True	True	False	False	False
12	12	False	False	True	True	False	True	False	False	True
13	13	False	True	False	False	False	True	False	True	False
14	14	True	False	False	True	False	False	False	True	True
15	15	False	True	False	False	False	False	True	True	True
16	16	False	True	True	False	False	False	False	True	False
17	17	False	False	False	True	False	True	True	False	True
18	18	False	True	True	False	False	True	True	False	False
19	19	False	True	True	False	True	False	False	False	False
20	20	False	True	True	True	False	True	False	False	False
21	21	True	True	False	False	False	True	True	True	False
22	22	False	True	False	True	False	False	False	True	True
23	23	False	True	True	False	False	False	True	False	True
24	24	False	False	False	False	True	False	False	True	False
25	25	True	False	False	False	False	False	False	True	True
26	26	True	True	False	False	False	True	False	True	True
27	27	False	True	False	True	False	True	True	False	True
28	28	True	True	False	False	False	True	False	True	False
29	29	False	False	True	False	False	False	True	False	False

- Bảng Dim\_Location

	Location_ID	Street	City	County	State
1	1		Williamston	Anderson	SC
2	2	1 1/2 Ave	Hanford	Kings	CA
3	3	1 Mile Rd	Hesperia	Newaygo	MI
4	4	1/2 Ave	Corcoran	Kings	CA
5	5	1/2 Ave	Hanford	Kings	CA
6	6	1/2 Ave	Lemoore	Kings	CA
7	7	1/2 Ave SE	Jamestown	Stutsman	ND
8	8	1/2 Mile Rd	Caledonia	Racine	WI
9	9	1/2 Rd	De Beque	Mesa	CO
10	10	1/4 Ave	Hanford	Kings	CA
11	11	10 Mile Rd	Gualala	Mendocino	CA
12	12	10 Mile Rd	Marysville	Piute	UT
13	13	10 Mile Rd	Point Arena	Mendocino	CA
14	14	10 Mile Rd NE	Rockford	Kent	MI
15	15	10000 E	Fort Duchesne	Uintah	UT
16	16	100th Ave	Milaca	Mille Lacs	MN
17	17	100th Ave	Princeton	Mille Lacs	MN
18	18	100th Ave	Randall	Morrison	MN
19	19	100th Ave N	Saint Petersburg	Pinellas	FL
20	20	100th Ave SE	Chatfield	Olmsted	MN
21	21	100th Ave SE	Clara City	Chippewa	MN
22	22	100th Ave SE	Eyota	Olmsted	MN
23	23	100th Pl	Live Oak	Suwannee	FL
24	24	100th St	Brownston	McLeod	MN
25	25	100th St	Elbow Lake	Grant	MN
26	26	100th St	Fayette	Fayette	IA
27	27	100th St	Oelwein	Fayette	IA
28	28	100th St E	Lancaster	Los Ange...	CA
29	29	100th St E	Littlerock	Los Ange...	CA

- Bảng Dim\_Visibility

	Visibility_ID	Visibility
1	1	0.000
2	2	1.000
3	3	2.000
4	4	3.000
5	5	4.000
6	6	5.000
7	7	6.000
8	8	7.000
9	9	8.000
10	10	9.000
11	11	10.000
12	12	11.000
13	13	12.000
14	14	13.000
15	15	14.000
16	16	15.000
17	17	16.000
18	18	19.000
19	19	20.000
20	20	22.000
21	21	23.000
22	22	25.000
23	23	30.000
24	24	34.000
25	25	35.000
26	26	36.000

- Bảng Fact

	ID	Start_Time	Location_ID	Traffic_Condition_ID	Airport_ID	Weather_ID	Wind_ID	Temperature_ID	Visibility_ID	Severity	Distance
1	A-1	2016-02-08 00:37:08.0000000	261223	96	1441	56	867	78	11	3	3.000
2	A-10	2016-02-08 15:16:43.0000000	261244	96	415	102	1034	68	1	2	0.000
3	A-1000	2016-03-23 07:04:16.0000000	237834	96	1145	7	1	75	11	2	0.000
4	A-10000	2016-05-26 10:23:53.0000000	219223	93	455	79	852	99	11	2	0.000
5	A-100000	2016-05-20 16:18:53.0000000	262616	95	1121	7	1	108	11	2	0.000
6	A-100001	2016-05-20 16:33:20.0000000	243237	96	302	94	147	104	11	2	0.000
7	A-100002	2016-05-20 16:44:03.0000000	229793	94	544	7	201	110	11	4	0.000
8	A-100003	2016-05-20 16:44:03.0000000	251424	95	907	81	1136	109	11	2	1.000
9	A-100004	2016-05-20 16:44:03.0000000	244215	96	246	7	676	109	11	2	2.000
10	A-100005	2016-05-20 16:46:48.0000000	212273	96	1367	7	958	111	11	2	0.000
11	A-100006	2016-05-20 16:49:31.0000000	212396	80	1092	76	670	112	11	2	0.000
12	A-100007	2016-05-20 16:53:40.0000000	228436	94	1045	7	1218	109	11	2	0.000
13	A-100008	2016-05-20 17:05:15.0000000	209891	96	544	7	201	110	11	2	0.000
14	A-100009	2016-05-20 16:57:51.0000000	4299	77	1045	7	1218	109	11	2	0.000
15	A-10001	2016-05-26 10:22:29.0000000	240738	96	1686	79	960	102	11	2	0.000
16	A-100011	2016-05-20 16:59:11.0000000	282304	96	1557	76	958	109	11	2	0.000
17	A-100012	2016-05-20 16:56:28.0000000	242335	96	144	94	959	110	11	2	0.000
18	A-100013	2016-05-20 16:56:28.0000000	242335	96	144	94	959	110	11	2	0.000
19	A-100014	2016-05-20 17:10:18.0000000	190073	94	907	81	1136	109	11	4	0.000
20	A-100015	2016-05-20 17:08:59.0000000	279883	96	983	7	775	110	11	4	2.000
21	A-100016	2016-05-20 17:10:18.0000000	264958	96	144	94	959	110	11	2	0.000
22	A-100017	2016-05-20 17:25:48.0000000	224125	96	144	76	645	109	11	2	0.000
23	A-100018	2016-05-20 17:27:11.0000000	258147	96	1227	7	707	105	11	2	0.000
24	A-100019	2016-05-20 17:31:16.0000000	237011	96	249	7	1130	109	11	3	1.000
25	A-10002	2016-05-26 10:26:17.0000000	237107	96	448	81	958	106	11	2	0.000
26	A-100020	2016-05-20 17:35:22.0000000	244193	96	1852	7	1	110	11	3	0.000
27	A-100021	2016-05-20 17:38:00.0000000	244091	96	1557	76	791	106	11	2	0.000

## CHƯƠNG 3: QUÁ TRÌNH PHÂN TÍCH KHO DỮ LIỆU – SSAS VÀ NGÔN NGỮ TRUY VẤN MDX

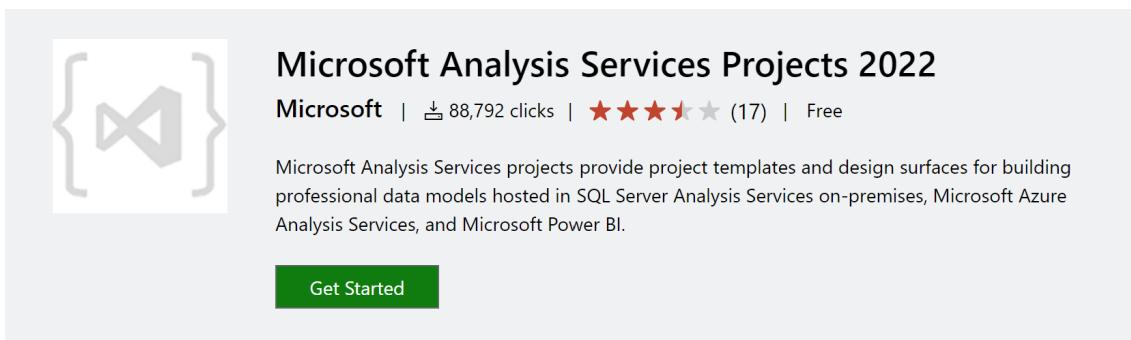
### 3.1 Định nghĩa Data Source View với Analysis Service Project

#### 3.1.1 Tải ứng dụng Microsoft Analysis Services Project 2022

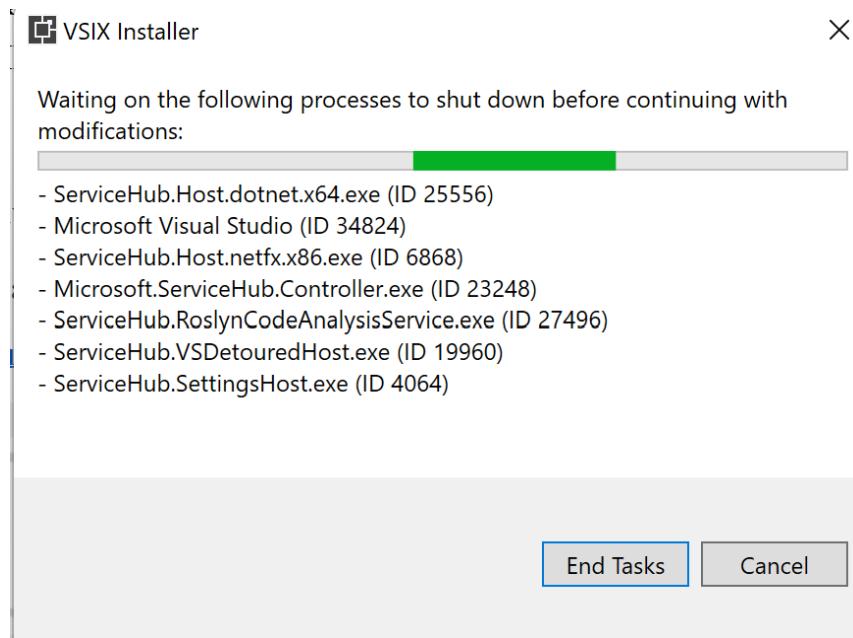
Bước 1: Truy cập vào đường link để tải ứng dụng

Link:

<https://marketplace.visualstudio.com/items?itemName=ProBITools.MicrosoftAnalysisServicesModelingProjects2022>

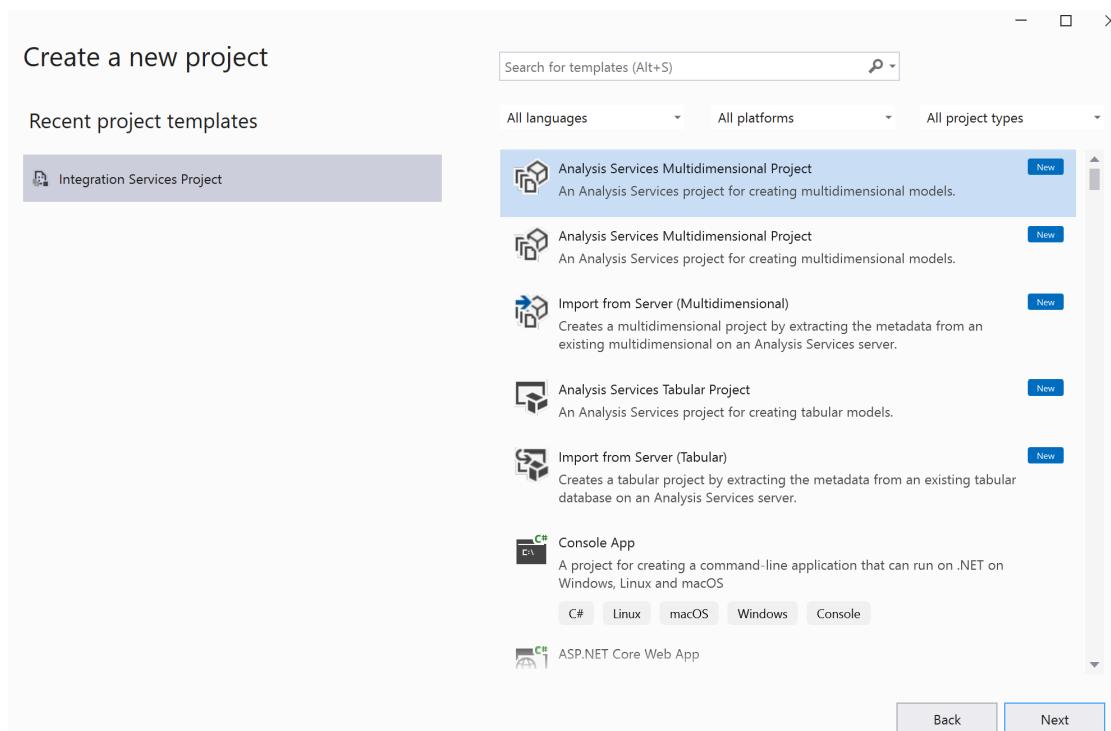


Bước 2: Sau khi tải xong và tiến hành cài đặt

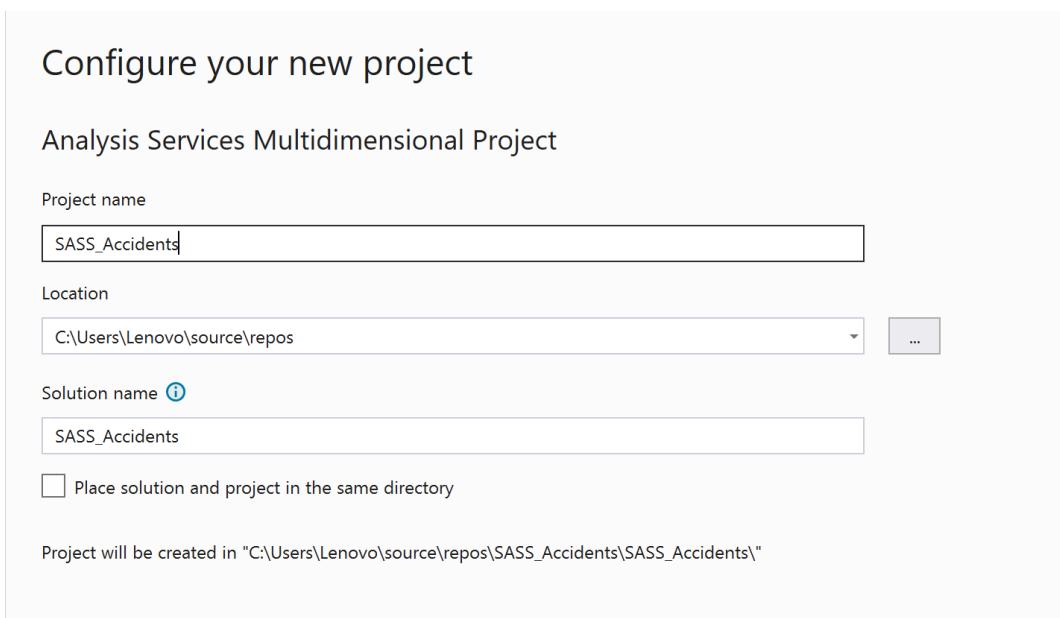


### 3.1.2 Tạo Analysis Service Project

Bước 1: Mở Visual Studio, chọn Create New Project, chọn Analysis Service Multidimensional Mining Project

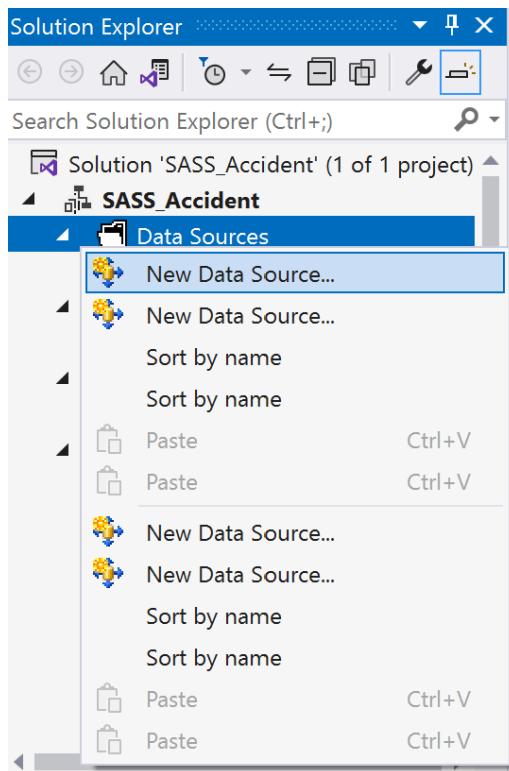


Bước 2: Tạo project mới, đặt tên project là **SSAS\_Accident**, ví trí lưu và nhấn OK

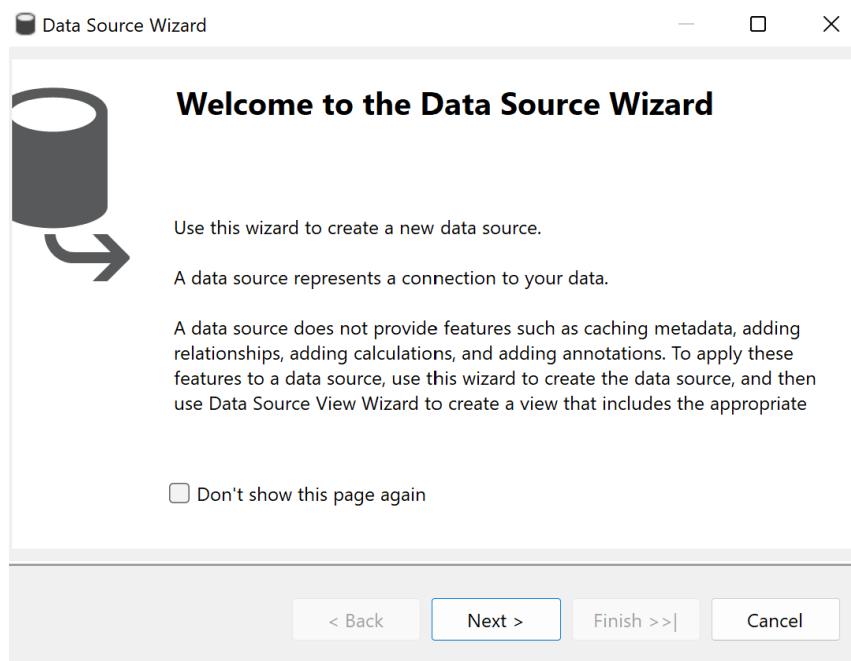


### 3.1.3 Định nghĩa Data Source

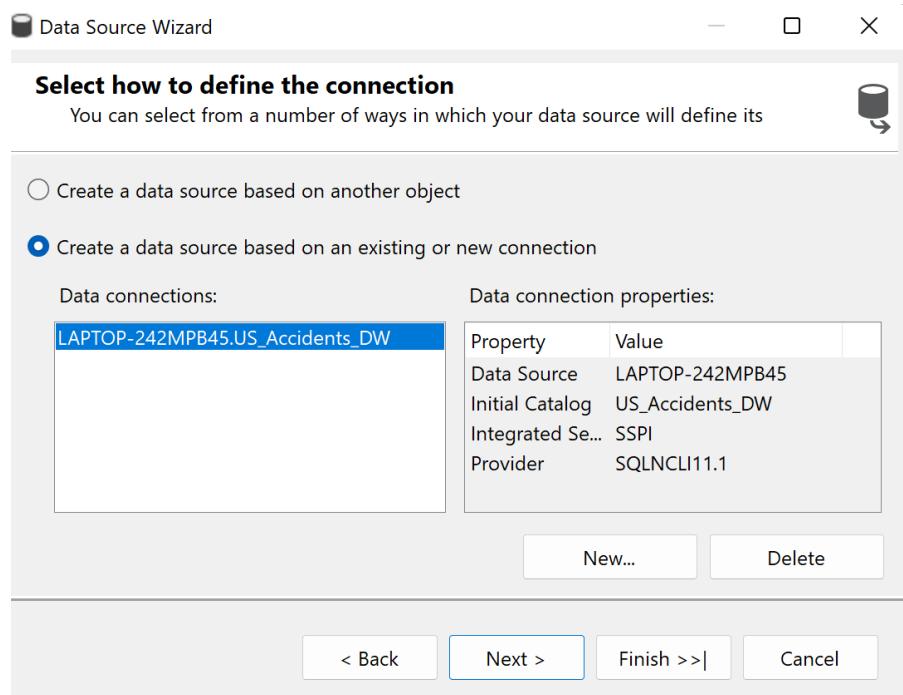
Bước 1: Tại Solution Explorer, nhập chuột phải vào Data Sources, chọn New Data Source



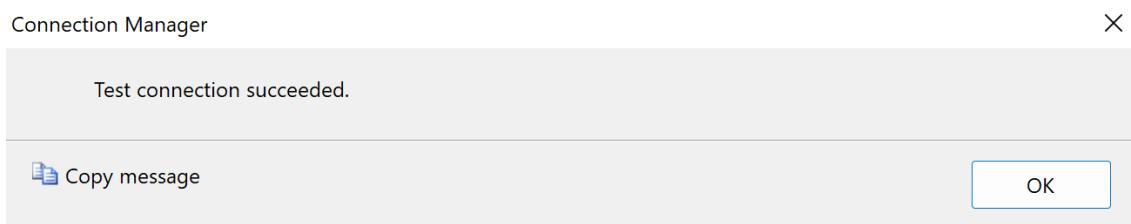
Bước 2: Cửa sổ Data Source Wizard hiện lên, nhấn Next



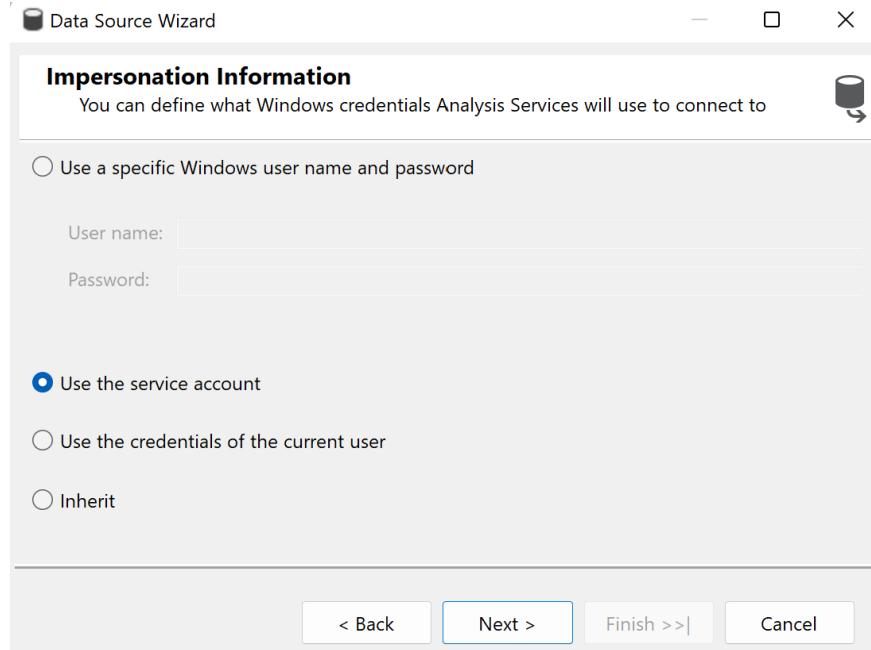
**Bước 3:** Tại trang **Select how to define the connection**, chọn **Create a data source based on an existing or new connection**, để tạo data source từ kết nối có sẵn. Nếu chưa hiện kết nối, bấm **New..** còn nếu đã hiện sẵn thì chọn **Connection**, nhấn **Next**



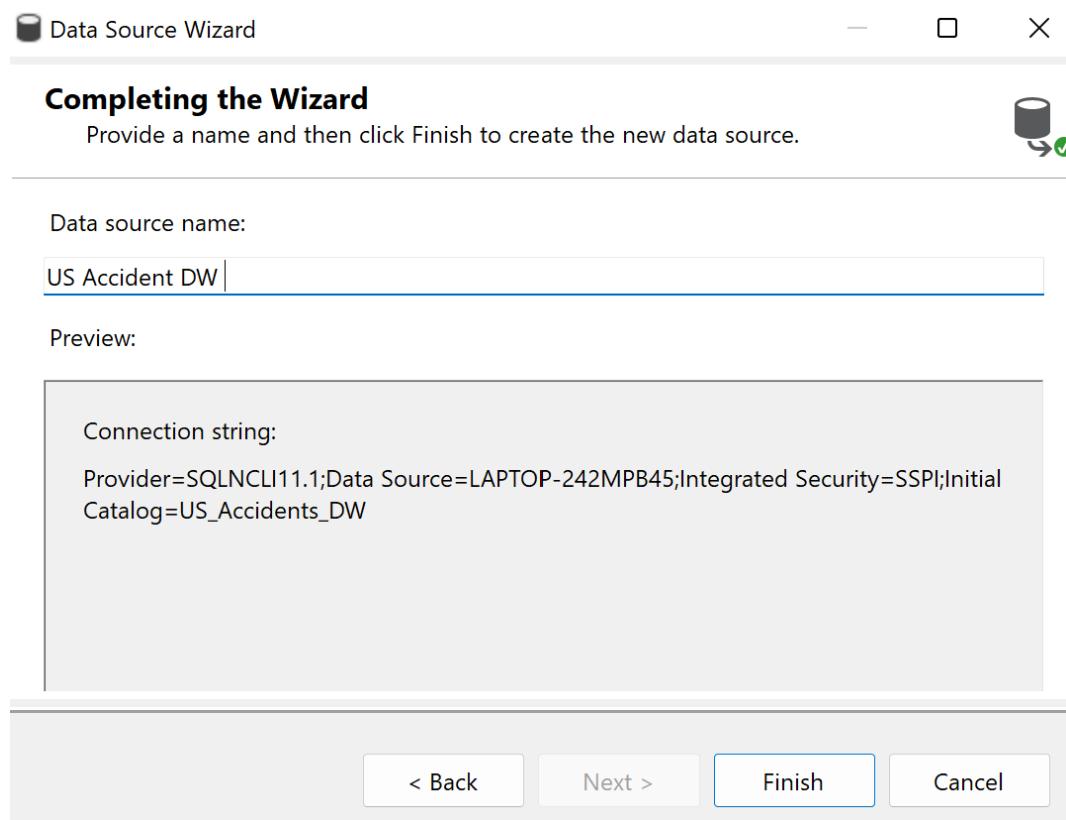
Test connection, nếu đã thành công thì hiện thông báo:



**Bước 4:** Tại trang Impersonation Information, chọn **Use the service account**. Nhấn **Next**

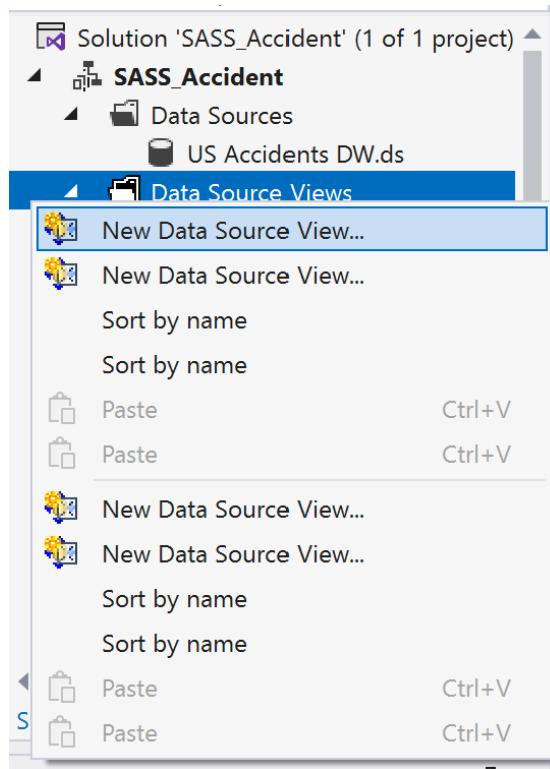


### Bước 5: Nhấn Finish để hoàn tất

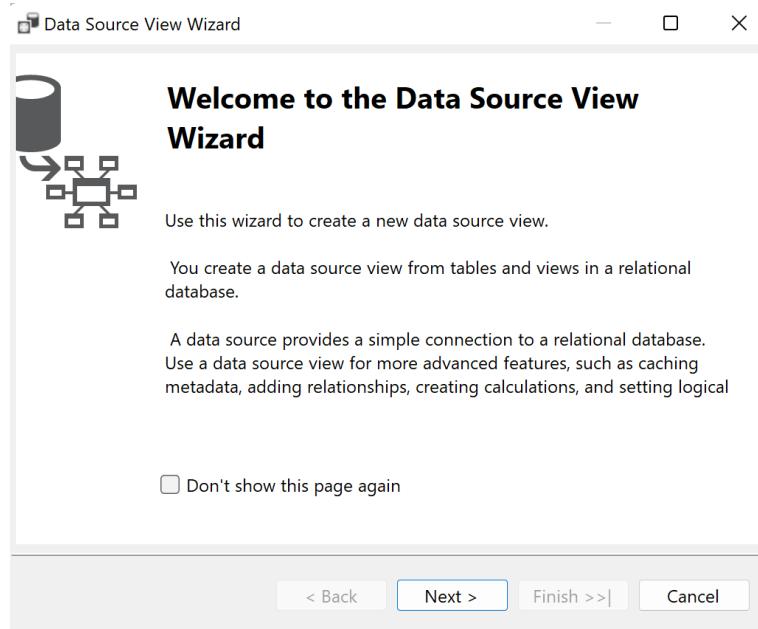


### 3.1.4 Định nghĩa Data Source View

**Bước 1:** Tại Solution Explorer, nhập chuột phải vào Data Source View, chọn New Data Source View

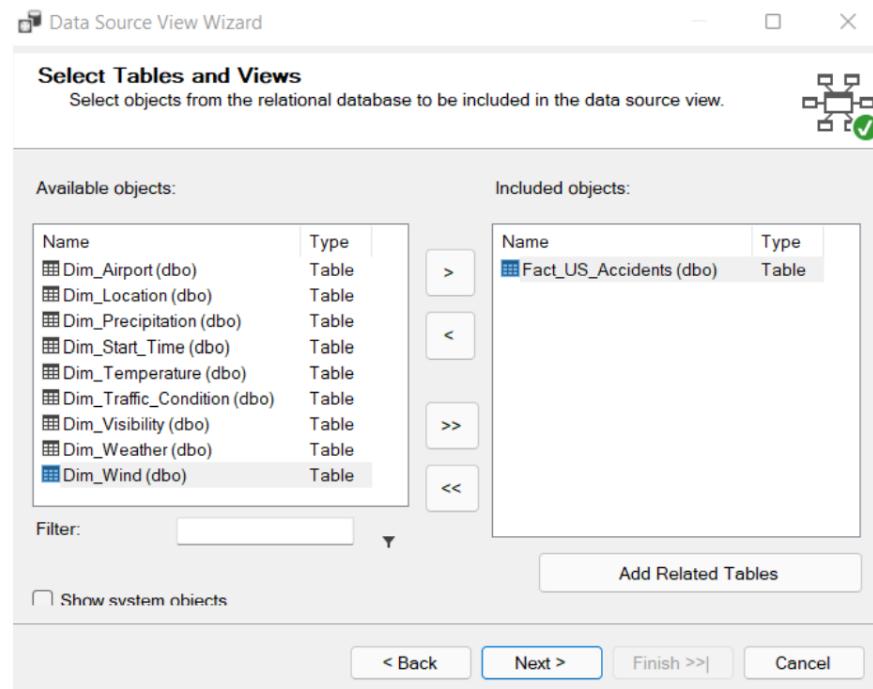


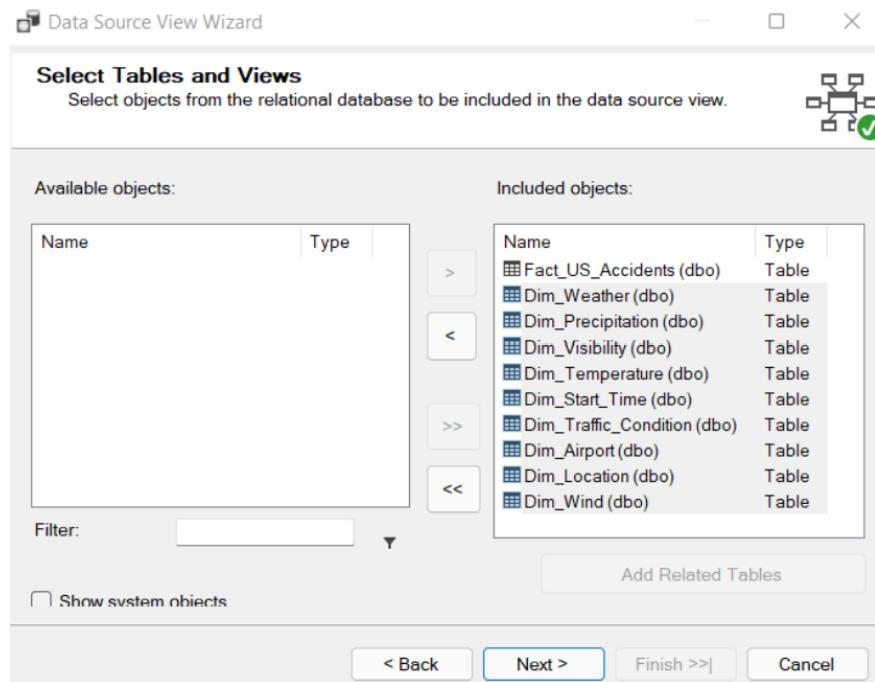
**Bước 2:** Cửa sổ Data Source View Wizard hiện lên, nhấn Next



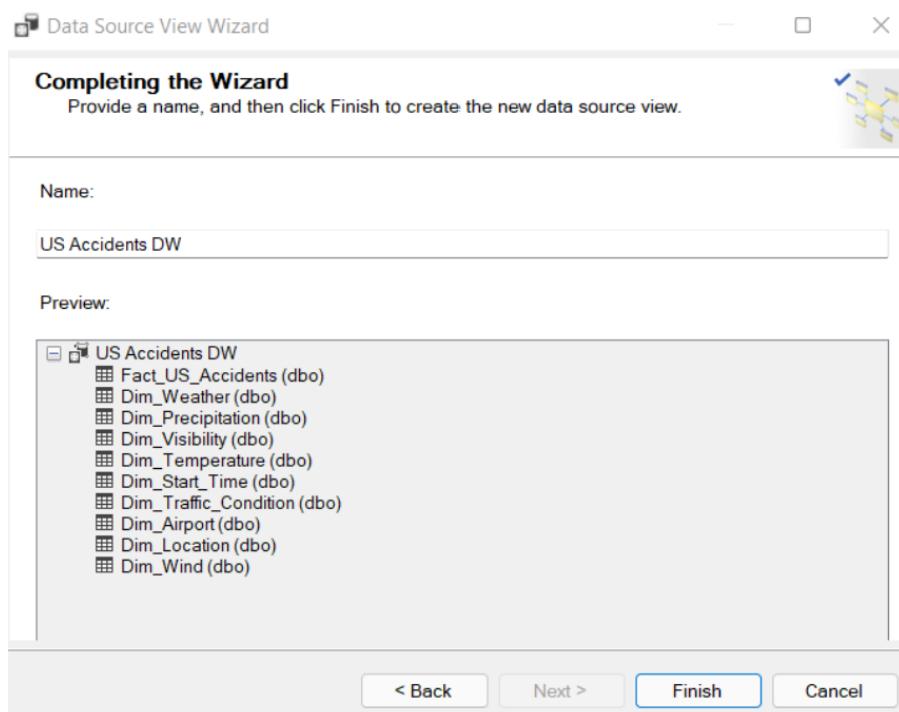
**Bước 3:** Tại trang Select a Data Source, chọn connection, nhấn Next

**Bước 4:** Tại trang Select Tables and Views, nhấn >> để chọn các bảng cần thiết ở Available objects, kết quả là các bảng nằm ở Included objects. Nhấn Next

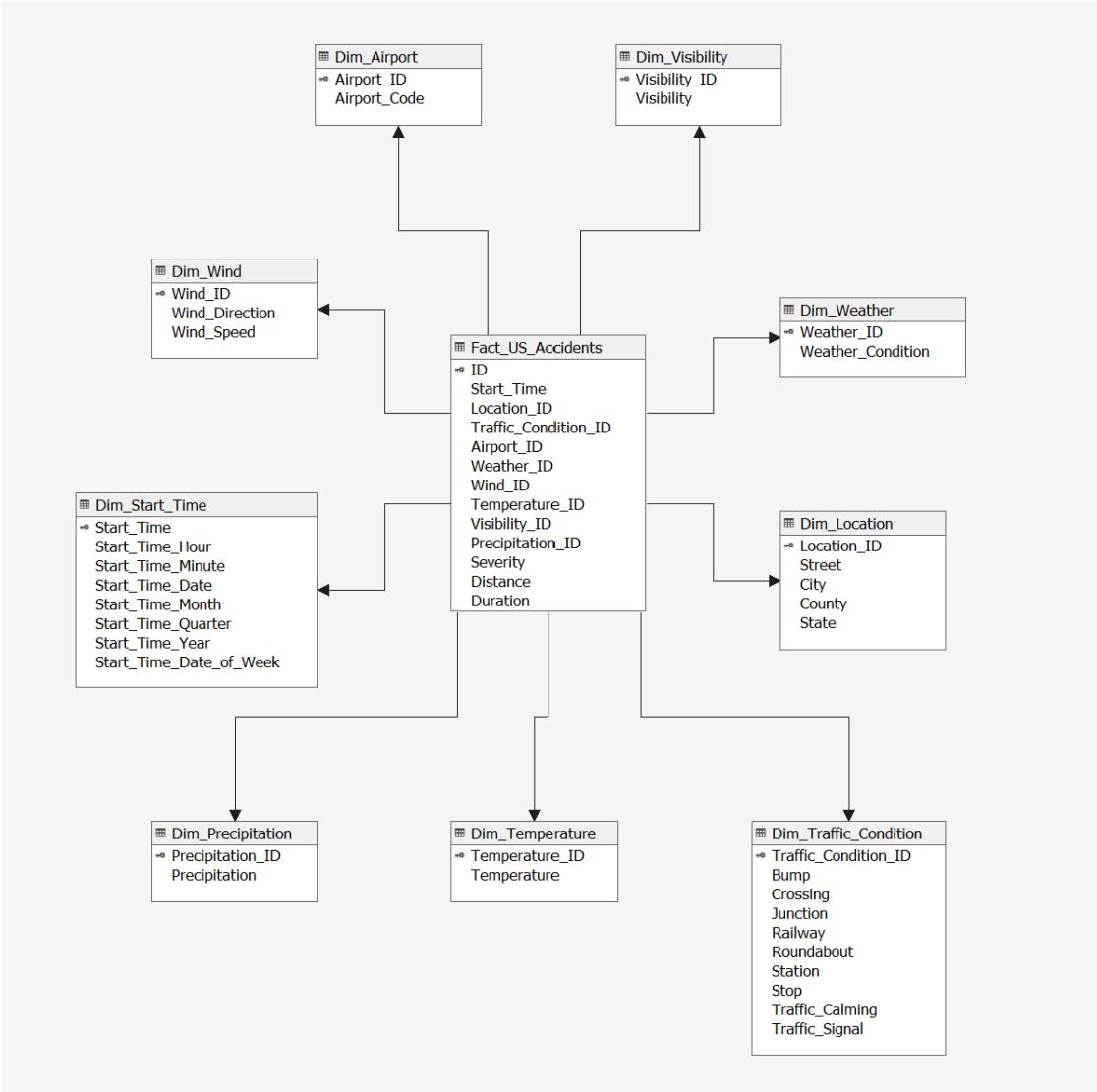




## Bước 5: Nhấn Finish để hoàn tất



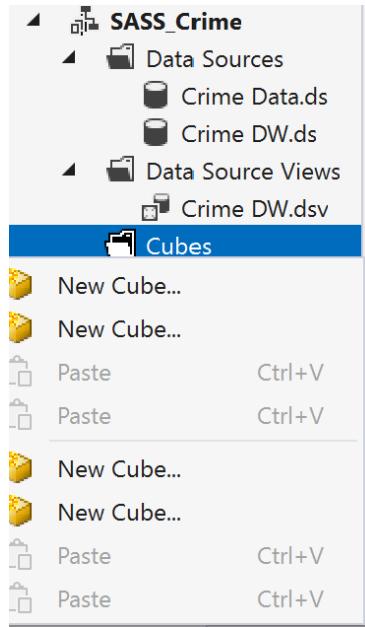
## Kết quả sau khi kết nối



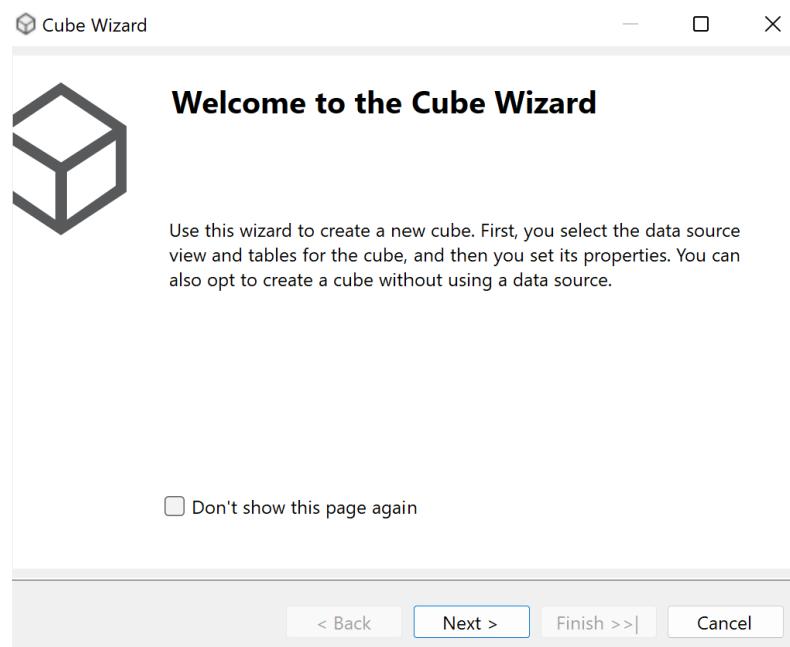
## 3.2 Định nghĩa và triển khai Cube

### 3.2.1 Định nghĩa Cube

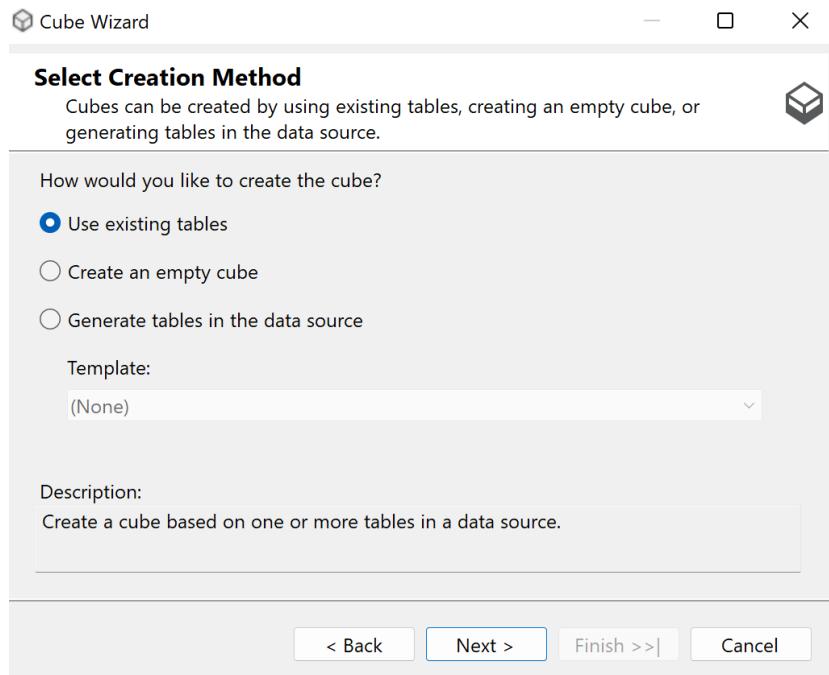
**Bước 1:** Tại Solution Explorer, nhấp chuột phải vào Cubes, chọn New Cube



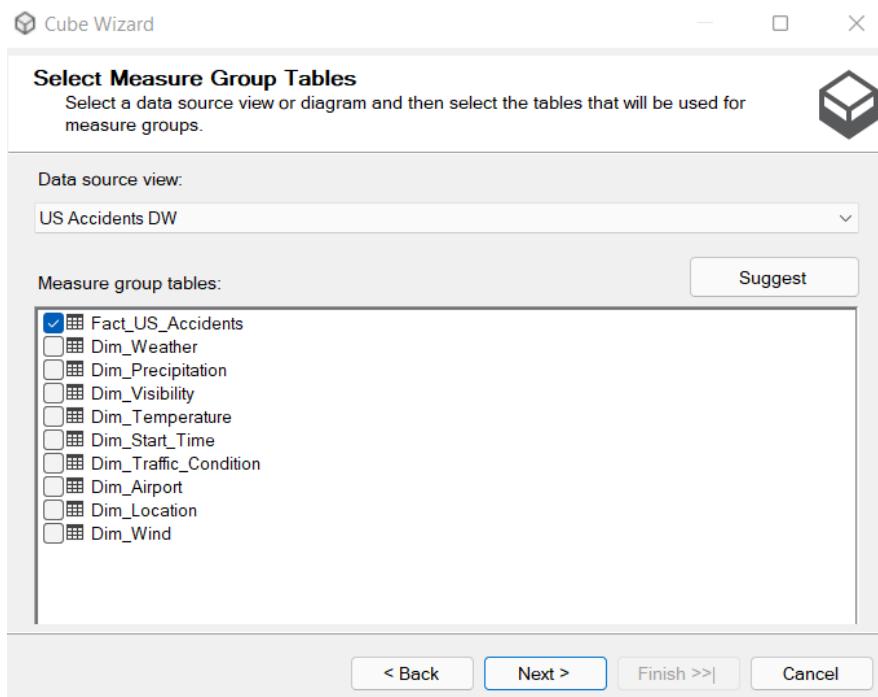
**Bước 2:** Cửa sổ Cube Wizard hiện lên, nhấn Next



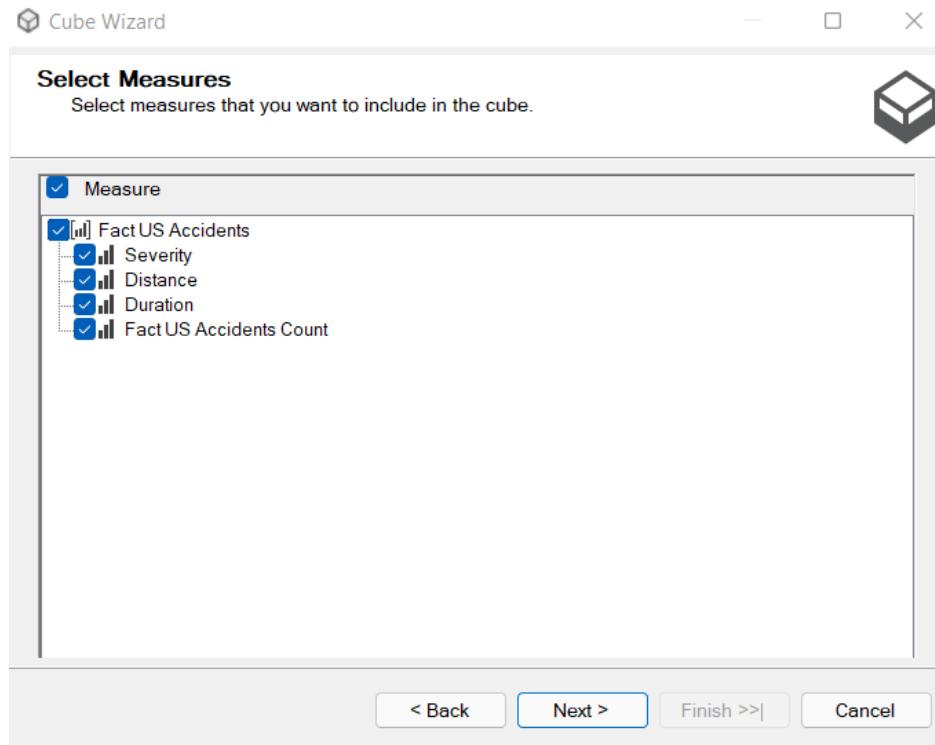
**Bước 3:** Tại trang Select Creation Method, chọn Use existing tables, nhấn Next



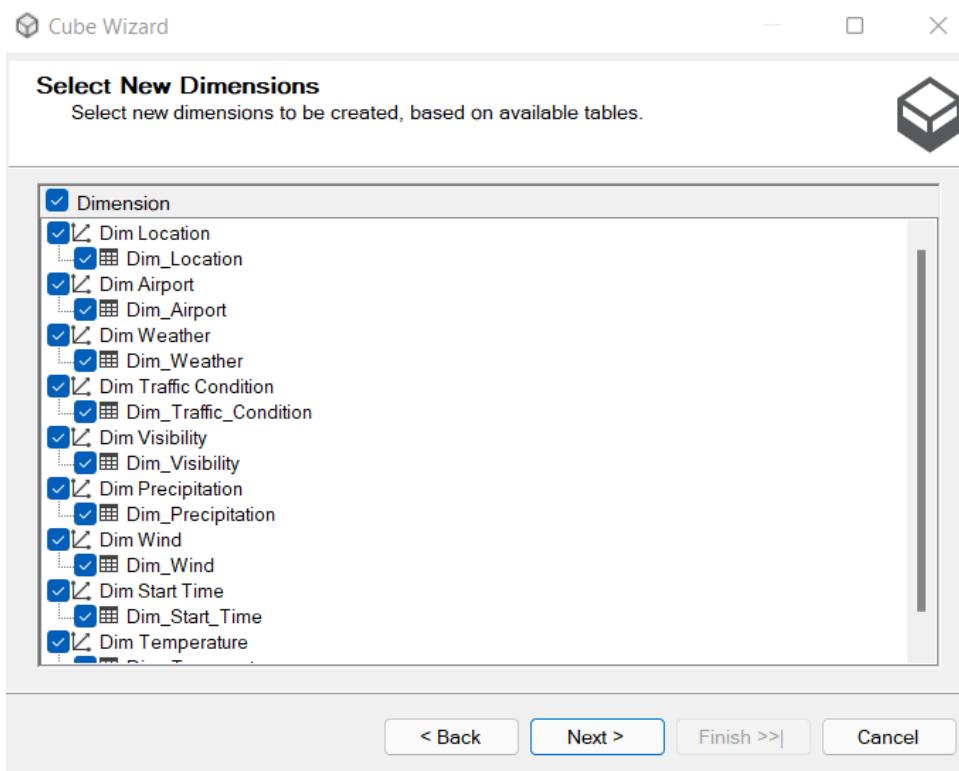
**Bước 4:** Tại trang **Select Measure Group Tables**, chọn **Data source view** và **Measure group tables**. Nhấn Next



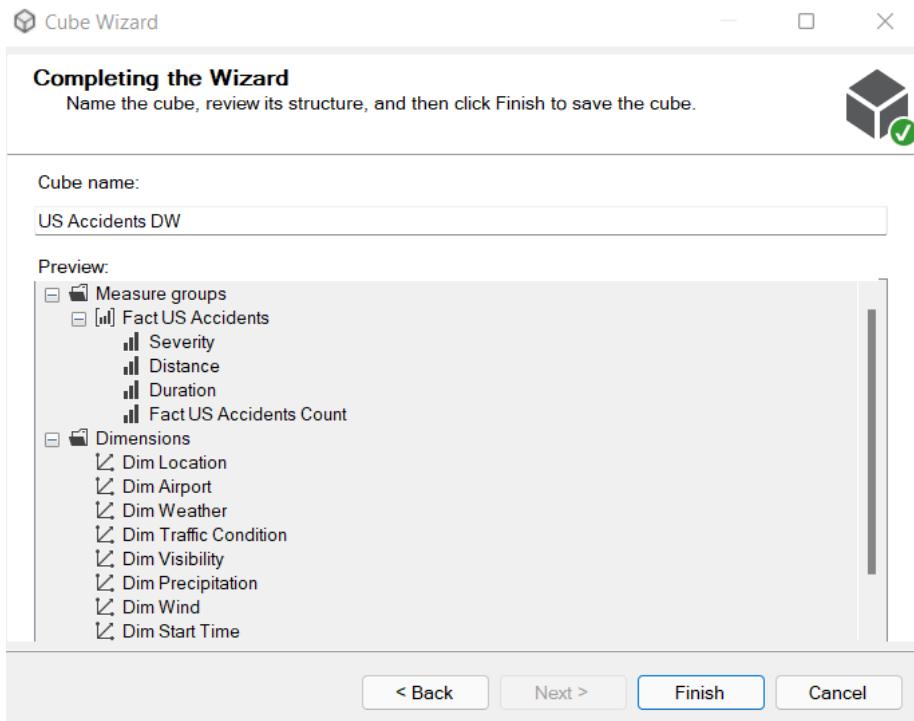
**Bước 5:** Tại trang Select Measure, chọn các độ đo cần thiết. Nhấn Next



**Bước 6:** Tại trang Select New Dimensions, chọn các bảng Dimensions. Nhấn Next

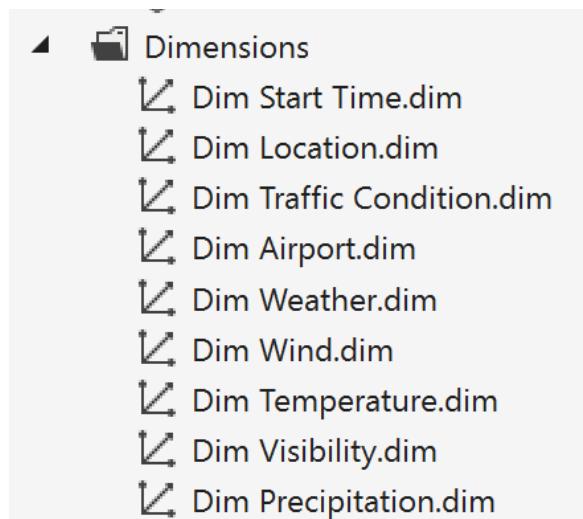


**Bước 7:** Tại trang Completing the Wizard, xem lại các độ đo và các bảng Dimension. Đặt tên Cube. Nhấn Finish để hoàn tất



### 3.2.2 Thêm thuộc tính vào bảng Dimension

Tại Solution Explorer, mở folder Dimensions và nhấp đúp chuột vào các bảng dimension để tiến hành thêm thuộc tính vào bảng



### a. Bảng Dim\_Start\_Time

Từ bảng Dim\_Start\_Time trong Data Source View, kéo thả các thuộc tính của bảng vào phần Attributes

The screenshot shows the 'Dimension Structure' interface. On the left, under 'Attributes', there is a tree view for the 'Dim Start Time' dimension. Underneath it, several attributes are listed: Start Time, Start Time Date, Start Time Date Of Week, Start Time Hour, Start Time Minute, Start Time Month, Start Time Quarter, and Start Time Year. On the right, under 'Data Source View', the 'Dim\_Start\_Time' table is selected. Inside the table, its attributes are listed: Start\_Time, Start\_Time\_Hour, Start\_Time\_Minute, Start\_Time\_Date, Start\_Time\_Month, Start\_Time\_Quarter, Start\_Time\_Year, and Start\_Time\_Date\_of\_Week.

### b. Bảng Dim\_Location

Từ bảng Dim\_Location trong Data Source View, kéo thả các thuộc tính của bảng vào phần Attributes

The screenshot shows the 'Dimension Structure' interface. On the left, under 'Attributes', there is a tree view for the 'Dim Location' dimension. Underneath it, attributes for City, County, Location ID, State, and Street are listed. On the right, under 'Data Source View', the 'Dim\_Location' table is selected. Inside the table, its attributes are listed: Location\_ID, Street, City, County, and State.

### c. Bảng Dim\_Weather

Từ bảng Dim\_Weather trong Data Source View, kéo thả các thuộc tính của bảng vào phần Attributes

The screenshot shows the 'Dimension Structure' interface. On the left, under 'Attributes', there is a tree view for the 'Dim Weather' dimension. Underneath it, attributes for Weather Condition and Weather ID are listed. In the center, under 'Hierarchies', there is a note: 'To create a new hierarchy... drag an attribute'. On the right, under 'Data Source View', the 'Dim\_Weather' table is selected. Inside the table, its attributes are listed: Weather\_ID and Weather\_Condition.

#### d. Bảng Dim\_Traffic\_Condition

Từ bảng Dim\_Traffic\_Condition trong Data Source View, kéo thả các thuộc tính của bảng vào phần Attributes

The screenshot shows the Data Source View interface for the Dim Traffic....dim [Design] cube. The top navigation bar includes tabs for 'Dim Traffic....dim [Design]\*', 'US Accidents....cube [Design]\*', 'Dim Start Time.dim [Design]\*', and 'Dim Location.dim [Design]\*'. Below the tabs are buttons for 'Dimension Structure', 'Attribute Relationships', 'Translations...', and 'Browse...'. The main area is divided into three sections: 'Attributes' (containing a list of traffic conditions like Bump, Crossing, Junction, Railway, Roundabout, Station, Stop, Traffic Calming, Traffic Condition ID, and Traffic Signal), 'Hierarchies' (with a placeholder for creating a new hierarchy), and 'Data Source View' (listing the Dim\_Traffic\_Condition dimension with its attributes: Traffic\_Condition\_ID, Bump, Crossing, Junction, Railway, Roundabout, Station, Stop, Traffic\_Calming, and Traffic\_Signal). A status bar at the bottom shows icons for search, refresh, and other database operations.

#### e. Bảng Dim\_Visibility

Từ bảng Dim\_Visibility trong Data Source View, kéo thả các thuộc tính của bảng vào phần Attributes

The screenshot shows the Data Source View interface for the Dim Visibility.dim [Design] cube. The top navigation bar includes tabs for 'Dim Traffic C...n.dim [Design]\*', 'Dim Start Time.dim [Design]\*', 'Dim Location.dim [Design]\*', and 'Dim Visibility.dim [Design]\*'. Below the tabs are buttons for 'Dimension Structure', 'Attribute Relationships', 'Translations...', and 'Browse...'. The main area is divided into three sections: 'Attributes' (containing Dim\_Visibility dimension with attributes Visibility and Visibility\_ID), 'Hierarchies' (with a placeholder for creating a new hierarchy), and 'Data Source View' (listing the Dim\_Visibility dimension with its attributes: Visibility\_ID and Visibility). A status bar at the bottom shows icons for search, refresh, and other database operations.

#### f. Bảng Dim\_Wind

Từ bảng Dim\_Wind trong Data Source View, kéo thả các thuộc tính của bảng vào phần Attributes

The screenshot shows the Data Source View interface for the Dim Wind.dim [Design] cube. The top navigation bar includes tabs for 'Dim Traffic....dim [Design]\*', 'Dim Start Time.dim [Design]\*', 'Dim Location.dim [Design]\*', and 'Dim Wind.dim [Design]\*'. Below the tabs are buttons for 'Dimension Structure', 'Attribute Relationships', 'Translations...', and 'Browse...'. The main area is divided into three sections: 'Attributes' (containing Dim\_Wind dimension with attributes Wind\_Direction, Wind\_ID, and Wind\_Speed), 'Hierarchies' (with a placeholder for creating a new hierarchy), and 'Data Source View' (listing the Dim\_Wind dimension with its attributes: Wind\_ID, Wind\_Direction, and Wind\_Speed). A status bar at the bottom shows icons for search, refresh, and other database operations.

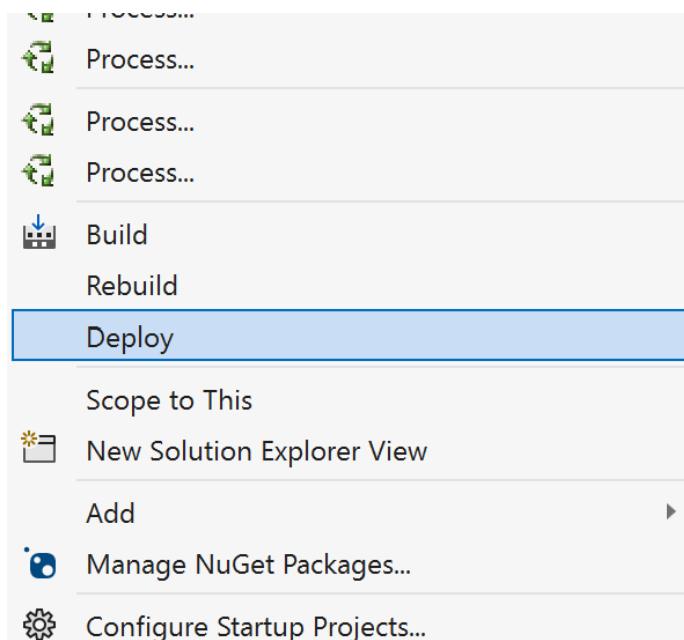
#### g. Bảng Dim\_Airpot

Từ bảng Dim\_Airpot trong Data Source View, kéo thả các thuộc tính của bảng vào phần Attributes

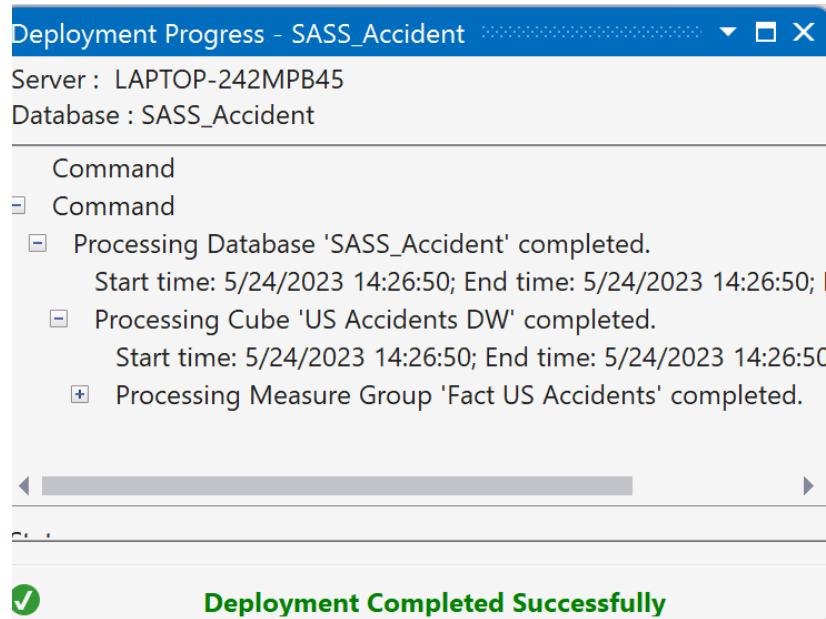
The screenshot shows the 'Dimension Structure' window in SQL Server Management Studio. On the left, under 'Attributes', there is a tree view with 'Dim\_Airport' expanded, showing 'Airport Code' and 'Airport ID'. In the center, under 'Hierarchies', there is a placeholder message: 'To create a new hierarchy... drag an attribute'. On the right, under 'Data Source View', a table named 'Dim\_Airport' is listed with columns 'Airport\_ID' and 'Airport\_Code'. The 'Dim\_Airport' table and its columns are highlighted with a blue selection bar.

### 3.2.3 Triển khai Analysis Service Project

Tại Solution Explorer, nhấp chuột phải vào project đang chạy, nhấn Deploy để triển khai



Kết quả sau khi deploy thành công



### 3.3 Tạo và sửa đổi độ đo, thuộc tính và phân cấp

#### 3.3.1 Xác định các độ đo (Measures)

##### 3.3.1.1 Đổi tên và thuộc tính các độ đo ban đầu

**Bước 1:** Tại khôi vừa tạo, chọn Show Measures Grid để hiện thị chi tiết các độ đo.

The screenshot shows the 'Cube Structure' tool interface. The title bar includes icons for 'Cube Structure' and 'Dimension Usa.' and a toolbar with various buttons. The main area is titled 'Measures' and displays the following hierarchy and details:

- US Accidents DW
  - Fact US Accidents
    - Average Severity
    - Sum Distance
    - Average Duration
    - Fact US Accidents Count
    - Average Distance
    - Min Distance
    - Max Distance

Chi tiết các độ đo sẽ hiển thị dưới dạng bảng, dễ dàng để tương tác.

Measures				
	Name	Measure Group	Data ...	Aggre...
■	Average Severity	Fact US Accidents	Unsig...	Avera...
■	Sum Distance	Fact US Accidents	Double	Sum
■	Average Duration	Fact US Accidents	Integer	Avera...

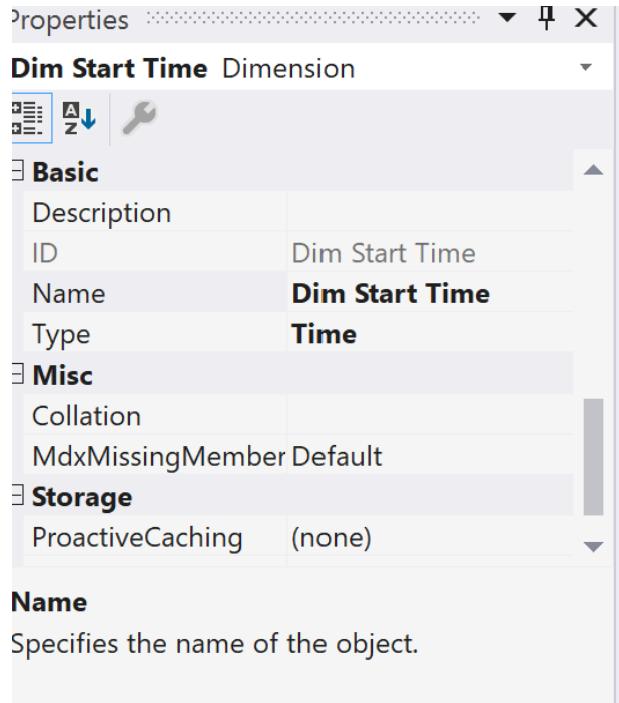
**Bước 2:** Ta đổi tên và thuộc các độ đo hiện tại theo các hàm tổng hợp (aggregation) như sau

- Severity (mức độ nghiêm trọng của vụ tai nạn) đổi thành Average Severity (mức độ nghiêm trọng trung bình của vụ tai nạn), sau đó đổi Aggregation từ Sum sang Average.
- Distance (chiều dài của đoạn đường bị ảnh hưởng bởi vụ tai nạn) đổi thành Sum Distance (tổng chiều dài của các đoạn đường bị ảnh hưởng bởi vụ tai nạn).
- Duration (Khoảng thời gian vụ tai nạn kéo dài) đổi thành Average Duration (Khoảng thời gian vụ tai nạn kéo dài trung bình), sau đó đổi Aggregation từ Sum sang Average.

Measures				
	Name	Measure Group	Data ...	Aggre...
■	Average Severity	Fact US Accidents	Unsig...	Avera...
■	Sum Distance	Fact US Accidents	Double	Sum
■	Average Duration	Fact US Accidents	Integer	Avera...
■	Fact US Accidents Count	Fact US Accidents	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.

**Bước 3:** Mở Dim Start Time.dim. Tại cửa sổ Properties, ta đổi kiểu bảng từ Regular sang Time.



Quá trình hoàn tất, ta có được các độ đo: Average Severity, Sum Distance, Average Duration và Fact US Accidents Count.

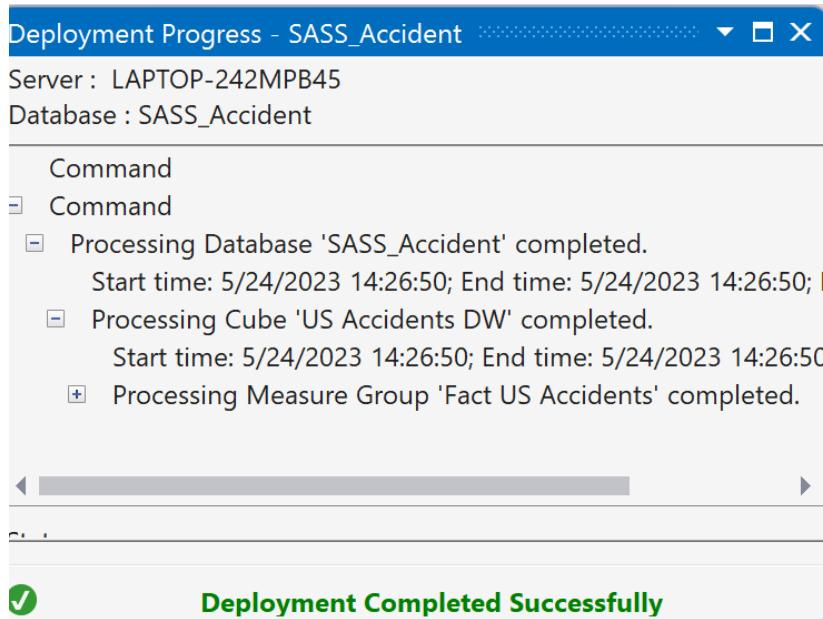
### 3.3.1.2. Tạo các độ đo mới

Để tạo ra các độ đo mới, ta chọn **Add new measure...**, sau đó tạo ra các độ đo mới như sau:

- Average Distance
- Min Distance
- Max Distance

	Name	Measure Group	Data Type	Aggregation
■	Average Severity	Fact US Accidents	UnsignedInt	AverageOfChildren
■	Sum Distance	Fact US Accidents	Double	Sum
■	Average Duration	Fact US Accidents	Integer	AverageOfChildren
■	Fact US Accidents Count	Fact US Accidents	Integer	Count
■	Average Distance	Fact US Accidents	Double	AverageOfChildren
■	Min Distance	Fact US Accidents	Double	Min
■	Max Distance	Fact US Accidents	Double	Sum

Cuối cùng, ta deploy lại project để chắc chắn rằng không xảy ra lỗi



### 3.3.1.3. Tổng kết độ đo

Tổng kết lại ta có được các độ đo sau

ST T	Tên thuộc tính	Kiểu dữ liệu	Ý nghĩa
1	Fact US Accidents Count	Int	Đếm số vụ tai nạn.
2	Average Severity	Double	Mức độ nghiêm trọng trung bình của vụ tai nạn.
3	Sum Distance	Double	Tổng chiều dài của các đoạn đường bị ảnh hưởng bởi vụ tai nạn.
4	Average Distance	Double	Trung bình chiều dài của các đoạn đường bị ảnh hưởng bởi vụ tai nạn.
5	Max Distance	Double	Lấy ra đoạn đường dài nhất bị ảnh hưởng bởi vụ tai nạn.

6	Min Distance	Double	Lấy ra đoạn đường ngắn nhất bị ảnh hưởng bởi vụ tai nạn.
7	Average Duration	Double	Trung bình khoảng thời gian vụ tai nạn kéo dài.

### 3.3.2 Phân cấp trong bảng chiều

#### 3.3.2.1 Phân cấp bảng Dim\_Start\_Time

**Bước 1:** Kéo những thuộc tính cần phân cấp qua cửa sổ hierarchies

The screenshot shows the 'Dim Start Time.dim [Design]' tab selected in the top ribbon. The main area is divided into three panels: 'Attributes' (listing attributes like Start Time, Start Time Date, etc.), 'Hierarchies' (showing a hierarchy tree from Start Time Year to Start Time Minute), and 'Data Source View' (listing physical columns). A tooltip in the 'Hierarchies' panel provides instructions for creating a new hierarchy.

**Bước 2:** Sắp xếp lại các thuộc tính phân cấp theo thứ tự: Start Time Year □ Start Time Quarter □ Start Time Month □ Start Time Date □ Start Time Hour □ Start Time Minute và đổi tên Hierarchy thành Accidents Start Time.

This screenshot shows the same interface as above, but the hierarchy tree has been rearranged. The Start Time Year node is now the root, followed by Start Time Quarter, Start Time Month, Start Time Date, Start Time Hour, and Start Time Minute. The <new level> node remains at the bottom of the tree.

**Bước 3:** Tại panel Attribute Relationships, tạo mối quan hệ như sau



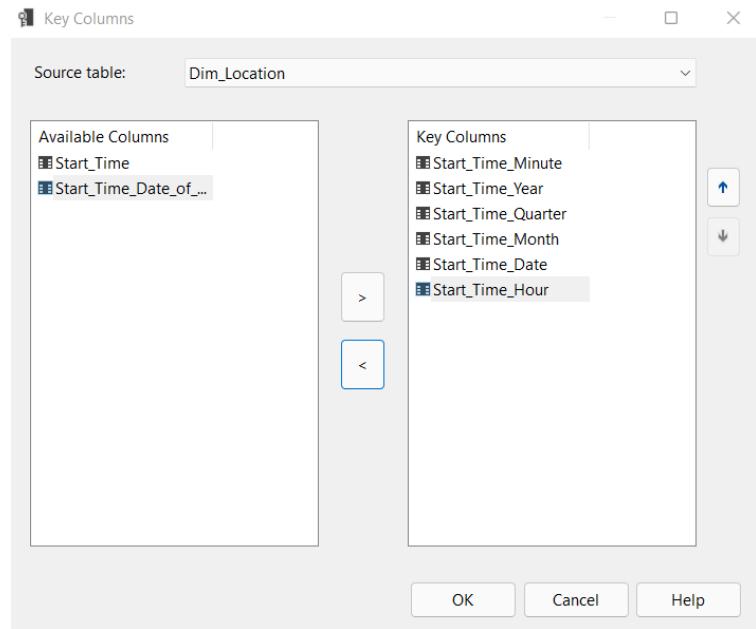
**Bước 4:** Chính khóa cột (KeyColumns) và tên cột (Name Column) của thuộc tính Start Time Minute. Vì thuộc tính Start Time Minute là thuộc tính cấp nhỏ nhất sẽ lấy khóa cột gồm chính nó và những thuộc tính cấp cao hơn.

Tại cửa sổ Properties của thuộc tính Start Time Minute, chọn KeyColumns.

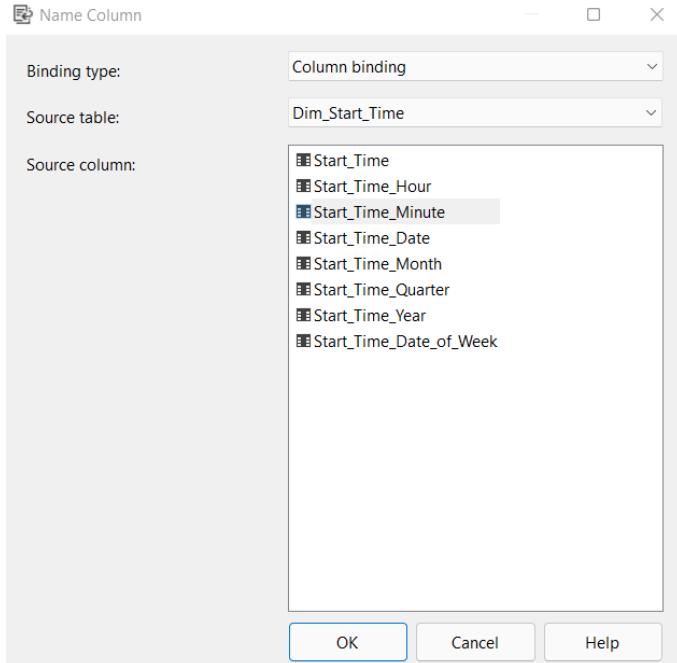
The screenshot shows the properties window for the 'Start Time Minute' dimension attribute. The 'KeyColumns' property is highlighted with a red border, indicating it is selected. The 'Name' property is also highlighted with a red border, showing its current value: 'Dim\_Start\_Time.Start\_Time\_Minute (WChar)'.

Property	Value
MembersWithDataCaption	
NamingTemplate	
RootMemberIf	ParentIsBlankSelfOrMissing
UnaryOperatorColumn	(none)
<b>Source</b>	
CustomRollupColumn	(none)
CustomRollupPropertiesColumn	(none)
<b>KeyColumns</b>	
Dim_Start_Time.Start_Time_Minute (WChar)	
<b>Name</b>	
Specifies the name of the object.	

Thêm các những thuộc tính cấp cao hơn vào KeyColumns, sau đó chọn OK để hoàn tất.



Tại cửa sổ Properties của thuộc tính Start Time Minute, ta chọn Name Column và chọn tên thuộc tính là Start Time Minute.



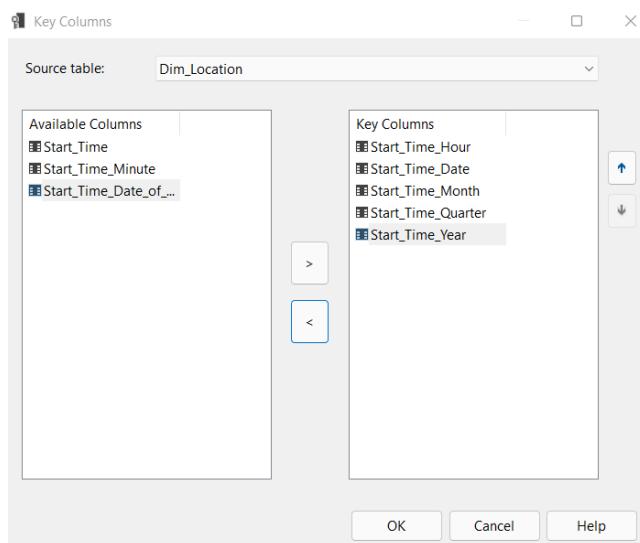
**Bước 5:** Chỉnh khóa cột (KeyColumns) và tên cột (Name Column) của thuộc tính Start Time Hour. Vì thuộc tính Start Time Hour là thuộc tính cấp nhỏ hơn Start Time Year, Start Time

Quarter, Start Time Month, Start Time Date nên sẽ lấy khóa cột gồm chính nó và những thuộc tính cấp cao hơn.

Tại cửa sổ Properties của thuộc tính Start Time Hour, chọn KeyColumns.

The screenshot shows the 'Start Time Hour' DimensionAttribute properties window. In the 'KeyColumns' section, the entry 'Dim\_Start\_Time.Start\_Time\_Hour (WChar)' is highlighted with a red rectangle. Other entries in the list include 'CustomRollupColumn', 'CustomRollupPropertiesColumn', and 'ValueColumn'. The 'Name' section specifies the name of the object.

Thêm các những thuộc tính cấp cao hơn vào KeyColumns, sau đó chọn OK để hoàn tất.



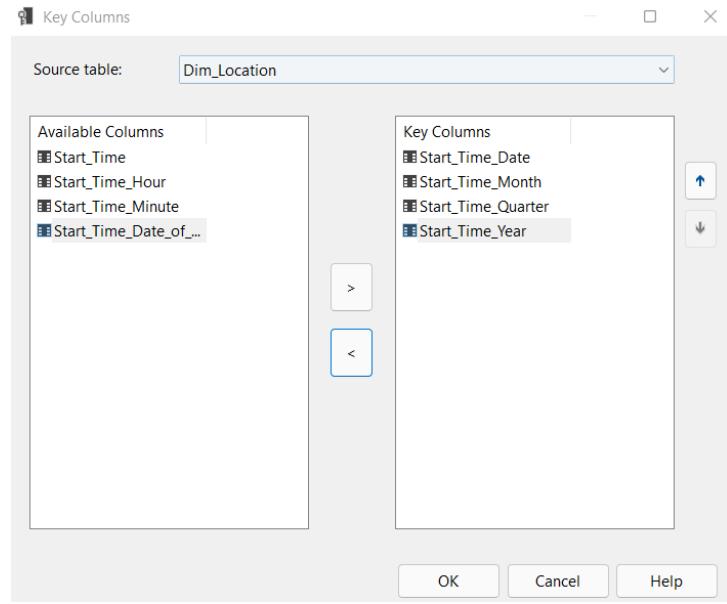
Tại cửa sổ Properties của thuộc tính Start Time Hour, ta chọn Name Column và chọn tên thuộc tính là Start Time Hour.

**Bước 6:** Chỉnh khóa cột (KeyColumns) và tên cột (Name Column) của thuộc tính Start Time Date. Vì thuộc tính Start Time Date là thuộc tính cấp nhỏ hơn Start Time Year, Start Time Quarter, Start Time Month nên sẽ lấy khóa cột gồm chính nó và những thuộc tính cấp cao hơn.

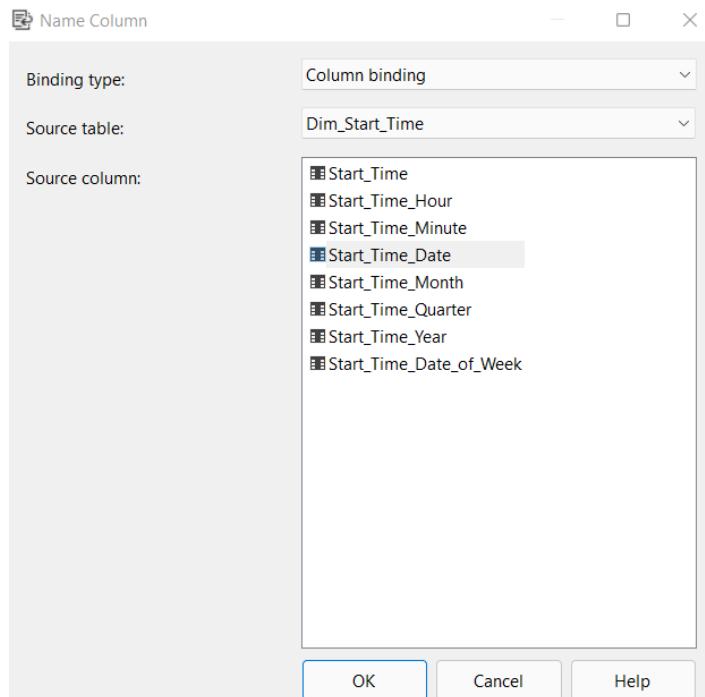
Tại cửa sổ Properties của thuộc tính Start Time Date, chọn KeyColumns.

The screenshot shows the 'Start Time Date' DimensionAttribute Properties window. The 'KeyColumns' section is highlighted with a red box, containing the value 'Dim\_Start\_Time.Start\_Time\_Date (WChar)'. Other visible properties include 'MembersWithDataCaption', 'NamingTemplate', 'RootMemberIf', 'UnaryOperatorColumn', 'Source', 'CustomRollupColumn', 'CustomRollupPropertiesColumn', 'NameColumn', and 'ValueColumn'. The 'Name' section at the bottom specifies the name of the object.

Thêm các những thuộc tính cấp cao hơn vào KeyColumns, sau đó chọn OK để hoàn tất.



Tại cửa sổ Properties của thuộc tính Start Time Date, ta chọn Name Column và chọn tên thuộc tính là Start Time Date.



**Bước 7:** Chỉnh khóa cột (KeyColumns) và tên cột (Name Column) của thuộc tính Start Time Month. Vì thuộc tính Start Time Month là thuộc tính cấp nhỏ hơn Start Time Year, Start Time Quarter nên sẽ lấy khóa cột gồm chính nó và những thuộc tính cấp cao hơn.

Tại cửa sổ Properties của thuộc tính Start Time Date, chọn KeyColumns.

Start Time Month DimensionAttribute

MembersWithCaption  
NamingTemplate  
RootMemberId  
UnaryOperatorColumn

ParentIsBlankSelfOrMissing  
(none)

Source

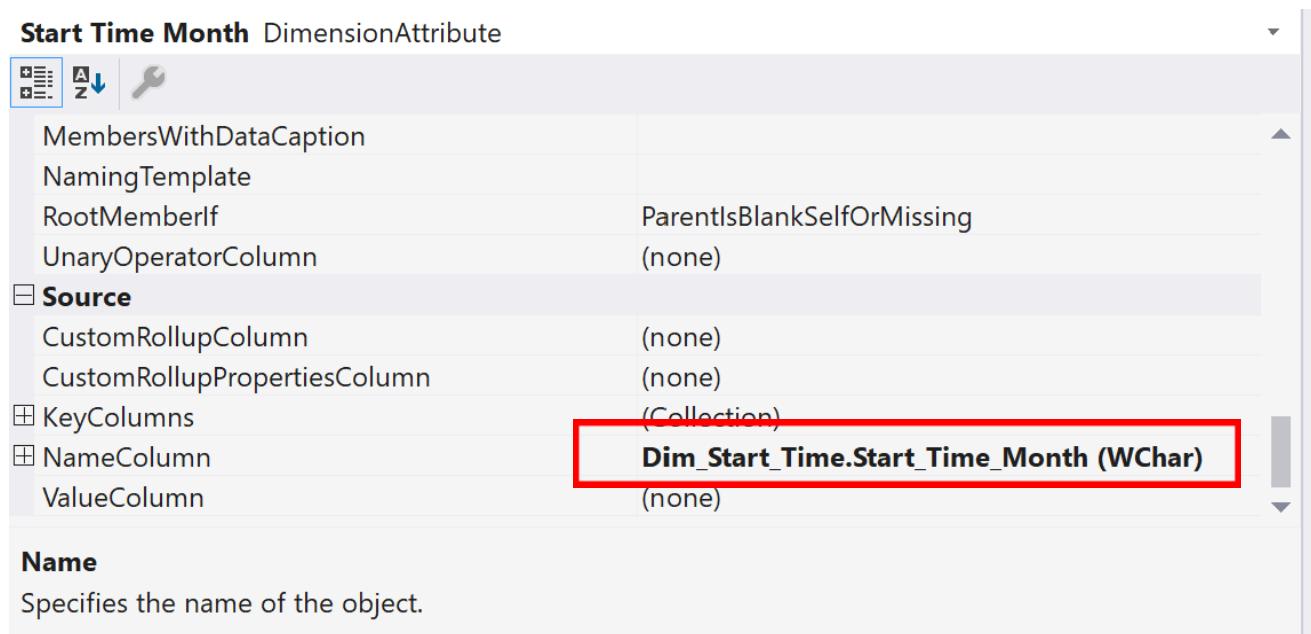
CustomRollupColumn  
CustomRollupPropertiesColumn

KeyColumns  
**Dim\_Start\_Time.Start\_Time\_Month (WChar)**

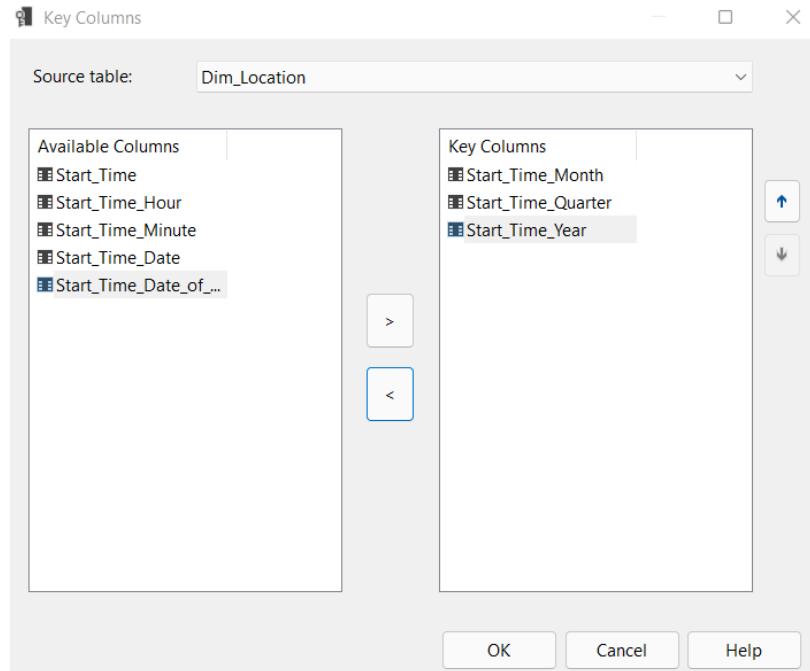
NameColumn  
ValueColumn

(Collection)  
(none)

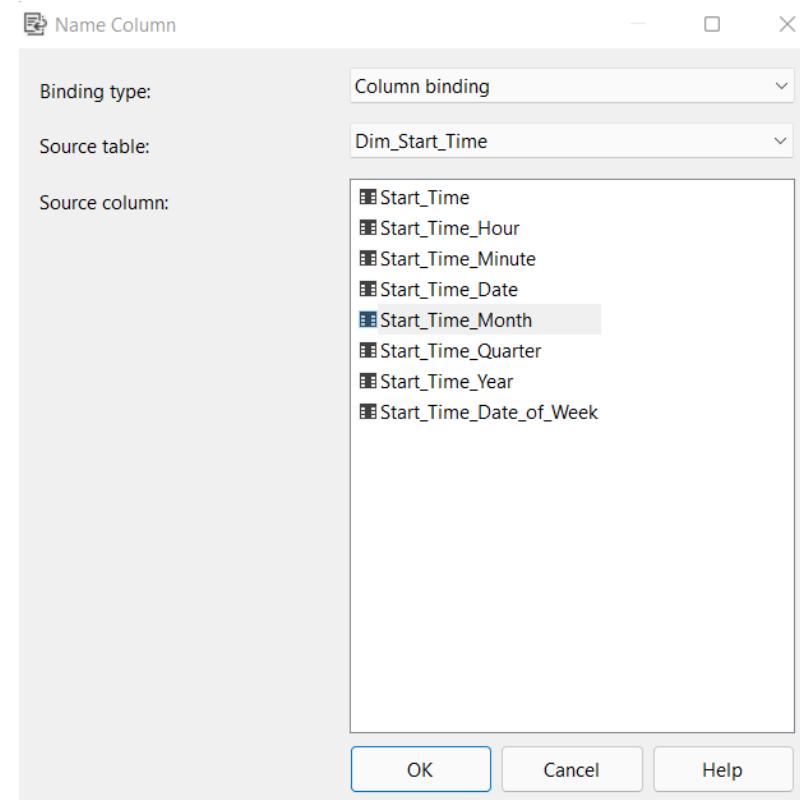
**Name**  
Specifies the name of the object.



Thêm các những thuộc tính cấp cao hơn vào KeyColumns, sau đó chọn OK để hoàn tất.



Tại cửa sổ Properties của thuộc tính Start Time Month, ta chọn Name Column và chọn tên thuộc tính là Start Time Month.



**Bước 7:** Chỉnh khóa cột (KeyColumns) và tên cột (Name Column) của thuộc tính Start Time Quarter. Vì thuộc tính Start Time Quarter là thuộc tính cấp nhỏ hơn Start Time Year nên sẽ lấy khóa cột gồm chính nó và thuộc tính cấp cao hơn.

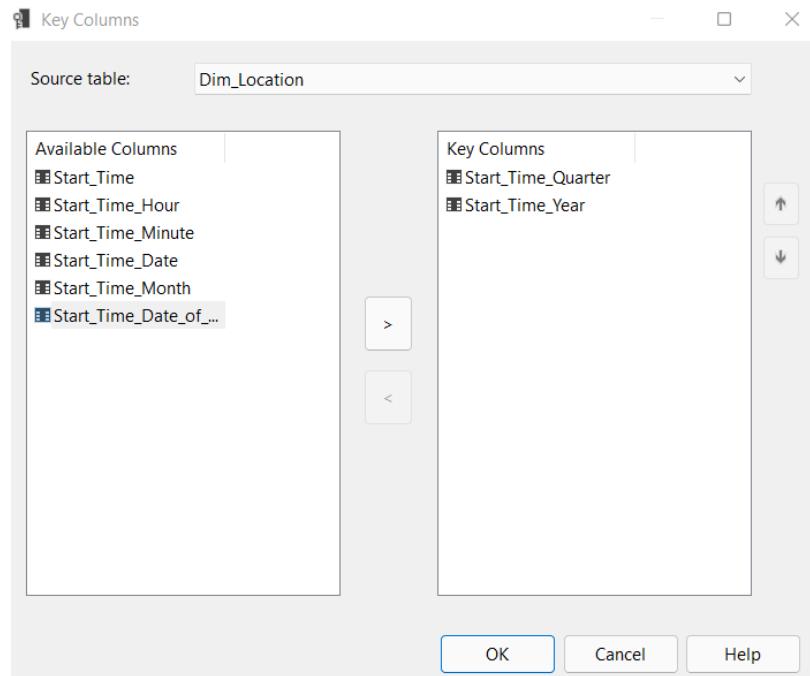
Tại cửa sổ Properties của thuộc tính Start Time Quarter, chọn KeyColumns.

Start Time Quarter DimensionAttribute

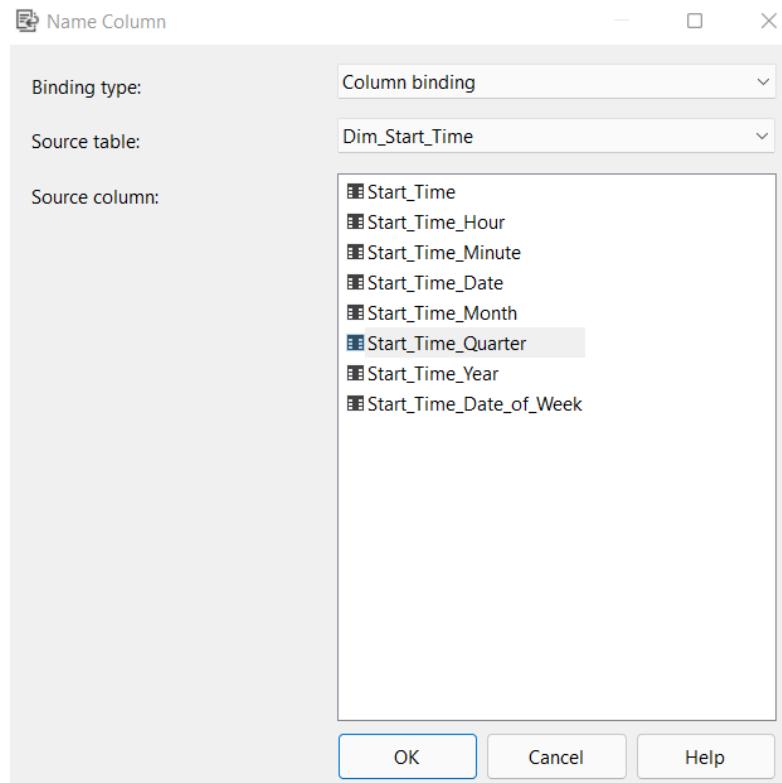
The screenshot shows the properties window for the 'Start Time Quarter' dimension attribute. The 'KeyColumns' property is highlighted with a red border. The window contains the following information:

Property	Value
MembersWithDataCaption	
NamingTemplate	
RootMemberId	
UnaryOperatorColumn	
Source	
CustomRollupColumn	(none)
CustomRollupPropertiesColumn	(none)
KeyColumns	(Collection)
NameColumn	Dim_Start_Time.Start_Time_Quarter (WChar)
ValueColumn	(none)
Name	Specifies the name of the object.

Thêm các những thuộc tính cấp cao hơn vào KeyColumns, sau đó chọn OK để hoàn tất.



Tại cửa sổ Properties của thuộc tính Start Time Quarter, ta chọn Name Column và chọn tên thuộc tính là Start Time Quarter.



### 3.3.2.2. Phân cấp bảng Dim\_Location

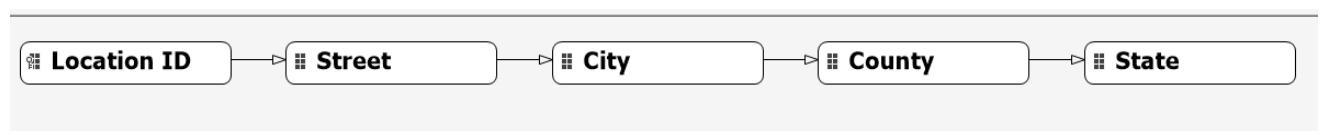
Bước 1: Kéo những thuộc tính cần phân cấp qua cửa sổ hierarchies

The screenshot shows the 'Dim Location.dim [Design]' tab selected in the top bar. The left pane, 'Attributes', lists the dimension's attributes: Dim Location, City, County, Location ID, State, and Street. The middle pane, 'Hierarchies', displays a hierarchy structure with levels: State, County, City, and Street, with '<new level>' indicated. A tooltip on the right says 'To create a new hierarchy... drag an attribute'. The right pane, 'Data Source View', shows the table 'Dim\_Location' with columns: Location\_ID, Street, City, County, and State.

Bước 2: Sắp xếp lại các thuộc tính phân cấp theo thứ tự: State □ County □ City □ Street và đổi tên Hierarchies thành Accidents Location.

This screenshot shows the same Dimension Studio interface as above, but the hierarchy structure in the 'Hierarchies' pane has been reordered: State, County, City, and Street, with '<new level>' still present. The tooltip 'To create a new hierarchy... drag an attribute' remains visible.

Bước 3: Tại panel Attribute Relationships, tạo mối quan hệ như sau

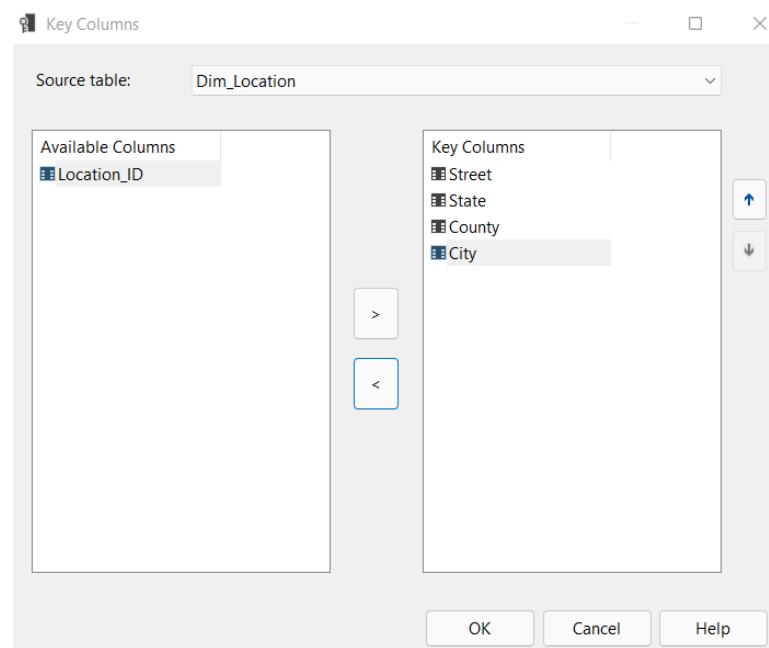


Bước 4: Chính khóa cột (KeyColumns) và tên cột (Name Column) của thuộc tính Street. Vì thuộc tính Street là thuộc tính cấp thấp nhất sẽ lấy khóa cột gồm chính nó và những thuộc tính cấp cao hơn.

Tại cửa sổ Properties của thuộc tính Street, chọn KeyColumns.

The screenshot shows the 'Street' DimensionAttribute properties. The 'KeyColumns' section is expanded, showing 'Dim\_Location.Street (WChar)' selected as the key column. Other properties like NamingTemplate, RootMemberIf, and UnaryOperatorColumn are also visible.

Thêm các những thuộc tính cấp cao hơn vào KeyColumns, sau đó chọn OK để hoàn tất.



Tại cửa sổ Properties của thuộc tính Street, ta chọn Name Column và chọn tên thuộc tính là Street.

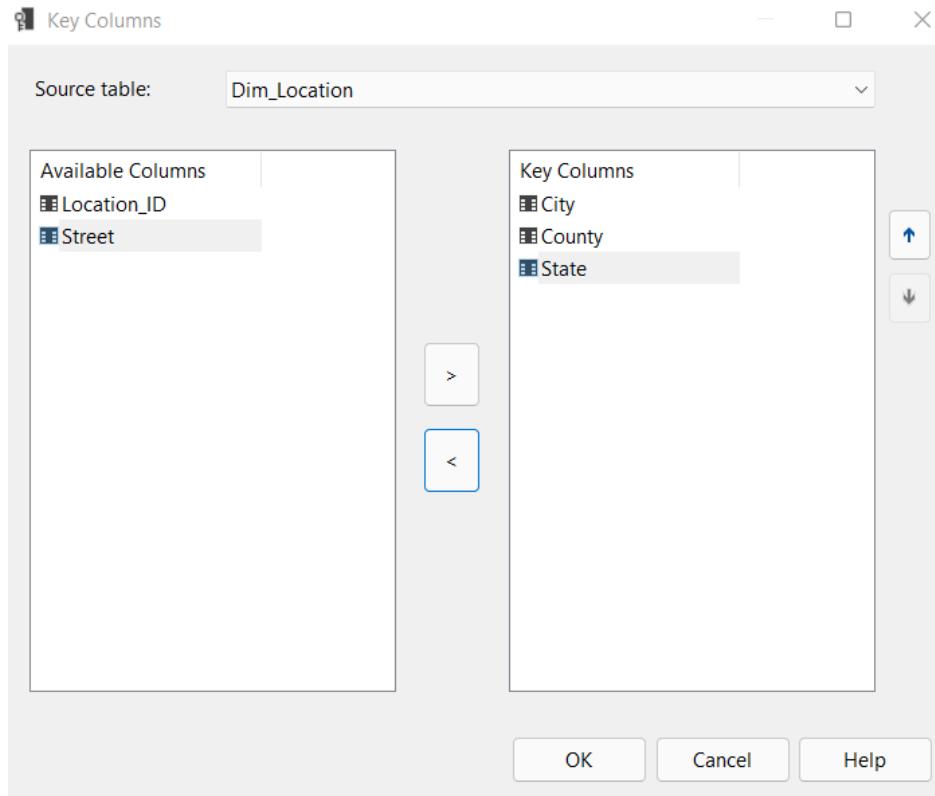
**Bước 5:** Chính khóa cột (KeyColumns) và tên cột (Name Column) của thuộc tính City. Vì thuộc tính City là thuộc tính cấp nhỏ hơn State, County nên sẽ lấy khóa cột gồm chính nó và những thuộc tính cấp cao hơn.

Tại cửa sổ Properties của thuộc tính City, chọn KeyColumns.

The screenshot shows the 'City' DimensionAttribute properties dialog. At the top, there are three icons: a grid for MembersWithCaption, a downward arrow for NamingTemplate, and a key for KeyColumns. Below these are several properties:

MembersWithCaption	
NamingTemplate	
RootMemberIf	ParentIsBlankSelfOrMissing
UnaryOperatorColumn	(none)
<b>Source</b>	
CustomRollupColumn	(none)
CustomRollupPropertiesColumn	(none)
<b>KeyColumns</b>	
NameColumn	(Collection)
ValueColumn	<b>Dim_Location.City (WChar)</b>
<b>Name</b>	
Specifies the name of the object.	

Thêm các những thuộc tính cấp cao hơn vào KeyColumns, sau đó chọn OK để hoàn tất.



Tại cửa sổ Properties của thuộc tính City, ta chọn Name Column và chọn tên thuộc tính là City.

**Bước 6:** Chính khóa cột (KeyColumns) và tên cột (Name Column) của thuộc tính County. Vì thuộc tính County là thuộc tính cấp nhỏ hơn State nên sẽ lấy khóa cột gồm chính nó và thuộc tính cấp cao hơn.

Tại cửa sổ Properties của thuộc tính City, chọn KeyColumns.

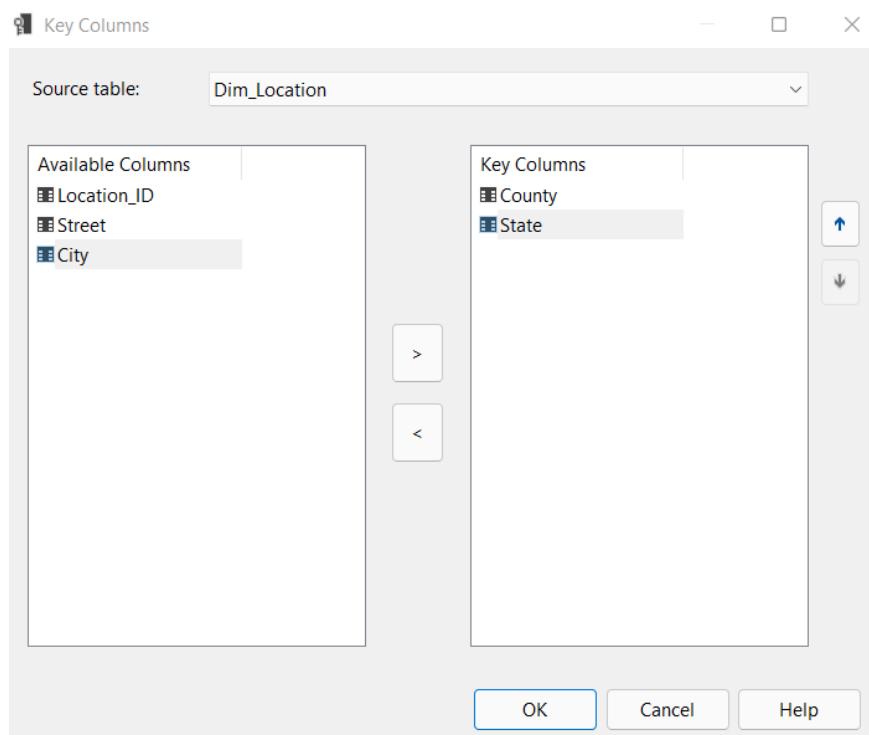
**County** DimensionAttribute

MembersWithDataCaption  
NamingTemplate  
RootMemberIf  
UnaryOperatorColumn  
**Source**  
CustomRollupColumn  
CustomRollupPropertiesColumn  
**KeyColumns**  
**NameColumn**  
ValueColumn

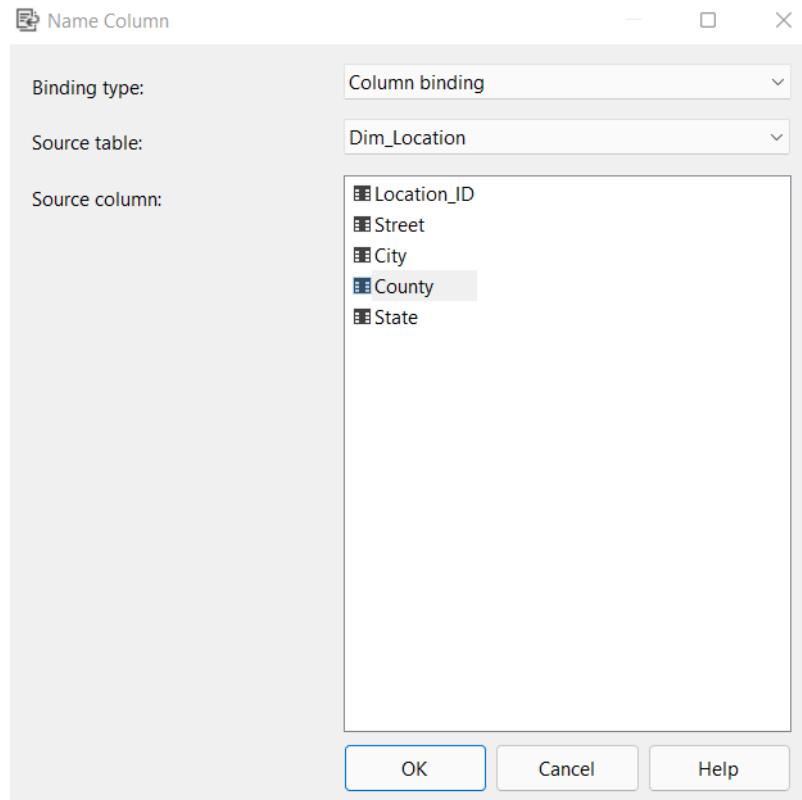
ParentIsBlankSelfOrMissing  
(none)

**Dim\_Location.County (WChar)**

Thêm các những thuộc tính cấp cao hơn vào KeyColumns, sau đó chọn OK để hoàn tất.



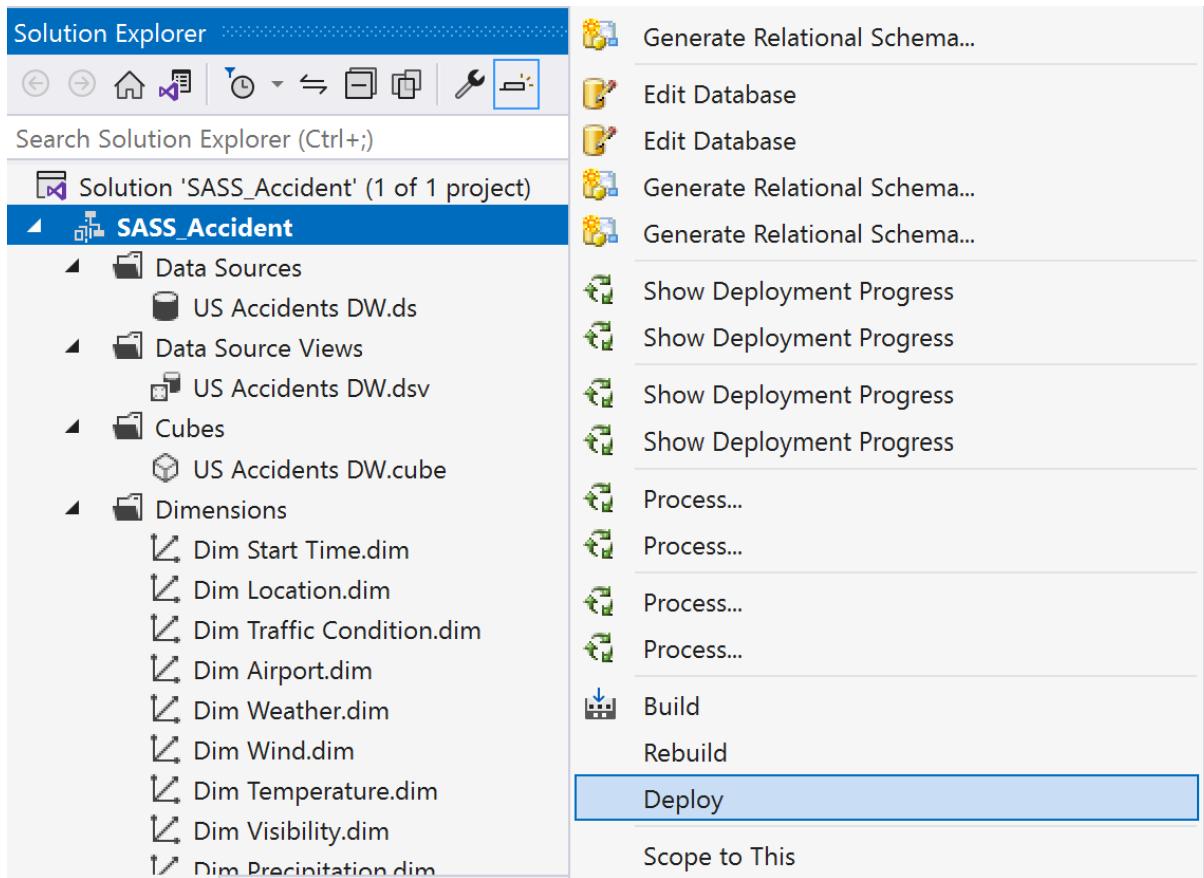
Tại cửa sổ Properties của thuộc tính County, ta chọn Name Column và chọn tên thuộc tính là County.



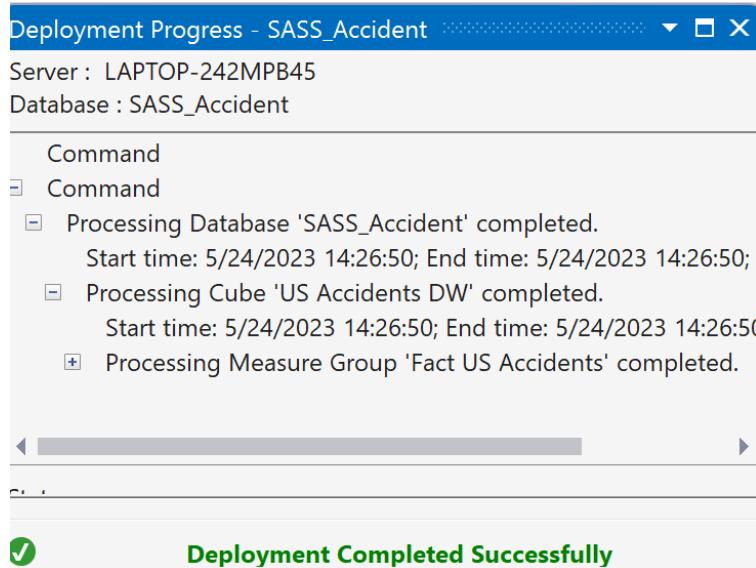
### 3.3.2.3 Deploy project

Sau khi quá trình phân cấp cho các bảng chiều hoàn tất, ta thực hiện deploy project để đảm bảo không có lỗi xảy ra sau quá trình phân cấp.

Nhấn chuột phải vào tên project (SSAS\_US\_Accidents) và nhấn Deploy như hình sau



Khi deploy thành công, hệ thống sẽ hiển thị như hình sau và chúng ta bắt đầu thực hiện các câu truy vấn.



### 3.4 Thực hiện truy vấn bằng ngôn ngữ truy vấn kho dữ liệu MDX

**MDX** là ngôn ngữ truy vấn cho cơ sở dữ liệu nhiều chiều, nó tương tự ngôn ngữ SQL cho cơ sở dữ liệu dạng quan hệ, tuy nhiên đây là ngôn ngữ tính toán vì thế nó tương tự có cú pháp giống công thức của bảng tính.

Cấu trúc của MDX giống như SQL nhưng mở rộng hơn để thao tác với cơ sở dữ liệu nhiều chiều. Câu truy vấn MDX có cấu trúc như sau:

Mệnh đề **SELECT** dùng để xác định các chiều của tập hợp kết quả.

Mệnh đề **FROM** xác định nguồn dữ liệu (cube) dùng để lấy dữ liệu

Mệnh đề **WHERE** dùng để xác định chiều cắt dữ liệu , nhằm lọc dữ liệu đầu ra.

**Câu 1: Liệt kê 10 quận xảy ra nhiều vụ tai nạn nhất và sắp xếp theo thứ tự tăng dần.**

#### a. Thực hiện manual (thao tác bằng tay) bằng Visual Studio

**Giải thích:** TAIL là một hàm trong ngôn ngữ truy vấn đa chiều (MDX) được sử dụng trong các hệ thống OLAP (Online Analytical Processing) để lấy một tập hợp con cuối cùng của các thành viên trong một tập hợp dữ liệu.

**Ý nghĩa trong câu truy vấn:** hàm TAIL được sử dụng để lấy 10 thành viên cuối cùng trong tập hợp các thành viên của đối tượng [Dim Location].[County], được sắp xếp theo giá trị của [Measures].[Fact US Accidents Count] theo thứ tự tăng dần (ASC)

- Sử dụng Named set

TAIL(

```
    ORDER([Dim Location].[County].MEMBERS,  
          [Measures].[Fact US Accidents Count], ASC), 10)
```

**Đổi sang chế độ Script View:**

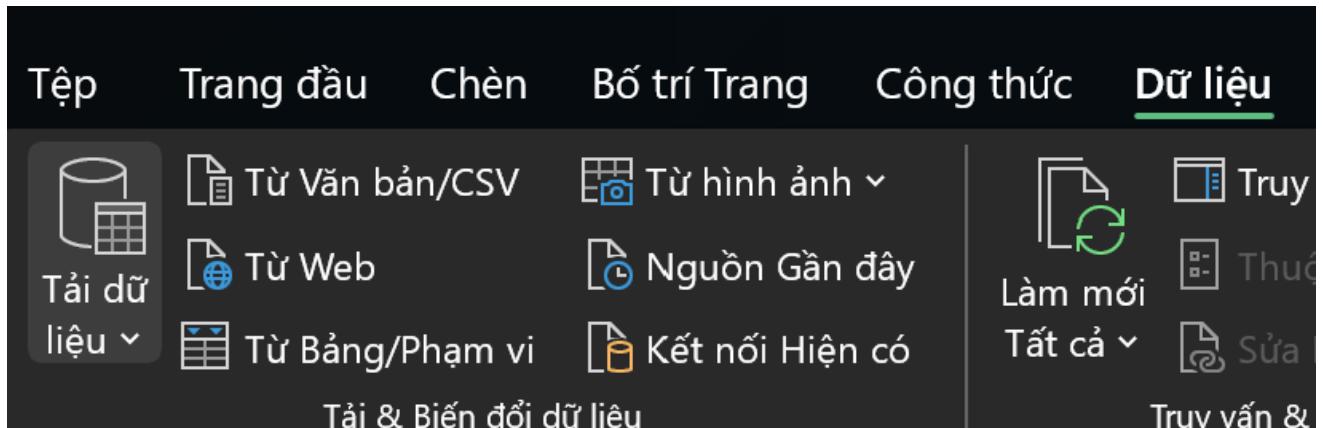
**Kết quả :**

```
SELECT {[Measures].[Fact US Accidents Count]} ON COLUMNS,  
NON EMPTY [Cau1] ON ROWS  
FROM [US Accidents DW];
```

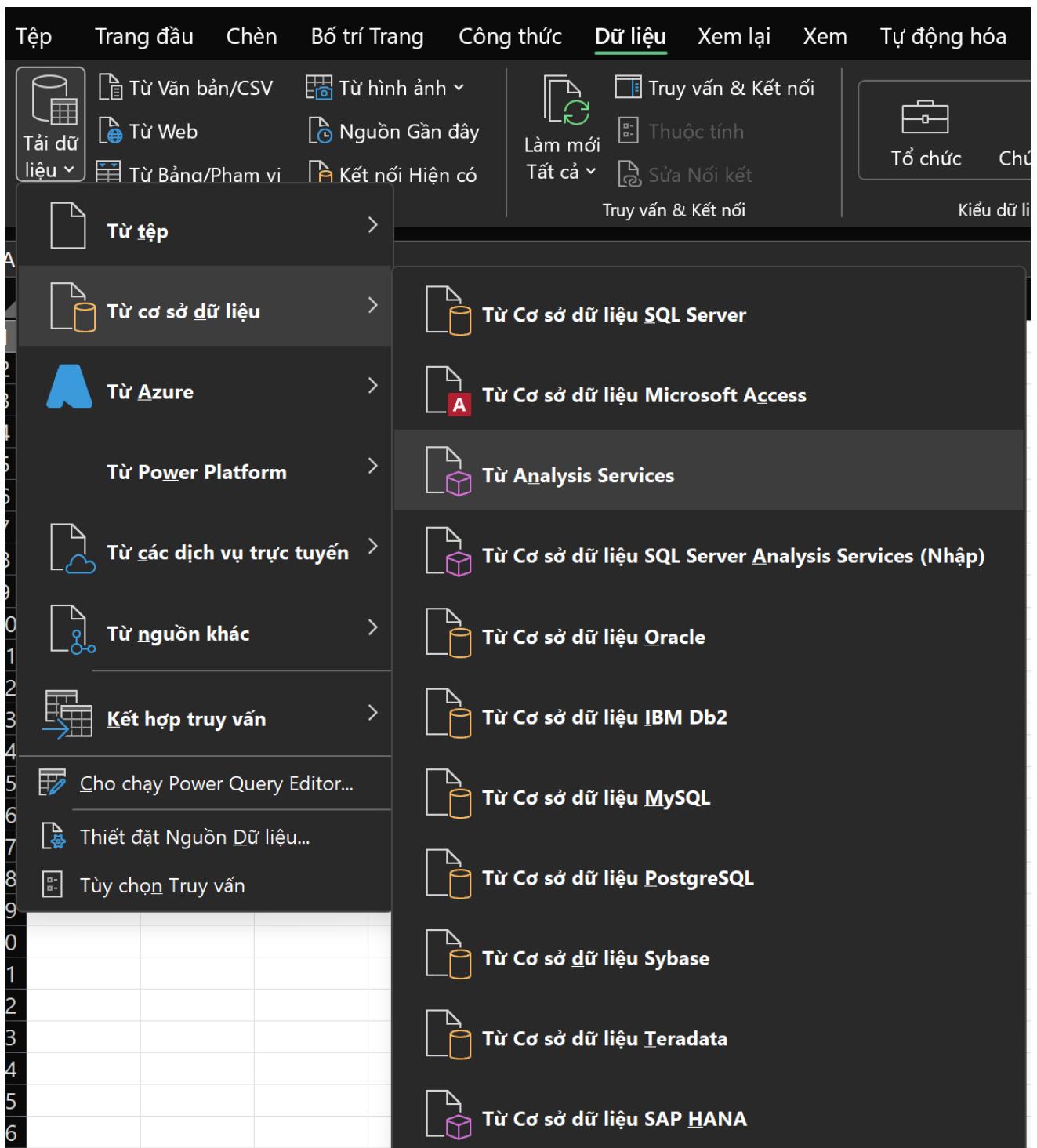
County	Fact US Accidents Count
Riverside	40665
Harris	42000
Sacramento	46479
San Diego	47016
Dallas	49331
Orange	50788
San Bernardino	54179
Orange	60206
Miami-Dade	142910
Los Angeles	225192

### b. Thực hiện bằng Pivot Excel

**Bước 1:** Vào Excel, sau đó nhấn vào tab “Dữ liệu”, chọn “Tải dữ liệu”



**Bước 2:** Chọn “Tù cơ sở dữ liệu”, sau đó chọn tiếp “ Analysis Services ”



**Bước 3:** Màn hình Excel sẽ hiện trang kết nối đến máy chủ SSMS, điền connect và chọn OK

### Kết nối tới máy chủ Cơ sở dữ liệu

Nhập thông tin cần thiết để kết nối tới máy chủ cơ sở dữ liệu.

1. Tên máy chủ: LAPTOP-242MPB45

2. Ủy nhiệm đăng nhập

Dùng Xác thực của Windows

Sử dụng tt>he Tên Người dùng và Mật khẩu sau

Tên Người dùng:

Mật khẩu:

Hủy bỏ

< Lùi lại

Tiếp >

Kết thúc

**Bước 4:** Chọn SSAS tùy muñn, sau đó chọn Fact,Cube phù hợp với yêu cầu truy vấn nếu dữ liệu quá lớn

**Chọn Cơ sở dữ liệu và Bảng**

Chọn Cơ sở dữ liệu và Bảng/Khối vuông chứa dữ liệu bạn muốn.

Chọn cơ sở dữ liệu chứa dữ liệu bạn cần:

Kết nối tới Khối vuông hay bảng đặc thù:

Tên	Mô tả	Đã chỉnh sửa	Đã tạo	Loại
 US Accidents DW		5/27/2023 13:55:40		CUBE

Hủy bỏ < Lùi lại Tiếp > Kết thúc

**Bước 5:** Chọn tên tệp để lưu, sau đó nhấn Kết Thúc để Excel load dữ liệu và cube từ SSAS

**Lưu Tệp Kết nối Dữ liệu và Kết thúc**

Nhập tên và mô tả cho tệp Kết nối Dữ liệu mới của bạn và nhấn Kết thúc để lưu.

Tên Tệp:

  Lưu mật khẩu trong tệp

Mô tả:

(Để trợ giúp người khác hiểu kết nối dữ liệu của bạn chỉ tới gì)

Tên Thân thiện:

Từ Khoá Tim Kiếm:

 Luôn cố gắng sử dụng tệp này để làm mới dữ liệuExcel Services: **Màn hình Excel hiện ra giao diện như sau:**

**Bước 6:** Sau đó thực hiện Pivot Table, kéo thả thuộc tính County trong bảng Dim Location vào cửa sổ Rows trong PivotTable. Sau đó, chọn độ đo Fact US Accidents Count

Nhãn Hàng	Tổng của Fact US Accidents Count
Abbeville	298
Acadia	83
Accomack	211
Ada	3969
Adair	181
Adams	4539
Addison	22
Aiken	2588
Aitkin	394
Alachua	2383
Alamance	460
Alameda	38303
Alamosa	6
Albany	4563
Albemarle	1815
Alcona	5
Alcorn	5
Alexander	56
Alexandria	1534
Alexandria City	9
Allamakee	15
Allegan	432
Allegany	399
Allegheny	183
Allegheny	5823

Sau đó chọn “Nhân Hàng” để tiến hành sắp xếp số lượng tai nạn xảy ra từ cao đến thấp theo top 10 bằng cách chọn “Bộ lọc Giá trị” và chọn “10 trên cùng”

The screenshot shows a dialog box titled "Nhân Hàng" (Sort) with the sub-titile "Tổng của Fact US Accidents Count". On the left, there are sorting options: "Sắp xếp từ A đến Z" (A to Z), "Sắp xếp từ Z đến A" (Z to A), and "Các tùy chọn sắp xếp khác..." (Other sorting options). Below these are filter options: "Bộ lọc Nhân" (Filter by Category) and "Bộ lọc Giá trị" (Filter by Value). A search bar "Tim kiếm" (Search) is also present. The main area lists county names with checkboxes, including Abbeville, Acadia, Accomack, Ada, Adair, Adams, Addison, Aiken, Aitkin, Alachua, Alamance, Alameda, Alamosa, and Albermarle. On the right, a list of accident counts is shown, with the value "54179" highlighted. A dropdown menu for filtering values includes "Bằng...", "Không Bằng...", "Lớn Hơn...", "Lớn Hơn Hoặc Bằng...", "Nhỏ Hơn...", "Nhỏ Hơn Hoặc Bằng...", "Giữa...", "Ngoài khoảng...", and "10 Trên cùng...". The value "19226" is also highlighted in this list.

Tổng của Fact US Accidents Count
225192
142910
113365
<b>54179</b>
49746
47016
46479
42080

Bộ lọc Giá trị
19832
<b>19588</b>
19468
<b>19226</b>

Nhãn Hàng	Fact US Accidents Count
Los Angeles	225192
Miami-Dade	142910
Orange	60206
San Bernardino	54179
Orange	50788
Dallas	49331
San Diego	47016
Sacramento	46479
Harris	42000
Riverside	40665
<b>Tổng Cuối</b>	<b>758766</b>

Câu 1

The screenshot shows a PivotTable in Excel with the following data:

Nhãn Hàng	Fact US Accidents Count
Los Angeles	225192
Miami-Dade	142910
Orange	60206
San Bernardino	54179
Orange	50788
Dallas	49331
San Diego	47016
Sacramento	46479
Harris	42000
Riverside	40665
<b>Tổng Cuối</b>	<b>758766</b>

A yellow box highlights the value **758766** in the 'Tổng Cuối' row.

The PivotTable settings pane on the right shows the following configuration:

- Trường PivotTable**: Chọn các trường để thêm vào báo cáo: (Dim Location, County, Location ID, State, Street)
- Tim kiếm**: (Search bar)
- Hierarchy**: Dim Location, Thêm Trường (County checked)
- Bộ lọc**: (Filter) - Cột: Hàng (County), Giá trị (Fact US Accidents Count)
- Cập nhật**: (Update button)

### c. Thực hiện bằng MDX

```

SELECT [Measures].[Fact US Accidents Count] ON COLUMNS,
TAIL(
    ORDER([Dim Location].[County].MEMBERS,
        [Measures].[Fact US Accidents Count], ASC), 10) ON ROWS
FROM [US Accidents DW];

```

Ta thu được kết quả như đã lọc trong Pivot table Excel

```

--Cau1
SELECT [Measures].[Fact US Accidents Count] ON COLUMNS,
TAIL(
    ORDER([Dim Location].[County].MEMBERS,
        [Measures].[Fact US Accidents Count], ASC), 10) ON ROWS
FROM [US Accidents DW];

```

County	Fact US Accidents Count
Riverside	40665
Harris	42000
Sacramento	46479
San Diego	47016
Dallas	49331
Orange	50788
San Bernardino	54179
Orange	60206
Miami-Dade	142910
Los Angeles	225192

Câu 2: Thông kê số vụ tai nạn xảy ra tại đường ray xe lửa ở các bang từ năm 2020 đến 2021

- a. Thực hiện manual bằng Visual Studio

**Bước 1:** Kéo thả thuộc tính Start Time Year, State và độ đo Fact US Accidents Count vào cửa sổ thực thi câu truy vấn.

**Bước 2:** Kéo thả dimension Dim Start Time và Dim Traffic Condition vào cửa sổ lọc. Sau đó lọc ra các vụ tai nạn xảy ra tại đường ray xe lửa từ năm 2020 đến 2021.

**Bước 3:** Thực thi câu truy vấn, ta thu được kết quả như sau

### Kết quả

Dimension	Hierarchy	Operator	Filter Expression	Param...
Dim Start Time	# Start Time Year	Equal	{ 2021, 2020 }	<input type="checkbox"/> <input type="checkbox"/>
Dim Traffic Condition	# Railway	Equal	{ True }	<input type="checkbox"/> <input type="checkbox"/>
<Select dimension>				<input type="checkbox"/> <input type="checkbox"/>
Start Time Year			Fact US Accidents Count	
2020	OK	39		
2020	OR	426		
2020	PA	216		
2020	SC	186		
2020	SD	3		
2020	TN	77		
2020	TX	196		
2020	UT	58		
2020	VA	98		
2020	WA	94		
2020	WI	1		
2020	WV	10		
2021	AL	13		
2021	AR	35		
2021	AZ	271		
2021	CA	3051		
2021	CO	57		
2021	CT	41		
2021	DC	128		
2021	DE	18		
2021	FL	1516		
2021	GA	46		
2021	IA	14		
2021	ID	22		
2021	IL	52		

### b. Câu lệnh truy vấn bằng Pivot Excel

**Trường PivotTable**

Chọn các trường để thêm vào báo cáo:

Tìm kiếm

Railway

Bump  
Crossing  
Junction  
 Railway  
Roundabout  
Station  
Stop

Kéo trường giữa các vùng bên dưới:

<b>Bộ lọc</b> Railway	<b>Cột</b>
<b>Hàng</b> Start Time Year State	<b>Giá trị</b> Fact US Accidents Count

Trì hoãn Cập nhật Bố trí      **Cập nhật**

100%

The screenshot shows the PivotTable Fields pane open in Excel. On the left, there's a list of fields: Nhãn Hàng (Fact US Accidents Count), 2021, AL, AR, AZ, CA, CO, CT, DC, DE, FL, GA, IA, ID, IL, IN, KS, KY, LA, MA, MD, ME, MI, MN, MO, MS, MT, NC. The 'Railway' field is highlighted in green and has a dropdown arrow next to it. On the right, under 'Chọn các trường để thêm vào báo cáo:', 'Railway' is checked with a green checkmark. Below that is a list of other categories: Bump, Crossing, Junction, Roundabout, Station, Stop. The 'Bộ lọc' (Filter) section contains 'Railway'. The 'Hàng' (Rows) section contains 'Start Time Year' and 'State'. The 'Giá trị' (Values) section contains 'Fact US Accidents Count'. At the bottom, there's a checkbox for 'Trì hoãn Cập nhật Bố trí' (Delay Refresh) and a 'Cập nhật' (Update) button.

**Kết quả sau khi thực hiện Pivot Table trên Excel**

Railway	True	
<b>Nhãn Hàng</b> Fact US Accidents Count		
2021		
AL	13	▼
AR	35	
AZ	271	
CA	3051	
CO	57	
CT	41	
DC	128	
DE	18	
FL	1516	
GA	46	
IA	14	
ID	22	
IL	52	
IN	31	
KS	17	
KY	3	
LA	1026	
MA	3	
MD	94	
ME	1	
MI	65	
MN	167	
MO	52	
MS	4	
MT	82	
NC	380	
2020		
AL	9	▲
AR	11	
AZ	254	
CA	1927	
CO	42	
CT	7	
DC	43	
DE	19	
FL	435	
GA	35	
IA	6	
ID	7	
IL	268	
IN	6	
KS	5	
KY	14	
LA	379	
MA	28	
MD	29	
MI	29	
MN	70	
MO	24	
MT	22	
NC	289	
NF	1	

### c. Thực hiện bằng MDX ( sử dụng Drill Down)

```
SELECT [Measures].[Fact US Accidents Count] ON COLUMNS,  
NONEMPTY(DrilldownLevel(  
    DrilldownLevel(  
        CrossJoin(  
            {[Dim Start Time].[Start Time Year].[Start Time Year].&[2020], [Dim Start  
Time].[Start Time Year].[Start Time Year].&[2021]},  
            [Dim Location].[State].[State]  
        )  
    )  
) ON ROWS  
FROM [US Accidents DW]  
WHERE [Dim Traffic Condition].[Railway].&[TRUE];
```

The screenshot shows the SSMS Results pane with the following content:

```
SELECT [Measures].[Fact US Accidents Count] ON 0,  
NONEMPTY(CROSSJOIN(  
    {[Dim Start Time].[Start Time Year].[Start Time Year].&[2020], [Dim Start Time].[Start Time Year].[Start Time Year].&[2021]},  
    [Dim Location].[State].[State]  
) ON 1  
FROM [US Accidents DW]  
WHERE [Dim Traffic Condition].[Railway].&[TRUE];
```

Messages

Fact US Accidents Count		
2020	TX	196
2020	UT	58
2020	VA	98
2020	WA	94
2020	WI	1
2020	WV	10
2021	AL	13
2021	AR	35
2021	AZ	271
2021	CA	3051
2021	CO	57
2021	CT	41
2021	DC	128
2021	DE	18
2021	FL	1516
2021	GA	46
2021	IA	14
2021	ID	22
...	...	...

Kết quả: Kết quả giữa PivotTable và MDX giống nhau

**Câu 3:** Cho biết số lượng các vụ tai nạn xảy ra của bốn hướng gió: Đông, Tây, Nam, Bắc thuộc các thành phố tại các tiểu bang với tốc độ gió lớn hơn 10 dặm/ giờ.

**a. Thực hiện manual bằng Visual Studio**

Tạo Calculated Member:

```
[Measures].[Fact US Accidents Count] / Sum([Dim Wind].[Wind Direction].CurrentMember.Parent, [Measures].[Fact US Accidents Count])
```

Sử dụng Named Set:

FILTER(

```
[Dim Location].[State].[State] *  
[Dim Location].[County].[County] *  
[Dim Location].[City].[City] *  
{[Dim Wind].[Wind Direction].[E], [S], [W], [N]},  
[Dim Wind].[Wind Speed] > 10 AND  
Left([Dim Location].[State].CurrentMember.Name, 1) = "M")
```

Chuyển sang Script Mode

```
SELECT {[Measures].[Fact US Accidents Count],  
[Measures].[AccidentsPercentageByWindDirection]} ON COLUMNS,  
NON EMPTY [Cau3] ON ROWS  
FROM [US Accidents DW];
```

**Kết quả:**

```

SELECT {[Measures].[Fact US Accidents Count], [Measures].[AccidentsPercentageByWindDirection]} ON COLUMNS,
NON EMPTY [Cau3] ON ROWS
FROM [US Accidents DW];

```

State	County	City	Wind Direction	Fact US Accidents Count	AccidentsPercentageBy...
MA	Middlesex	Ayer	W	4	0.571428571428571
MA	Suffolk	East Boston	S	5	0.277777777777778
MD	Carroll	Hampstead	S	6	0.157894736842105
MD	Cecil	Rising Sun	N	5	0.102040816326531
MD	Charles	Nanjemoy	W	7	0.777777777777778
MD	Kent	Millington	E	6	0.157894736842105
MI	Menominee	Menominee	N	4	1
MO	Dallas	Buffalo	S	4	0.266666666666667

### b. Thực hiện bằng Pivot Table

The screenshot shows the Microsoft Excel ribbon with the 'Trường PivotTable' (PivotTable Fields) tab selected. The main area displays a PivotTable with data from the 'US Accidents DW' database. The data includes columns for State, County, City, Wind Direction, Fact US Accidents Count, and AccidentsPercentageByWindDirection. The PivotTable is currently set up with 'Nhân Hàng' (Row Labels) and 'Fact US Accidents Count' (Value). The 'Cột' (Column) section is set to sum the values. The 'Bộ lọc' (Filter) section shows 'Cau3' selected. The 'Trường' (Fields) pane on the right lists various dimensions and measures, with 'Cau3' checked as a filter field.

AA	AB	AC	AD
7	2.071428571	2.5	2.468085106
<b>Nhân Hàng</b>	<b>Fact US Accidents Count</b>	<b>AccidentsPercentageByWindDirection</b>	
<b>MA</b>			
Middlesex			
Ayer	4	0.571428571	
W			
Suffolk			
East Boston	5	0.277777778	
S			
<b>MD</b>			
Carroll			
Hampstead	6	0.157894737	
S			
Cecil			
Rising Sun	5	0.102040816	
N			
Charles			
Nanjemoy	7	0.777777778	
W			
Kent			
Millington	6	0.157894737	
E			
<b>MI</b>			
Menominee			
Menominee	4	1	
N			
<b>MO</b>			
Dallas			
Buffalo	4	0.266666667	
S			

**Kết quả:**

Nhân Hàng	Fact US Accidents Count	AccidentsPercentageByWindDirection
<b>MA</b>		
<b>Middlesex</b>		
Ayer		
W	4	0.571428571
<b>Suffolk</b>		
East Boston		
S	5	0.277777778
<b>MD</b>		
<b>Carroll</b>		
Hampstead		
S	6	0.157894737
<b>Cecil</b>		
Rising Sun		
N	5	0.102040816
<b>Charles</b>		
Nanjemoy		
W	7	0.777777778
<b>Kent</b>		
Millington		
E	6	0.157894737
<b>MI</b>		
<b>Menominee</b>		
Menominee		
N	4	1
<b>MO</b>		
<b>Dallas</b>		
Buffalo		
S	4	0.266666667

### c. Thực hiện bằng MDX

```

WITH MEMBER [Measures].[Accidents Percentage by Wind Direction]
AS ([Measures].[AccidentsPercentageByWindDirection]),
FORMAT_STRING = 'Percent'

SELECT {[Measures].[Fact US Accidents Count], [Measures].[Accidents
Percentage by Wind Direction]} ON COLUMNS,
NON EMPTY
FILTER(      [Dim Location].[State].[State] *

```

```

[Dim Location].[County].[County] *
[Dim Location].[City].[City] *
{[Dim Wind].[Wind Direction].[E], [S], [W], [N]},
[Dim Wind].[Wind Speed] > 10 AND
Left([Dim Location].[State].CurrentMember.Name, 1) = "M") ON
ROWS
FROM [US Accidents DW];

```

--Cau3

```

WITH MEMBER [Measures].[Accidents Percentage by Wind Direction]
AS ([Measures].[AccidentsPercentageByWindDirection]),
FORMAT_STRING = 'Percent'
SELECT {[Measures].[Fact US Accidents Count], [Measures].[Accidents Percentage by Wind Dir
NON EMPTY
FILTER( [Dim Location].[State].[State] *
[Dim Location].[County].[County] *
[Dim Location].[City].[City] *
{[Dim Wind].[Wind Direction].[E], [S], [W], [N]},
[Dim Wind].[Wind Speed] > 10 AND
Left([Dim Location].[State].CurrentMember.Name, 1) = "M") ON ROWS
FROM [US Accidents DW];

```

100 %

				Fact US Accidents Count	Accidents Percentage by Wind Direction
MA	Middlesex	Ayer	W	4	57.14%
MA	Suffolk	East Boston	S	5	27.78%
MD	Carroll	Hampstead	S	6	15.79%
MD	Cecil	Rising Sun	N	5	10.20%
MD	Charles	Nanjemoy	W	7	77.78%
MD	Kent	Millington	E	6	15.79%
MI	Menominee	Menominee	N	4	100.00%
MO	Dallas	Buffalo	S	4	26.67%

**Câu 4: Thống kê theo năm, top 5 loại thời tiết có nhiều vụ tai nạn xảy ra nhất trong năm 2020 và năm 2021**

a. Thực hiện manual bằng Visual Studio

Sử dụng Named Set:

```
NONEEMPTY(
```

```

CROSSJOIN([Dim Start Time].[Start Time Year].[Start Time Year].members,
{TOPCOUNT
([Dim Weather].[Weather Condition].[Weather Condition],5,[Measures].[Fact
US Accidents Count]))}

```

**Kết quả:**

2020	Fair	281512
2020	Mostly Cloudy	78429
2020	Cloudy	100804
2020	Partly Cloudy	50223
2021	Fair	715713
2021	Mostly Cloudy	190838
2021	Cloudy	213405
2021	Partly Cloudy	128677

### b. Thực hiện bằng Pivot Excel

Chọn các trường tương ứng bỏ vào Hàng, Cột, Giá trị. Ta được kết quả như sau

The screenshot shows the Power BI PivotTable editor interface. On the left is a table view with data grouped by year (2021 and 2020) and weather conditions. The total row at the bottom is highlighted with a green border. On the right, the 'Trường PivotTable' pane is open, displaying the structure of the table. It shows columns for 'Visibility', 'Visibility ID', 'Dim Weather' (with 'Weather Condition' checked), 'Weather ID', 'Dim Wind', and 'Wind Direction'. Below this, there's a section for dragging fields between rows and columns, with 'Start Time Year' and 'Weather Condition' in the 'Row' section and 'Fact US Accidents Count' in the 'Value' section.

Nhãn Hàng	Fact US Accidents Count
<b>2021</b>	
Fair	715713
Cloudy	213405
Mostly Cloudy	190838
Partly Cloudy	128677
Light Rain	62980
<b>2020</b>	
Fair	281512
Cloudy	100804
Mostly Cloudy	78429
Partly Cloudy	50223
Light Rain	31542
<b>Tổng Cuối</b>	<b>1854123</b>

## Kết quả sau khi thực hiện Pivot Table

Nhãn Hàng	Fact US Accidents Count
<b>2021</b>	
Fair	715713
Cloudy	213405
Mostly Cloudy	190838
Partly Cloudy	128677
Light Rain	62980
<b>2020</b>	
Fair	281512
Cloudy	100804
Mostly Cloudy	78429
Partly Cloudy	50223
Light Rain	31542
<b>Tổng Cuối</b>	<b>1854123</b>

### c. Thực hiện bằng MDX

```
SELECT [Measures].[Fact US Accidents Count] ON 0,  
NONEMPTY(CROSSJOIN({[Dim Start Time].[Start Time Year].[2021], [Dim  
Start Time].[Start Time Year].[2020]}, {TOPCOUNT([Dim Weather].[Weather  
Condition].[Weather Condition],5,[Measures].[Fact US Accidents Count])})) ON 1  
FROM [US Accidents DW];
```

Kết quả sau khi truy vấn MDX giống với Pivot Table :

		Fact US Accidents Count
2021	Fair	715713
2021	Mostly Cloudy	190838
2021	Cloudy	213405
2021	Partly Cloudy	128677
2020	Fair	281512
2020	Mostly Cloudy	78429
2020	Cloudy	100804
2020	Partly Cloudy	50223

**Câu 5:** Thống kê tổng chiều dài của đoạn đường bị ảnh hưởng bởi các vụ tai nạn xảy ra tại nơi vừa có tín hiệu giao thông (Traffic Signal), vừa có trạm xe (Station) theo năm, quý, tiểu bang, quận, thành phố trong năm 2020, 2021.

#### a. Thực hiện Manual

Đầu tiên tạo Named Set giữa cho Cột (Column) và Hàng (Row) tương ứng với Dim Start Time và Dim Location

#### Cau5\_Time

```
{ {[Dim Start Time].[Start Time Year].[2020],  
[Dim Start Time].[Start Time Year].[2021]} * }
```

```
[Dim Start Time].[Start Time Quarter].[Start Time Quarter] * {  
[Measures].[Average Severity]} }
```

### Cau5\_Station

```
{[Dim Location].[State].[State] *  
[Dim Location].[County].[County] *  
[Dim Location].[City].[City]}
```

### Chuyển sang chế độ Script Mode

```
SELECT {[Measures].[Average Severity]} ON COLUMNS,  
NON EMPTY  
    [Cau5_Location] * { {[Dim Start Time].[Start Time Year].[2020], [Dim Start  
Time].[Start Time Year].[2021]} *  
        [Dim Start Time].[Start Time Quarter].[Start Time Quarter] } ON ROWS  
FROM [US Accidents DW]  
WHERE ([Dim Traffic Condition].[Station].[TRUE],  
[Dim Traffic Condition].[Traffic Signal].[TRUE]);
```

Kết quả:

State	County	City	Start Time Year	Start Time Quarter	Average Severity
AL	Jefferson	Birmingham	2020	2	1.5
AL	Jefferson	Birmingham	2021	1	2
AL	Jefferson	Birmingham	2021	2	2
AL	Jefferson	Birmingham	2021	3	2
AL	Madi...	Huntsville	2020	2	2.33333333333...
AL	Madi...	Huntsville	2020	4	2
AL	Madi...	Huntsville	2021	3	2
AL	Madi...	Huntsville	2021	4	2.66666666666...
AL	Tuscaloosa	Tuscaloosa	2021	4	2
AZ	Maricopa	Cave Creek	2021	1	2
AZ	Maricopa	Cave Creek	2021	2	2
AZ	Maricopa	Cave Creek	2021	4	2
AZ	Maricopa	Gilbert	2020	3	2
AZ	Maricopa	Gilbert	2021	1	2
AZ	Maricopa	Gilbert	2020	1	4
AZ	Maricopa	Gilbert	2020	2	1
AZ	Maricopa	Gilbert	2020	3	4
AZ	Maricopa	Gilbert	2020	4	2

## b. Thực hiện bằng Pivot Table

The screenshot shows the PivotTable builder interface with the following details:

- PivotTable Structure:**
  - Rows: Nhóm Hàng (Nhân Hàng), Năm (2020, 2021).
  - Columns: Nhóm cột (Nhân cột), Nhóm Tỉnh (AL, Autauga, Autaugaville, Billingsley, Deatsville, Marbury, Prattville, Selma, Bay Minette, Baldwin, Daphne, Elberta, Fairhope, Foley, Loxley, Magnolia Springs).
  - Data: Average Severity.
- Data Source (Trường PivotTable):**
  - Chọn các trường để thêm vào báo cáo: Tim kiếm.
  - Các tập hợp (Groups):
    - Dim Airport: Airport Code, Airport ID.
    - Dim Location: Hierarchy, Dim Location.
  - Kéo trường giữa các vùng bên dưới:
    - Bộ lọc (Filters): Cau5\_Location.
    - Cột (Columns): Cau5\_Time.
    - Hàng (Rows): Cau5\_Hierarchy.
    - Giá trị (Values): Σ Giá trị.

Kết quả

Câu 5																			
Average Severity	Nhân cột	AL	Autauga	Baldwin	Selma	Bay Minette	Daphne	Elberta	Fairhope	Foley	Loxley	Magnolia Springs	Orange Beach	Perdido	Robertsdale	Seminole	Silverhill	Spanish Fort	Si
Nhân Hàng	Autaugaville	Billingsley	Deatsville	Marbury	Prattville														
2020	1	Average Severity																	
	2	Average Severity																	
	3	Average Severity																	
	4	Average Severity																	
	1	Average Severity																	
	2	Average Severity																	
	3	Average Severity																	
	4	Average Severity																	
2021	1	Average Severity																	
	2	Average Severity																	
	3	Average Severity																	
	4	Average Severity																	
	1	Average Severity																	
	2	Average Severity																	
	3	Average Severity																	
	4	Average Severity																	

### c. Thực hiện bằng MDX

```

SELECT {[Dim Start Time].[Start Time Year].[2020],
          [Dim Start Time].[Start Time Year].[2021]}
        * {[Dim Start Time].[Start Time Quarter].[Start Time Quarter]}
        * {[Measures].[Average Severity]} } ON COLUMNS,
NON EMPTY {
          [Dim Location].[State].[State] *
          [Dim Location].[County].[County] *
          [Dim Location].[City].[City]} ON ROWS
FROM [US Accidents DW]
WHERE ([Dim Traffic Condition].[Station].[TRUE],
[Dim Traffic Condition].[Traffic Signal].[TRUE]);

```

```

SELECT {[Dim Start Time].[Start Time Year].[2020],
         [Dim Start Time].[Start Time Year].[2021]}
       * {[Dim Start Time].[Start Time Quarter].[Start Time Quarter]
           * {[Measures].[Average Severity]} } ON COLUMNS,
NON EMPTY {
    [Dim Location].[State].[State] *
    [Dim Location].[County].[County] *
    [Dim Location].[City].[City]} ON ROWS
FROM [US Accidents DW]
WHERE ([Dim Traffic Condition].[Station].[TRUE],
       [Dim Traffic Condition].[Traffic Signal].[TRUE]);

```

10 % ▾

Messages Results

		2020	2020	2020	2020	2021	2021	Average Severity
		1	2	3	4	1	2	Average Severity
IL	Jefferson	Birmingham	(null)	1.5	(null)	(null)	2	2
IL	Madison	Huntsville	(null)	2.33333333333333	(null)	(null)	(null)	(null)
IL	Tuscaloosa	Tuscaloosa	(null)	(null)	(null)	(null)	(null)	(null)
AZ	Maricopa	Chandler	(null)	(null)	(null)	(null)	2	2
AZ	Maricopa	Gilbert	(null)	(null)	2	(null)	2	(null)
AZ	Maricopa	Glendale	4	1	4	2	2	8
AZ	Maricopa	Mesa	(null)	1	(null)	(null)	2	2
AZ	Maricopa	Paradise Valley	(null)	2	(null)	(null)	(null)	(null)
AZ	Maricopa	Peoria	(null)	2	(null)	(null)	2	(null)
AZ	Maricopa	Phoenix	1.6	1.13333333333333	2	2	2	2.33333333333333
AZ	Maricopa	Scottsdale	4	1	(null)	2.4	2	2
AZ	Maricopa	Sun City	(null)	4	(null)	(null)	(null)	(null)
AZ	Maricopa	Tempe	1	1	(null)	2	2	2
AZ	Maricopa	Tolleson	(null)	(null)	(null)	2	(null)	(null)
AZ	Pima	Tucson	1.3125	1.23809523809524	2	2.11764705882353	2.125	2
AZ	Yavapai	Sedona	(null)	(null)	(null)	(null)	(null)	(null)
CA	Alameda	Albany	(null)	(null)	(null)	(null)	(null)	2
CA	Alameda	Berkeley	2	(null)	(null)	4	2	2
CA	Alameda	Castro Valley	2	2	(null)	2.66666666666667	3	(null)
CA	Alameda	Emeryville	4	2	(null)	3	2	2.28571428571429
CA	Alameda	Fremont	(null)	(null)	(null)	(null)	(null)	2
CA	Alameda	Hayward	2	1.83333333333333	4	2	2.4	2
CA	Alameda	Oakland	2	3.24	2	2.5	2.53333333333333	2.85714285714286
CA	Alameda	San Leandro	1571428571429	?	(null)	2.33333333333333	2.85714285714286	3.5

**Câu 6:** Thống kê theo năm, tháng tổng số vụ tai nạn, chỉ số nghiêm trọng mà bị ảnh hưởng tại những con đường có tai nạn gờ giảm tốc khi nhiệt độ trong ngày trong khoảng -10 F đến 100 F.

Đầu tiên ta tạo ra Named Set [Cau06] với biểu thức như sau:

#### a. Thực hiện manual bằng Visual Studio

#### Named Set:

```

FILTER(NONEMPTY(CROSSJOIN(
[Dim Start Time].[Start Time Year].[Start Time Year],
[Dim Start Time].[Start Time Month].[Start Time Month]* 
[Dim Location].[Street].[Street])),
[Dim Temperature].[Temperature]>-10

```

```
    AND [Dim Temperature].[Temperature]<100
```

```
)
```

^ Expression

```
FILTER(NONEMPTY(CROSSJOIN(
    [Dim Start Time].[Start Time Year].[Start Time Year],
    [Dim Start Time].[Start Time Month].[Start Time Month]* 
    [Dim Location].[Street].[Street])),|
    [Dim Temperature].[Temperature]>-10
    AND [Dim Temperature].[Temperature]<100
```

✓ No issues found

Ln: 4 Ch: 37 Col: 40 TABS CRLF

Chuyển sang Script Mode

```
SELECT {[Measures].[Fact US Accidents Count], [Measures].[Average Severity] } ON
COLUMNS,
NON EMPTY [Cau06] ON ROWS
FROM [US Accidents DW]
WHERE [Dim Traffic Condition].[Bump].&[TRUE];
```

Edit as Text Import... MDX

Design Mode

US Accidents DW

Metadata Functions

Search Model

Measure Group:

<All>

US Accidents DW

- Measures
  - Fact US Accidents
    - Average Distance
    - Average Duration
    - Average Severity
    - Fact US Accidents Count
    - Max Distance
    - Min Distance
    - Sum Distance
- KPIs
- Dim Airport
- Dim Location
- Dim Precipitation
- Dim Start Time
  - Start Time
  - Members
  - Start Time
  - Start Time Date
  - Start Time Date Of Week
  - Start Time Hour
  - Start Time Minute
  - Start Time Month
  - Start Time Quarter

Start Time Year Start Time Month Street Fact US Accidents Count Average Severity

2016	12	I-5 N	1	2
2016	3	Cha...	1	2
2016	4	Tran...	1	2
2016	5	Wo...	1	2
2016	6	E St	1	2
2016	8	I-5 N	2	2
2016	9	I-5 N	1	2
2017	10	I-5 N	3	2
2017	11	S O...	1	4
2017	11	I-5 N	1	2
2017	12	I-5 N	1	2
2017	2	I-5 N	1	2
2017	3	E O...	1	2
2017	3	US-1...	1	2
2017	5	Holl...	1	2

## b. Thực hiện bằng Pivot Table trong Excel

Bump True

Nhân Hàng Fact US Accidents Count Average Severity Câu 6

Câu 8 Câu 9

Nhân Hàng Fact US Accidents Count Start Time Year 2020

Nhân Hàng Fact US Accidents Count

1 CA 61449  
MN 8680  
OR 8755

2 CA 50889  
FL 9228  
NC 9072

3 CA 5782  
FL 8603  
VA 2216

4 CA 70481  
FL 53699  
TX 16717

Visibility All

Nhân Hàng Max Distance Average Severity Câu 10

2016

10 CA 8915 2.710550319  
WA 1120 2.660332542

11 CA 8591 2.683010618  
WA 1133 2.704057279

12 CA 8419 2.704465146  
WA 1408 2.832997988

3 CA 2151 2.629584352

4 CA 6865 2.603337126

Trường PivotTable

Chọn các trường để thêm vào báo cáo:

Tìm kiếm

Average Distance  
Average Duration  
 Average Severity  
 Fact US Accidents Count  
Max Distance  
Min Distance  
Sum Distance

Kéo trường giữa các vùng bên dưới:

Bộ lọc Cột

Bump Σ Giá trị

Hàng Σ Giá trị

Cau06 Fact US Accidents Count  
Average Severity

## Kết quả:

Nhân Hàng	Fact US Accidents Count	Average Severity	Câu hỏi
<b>2016</b>			
<b>12</b>			
I-5 N	1	2	
<b>3</b>			
Charles River Dam Rd	1	2	
<b>4</b>			
Trans-Manhattan Expy W	1	2	
<b>5</b>			
Woodland Ter	1	2	
<b>6</b>			
E St	1	2	
<b>8</b>			
I-5 N	2	2	
<b>9</b>			
I-5 N	1	2	
<b>2017</b>			
<b>10</b>			
I-5 N	3	2	
<b>11</b>			
S Othello St	1	4	
I-5 N	1	2	
<b>12</b>			
I-5 N	1	2	
<b>2</b>			
I-5 N	1	2	
<b>3</b>			
E Oltorf St	1	2	
US-101 N	1	2	
<b>5</b>			
Hollywood Fwy S	1	2	
I-5 N	2	2.5	
US-101 S	1	2	
<b>6</b>			
E Platte Ave	1	2	
<b>8</b>			
I-5 N	1	2	
<b>9</b>			
Highway 87 Bikeway	1	2	

## c. Truy vấn bằng MDX

```

SELECT {[Measures].[Fact US Accidents Count],[Measures].[Average Severity]} ON 0,
FILTER(NONEMPTY(CROSSJOIN(
    [Dim Start Time].[Start Time Year].[Start Time Year],
    [Dim Start Time].[Start Time Month].[Start Time Month]* 
    [Dim Location].[Street].[Street])),
    [Dim Temperature].[Temperature]>-10
    AND [Dim Temperature].[Temperature]<100
) ON 1
FROM [US Accidents DW]
WHERE [Dim Traffic Condition].[Bump].&[TRUE];

```

100 %

		Messages	Results	
		Fact US Accidents Count	Average Severity	
2016	12	I-5 N	1	2
2016	3	Charles River Dam Rd	1	2
2016	4	Trans-Manhattan Expy W	1	2
2016	5	Woodland Ter	1	2
2016	6	E St	1	2
2016	8	I-5 N	2	2
2016	9	I-5 N	1	2
2017	10	I-5 N	3	2
2017	11	S Othello St	1	4
2017	11	I-5 N	1	2
2017	12	I-5 N	1	2
2017	2	I-5 N	1	2
2017	3	E Olton St	1	2
2017	3	US-101 N	1	2
2017	5	Hollywood Fwy S	1	2
2017	5	I-5 N	2	2.5
2017	5	US-101 S	1	2
2017	6	E Platte Ave	1	2
2017	8	I-5 N	1	2
2017	9	Highway 87 Bikeway	1	2
2017	9	Hutchinson River Pkwy N	1	4
2017	9	I-5 N	1	2
2018	1	Guadalupe Fwy N	1	2
2018	10	Devonshire St	1	2
2018	10	W 11th Ave	1	2
2018	10	SW Kelly Ave	1	2
2018	11	E Redfield Rd	1	4
2018	11	NW Wilson River Hwy	1	2

**Câu 7:** Liệt kê tổng số vụ tai nạn, mức độ nghiêm trọng trung bình, và thời gian trung bình của các vụ tai nạn giao thông theo quý trong năm.

- Thực hiện bằng Manual

Start Time Year	Start Time Quarter	Fact US Accidents Count	Average Severity
2016	1	2886	2.95355528154...
2016	2	22131	2.80185574796...
2016	3	43879	2.89342981662...
2016	4	50214	2.99665891204...
2017	1	46541	2.95529905561...
2017	2	41650	2.93707761769...
2017	3	36221	3.05159829695...
2017	4	35016	3.13310082209...
2018	1	38592	3.09113173498...
2018	2	30778	3.08134216544...
2018	3	36448	2.87621508154...
2018	4	52373	2.98915000692...
2019	1	47856	2.93849348594...
2019	2	41334	2.99820378364...
2019	3	63109	2.83679683638...
2019	4	97337	2.95821043675...
2020	1	115562	2.80404134273...
2020	2	168236	2.59862921745...
2020	3	35611	3.23657525791...
2020	4	283177	3.55147581078...
2021	1	283090	3.50353176149...
2021	2	259028	3.341777129
2021	3	345046	3.486645009
2021	4	580956	4.39076281287...

## b. Thực hiện bằng Pivot Table

The screenshot shows the Microsoft Power BI PivotTable editor interface. On the left, there is a large preview grid of the data. On the right, there are several panes:

- Chọn các trường để thêm vào báo cáo:** A list of available fields: Fact US Accidents, Average Distance, Average Duration, Average Severity, Fact US Accidents Count, Max Distance, and Min Distance.
- Tìm kiếm:** A search bar.
- Kết quả:** A summary table showing counts for each state (MA, MD, ME, MI, MN, MO, MS, MT, NC, ND, NE, NJ, NM, NV, NY, OH, OK, OR, PA, RI, SC, TN, TX, UT, VA, WA, WI, WV) and the total count for 2020 (29).
- Bộ lọc:** Filter settings for Start Time Year and Start Time Quarter.
- Cột:** Column settings for the PivotTable.
- Hàng:** Row settings for the PivotTable.

**Kết quả:**

Nhận Hàng	Fact US Accidents Count	Average Severity	Average Distance
<b>2021</b>			
1	283090	3.503531761	3.503531761
2	259028	3.341777129	3.341777129
3	345046	3.486645009	3.486645009
4	580956	4.390762813	4.390762813
<b>2020</b>			
1	115562	2.804041343	2.804041343
2	168236	2.598629217	2.598629217
3	35611	3.236575258	3.236575258
4	283177	3.551475811	3.551475811
<b>2019</b>			
1	47856	2.938493486	2.938493486
2	41334	2.998203784	2.998203784
3	63109	2.836796836	2.836796836
4	97337	2.958210437	2.958210437
<b>2018</b>			
1	38592	3.091131735	3.091131735
2	30778	3.081342165	3.081342165
3	36448	2.876215082	2.876215082
4	52373	2.989150007	2.989150007
<b>2017</b>			
1	46541	2.955299056	2.955299056
2	41650	2.937077618	2.937077618
3	36221	3.051598297	3.051598297
4	35016	3.133100822	3.133100822
<b>2016</b>			
1	2886	2.953555282	2.953555282
2	22131	2.801855748	2.801855748
3	43879	2.893429817	2.893429817
4	50214	2.996658912	2.996658912
<b>Tổng Cuối</b>	<b>2757071</b>	<b>3.345702772</b>	<b>3.345702772</b>
		Câu 7	

c. Thực hiện bằng MDX

```

SELECT {[Measures].[Fact US Accidents Count],
          [Measures].[Average Severity],
          [Measures].[Average Distance],
          [Measures].[Average Duration]} ON COLUMNS,
NON EMPTY CROSSJOIN([Dim Start Time].[Start Time Year].MEMBERS, [Dim Start
Time].[Start Time Quarter].MEMBERS) ON ROWS
FROM [US Accidents DW];

```

```

SELECT {[Measures].[Fact US Accidents Count],
         [Measures].[Average Severity],
         [Measures].[Average Distance],
         [Measures].[Average Duration]} ON COLUMNS,
NON EMPTY CROSSJOIN([Dim Start Time].[Start Time Year].MEMBERS, [Dim Start Time].[Start Time Quarter].MEMBERS) ON ROWS
FROM [US Accidents DW];

```

100 %

	Fact US Accidents Count	Average Severity	Average Distance	Average Duration
2016 3	43879	2.89342981662692	2.89342981662692	0
2016 4	50214	2.99665891204277	2.99665891204277	0
2017 All	159428	3.01052945951327	3.01052945951327	0
2017 1	46541	2.95529905561385	2.95529905561385	0
2017 2	41650	2.93707761769189	2.93707761769189	0
2017 3	36221	3.05159829695208	3.05159829695208	0
2017 4	35016	3.13310082209331	3.13310082209331	0
2018 All	158191	3.00449964416079	3.00449964416079	0
2018 1	38592	3.09113173498324	3.09113173498324	0
2018 2	30778	3.08134216544437	3.08134216544437	0
2018 3	36448	2.87621508154368	2.87621508154368	0
2018 4	52373	2.98915000692553	2.98915000692553	0
2019 All	249636	2.92986438032673	2.92986438032673	0
2019 1	47856	2.93849348594596	2.93849348594596	0
2019 2	41334	2.99820378364284	2.99820378364284	0
2019 3	63109	2.83679683638161	2.83679683638161	0
2019 4	97337	2.95821043675553	2.95821043675553	0
2020 All	602586	3.04962158144204	3.04962158144204	0
2020 1	115562	2.80404134273216	2.80404134273216	0
2020 2	168236	2.59862921745275	2.59862921745275	0
2020 3	35611	3.23657525791376	3.23657525791376	0
2020 4	283177	3.55147581078494	3.55147581078494	0
2021 All	1468120	3.76801177682165	3.76801177682165	0
2021 1	283090	3.50353176149649	3.50353176149649	0
2021 2	259028	3.34177712944758	3.34177712944758	0
2021 3	345046	3.48664500947709	3.48664500947709	0
2021 4	580956	4.39076281287247	4.39076281287247	0

**Câu 8:** Cho biết tên con đường có số vụ tai nạn giao thông cao nhất và tên con đường có chỉ số nghiêm trọng tai nạn giao thông cao nhất.

a. Thực hiện bằng tay Manual bằng Script Mode

Đầu tiên tạo Named Set [Cau8]

```

UNION(
    TopCount({[Dim Location].[Street].Children},1,
              [Measures].[Fact US Accidents Count]),
    TopCount({[Dim Location].[Street].Children},1,
              [Measures].[Average Severity]))

```

Sau đó thực hiện Script Mode

```
SELECT {[Measures].[Fact US Accidents Count]} ON COLUMNS,  
NON EMPTY [Cau8]  
ON ROWS  
FROM [US Accidents DW];
```

Kết quả:

```
SELECT {[Measures].[Fact US Accidents Count]} ON COLUMNS,  
NON EMPTY [Cau8]  
ON ROWS  
FROM [US Accidents DW]; |
```

Street	Fact US Accidents Count
I-95 S	5111
N East Ave	12

## b. Thực hiện bằng Pivot Table

**Câu 8**

Nhãn Hàng	Fact US Accidents Count
N East Ave	12
I-95 S	5111

**Trường PivotTable**

Chọn các trường để thêm vào báo cáo:

Tìm kiếm

↓  $\sum$  Fact US Accidents

- Average Distance
- Average Duration
- Average Severity
- Fact US Accidents Count
- Max Distance
- Min Distance

Kéo trường giữa các vùng bên dưới:

<b>T</b> Bộ lọc	<b>C</b> Cột
<b>Hàng</b>	<b>Giá trị</b>
Cau8	Fact US Accidents Count

Kết quả:

Nhãn Hàng	Fact US Accidents Count
N East Ave	12
I-95 S	5111

### c. Thực hiện bằng MDX

```

SELECT {[Measures].[Fact US Accidents Count]} ON COLUMNS,
NON EMPTY
UNION(
    TopCount({[Dim Location].[Street].Children},1,
        [Measures].[Fact US Accidents Count]),
    TopCount({[Dim Location].[Street].Children},1,
        [Measures].[Average Severity]))
ON ROWS
FROM [US Accidents DW];

```

The screenshot shows a DAX editor interface. The top pane displays the DAX query:

```

SELECT {[Measures].[Fact US Accidents Count]} ON COLUMNS,
NON EMPTY
UNION(
    TopCount({[Dim Location].[Street].Children},1,
        [Measures].[Fact US Accidents Count]),
    TopCount({[Dim Location].[Street].Children},1,
        [Measures].[Average Severity]))
ON ROWS
FROM [US Accidents DW];

```

The bottom pane shows the results of the query, which is a table titled "Fact US Accidents Count". The table contains two rows:

	Fact US Accidents Count
I-95 S	5111
N East Ave	12

**Câu 9:** Lấy ra 3 tiêu bang xảy ra nhiều vụ tai nạn nhất với mỗi quý trong năm 2020.

- Thực hiện Manual

Đầu tiên, tạo Named Set như sau:

```
{Generate(
    [Dim Start Time].[Start Time Quarter].Children,
    TopCount(
        {[Dim Start Time].[Start Time Quarter].CurrentMember*
        [Dim Location].[State].Children}, 3,
        [Measures].[Fact US Accidents Count]))}
```

Sau đó mở Script Mode:

```
SELECT {[Measures].[Fact US Accidents Count]} ON COLUMNS,
NON EMPTY [Cau9] ON ROWS
FROM [US Accidents DW]
WHERE [Dim Start Time].[Start Time Year].[2020];
```

Start Time Quarter	State	Fact US Accidents Count
1	CA	61449
1	OR	8755
1	MN	8680
2	CA	50889
2	FL	9228
2	NC	9072
3	FL	8603
3	CA	5782
3	VA	2216
4	CA	70481
4	FL	53699
4	TX	16717

b. Thực hiện bằng Pivot Table Excel

**Câu 9**

Start Time Year: 2020

Nhân Hàng	Fact US Accidents Count
1	CA 61449 MN 8680 OR 8755
2	CA 50889 FL 9228 NC 9072
3	CA 5782 FL 8603 VA 2216
4	CA 70481 FL 53699 TX 16717

**Trường PivotTable**

Chọn các trường để thêm vào báo cáo:

Tìm kiếm

- Start Time Minute
- Start Time Month
- Start Time Quarter
- Start Time Year

Dim Temperature

- Temperature
- Temperature ID

Kéo trường giữa các vùng bên dưới:

<b>Bộ lọc</b> Start Time Year	<b>Cột</b>
<b>Hàng</b> Cau9	<b>Giá trị</b> Fact US Accidents Count

**Kết quả:**

Start Time Year 2020	
Nhãn Hàng	Fact US Accidents Count
<b>1</b>	
CA	61449
MN	8680
OR	8755
<b>2</b>	
CA	50889
FL	9228
NC	9072
<b>3</b>	
CA	5782
FL	8603
VA	2216
<b>4</b>	
CA	70481
FL	53699
TX	16717

c. Thực hiện bằng MDX

```

SELECT {[Measures].[Fact US Accidents Count]} ON COLUMNS,
NON EMPTY
{Generate(
    [Dim Start Time].[Start Time Quarter].Children,
    TopCount(
        {[Dim Start Time].[Start Time Quarter].CurrentMember*}
        [Dim Location].[State].Children}, 3,
        [Measures].[Fact US Accidents Count]))} ON ROWS
FROM [US Accidents DW]
WHERE [Dim Start Time].[Start Time Year].[2020];

```

```

SELECT {[Measures].[Fact US Accidents Count]} ON COLUMNS,
NON EMPTY
{Generate(
    [Dim Start Time].[Start Time Quarter].Children,
    TopCount(
        {[Dim Start Time].[Start Time Quarter].CurrentMember*}
        [Dim Location].[State].Children}, 3,
        [Measures].[Fact US Accidents Count]))} ON ROWS
FROM [US Accidents DW]
WHERE [Dim Start Time].[Start Time Year].[2020];

```

100 %

	Fact US Accidents Count
1 CA	61449
1 OR	8755
1 MN	8680
2 CA	50889
2 FL	9228
2 NC	9072
3 FL	8603
3 CA	5782
3 VA	2216
4 CA	70481
4 FL	53699
4 TX	16717

### Kết quả sau khi truy vấn MDX giống với Pivot Table

**Câu 10:** Liệt kê chiều dài lớn nhất của đoạn đường và mức độ nghiêm trọng bị ảnh hưởng bởi các vụ tai nạn giao thông có tầm nhìn xa trên 2 dặm tại tiểu bang California (CA) và Washington (WA)

#### a. Thực hiện bằng tay

Tạo Named Set:

```

FILTER(
CROSSJOIN([Dim Start Time].[Start Time Year].[Start Time Year] *
    [Dim Start Time].[Start Time Month].[Start Time Month],
    {[Dim Location].[State].[CA], [Dim Location].[State].[WA]}),
[Dim Visibility].[Visibility] > 2)

```

Chọn chế độ Script Mode:

```
SELECT {[Measures].[Max Distance], [Measures].[Average Severity]} ON COLUMNS,  
NON EMPTY [Cau10] ON ROWS  
FROM [US Accidents DW];
```

Kết quả:

```
SELECT {[Measures].[Max Distance], [Measures].[Average Severity]} ON COLUMNS,  
NON EMPTY [Cau10] ON ROWS  
FROM [US Accidents DW];
```

Start Time Year	Start Time Month	State	Max Distance	Average Severity
2016	10	CA	8915	2.71055031924...
2016	10	WA	1120	2.66033254156...
2016	11	CA	8591	2.68301061836...
2016	11	WA	1133	2.70405727923...
2016	12	CA	8419	2.70446514616...
2016	12	WA	1408	2.83299798792...
2016	3	CA	2151	2.62958435207...
2016	4	CA	6865	2.60333712552...
2016	5	CA	9881	2.72579310344...
2016	6	CA	7145	2.53278979085...
2016	6	WA	363	2.55633802816...
2016	7	CA	5387	2.54343720491...
2016	7	WA	827	2.65916398713...
2016	8	CA	7051	2.55841799709...
2016	8	WA	1159	2.56984478935...
...	...	...	...	...

```

SELECT {[Measures].[Max Distance], [Measures].[Average Severity]} ON COLUMNS,
NON EMPTY [Cau10] ON ROWS
FROM [US Accidents DW];

```

Start Time Year	Start Time Month	State	Max Distance	Average Severity
2016	10	CA	8915	2.71055031924597
2016	10	WA	1120	2.6603325415677
2016	11	CA	8591	2.68301061836352
2016	11	WA	1133	2.70405727923628
2016	12	CA	8419	2.70446514616126
2016	12	WA	1408	2.83299798792757
2016	3	CA	2151	2.62958435207824
2016	4	CA	6865	2.60333712552143
2016	5	CA	9881	2.72579310344828
2016	6	CA	7145	2.53278979085431
2016	6	WA	363	2.55633802816901
2016	7	CA	5387	2.54343720491029
2016	7	WA	827	2.65916398713826
2016	8	CA	7051	2.55841799709724
2016	8	WA	1159	2.56984478935698
2016	9	CA	7858	2.61846051316228
2016	9	WA	753	2.46885245901639
2017	1	CA	7717	2.72204585537919
2017	1	WA	773	2.62925170068027
2017	10	CA	4029	2.59266409266409

### b. Thực hiện bằng Pivot Table

Kết quả:

Trường PivotTable

Chọn các trường để thêm vào báo cáo:

Tìm kiếm

Fact US Accidents Count

Max Distance

Min Distance

Sum Distance

Các tập hợp

Cau10

> Dim Airport

Kéo trường giữa các vùng bên dưới:

Bộ lọc	Cột
Visibility	$\sum$ Giá trị
Hàng	$\sum$ Giá trị
Cau10	Max Distance
	Average Severity

Visibility All

Nhân Hàng Max Distance Average Severity Câu 10

2016

10 CA 8915 2.710550319  
WA 1120 2.660332542

11 CA 8591 2.683010618  
WA 1133 2.704057279

12 CA 8419 2.704465146  
WA 1408 2.832997988

3 CA 2151 2.629584352

4 CA 6865 2.603337126

5 CA 9881 2.725793103

6 CA 7145 2.532789791  
WA 363 2.556338028

7 CA 5387 2.543437205  
WA 827 2.659163987

8 CA 7051 2.558417997  
WA 1159 2.569844789

9 CA 7858 2.618460513  
WA 753 2.468852459

2017

1 CA 7717 2.722045855  
WA 773 2.629251701

10 CA 4029 2.592664093  
WA 1051 2.863760218

11 CA 5658 2.623087622

Visibility All

Nhân Hàng Max Distance Average Severity Câu 10

2016

10 CA 8915 2.710550319  
WA 1120 2.660332542

11 CA 8591 2.683010618  
WA 1133 2.704057279

12 CA 8419 2.704465146  
WA 1408 2.832997988

3 CA 2151 2.629584352

4 CA 6865 2.603337126

5 CA 9881 2.725793103

6 CA 7145 2.532789791  
WA 363 2.556338028

7 CA 5387 2.543437205  
WA 827 2.659163987

8 CA 7051 2.558417997  
WA 1159 2.569844789

9 CA 7858 2.618460513  
WA 753 2.468852459

2017

### c. Thực hiện bằng MDX

```
SELECT {[Measures].[Max Distance],  
        [Measures].[Average Severity]} ON COLUMNS,  
NON EMPTY  
FILTER(  
CROSSJOIN([Dim Start Time].[Start Time Year].[Start Time Year] *  
          [Dim Start Time].[Start Time Month].[Start Time Month],  
          {[Dim Location].[State].[CA], [Dim Location].[State].[WA]}),  
[Dim Visibility].[Visibility] > 2) ON ROWS  
FROM [US Accidents DW];
```

The screenshot shows a SQL Server Management Studio (SSMS) interface. In the top pane, there is a code editor containing an MDX query. The query selects measures for Max Distance and Average Severity across non-empty rows filtered by start time year and month, and visibility levels greater than 2, from the US Accidents DW database. In the bottom pane, there is a results grid displaying the output of the query. The results grid has columns for Year, Month, State, Max Distance, and Average Severity. The data spans from 2016 to 2017, covering all months and both CA and WA states.

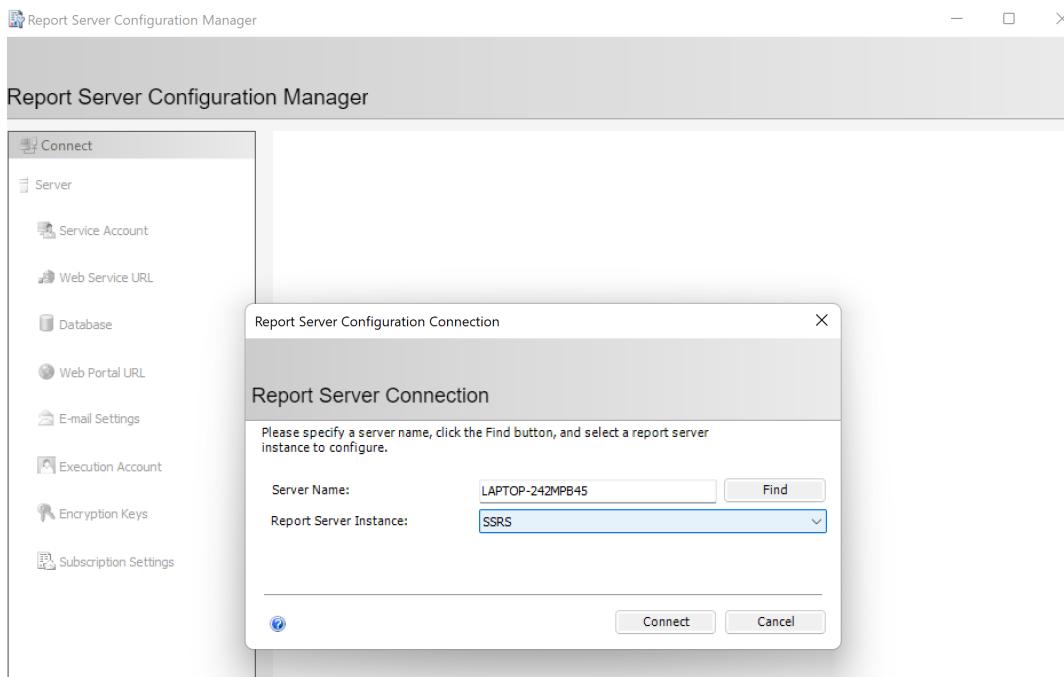
			Max Distance	Average Severity
2016	10	CA	8915	2.71055031924597
2016	10	WA	1120	2.6603325415677
2016	11	CA	8591	2.68301061836352
2016	11	WA	1133	2.70405727923628
2016	12	CA	8419	2.70446514616126
2016	12	WA	1408	2.83299798792757
2016	3	CA	2151	2.62958435207824
2016	4	CA	6865	2.60333712552143
2016	5	CA	9881	2.72579310344828
2016	6	CA	7145	2.53278979085431
2016	6	WA	363	2.55633802816901
2016	7	CA	5387	2.54343720491029
2016	7	WA	827	2.65916398713826
2016	8	CA	7051	2.55841799709724
2016	8	WA	1159	2.56984478935698
2016	9	CA	7858	2.61846051316228
2016	9	WA	753	2.46885245901639
2017	1	CA	7717	2.72204585537919
2017	1	WA	773	2.62925170068027
2017	10	CA	4029	2.59266409266409
2017	10	WA	1051	2.86376021798365
2017	11	CA	5658	2.6230876216968
2017	11	WA	1026	2.74331550802139
2017	12	CA	5569	2.70339805825243
2017	12	WA	966	2.99071207430341
2017	2	CA	8272	2.63188036907413
2017	2	WA	1163	2.77565632458234
2017	3	CA	7741	2.5898293743727

## **CHƯƠNG 4: QUÁ TRÌNH LẬP BÁO BIỂU – SSRS (THAM KHẢO LÀM THÊM)**

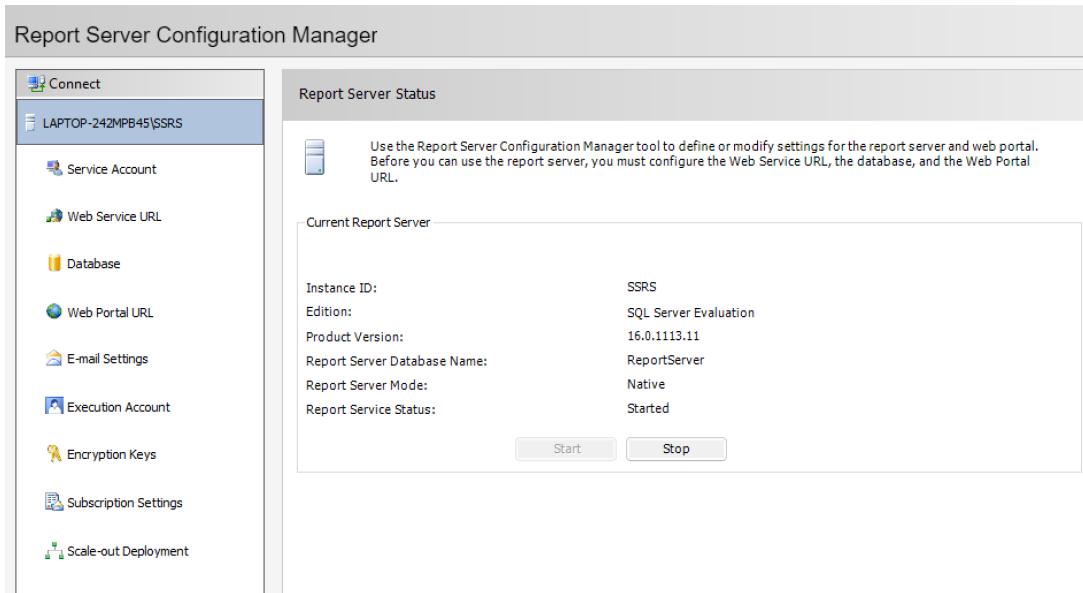
### **4.1 Tạo report với Visual Studio**

#### **4.1.1 Cấu hình Report Server Configuration Manager**

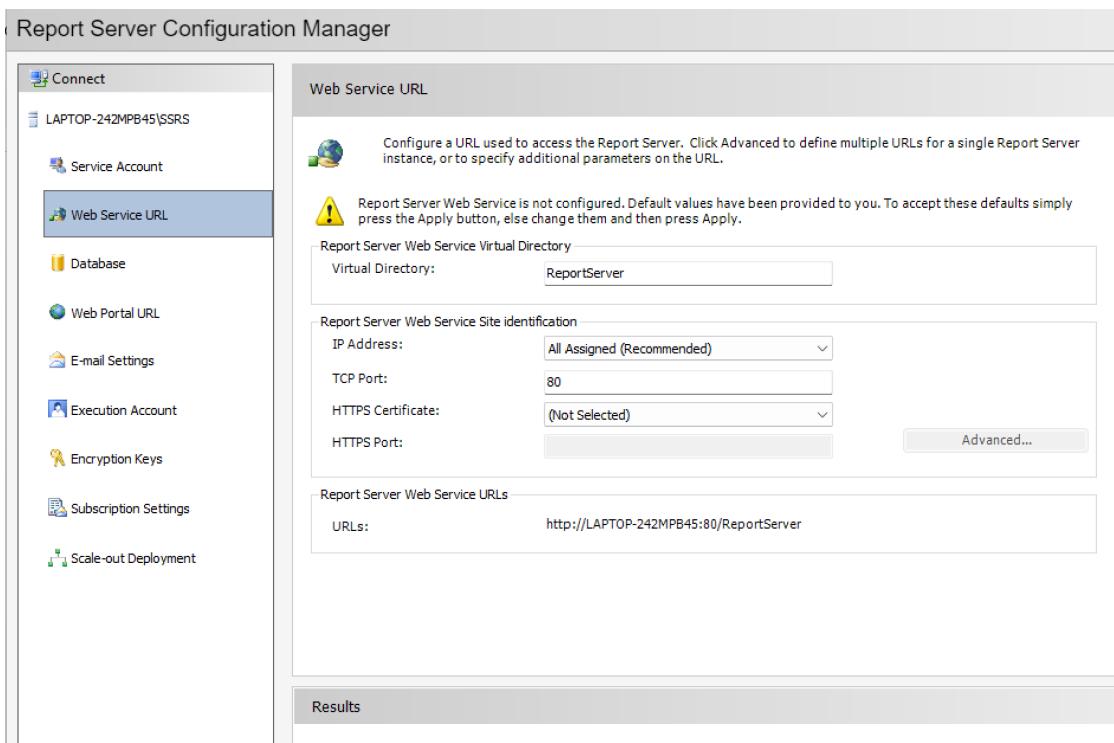
**Bước 1:** Mở Report Server Configuration Manager lên, hộp thoại mở ra như hình dưới, nhấn Connect



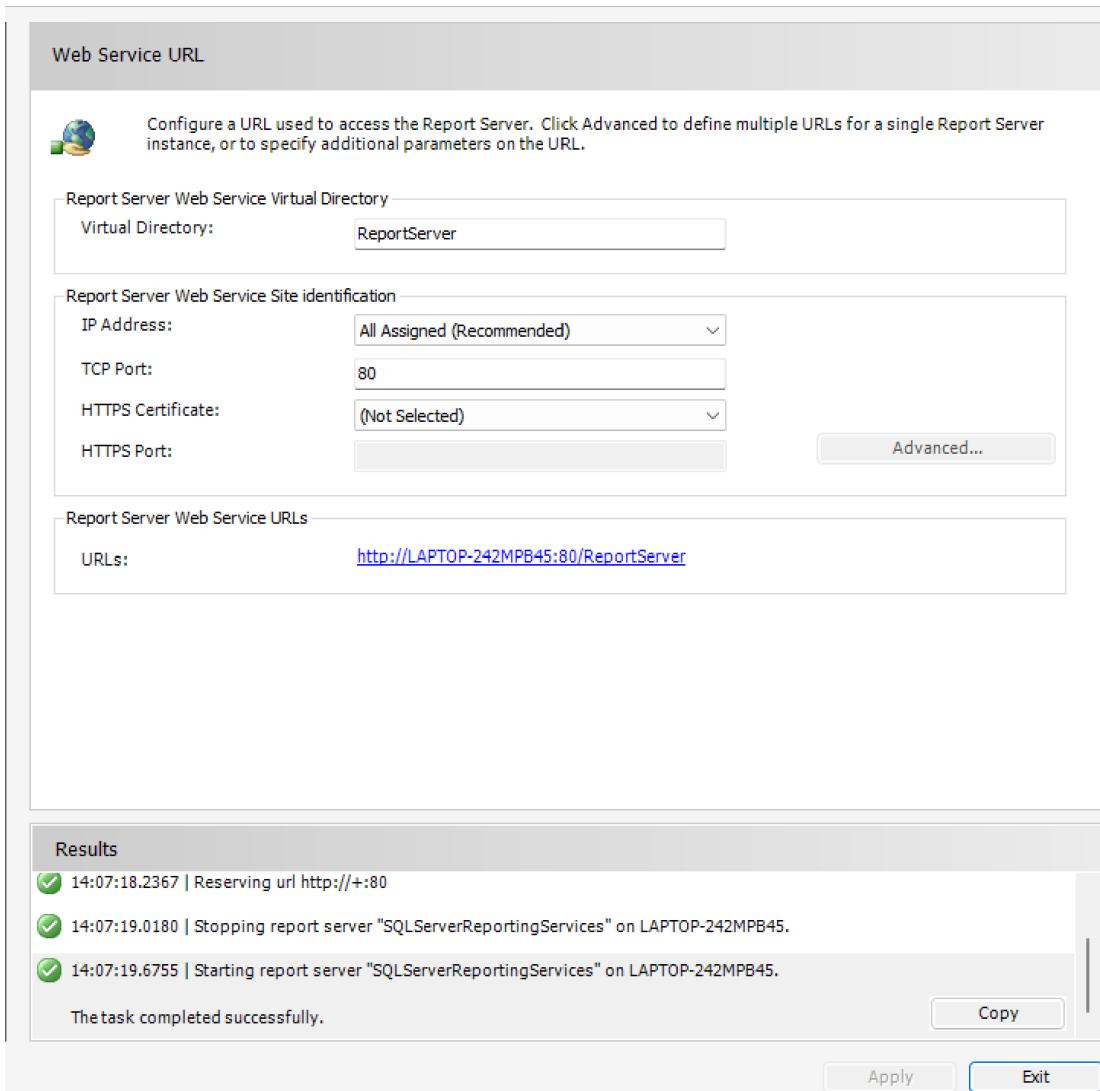
Cửa sổ công cụ Report Server Configuration Manager



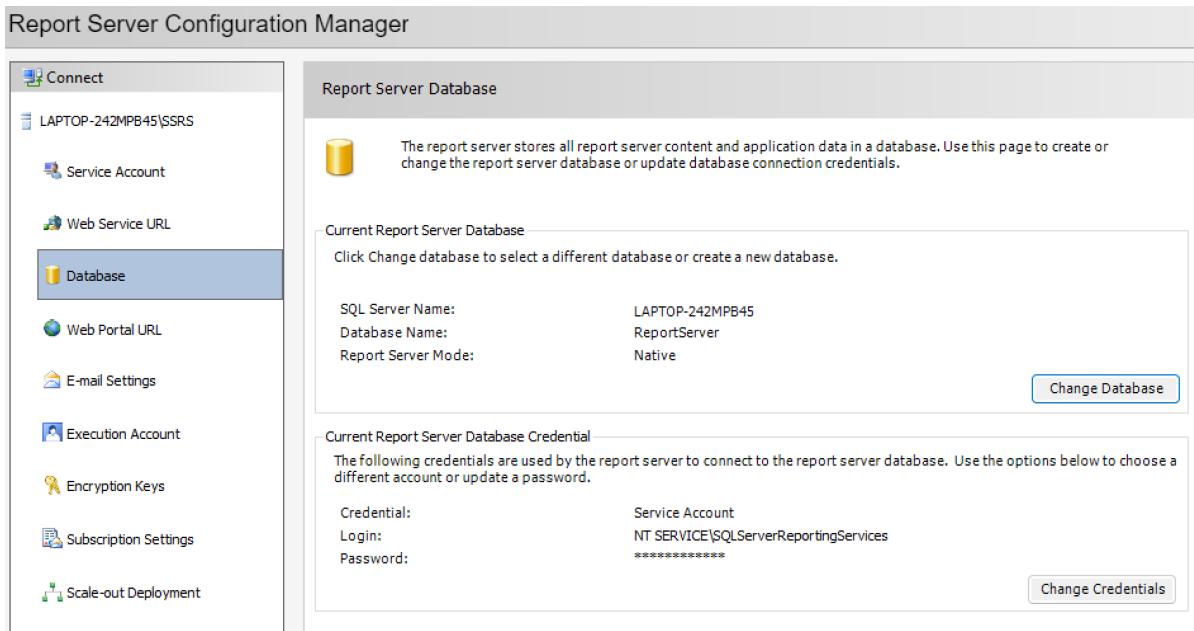
## Bước 2: Kiểm tra Web Service URL -> Chọn Apply để tạo URLs



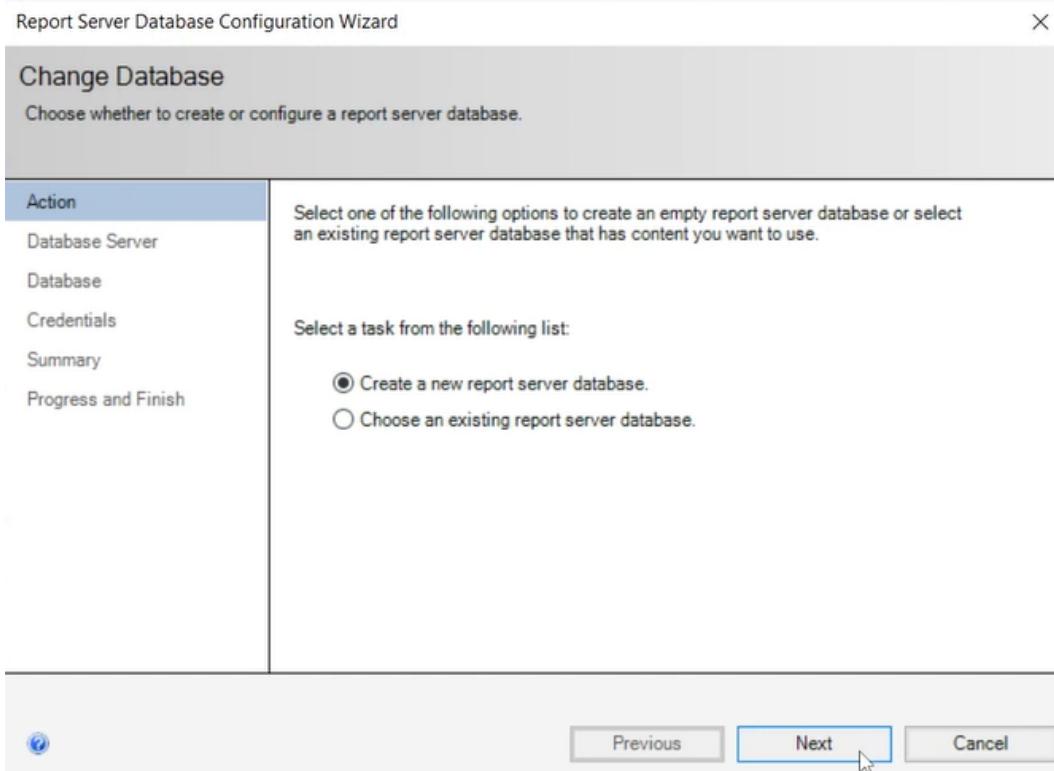
## Bước 3: Sau khi tạo thành công, sẽ hiện lên thông báo như sau



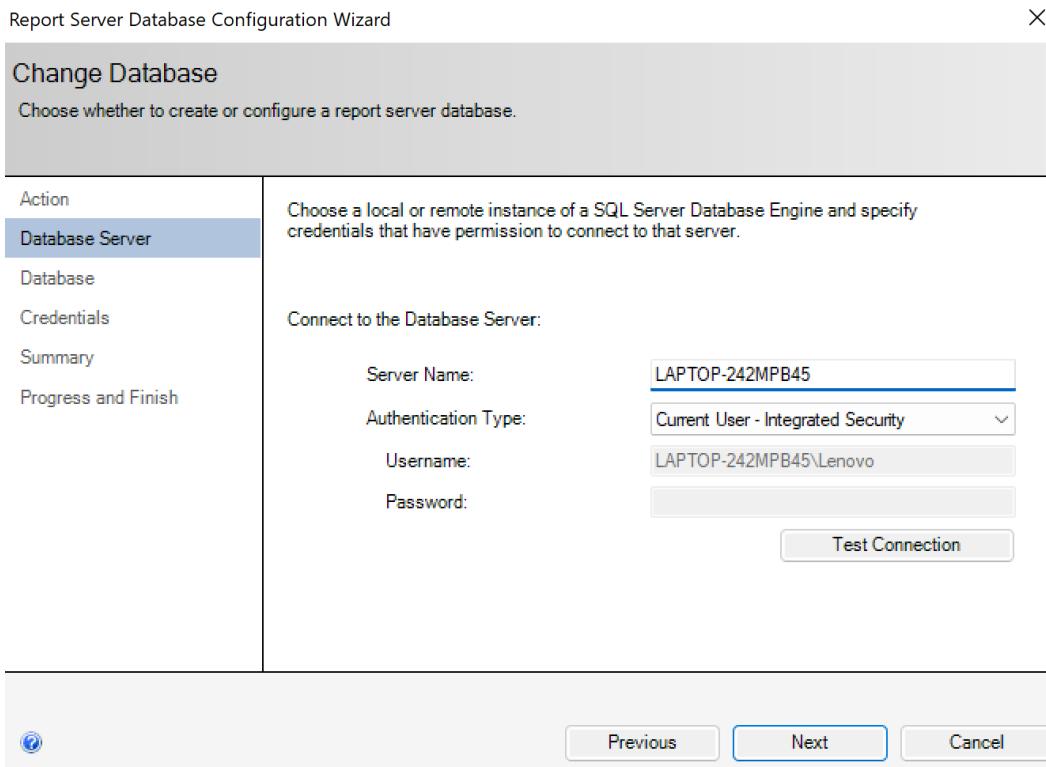
**Bước 4:** Mở sang tab Database, nhấn vào Change Database



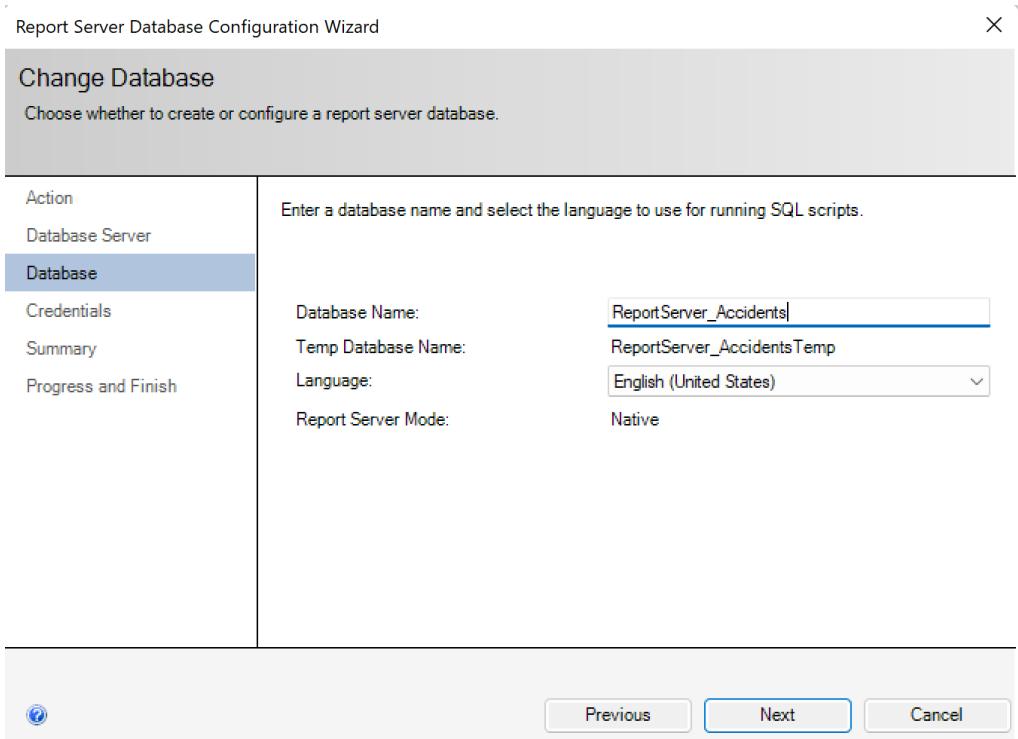
**Bước 5:** Hộp Report Server Database Configuration Wizard hiện ra, tick chọn Create a new report server database



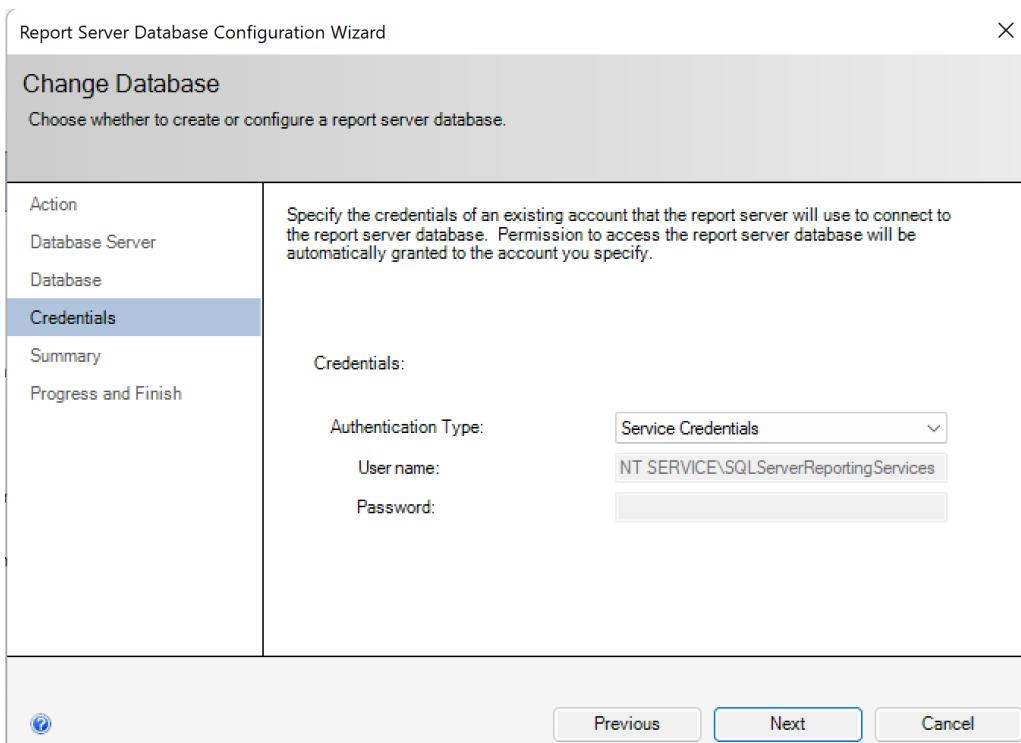
**Bước 6:** Nhấn Next



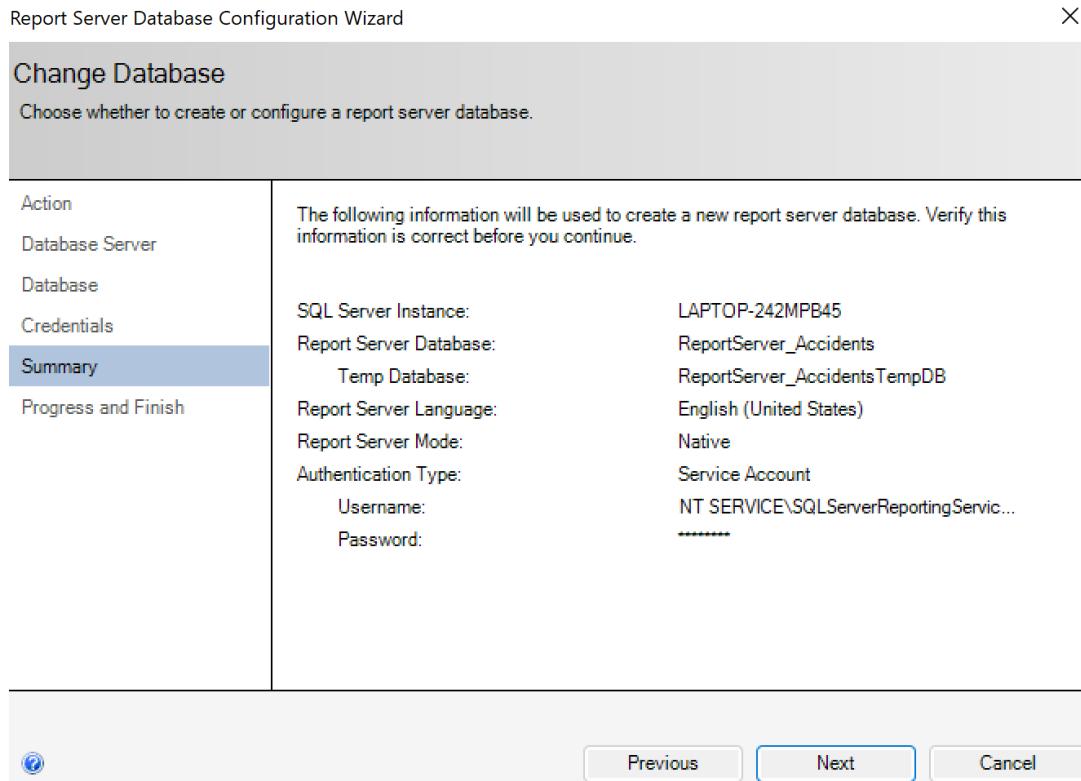
### Bước 7: Đổi tên database. Nhấn Next



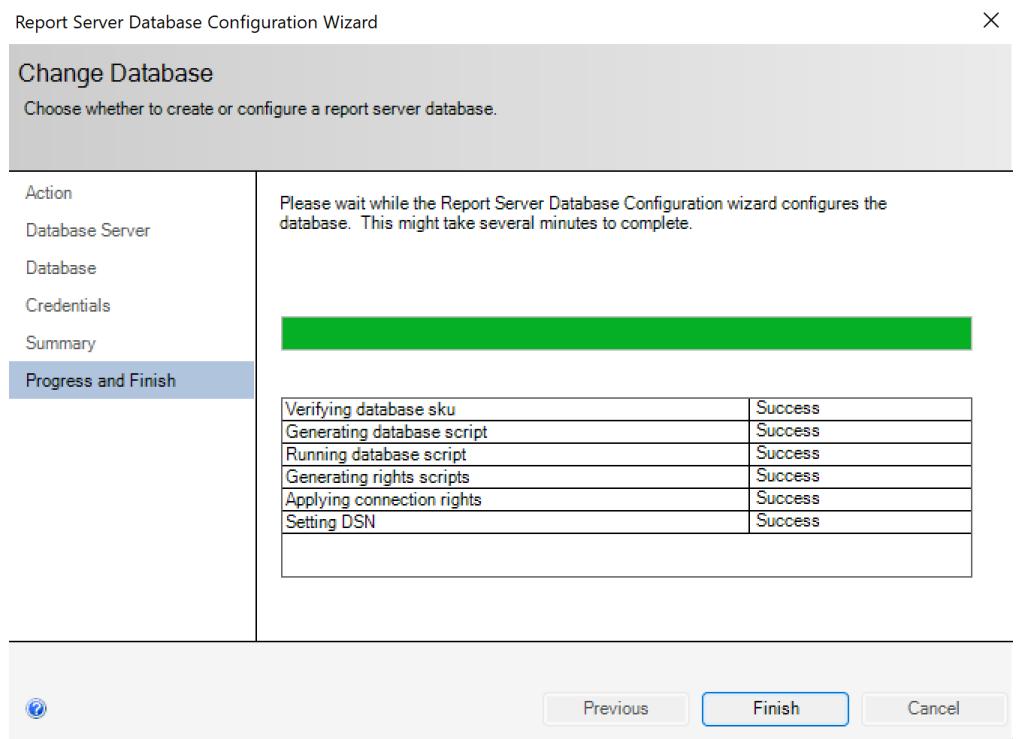
## Bước 8: Nhấn Next



## Bước 9: Nhấn Next

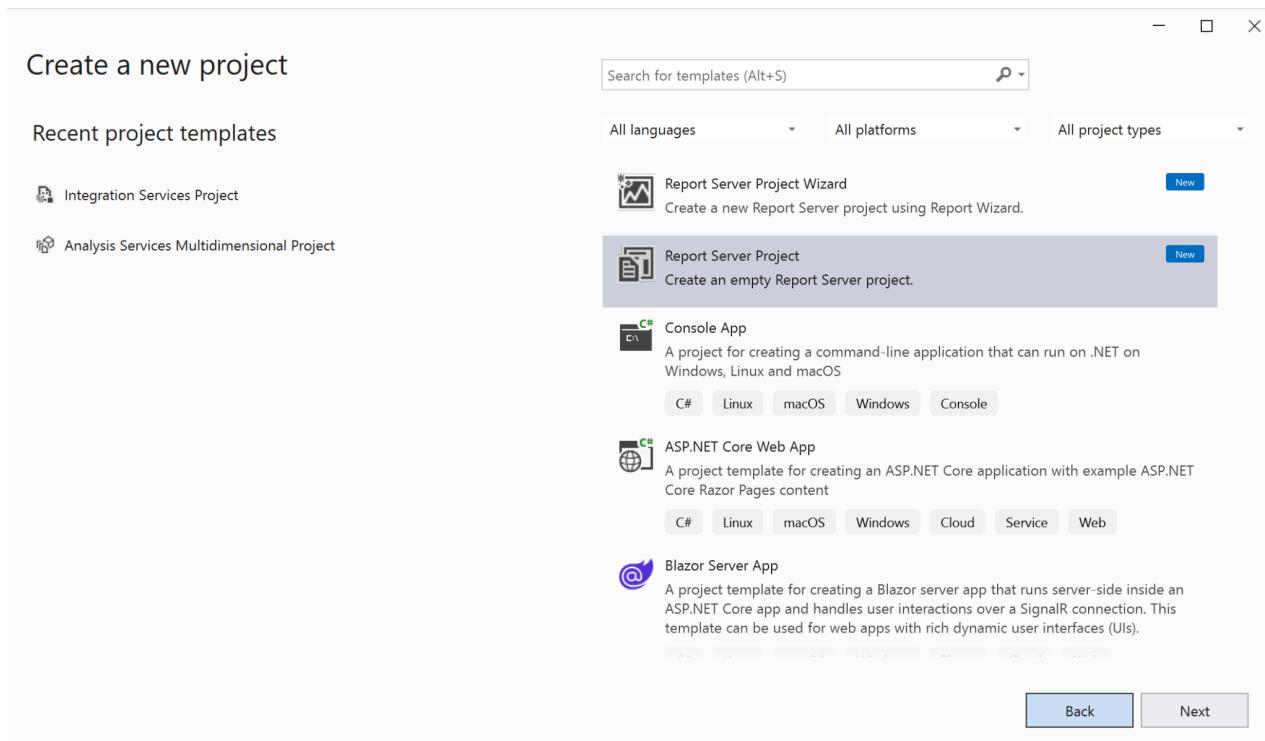


## Bước 10: Quá trình tạo database hoàn tất. Nhấn Finish



## 4.1.2 Tạo Report Service Project

### Bước 1: Tạo Project mới với Report Server Project



### Bước 2: Điền tên, vị trí lưu trữ project. Nhấn Create

## Configure your new project

### Report Server Project

Project name

Location

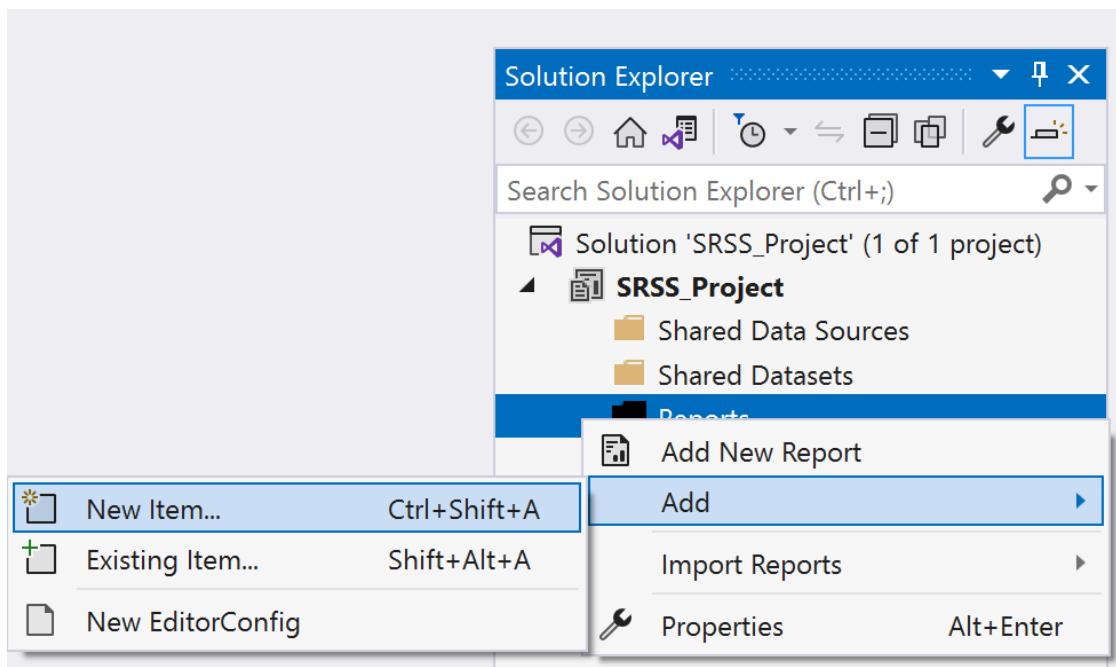
 

Solution name (i)

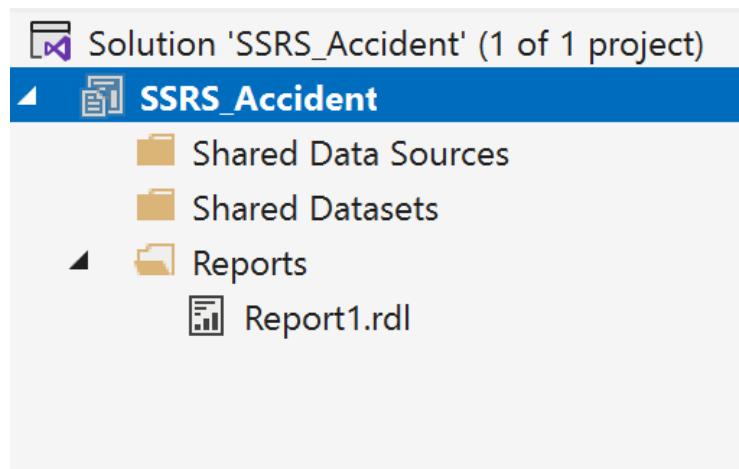
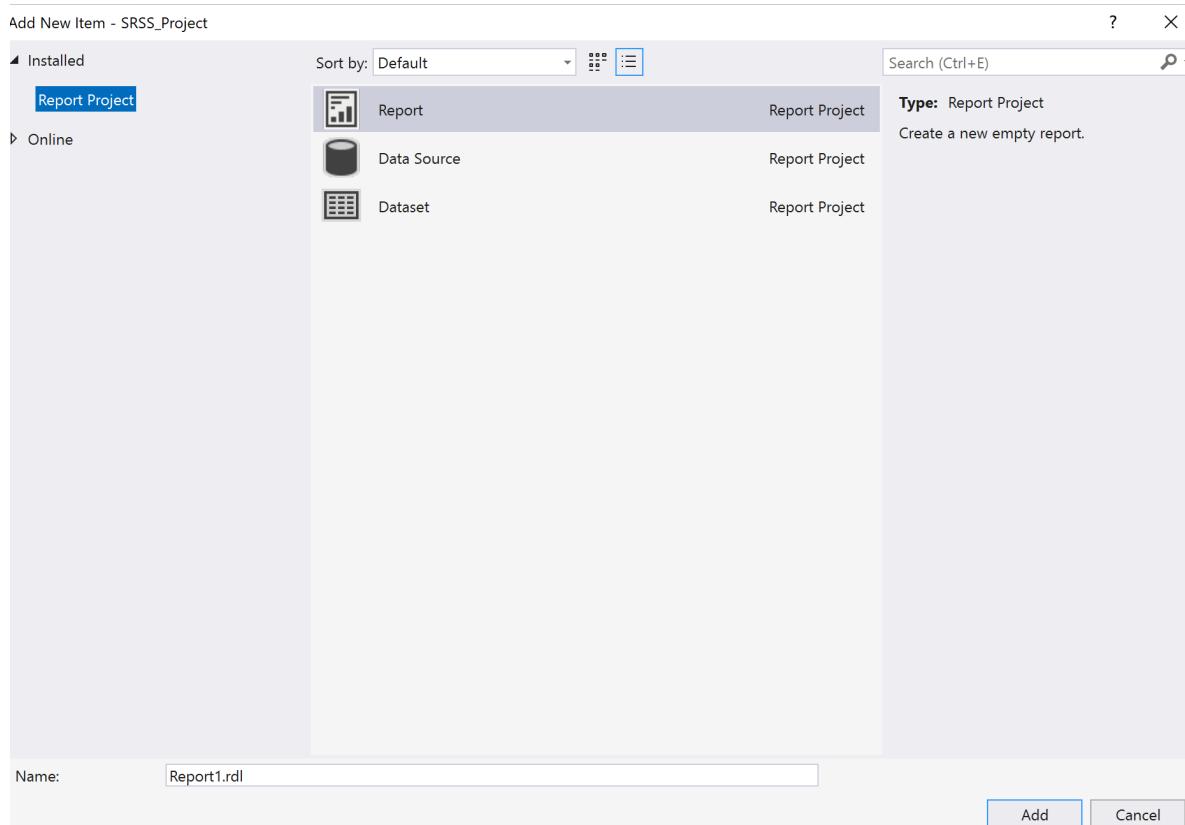
Place solution and project in the same directory

Project will be created in "C:\Users\Lenovo\source\repos\SRSS\_Project\SRSS\_Project\"

**Bước 3:** Tại tab Solution Explorer, nhấp chuột phải tại Reports, chọn Add, chọn New Item..

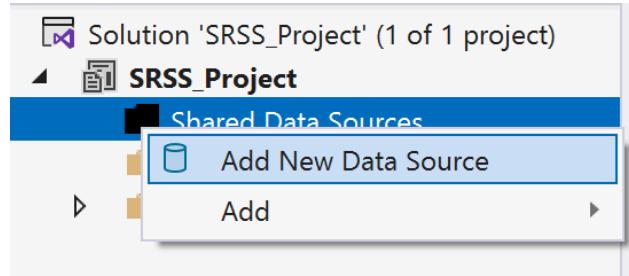


**Bước 4:** Cửa sổ Add New Item hiện ra, chọn Report, điền tên Report và nhấn nút Add

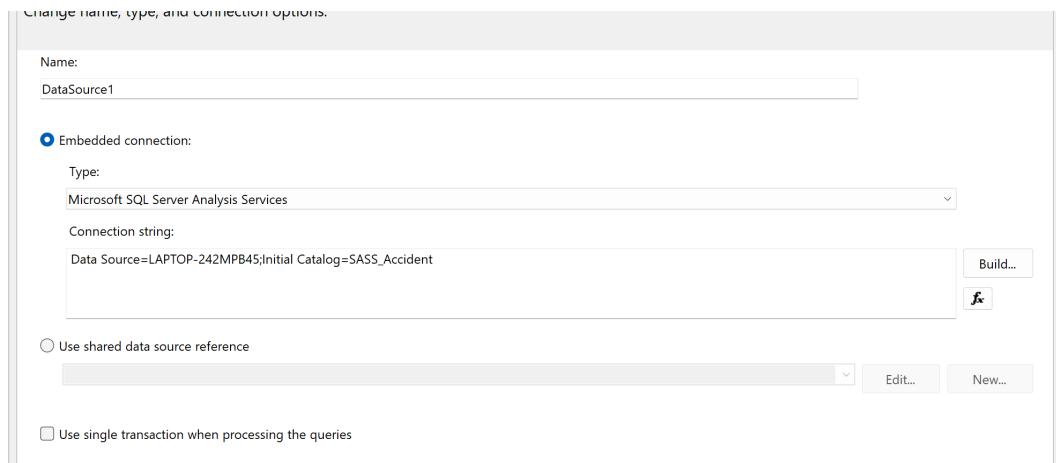


#### 4.1.2 Tạo và cấu hình Share Data Sources

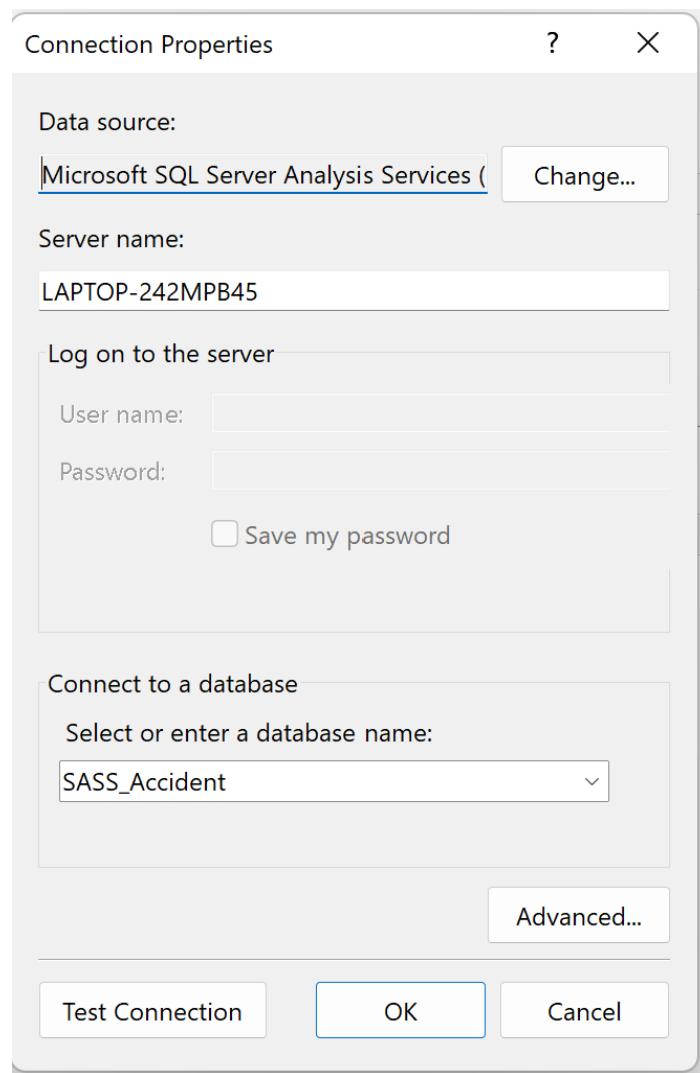
**Bước 1:** Tại tab Solution Explorer, nhấp chuột phải vào Shared Data Sources, chọn Add New Data Source



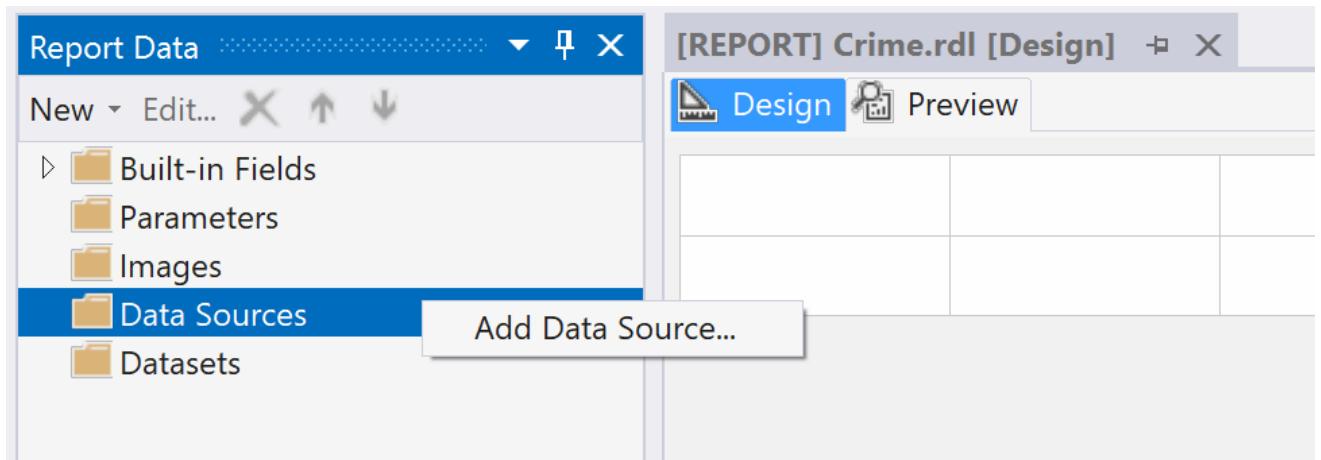
**Bước 2:** Cửa sổ Shared Data Source Properties hiện lên, điền trường tên, click vào mục Use shared data source reference và chọn dataset. Nhấn OK.



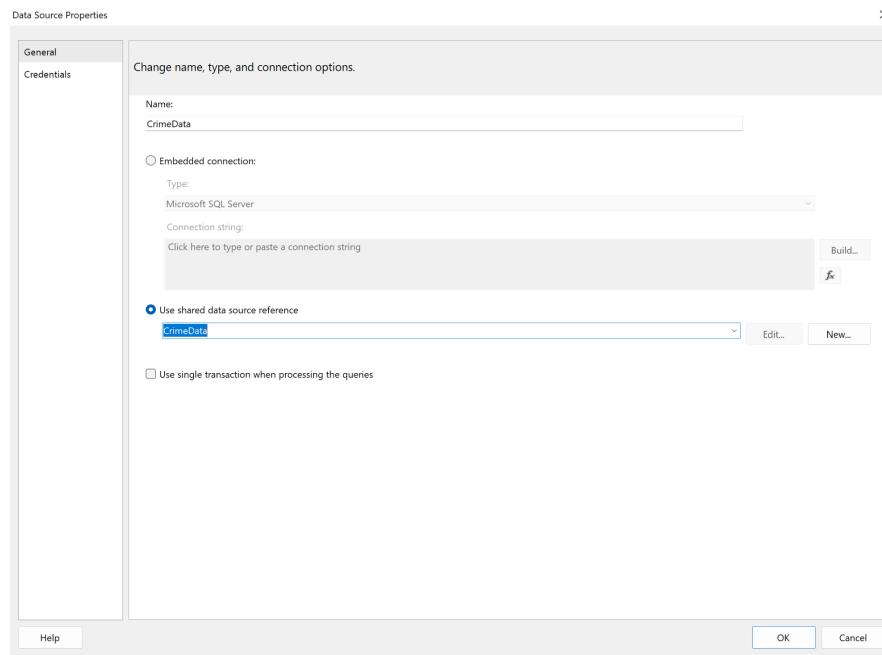
**Bước 3:** Bấm Build, hộp thoại Connection Properties mở ra. Điền tên server và chọn database name ở trường Select or enter a database name. Nhấn Test Connection để kiểm tra kết nối. Nhấn OK



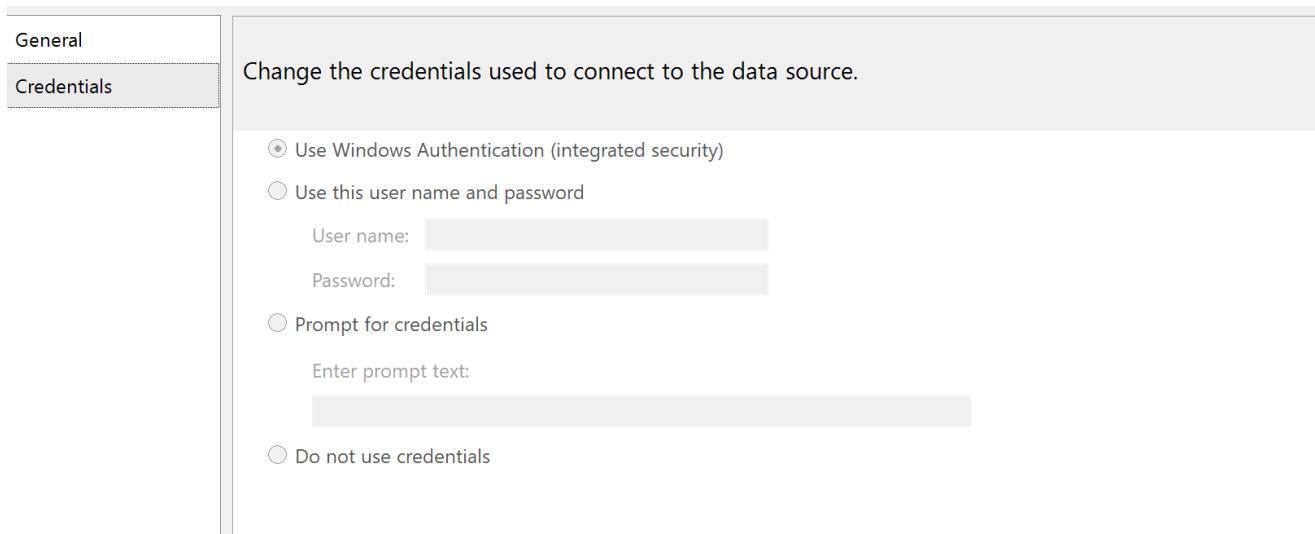
**Bước 5:** Tại thanh Report Data, tại phần Data Sources, nhấn chuột phải, chọn Add Data Sources



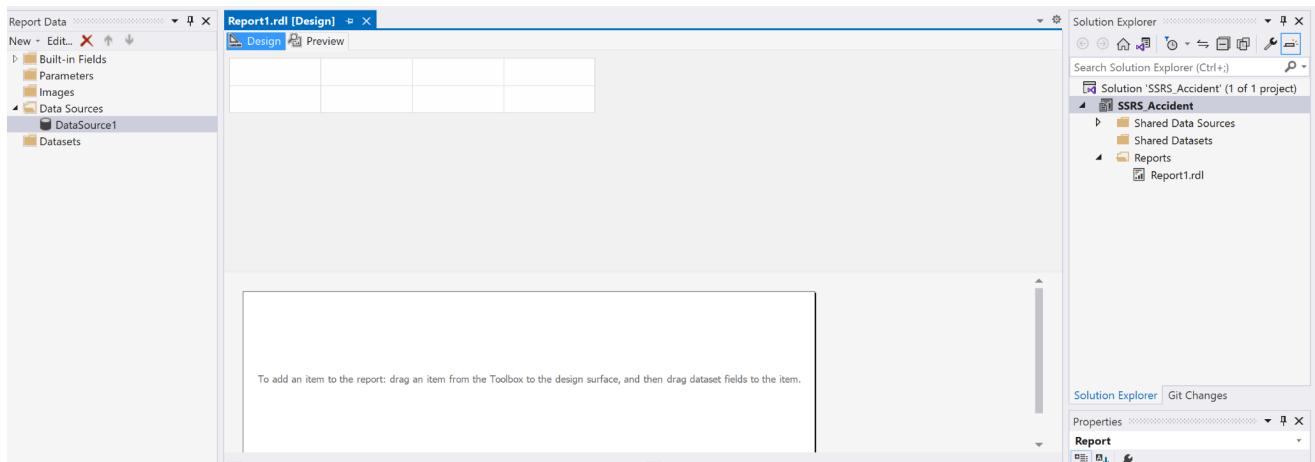
**Bước 6:** Cửa sổ Data Source Properties mở ra. Điền vào trường tên, và tick vào ô Use shared data source reference, chọn thông tin data source. Nhấn OK



**Bước 7:** File rds được tạo. Nhấn vào để mở cửa sổ Shared Data Source Properties. Tick vào ô Use Windows Authentication



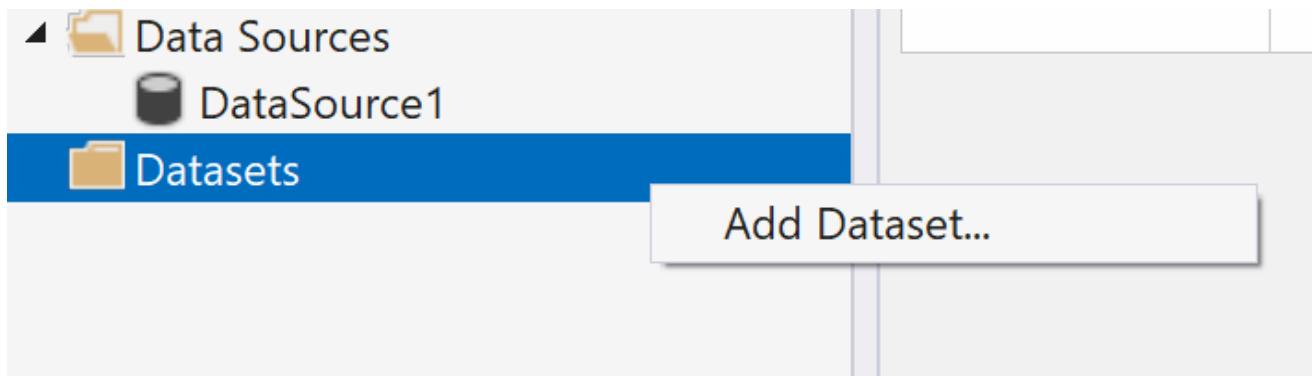
Sau khi đã hoàn tất, màn hình giao diện hiện lên như sau:



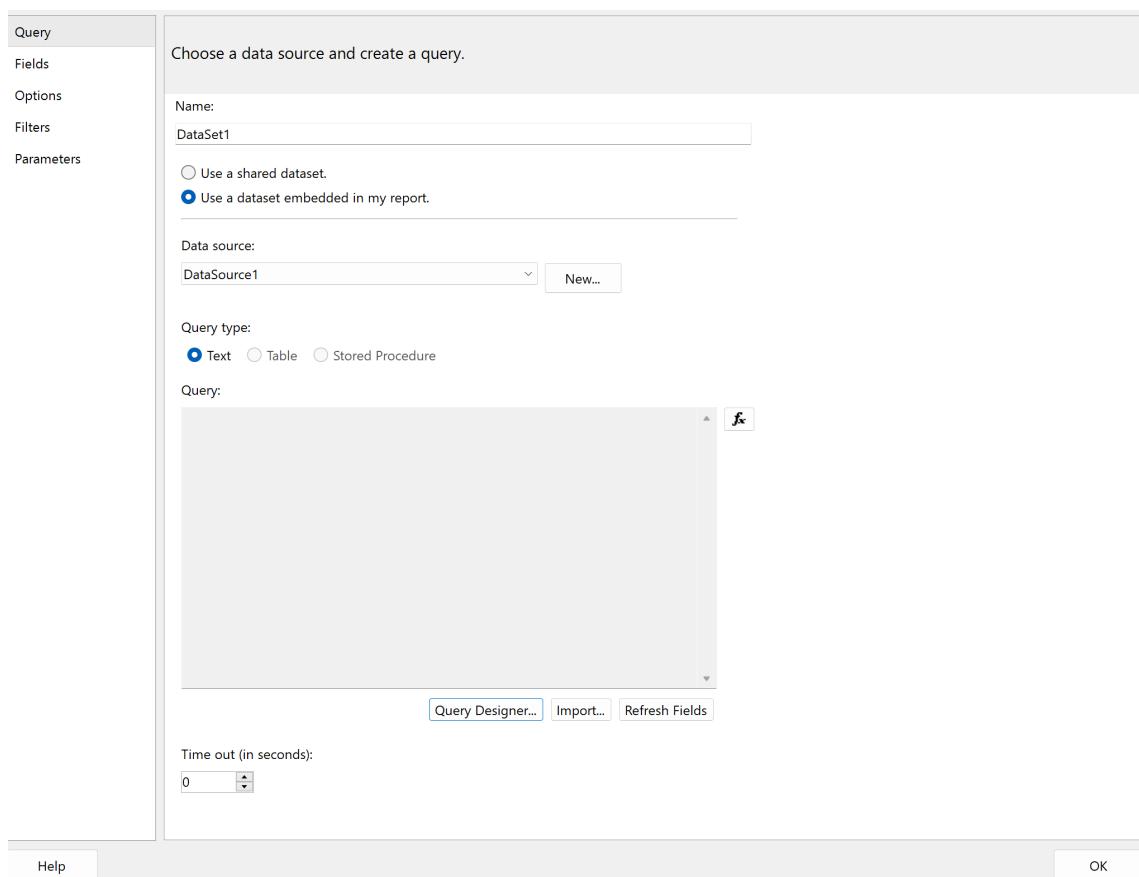
Trong đó

- Tab Design dùng để kéo thả và thiết kế báo cáo.
- Tab Preview dùng để hiển thị kết quả báo cáo, in báo cáo.
- Row Groups dùng để tạo và thực hiện các phép gom nhóm, tổng hợp giữa các dòng dữ liệu.
- Column Groups dùng để tạo và thực hiện các phép gom nhóm, tổng hợp giữa các cột dữ liệu.

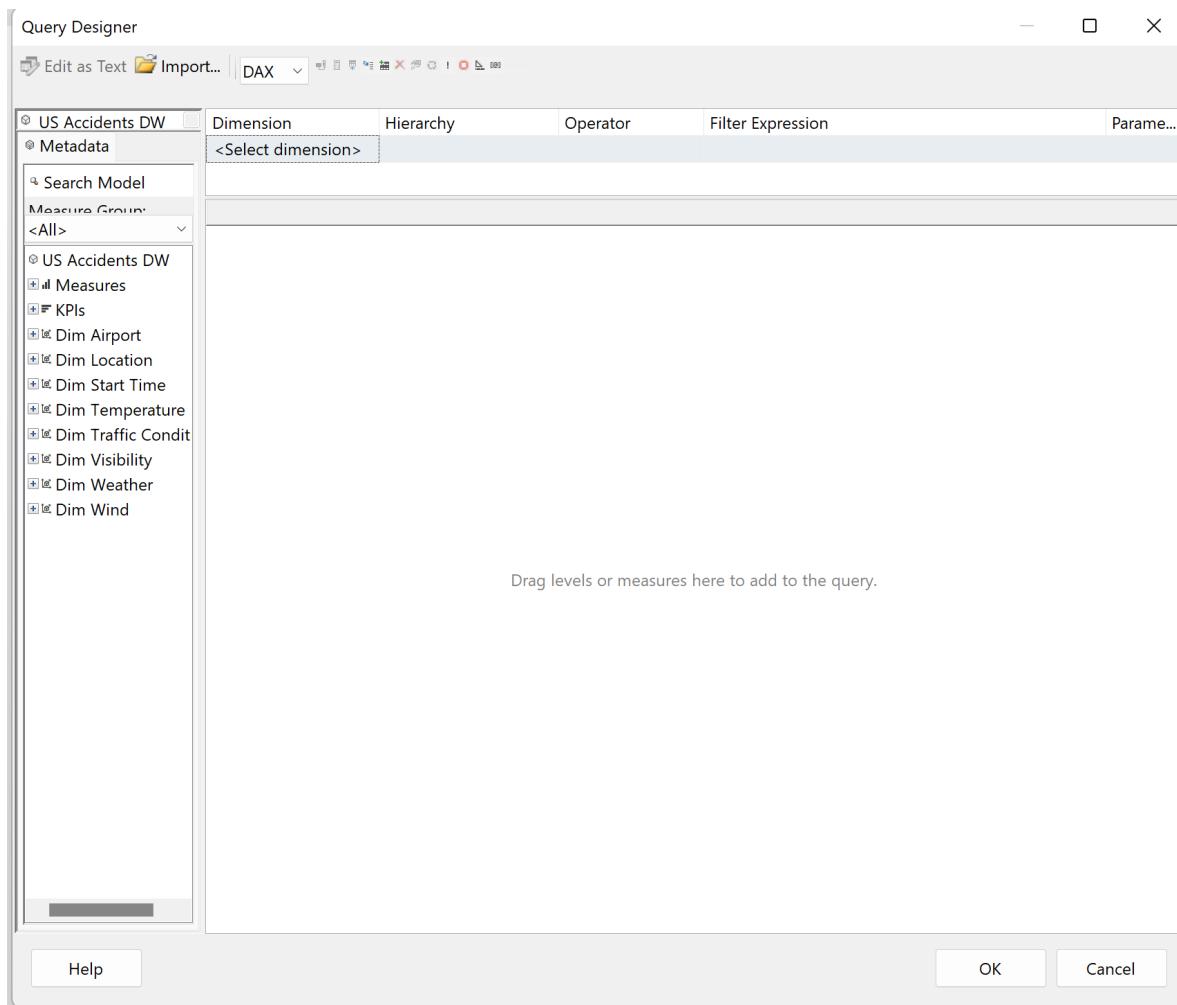
**Bước 8:** Tại cửa sổ Report Data, click chuột phải vào folder Datasets. Sau đó chọn Add Dataset... để thêm bộ dữ liệu cho báo biểu



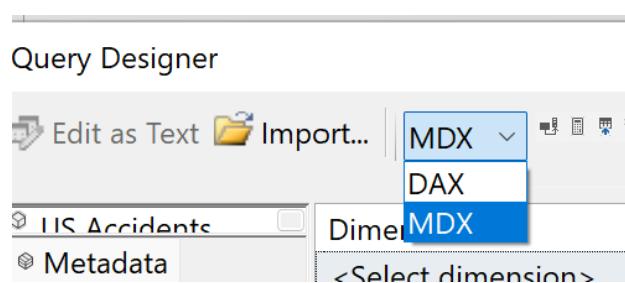
**Bước 9:** Hộp thoại Dataset Properties xuất hiện, ta nhập tên dataset, chọn chế độ Use a dataset embedded in my report và chọn dữ liệu nguồn là DataSource1 mà ta mới tạo ở các bước trên. Sau đó chọn Query Designer để thiết kế dữ liệu cho báo biểu.



**Bước 10:** Cửa sổ Query Designer xuất hiện cùng với bảng Fact và các bảng Dim mà ta đã tạo trong quá trình SSAS.



**Chú ý:** Nhớ đổi sang tab MDX thay vì DAX



**Bước 11:** Ta kéo thả các thuộc tính, độ đo cần thiết cho câu truy vấn. Sau đó, thực thi câu truy vấn để kiểm tra dữ liệu đầu vào cho báo biểu.

Query Designer

Edit as Text Import... MDX

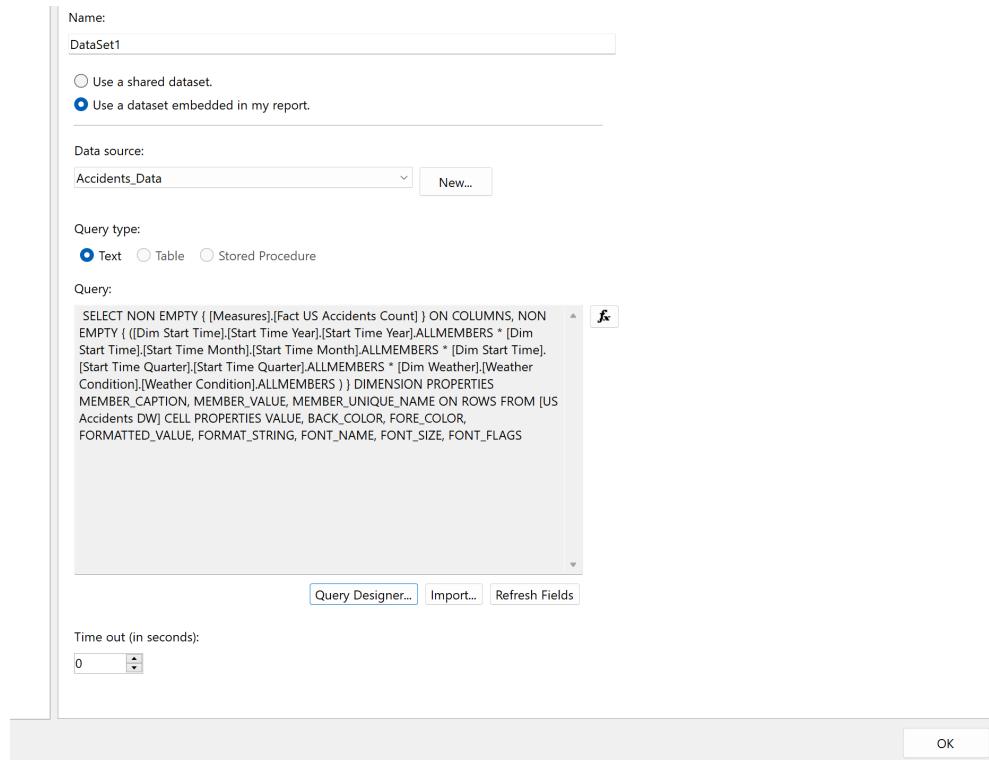
Dimension Hierarchy Operator Filter Expression Parameter

<Select dimension>

Start Time Year	Start Time Month	Start Time Quarter	Weather Condition	Fact US Accidents Count
2016	1	1	Fair	7
2016	10	4	Clear	6450
2016	10	4	Cloudy	21
2016	10	4	Drizzle	8
2016	10	4	Fair	92
2016	10	4	Fog	75
2016	10	4	Haze	99
2016	10	4	Heavy Drizzle	1
2016	10	4	Heavy Rain	50
2016	10	4	Heavy Thundersto...	12
2016	10	4	Light Drizzle	45
2016	10	4	Light Ice Pellets	1
2016	10	4	Light Rain	769
2016	10	4	Light Rain Showers	1
2016	10	4	Light Snow	11
2016	10	4	Light Thunderstor...	31
2016	10	4	Mist	7
2016	10	4	Mostly Cloudy	2251
2016	10	4	Overcast	2287
2016	10	4	Partly Cloudy	1846
2016	10	4	Patches of Fog	8
2016	10	4	Rain	175
2016	10	4	Scattered Clouds	1564
2016	10	4	Shallow Fog	18
2016	10	4	Smoke	1
2016	10	4	Snow	1
2016	10	4	Thunderstorm	9

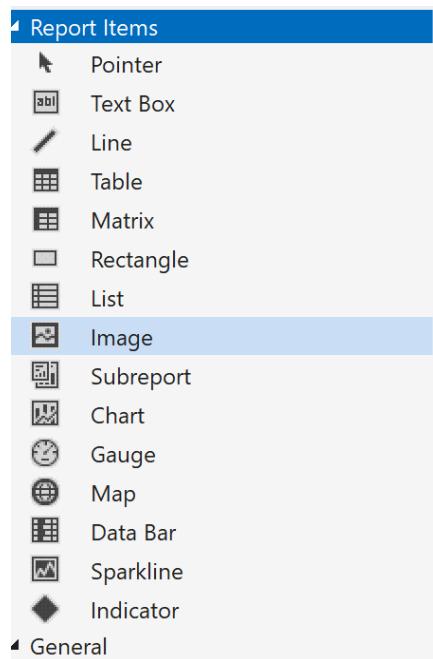
Help OK Cancel

Bước 12: Chọn OK để hoàn tất quá trình tạo dataset

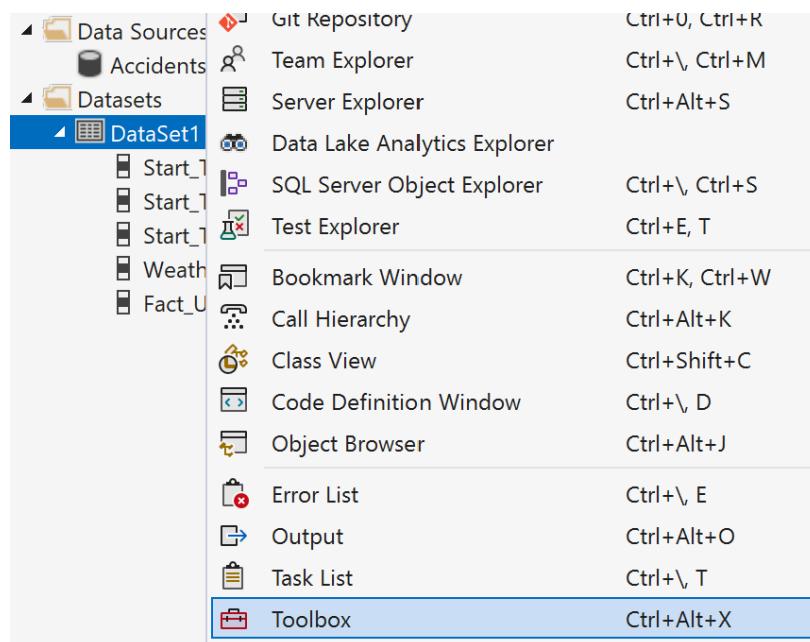


Dữ liệu vừa được tạo sẽ được hiển thị trong cửa sổ Report Data như sau

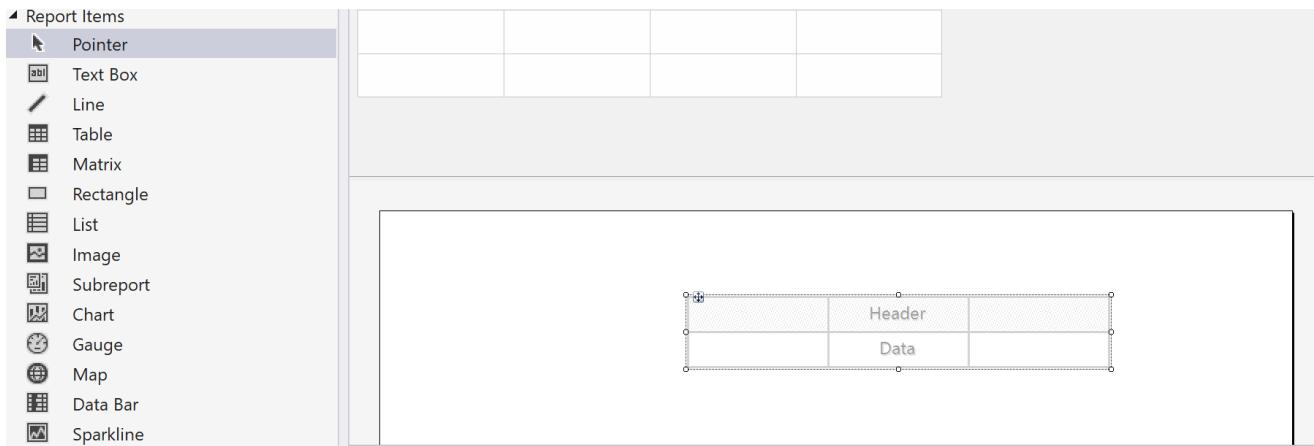
Danh sách Toolbox sẽ hiện ra như sau:



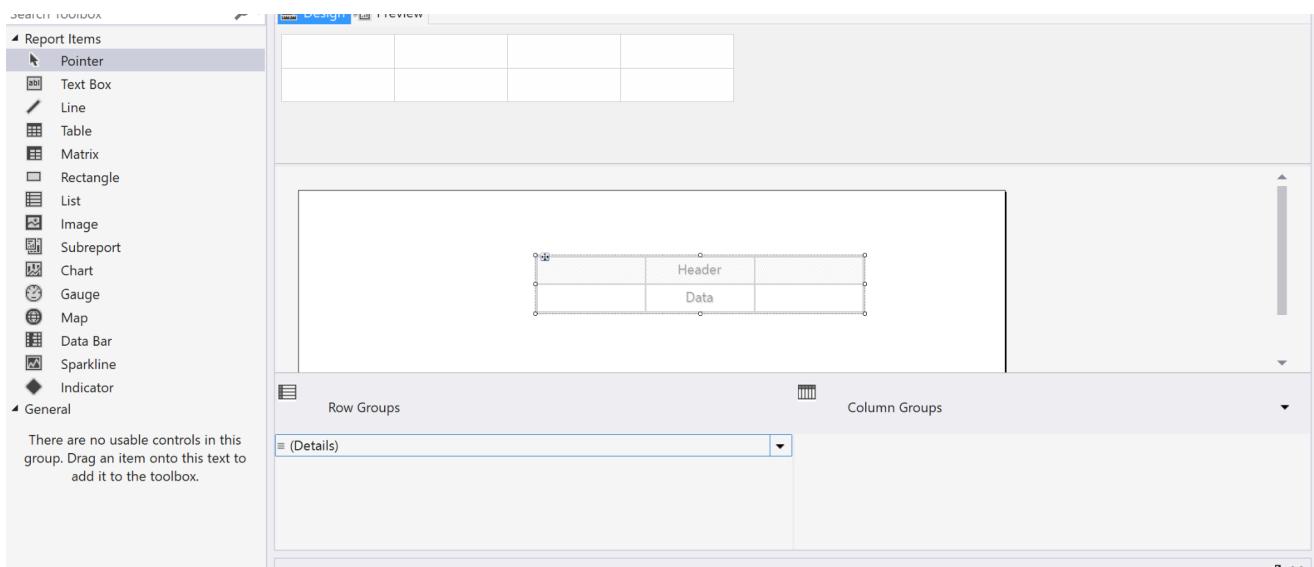
Bước 13: Chọn Tab View □ Toolbox để mở cửa sổ Toolbox



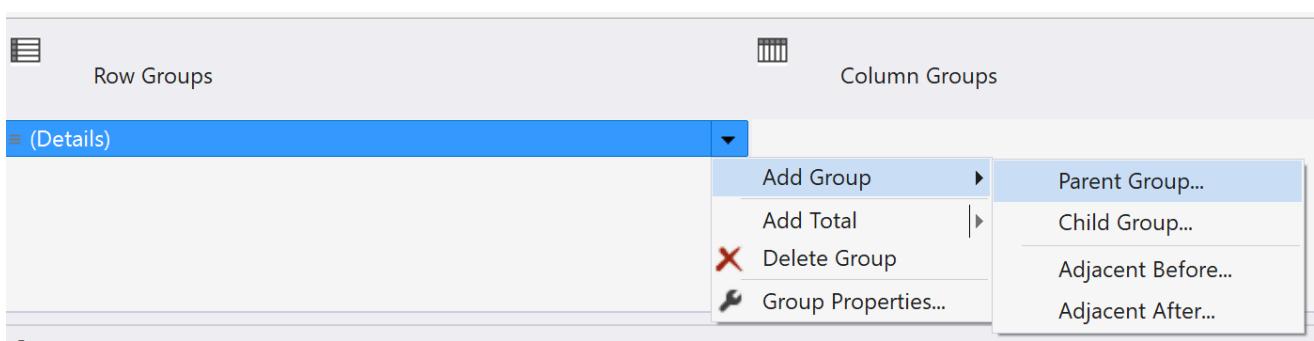
Bước 14: Kéo thả một table từ toolbox vào phần design

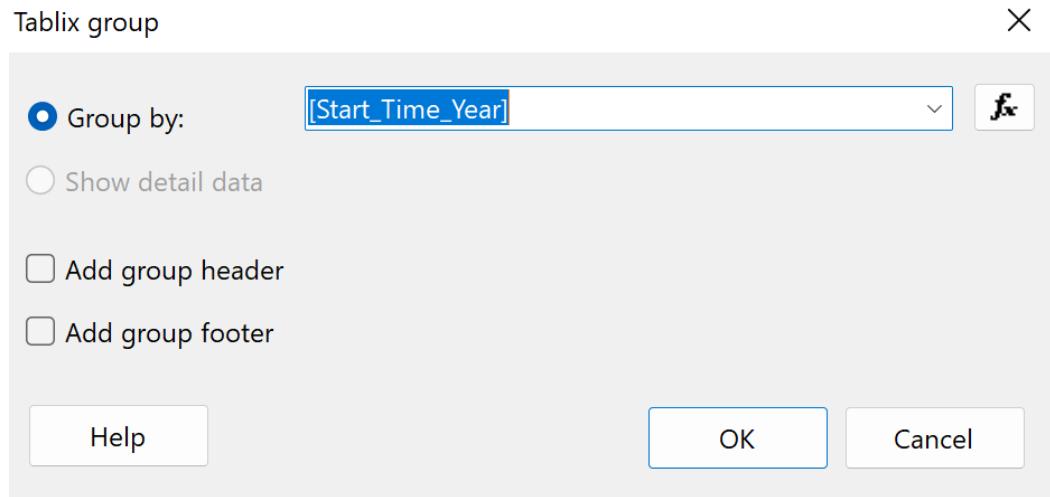


**Bước 15:** Sau đó ta kéo các thuộc tính lần lượt vào bảng

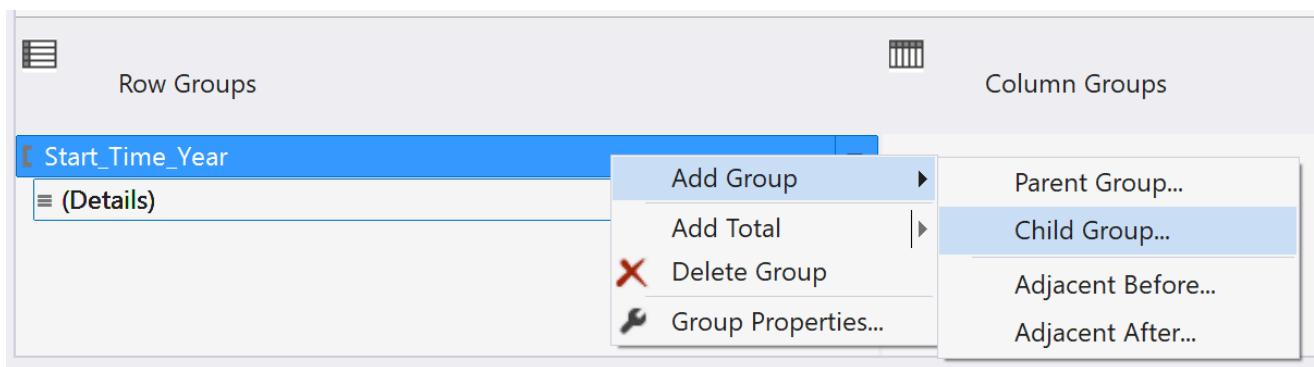


**Bước 16:** Tại cửa sổ Row Groups, ta click chuột phải vào Details. Sau đó, chọn Add Group □ Parent Group.

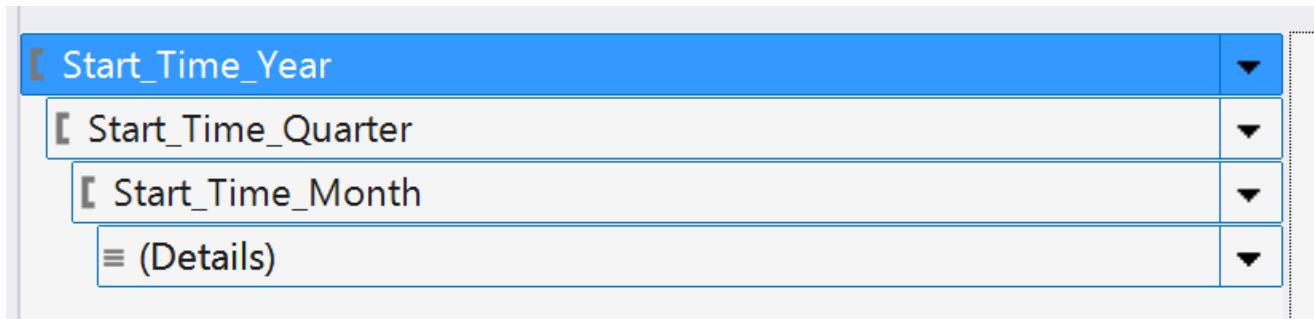




**Bước 17:** Tại cửa sổ Row Groups, ta click chuột phải vào Details. Sau đó, chọn Add Group □ Child Group.



**Kết quả:**



**Bước 18 :** Thiết kế, chỉnh sửa report sau đó nhấn “Preview” để xem trước

Number of Accidents/Weather by Years

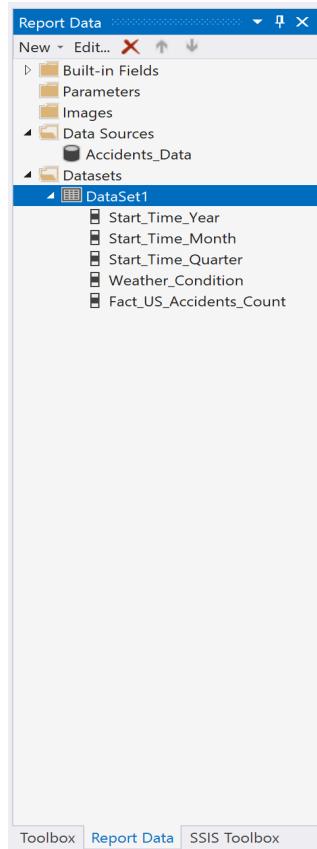
Start Time Year	Start Time Quarter	Start Time Month	Weather Condition	Fact US Accidents Count
[Start_Time_Year]	[Start_Time_Quarter]	[Start_Time_Month]	[Weather_Condition]	[ct_US_Accidents_Count]

Sau khi nhấn Preview:

Number of Accidents/Weather by Years

Start Time Year	Start Time Quarter	Start Time Month	Weather Condition	Fact US Accidents Count
2016	1	1	Fair	7
			Clear	102
			Cloudy	2
			Haze	7
			Light Drizzle	1
			Light Freezing Drizzle	4
			Light Rain	19
			Light Snow	97
			Mostly Cloudy	53
			Overcast	190
			Partly Cloudy	20
			Rain	14
			Scattered Clouds	22
			Snow	6
		3	Clear	965
			Cloudy	7
			Drizzle	4
			Fair	8

Sau đó, chuyển sang tab Report Data để tiến hành tạo thêm nhiều bảng biểu khác



#### 4.1.3 Tạo bảng biểu SSRS

**Nội dung bảng biểu 1: Thống kê các vụ tai nạn theo điều kiện trực tiếp**

← → ⌂ ⓘ localhost/ReportServer/Pages/ReportViewer.aspx?%2fSSRS\_Accidents%2fAccidents

|◀ < 1 of 2 ? > ▶| ⌂ 100% ⌄ | ↻ | ⌂ | 🔍 | ↗|

Number of Accidents/Weather by Years

Start Time Year	Start Time Quarter	Start Time Month	Weather Condition	Fact US Accidents Count
2016	1	1	Fair	7
		2	Clear	102
			Cloudy	2
			Haze	7
			Light Drizzle	1
			Light Freezing Drizzle	4
			Light Rain	19
			Light Snow	97
			Mostly Cloudy	53
			Overcast	190
			Partly Cloudy	20
			Rain	14
			Scattered Clouds	22
			Snow	6
	3	3	Clear	965
			Cloudy	7
			Drizzle	4
			Fair	8
			Fog	2
			Haze	10
			Heavy Rain	8
			Heavy Snow	1
			Light Drizzle	12
			Light Freezing Fog	2
			Light Rain	123
			Light Snow	26
			Light Thunderstorms and Rain	3
			Mist	1
			Mostly Cloudy	325
			Overcast	370
			Partly Cloudy	205
			Patches of Fog	1
			Rain	34
			Scattered Clouds	232
			Snow	2
			Thunderstorms and Rain	1
	2	4	Clear	2494
			Cloudy	6

Nội dung bảng biểu 2:

Query Designer

Query Designer

Edit as Text Import... MDX

Dimension	Hierarchy	Operator	Filter Expression	Param.
<Select dimension>				
State	County	City	Average Severity	Fact US Accidents Count
AL	Autau...	Au...	3.14285714285...	9
AL	Autau...	Bill...	2.5	4
AL	Autau...	De...	2.76923076923...	49
AL	Autau...	Ma...	2.46153846153...	30
AL	Autau...	Pra...	2.5748031496063	147
AL	Autau...	Sel...	4	3
AL	Baldwin	Ba...	2.44117647058...	78
AL	Baldwin	Da...	2.48257372654...	454
AL	Baldwin	Elb...	3.5	6
AL	Baldwin	Fai...	3.25	25
AL	Baldwin	Fol...	3.71428571428...	9
AL	Baldwin	Lo...	2.3125	87
AL	Baldwin	Ma...	2	1
AL	Baldwin	Or...	2	2
AL	Baldwin	Per...	2.71428571428...	15
AL	Baldwin	Ro...	2.54545454545...	90
AL	Baldwin	Se...	4	7
AL	Baldwin	Sil...	2	1
AL	Baldwin	Sp...	2.44444444444...	9
AL	Baldwin	Sta...	2	2
AL	Barbour	Clio	2.66666666666...	3
AL	Barbour	Euf...	2	6
AL	Barbour	Mi...	2	1
AL	Bibb	Bre...	3.5	5
AL	Bibb	Bri...	2	2
AL	Bibb	Ce...	2	8
AL	Bibb	La...	2	2
AL	Bibb	Re...	2	2

Help OK Cancel

## Cài đặt gom nhóm và phân cấp

Row Groups

- State
- County

Kết quả:

**Number Accidents & Average Severity by State/Country**

State	County	Fact US Accidents Count	Average Severity
AL	Autauga	242	2.681592039801
	Baldwin	786	2.54141104294479
	Barbour	10	2.2
	Bibb	31	2.96
	Blount	433	2.73939393939394
	Bullock	32	4
	Butler	123	2.77227722772277
	Calhoun	102	2.50561797752809
	Chambers	40	2.875
	Cherokee	14	4.22222222222222
	Chilton	733	2.39197530864198
	Choctaw	12	5.33333333333333
	Clarke	9	3.71428571428571

**Chế độ Preview**

The screenshot shows a software interface with a toolbar at the top. The 'Preview' tab is selected. Below the toolbar, there are navigation buttons (Back, Forward, Home, etc.) and a search bar labeled 'Find' and 'Next'. The main area displays a table titled 'Number Accidents & Average Severity by State/Country'. The table has three columns: 'State', 'County', and 'Fact US Accidents Count'. The data is grouped by state (AL) and then by county. The table shows the following data:

Number Accidents & Average Severity by State/Country		
State	County	Fact US Accidents Count
AL	Autauga	242
	Baldwin	786
	Barbour	10
	Bibb	31
	Blount	433
	Bullock	32
	Butler	123
	Calhoun	102
	Chambers	40
	Cherokee	14
	Chilton	733
	Choctaw	12
	Clarke	9
	Clay	9
	Cleburne	146
	Coffee	22
	Colbert	34
	Conecuh	75
	Coosa	18
	Covington	9
	Crenshaw	8

### Nội dung bảng biểu 3:

#### Bước 1: Chọn chế độ Script Mode

Đầu tiên, tạo mới dataset sau đó tiến hành lấy ra top 7 thời tiết có nhiều vụ tai nạn nhất

```
SELECT {[Measures].[Fact US Accidents Count]} ON 0,
TOPCOUNT( [Dim Weather].[Weather Condition].[Weather Condition],7,[Measures].[Fact US Accidents Count]) ON 1
FROM [US Accidents DW]
```

Query Designer

Edit as Text Import... MDX

```
SELECT {[Measures].[Fact US Accidents Count]} ON 0,
TOPCOUNT( [Dim Weather].[Weather Condition].[Weather Condition],7,[Measures].[Fact US Accidents Count]) ON 1
FROM [US Accidents DW]
```

US Accidents DW

Measure

Fact US Accidents

- Average Distance
- Average Duration
- Average Speed
- Fact US Accidents
- Max Distance
- Min Distance
- Sum Distance

KPIs

Dim Airport

Dim Location

Dim Precipitation

Dim Start Time

Dim Temperature

Dim Traffic Conditions

Dim Visibility

Dim Weather

- Weather Condition
- Weather ID

Dim Wind

Weather Condition	Fact US Accidents Count
Fair	1096922
Mostly Cloudy	362526
Cloudy	346461
Partly Cloudy	248496
Clear	173815
Light Rain	127871
Overcast	84878

Help OK Cancel

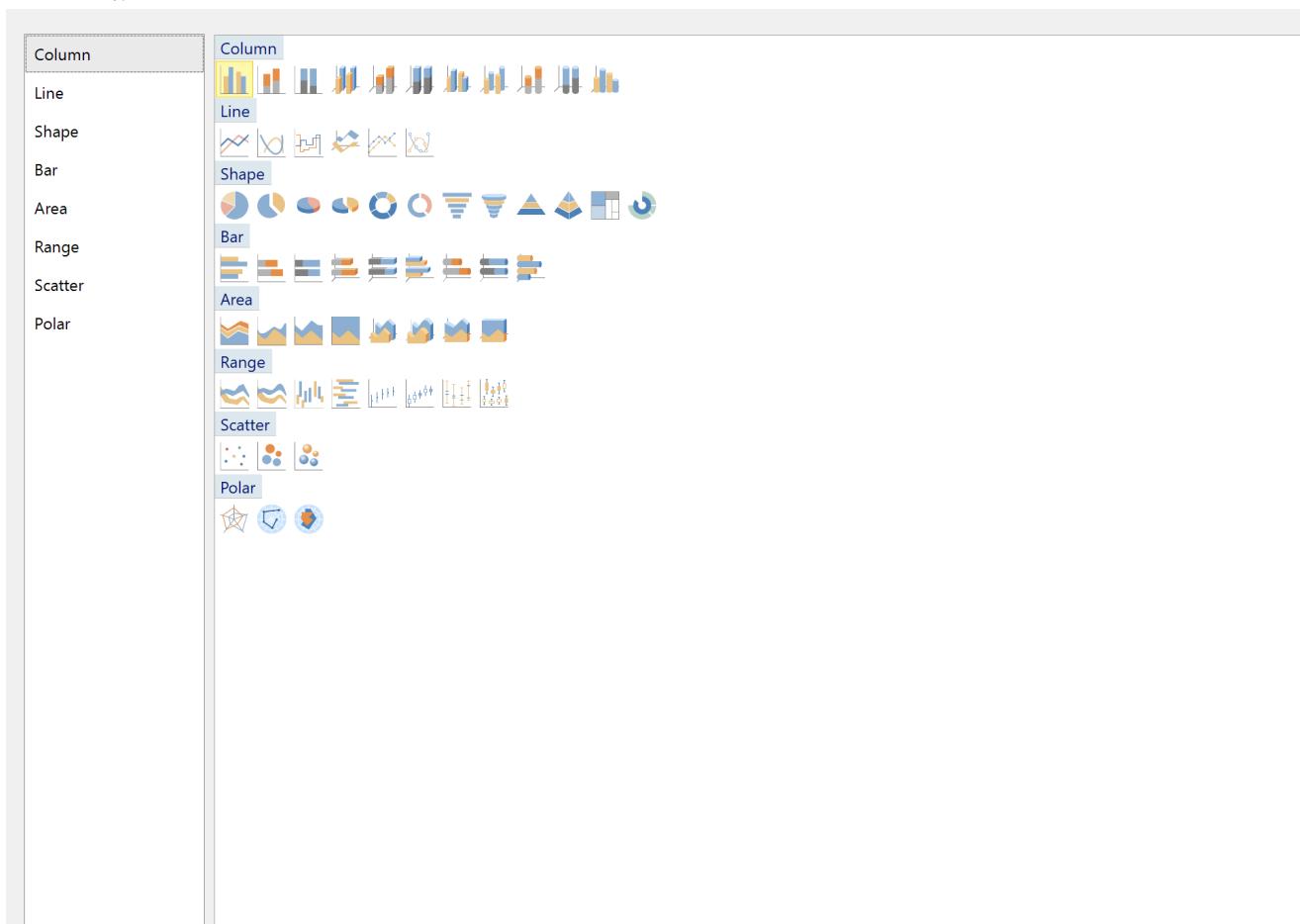
The screenshot shows the Microsoft Analysis Services Query Designer window. The MDX tab is selected. The query pane contains the following MDX code:

```
SELECT {[Measures].[Fact US Accidents Count]} ON 0,
TOPCOUNT( [Dim Weather].[Weather Condition].[Weather Condition],7,[Measures].[Fact US Accidents Count]) ON 1
FROM [US Accidents DW]
```

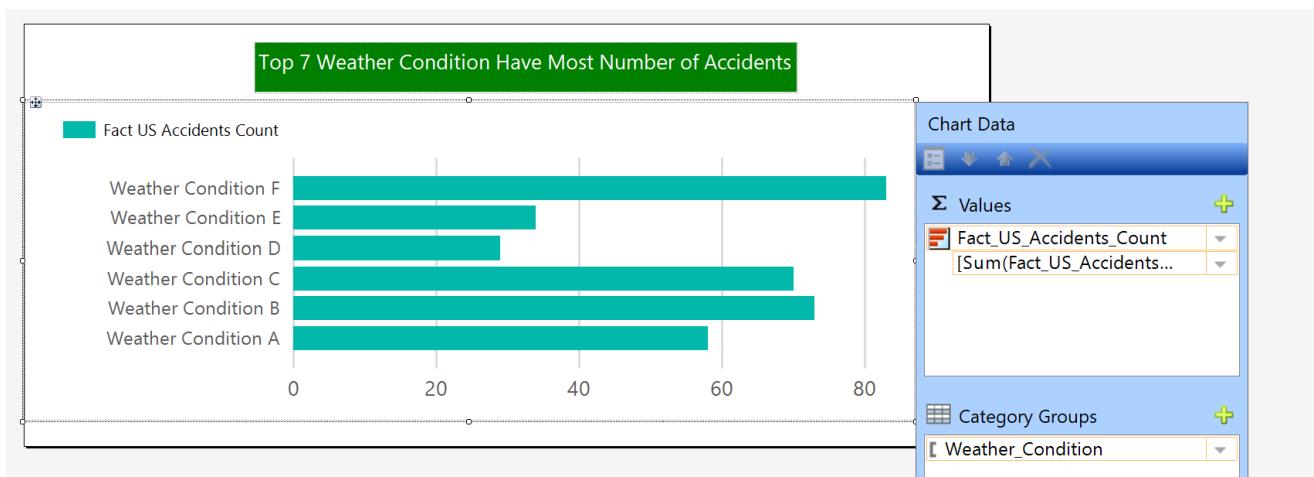
The left pane displays the cube structure with dimensions like US Accidents DW, Measures, Fact US Accidents, and various weather-related dimensions. A results grid is shown on the right, listing the top 7 weather conditions by accident count. The 'Fair' condition has the highest count at 1,096,922.

## Bước 2: Đổi tab Toolbox, chọn biểu đồ Chart tùy chỉnh

Select Chart Type



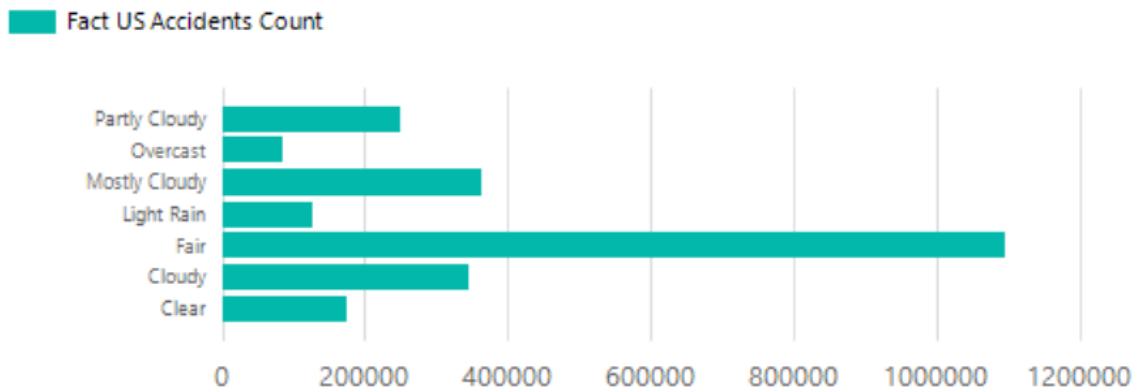
Sau đó chọn các giá trị và thẻ loại



Kết quả:

---

### Top 7 Weather Condition Have Most Number of Accidents



## 4.2 Tạo report với Power BI

### 4.2.1 Tạo project với Power BI

**Bước 1:** Mở Power BI, từ Home, chọn Get Data, chọn Analysis Services

The screenshot shows the Microsoft Power BI ribbon interface. The 'Get data' tab is currently selected, indicated by a grey background. The ribbon tabs include 'Get data', 'Excel', 'Data', 'SQL', 'Enter data', 'Dataverse', 'Recent sources', 'Transform', 'Refresh data', 'Queries', 'New visual', 'Text box', and 'More visuals'. Below the ribbon, a 'Common data sources' dropdown menu is open, listing various options: 'Excel workbook', 'Power BI datasets', 'Dataflows', 'Dataverse', 'SQL Server', 'Analysis Services', 'Text/CSV', 'Web', 'OData feed', 'Blank query', 'Power BI Template Apps', and 'More...'. A tooltip for the 'SQL Server' icon is displayed, stating: 'Live connect or import data from a SQL Server Analysis Services database. Once loaded, your data will appear in the Data pane'. At the bottom right, there is a link 'Get data from another source →'.

Get data

Excel

Data

SQL

Enter data

Dataverse

Recent sources

Transform

Refresh data

Queries

New visual

Text box

More visuals

Common data sources

- Excel workbook
- Power BI datasets
- Dataflows
- Dataverse
- SQL Server
- Analysis Services
- Text/CSV
- Web
- OData feed
- Blank query
- Power BI Template Apps
- More...

Live connect or import data from a SQL Server Analysis Services database.  
Once loaded, your data will appear in the Data pane

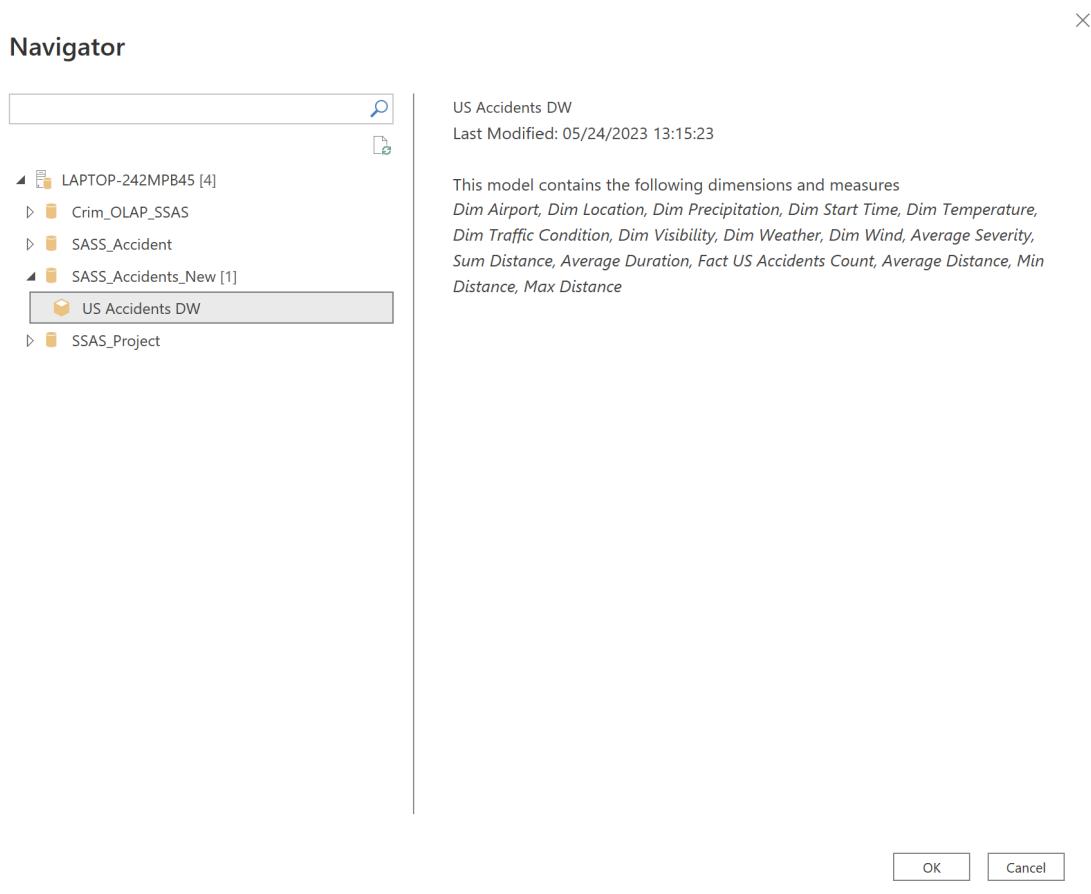
Import data from SQL Server

Paste data into a blank table

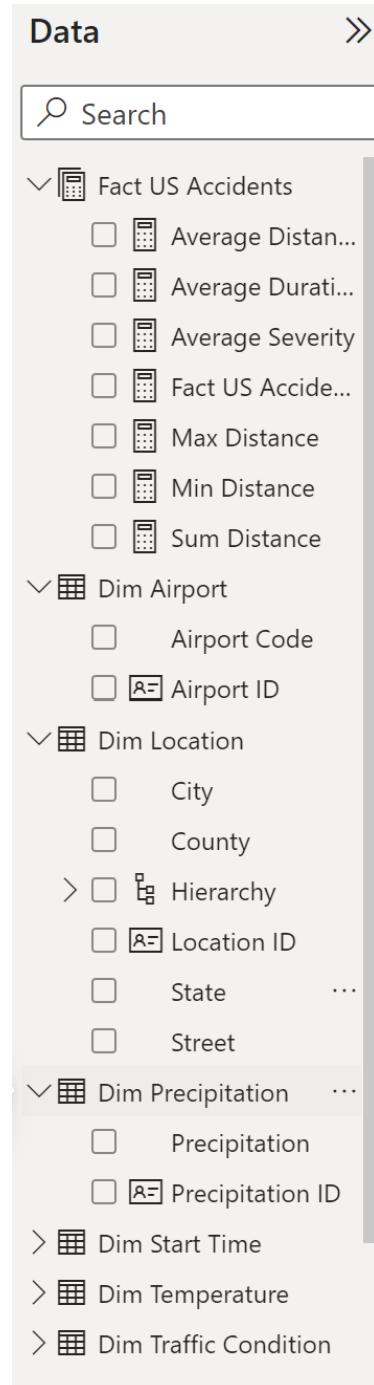
Get data from another source →



**Bước 2:** Cửa sổ SQL Server Analysis Services data mở ra, điền tên server, database và nhấn OK

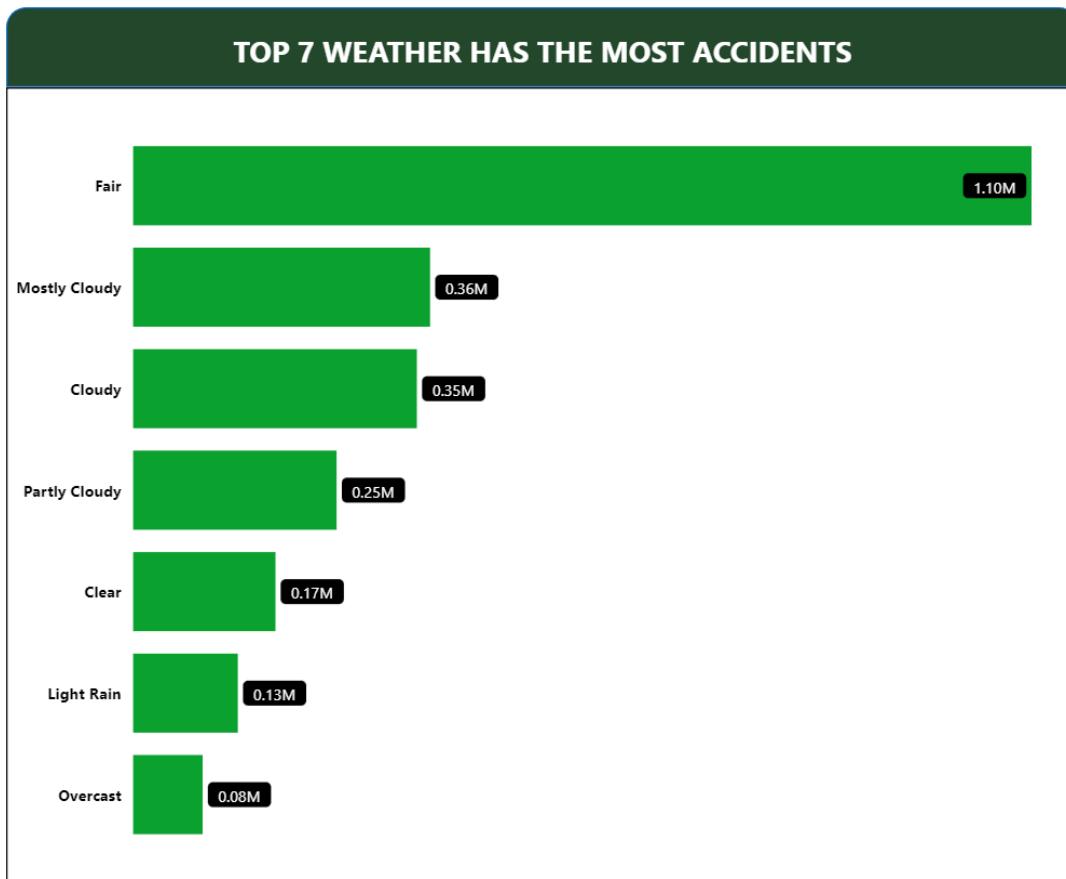


**Bước 3:** Cửa sổ Navigator mở ra, chọn database và nhấn OK



#### 4.2.2 Tạo report với 3 report dạng Table, Matrix, Chart.

a . Report 1: Thống kê top 7 số lượng tai nạn nhiều nhất theo kiểu thời tiết ( Biểu đồ dạng Chart)



b . Report 2: Thông kê top 5 số lượng tai nạn nhiều nhất theo kiểu thời tiết / năm (Report dạng Matrix )

TOP 5 AVERAGE SEVERITY AND NUMBER OF ACCIDENTS BY YEARS/STATE											
State		CA		FL		OR		TX		VA	
Start Time	Year	Average Severity	Fact US Accidents Count								
<b>2019</b>											
1	2019	2.9	6480	3.3	2659	2.2	11355	3.1	1952	3.0	1345
2	2019	2.9	6500	3.3	3000	2.1	5493	3.0	2275	3.0	1289
3	2019	2.8	30435	3.3	2399	2.1	8920	3.0	1545	3.0	792
4	2019	2.8	60513	3.4	1668	2.1	8308	3.2	796	3.4	579
<b>2020</b>											
1	2020	2.7	61449	2.8	3504	2.1	8755	3.0	1745	2.7	1605
2	2020	2.4	50889	2.3	9228	2.1	8245	2.7	6034	2.4	6585
3	2020	2.6	5782	3.5	8603	3.8	1395	2.3	1538	2.6	2216
4	2020	3.6	70481	3.6	53699	2.9	9313	2.4	16717	2.6	12550
<b>2021</b>											
1	2021	3.7	75904	3.7	49957	3.0	11330	2.4	15773	2.6	11021
2	2021	3.3	66268	3.4	44561	2.7	7041	2.5	16767	2.7	13398
3	2021	3.5	87514	3.4	62669	2.8	9467	2.5	22382	2.8	16060
4	2021	4.4	147097	4.3	117581	3.1	16243	2.7	26610	3.1	26479

c. Report 3: Thông kê các khu vực có khoảng cách xảy ra tai nạn nhiều nhất và số tai nạn theo tên đường với filter là năm ( Report kiểu Table )

2016	2017	2018	2019	2020	2021
------	------	------	------	------	------

MAX DISTANCE AND N. OF ACCIDENTS BY STREET		
Street	Max Distance	Number of Accidents
I-95 S	11,407	5,111
I-95 N	11,049	4,909
S Orange Blossom Trl	9,441	4,691
I-10 W	7,350	3,559
SW 137th Ave	7,108	3,554
I-10 E	7,076	3,412
I-10 E	6,480	3,182
I-635 W	6,408	2,836
S Dixie Hwy	6,319	3,108
I-45 N	6,224	2,582
I-5 N	5,747	2,783
SW 88th St	5,662	2,830
CA-91 E	5,475	2,685
SW 87th Ave	5,132	2,566
Brooklyn Queens Expy	5,077	2,425
Golden State Fwy S	5,054	2,450
I-95 S	5,008	2,288
SW 8th St	4,815	2,398
Harbor Fwy N	4,746	2,294
<b>Total</b>	<b>3,199,801</b>	<b>1,540,671</b>

Kết quả sau khi tạo report trên PowerBI:

## USA ACCIDENTS OVERVIEW FROM 2016 TO 2021

**2.76M**

Fact US Accidents Count

**5.89M**

Max Distance

**1.00**

Min Distance

**1.14M**

Sum Distance

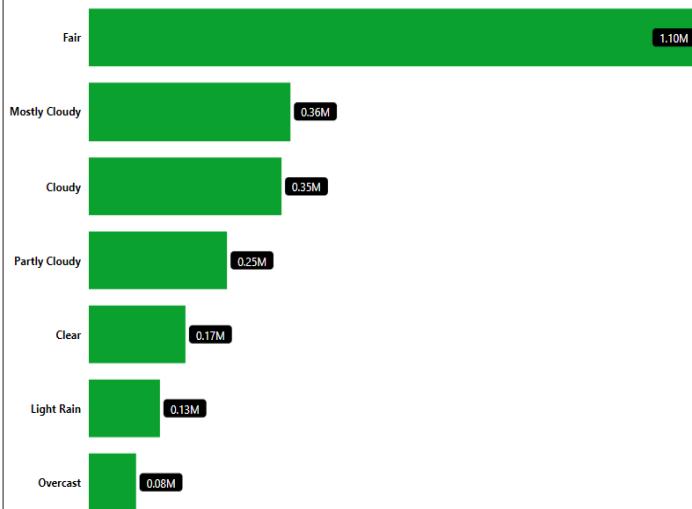
**3.35**

Average Distance

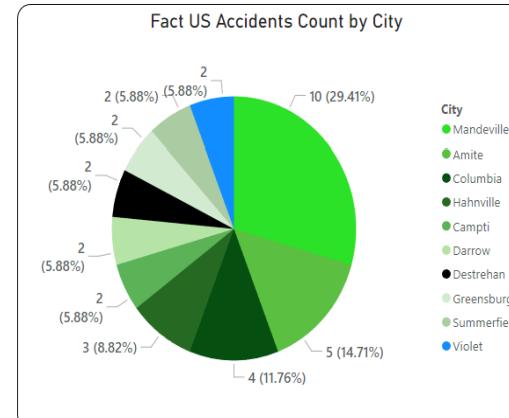
**3.35**

Average Severity

### TOP 7 WEATHER HAS THE MOST ACCIDENTS



### NUMBER OF ACCIDENTS BY CITY THAT AVERAGE SEVERITY > 5



→ Go to the following page

## USA ACCIDENTS OVERVIEW FROM 2016 TO 2021

**FILTER**

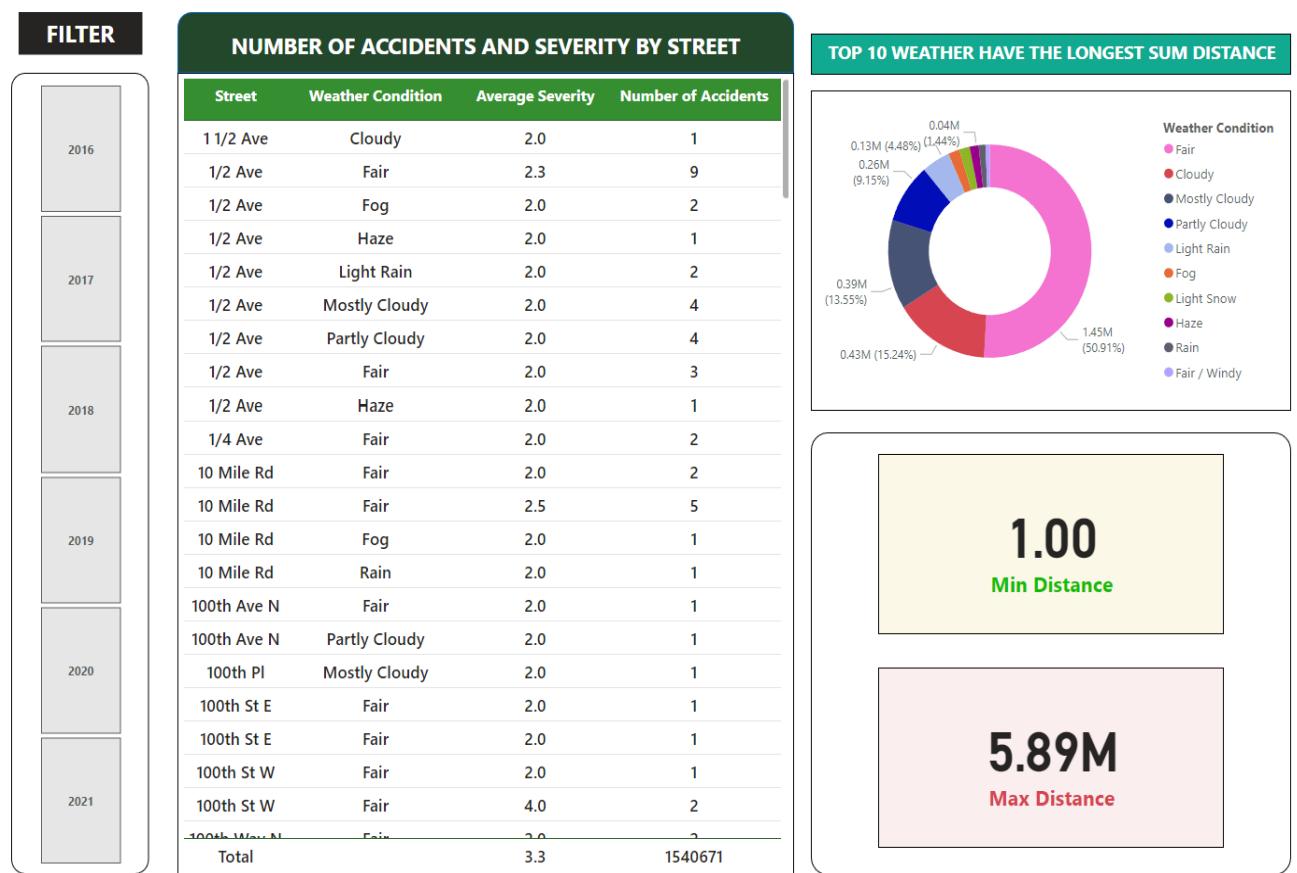
Start Time Year  
All

Wind Direction  
All

### TOP 7 WEATHER HAS THE MOST ACCIDENTS

State	CA		FL		OR		TX		VA		Total	
Start Time Year	Average Severity	Fact US Accidents Count										
□ 2019	<b>2.8</b>	<b>103928</b>	<b>3.3</b>	<b>9726</b>	<b>2.1</b>	<b>34076</b>	<b>3.0</b>	<b>6568</b>	<b>3.1</b>	<b>4005</b>	<b>2.7</b>	<b>158303</b>
1	2.9	6480	3.3	2659	2.2	11355	3.1	1952	3.0	1345	<b>2.6</b>	<b>23791</b>
2	2.9	6500	3.3	3000	2.1	5493	3.0	2275	3.0	1289	<b>2.7</b>	<b>18557</b>
3	2.8	30435	3.3	2399	2.1	8920	3.0	1545	3.0	792	<b>2.6</b>	<b>44091</b>
4	2.8	60513	3.4	1668	2.1	8308	3.2	796	3.4	579	<b>2.7</b>	<b>71864</b>
□ 2020	<b>2.9</b>	<b>188601</b>	<b>3.3</b>	<b>75034</b>	<b>2.4</b>	<b>27708</b>	<b>2.5</b>	<b>26034</b>	<b>2.6</b>	<b>22956</b>	<b>3.0</b>	<b>340333</b>
1	2.7	61449	2.8	3504	2.1	8755	3.0	1745	2.7	1605	<b>2.6</b>	<b>77058</b>
2	2.4	50889	2.3	9228	2.1	8245	2.7	6034	2.4	6585	<b>2.4</b>	<b>80981</b>
3	2.6	5782	3.5	8603	3.8	1395	2.3	1538	2.6	2216	<b>3.2</b>	<b>19534</b>
4	3.6	70481	3.6	53699	2.9	9313	2.4	16717	2.6	12550	<b>3.7</b>	<b>162760</b>
□ 2021	<b>3.8</b>	<b>376783</b>	<b>3.8</b>	<b>274768</b>	<b>2.9</b>	<b>44081</b>	<b>2.6</b>	<b>81532</b>	<b>2.9</b>	<b>66958</b>	<b>3.9</b>	<b>844122</b>
1	3.7	75904	3.7	49957	3.0	11330	2.4	15773	2.6	11021	<b>3.7</b>	<b>163985</b>
2	3.3	66268	3.4	44561	2.7	7041	2.5	16767	2.7	13398	<b>3.4</b>	<b>148035</b>
3	3.5	87514	3.4	62669	2.8	9467	2.5	22382	2.8	16060	<b>3.5</b>	<b>198092</b>
4	4.4	147097	4.3	117581	3.1	16243	2.7	26610	3.1	26479	<b>4.6</b>	<b>334010</b>
Total	<b>3.3</b>	<b>669312</b>	<b>3.7</b>	<b>359528</b>	<b>2.5</b>	<b>105865</b>	<b>2.6</b>	<b>114134</b>	<b>2.8</b>	<b>93919</b>	<b>3.4</b>	<b>1342758</b>

## USA ACCIDENTS OVERVIEW FROM 2016 TO 2021



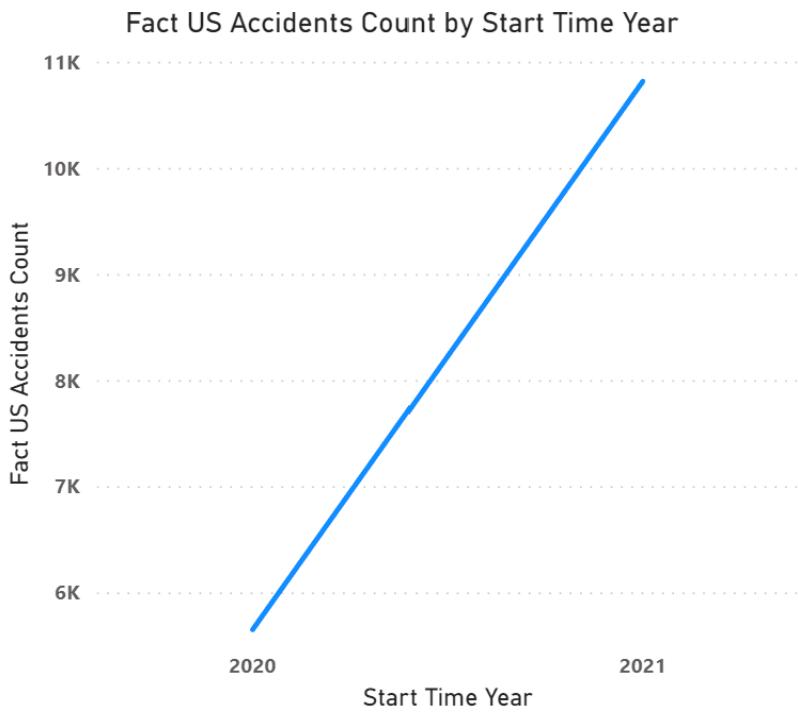
+ Thực hiện từng câu truy vấn trên Power BI

Câu 1: Liệt kê 10 quận xảy ra nhiều vụ tai nạn nhất và sắp xếp theo thứ tự tăng dần.

County	Fact US Accidents Count
Riverside	40665
Harris	42000
Sacramento	46479
San Diego	47016
Dallas	49331
Orange	50788
San Bernardino	54179
Orange	60206
Miami-Dade	142910
Los Angeles	225192
<b>Total</b>	<b>758766</b>

**Ý nghĩa:** Tìm ra 10 country có số lượng tai nạn lớn nhất và được sắp xếp từ thấp đến cao, với tổng số lượng các vụ tai nạn 758766. Từ đó chú ý đến những thành phố có Top 10 vụ tai nạn và để người dùng, người phân tích chú ý hơn vào những thành phố này để ngăn ngừa tai nạn,..

**Câu 2:** Thống kê số vụ tai nạn xảy ra tại đường ray xe lửa ở các bang từ năm 2020 đến 2021



**Ý nghĩa:** Nhìn vào đồ thị dạng Line Chart, người dùng có thể thấy được số vụ tai nạn bởi tàu ray xe lửa vào năm 2020 và 2021. Với sự tăng trưởng đi lên về số lượng của các vụ tai nạn (2021 gấp đôi 2020) để từ đó người dùng có thể thấy sự khác biệt và tăng lên nhanh chóng giữa 2 năm để rút ra những giải pháp hợp lý, hạn chế các vụ tai nạn

**Câu 3:** Cho biết số lượng các vụ tai nạn xảy ra của bốn hướng gió: Đông, Tây, Nam, Bắc thuộc các thành phố tại các tiểu bang với tốc độ gió lớn hơn 10 dặm/ giờ.

State	2016	2017	2018	2019	2020	2021	Total
AL		1		43	312	750	1106
AR		1	2	55	170	684	912
AZ			1	151	980	1631	2763
CA	30	39	48	7171	11781	25883	44952
CO	2	1	1	530	772	801	2107
CT				116	579	1381	2076
DC				30	227	613	870
DE				13	103	230	346
FL	12	14	19	862	10680	32284	43871
GA	6	12	7	184	658	443	1310
IA				67	191	882	1140
ID	1	5	1	49	83	391	530
IL				553	1210	1066	2829
IN	2	1	2	129	361	1068	1563
KS				60	184	694	938
KY				33	247	100	380
LA	11	15	15	60	1037	3624	4762
MA				49	415	105	569
MD	2	1	2	112	724	1624	2460
Total	111	152	222	17402	56909	133408	208204

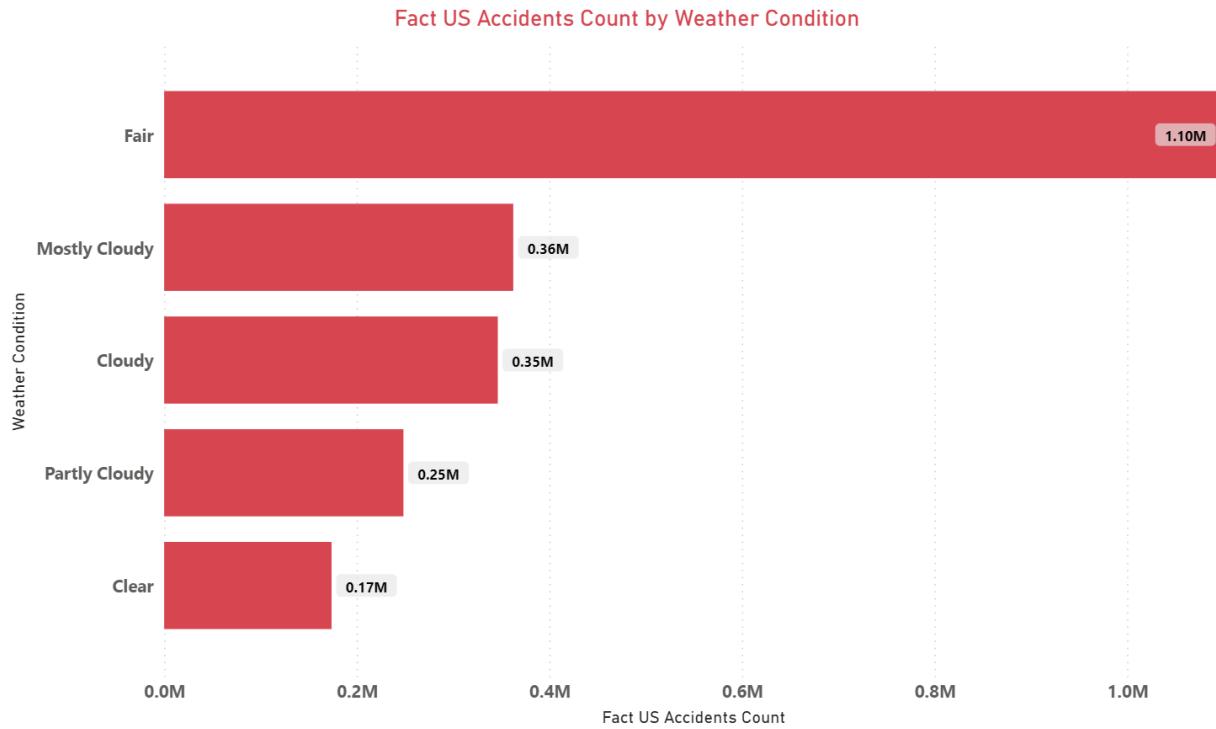
**Ý nghĩa :** Nhìn vào biểu đồ dạng Matrix trên, ta thấy được số lượng các vụ tai nạn có hướng gió là N,E,W,S và tốc độ gió lớn hơn 10 ( $\geq 10$ ) được nhóm bởi các cột State và City, trong đó State là phân cấp lớn hơn City.

Sau khi nhấn vào dấu cộng (+), biểu đồ dạng Matrix sẽ thả các City như sau:

State	2016	2017	2018	2019	2020	2021	Total
AL	1		43	312	750	1106	
Alabaster					9	9	
Alexander City				4	1	5	
Altoona					1	1	
Andalusia					2	2	
Anniston				3	5	8	
Arab					2	2	
Ardmore			1			1	
Ardmore					1	1	
Athens		1		8	4	13	
Atmore		1		1	7	9	
Auburn					1	1	
Axis				1		1	
Bay Minette				2	5	7	
Bessemer				1	3	4	
Birmingham		11		55	174	240	
Birmingham				7	16	23	
Blountsville				3		3	
Total	111	152	222	17402	56909	133408	208204

Từ đó, người dùng sẽ biết được các State và gồm các thành phố nào, có bao nhiêu vụ tai nạn. Và cách sử dụng biểu đồ như thế sẽ giúp người dùng có cái nhìn khái quát hơn, tổng quát hơn về Location

**Câu 4:** Thống kê theo năm, top 5 loại thời tiết có nhiều vụ tai nạn xảy ra nhất trong năm 2020 và năm 2021



**Ý nghĩa:** Nhìn vào biểu đồ cột trên, ta có thể thấy rằng đây là top 5 loại thời tiết gây ra các vụ tai nạn nhiều nhất gồm: Fair, Mostly Cloudy, Cloudy, Partly Cloudy, Clear. Với ở cuối chart là số lượng các vụ tai nạn và các loại thời tiết được sắp xếp theo từ lớn nhất là 1.10M đến thấp nhất là 0.17M. Với loại biểu đồ này, dữ liệu sẽ không thể hiện dưới dạng tổng mà sẽ được trực quan hóa với hình ảnh để người dùng biết rõ chi tiết về loại thời tiết/ số lượng

**Câu 5:** Thống kê tổng chiều dài của đoạn đường bị ảnh hưởng bởi các vụ tai nạn xảy ra tại nơi vừa có tín hiệu giao thông (Traffic Signal), vừa có trạm xe (Station) theo năm, quý, tiểu bang, quận, thành phố trong năm 2020, 2021.

State	2016	2017	2018	2019	2020	2021	Total
AL		5.00	0.00	1.00	17.00	3.00	<b>26.00</b>
AR			0.00	14.00		2.00	<b>16.00</b>
AZ	29.00	47.00	83.00	16.00	69.00	189.00	<b>433.00</b>
CA	91.00	61.00	62.00	92.00	622.00	1,935.00	<b>2,863.00</b>
CO	121.00	109.00	122.00	381.00	26.00	84.00	<b>843.00</b>
CT	0.00	28.00	1.00	10.00	15.00	7.00	<b>61.00</b>
DC	0.00	1.00	0.00	0.00	0.00	2.00	<b>3.00</b>
DE	0.00	2.00	0.00		3.00	7.00	<b>12.00</b>
FL	7.00	8.00	1.00	0.00	18.00	121.00	<b>155.00</b>
GA	17.00	7.00	2.00	6.00	10.00	11.00	<b>53.00</b>
IA	26.00	77.00	10.00	38.00	15.00	14.00	<b>180.00</b>
ID	1.00		1.00	0.00	36.00	86.00	<b>124.00</b>
IL	12.00	3.00	5.00	0.00	3.00	2.00	<b>25.00</b>
IN	2.00	19.00	2.00	1.00	0.00	22.00	<b>46.00</b>
KS	17.00			43.00	1.00	233.00	<b>294.00</b>
KY	1.00	2.00	1.00	0.00	1.00	1.00	<b>6.00</b>
LA		0.00	0.00		0.00	0.00	<b>0.00</b>
MA	0.00	0.00	0.00	0.00	0.00	1.00	<b>1.00</b>
MD	4.00	6.00	12.00	3.00	23.00	30.00	<b>78.00</b>
<b>Total</b>	<b>562.00</b>	<b>583.00</b>	<b>491.00</b>	<b>930.00</b>	<b>2,116.00</b>	<b>5,106.00</b>	<b>9,788.00</b>

**Câu 6:** Thống kê theo năm, tháng tổng số vụ tai nạn, chỉ số nghiêm trọng mà bị ảnh hưởng tại những con đường có tai nạn gờ giảm tốc khi nhiệt độ trong ngày trong khoảng -10 F đến 100 F

Start Time Year	2016																Drill on	Rows	...
City	1	10	11	12	2	3	4	5	6	7	8	9	Total	1	10				
[+] Abbeville																			
[+] Abbeville																			
[+] Abbeville																			
[+] Abbotsford																			
[+] Abbotsford																			
[+] Abbottstown													1.00		1.00				
[+] Abbottstown																			
[+] Aberdeen																			
[+] Aberdeen	1.00		1.00					1.00	3.00	11.00			17.00	2.00	3.00				
[+] Aberdeen																			
[+] Aberdeen	0.00												0.00		0.00				
[+] Abilene																			
[+] Abingdon			0.00											0.00		0.00			
[+] Abingdon																			
[+] Abington																			
[+] Absarokee																			
[+] Absecon	0.00		0.00				0.00		2.00	0.00	0.00	0.00	2.00	0.00	0.00				
<b>Total</b>	<b>0.00</b>	<b>5,385.00</b>	<b>6,167.00</b>	<b>7,598.00</b>	<b>677.00</b>	<b>1,259.00</b>	<b>1,990.00</b>	<b>2,112.00</b>	<b>3,782.00</b>	<b>5,790.00</b>	<b>6,246.00</b>	<b>4,903.00</b>	<b>45,909.00</b>	<b>5,746.00</b>	<b>4,328.00</b>				

**Ý nghĩa:** Nhìn vào Matrix, ta thấy được các thành phố (với dấu + là sẽ hiện ra các đường/Street) được thống kê theo năm và theo tháng. Với dữ liệu là tổng quãng đường xảy ra các vụ tai nạn. Với cách trực quan hóa dữ liệu như trên, người dùng có thể biết được năm và tháng, cũng như thành phố và con đường ở thành phố đó xuất hiện tai nạn

**Câu 7:** Liệt kê tổng số vụ tai nạn, mức độ nghiêm trọng trung bình, và thời gian trung bình của các vụ tai nạn giao thông theo quý trong năm.

Start Time Year	Average Severity	Number of Accidents	Average Distance
2016	2.7	58419	2.69
1	2.7	1227	2.66
2	2.6	14260	2.64
3	2.6	19943	2.65
4	2.8	22989	2.77
2017	2.7	70315	2.74
1	2.7	21341	2.73
2	2.7	18423	2.68
3	2.8	15631	2.77
4	2.8	14920	2.80
2018	2.7	77817	2.71
1	2.8	16820	2.80
2	2.8	14072	2.79
3	2.6	19818	2.58
4	2.7	27107	2.71
2019	2.7	160969	2.70
2020	3.0	340740	3.02
2021	3.9	832411	3.91
Total	3.3	1540671	3.32

**Ý nghĩa:** Nhìn vào Matrix ở hình trên, ta thấy rằng dữ liệu được thể hiện như sau: Cột Time Year sẽ gồm năm, và (+) là các quý trong năm, vì Year có phân cấp cao hơn Quarter. Từ cột Start Time Year, dữ liệu về độ nghiêm trọng (Severity, Number of Accidents, Distance) sẽ được thể hiện ở cột ngang (Rows). Và cột **total** của các quý trong năm sẽ được thể hiện ở dòng ngang có năm để người dùng có thể có cái nhìn tổng quát đến chi tiết về dữ liệu/số liệu về tai nạn

☒ 2018	2.7	77817	2.71
1	2.8	16820	2.80
2	2.8	14072	2.79
3	2.6	19818	2.58
4	2.7	27107	2.71

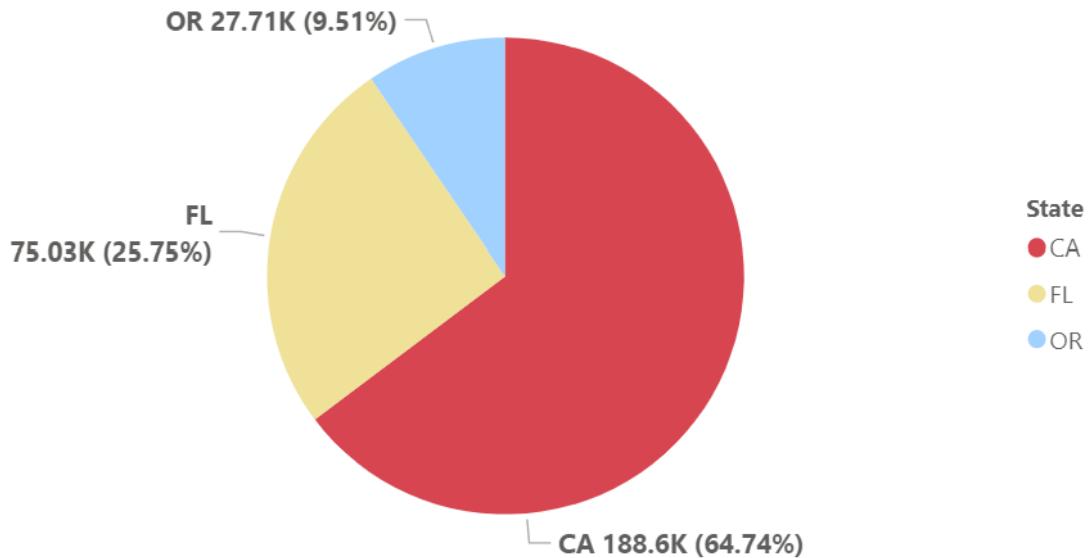
**Câu 8:** Cho biết tên con đường có số vụ tai nạn giao thông cao nhất và tên con đường có chỉ số nghiêm trọng tai nạn giao thông cao nhất.

Street	Fact US Accidents Count	Street	Average Severity
I-95 S	5111	N East Ave	48
<b>Total</b>	<b>5111</b>	<b>Total</b>	<b>48</b>

**Ý nghĩa:** Sau khi lọc điều kiện TOP 1 trong PowerBI ta được 2 bảng như trên, với yêu cầu là tìm con đường xảy ra nhiều vụ nhất và con đường có độ nghiêm trọng tai nạn xảy ra nhiều nhất. Sau khi đã kéo thả xong, người dùng có thể xác định rõ ràng con đường và dữ liệu về tai nạn của con đường đó

**Câu 9:** Lấy ra 3 tiêu bang xảy ra nhiều vụ tai nạn nhất với mỗi quý trong năm 2020.

### Fact US Accidents Count by State



**Ý nghĩa:** Nhìn vào Pie Chart, ta thấy rằng top 3 tiểu bang có tổng số lượng vụ án lớn nhất vào năm 2020 là OR, FL, CA. Với cách thể hiện dữ liệu dưới dạng Pie Chart, người dùng có thể quan sát dữ liệu chi tiết hơn gồm: Tên tiểu bang - Số lượng vụ án - Phần trăm số lượng vụ án trên tổng 100%. Từ đó suy ra, tiểu bang nào chiếm nhiều phần trăm nhất là CA, ít phần trăm nhất là OR

**Câu 10:** Liệt kê chiều dài lớn nhất của đoạn đường và mức độ nghiêm trọng bị ảnh hưởng bởi các vụ tai nạn giao thông có tầm nhìn xa trên 2 dặm tại tiểu bang California (CA) và Washington (WA)

Start Time Year	CA	WA	Total
2	28544	545	<b>29089</b>
8	28165	588	<b>28753</b>
7	27024	621	<b>27645</b>
5	18829	445	<b>19274</b>
4	18735	352	<b>19087</b>
3	17513	353	<b>17866</b>
2020	188601	5243	<b>193844</b>
12	29256	507	<b>29763</b>
11	28685	410	<b>29095</b>
1	21353	412	<b>21765</b>
3	20789	514	<b>21303</b>
2	19307	247	<b>19554</b>
6	18399	1100	<b>19499</b>
4	16405	603	<b>17008</b>
5	16085	859	<b>16944</b>
10	12540	391	<b>12931</b>
9	5626	191	<b>5817</b>
8	103	9	<b>112</b>
7	53		<b>53</b>
2019	103928	4266	<b>108194</b>
10	25451	406	<b>25857</b>
<b>Total</b>	<b>770977</b>	<b>32181</b>	<b>803158</b>

**Ý nghĩa:** Nhìn vào hình trên, ta thấy được rằng dữ liệu đang được trực quan hóa dưới dạng Matrix để cho người đọc dễ dàng nắm bắt được các số liệu chính xác, với column là gồm các năm và tháng, rows là số lượng các vụ tai nạn của 2 tiểu bang CA, WA vào năm 2020. Cột total thể hiện tổng số lượng các vụ tai nạn của 2 thành phố. Từ đó người đọc có thể nhìn chi tiết đến bao quát tổng số lượng vụ tai nạn của 2 thành phố

## CHƯƠNG 5: KHAI THÁC DỮ LIỆU – DATA MINING

Từ bộ dữ liệu trên, nhóm chúng em tiến hành khai thác dữ liệu bằng các thuật toán sau

- **Thuật toán DBScan**

### Định nghĩa:

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) là một thuật toán phân cụm không giám sát trong lĩnh vực khai phá dữ liệu (data mining) và học máy (machine learning). Thuật toán này được sử dụng để phân cụm các điểm dữ liệu trong không gian dựa trên mật độ (density)

Ý tưởng chính của thuật toán DBSCAN là xác định các cụm dữ liệu dựa trên sự tập trung của các điểm trong không gian. Các cụm này được xác định dựa trên hai tham số quan trọng: bán kính  $\epsilon$  (epsilon) và số lượng điểm tối thiểu MinPts.

Cách hoạt động của DBSCAN như sau:

- + Chọn một điểm dữ liệu ngẫu nhiên chưa được gán vào cụm nào.
- + Xác định xem điểm này có đủ số lượng điểm láng giềng trong bán kính  $\epsilon$  (epsilon) hay không. Nếu có, điểm này sẽ được gán vào một cụm.
- + Mở rộng cụm bằng cách tìm kiếm và gán tất cả các điểm láng giềng của điểm đã được xác định vào cùng một cụm.
- + Lặp lại quá trình trên cho tất cả các điểm trong cụm vừa được mở rộng, cho đến khi không còn điểm nào có thể được thêm vào cụm hoặc không còn điểm dữ liệu chưa được gán vào cụm nào.

Đặc điểm quan trọng của DBSCAN là khả năng xác định các cụm có hình dạng và kích thước không đều, có thể xử lý các điểm nhiễu (noise) không thuộc vào bất kỳ cụm nào. Ngoài ra, DBSCAN cũng không yêu cầu sự chuẩn bị dữ liệu trước, chẳng hạn việc xác định số lượng cụm trước.

DBSCAN đã được áp dụng rộng rãi trong nhiều lĩnh vực, bao gồm xử lý hình ảnh, phân tích dữ liệu địa lý, phân tích tín hiệu, và nhiều ứng dụng khác trong lĩnh vực trí tuệ nhân tạo và khai phá dữ liệu.

## Bước 1: Import thư viện và đọc file

```
▼ Import libraries and read file

[50] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
    from sklearn.preprocessing import StandardScaler
    from sklearn.preprocessing import OneHotEncoder
    from sklearn.cluster import DBSCAN

[51] df = pd.read_csv("US_Accidents_Dec21_updated.csv")
    data = df
```

## Bước 2: Mô tả dữ liệu để xem từng cột có kiểu dữ liệu là gì sau đó kiểm tra trong mẫu dữ liệu có cột, hàng nào bị NULL hay không

```
[3] print(data.shape)
(2500, 23)

[4] data.head()

   Source  Severity      Start_Time        End_Time  Start_Lat  Start_Lng  Distance(mi)  County
0  Source1       1  2020-04-17 09:29:30  2020-04-17 10:29:30  26.706900 -80.119360      0.000  Palm Beach
1  Source1       2  2022-04-21 10:01:00.000000000  2022-04-21 11:44:08.000000000  38.781024 -121.265820      0.045  Placer
2  Source3       3  2016-08-12 16:45:00  2016-08-12 17:15:00  33.985249 -84.269348      0.000   Fulton
3  Source2       3  2019-09-20 15:22:16  2019-09-20 15:56:00  47.118706 -122.556908      0.000   Pierce
4  Source2       2  2019-06-03 16:55:43  2019-06-03 18:12:09  33.451355 -111.890343      0.000 Maricopa
```

5 rows × 23 columns

```

▼ Check missing values

▶ missing_values = data.isna().sum().sort_values(ascending=False)
missing_percentage = missing_values[missing_values!=0]/len(data)*100
print(" Percentage of Missing Values \n", missing_percentage)

▷ Percentage of Missing Values
End_Lat      100.00
End_Lng       100.00
Precipitation(in)   90.60
Wind_Chill(F)    77.00
Wind_Speed(mph)   8.35
Visibility(mi)    1.00
Weather_Condition 0.45
Humidity(X)      0.35
Temperature(F)    0.35
Wind_Direction     0.30
Weather_Timestamp  0.30
Pressure(in)       0.30
dtype: float64

▼ Getting List of Columns Having Null Values

[ ] null_cols = [i for i in data.columns if data[i].isnull().any()]
print(null_cols)

['End_Lat', 'End_Lng', 'Weather_Timestamp', 'Temperature(F)', 'Wind_Chill(F)', 'Humidity(X)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Direction', 'Wind_Speed(mph)', 'Precipitation(in)']

```

### Bước 3: Encoding data

One-hot encoding là một kỹ thuật được sử dụng để biểu diễn dữ liệu phân loại dưới dạng số. Thường được áp dụng trong machine learning và phân tích dữ liệu khi làm việc với các biến phân loại.

```

[7] #Encoding data
encoder = OneHotEncoder()
encoded_data = encoder.fit_transform(data)

[8] encoded_data.shape

(2500, 11399)

```

Sau đó, đếm số lần xuất hiện của mỗi nhãn cụm (cluster label) và tạo biểu đồ cột (bar chart) hiển thị phân phối của các nhãn cụm.

```

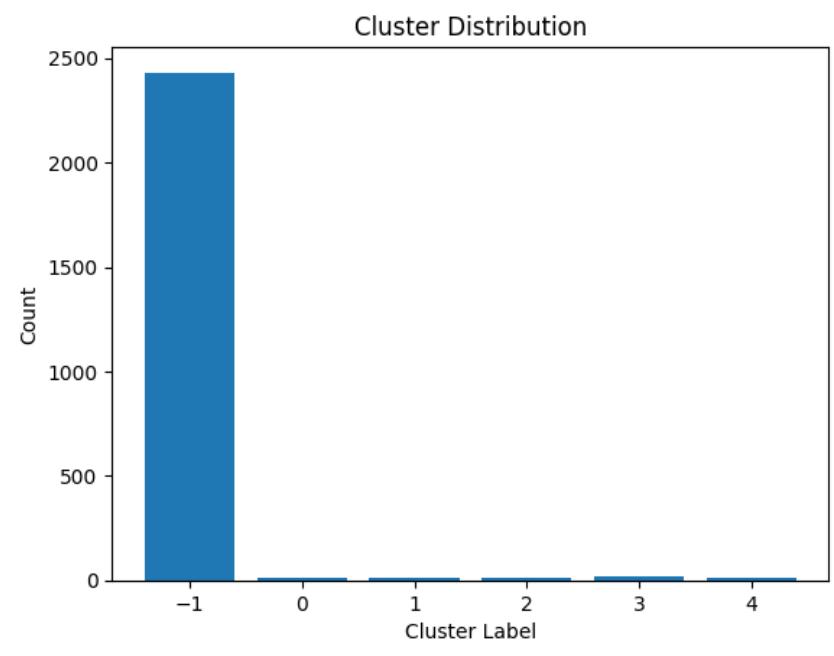
unique_labels, label_counts = np.unique(pred, return_counts=True)

# Cr Loading... chart of the cluster distribution
plt.bar(unique_labels, label_counts)

# Add labels and title
plt.xlabel('Cluster Label')
plt.ylabel('Count')
plt.title('Cluster Distribution')

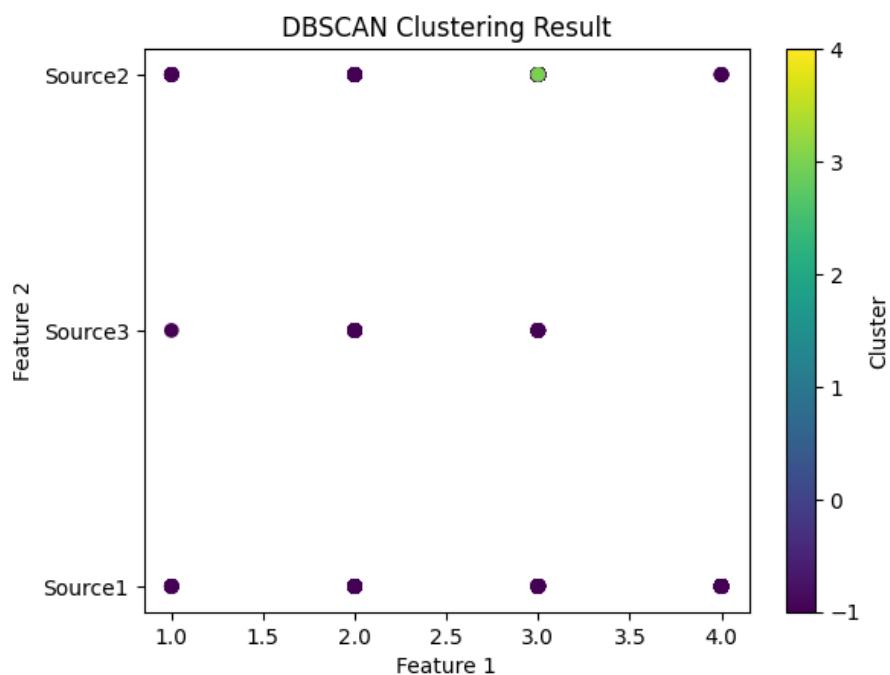
# Show the plot
plt.show()

```



Bước 4: Tạo biểu đồ phân tán (scatter plot) của các điểm dữ liệu và sử dụng màu sắc để đại diện cho nhãn cụm (cluster label) được dự đoán bằng thuật toán DBSCAN.

```
plt.scatter(data['Severity'], data['Source'], c = pred)
plt.xlabel('Feature 1')
plt.ylabel('Feature 2')
plt.title('DBSCAN Clustering Result')
plt.colorbar(label='Cluster')
plt.show()
```



- Thuật toán Mean Shift là một thuật toán phân cụm không giám sát trong lĩnh vực khai phá dữ liệu (data mining) và học máy (machine learning). Nó được sử dụng để tìm các cụm dữ liệu tự động dựa trên mật độ (density).
- Ý tưởng chính của thuật toán Mean Shift là tìm điểm cực đại của một hàm mật độ (density function) trong không gian đặc trưng (feature space). Điểm cực đại này thường tương ứng với trung tâm của một cụm dữ liệu. Thuật toán tiếp tục dịch chuyển điểm cực đại này trong không gian đặc trưng dựa trên mật độ, cho đến khi nó đạt đến một điểm cực đại mới hoặc hội tụ tại một điểm ổn định.
- **Các bước chính của thuật toán Mean Shift như sau:**
  - + Khởi tạo các điểm dữ liệu trong không gian đặc trưng.
  - + Đối với mỗi điểm dữ liệu, tính toán vectơ gradient của hàm mật độ tại điểm đó.
  - + Dịch chuyển mỗi điểm dữ liệu theo hướng của vectơ gradient, cập nhật vị trí mới
  - + Lặp lại các bước 2 và 3 cho đến khi điểm dữ liệu không còn thay đổi đáng kể hoặc đạt đến một điểm cực đại.
  - + Gom nhóm các điểm dữ liệu dựa trên các điểm cực đại mà thuật toán tìm được.

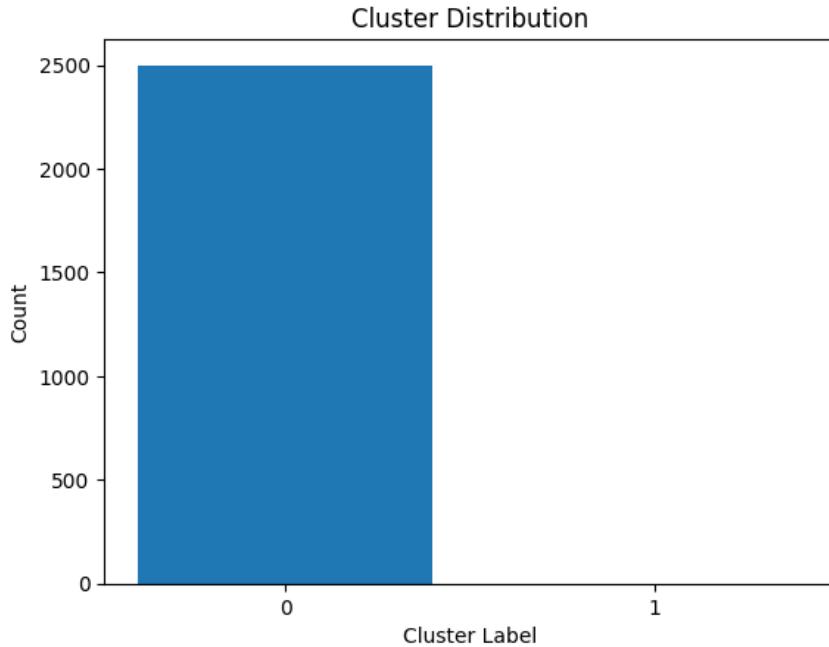
Thuật toán Mean Shift có thể xử lý hiệu quả các cụm dữ liệu có hình dạng và kích thước không đều, và không yêu cầu thông số cụ thể về số lượng cụm. Nó đã được áp dụng trong nhiều lĩnh vực, bao gồm phân tích hình ảnh, phân tích dữ liệu địa lý, nhận dạng đối tượng, và nhiều ứng dụng khác trong lĩnh vực trí tuệ nhân tạo và khai phá dữ liệu thuật toán Mean Shift

#### **Bước 6: Tạo biểu đồ phân tán đếm số lần xuất hiện của mỗi nhãn cụm (cluster label) và tạo biểu đồ cột (bar chart) hiển thị phân phối của các nhãn cụm kết quả của thuật toán Mean Shift**

```
# Encoding data using OneHotEncoder
encoder = OneHotEncoder()
encoded_data = encoder.fit_transform(data)

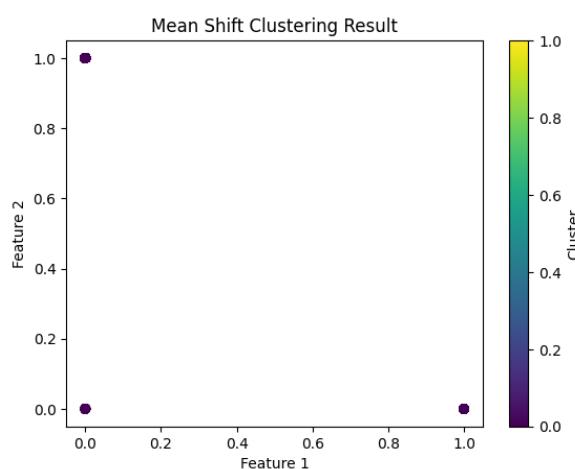
# Converting encoded_data to a numpy array
encoded_data_array = encoded_data.toarray()

# Applying Mean Shift clustering
clustering = MeanShift().fit(encoded_data_array)
```



**Bước 7: Tạo biểu đồ phân tán (scatter plot) của các điểm dữ liệu và sử dụng màu sắc để đại diện cho nhãn cụm (cluster label) được dự đoán bằng thuật toán Mean Shift.**

```
# Create a scatter plot of the encoded data with color-coded clusters
plt.scatter(encoded_data_array[:, 0], encoded_data_array[:, 1], c=cluster_labels)
plt.xlabel('Feature 1')
plt.ylabel('Feature 2')
plt.title('Mean Shift Clustering Result')
plt.colorbar(label='Cluster')
plt.show()
```



data.describe().T								
	count	mean	std	min	25%	50%	75%	max
<b>Severity</b>	2500.0	2.217600	0.484917	1.000000	2.000000	2.000000	2.000000	4.000000
<b>Start_Lat</b>	2500.0	36.216774	5.039209	25.232961	33.411743	36.006747	40.112700	48.596352
<b>Start_Lng</b>	2500.0	-94.413786	17.402864	-124.384621	-117.157848	-87.357115	-80.354199	-70.119828
<b>Distance(mi)</b>	2500.0	0.599786	1.886305	0.000000	0.000000	0.032000	0.495500	55.312000

**Nhận xét:** Để dàng nhận thấy khi mô tả dữ liệu bằng hàm describe() thì cột dữ liệu “Severity” đều có giá trị bằng 2 do đó nó ảnh hưởng trực tiếp đến quá trình phân cụm (sẽ luôn có 1 cụm với giá trị Severity = 2 lớn nhất)

- Với thuật toán DBSCAN, sau khi chạy thực nghiệm sẽ có 6 cụm, trong đó cụm (-1) là chiếm nhiều nhất với giá trị Severity = 2.

Nhìn vào biểu đồ phân tán, đa số các giá trị đều là -1, chỉ có 1 giá trị là 4

- Với thuật toán Mean Shift, sau khi chạy thực nghiệm sẽ có 2 cụm, trong đó cụm 0 với Severity = 2, thông trại hầu hết các cụm còn lại

Nhìn vào biểu đồ phân tán, tất cả các giá trị đều là 0