ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA HỆ THỐNG THÔNG TIN

BÁO CÁO ĐỒ ÁN

KHO DỮ LIỆU VÀ OLAP

PHÂN TÍCH DỮ LIỆU TAI NẠN GIAO THÔNG Ở CHICAGO

GVHD: Đỗ Thị Minh Phụng

Lớp: IS217.O13

Thành viên nhóm: 22521704 – Lê Châu Khánh Vy

22521166 – Nguyễn Hạ Phương

Thành Phố Hồ Chí Minh, 09/2024

MỤC LỤC

MỤC LỤC **1**
LỜI CẢM ƠN **4**
NHẬN XÉT CỦA GIẢNG VIÊN **5**
LÝ DO CHỌN ĐỀ TÀI **6**
CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ DỮ LIỆU **7**
1.1 Phát biểu về dữ liệu: 7
1.1.1 Mô tả dữ liệu *7*
1.1.2 Tiền xử lý: *7*
1.1.3 Thuộc tính của dữ liệu: *16*
1.1.4 Hướng chủ đề: *17*
1.2 XÂY DỰNG KHO DỮ LIỆU: 18
1.2.1 Thiết kế lược đồ: *18*
1.2.2 Mô tả chi tiết các bảng: *18*
1.3 CÁC CÂU TRUY VẤN: 22
CHƯƠNG 2: TÍCH HỢP DỮ LIỆU VÀO KHO DỮ LIỆU (SSIS) **23**
2.1 Chuẩn bị công cụ và cơ sở dữ liệu 23
2.1.1 Chuẩn bị công cụ *23*
2.1.2 Chuẩn bị cơ sở dữ liệu *24*
2.2 Tạo project SSIS 25
2.3 Tạo các bảng Dim và bảng Fact 27
2.3.1 Tạo bảng Dim_Time *36*
2.3.2 Tạo bảng Dim_ LightingConditions *46*
2.3.3 Tạo bảng Dim_TrafficControl *50*
2.3.4 Tạo bảng Dim_Street *55*
2.3.5 Tạo bảng Dim_Damages *59*
2.3.6 Tạo bảng Dim_ FirstCrashTypes *62*
2.3.7 Tạo bảng Dim_WeatherConditions *64*
2.3.8 Tạo bảng Dim_RoadAlignment *68*
2.3.9 Tạo bảng Dim_PrimaryCauses *71*
2.3.10 Tạo bảng Dim_LS *74*
2.3.11 Tạo bảng Fact_Raw *77*
2.3.12 Tạo bảng Dim_LocationStreets *82*
2.3.13 Tạo bảng Fact_TrafficCrashRecords *91*
2.3.14 Tạo khóa ngoại từ bảng Fact đến các Dimension. *196*
2.4 Chạy dự án SSIS 206
2.5 Kiểm tra dữ liệu các bảng 209
2.6 Lược đồ cơ sở dữ liệu sau khi hoàn thành 213
CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU TRỰC TUYẾN (SSAS) **215**

3.1 Chuẩn bị các công cụ 215
3.1.1 Cài đặt Microsoft Analysis Services Projects 215
3.1.2 Cài đặt Analysis Services 216
3.2 Tạo mới Project SSAS 226
3.3 Xác định dữ liệu nguồn (Data Sources) 229
3.4 Xác định khung nhìn dữ liệu nguồn (Data Source Views) 234
3.5 Xây dựng các khối (Cube) và deploy Cube 238
3.5.1 Tạo Cube và Dimension 238
3.5.2 Thêm thuộc tính vào Dimension 242
3.5.3 Deploy Analysis Services Project 247
3.5.4 Tạo mới các Measures 251
3.5.5 Tạo Hierachies và định nghĩa các Attribute Relationship 255
3.6 Thực hiện 15 câu truy vấn (MDX, Manual và Pivot Excel) 267
3.6.1 Câu truy vấn 1 267
3.6.2 Câu truy vấn 2 271
3.6.3 Câu truy vấn 3 275
3.6.4 Câu truy vấn 4 278
3.6.5 Câu truy vấn 5 283
3.6.6 Câu truy vấn 6 287
3.6.7 Câu truy vấn 7 291
3.6.8 Câu truy vấn 8 299
3.6.9 Câu truy vấn 9 302
3.6.10 Câu truy vấn 10 306
3.6.11 Câu truy vấn 11 310
3.6.12 Câu truy vấn 12 315
3.6.13 Câu truy vấn 13 320
3.6.14 Câu truy vấn 14 325
3.6.15 Câu truy vấn 15 329
CHƯƠNG 4: QUÁ TRÌNH TẠO BÁO CÁO DỮ LIỆU(SSRS) 334
4.1 Tạo project SSRS trên Power BI 334
4.1.1 Chuẩn bị công cụ Power BI 334
4.1.2 Thực hiện tạo báo cáo trên Power BI 334
4.1.3 Public các report đã tạo 350
4.2 Tạo project SSRS trên Locker 353
4.2.1 Trích xuất dữ liệu từ kho dữ liệu dưới dạng .csv 353
4.2.2 Upload và lọc dữ liệu trên Locker Studio 355
4.2.3 Thực hiện tạo báo cáo trên Locker Studio 359
CHƯƠNG 5: QUÁ TRÌNH KHAI THÁC DỮ LIỆU (DATA MINING) 369
5.1 Mô tả bộ dữ liệu 369

5.1.1 Tổng quan dữ liệu *369*
5.1.2 Mô tả bộ dữ liệu *369*
5.2 Phân tích tổng quan bộ dữ liệu 371
5.2.1 Thống kê mô tả *371*
5.2.2 Kiểm tra dữ liệu ban đầu *373*
5.3 Khám phá dữ liệu 376
5.3.1 Loại bỏ thuộc tính không cần thiết *377*
5.3.2 Lựa chọn thuộc tính *378*
5.3.3 Tổng kết *396*
5.4 Thuật toán cho mô hình 397
5.4.1 Cơ sở lý thuyết *397*
5.4.2 Thuật toán Logistic Regression *402*
5.4.3 Thuật toán Random Forest *405*
5.4.4 Thuật toán XGBoost *407*
5.4.5 Thuật toán LightGBM *413*
5.4.6 Tổng kết *418*
TÀI LIỆU THAM KHẢO 422

LỜI CẢM ƠN

Lời đầu tiên, nhóm em xin bày tỏ lòng biết ơn chân thành tới toàn thể quý Cô và Thầy tại Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM, cũng như quý Thầy Cô thuộc Khoa Hệ thống thông tin. Những kiến thức mà chúng em đã học từ quý Thầy Cô là nền tảng quan trọng, giúp chúng em hiểu sâu hơn về môn học này.

Nhóm em cũng xin chân thành cảm ơn Cô Đỗ Thị Minh Phụng đã hỗ trợ chúng em suốt hành trình làm đồ án. Sự nhiệt huyết và hướng dẫn tận tình của Cô đã là động lực lớn lao giúp chúng em vượt qua những khó khăn trong nghiên cứu. Cô đã tạo điều kiện và cơ hội để nhóm có thể trải nghiệm môi trường học tập từ những kinh nghiệm, kiến thức và tư duy chuyên môn của cô, tất cả những điều đó đều rất quý giá với chúng em.

Dưới sự hướng dẫn của Cô, chúng em đã có cơ hội áp dụng những kiến thức đã học vào thực tế. Nhóm em nhận thức rằng có nhiều điểm cần được hoàn thiện. Chúng em trân trọng sự góp ý và hướng dẫn của Cô để từ những thiếu sót, chúng em có thể rút kinh nghiệm và phát triển.

Chúng em hi vọng rằng những kiến thức và kinh nghiệm mình học được sẽ là bước đệm vững chắc cho tương lai học tập và làm việc của mình.

Trân trọng cảm ơn!

NHẬN XÉT CỦA GIẢNG VIÊN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

LÝ DO CHỌN ĐỀ TÀI

Trong bối cảnh đô thị hóa ngày càng gia tăng, tai nạn giao thông đã trở thành một vấn đề nghiêm trọng tại nhiều thành phố lớn. Những vụ tai nạn không chỉ gây thiệt hại về người và tài sản mà còn ảnh hưởng đến chất lượng cuộc sống của cư dân. Tai nạn giao thông có sự tác động của nhiều yếu tố từ yếu tố khách quan như điều kiện đường xá, thời tiết, môi trường, ánh sáng và sự hoạt động của các thiết bị điều khiển giao thông đến yếu tố chủ quan như các lỗi vi phạm của người lưu thông.

Với sự phát triển của công nghệ thông tin, việc thu thập và phân tích dữ liệu về tai nạn giao thông ngày càng trở nên khả thi và cần thiết. Tuy nhiên, việc quản lý thông tin và phân tích dữ liệu về tai nạn giao thông vẫn gặp nhiều thách thức, từ việc xác định nguyên nhân đến việc phát hiện các mô hình và xu hướng.

Chính vì vậy, nghiên cứu về phân tích dữ liệu về tác nhân và thiệt hại của tai nạn giao thông không chỉ nhằm mục tiêu cung cấp thực trạng về tình hình an toàn giao thông mà còn góp phần đề xuất các giải pháp cải thiện hệ thống giao thông trong khu vực. Thông qua việc **xây dựng và khai thác một kho dữ liệu tai nạn giao thông tại Chicago**, chúng em hy vọng sẽ mang lại thông tin hữu ích cho các nhà quản lý, giúp họ đưa ra các quyết định hiệu quả nhằm giảm thiểu tai nạn và nâng cao an toàn cho cộng đồng. Đó chính là lý do nhóm em quyết định chọn đề tài này.

CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ DỮ LIỆU

1.1 Phát biểu về dữ liệu:

1.1.1 Mô tả dữ liệu

- Tên bộ dữ liệu: **Chicago Traffic Crashes**

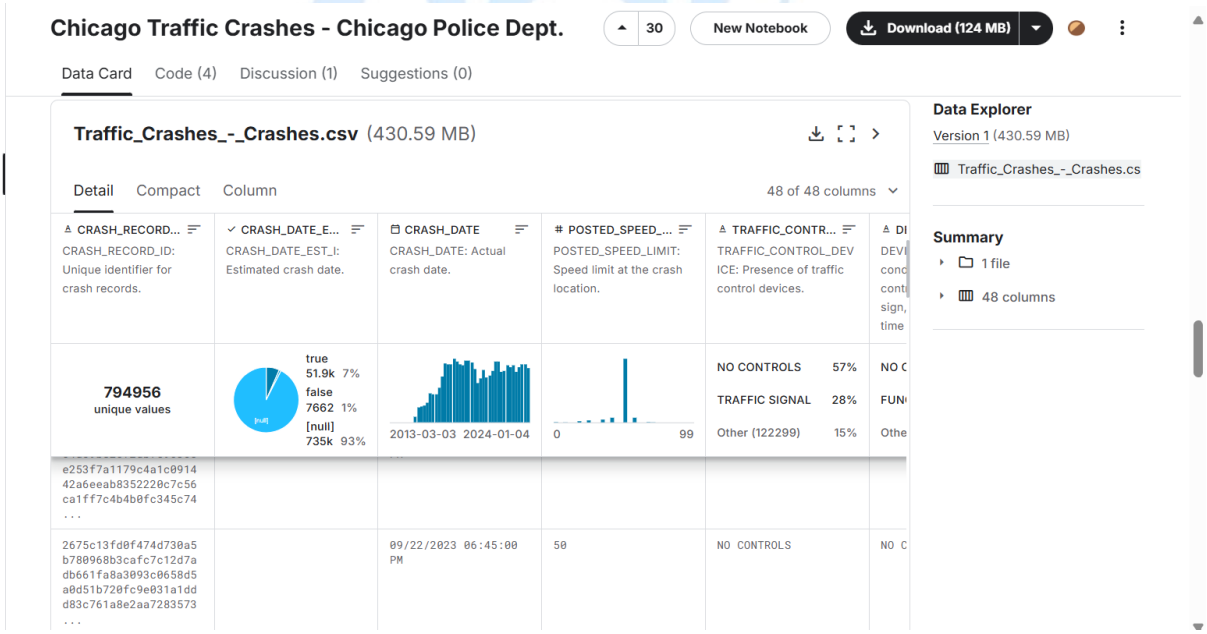
- Tác giả: **Anoop Johny**

- Bộ dữ liệu của ông chứa thông tin về các vụ tai nạn giao thông trên đường phố trong giới hạn Thành phố Chicago, thuộc thẩm quyền của Sở Cảnh sát Chicago (CPD). Dữ liệu có nguồn gốc từ hệ thống báo cáo sự cố điện tử (E-Crash) tại CPD, không bao gồm bất kỳ thông tin nhận dạng cá nhân nào. Các hồ sơ bao gồm chi tiết về ngày, giờ, địa điểm, điều kiện và các yếu tố góp phần.

- Dữ liệu được thu thập từ ngày 03/03/2013. Tập dữ liệu này được làm mới hàng ngày với tất cả những thay đổi mới nhất, đảm bảo có dữ liệu mới nhất. Ngày cập nhật gần nhất: 05/01/2024

- Kho dữ liệu gồm 794956 dòng và 48 thuộc tính

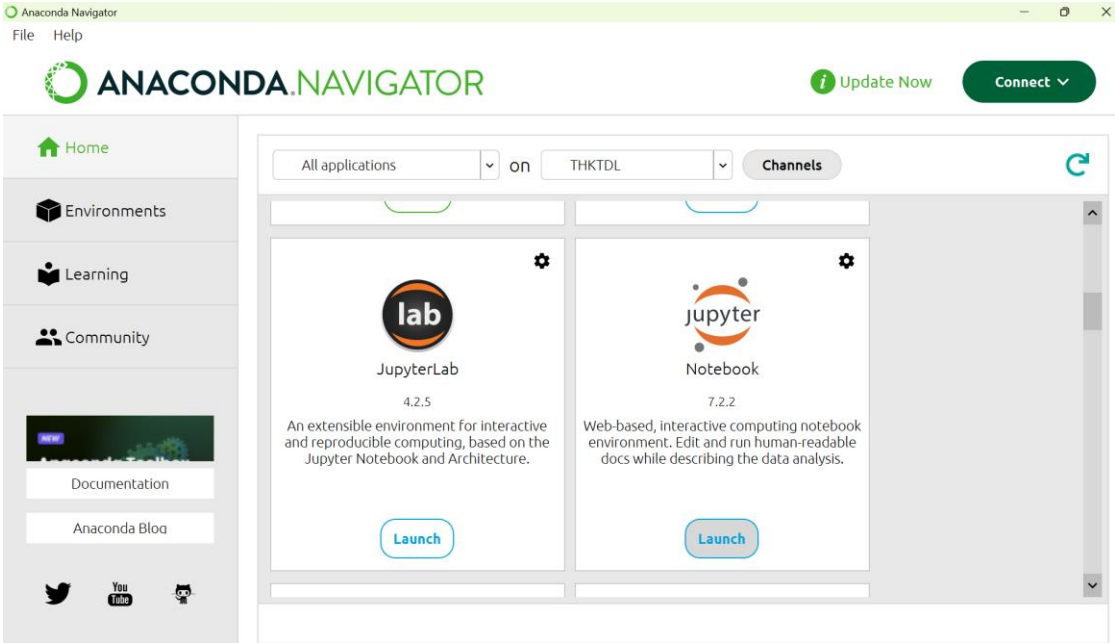
- Link dataset: https://www.kaggle.com/datasets/anoopjohny/traffic-crashes-crashes/data



1.1.2 Tiền xử lý:

1.1.2.2 Công cụ sử dụng:

- Cài đặt Notebook của Anaconda Navigator.



1.1.2.3 Cài đặt:

Download bộ dữ liệu và tiến hành làm sạch dữ liệu bằng ngôn ngữ Python qua các bước:

- Tiến hành import thư viện và đọc file csv.

```
[1]: #import necessary libraries
     import numpy as np
     import pandas as pd
```

```
[7]: #Load the dataset
     df = pd.read_csv('Traffic_Crashes_-_Crashes.csv')
     df
```

[7]:

	CRASH_RECORD_ID	CRASH_DATE_EST_I	CRASH_DATE	POSTED_SPEED_LIMIT	TRAFFIC_CONTROL_DEVICE	DEVICE_CONDITION
0	23a79931ef555d54118f64dc9be2cf2dbf59636ce253f7...	NaN	09/05/2023 07:05:00 PM	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY
1	2675c13fd0f474d730a5b780968b3cafc7c12d7adb661f...	NaN	09/22/2023 06:45:00 PM	50	NO CONTROLS	NO CONTROLS
2	5f54a59fcb087b12ae5b1acff96a3caf4f2d37e79f8db4...	NaN	07/29/2023 02:45:00 PM	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY
3	7ebf015016f83d09b321afd671a836d6b148330535d5df...	NaN	08/09/2023 11:00:00 PM	30	NO CONTROLS	NO CONTROLS
4	6c1659069e9c6285a650e70d6f9b574ed5f64c12888479...	NaN	08/18/2023 12:50:00 PM	15	OTHER	FUNCTIONING PROPERLY
...	...	...	...	...	...	...
794951	b4944ab32a9f692cf95c2645b6683f032e74932532947d...	NaN	11/06/2023 07:00:00 AM	30	NO CONTROLS	NO CONTROLS
794952	d4bca3b24c17b9af828558fc95699da3c57fe7ea07daf5...	NaN	10/31/2023 06:30:00 PM	30	NO CONTROLS	NO CONTROLS
794953	ef84ea51375f9e29d55d99fb2b1dfaf8ea9d1fbc0f667e...	NaN	11/04/2023 11:00:00 PM	30	NO CONTROLS	NO CONTROLS
794954	9ba7b0e570841fc865b62892fb4931684ce4abe8162ee1...	NaN	11/06/2023 06:35:00 PM	30	NO CONTROLS	NO CONTROLS
794955	b2bd97994325bbc4c059742f1d87453b6364e6eb245809...	NaN	11/06/2023 07:00:00 PM	30	NO CONTROLS	NO CONTROLS

794956 rows × 48 columns

1.1.2.4 Một số bước tiền xử lý:

Bước 1: Đọc, tìm hiểu và phân tích bộ dữ liệu trước khi xử lý.

- ❖ Xem các thông tin cơ bản của bộ dữ liệu:

```
[38]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 794956 entries, 0 to 794955
Data columns (total 48 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   CRASH_RECORD_ID               794956 non-null  object
 1   CRASH_DATE_EST_I              59606 non-null   object
 2   CRASH_DATE                    794956 non-null  object
 3   POSTED_SPEED_LIMIT            794956 non-null  int64
 4   TRAFFIC_CONTROL_DEVICE        794956 non-null  object
 5   DEVICE_CONDITION              794956 non-null  object
 6   WEATHER_CONDITION             794956 non-null  object
 7   LIGHTING_CONDITION            794956 non-null  object
 8   FIRST_CRASH_TYPE              794956 non-null  object
 9   TRAFFICWAY_TYPE               794956 non-null  object
 10  LANE_CNT                      199006 non-null  float64
 11  ALIGNMENT                     794956 non-null  object
 12  ROADWAY_SURFACE_COND          794956 non-null  object
 13  ROAD_DEFECT                   794956 non-null  object
 14  REPORT_TYPE                   771698 non-null  object
 15  CRASH_TYPE                    794956 non-null  object
 16  INTERSECTION_RELATED_I        182215 non-null  object
 17  NOT_RIGHT_OF_WAY_I            36727 non-null   object
 18  HIT_AND_RUN_I                 248626 non-null  object
 19  DAMAGE                        794956 non-null  object
 20  DATE_POLICE_NOTIFIED          794956 non-null  object
 21  PRIM_CONTRIBUTORY_CAUSE       794956 non-null  object
 22  SEC_CONTRIBUTORY_CAUSE        794956 non-null  object
 23  STREET_NO                     794956 non-null  int64
 24  STREET_DIRECTION              794952 non-null  object
 25  STREET_NAME                   794955 non-null  object
 26  BEAT_OF_OCCURRENCE            794951 non-null  float64
 27  PHOTOS_TAKEN_I                10371 non-null   object
 28  STATEMENTS_TAKEN_I            17557 non-null   object
 29  DOORING_I                     2461 non-null    object
 30  WORK_ZONE_I                   4603 non-null    object
 31  WORK_ZONE_TYPE                3574 non-null    object
 32  WORKERS_PRESENT_I             1178 non-null    object
 33  NUM_UNITS                     794956 non-null  int64
 34  MOST_SEVERE_INJURY            793204 non-null  object
 35  INJURIES_TOTAL                793215 non-null  float64
 36  INJURIES_FATAL                793215 non-null  float64
 37  INJURIES_INCAPACITATING       793215 non-null  float64
 38  INJURIES_NON_INCAPACITATING   793215 non-null  float64
 39  INJURIES_REPORTED_NOT_EVIDENT 793215 non-null  float64
 40  INJURIES_NO_INDICATION        793215 non-null  float64
 41  INJURIES_UNKNOWN              793215 non-null  float64
 42  CRASH_HOUR                    794956 non-null  int64
 43  CRASH_DAY_OF_WEEK             794956 non-null  int64
 44  CRASH_MONTH                   794956 non-null  int64
 45  LATITUDE                      789567 non-null  float64
 46  LONGITUDE                     789567 non-null  float64
 47  LOCATION                      789567 non-null  object
dtypes: float64(11), int64(6), object(31)
memory usage: 291.1+ MB
```

❖ Kiểm tra giá trị null.

```
[43]: #check for missing values
print("\nMissing Values:")
print(df.isnull().sum())
```

```
Missing Values:
CRASH_RECORD_ID                    0
CRASH_DATE                         0
POSTED_SPEED_LIMIT                 0
TRAFFIC_CONTROL_DEVICE             0
DEVICE_CONDITION                   0
WEATHER_CONDITION                  0
LIGHTING_CONDITION                 0
FIRST_CRASH_TYPE                   0
TRAFFICWAY_TYPE                    0
ALIGNMENT                          0
DAMAGE                             0
PRIM_CONTRIBUTORY_CAUSE            0
STREET_NO                          0
STREET_DIRECTION                   4
STREET_NAME                        1
NUM_UNITS                          0
INJURIES_TOTAL                  1725
INJURIES_FATAL                  1725
INJURIES_INCAPACITATING         1725
INJURIES_NON_INCAPACITATING     1725
INJURIES_NO_INDICATION          1725
CRASH_HOUR                         0
CRASH_DAY_OF_WEEK                  0
CRASH_MONTH                        0
dtype: int64
```

Bước 2: Xóa các cột không sử dụng.

```
[39]: #delete unused columns
df = df.drop(columns= ['CRASH_DATE_EST_I'])
df = df.drop(columns = ['LANE_CNT'])
df = df.drop(columns = ['ROADWAY_SURFACE_COND'])
df = df.drop(columns = ['ROAD_DEFECT'])
df = df.drop(columns = ['REPORT_TYPE'])

df = df.drop(columns = ['CRASH_TYPE'])
df = df.drop(columns = ['INTERSECTION_RELATED_I'])
df = df.drop(columns = ['NOT_RIGHT_OF_WAY_I'])
df = df.drop(columns = ['HIT_AND_RUN_I'])
df = df.drop(columns = ['SEC_CONTRIBUTORY_CAUSE'])

df = df.drop(columns = ['DATE_POLICE_NOTIFIED'])
df = df.drop(columns = ['BEAT_OF_OCCURRENCE'])
df = df.drop(columns = ['PHOTOS_TAKEN_I'])
df = df.drop(columns = ['STATEMENTS_TAKEN_I'])
df = df.drop(columns = ['DOORING_I'])

df = df.drop(columns = ['WORK_ZONE_I'])
df = df.drop(columns = ['WORK_ZONE_TYPE'])
df = df.drop(columns = ['WORKERS_PRESENT_I'])
df = df.drop(columns = ['INJURIES_REPORTED_NOT_EVIDENT'])
df = df.drop(columns = ['INJURIES_UNKNOWN'])

df = df.drop(columns = ['MOST_SEVERE_INJURY'])
df = df.drop(columns = ['LATITUDE'])
df = df.drop(columns = ['LONGITUDE'])
df = df.drop(columns = ['LOCATION'])
```

Bước 3: Chỉ định rõ định dạng của ngày tháng bằng cách sử dụng tham số format trong hàm pd.to_datetime(). Xóa các dòng có năm xảy ra tai nạn = {2013, 2014, 2024} vì dữ liệu ít, không thể dùng để thống kê và đánh giá số liệu trong một năm.

```
[41]: # Specify an explicit date format
df['CRASH_DATE'] = pd.to_datetime(df['CRASH_DATE'], format='%m/%d/%Y %I:%M:%S %p')
#Delete year 2013, 2014, 2024
df = df[~df['CRASH_DATE'].dt.year.isin([2013, 2014, 2015, 2024])]
```

Bước 4: Kiểm tra độ dài của giá trị dài nhất trong mỗi cột đã liệt kê và liệt kê các giá trị khác nhau trong cột đó.

[44]:

```python
columns_to_check = [
    'TRAFFIC_CONTROL_DEVICE',
    'DEVICE_CONDITION',
    'WEATHER_CONDITION',
    'LIGHTING_CONDITION',
    'FIRST_CRASH_TYPE',
    'TRAFFICWAY_TYPE',
    'ALIGNMENT',
    'DAMAGE',
    'PRIM_CONTRIBUTORY_CAUSE'
]

for column in columns_to_check:
    print(f"Thuộc tính: {column}")
    # Calculate the length of the longest value
    max_length = df[column].apply(len).max()
    print(f"Độ dài giá trị dài nhất: {max_length}")
    # Lists the different values in the column
    print(f"Các giá trị khác nhau: {df[column].unique()}")
    print("\n")
```

```
Thuộc tính: TRAFFIC_CONTROL_DEVICE
Độ dài giá trị dài nhất: 24
Các giá trị khác nhau: ['TRAFFIC SIGNAL' 'NO CONTROLS' 'OTHER' 'UNKNOWN' 'OTHER WARNING SIGN'
 'STOP SIGN/FLASHER' 'PEDESTRIAN CROSSING SIGN' 'OTHER REG. SIGN' 'YIELD'
 'LANE USE MARKING' 'RAILROAD CROSSING GATE' 'FLASHING CONTROL SIGNAL'
 'SCHOOL ZONE' 'POLICE/FLAGMAN' 'DELINEATORS' 'OTHER RAILROAD CROSSING'
 'RR CROSSING SIGN' 'NO PASSING' 'BICYCLE CROSSING SIGN']


Thuộc tính: DEVICE_CONDITION
Độ dài giá trị dài nhất: 24
Các giá trị khác nhau: ['FUNCTIONING PROPERLY' 'NO CONTROLS' 'FUNCTIONING IMPROPERLY' 'UNKNOWN'
 'OTHER' 'NOT FUNCTIONING' 'MISSING' 'WORN REFLECTIVE MATERIAL']


Thuộc tính: WEATHER_CONDITION
Độ dài giá trị dài nhất: 24
Các giá trị khác nhau: ['CLEAR' 'SNOW' 'RAIN' 'UNKNOWN' 'CLOUDY/OVERCAST' 'FOG/SMOKE/HAZE'
 'BLOWING SNOW' 'FREEZING RAIN/DRIZZLE' 'OTHER' 'SEVERE CROSS WIND GATE'
 'SLEET/HAIL' 'BLOWING SAND, SOIL, DIRT']


Thuộc tính: LIGHTING_CONDITION
Độ dài giá trị dài nhất: 22
Các giá trị khác nhau: ['DUSK' 'DARKNESS, LIGHTED ROAD' 'DAYLIGHT' 'DARKNESS' 'UNKNOWN' 'DAWN']


Thuộc tính: FIRST_CRASH_TYPE
Độ dài giá trị dài nhất: 28
Các giá trị khác nhau: ['ANGLE' 'REAR END' 'PARKED MOTOR VEHICLE' 'SIDESWIPE SAME DIRECTION'
 'PEDESTRIAN' 'FIXED OBJECT' 'TURNING' 'SIDESWIPE OPPOSITE DIRECTION'
 'REAR TO FRONT' 'HEAD ON' 'REAR TO SIDE' 'PEDALCYCLIST' 'OTHER OBJECT'
 'ANIMAL' 'REAR TO REAR' 'OTHER NONCOLLISION' 'OVERTURNED' 'TRAIN']


Thuộc tính: TRAFFICWAY_TYPE
Độ dài giá trị dài nhất: 31
Các giá trị khác nhau: ['FIVE POINT, OR MORE' 'DIVIDED - W/MEDIAN BARRIER'
 'DIVIDED - W/MEDIAN (NOT RAISED)' 'NOT DIVIDED' 'OTHER' 'ONE-WAY'
 'PARKING LOT' 'T-INTERSECTION' 'RAMP' 'FOUR WAY' 'UNKNOWN' 'ALLEY'
 'UNKNOWN INTERSECTION TYPE' 'DRIVEWAY' 'TRAFFIC ROUTE' 'NOT REPORTED'
 'CENTER TURN LANE' 'L-INTERSECTION' 'Y-INTERSECTION' 'ROUNDABOUT']

Thuộc tính: ALIGNMENT
Độ dài giá trị dài nhất: 21
Các giá trị khác nhau: ['STRAIGHT AND LEVEL' 'CURVE ON GRADE' 'CURVE, LEVEL' 'STRAIGHT ON GRADE'
 'STRAIGHT ON HILLCREST' 'CURVE ON HILLCREST']


Thuộc tính: DAMAGE
Độ dài giá trị dài nhất: 13
Các giá trị khác nhau: ['OVER $1,500' '$501 - $1,500' '$500 OR LESS']


Thuộc tính: PRIM_CONTRIBUTORY_CAUSE
Độ dài giá trị dài nhất: 80
Các giá trị khác nhau: ['UNABLE TO DETERMINE' 'FOLLOWING TOO CLOSELY'
 'FAILING TO REDUCE SPEED TO AVOID CRASH' 'FAILING TO YIELD RIGHT-OF-WAY'
 'NOT APPLICABLE' 'WEATHER' 'IMPROPER BACKING'
 'IMPROPER TURNING/NO SIGNAL' 'DRIVING SKILLS/KNOWLEDGE/EXPERIENCE'
 'IMPROPER LANE USAGE'
 'VISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.)'
 'ROAD ENGINEERING/SURFACE/MARKING DEFECTS'
 'EQUIPMENT - VEHICLE CONDITION' 'IMPROPER OVERTAKING/PASSING'
 'RELATED TO BUS STOP' 'DISREGARDING TRAFFIC SIGNALS'
 'DRIVING ON WRONG SIDE/WRONG WAY' 'DISREGARDING ROAD MARKINGS'
 'DISTRACTION - FROM INSIDE VEHICLE' 'ANIMAL'
 'ROAD CONSTRUCTION/MAINTENANCE' 'TEXTING'
 'CELL PHONE USE OTHER THAN TEXTING' 'DISREGARDING OTHER TRAFFIC SIGNS'
 'DISREGARDING STOP SIGN' 'EXCEEDING AUTHORIZED SPEED LIMIT'
 'OPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER'
 'DISTRACTION - FROM OUTSIDE VEHICLE' 'PHYSICAL CONDITION OF DRIVER'
 'EXCEEDING SAFE SPEED FOR CONDITIONS' 'DISREGARDING YIELD SIGN'
 'TURNING RIGHT ON RED'
 'UNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED)'
 'EVASIVE ACTION DUE TO ANIMAL, OBJECT, NONMOTORIST'
 'HAD BEEN DRINKING (USE WHEN ARREST IS NOT MADE)'
 'DISTRACTION - OTHER ELECTRONIC DEVICE (NAVIGATION DEVICE, DVD PLAYER, ETC.)'
 'OBSTRUCTED CROSSWALKS' 'BICYCLE ADVANCING LEGALLY ON RED LIGHT'
 'PASSING STOPPED SCHOOL BUS' 'MOTORCYCLE ADVANCING LEGALLY ON RED LIGHT']
```

Bước 5: Lặp qua mỗi cột thuộc tính, thay ký tự , thành ; giúp bảo toàn tính toàn vẹn của dữ liệu khi đưa vào Flat File Source, vì các giá trị trong các cột sẽ không bị chia tách thành nhiều cột khác nhau.

```
[45]: # Replace , to ;
      for col in df.columns:
          if df[col].dtype == object:
              df[col] = df[col].str.replace(',', ';')
```

Bước 6: Xóa các kí tự '$' trong bảng DAMAGE, đổi kí tự ';' thành '.'.

```
[46]: # Delete $ in column 'DAMAGE', replace ; to .
      df['DAMAGE'] = df['DAMAGE'].str.replace('$', '', regex = False)
      df['DAMAGE'] = df['DAMAGE'].str.replace(';', '.', regex = False)
```

Bước 7: Kiểm tra và xóa các dòng dữ liệu null.

```
[65]: # Delete rows have value null
      df.dropna(inplace=True)
```

Bước 8: Sắp xếp và khởi tạo lại giá trị (kiểu int tăng dần theo thời gian) cho thuộc tính CRASH_RECORD_ID.

```
[56]: df.sort_values(by='CRASH_DATE', ascending=True, inplace=True)
```

```
[60]: total_rows = len(df)

      # Create a sequence of numbers from 1 to total_rows
      df['CRASH_RECORD_ID'] = range(1, total_rows + 1)
```

Kết quả:

[64]:

	CRASH_RECORD_ID	CRASH_DATE	POSTED_SPEED_LIMIT	TRAFFIC_CONTROL_DEVICE	DEVICE_CONDITION	WEATHER_CONDITION	LIGHTING_CONDITION	FIRST
655591	1	2016-01-01 00:01:00	30	NO CONTROLS	NO CONTROLS	CLEAR	DARKNESS; LIGHTED ROAD	
480983	2	2016-01-01 00:08:00	30	NO CONTROLS	NO CONTROLS	CLEAR	DARKNESS; LIGHTED ROAD	
598874	3	2016-01-01 00:15:00	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	CLEAR	DARKNESS; LIGHTED ROAD	
781860	4	2016-01-01 00:36:00	30	UNKNOWN	UNKNOWN	CLEAR	DARKNESS; LIGHTED ROAD	SI
637778	5	2016-01-01 00:55:00	35	NO CONTROLS	NO CONTROLS	CLEAR	DARKNESS; LIGHTED ROAD	
...	...	...	...	...	...	...	...	
25001	782877	2023-12-31 23:25:00	30	TRAFFIC SIGNAL	FUNCTIONING PROPERLY	CLEAR	DARKNESS; LIGHTED ROAD	SI
25089	782878	2023-12-31 23:29:00	30	NO CONTROLS	NO CONTROLS	CLEAR	DARKNESS	
25075	782879	2023-12-31 23:30:00	5	NO CONTROLS	NO CONTROLS	UNKNOWN	UNKNOWN	P
28632	782880	2023-12-31 23:30:00	25	NO CONTROLS	NO CONTROLS	CLEAR	DARKNESS; LIGHTED ROAD	P
25066	782881	2023-12-31 23:50:00	30	NO CONTROLS	NO CONTROLS	RAIN	DARKNESS; LIGHTED ROAD	P

782881 rows × 24 columns

Các cột dữ liệu dạng chuỗi:

```
Thuộc tính: TRAFFIC_CONTROL_DEVICE
Độ dài giá trị dài nhất: 24
Các giá trị khác nhau: ['NO CONTROLS' 'TRAFFIC SIGNAL' 'UNKNOWN' 'STOP SIGN/FLASHER' 'OTHER'
 'OTHER REG. SIGN' 'RAILROAD CROSSING GATE' 'LANE USE MARKING' 'YIELD'
 'DELINEATORS' 'OTHER RAILROAD CROSSING' 'OTHER WARNING SIGN'
 'SCHOOL ZONE' 'POLICE/FLAGMAN' 'NO PASSING' 'RR CROSSING SIGN'
 'PEDESTRIAN CROSSING SIGN' 'FLASHING CONTROL SIGNAL'
 'BICYCLE CROSSING SIGN']

Thuộc tính: DEVICE_CONDITION
Độ dài giá trị dài nhất: 24
Các giá trị khác nhau: ['NO CONTROLS' 'FUNCTIONING PROPERLY' 'UNKNOWN' 'OTHER'
 'FUNCTIONING IMPROPERLY' 'NOT FUNCTIONING' 'WORN REFLECTIVE MATERIAL'
 'MISSING']

Thuộc tính: WEATHER_CONDITION
Độ dài giá trị dài nhất: 24
Các giá trị khác nhau: ['CLEAR' 'UNKNOWN' 'SNOW' 'CLOUDY/OVERCAST' 'RAIN' 'OTHER'
 'FOG/SMOKE/HAZE' 'SLEET/HAIL' 'SEVERE CROSS WIND GATE'
 'FREEZING RAIN/DRIZZLE' 'BLOWING SNOW' 'BLOWING SAND; SOIL; DIRT']

Thuộc tính: LIGHTING_CONDITION
Độ dài giá trị dài nhất: 22
Các giá trị khác nhau: ['DARKNESS; LIGHTED ROAD' 'DARKNESS' 'UNKNOWN' 'DAYLIGHT' 'DUSK' 'DAWN']

Thuộc tính: FIRST_CRASH_TYPE
Độ dài giá trị dài nhất: 28
Các giá trị khác nhau: ['OTHER OBJECT' 'REAR END' 'SIDESWIPE SAME DIRECTION'
 'PARKED MOTOR VEHICLE' 'OTHER NONCOLLISION' 'TURNING' 'ANGLE' 'HEAD ON'
 'SIDESWIPE OPPOSITE DIRECTION' 'FIXED OBJECT' 'PEDALCYCLIST' 'ANIMAL'
 'PEDESTRIAN' 'OVERTURNED' 'TRAIN' 'REAR TO SIDE' 'REAR TO FRONT'
 'REAR TO REAR']

Thuộc tính: TRAFFICWAY_TYPE
Độ dài giá trị dài nhất: 31
Các giá trị khác nhau: ['ONE-WAY' 'NOT DIVIDED' 'DIVIDED - W/MEDIAN (NOT RAISED)' 'PARKING LOT'
 'OTHER' 'ALLEY' 'DIVIDED - W/MEDIAN BARRIER' 'UNKNOWN' 'CENTER TURN LANE'
 'DRIVEWAY' 'RAMP' 'FOUR WAY' 'T-INTERSECTION' 'UNKNOWN INTERSECTION TYPE'
 'Y-INTERSECTION' 'FIVE POINT; OR MORE' 'TRAFFIC ROUTE' 'ROUNDABOUT'
 'NOT REPORTED' 'L-INTERSECTION']
```

```
Thuộc tính: ALIGNMENT
Độ dài giá trị dài nhất: 21
Các giá trị khác nhau: ['STRAIGHT AND LEVEL' 'CURVE; LEVEL' 'STRAIGHT ON GRADE' 'CURVE ON GRADE'
 'STRAIGHT ON HILLCREST' 'CURVE ON HILLCREST']

Thuộc tính: DAMAGE
Độ dài giá trị dài nhất: 11
Các giá trị khác nhau: ['500 OR LESS' '501 - 1.500' 'OVER 1.500']

Thuộc tính: PRIM_CONTRIBUTORY_CAUSE
Độ dài giá trị dài nhất: 80
Các giá trị khác nhau: ['NOT APPLICABLE' 'FAILING TO REDUCE SPEED TO AVOID CRASH'
 'IMPROPER LANE USAGE' 'UNABLE TO DETERMINE'
 'UNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED)'
 'FOLLOWING TOO CLOSELY' 'DISREGARDING STOP SIGN'
 'EXCEEDING AUTHORIZED SPEED LIMIT' 'FAILING TO YIELD RIGHT-OF-WAY'
 'IMPROPER OVERTAKING/PASSING' 'IMPROPER BACKING'
 'IMPROPER TURNING/NO SIGNAL' 'EXCEEDING SAFE SPEED FOR CONDITIONS'
 'OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER'
 'DRIVING ON WRONG SIDE/WRONG WAY' 'WEATHER'
 'VISION OBSCURED (SIGNS; TREE LIMBS; BUILDINGS; ETC.)'
 'DISREGARDING TRAFFIC SIGNALS' 'ROAD CONSTRUCTION/MAINTENANCE'
 'DISREGARDING ROAD MARKINGS' 'PHYSICAL CONDITION OF DRIVER'
 'DISTRACTION - FROM INSIDE VEHICLE' 'DRIVING SKILLS/KNOWLEDGE/EXPERIENCE'
 'DISTRACTION - FROM OUTSIDE VEHICLE' 'ANIMAL'
 'EQUIPMENT - VEHICLE CONDITION'
 'EVASIVE ACTION DUE TO ANIMAL; OBJECT; NONMOTORIST'
 'TURNING RIGHT ON RED' 'ROAD ENGINEERING/SURFACE/MARKING DEFECTS'
 'BICYCLE ADVANCING LEGALLY ON RED LIGHT'
 'DISREGARDING OTHER TRAFFIC SIGNS' 'DISREGARDING YIELD SIGN'
 'CELL PHONE USE OTHER THAN TEXTING'
 'HAD BEEN DRINKING (USE WHEN ARREST IS NOT MADE)'
 'PASSING STOPPED SCHOOL BUS'
 'DISTRACTION - OTHER ELECTRONIC DEVICE (NAVIGATION DEVICE; DVD PLAYER; ETC.)'
 'TEXTING' 'MOTORCYCLE ADVANCING LEGALLY ON RED LIGHT'
 'RELATED TO BUS STOP' 'OBSTRUCTED CROSSWALKS']
```

<u>Bước 11:</u> Lưu dataset đã lọc thành file CSV "DataOL14.csv" để xây dựng và phân tích kho dữ liệu sau này.

```
[86]: df.to_csv('dataOL14.csv', index=False)
```

File dữ liệu sau khi làm sạch: Gồm 24 thuộc tính và 782881 dòng.

1.1.3 Thuộc tính của dữ liệu:

STT	Tên thuộc tính	Kiểu	Ý nghĩa
1	CRASH_RECORD_ID	int	Mã định danh duy nhất cho mỗi bản ghi tai nạn.
2	CRASH_DATE	timestamp	Ngày tai nạn xảy ra.
3	POSTED_SPEED_LIMIT	int	Giới hạn tốc độ được quy định tại địa điểm xảy ra tai nạn.
4	TRAFFIC_CONTROL_DEVICE	varchar	Thiết bị kiểm soát giao thông tại địa điểm xảy ra tai nạn.
5	DEVICE_CONDITION	varchar	Mô tả trạng thái hoặc điều kiện của các thiết bị kiểm soát giao thông tại thời điểm xảy ra tai nạn
6	WEATHER_CONDITION	varchar	Điều kiện thời tiết tại thời điểm xảy ra tai nạn.
7	LIGHTING_CONDITION	varchar	Điều kiện chiếu sáng tại thời điểm xảy ra tai nạn.
8	FIRST_CRASH_TYPE	varchar	Loại sự cố chính.

9	TRAFFICWAY_TYPE	varchar	Loại đường giao thông nơi xảy ra sự cố.
10	ALIGNMENT	varchar	Cách mà đường hoặc tuyến đường được bố trí hoặc thiết kế theo một chiều cụ thể (STRAIGHT AND LEVEL (Thẳng và bằng phẳng), CURVE ON GRADE (Đường cong có độ dốc), CURVE, LEVEL (cong, bằng phẳng))
11	DAMAGE	varchar	Mức độ thiệt hại.
12	PRIM_CONTRIBUTORY _CAUSE	varchar	Nguyên nhân trực tiếp và quan trọng nhất dẫn đến vụ tai nạn.
13	STREET_NO	int	Số nhà hoặc số đường nơi tai nạn xảy ra
14	STREET_DIRECTION	varchar(1)	Hướng đi của đường, chẳng hạn như N (Bắc), S (Nam), E (Đông), hoặc W (Tây).
15	STREET_NAME	varchar	Tên con đường nơi tai nạn xảy ra
16	NUM_UNITS	int	Số lượng phương tiện hoặc đơn vị tham gia vào vụ tai nạn.
17	MOST_SEVERE_INJURY	varchar	Chấn thương nghiêm trọng nhất được báo cáo trong vụ tai nạn
18	INJURIES_TOTAL	int	Tổng số lượng chấn thương trong vụ tai nạn
19	INJURIES_FATAL	int	Số lượng chấn thương dẫn đến cái chết
20	INJURIES_INCAPACITA TING	int	Số lượng chấn thương khiến nạn nhân không thể thực hiện các hoạt động bình thường trong một khoảng thời gian
21	INJURIES_NON_INCAP ACITATING	int	Số lượng chấn thương không ảnh hưởng nghiêm trong đến khả năng hoạt động của người bị thương.
22	INJURIES_NO_INDICATI ON	int	Số lượng chấn thương không có biểu hiện đáng chú ý (không có chấn thương)
23	CRASH_HOUR	int	Giờ xảy ra tai nạn
24	CRASH_DAY_OF_WEEK	int	Thứ tự ngày trong tuần của ngày xảy ra tai nạn
25	CRASH_MONTH	int	Tháng xảy ra tai nạn
26	CRASH_DAY	int	Ngày xảy ra tai nạn
27	CRASH_YEAR	int	Năm xảy ra tai nạn

1.1.4 Hướng chủ đề:

Tai nạn giao thông (Traffic Crash).

Hướng khai thác:

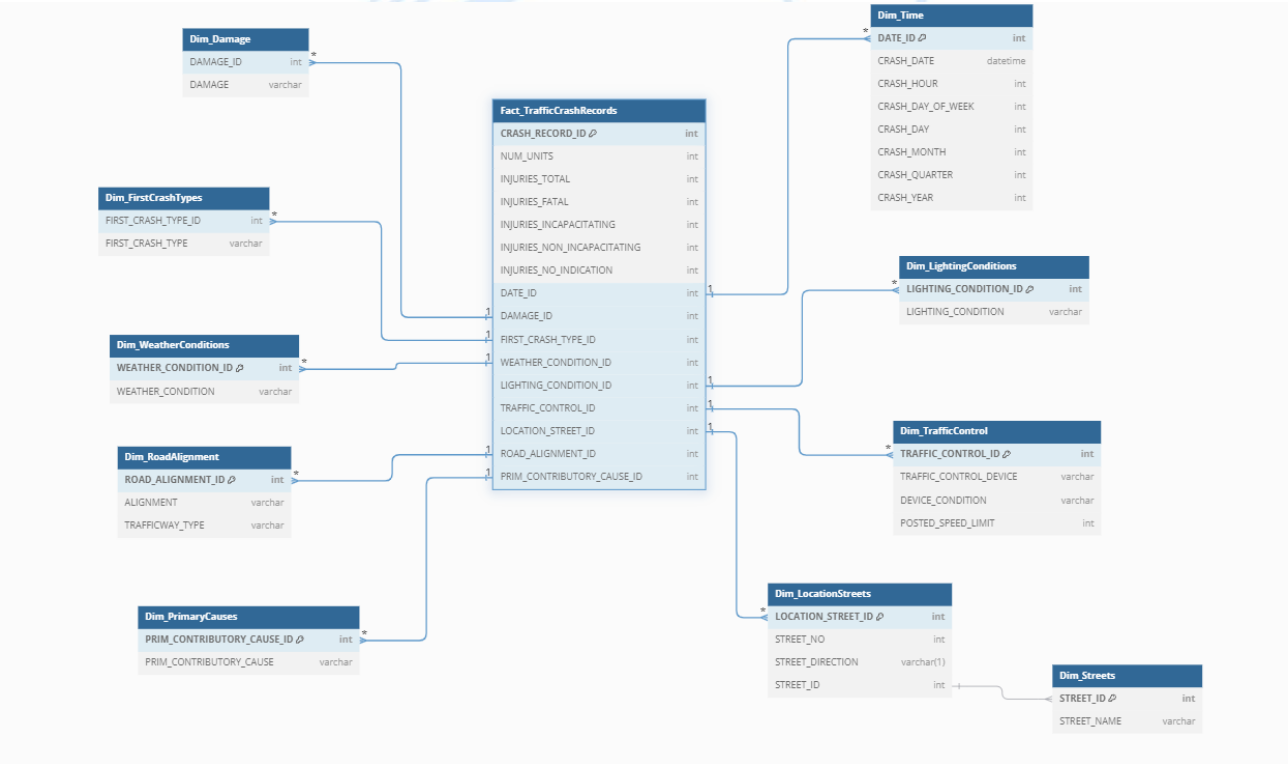
- ❖ Quản lý và phân tích các vụ tai nạn giao thông. Các bảng dữ liệu đều xoay quanh các yếu tố có thể ảnh hưởng đến tai nạn giao thông, giúp thu thập và phân tích thông tin chi tiết về các vụ tai nạn.
- ❖ Khai thác và tìm hiểu mối tương quan giữa các bộ dữ liệu về: nguyên nhân tai nạn giao thông, thiệt hại, thương vong, thời gian xảy ra tai nạn, điều kiện giao thông, đường sá, khu vực xảy ra, loại tai nạn.

Mục đích:

- ❖ Dựa trên các phân tích dữ liệu lịch sử, dự đoán xu hướng tai nạn có thể xảy ra trong tương lai, giúp đưa ra các biện pháp phòng ngừa.
- ❖ Phân tích dữ liệu để phát hiện các yếu tố tiềm ẩn có thể dẫn đến tai nạn, từ đó đề xuất biện pháp cải thiện giao thông.

1.2 XÂY DỰNG KHO DỮ LIỆU:

1.2.1 Thiết kế lược đồ:



1.2.2 Mô tả chi tiết các bảng:

1.2.2.1 Bảng Fact_TrafficCrashRecords:

Ràng buộc	Tên thuộc tính	Kiểu dữ liệu	Mô tả thuộc tính
Khóa chính	CRASH_RECORD_ID	int	Mã định danh duy nhất cho mỗi bản ghi tai nạn

	NUM_UNITS	int	Số lượng phương tiện hoặc đơn vị tham gia vào vụ tai nạn.
	INJURIES_TOTAL	int	Tổng số lượng chấn thương trong vụ tai nạn
	INJURIES_FATAL	int	Số lượng chấn thương dẫn đến cái chết
	INJURIES_INCAPACITATING	int	Số lượng chấn thương khiến nạn nhân không thể thực hiện các hoạt động bình thường trong một khoảng thời gian
	INJURIES_NON_INCAPACITATING	int	Số lượng chấn thương không ảnh hưởng nghiêm trọng đến khả năng hoạt động của người bị thương.
	INJURIES_NO_INDICATION	int	Số lượng chấn thương không có biểu hiện đáng chú ý (không có chấn thương).
Khóa ngoại	DATE_ID	int	Mã ngày xảy ra tai nạn.
Khóa ngoại	DAMAGE_ID	int	Mã mức độ thiệt hại.
Khóa ngoại	FIRST_CRASH_TYPE_ID	int	Mã loại sự cố chính.
Khóa ngoại	WEATHER_CONDITION_ID	int	Mã thời tiết.
Khóa ngoại	LIGHTING_CONDITION_ID	int	Mã điều kiện chiếu sáng.
Khóa ngoại	TRAFFIC_CONTROL_ID	int	Mã thiết bị kiểm soát giao thông.
Khóa ngoại	LOCATION_STREET_ID	int	Mã địa chỉ.
Khóa ngoại	ROAD_ALIGNMENT_ID	int	Mã căn chỉnh.
Khóa ngoại	PRIM_CONTRIBUTORY_CAUSE_ID	int	Mã nguyên nhân.

1.2.2.2 Bảng Dim_Time:

Ràng buộc	Tên thuộc tính	Kiểu	Mô tả thuộc tính
Khóa chính	CRASH_ID	int	Mã thời gian xảy ra tai nạn
	CRASH_DATE	datetime	Thời gian cụ thể tai nạn xảy ra.

	CRASH_HOUR	int	Giờ xảy ra tai nạn
	CRASH_DAY_OF_WEEK	int	Thứ tự ngày trong tuần của ngày xảy ra tai nạn
	CRASH_DAY	int	Ngày xảy ra tai nạn
	CRASH_MONTH	int	Tháng xảy ra tai nạn
	CRASH_QUARTER	int	Quý xảy ra tai nạn
	CRASH_YEAR	int	Năm xảy ra tai nạn

1.2.2.3 Bảng Dim_ DAMAGE_ID:

Ràng buộc	Tên thuộc tính	Kiểu	Mô tả thuộc tính
Khóa chính	DAMAGE_ID	int	Mã mức độ thiệt hại.
	DAMAGE	varchar	Mức độ thiệt hại.

1.2.2.4 Bảng Dim_FirstCrashTypes:

Ràng buộc	Tên thuộc tính	Kiểu	Mô tả thuộc tính
Khóa chính	FIRST_CRASH_TYPE_ID	int	Mã loại sự cố chính.
	FIRST_CRASH_TYPE	varchar	Tên loại sự cố chính.

1.2.2.5 Bảng Dim_WeatherConditions:

Ràng buộc	Tên thuộc tính	Kiểu	Mô tả thuộc tính
Khóa chính	WEATHER_CONDITION_ID	int	Mã thời tiết.
	WEATHER_CONDITION	varchar	Điều kiện thời tiết tại thời điểm xảy ra tai nạn.

1.2.2.6 Bảng Dim_LightingConditions:

Ràng buộc	Tên thuộc tính	Kiểu	Mô tả thuộc tính
Khóa chính	LIGHTING_CONDITION_ID	int	Mã điều kiện chiếu sáng.
	LIGHTING_CONDITION	varchar	Điều kiện chiếu sáng tại thời điểm xảy ra tai nạn.

1.2.2.7 Bảng Dim_TrafficControl

Ràng buộc	Tên thuộc tính	Kiểu	Mô tả thuộc tính

Khóa chính	TRAFFIC_CONTROL_ID	int	Mã thiết bị kiểm soát giao thông.
	TRAFFIC_CONTROL_DEVICE	varchar	Thiết bị kiểm soát giao thông tại địa điểm xảy ra tai nạn.
	DEVICE_CONDITION	varchar	Mô tả trạng thái hoặc điều kiện của các thiết bị kiểm soát giao thông tại thời điểm xảy ra tai nạn
	POSTED_SPEED_LIMIT	int	Giới hạn tốc độ được quy định tại địa điểm xảy ra tai nạn.

1.2.2.8 Bảng Dim_LocationStreets:

Ràng buộc	Tên thuộc tính	Kiểu	Mô tả thuộc tính
Khóa chính	LOCATION_STREET_ID	int	Mã địa chỉ.
	STREET_NO	int	Số nhà hoặc số đường nơi tai nạn xảy ra
	STREET_DIRECTION	varchar(1)	Hướng đi của đường, chẳng hạn như N (Bắc), S (Nam), E (Đông), hoặc W (Tây).
Khóa ngoại	STREET_ID	int	Mã tên đường

1.2.2.9 Bảng Dim_Streets:

Ràng buộc	Tên thuộc tính	Kiểu	Mô tả thuộc tính
Khóa chính	STREET_ID	int	Mã tên đường
	STREET_NAME	varchar	Tên con đường nơi tai nạn xảy ra

1.2.2.10 Bảng Dim_RoadAlignment:

Ràng buộc	Tên thuộc tính	Kiểu	Mô tả thuộc tính
Khóa chính	ROAD_ALIGNMENT_ID	int	Mã căn chỉnh.
	ALIGNMENT	varchar	Cách mà đường hoặc tuyến đường được bố trí hoặc thiết kế theo một chiều cụ thể (STRAIGHT AND LEVEL (Thẳng và bằng phẳng), CURVE ON GRADE (Đường cong có độ dốc), CURVE, LEVEL (cong, bằng phẳng))

	TRAFFICWAY_TYPE	varchar	Loại đường giao thông nơi xảy ra sự cố.

1.2.2.11 Bảng Dim_PrimaryCauses

Ràng buộc	Tên thuộc tính	Kiểu	Mô tả thuộc tính
Khóa chính	PRIM_CONTRIBUTORY_CAUSE_ID	int	Mã nguyên nhân.
	PRIM_CONTRIBUTORY_CAUSE	varchar	Nguyên nhân trực tiếp và quan trọng nhất dẫn đến vụ tai nạn.

1.2.2.12 Bảng Dim_Damages

Ràng buộc	Tên thuộc tính	Kiểu	Mô tả thuộc tính
Khóa chính	DAMAGE_ID	int	Mã thiệt hại.
	DAMAGE	varchar	Mức độ thiệt hại về tài sản mà vụ tai nạn gây ra.

1.3 CÁC CÂU TRUY VẤN:

1. Liệt kê top 10 vụ tai nạn giao thông có số lượng tử vong cao nhất (sắp xếp theo số lượng giảm dần).
2. Thống kê tỉ lệ tử vong và bị thương nặng của nạn nhân trong các nhân tố trực tiếp dẫn đến vụ tai nạn (FIRST_CRASH_TYPE).
3. Liệt kê các thông số về chấn thương do tai nạn giao thông từ 2016 đến 2023.
4. Trong 10 nguyên nhân gây va chạm giao thông phổ biến nhất ở mỗi năm, thống kê số lượng nạn nhân tử vong, bị thương nặng và bị thương nhẹ ở mỗi nguyên nhân từ năm 2021 đến năm 2023 khi điều kiện thời tiết là "CLEAR". Ở đây ta sẽ không tính nguyên nhân là "NOT APPLICABLE" và "UNABLE TO DETERMINE" vì đây là hai nguyên nhân không xác định.
5. Liệt kê top 10 con đường có nhiều vị trí xảy ra tai nạn giao thông nhất, cùng với hướng đường và số địa điểm đã xảy ra va chạm giao thông ở mỗi hướng đường của con đường đó.
6. Thống kê số lượng các vụ va chạm giao thông cho từng mức độ thiệt hại theo từng tháng trong năm 2023.
7. Thống kê số nạn nhân bị thương ở mỗi vụ tai nạn [(số nạn nhân tử vong + bị thương nặng + bị thương nhẹ) / số vụ tai nạn] xảy ra ở loại cấu trúc đường cong

(ALIGNMENT có kí tự “CURVE”) khi trời tối (LIGHTING CONDITION có kí tự “DARK” trong tên) qua từng năm (từ 2016 đến 2023).

8. Thống kê số lượng phương tiện va chạm nhiều nhất trong một vụ tai nạn giao thông ở các nguyên nhân.
9. Thống kê tổng số nạn nhân, số vụ tai nạn theo từng thiết bị điều khiển (TRAFFIC CONTROL DEVICE) và từng nguyên nhân (PRIM CONTRIBUTORY CAUSE bắt đầu bằng ký tự 'F').
10. Thống kê tổng số lượng vụ tai nạn và số lượng tử vong ở mỗi vụ tai nạn qua từng tháng và quý trong năm (từ 2016 đến 2023) do nguyên nhân ‘ANIMAL’.
11. Thống kê các nguyên nhân gây tai nạn ở mỗi loại đường (TRAFFICWAY TYPE) với số lượng tử vong lớn hơn hoặc bằng 6.
12. Liệt kê các loại đường giao thông (TRAFFICWAY TYPE) là top 5 loại đường xảy ra nhiều vụ tai nạn nhất trong cả hai năm 2022 và 2023.
13. Thống kê top 3 khung giờ xảy ra nhiều vụ tai nạn nhất của các ngày trong tuần.
14. Thống kê Top 5 nguyên nhân phổ biến nhất xuất hiện ở một trong các tập hợp: ({Nguyên nhân gây tử vong trung bình cao nhất}, {Nguyên nhân gây thiệt hại về tài sản cao nhất (OVER 1.500)}, {Nguyên nhân gây ra số lượng tử vong cao nhất khi tốc độ không vượt quá 40}) trong năm 2020.
15. Thống kê số vụ tai nạn và số lượng tử vong trong các điều kiện thời tiết ('SNOW', 'RAIN', 'CLOUDY/OVERCAST', 'CLEAR') ở từng quý từ năm 2020 đến 2023.

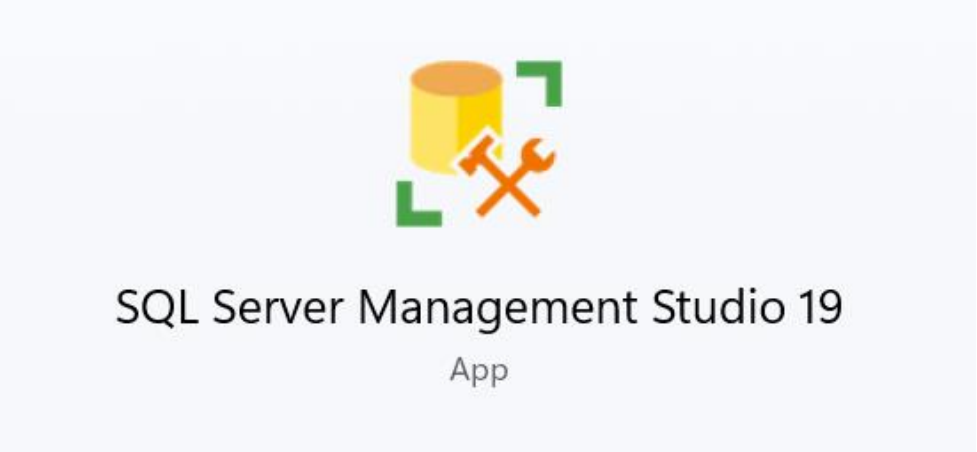
CHƯƠNG 2: TÍCH HỢP DỮ LIỆU VÀO KHO DỮ LIỆU (SSIS)

2.1 Chuẩn bị công cụ và cơ sở dữ liệu

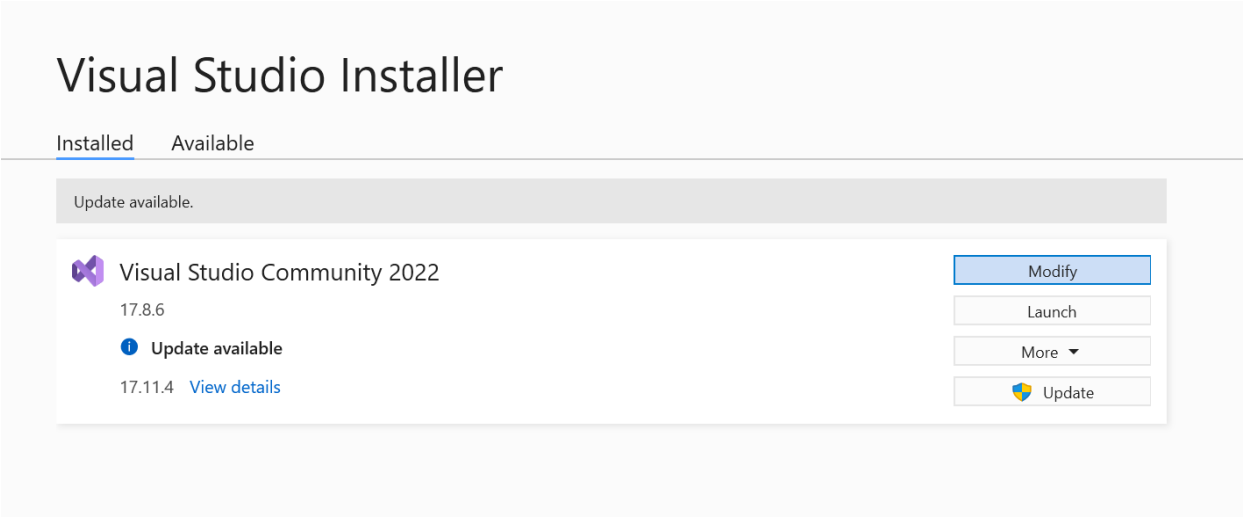
2.1.1 Chuẩn bị công cụ

Để thực hiện quá trình tích hợp dữ liệu vào kho dữ liệu - SSIS, ta cần cài đặt các công cụ:

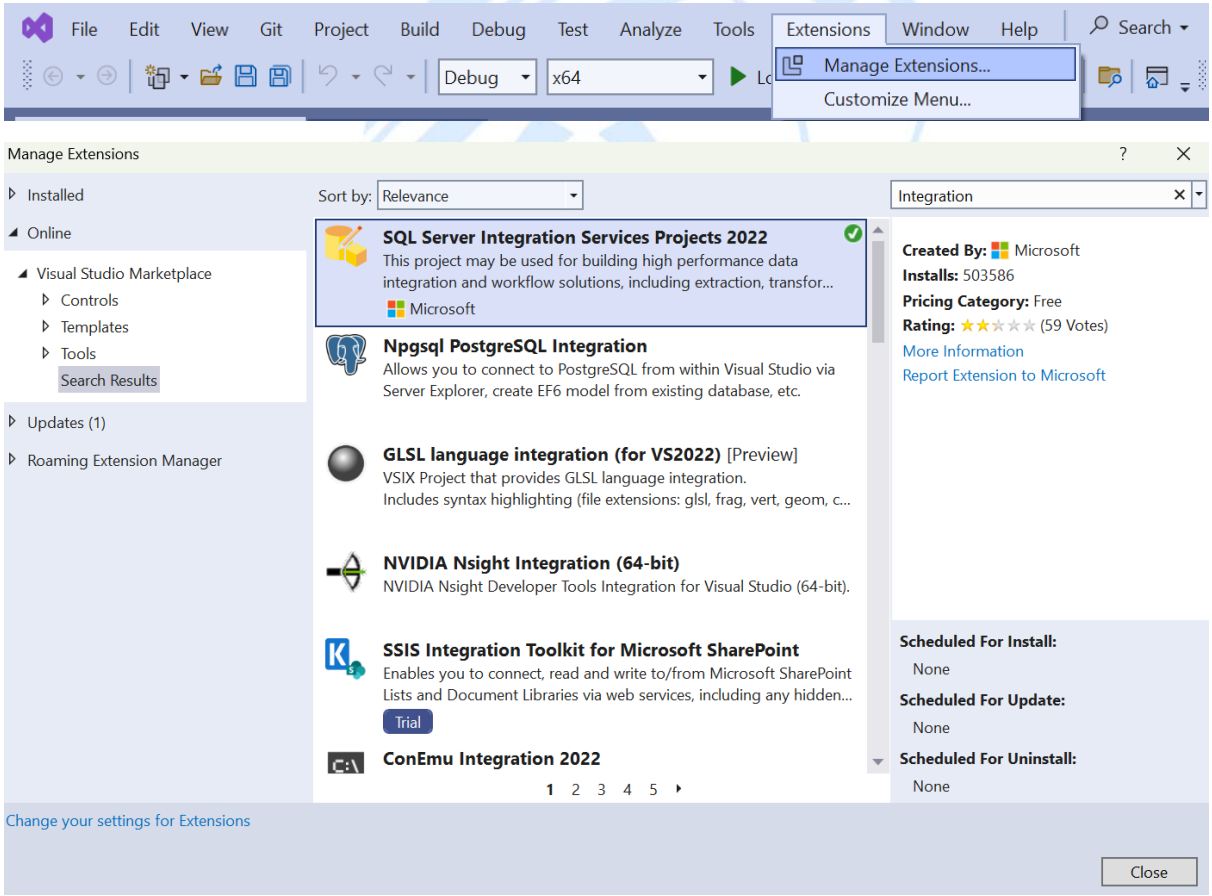
- **Microsoft SQL Server Management Studio**: tiến hành download và cài đặt công cụ.



- **Visual Studio Community 2022**: tiến hành download và cài đặt công cụ.

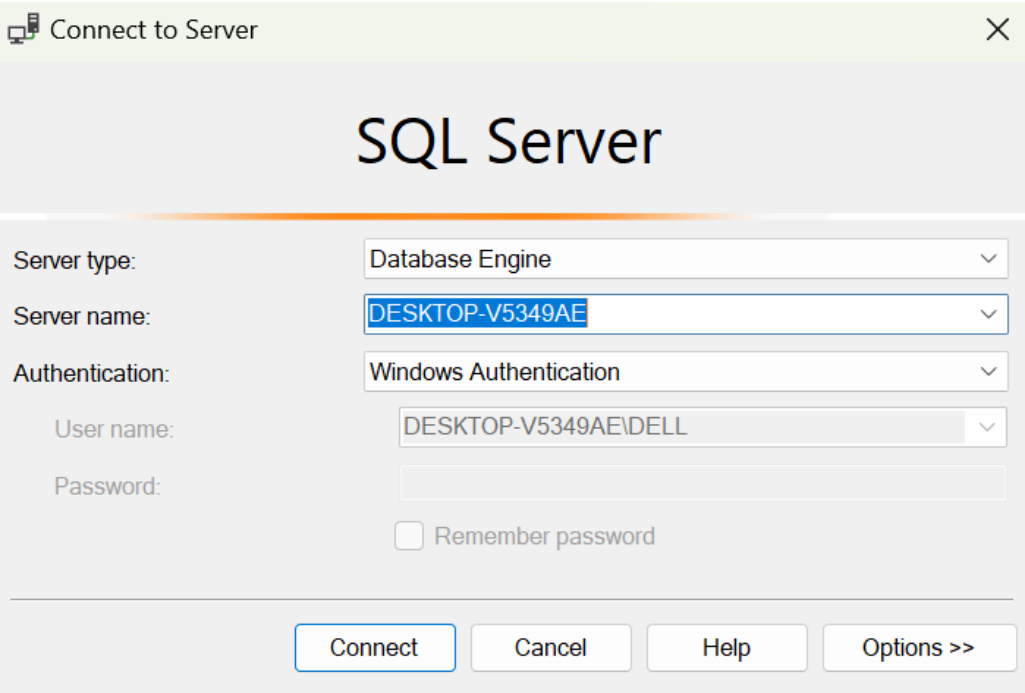


- **SQL Server Integration Services Project**: vào Extensions => Manage Extensions và tiến hành download công cụ.

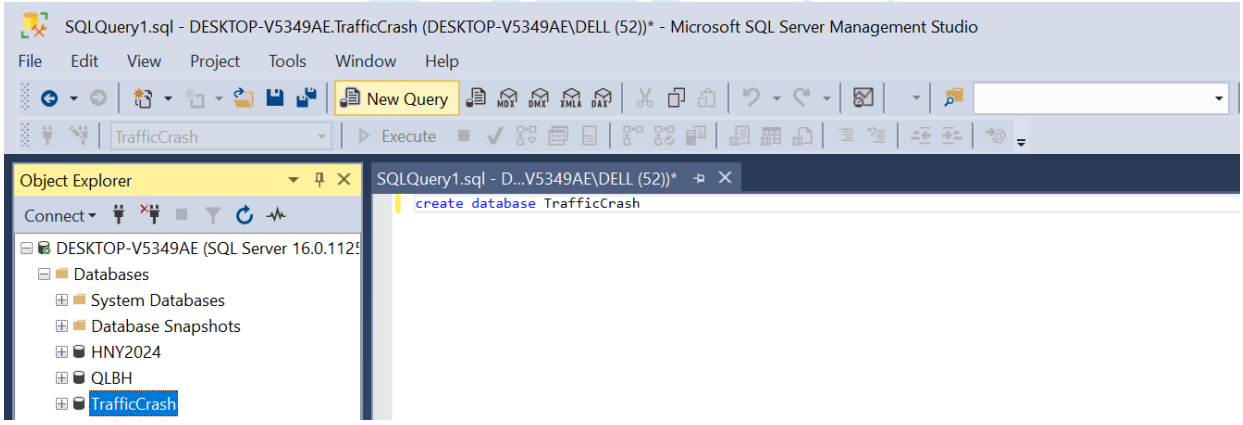


2.1.2 Chuẩn bị cơ sở dữ liệu

Bước 1: Mở SQL Server Management Studio 19 và thực hiện kết nối với server.

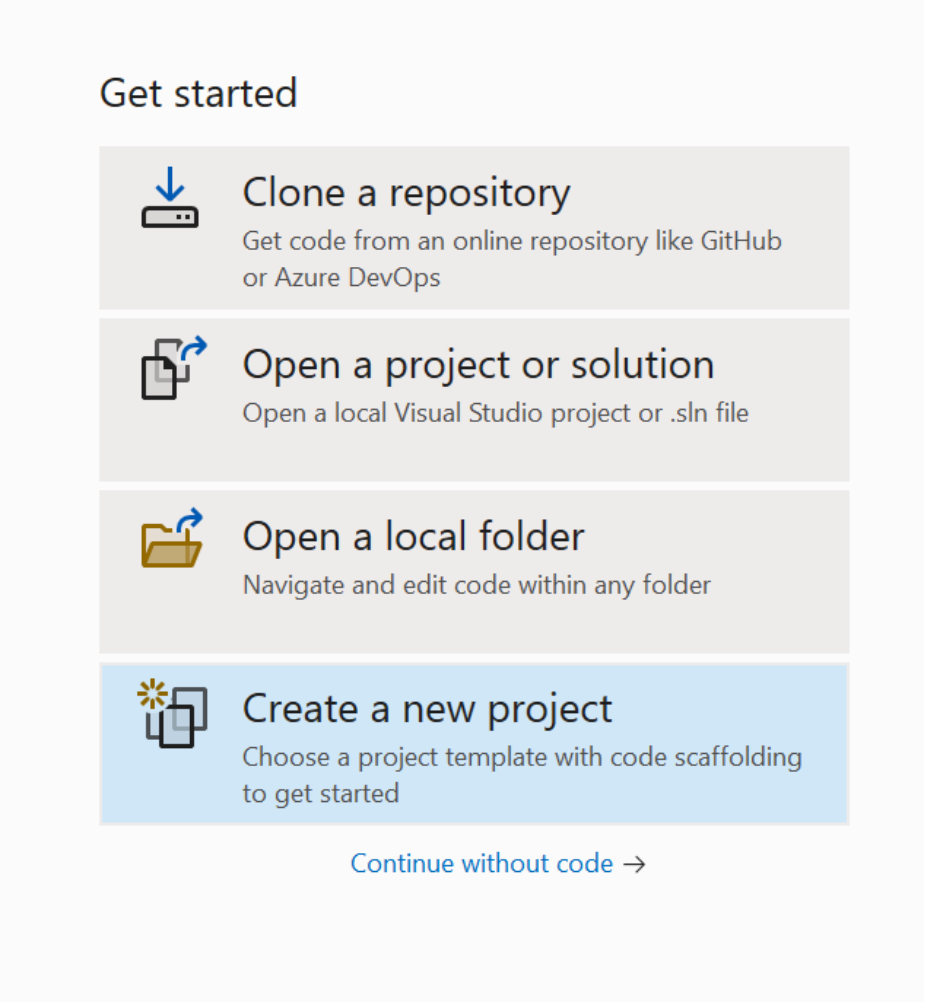


Bước 2: Tạo một database mới có tên TrafficCrash để lưu các bảng Dim, Fact chứa dữ liệu của bộ dữ liệu.

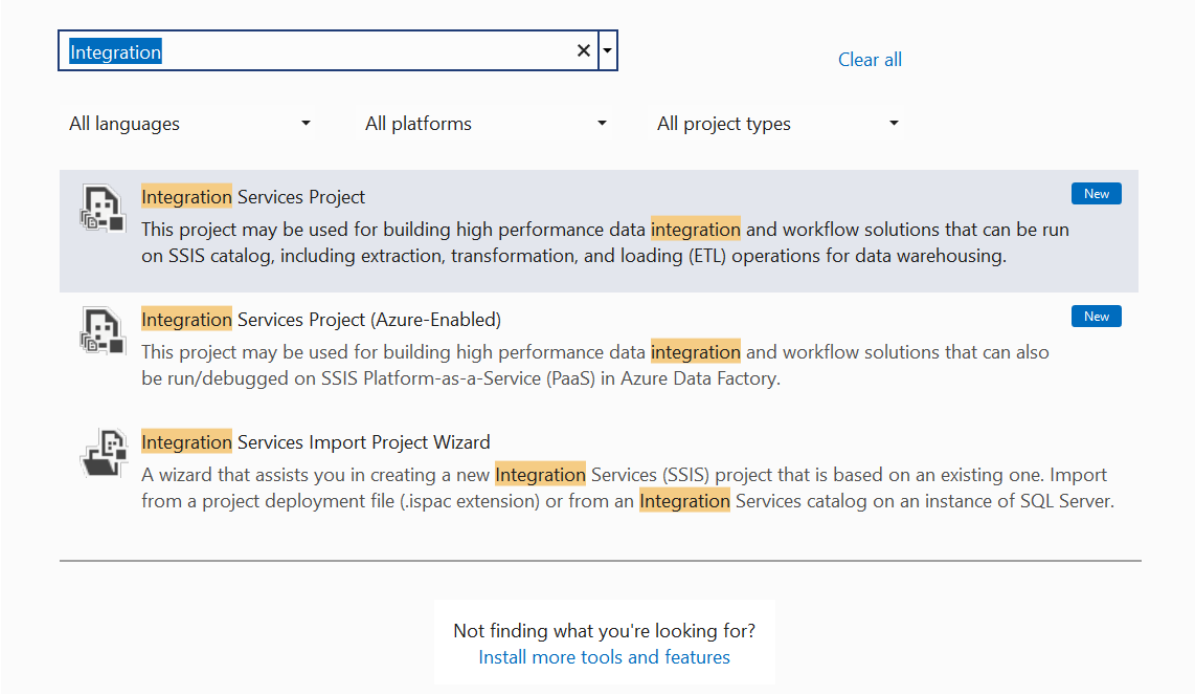


2.2 Tạo project SSIS

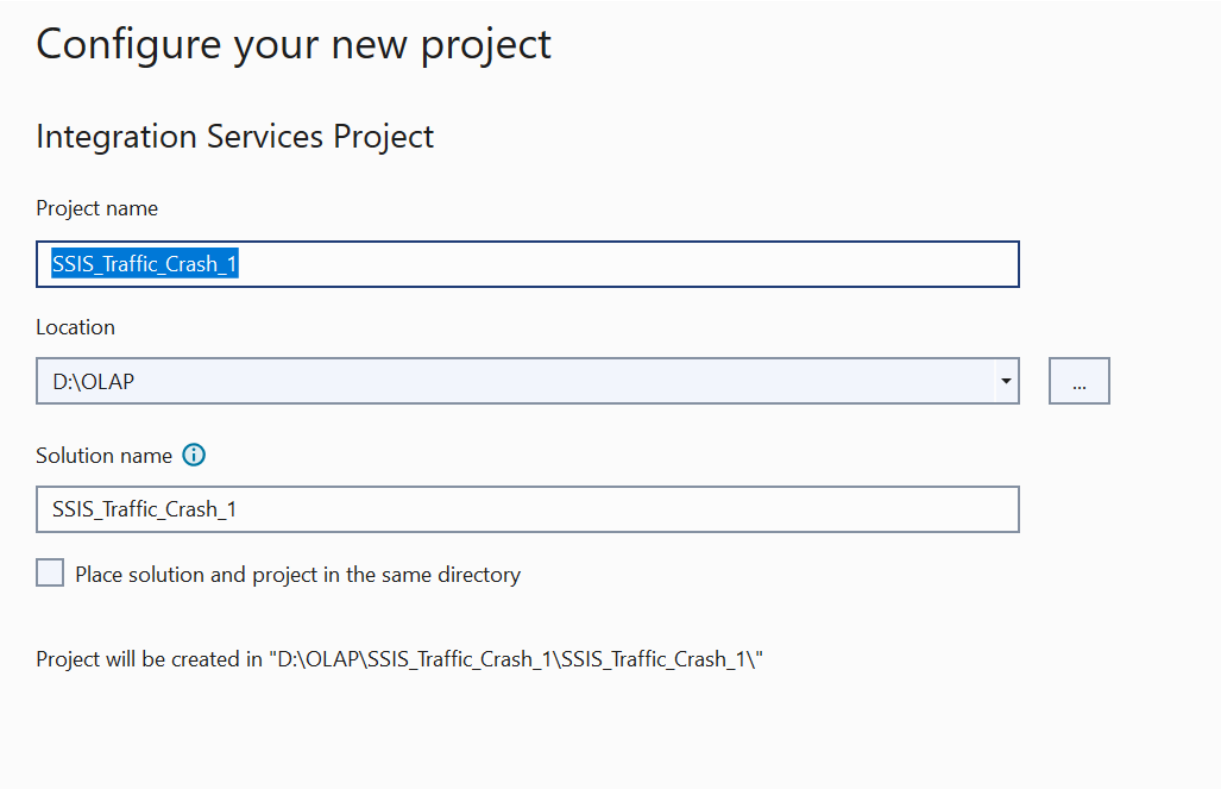
Bước 1: Vào Visual Studio, chọn “Create a new project”.



Bước 2: Tìm kiếm và chọn "Integration Services Project", sau đó nhấn "Next" để tiếp tục.

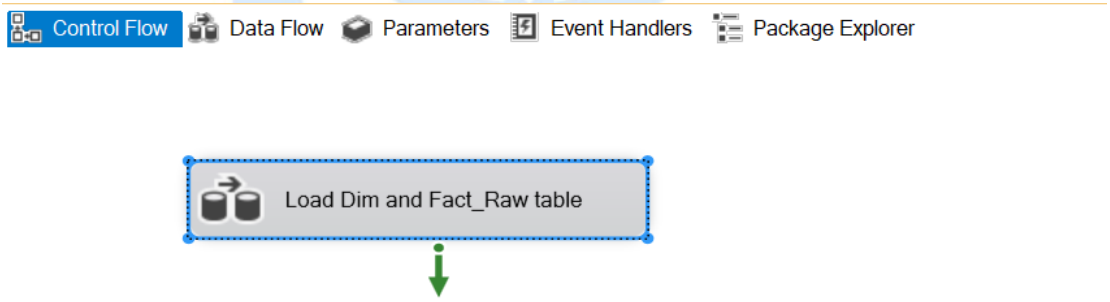


Bước 3: Đặt tên project và tạo đường dẫn và chọn "Create".

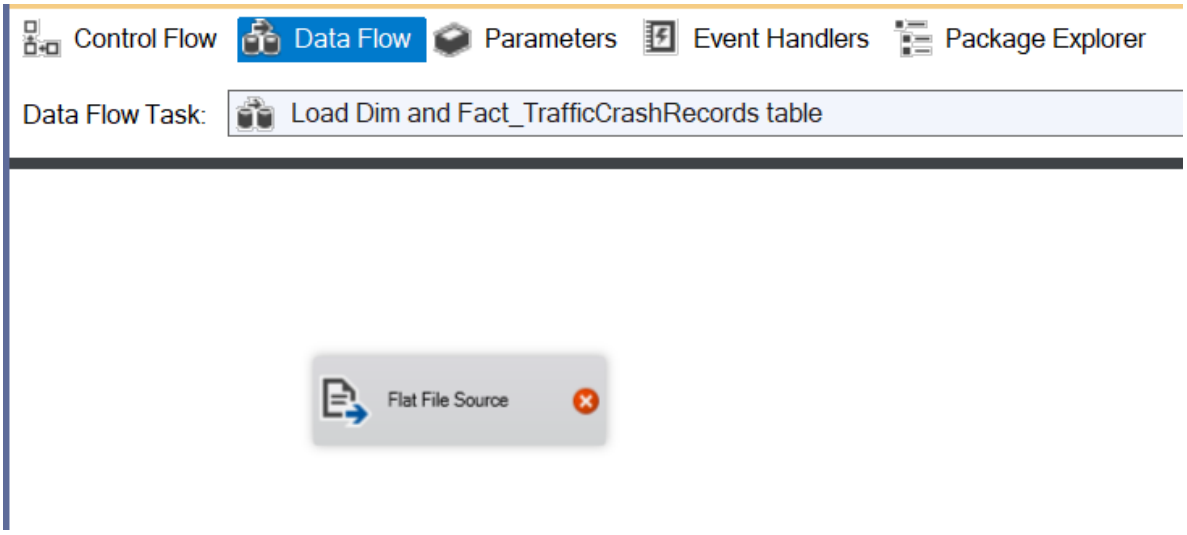


2.3 Tạo các bảng Dim và bảng Fact

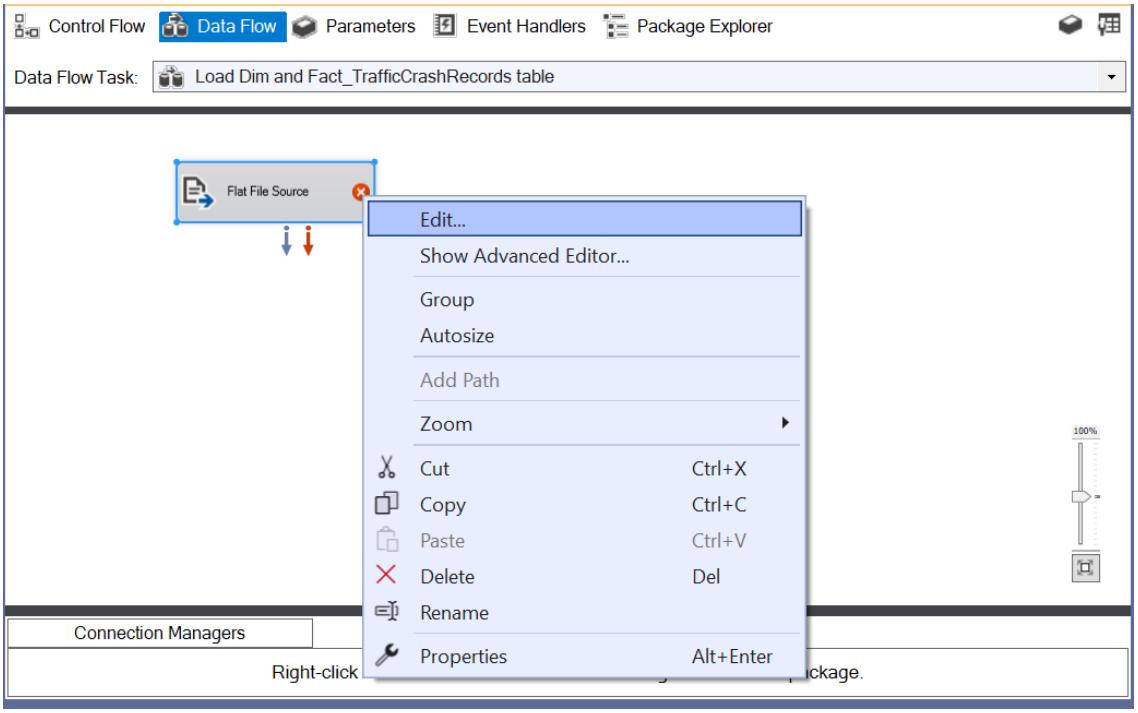
- Trong Control Flow, tạo Data Flow Task, đặt tên "Load Dim and Fact_Raw table" để chuẩn bị cho việc tạo các bảng Dim.



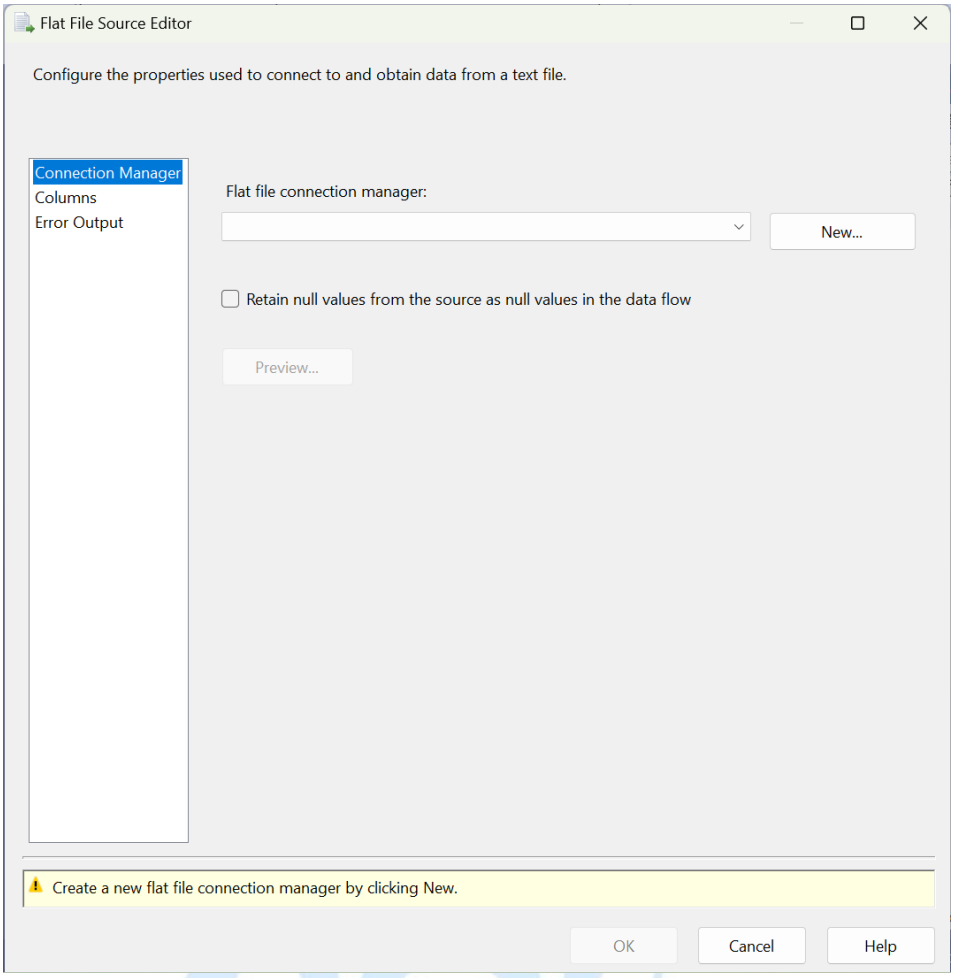
- Trong Data Flow, tạo mới một "Flat File Source".



- Nháy đúp hoặc nhấn chuột phải vào “Flat File Source”.

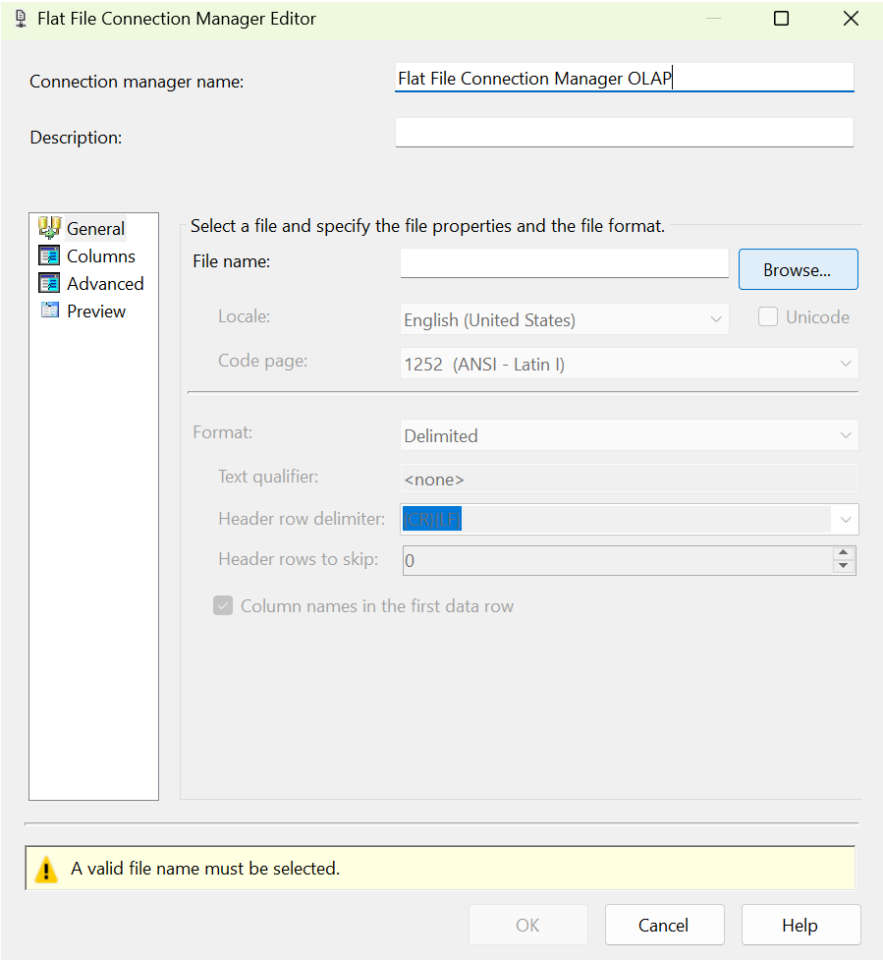


- Chọn “Edit” để hiển thị “Flat File Source Editor”.

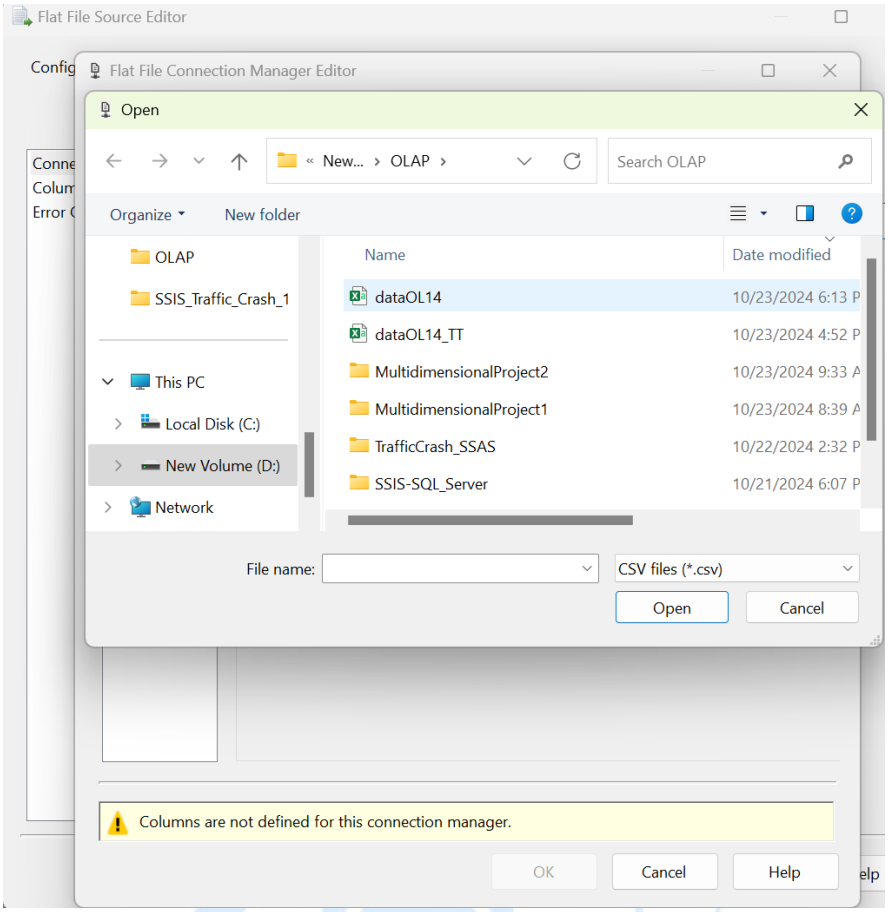


- Sau đó, chọn vào "New" để tạo mới một "Flat File Connection", kết nối đến file .csv chứa bộ dữ liệu.

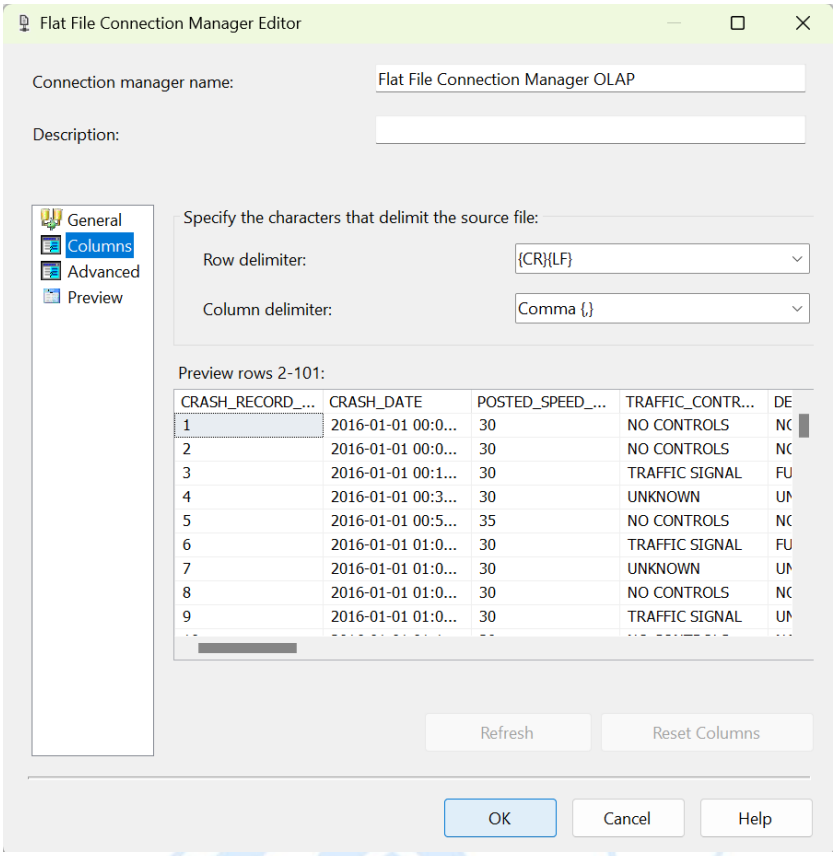
- Ở "Flat File Connection Manager Editor", chọn "Browse", chọn file .csv chứa bộ dữ liệu đã chuẩn bị.



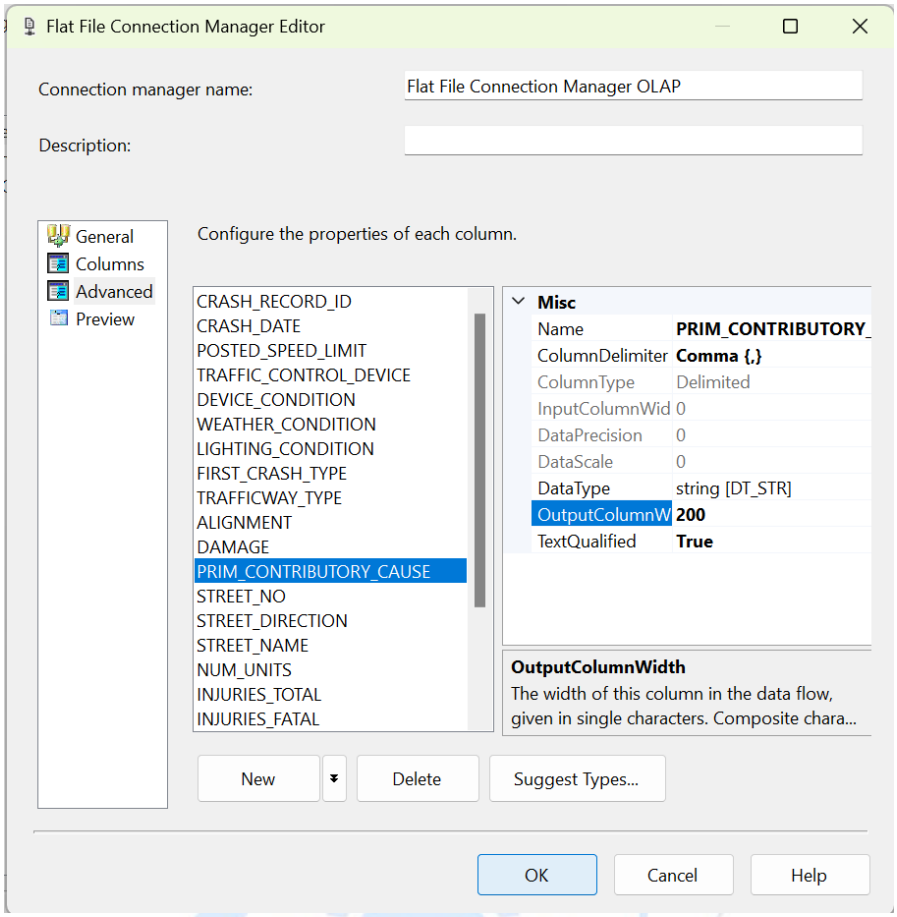
- Tiến hành chọn file dữ liệu dataOL14.csv và nhấn “Open”.



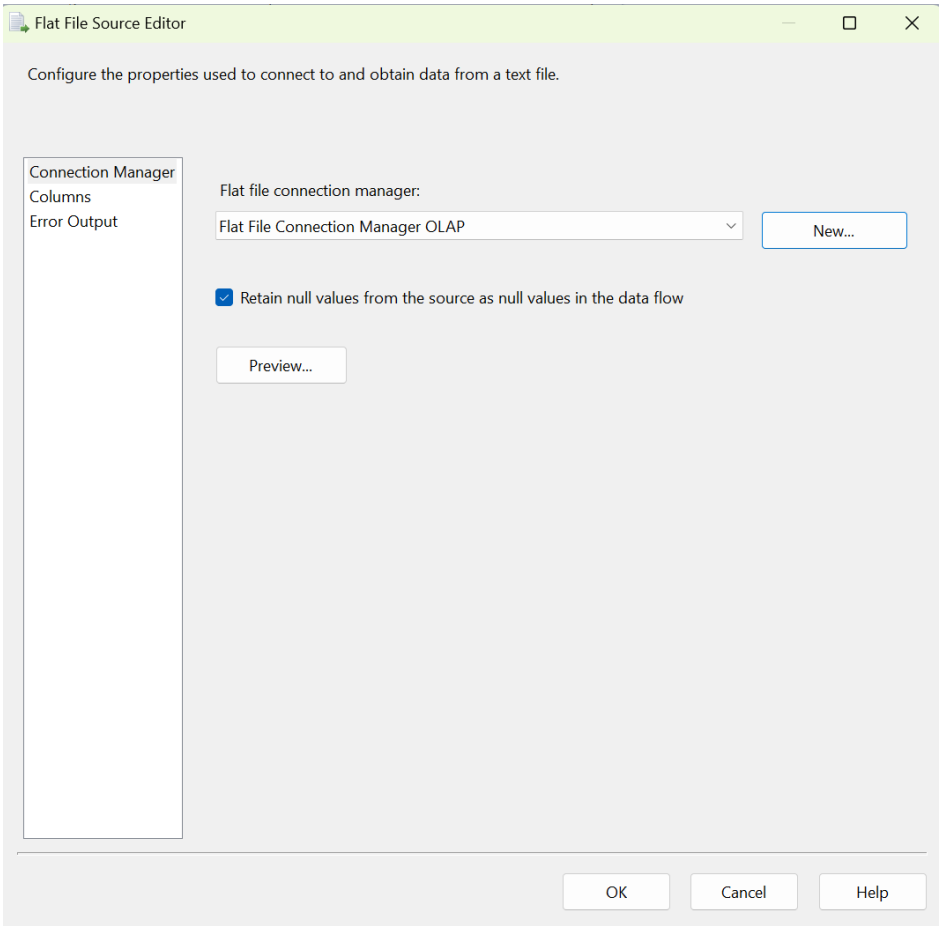
- Ở mục “Columns”, ta có thể kiểm tra các cột dữ liệu sau khi chọn file .csv.



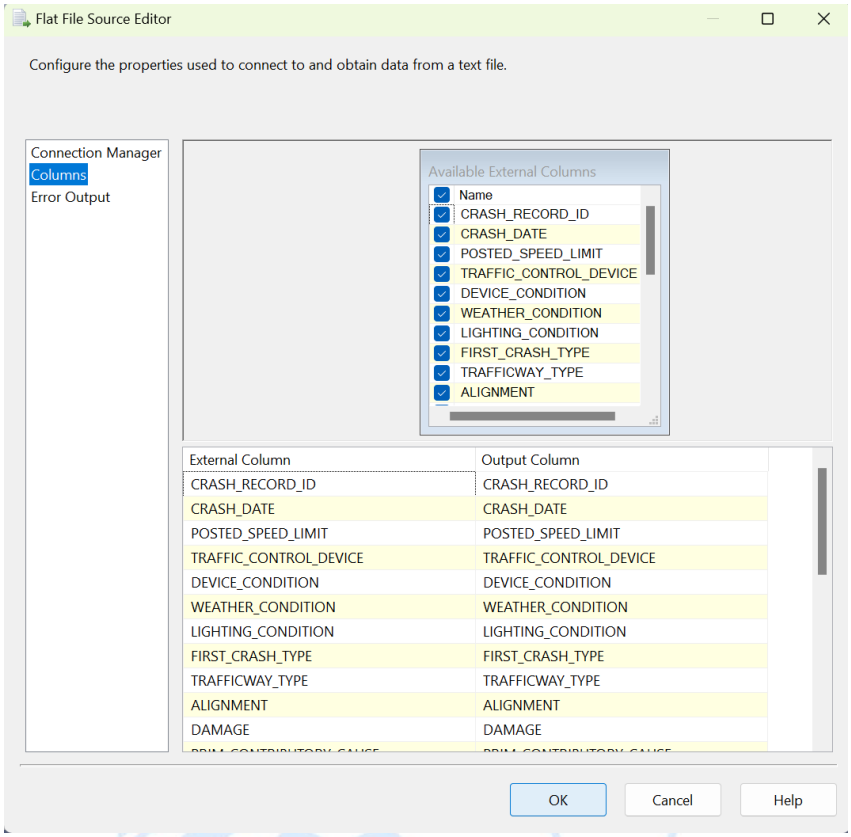
- Ở mục “Advanced”, để thay đổi kích thước ở các cột dữ liệu kiểu String có độ dài ký tự lớn hơn độ dài mặc định, ta gõ lại độ dài dữ liệu. Sau đó nhấn “OK”.



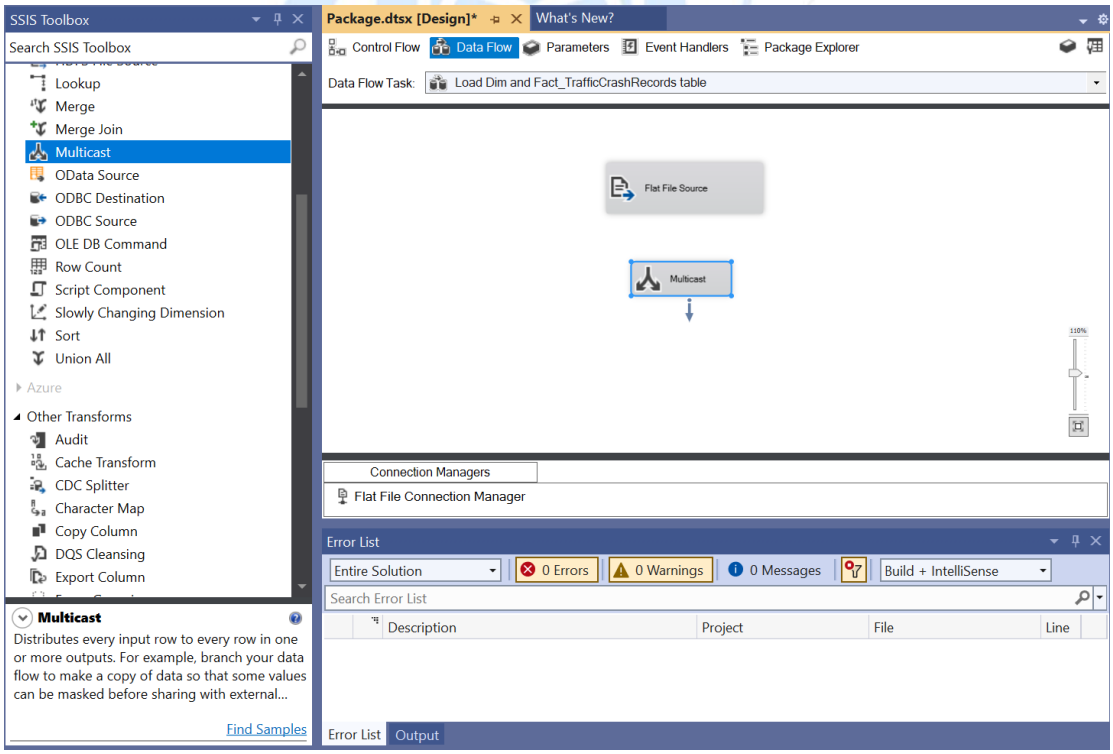
- Nếu muốn giữ lại những giá trị null cần thiết, tích vào "*Retain null values from the sources as null values in the data flow*".



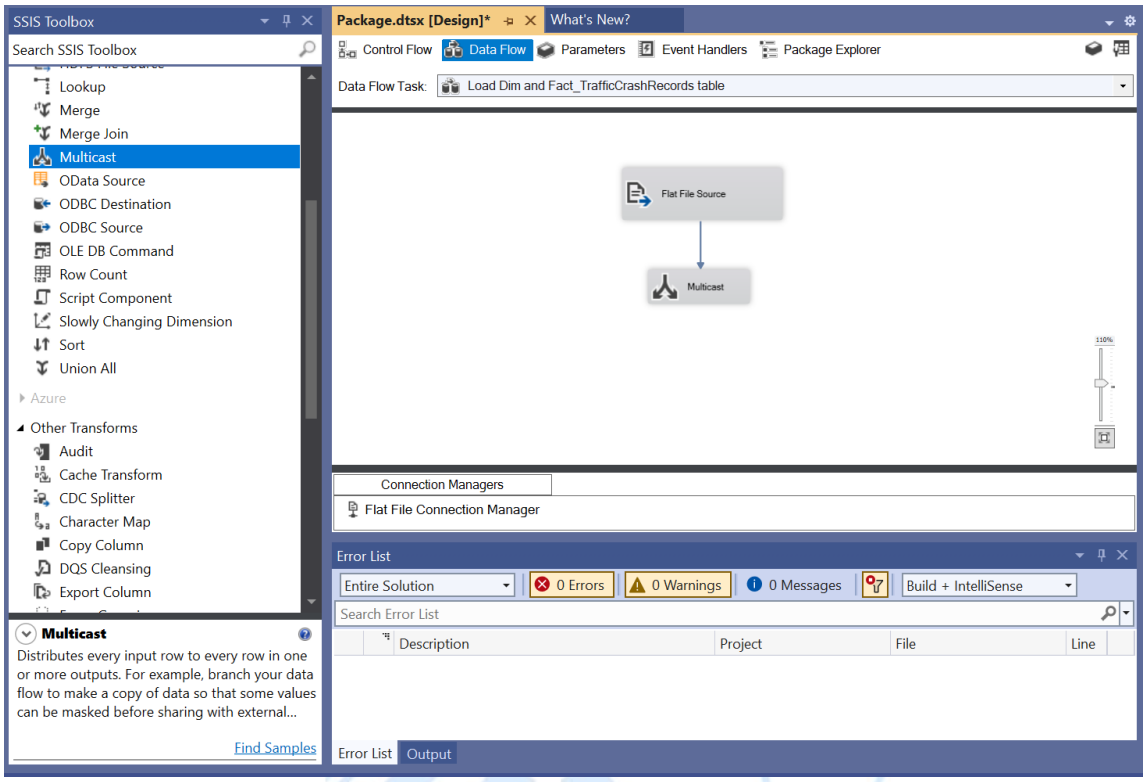
- Sau đó nhấn "OK" để hoàn tất việc import dữ liệu.



- Tạo Multicast để phân tán dữ liệu từ Flat File Source đến các bảng Dimension.

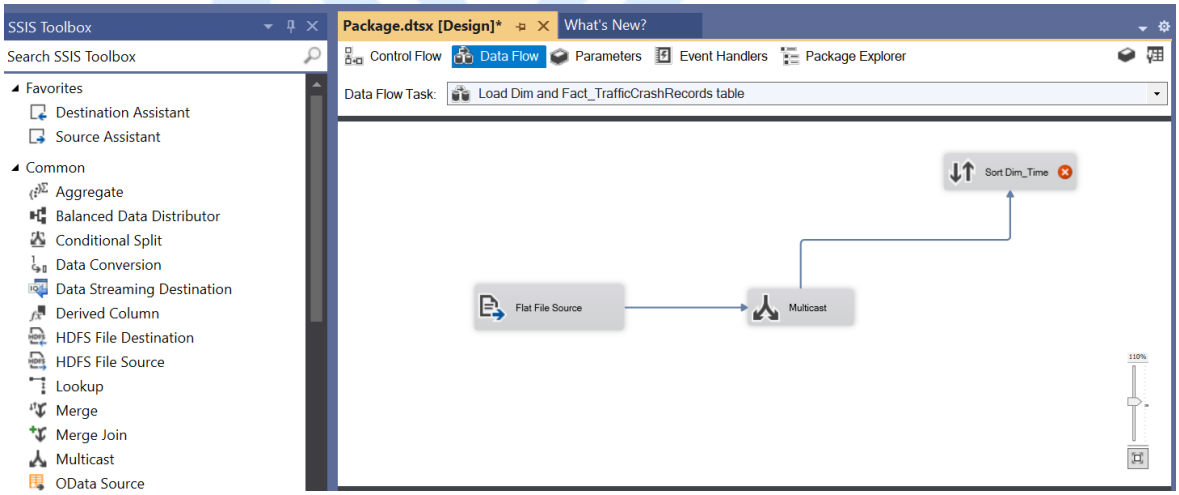


- Sau đó tiến hành kết nối Flat File Source và Multicast.

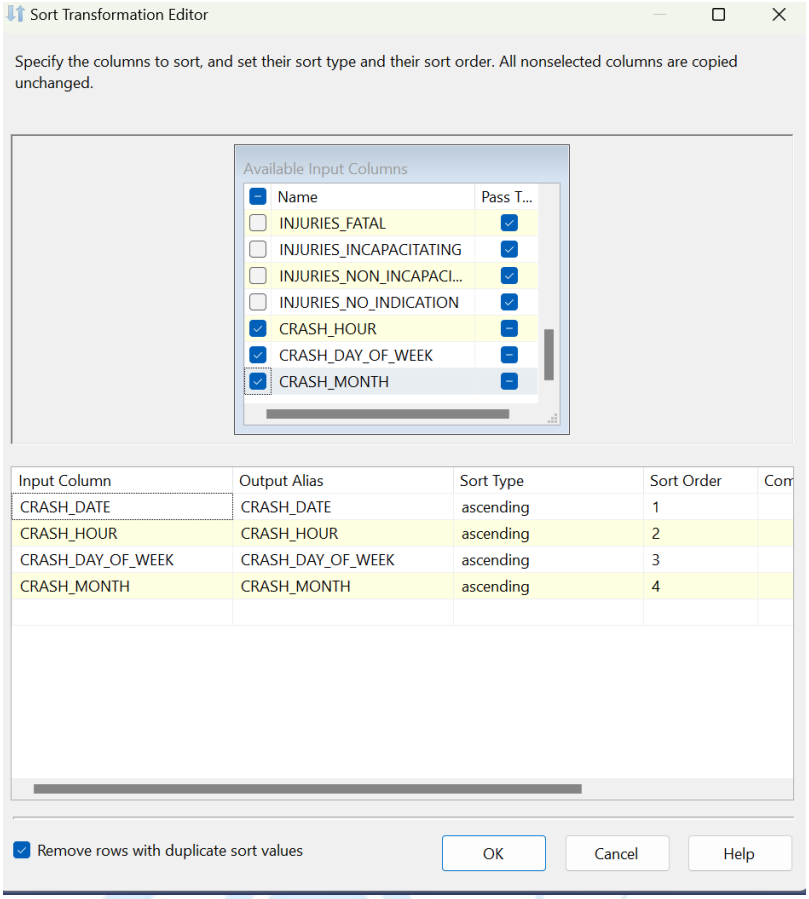


2.3.1 *Tạo bảng Dim_Time*

Bước 1: Tạo một "Sort", đặt tên là "Sort Dim_Time". Kết nối từ "Multicast" đến "Sort Dim_Time".

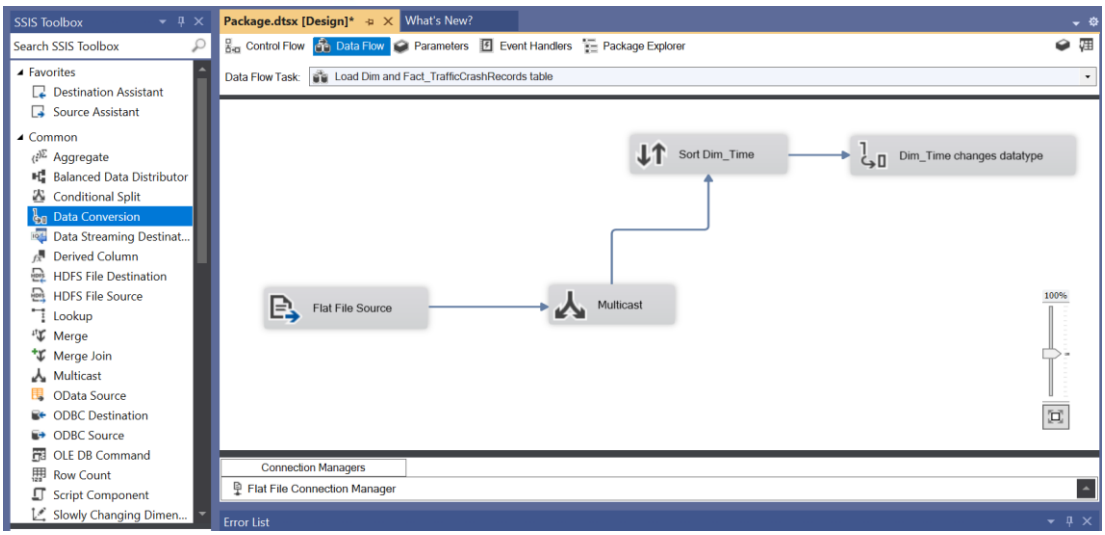


- Nhấn chuột phải vào "Sort Dim_Time", chọn "Edit".
- Chọn "CRASH_DATE" để lấy dữ liệu cho "Sort Dim_Time".
- Chọn vào "*Remove rows with duplicate sort values*" để xóa các dòng dữ liệu trùng nhau. Sau đó chọn "OK" để hoàn tất.

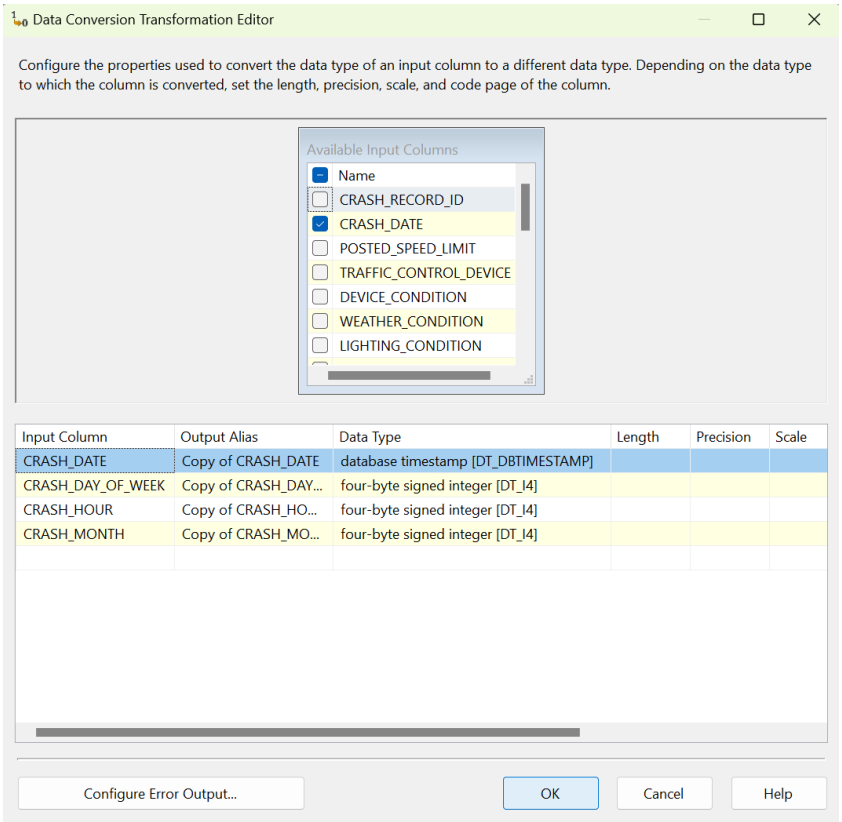


Bước 2: Đổi kiểu dữ liệu "CRASH_DATE"

- Đầu tiên, tạo một "Data Conversion", đổi tên thành "Dim_Time changes datatype". Trong đây sẽ chuyển CRASH_DATE kiểu String thành kiểu Date.
- Tạo kết nối từ "Sort Dim_Time" đến "Dim_Time changes datatype".

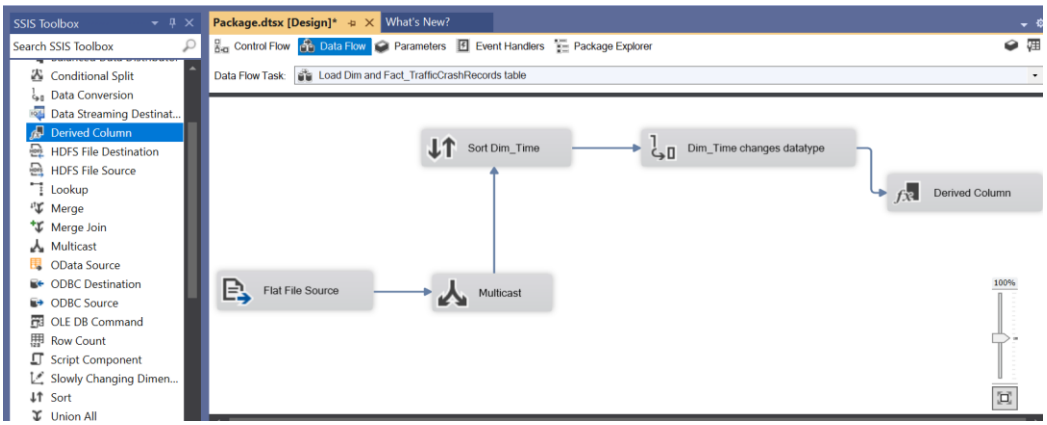


Bước 3: Nhấn chuột phải vào "Dim_Date changes datatype", chọn "Edit". Chọn "CRASH_DATE", "CRASH_HOUR", "CRASH_DAY_OF_WEEK", "CRASH_MONTH" cần chuyển đổi kiểu, chọn kiểu dữ liệu trong mục "Data Type". Sau đó nhấn "OK" để hoàn tất.



Bước 4: Có thể tách CRASH_DATE thành các cột day, month, year, hour.

- Đầu tiên, tạo một "Derived Column", tạo kết nối từ "Dim_Time changes datatype" đến "Derived Column".

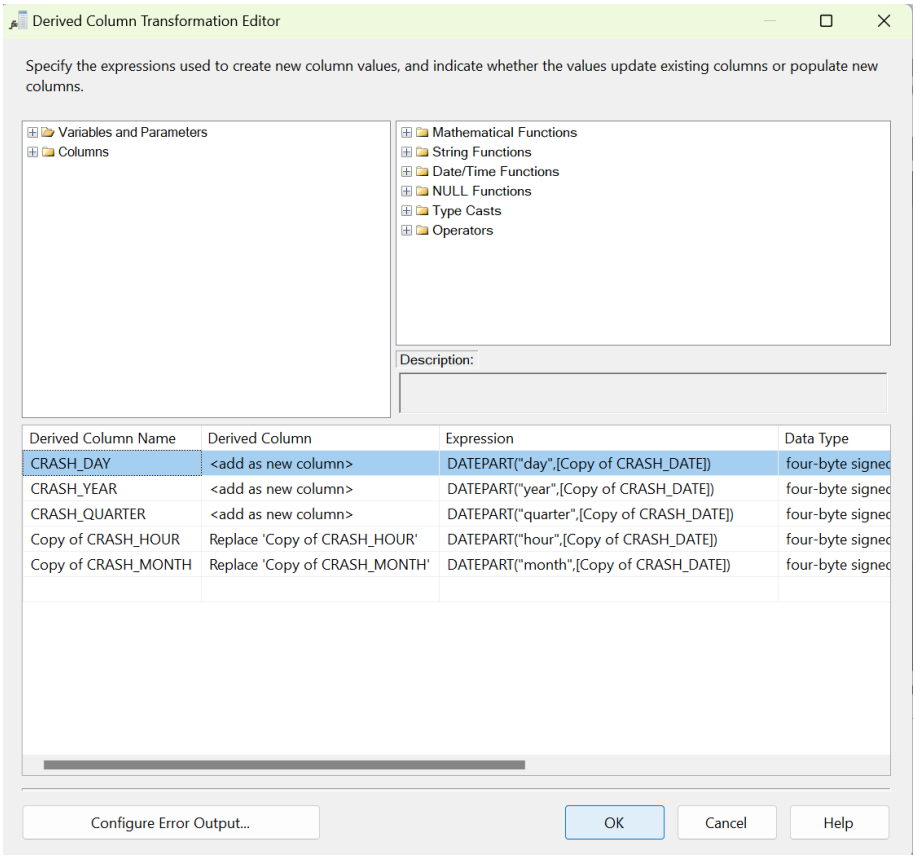


Bước 5: Tạo các cột CRASH_DAY, CRASH_MONTH, CRASH_YEAR, CRASH_HOUR, CRASH_QUARTER từ CRASH_DATE (Ở bộ dữ liệu đã có sẵn cột

CRASH HOUR, CRASH_MONTH nêm có thể chọn thay thế cột (phải có cùng kiểu dữ liệu) hoặc giữ nguyên).

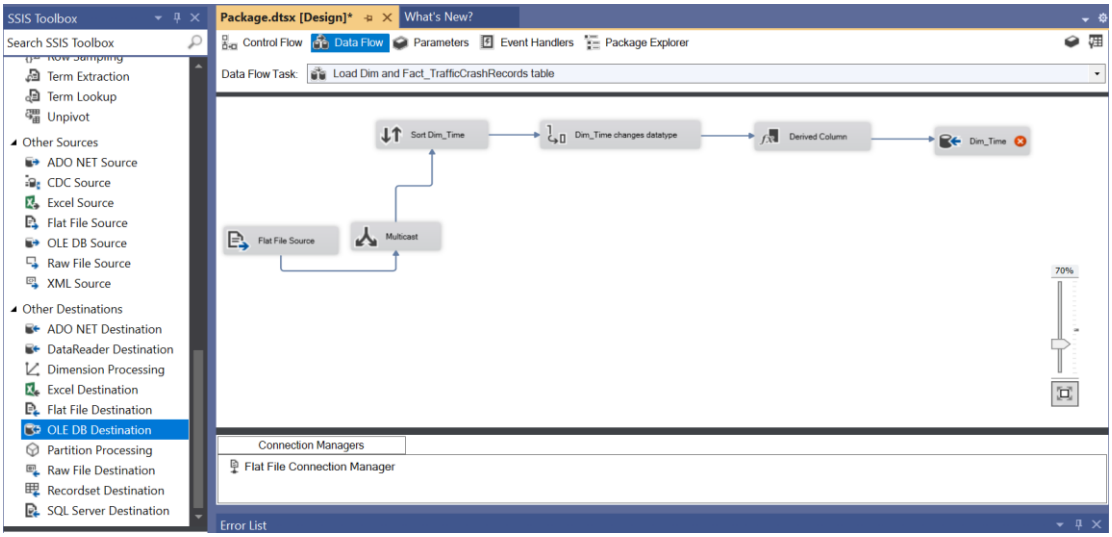
- Nhấn chuột phải vào “Derived Column”, chọn “Edit”.

- Chọn hàm DATEPART(<kiểu dữ liệu thời gian>, cột dữ liệu) trong mục “Date/Time Functions” để chia “DAY” từ “CRASH_DATE”, ta cài đặt: DATEPART(“day”, [Copy of CRASH_DATE]).
- Cột dữ liệu [Copy of CRASH_DATE] là cột dữ liệu kiểu DateTime được chuyển từ CRASH_DATE kiểu String đã thực hiện ở bước trước đó.
  “Derived Column Name” là tên cột sẽ được thêm mới hoặc thay thế.
- Tiến hành tương tự với “CRASH_MONTH”, “CRASH_YEAR” và “CRASH_HOUR”, “CRASH_QUARTER”.

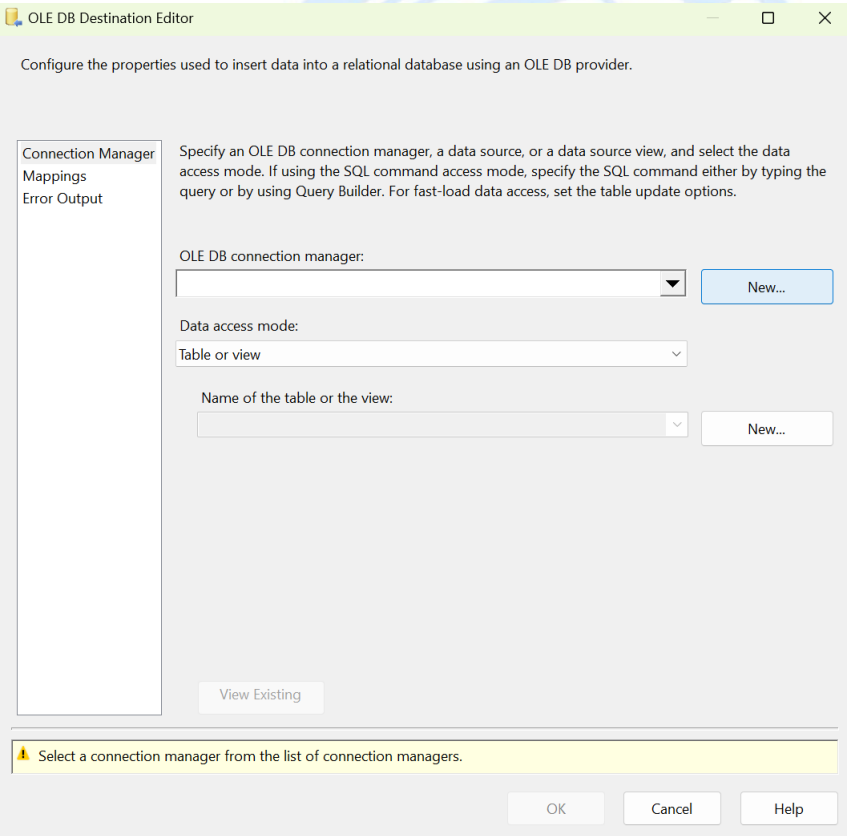


Bước 6: Tiến hành đổ dữ liệu bảng Dim_Time vào TrafficCrash.

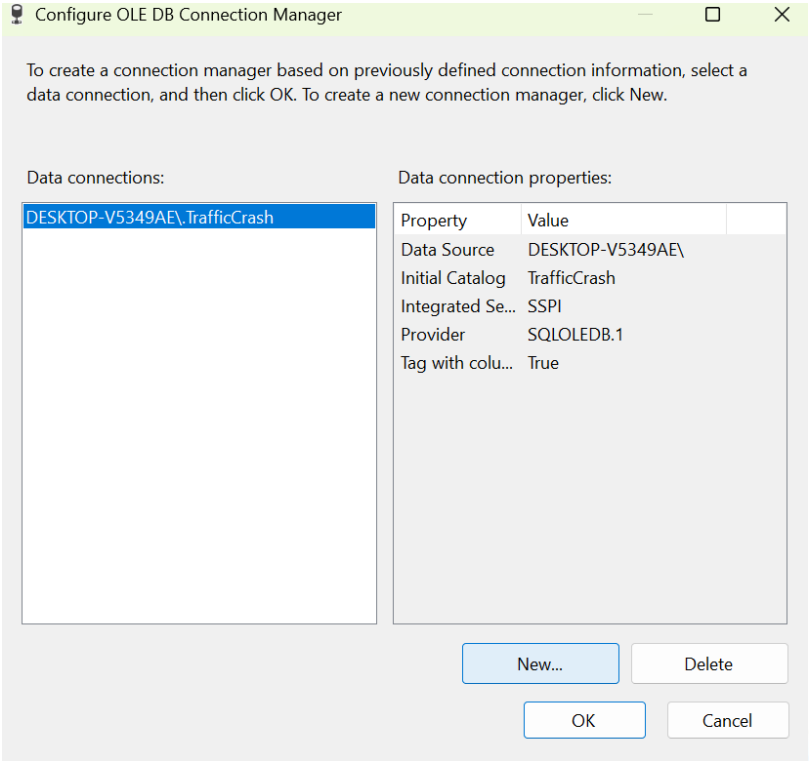
- Đầu tiên, tạo một OLE DB Destination, đặt tên là “Dim_Time”, kết nối từ “Derived Column” đến “Dim_Time”.



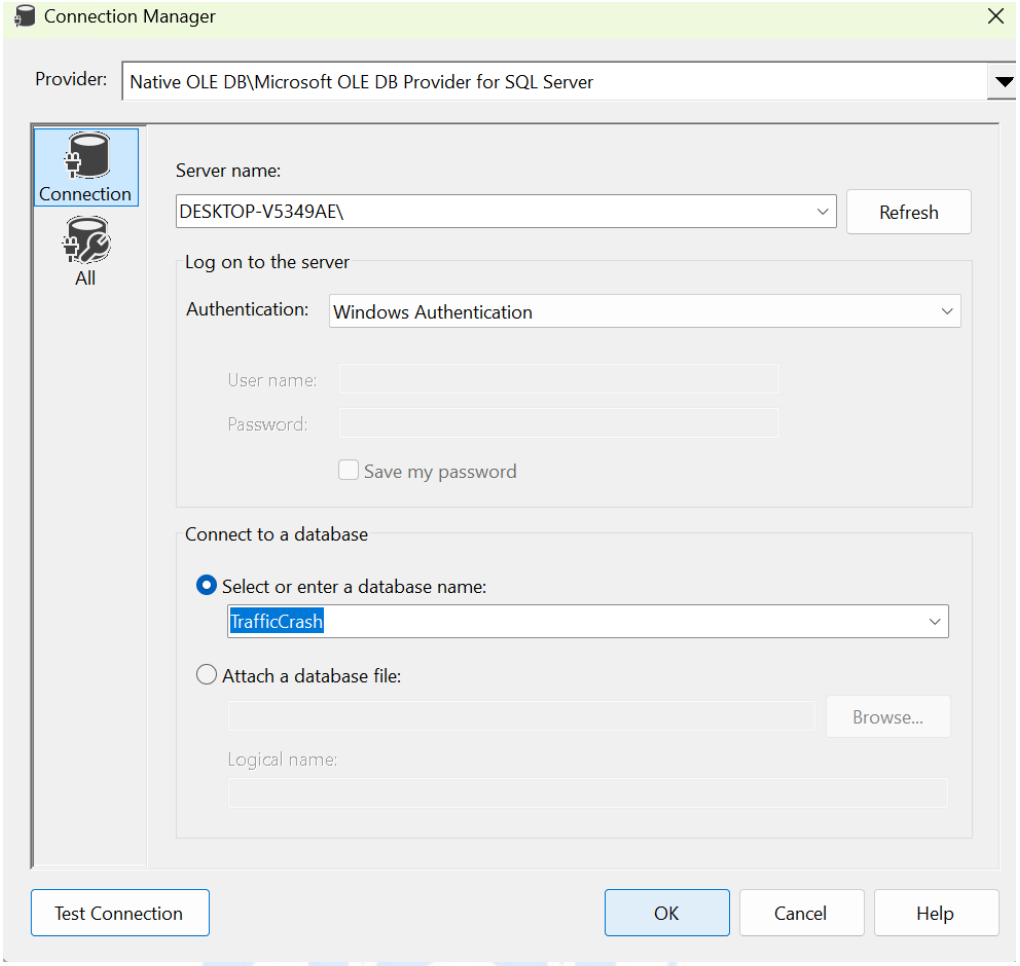
Bước 7: Nhấn chuột phải vào "Dim_Time", chọn "Edit". Sau đó chọn vào "New".



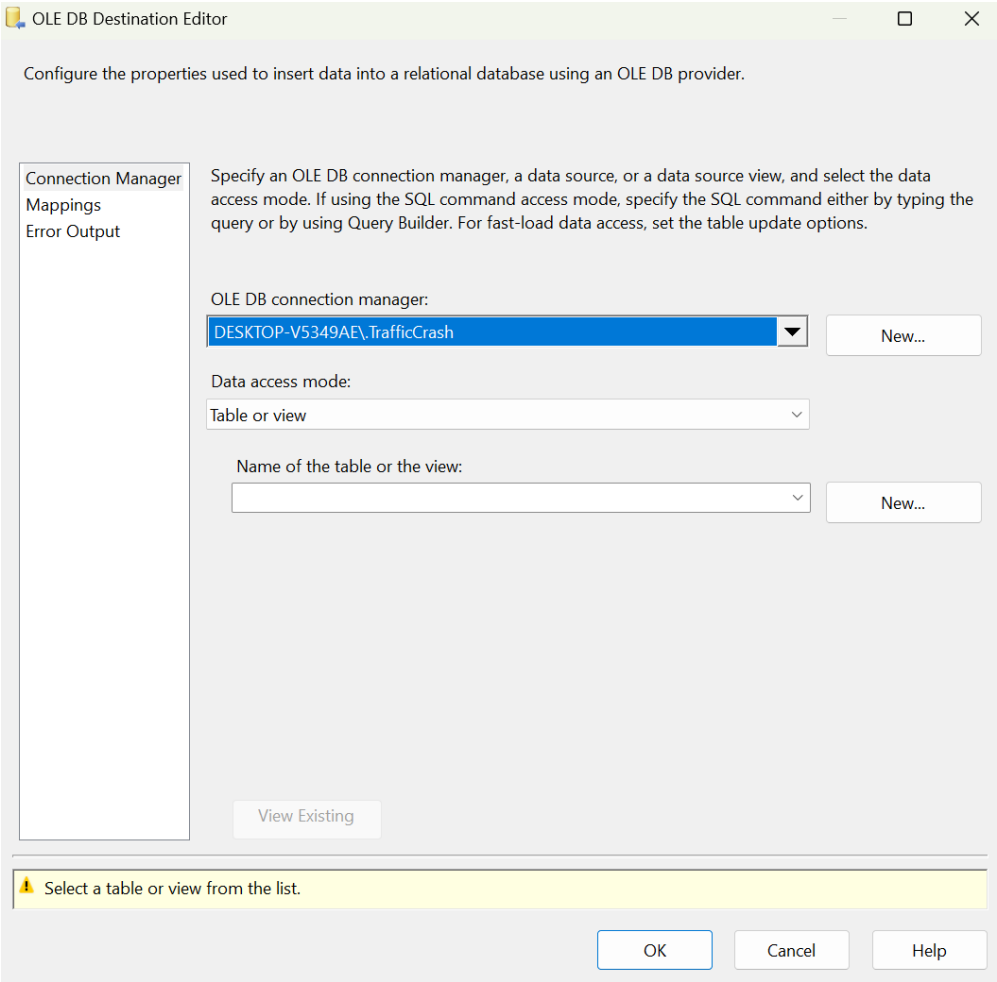
- Tiếp tục chọn vào "New" để tạo kết nối OLE DB mới.



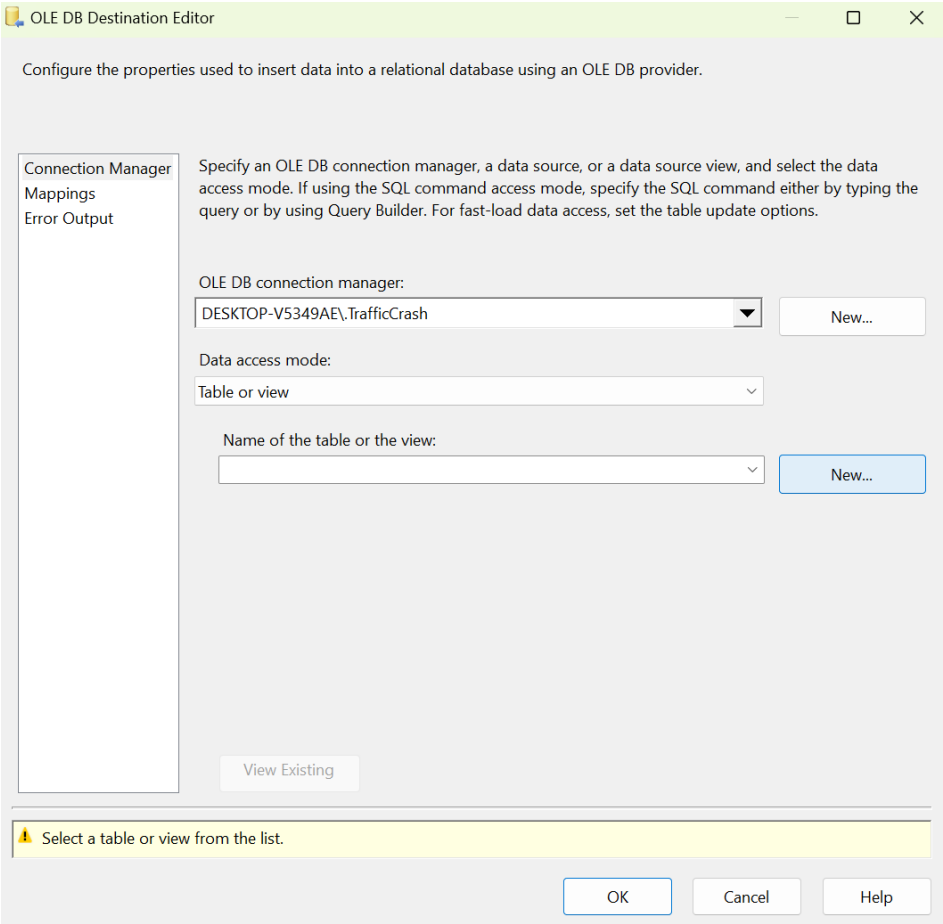
- Vào mục Provider để chọn provider thích hợp.
- Chọn Server name trùng với Server name MS SQL Server để có thể kết nối đến datawarehouse TrafficaCrash vừa tạo.
- Kết nối đến server bằng tài khoản window mặc định (Windows Authentication) Nhấn "Test Connection" để kiểm tra kết nối và sau đó nhấn "OK".



Bước 8: Chọn vào Connection vừa tạo.

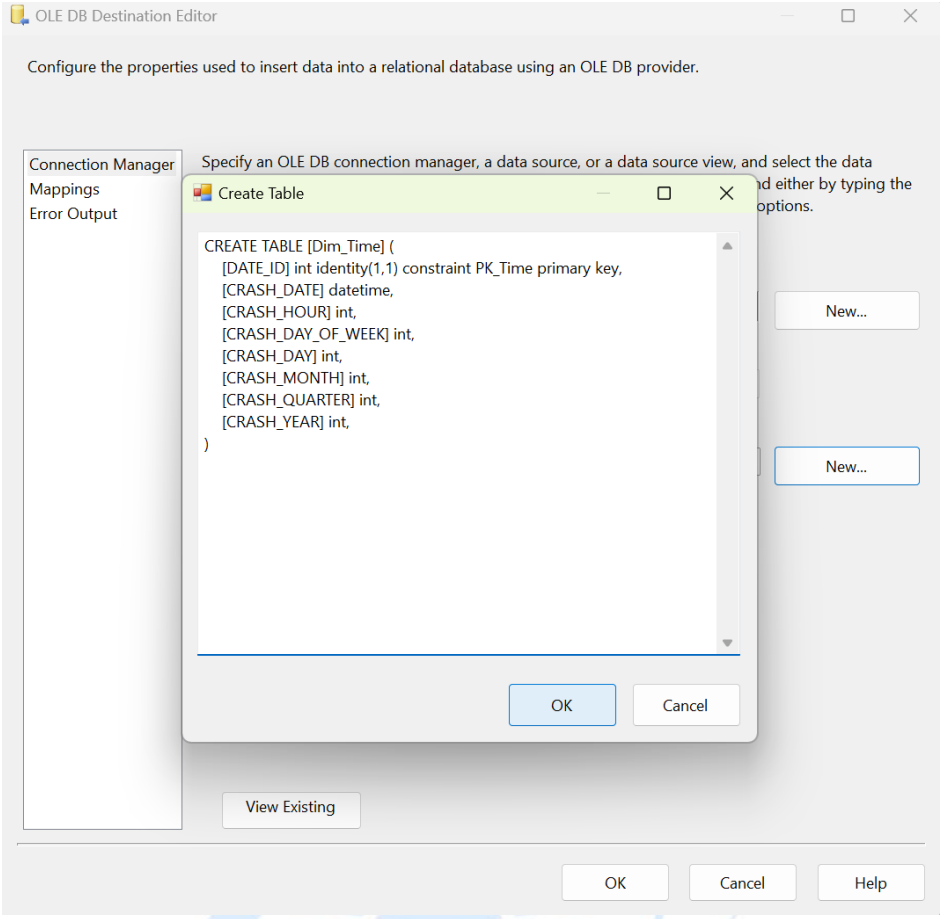


Bước 9: Chọn "New" để tạo bảng Dim_Date mới trong database.

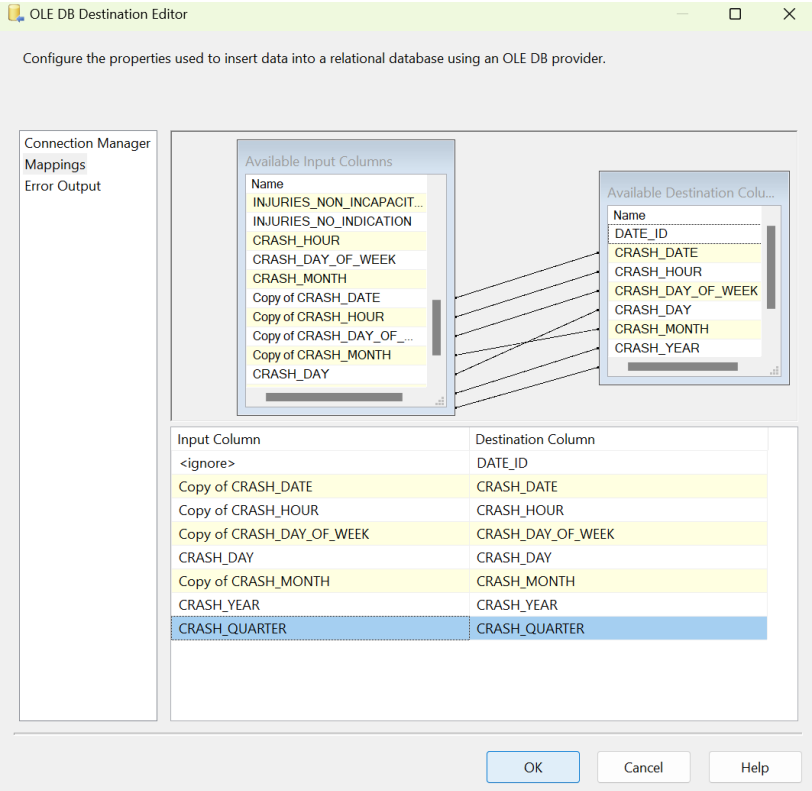


- Viết lệnh SQL để tạo bảng, sau đó nhấn "OK".

```
CREATE TABLE [Dim_Time] (
    [DATE_ID] int identity(1,1) constraint PK_Time primary key,
    [CRASH_DATE] datetime,
    [CRASH_HOUR] int,
    [CRASH_DAY_OF_WEEK] int,
    [CRASH_DAY] int,
    [CRASH_MONTH] int,
    [CRASH_QUARTER] int,
    [CRASH_YEAR] int,
)
```



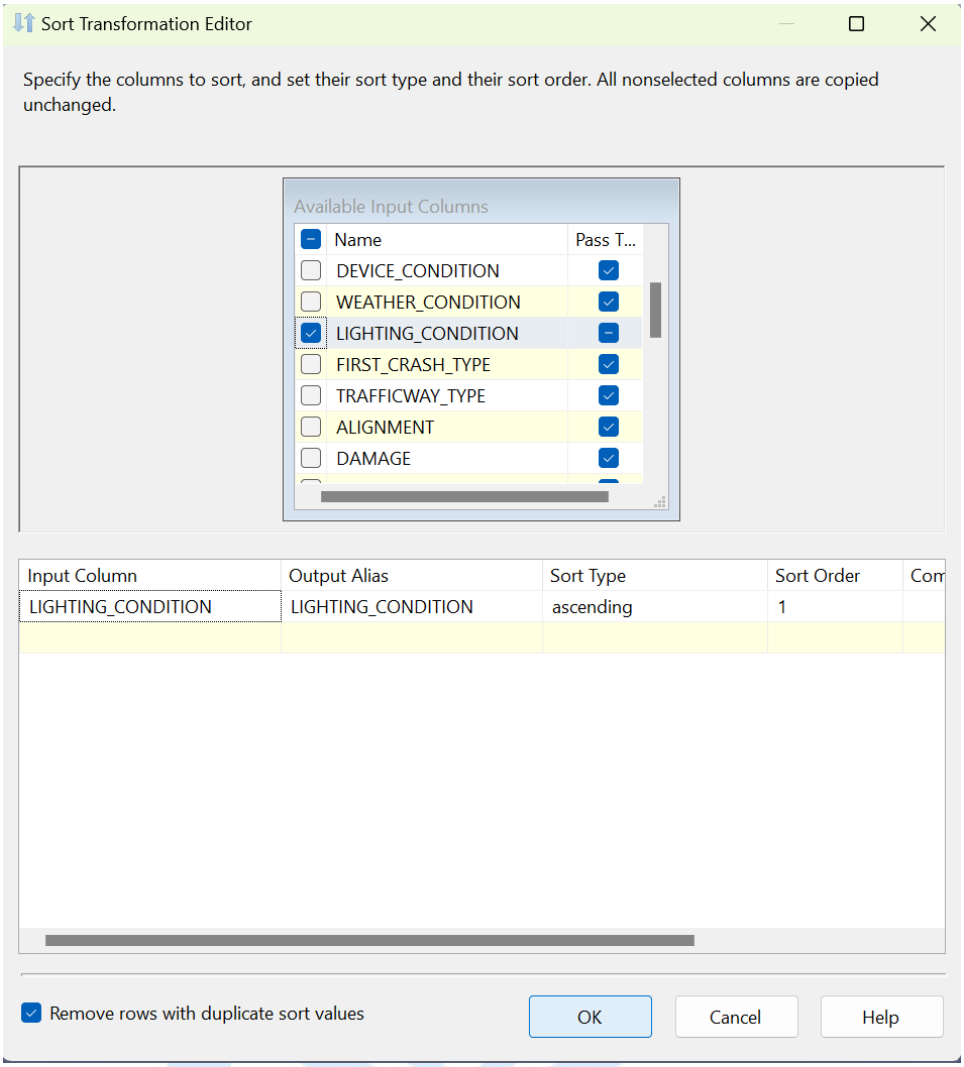
- Để kiểm tra việc ánh xạ các cột dữ liệu, ta vào mục “Mappings”. Đổi lại “Input Column” cho các cột. Sau đó nhấn “OK” để hoàn tất.



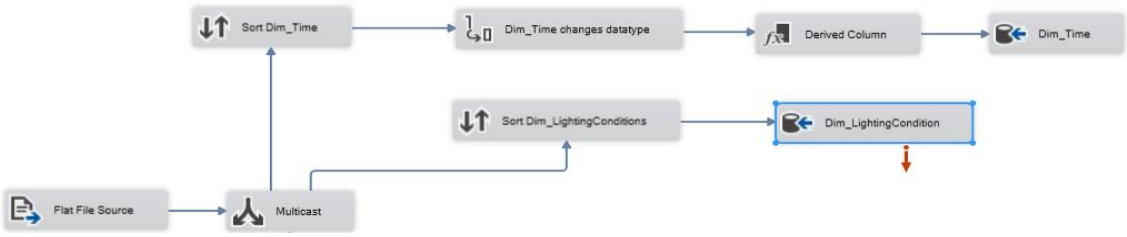
2.3.2 Tạo bảng Dim_ LightingConditions

Bước 1: Tạo một "Sort", đặt tên là "Sort Dim_ LightingConditions". Kết nối từ "Multicast" đến "Dim_LightingConditions".

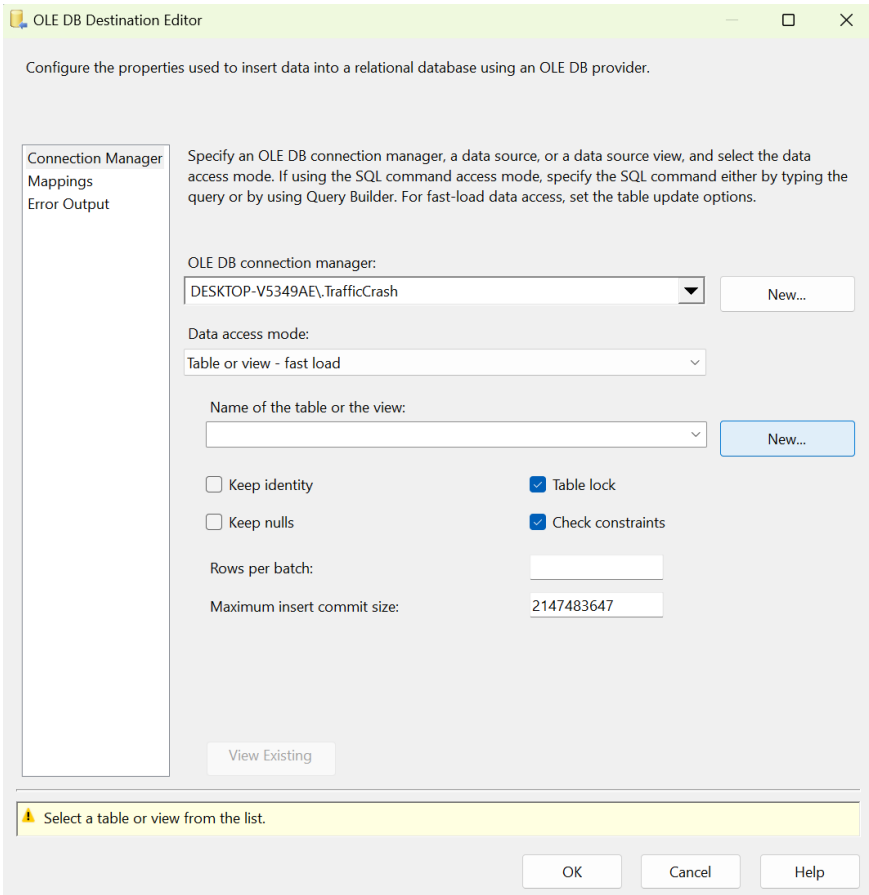
- Nhấn chuột phải vào "Sort Dim_ LightingConditions", chọn "Edit".
- Chọn "LIGHTING_CONDITION" để làm dữ liệu cho "Sort Dim_ LightingConditions".



Bước 2: Tiến hành đổ dữ liệu bảng Dim_LightingConditions vào TrafficCrash. Đầu tiên, tạo một OLE DB Destination, đặt tên là "Dim_LightingConditions".

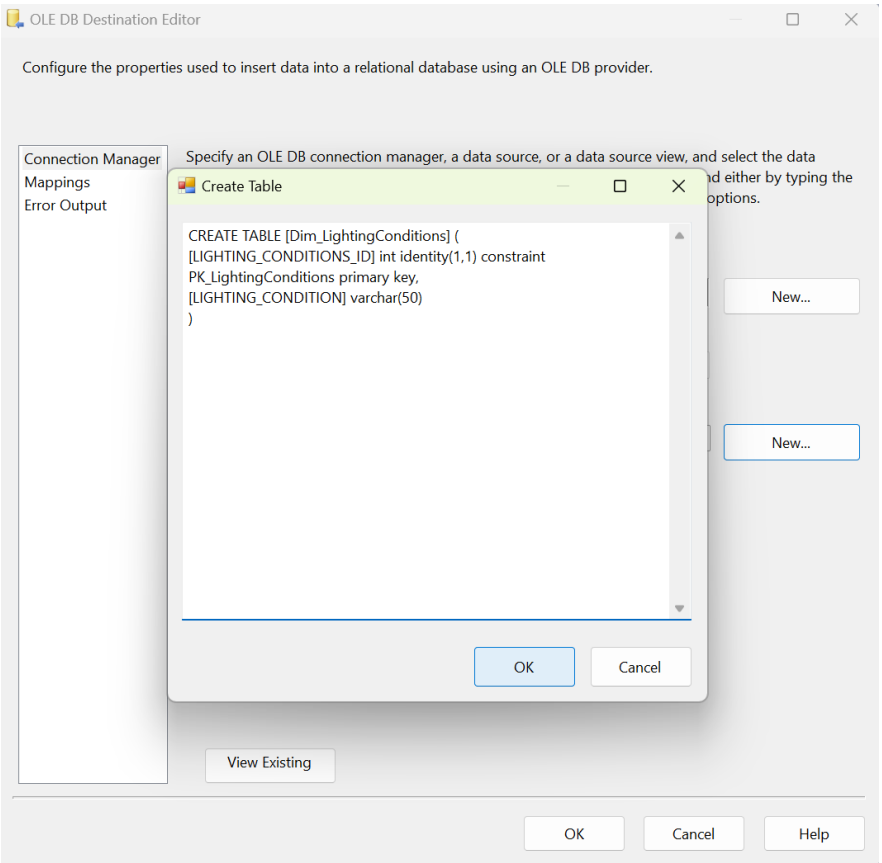


Bước 3: Nhấn chuột phải vào "Dim_ LightingConditions", chọn "Edit". Sau đó chọn "New" để tạo bảng Dim_ LightingConditions mới trong database.

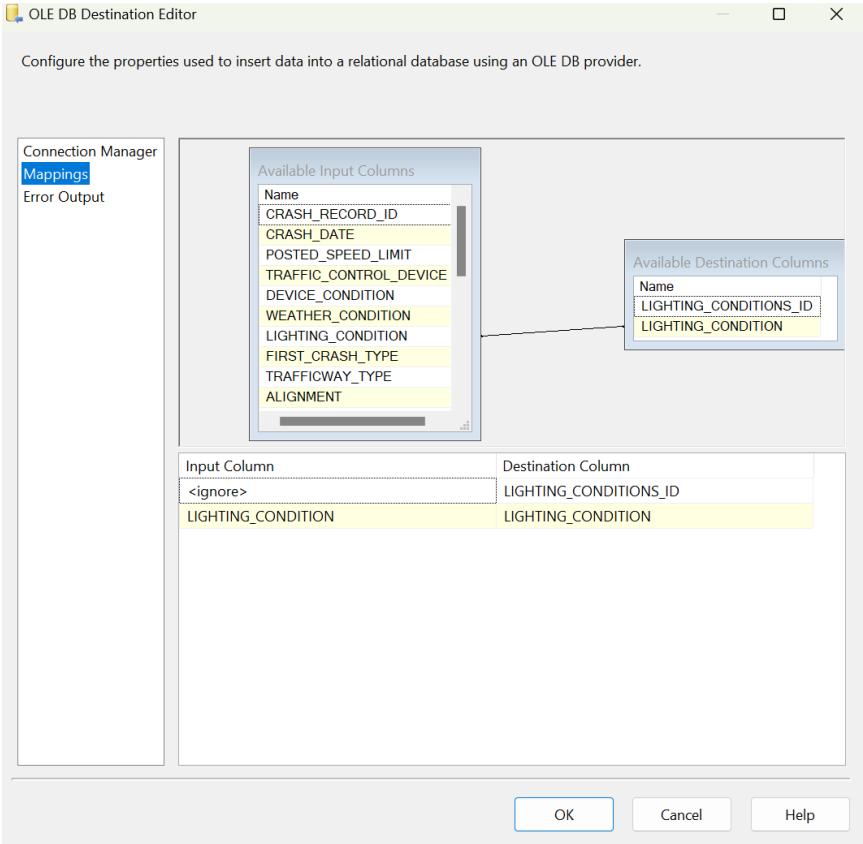


- Viết lệnh SQL để tạo bảng, nhấn "OK" để tạo.

```
CREATE TABLE [Dim_LightingConditions] (
[LIGHTING_CONDITION_ID] int identity(1,1) constraint PK_LightingConditions primary key,
[LIGHTING_CONDITION] varchar(50)
)
```

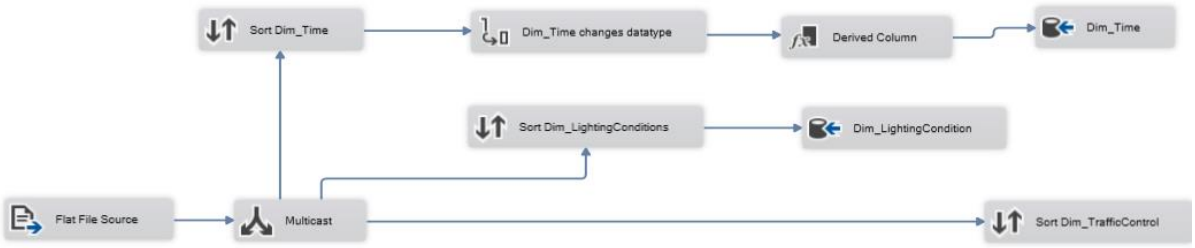


- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



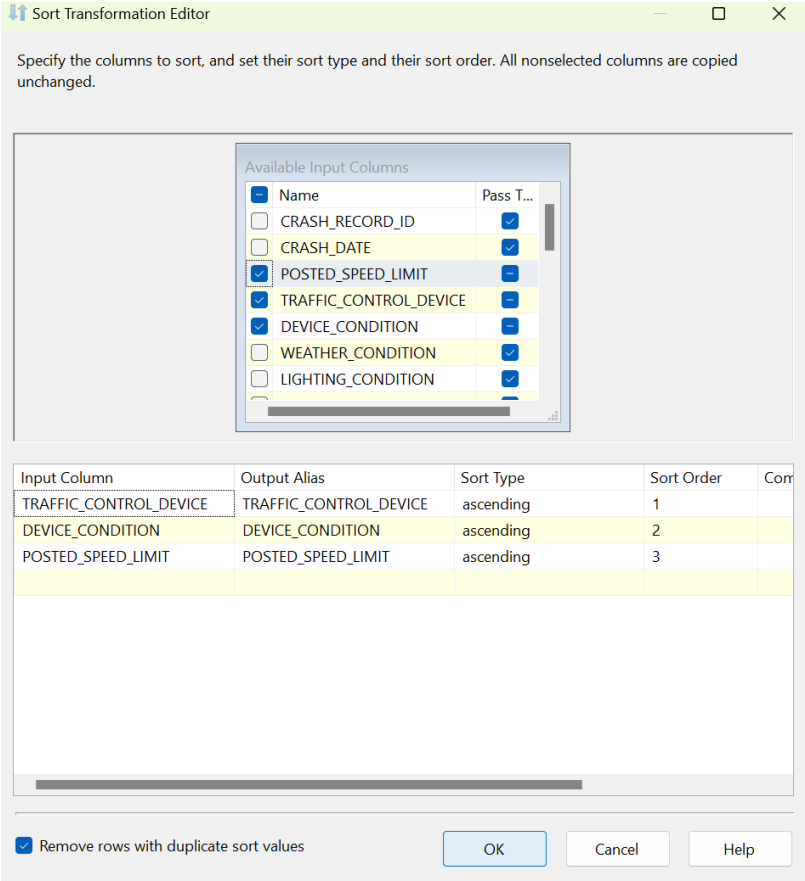
2.3.3 Tạo bảng Dim_TrafficControl

Bước 1: Tạo một "Sort", đặt tên là "Sort Dim_TrafficControl".

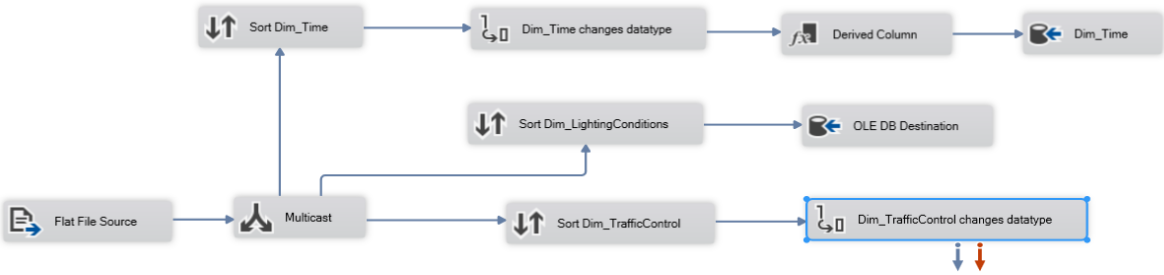


Bước 2: Nhấn chuột phải vào "Sort Dim_ TrafficControl", chọn "Edit".

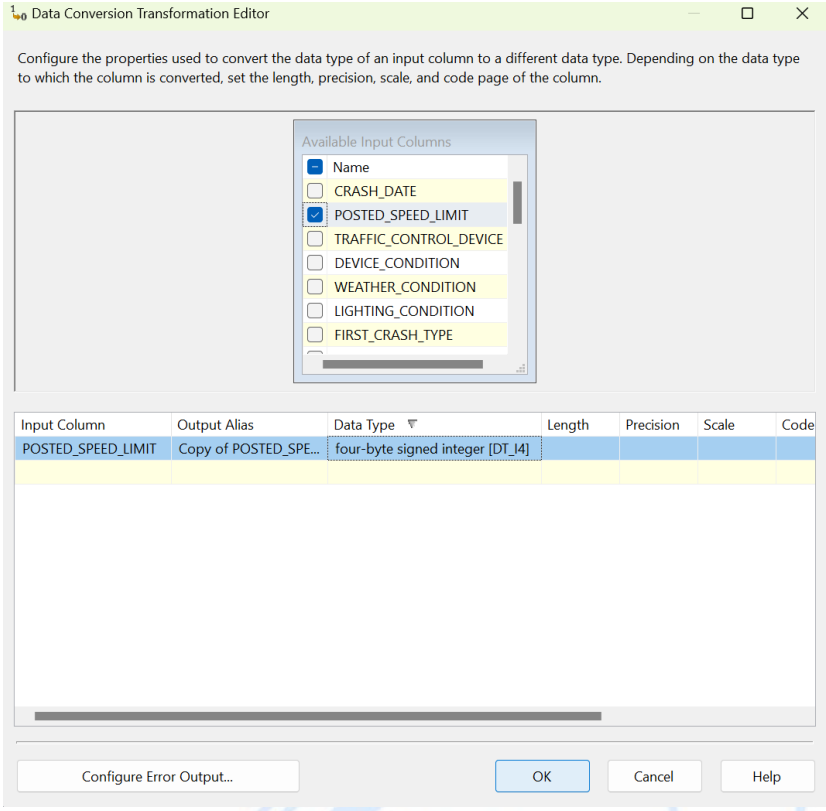
- Chọn các thuộc tính để làm dữ liệu cho "Sort Dim_ TrafficControl".
- Chọn vào "*Remove rows with duplicate sort values*" để xóa các dòng dữ liệu trùng nhau. Sau đó chọn "OK" để hoàn tất.



Bước 3: Tiến hành tạo một "Data Conversion" đặt tên là "Dim_TrafficControl changes datatype" để đổi kiểu dữ liệu của "POSTED_SPEED_LIMIT".

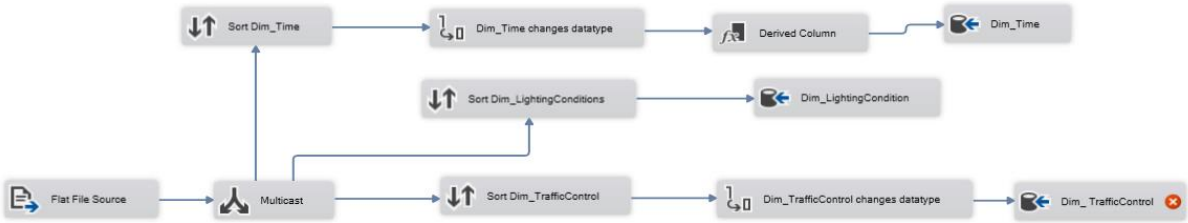


- Đổi kiểu dữ liệu của "POSTED_SPEED_LIMIT" thành Int. Sau đó nhấn "OK" để hoàn tất.

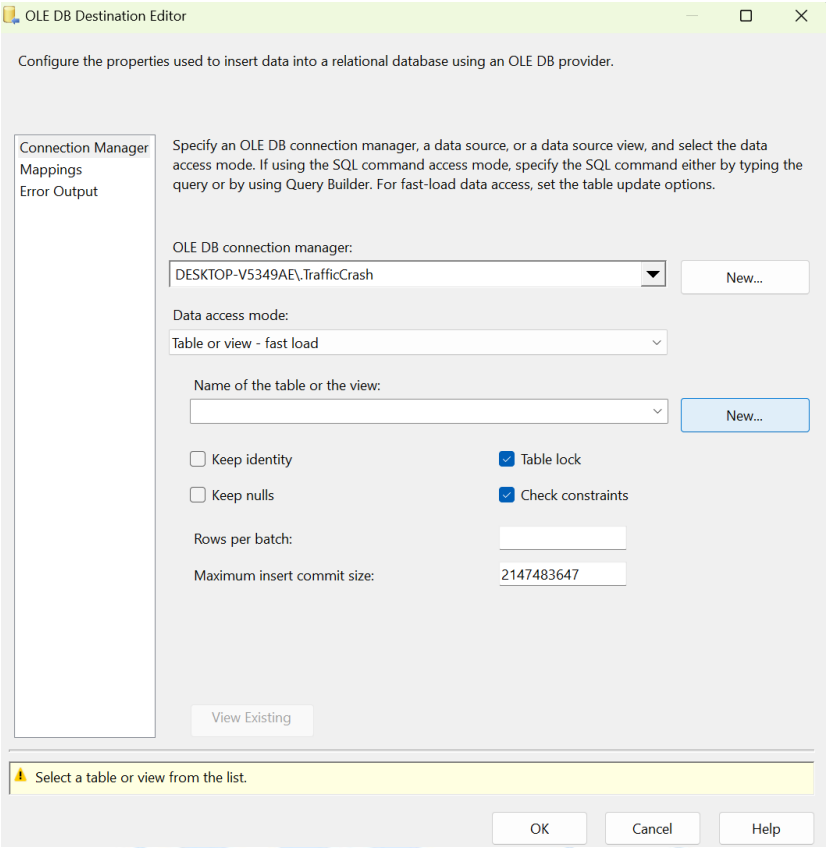


Bước 4: Thực hiện đổ dữ liệu từ bảng Dim_ TrafficControl vào TrafficCrash.

Đầu tiên, tạo một OLE DB Destination, đặt tên là "Dim_ TrafficControl".

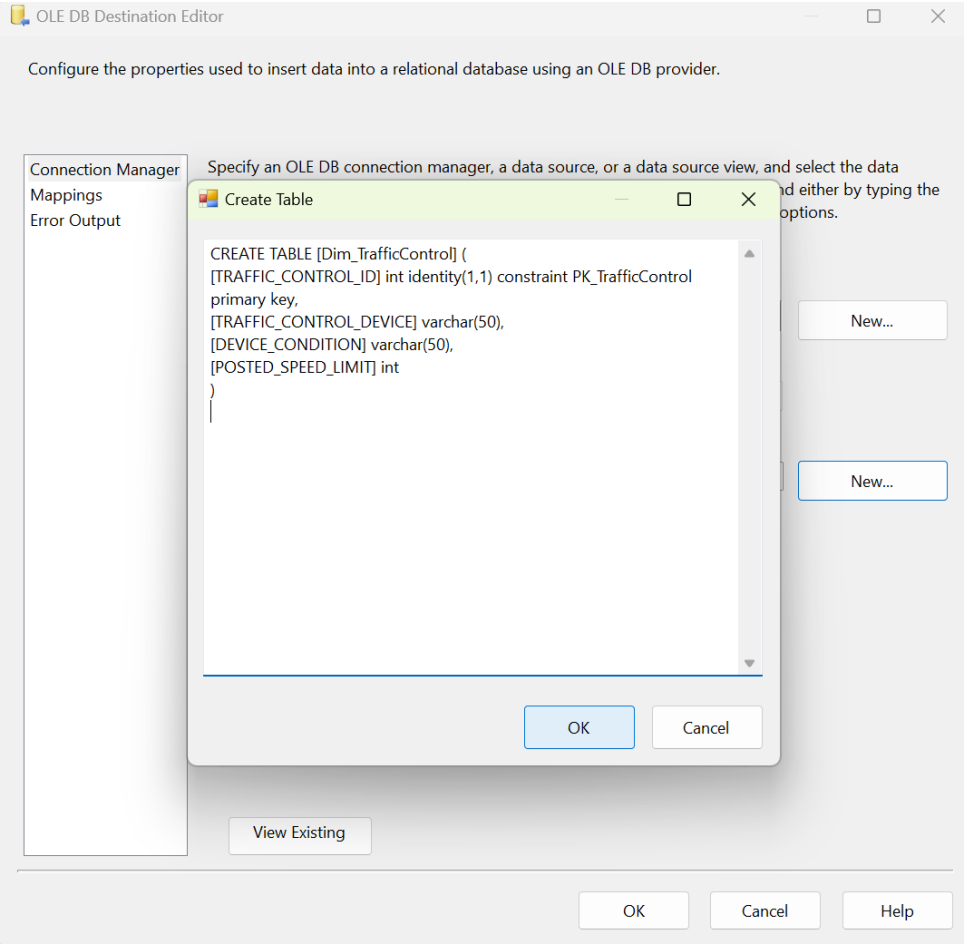


Bước 5: Nhấn chuột phải vào "Dim_ TrafficControl", chọn "Edit". Chọn "New" để tạo bảng Dim_ TrafficControl mới trong database.

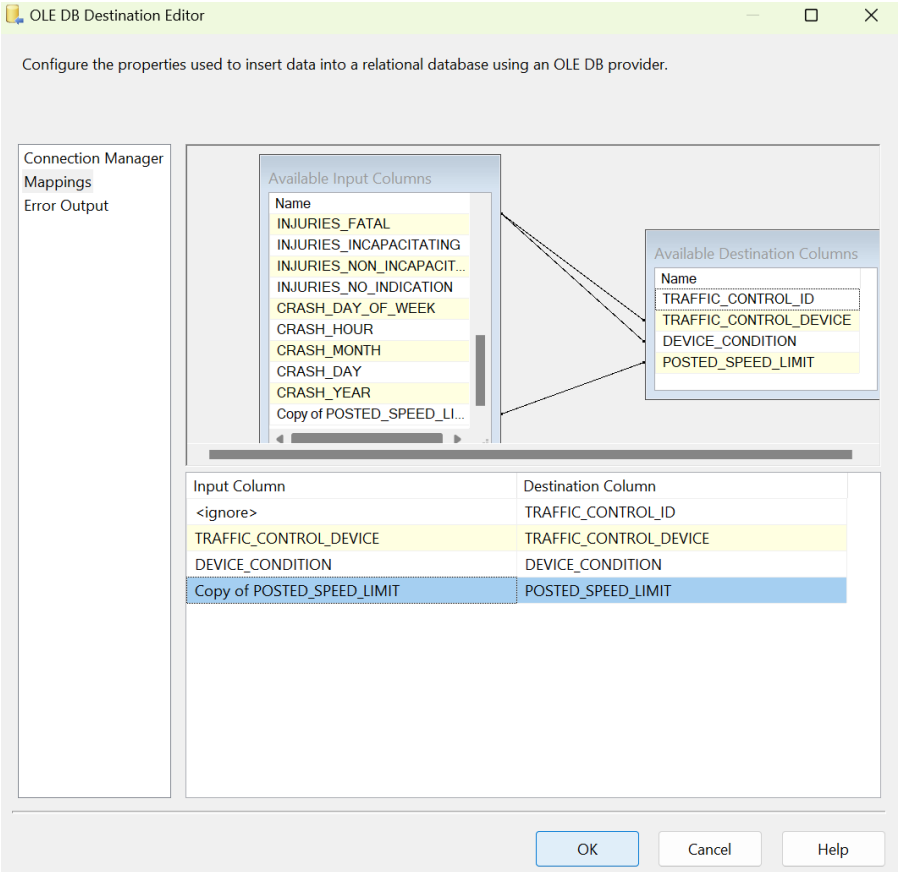


- Viết lệnh SQL để tạo bảng, nhấn "OK" để tạo.

```
CREATE TABLE [Dim_TrafficControl] (
[TRAFFIC_CONTROL_ID] int identity(1,1) constraint PK_TrafficControl primary key,
[TRAFFIC_CONTROL_DEVICE] varchar(50),
[DEVICE_CONDITION] varchar(50),
[POSTED_SPEED_LIMIT] int
)
```

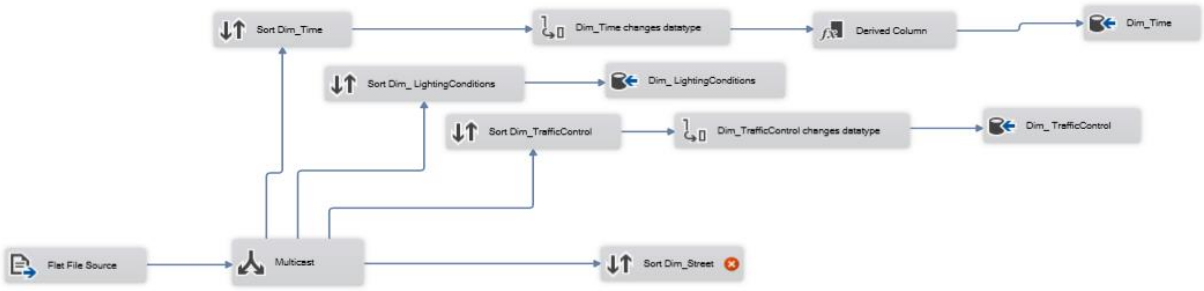


- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Đổi lại “Input Column” cho cột “POSTED_SPEED_LIMIT”. Nhấn “OK” để hoàn tất.



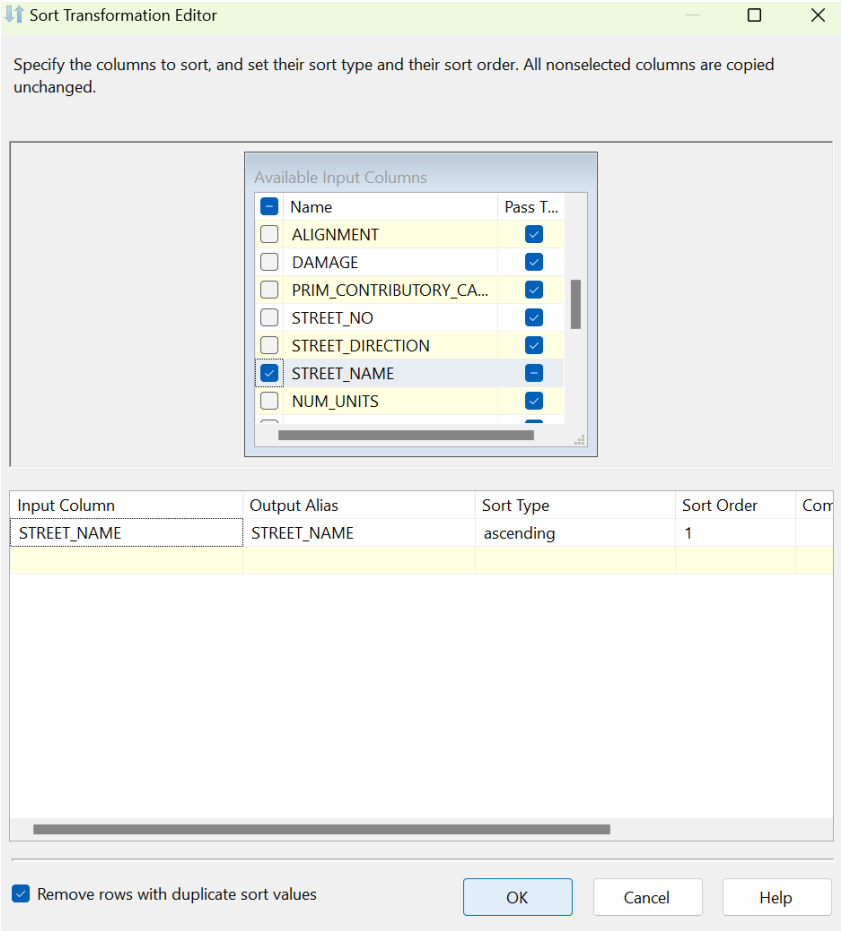
2.3.4 Tạo bảng Dim_Street

Bước 1: Tạo một "Sort", đặt tên là "Sort Dim_Street".

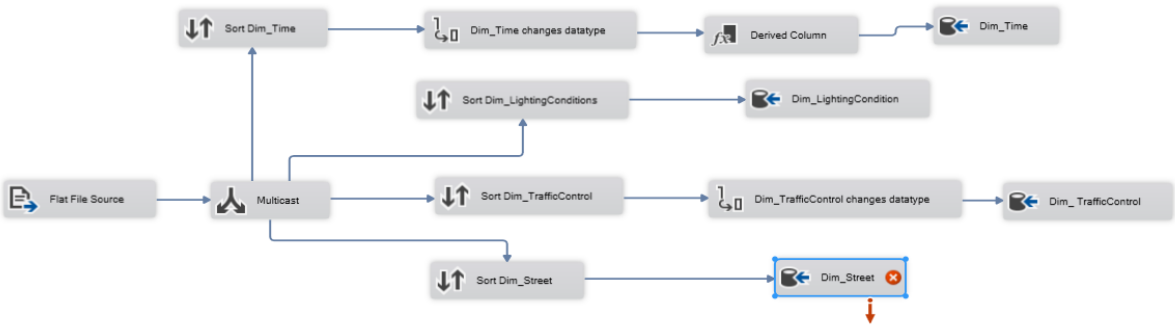


Bước 2: Nhấn chuột phải vào "Sort Dim_ Street", chọn "Edit".

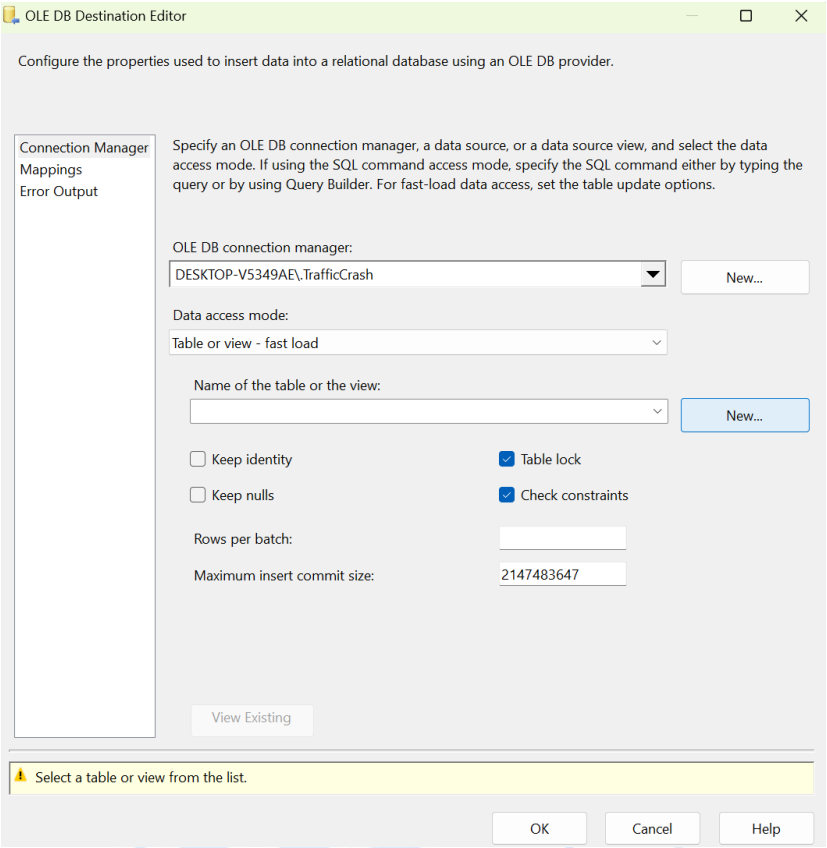
- Chọn các thuộc tính để làm dữ liệu cho "Sort Dim_Street".
- Chọn vào "*Remove rows with duplicate sort values*" để xóa các dòng dữ liệu trùng nhau. Sau đó chọn "OK" để hoàn tất.



Bước 3: Tiến hành đổ dữ liệu bảng Dim_Street vào TrafficCrash. Đầu tiên, tạo một OLE DB Destination, đặt tên là "Dim_Street".

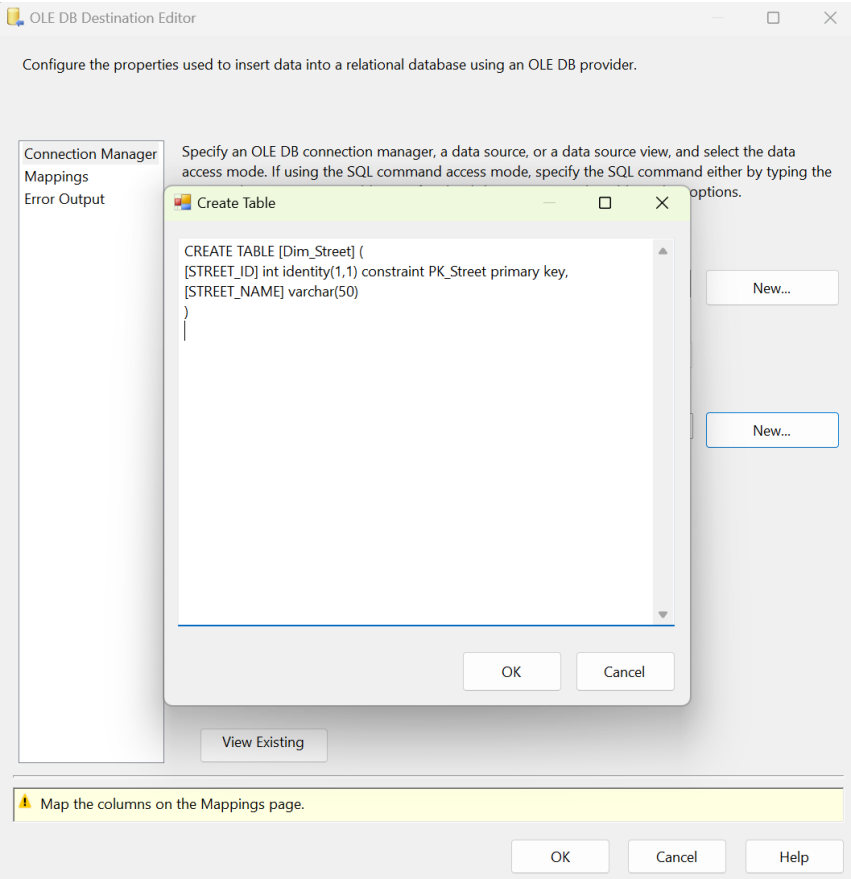


Bước 4: Nhấn chuột phải vào "Dim_Street", chọn "Edit". Chọn "New" để tạo bảng Dim_Street mới trong database.

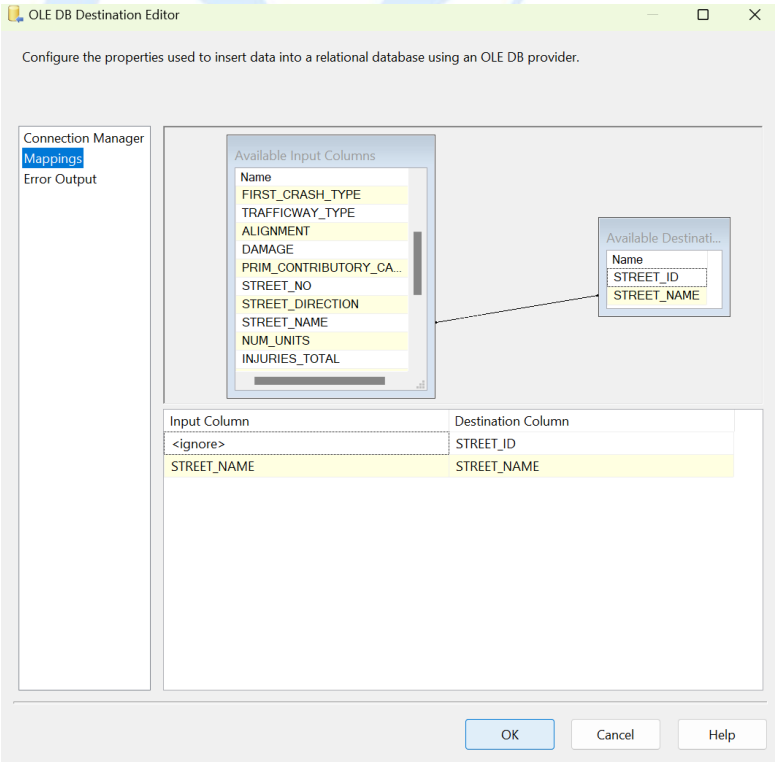


- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Dim_Street] (
[STREET_ID] int identity(1,1) constraint PK_Street primary key,
[STREET_NAME] varchar(50)
)
```

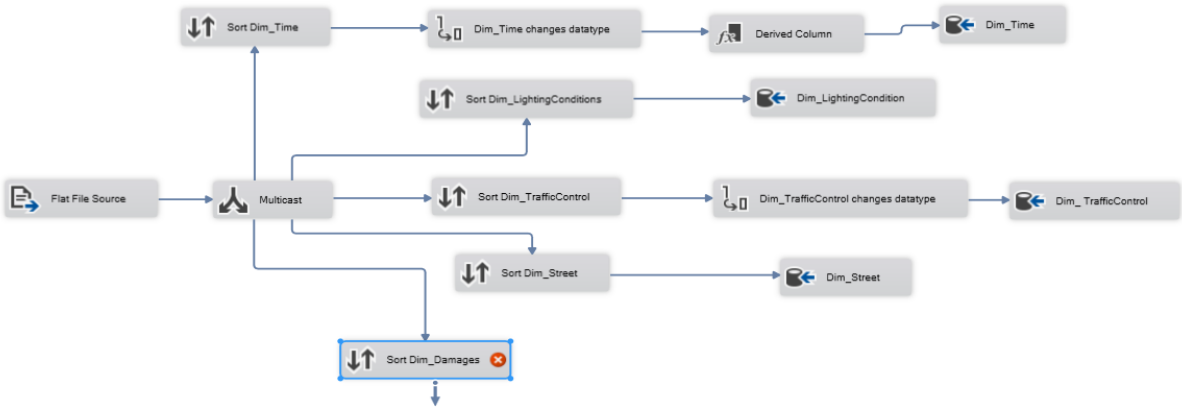


- Vào mục "Mappings" để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn "OK" để hoàn tất.



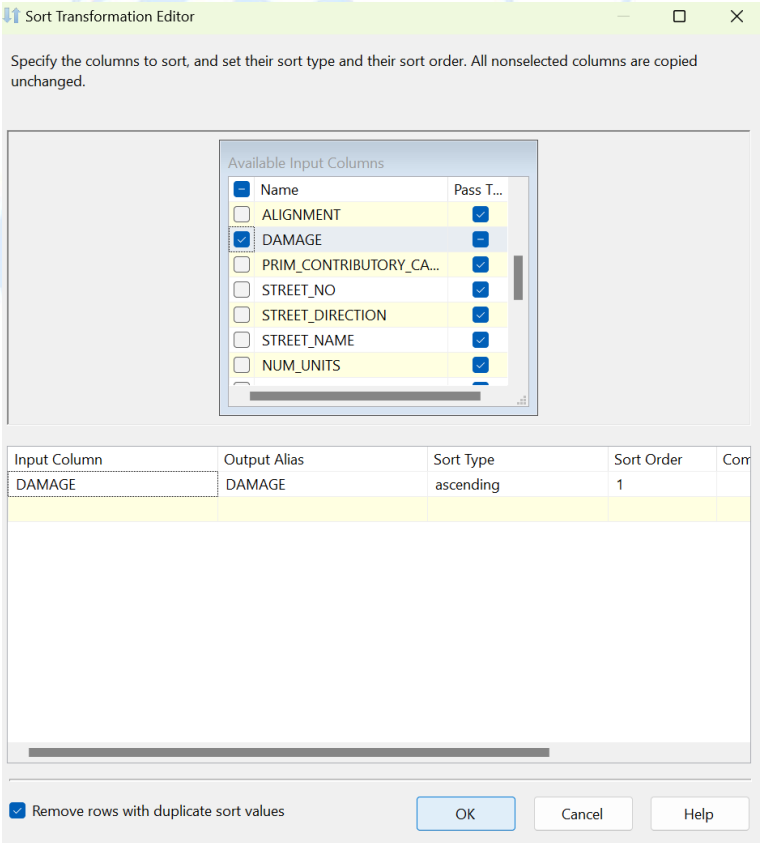
2.3.5 Tạo bảng Dim_Damages

Bước 1: Tạo một "Sort", đặt tên là "Sort Dim_Damages".

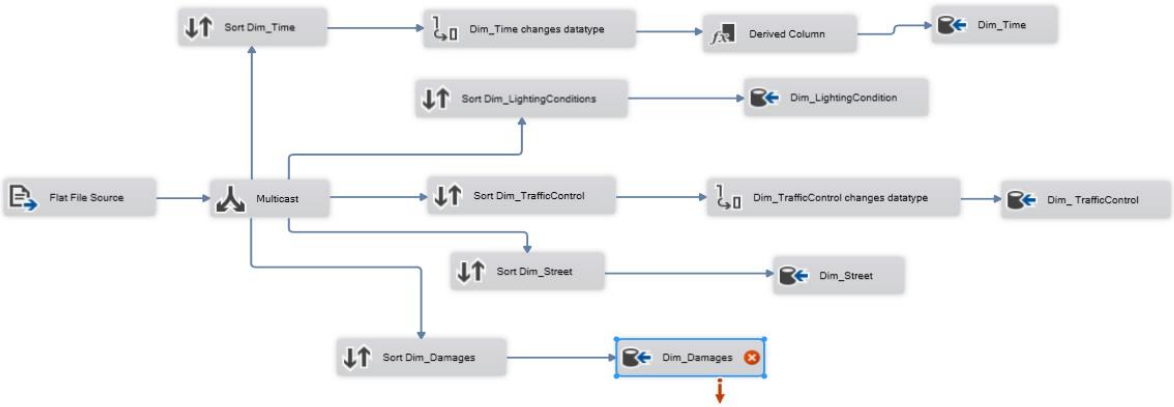


Bước 2: Nhấn chuột phải vào "Sort Dim_ Damages", chọn "Edit".

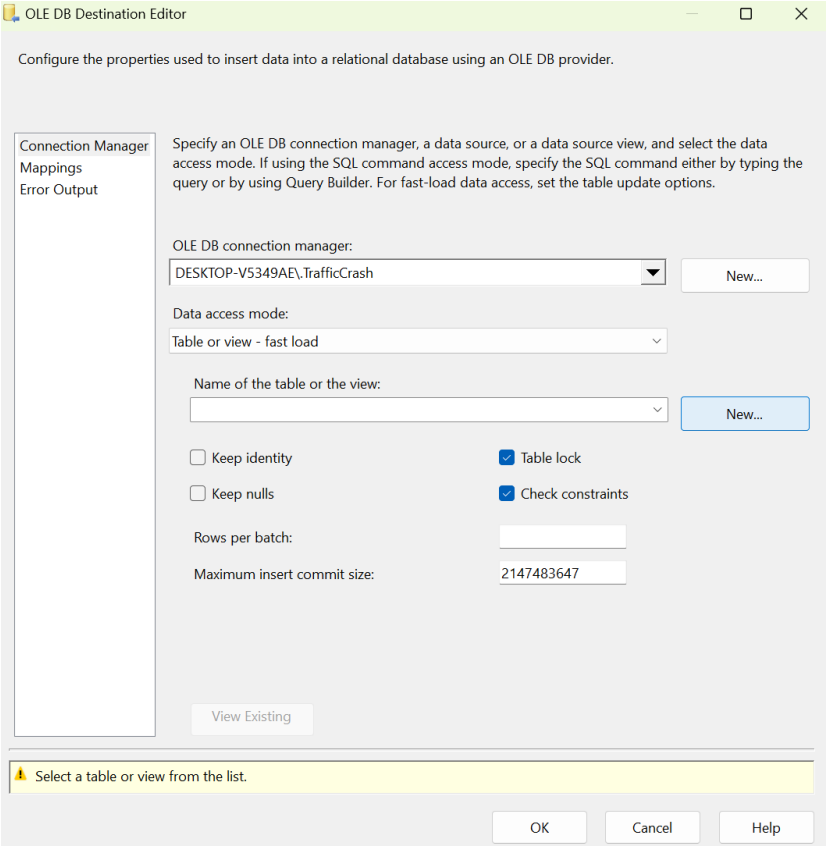
- Chọn các thuộc tính để làm dữ liệu cho "Sort Dim_ Damages".
- Chọn vào "*Remove rows with duplicate sort values*" để xóa các dòng dữ liệu trùng nhau. Sau đó chọn "OK" để hoàn tất.



Bước 3: Tiến hành đổ dữ liệu bảng Dim_Damages vào TrafficCrash. Đầu tiên, tạo một OLE DB Destination, đặt tên là "Dim_Damages".

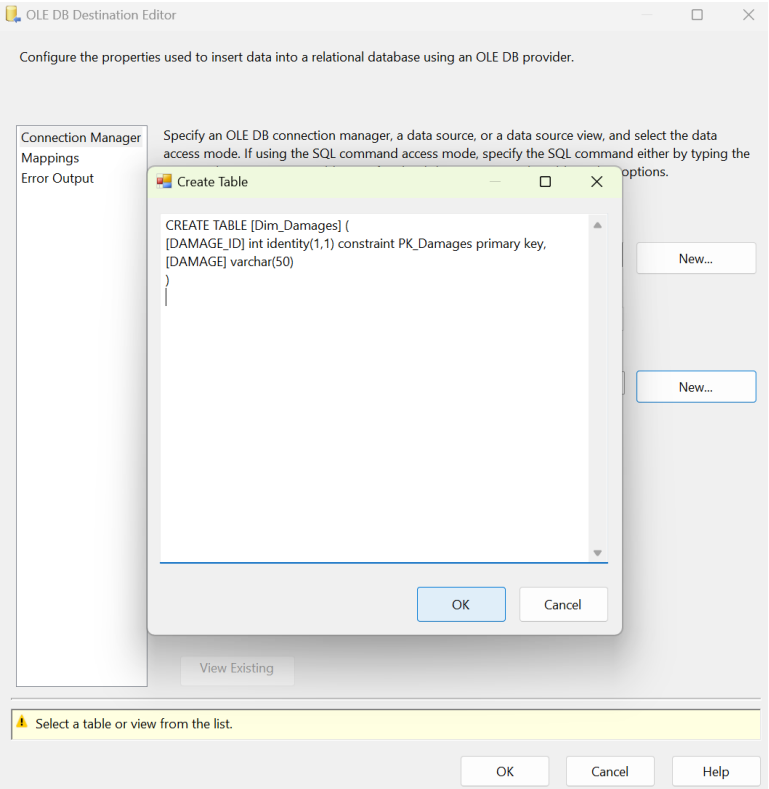


Bước 4: Nhấn chuột phải vào "Dim_Damages", chọn "Edit". Chọn "New" để tạo bảng Dim_Damages mới trong database.

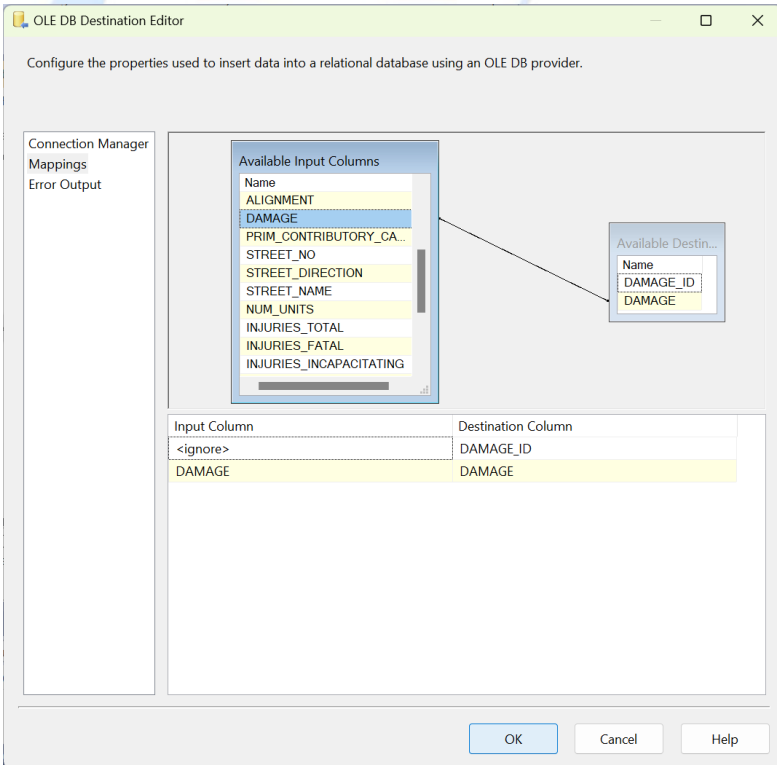


- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Dim_Damages] (
[DAMAGE_ID] int identity(1,1) constraint PK_Damages primary key,
[DAMAGE] varchar(50)
)
```

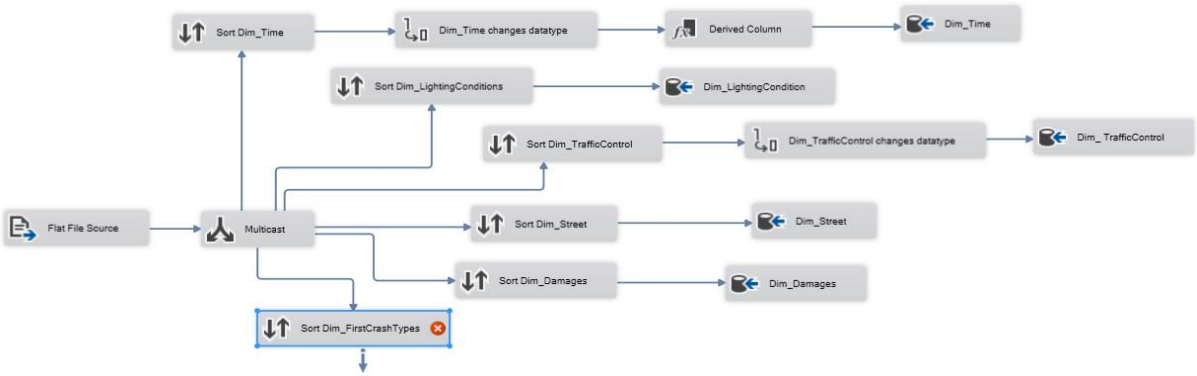


- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



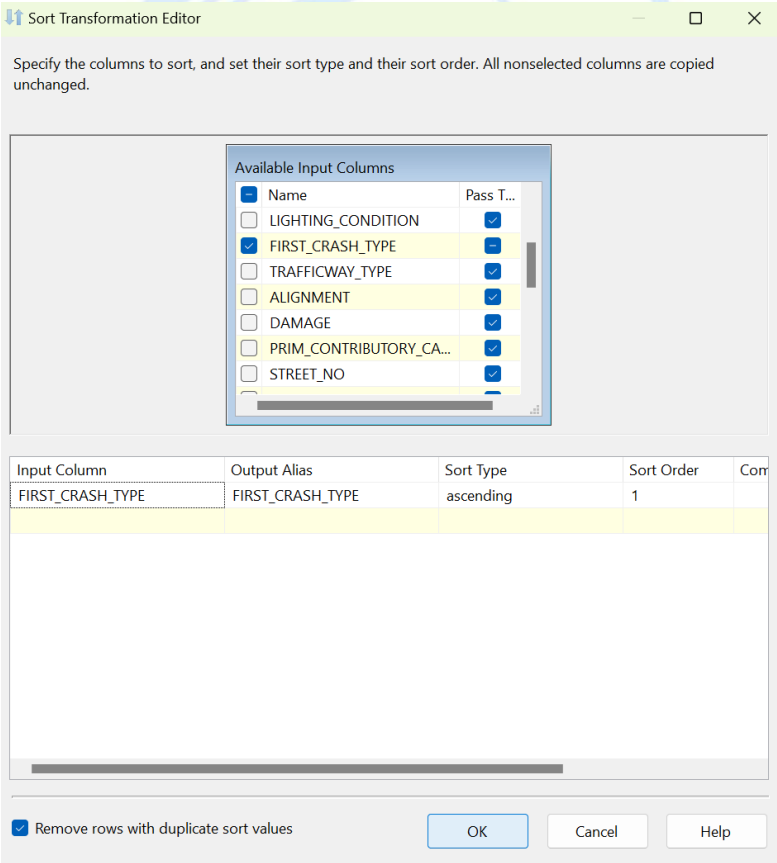
2.3.6 Tạo bảng Dim_ FirstCrashTypes

Bước 1: Tạo một "Sort", đặt tên là "Sort Dim_FirstCrashTypes".

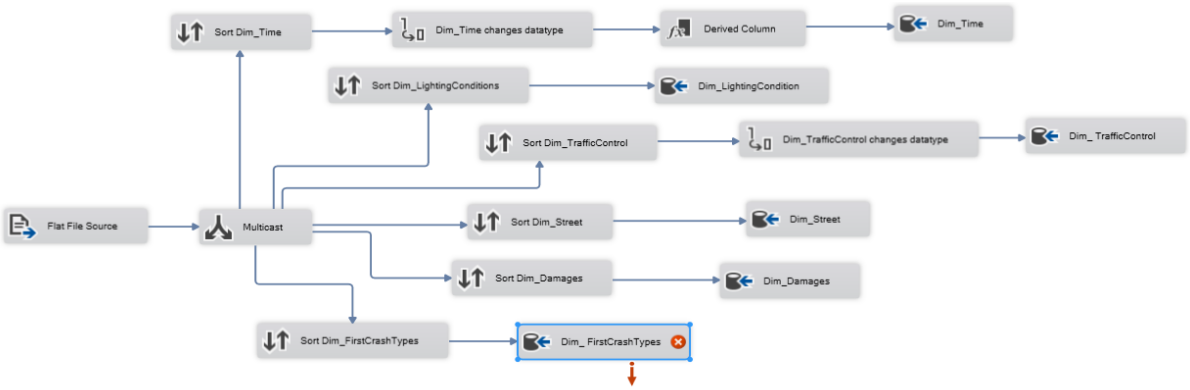


Bước 2: Nhấn chuột phải vào "Sort Dim_FirstCrashTypes", chọn "Edit".

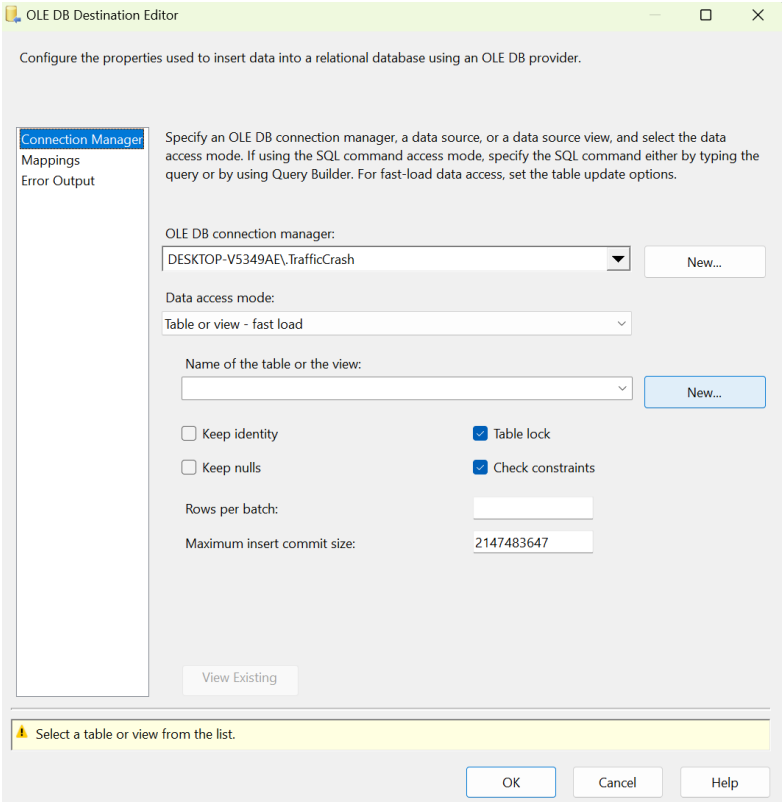
- Chọn các thuộc tính để làm dữ liệu cho "Sort Dim_FirstCrashTypes".
- Chọn vào "*Remove rows with duplicate sort values*" để xóa các dòng dữ liệu trùng nhau. Sau đó chọn "OK" để hoàn tất.



Bước 3: Tiến hành đổ dữ liệu bảng Dim_FirstCrashTypes vào TrafficCrash. Đầu tiên, tạo một OLE DB Destination, đặt tên là "Dim_FirstCrashTypes".



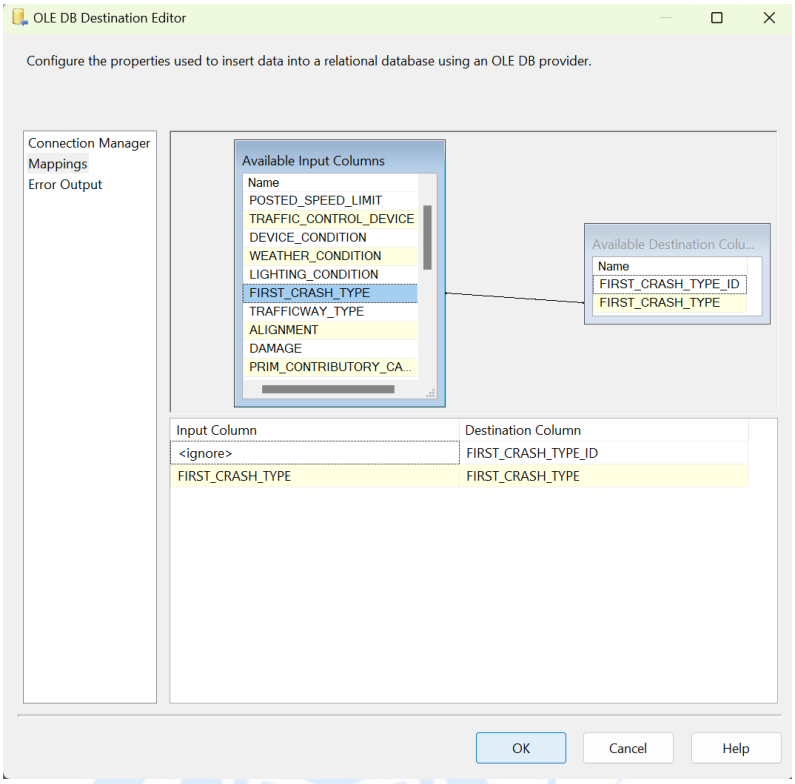
Bước 4: Nhấn chuột phải vào "Dim_FirstCrashTypes", chọn "Edit". Chọn "New" để tạo bảng Dim_ FirstCrashTypes mới trong database.



- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

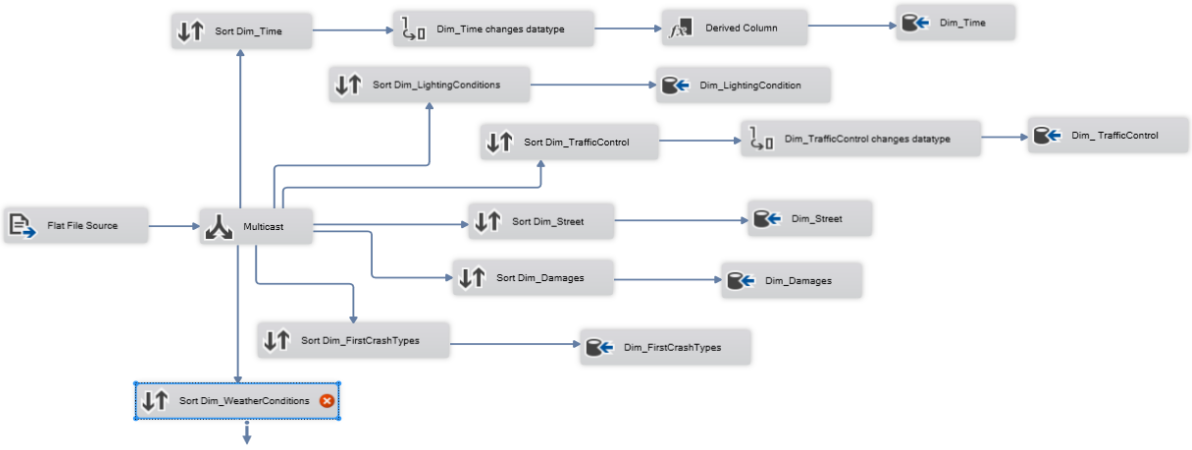
```
CREATE TABLE [Dim_FirstCrashTypes] (
[FIRST_CRASH_TYPE_ID] int identity(1,1) constraint
PK_firstcrashtypes primary key,
[FIRST_CRASH_TYPE] varchar(50)
)
```

- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



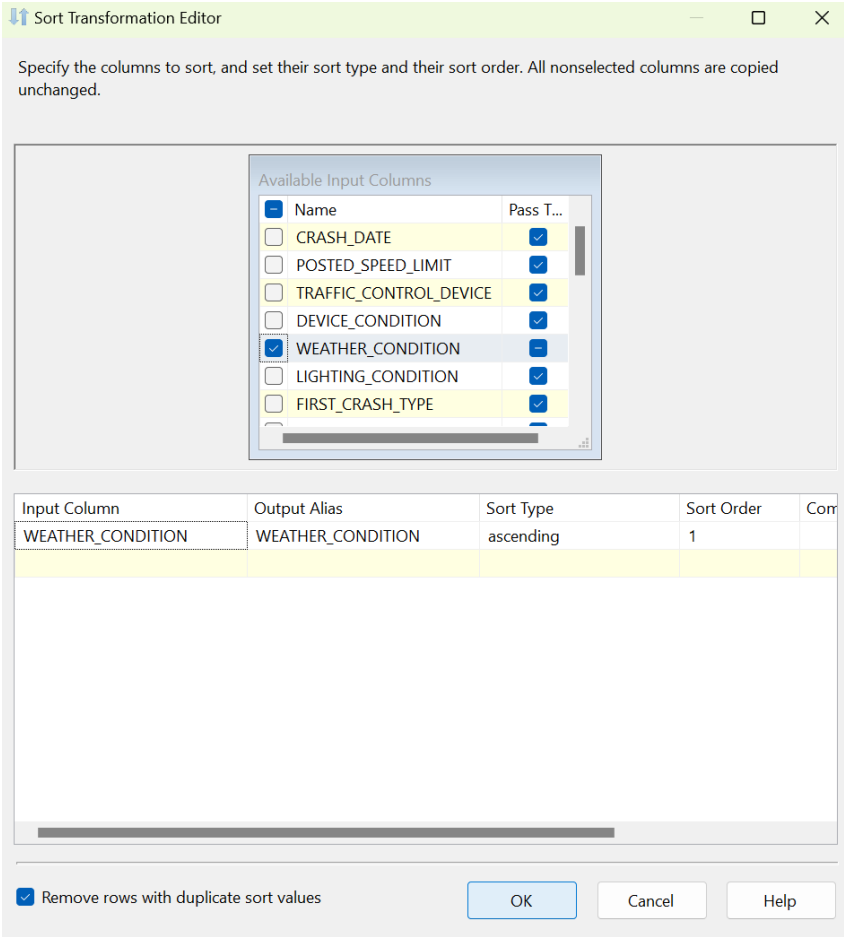
2.3.7 Tạo bảng Dim_WeatherConditions

Bước 1: Tạo một “Sort”, đặt tên là “Sort Dim_WeatherConditions”.

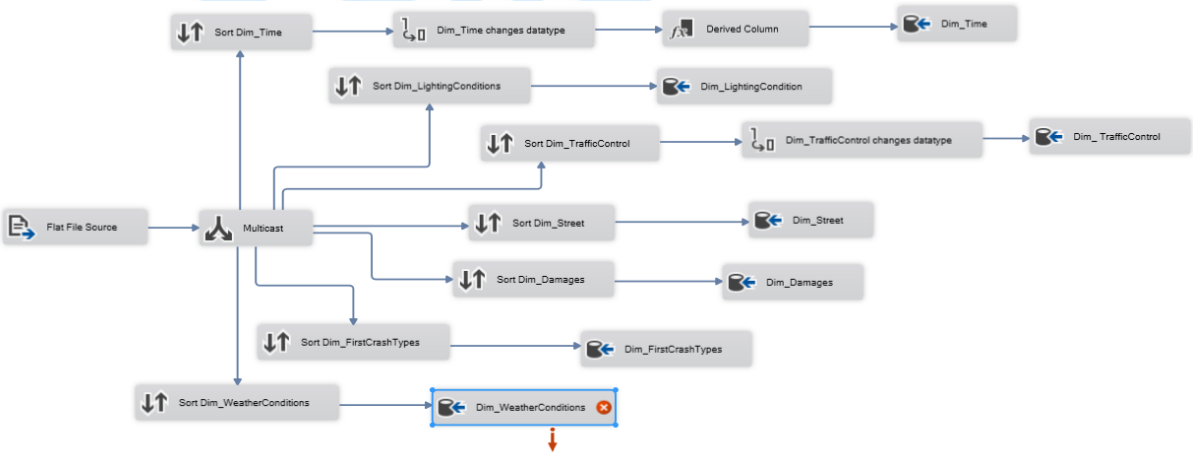


Bước 2: Nhấn chuột phải vào “Sort Dim_WeatherConditions”, chọn “Edit”.

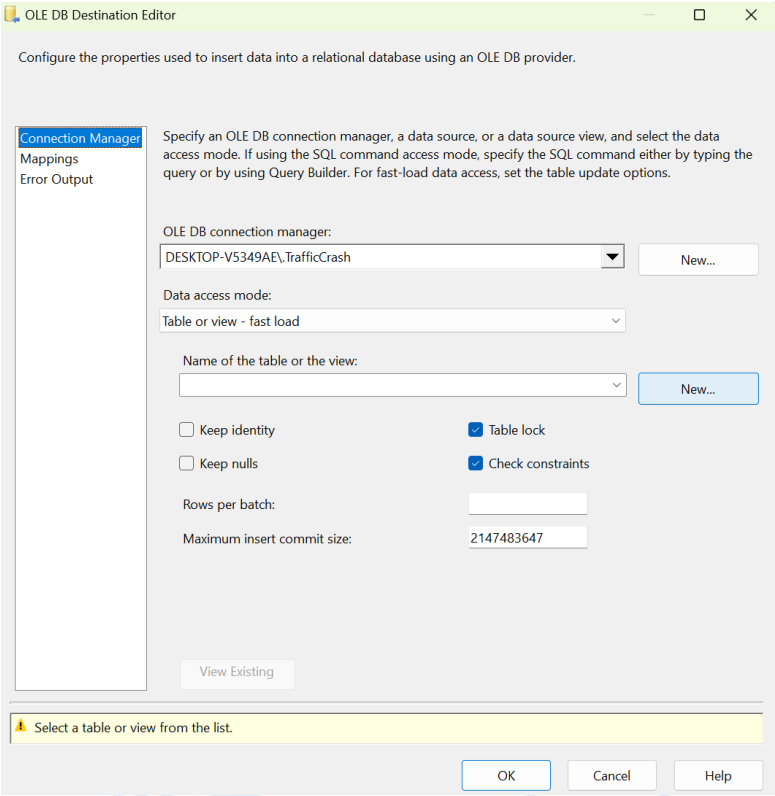
- Chọn các thuộc tính để làm dữ liệu cho “Sort Dim_WeatherConditions”.
- Chọn vào “*Remove rows with duplicate sort values*” để xóa các dòng dữ liệu trùng nhau. Sau đó chọn “OK” để hoàn tất.



Bước 3: Tiến hành đổ dữ liệu bảng Dim_WeatherConditions vào TrafficCrash. Đầu tiên, tạo một OLE DB Destination, đặt tên là "Dim_WeatherConditions".

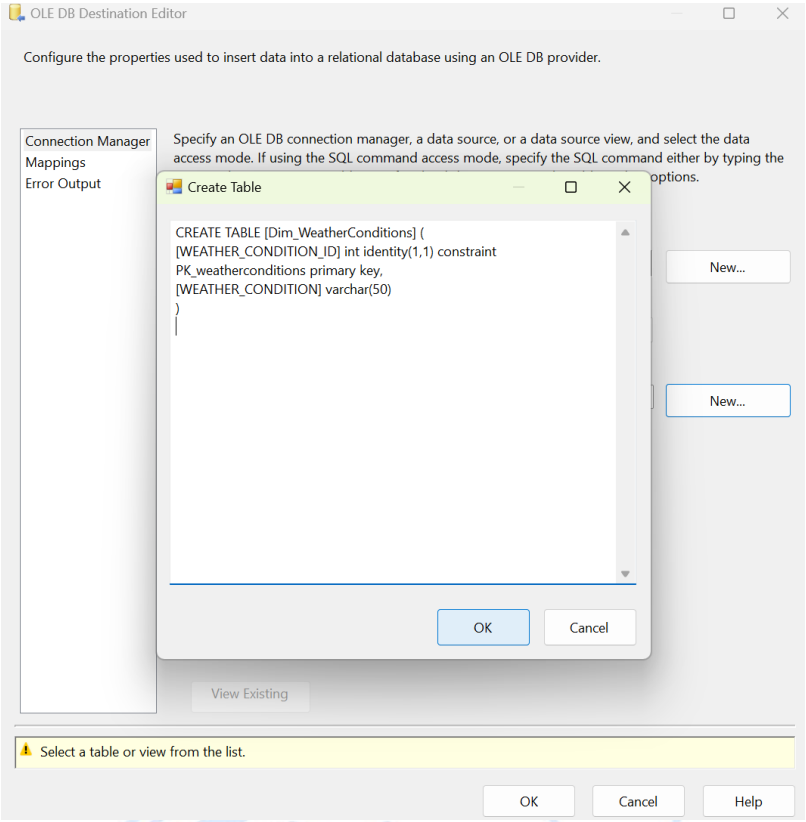


Bước 4: Nhấn chuột phải vào "Dim_WeatherConditions", chọn "Edit". Chọn "New" để tạo bảng Dim_WeatherConditions mới trong database.

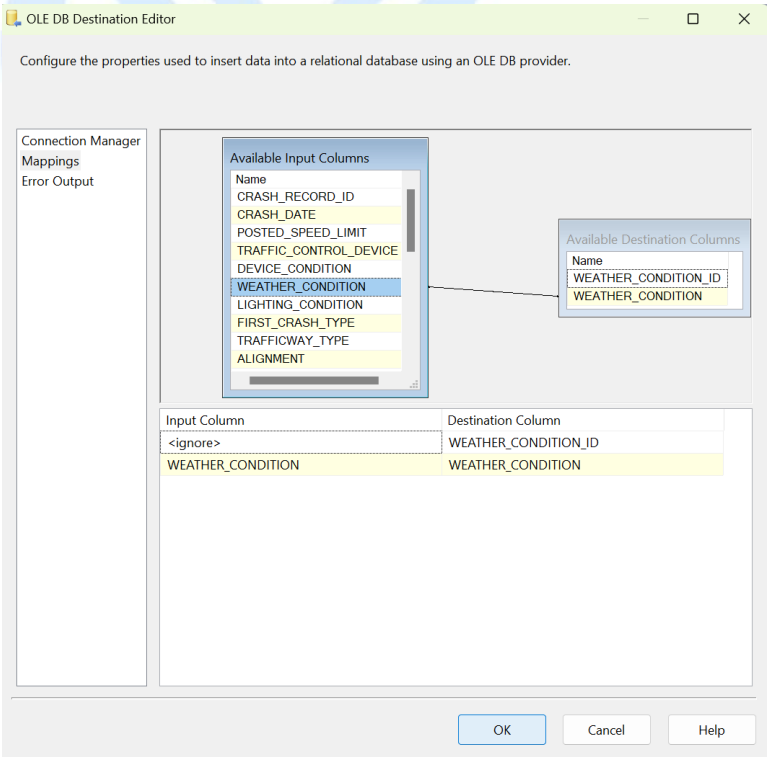


- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Dim_WeatherConditions] (
[WEATHER_CONDITION_ID] int identity(1,1) constraint
PK_weatherconditions primary key,
[WEATHER_CONDITION] varchar(50)
)
```

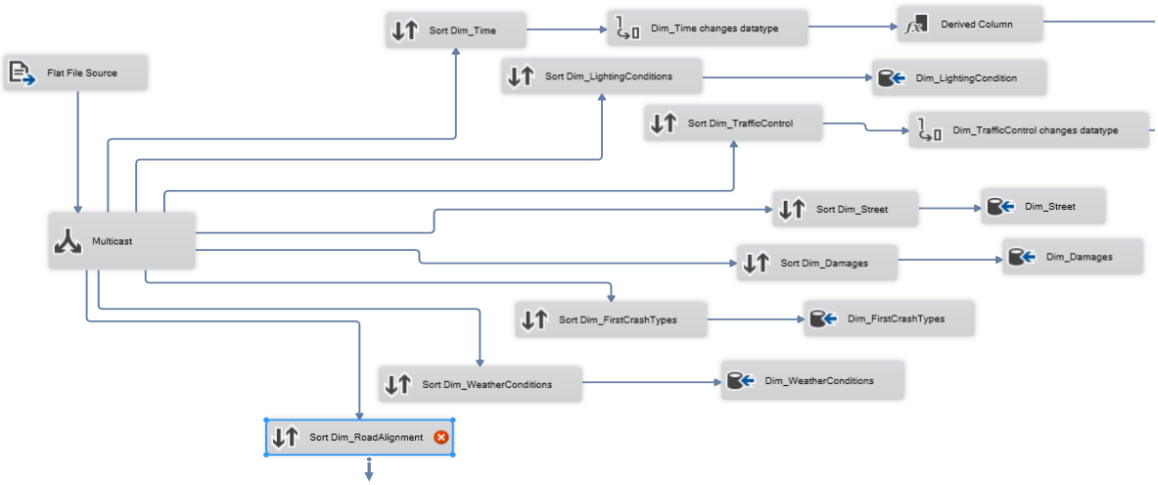


- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



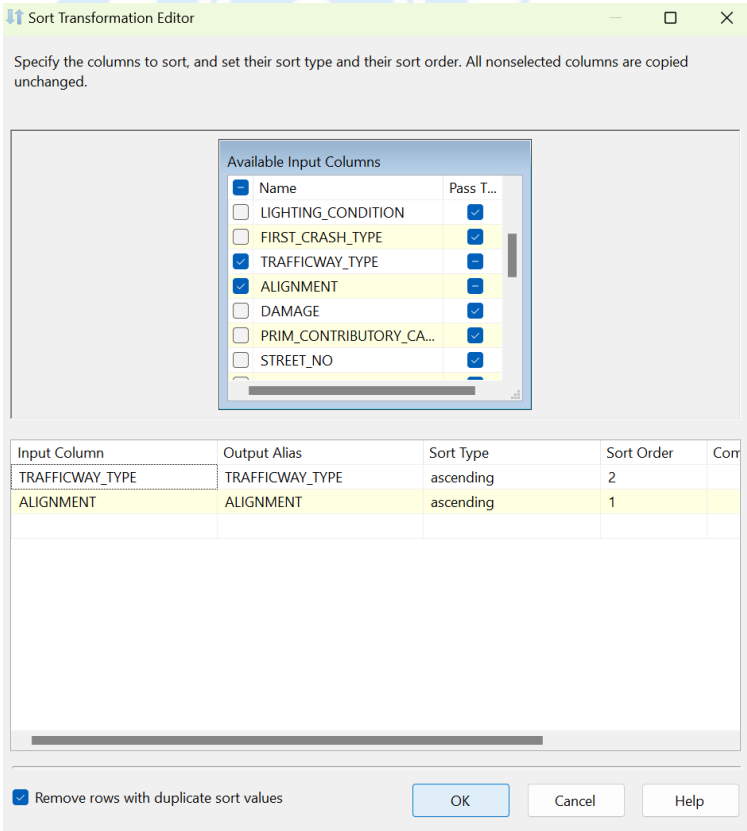
2.3.8 Tạo bảng Dim_RoadAlignment

Bước 1: Tạo một "Sort", đặt tên là "Sort Dim_RoadAlignment".

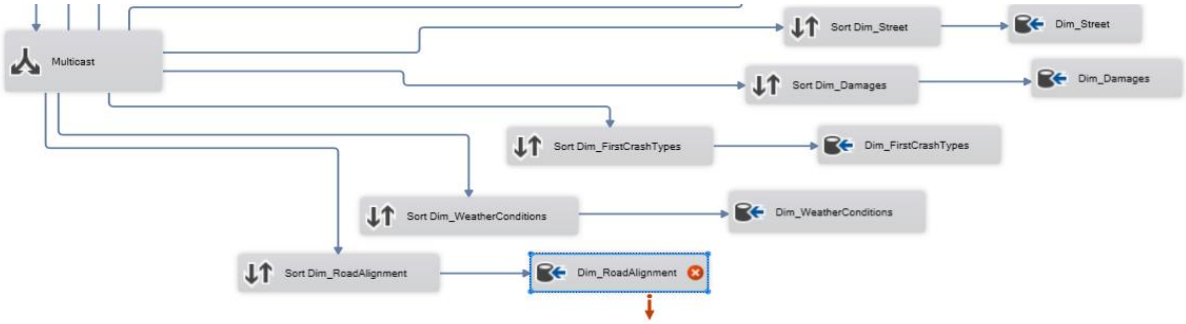


Bước 2: Nhấn chuột phải vào "Sort Dim_RoadAlignment", chọn "Edit".

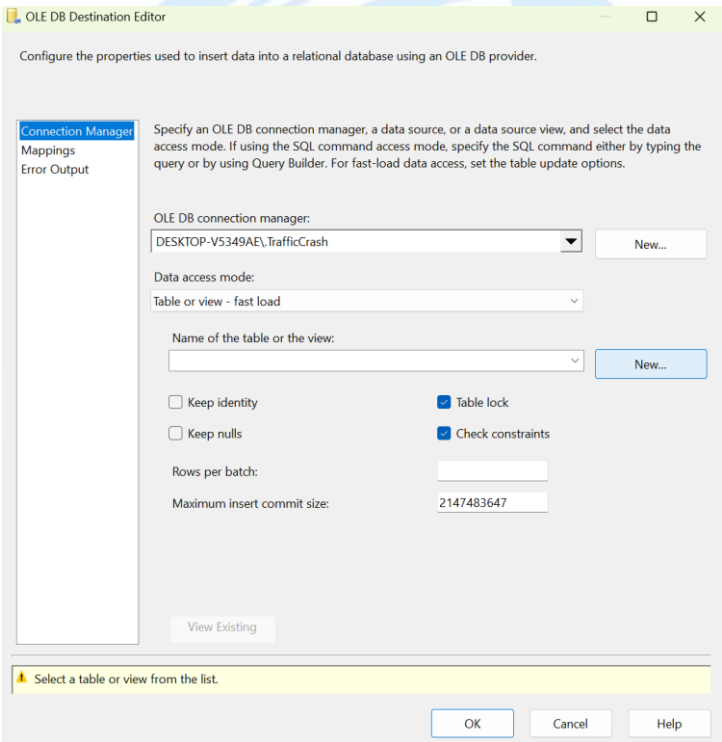
- Chọn các thuộc tính để làm dữ liệu cho "Sort Dim_RoadAlignment".
- Chọn vào "*Remove rows with duplicate sort values*" để xóa các dòng dữ liệu trùng nhau. Sau đó chọn "OK" để hoàn tất.



Bước 3: Tiến hành đổ dữ liệu bảng Dim_RoadAlignment vào TrafficCrash. Đầu tiên, tạo một OLE DB Destination, đặt tên là "Dim_RoadAlignment".

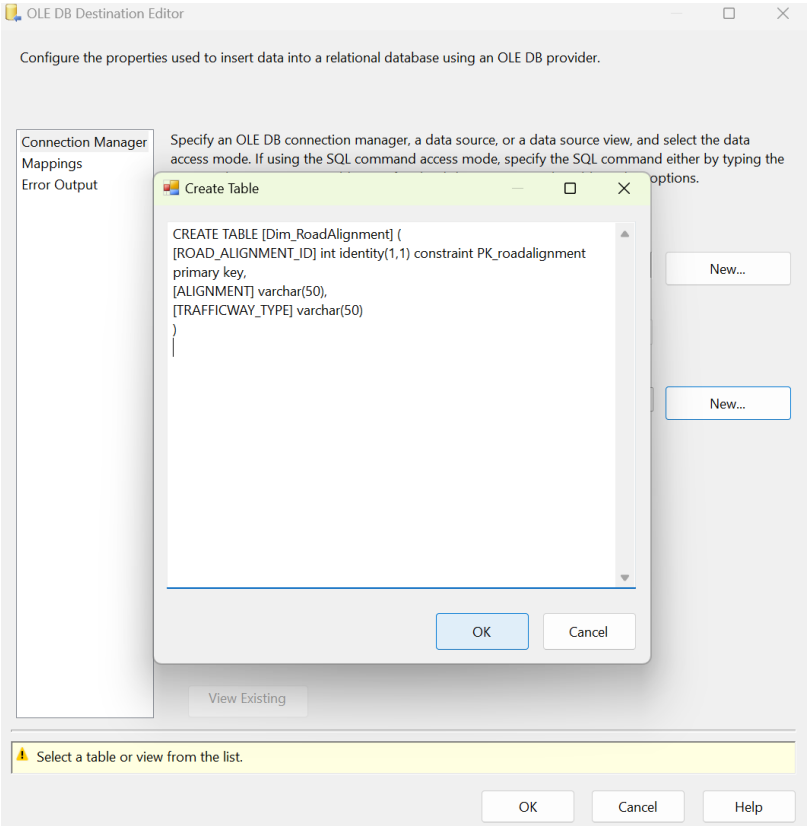


Bước 4: Nhấn chuột phải vào "Dim_RoadAlignment", chọn "Edit". Chọn "New" để tạo bảng Dim_RoadAlignment mới trong database.

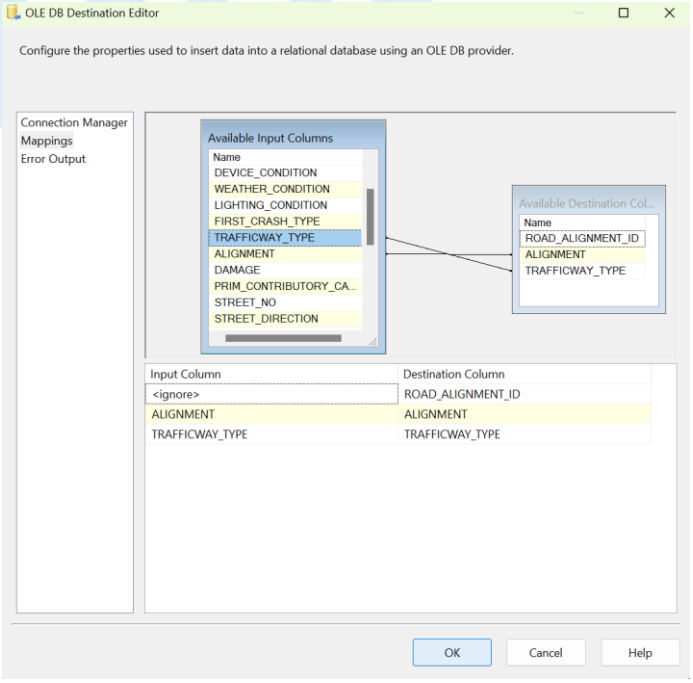


- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Dim_RoadAlignment] (
[ROAD_ALIGNMENT_ID] int identity(1,1) constraint
PK_roadalignment primary key,
[ALIGNMENT] varchar(50),
[TRAFFICWAY_TYPE] varchar(50)
)
```

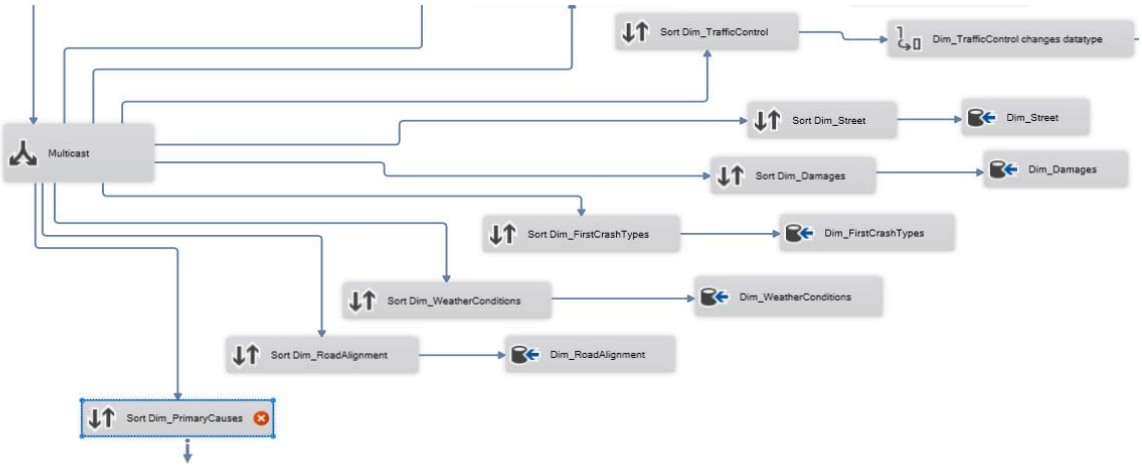


- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



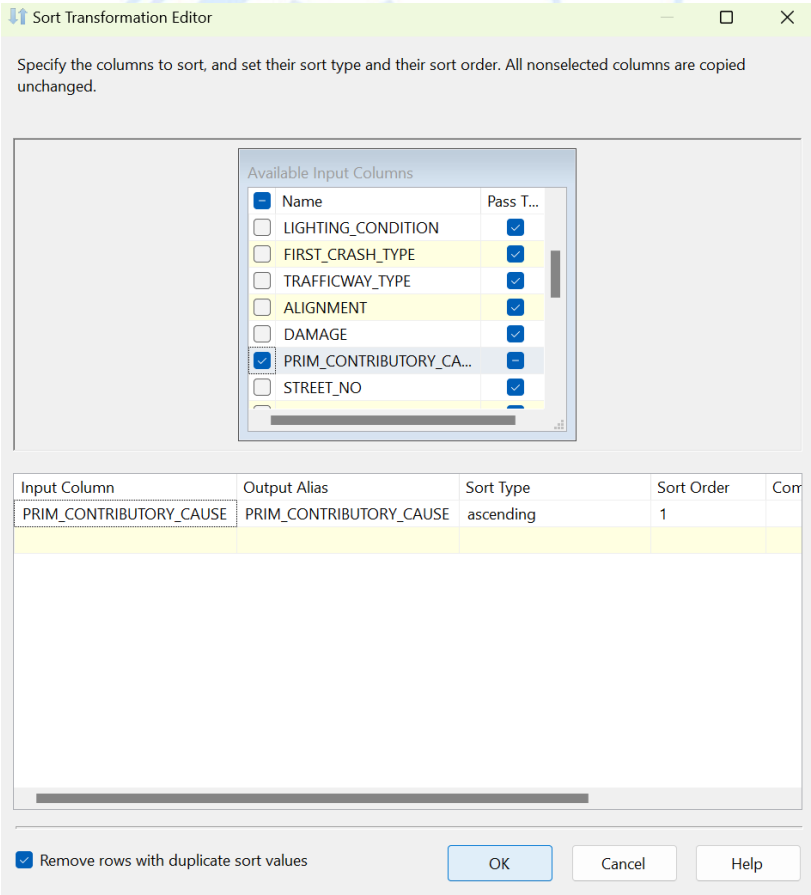
2.3.9 Tạo bảng Dim_PrimaryCauses

Bước 1: Tạo một "Sort", đặt tên là "Sort Dim_PrimaryCauses".

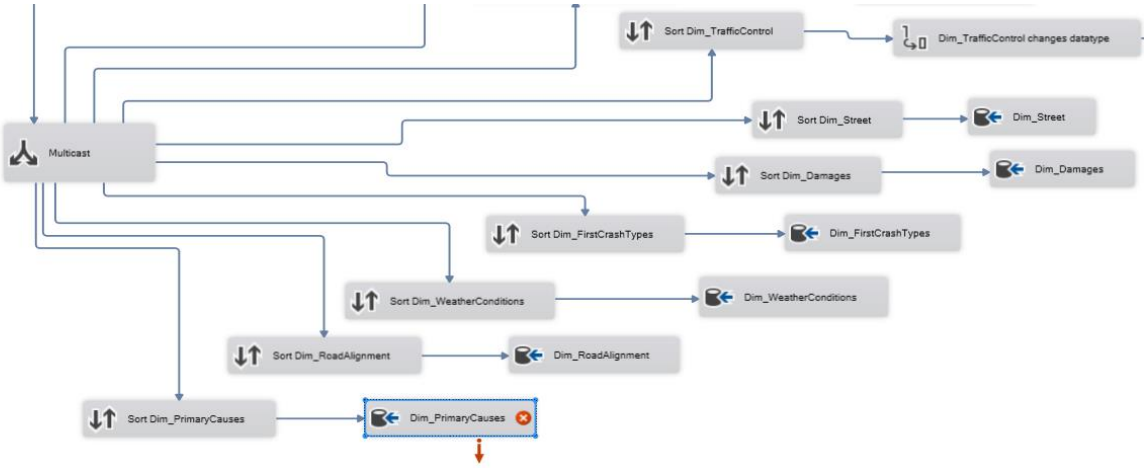


Bước 2: Nhấn chuột phải vào "Sort Dim_PrimaryCauses", chọn "Edit".

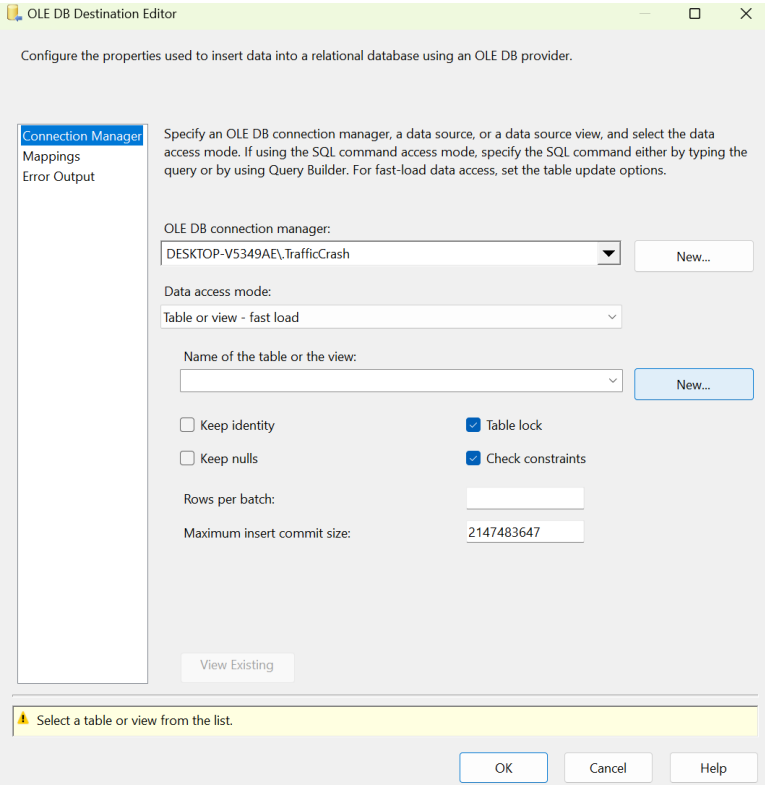
- Chọn các thuộc tính để làm dữ liệu cho "Sort Dim_PrimaryCauses".
- Chọn vào "*Remove rows with duplicate sort values*" để xóa các dòng dữ liệu trùng nhau. Sau đó chọn "OK" để hoàn tất.



Bước 3: Tiến hành đổ dữ liệu bảng Dim_PrimaryCauses vào TrafficCrash. Đầu tiên, tạo một OLE DB Destination, đặt tên là "Dim_PrimaryCauses".

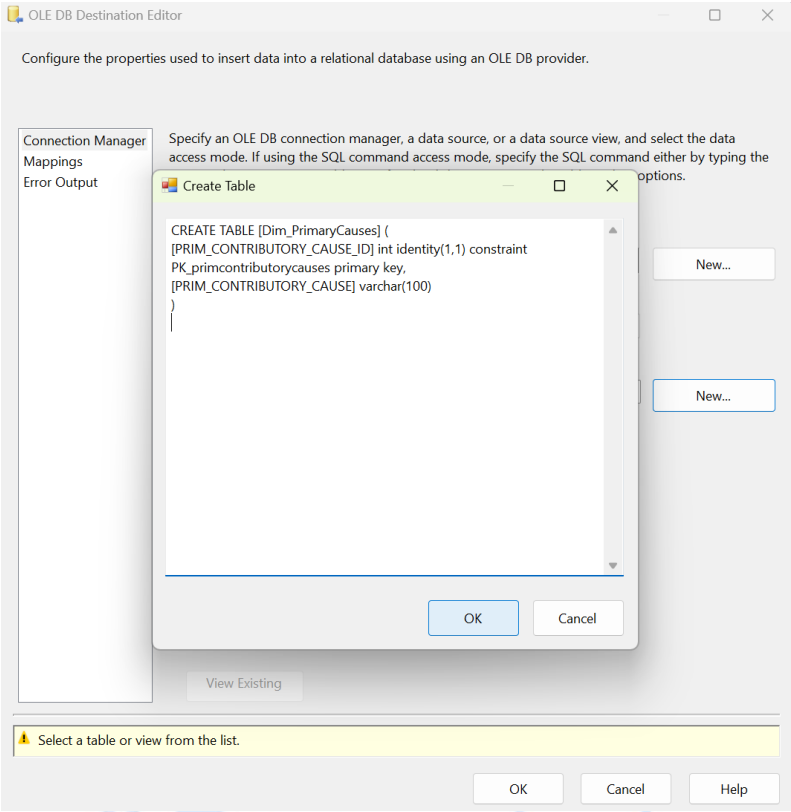


Bước 4: Nhấn chuột phải vào "Dim_PrimaryCauses", chọn "Edit". Chọn "New" để tạo bảng Dim_PrimaryCauses mới trong database.

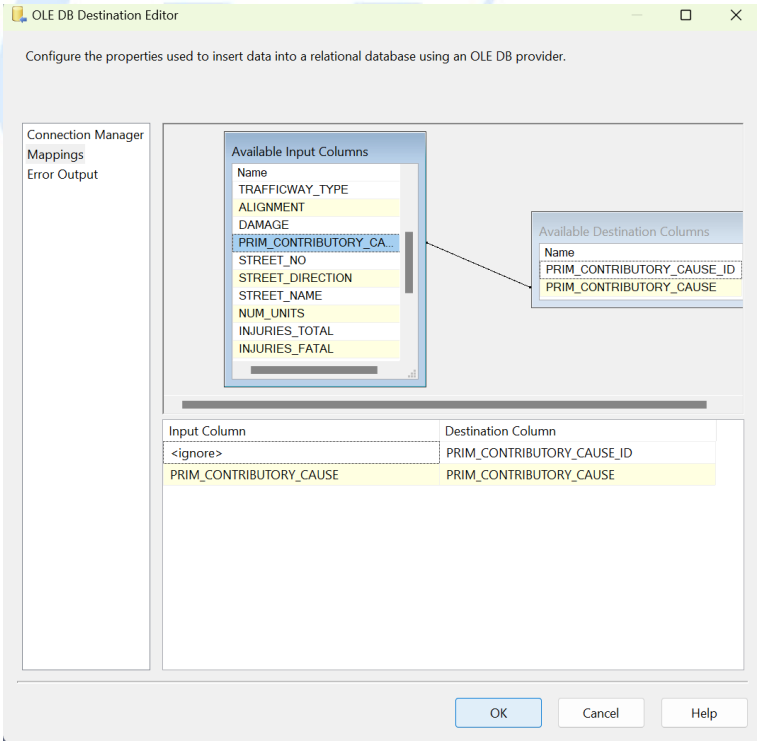


- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Dim_PrimaryCauses] (
[PRIM_CONTRIBUTORY_CAUSE_ID] int identity(1,1) constraint
PK_primcontributorycauses primary key,
[PRIM_CONTRIBUTORY_CAUSE] varchar(100)
)
```



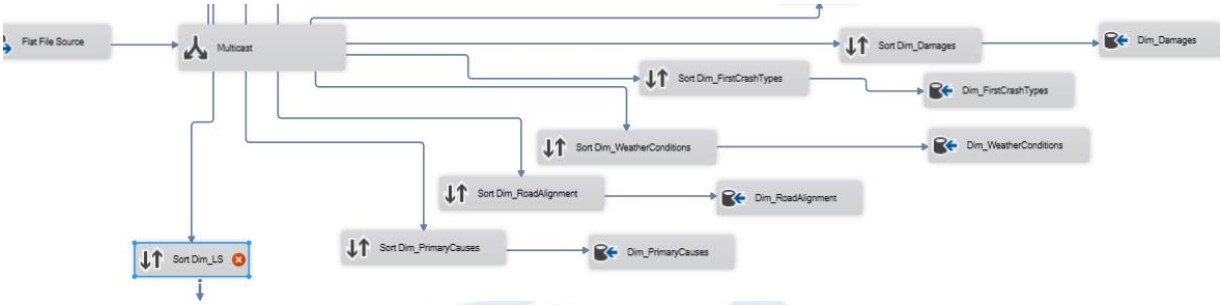
- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



2.3.10 Tạo bảng Dim_LS

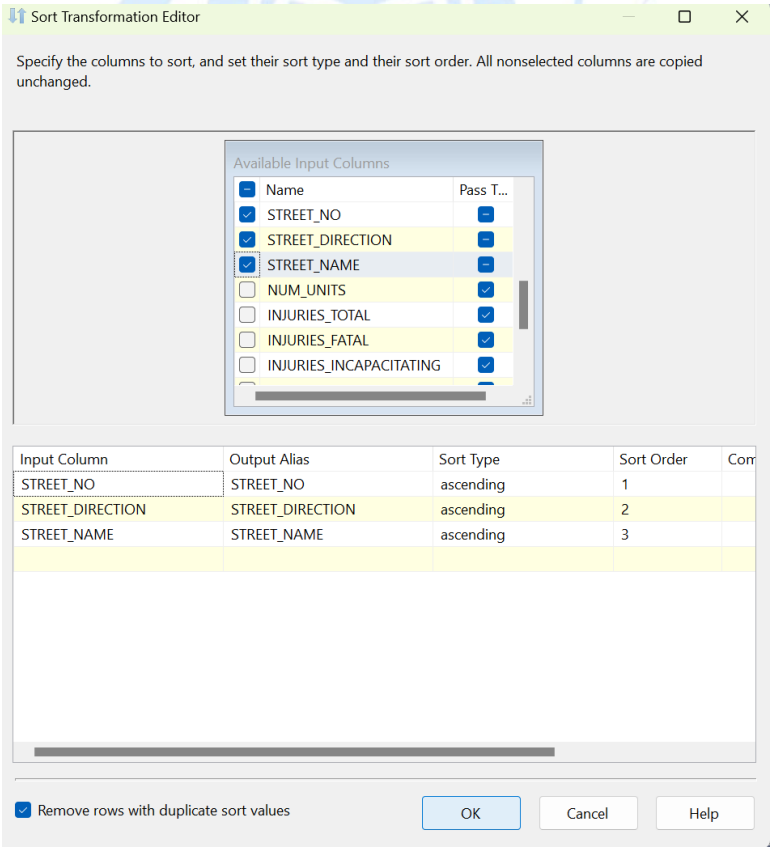
Bảng Dim_LS dùng để Merge với Dim_Street tạo thành Dim_LocationStreets.

Bước 1: Tạo một "Sort", đặt tên là "Sort Dim_LS".

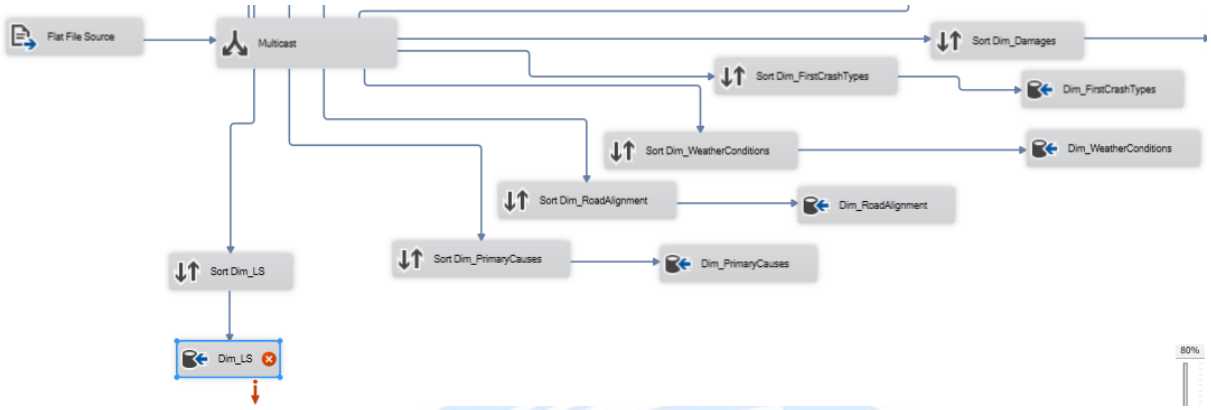


Bước 2: Nhấn chuột phải vào "Sort Dim_LS", chọn "Edit".

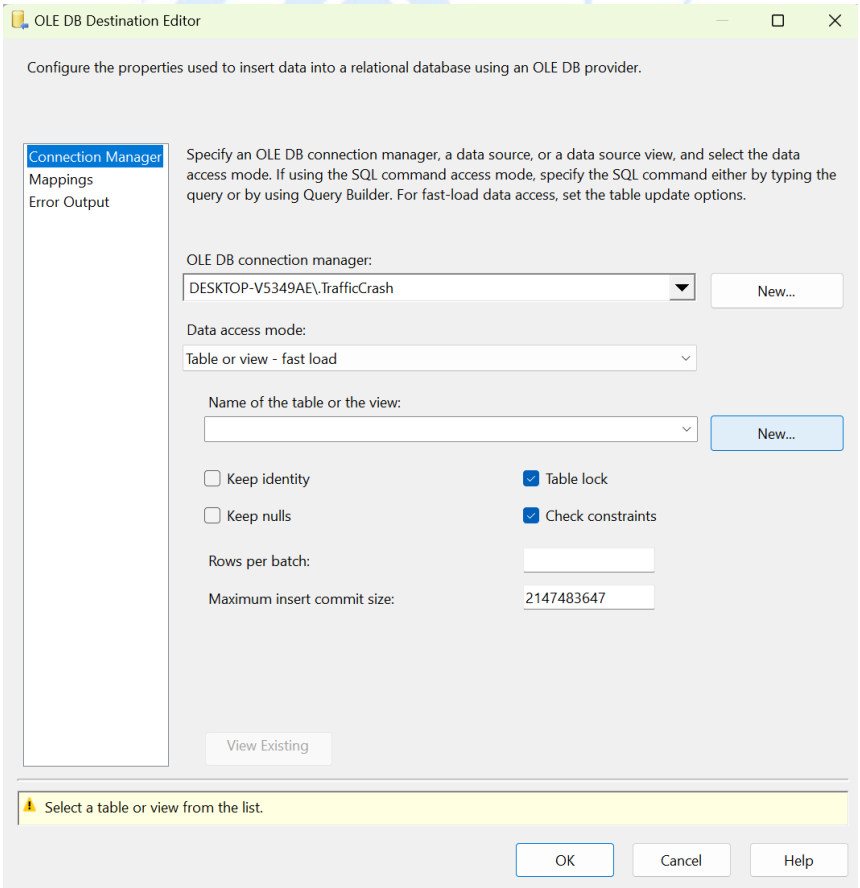
- Chọn các thuộc tính để làm dữ liệu cho "Sort Dim_LS".
- Chọn vào "*Remove rows with duplicate sort values*" để xóa các dòng dữ liệu trùng nhau. Sau đó chọn "OK" để hoàn tất.



Bước 3: Thực hiện đổ dữ liệu từ bảng Dim_LS vào TrafficCrash. Đầu tiên, tạo một OLE DB Destination, đặt tên là “Dim_LS”.

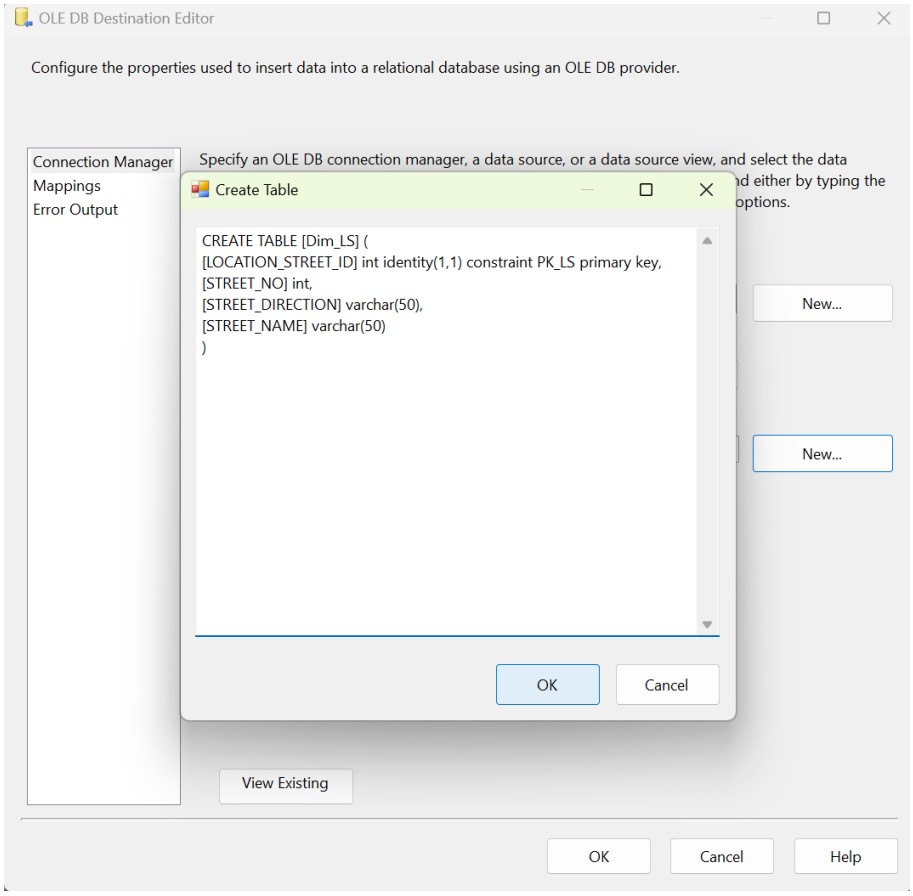


Bước 4: Nhấn chuột phải vào “Dim_LS”, chọn “Edit”. Chọn “New” để tạo bảng Dim_LS mới trong database.

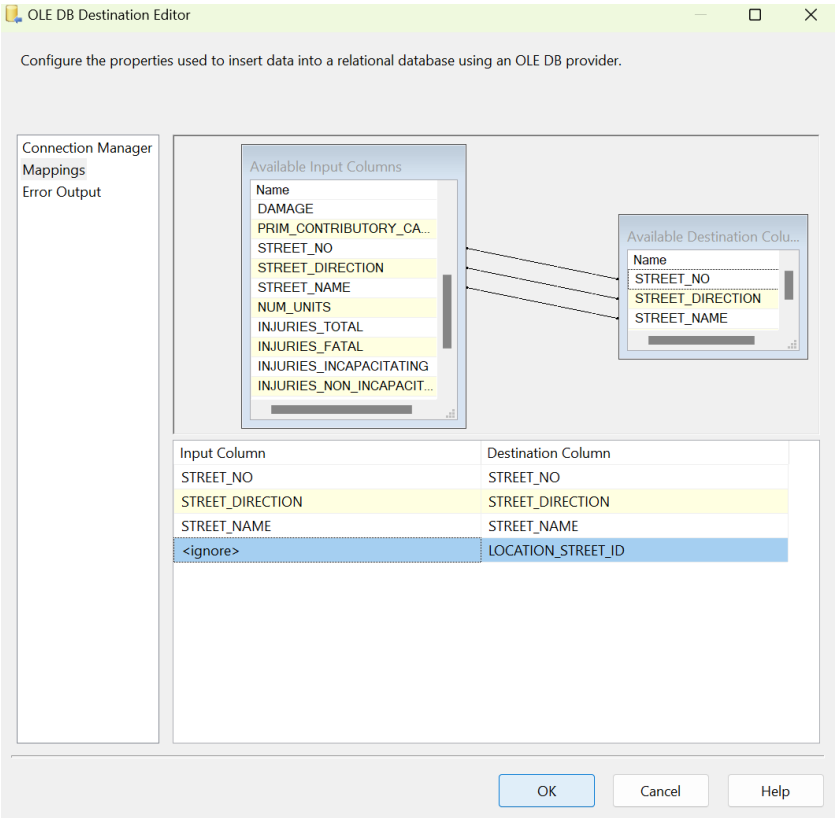


- Viết câu lệnh sql để tạo bảng. Sau đó nhấn “OK”.

```
CREATE TABLE [Dim_LS] (
[LOCATION_STREET_ID] int identity(1,1) constraint PK_LS primary key,
[STREET_NO] int,
[STREET_DIRECTION] varchar(50),
[STREET_NAME] varchar(50)
)
```

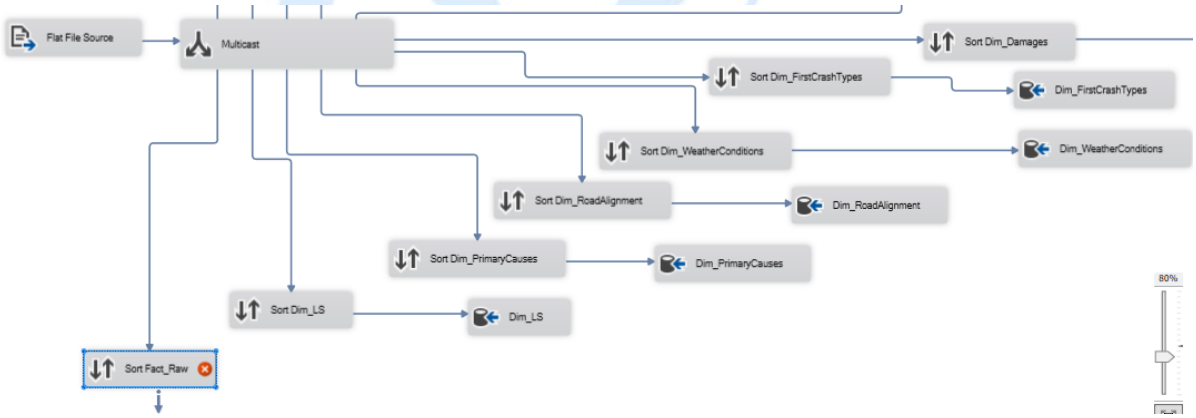


- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



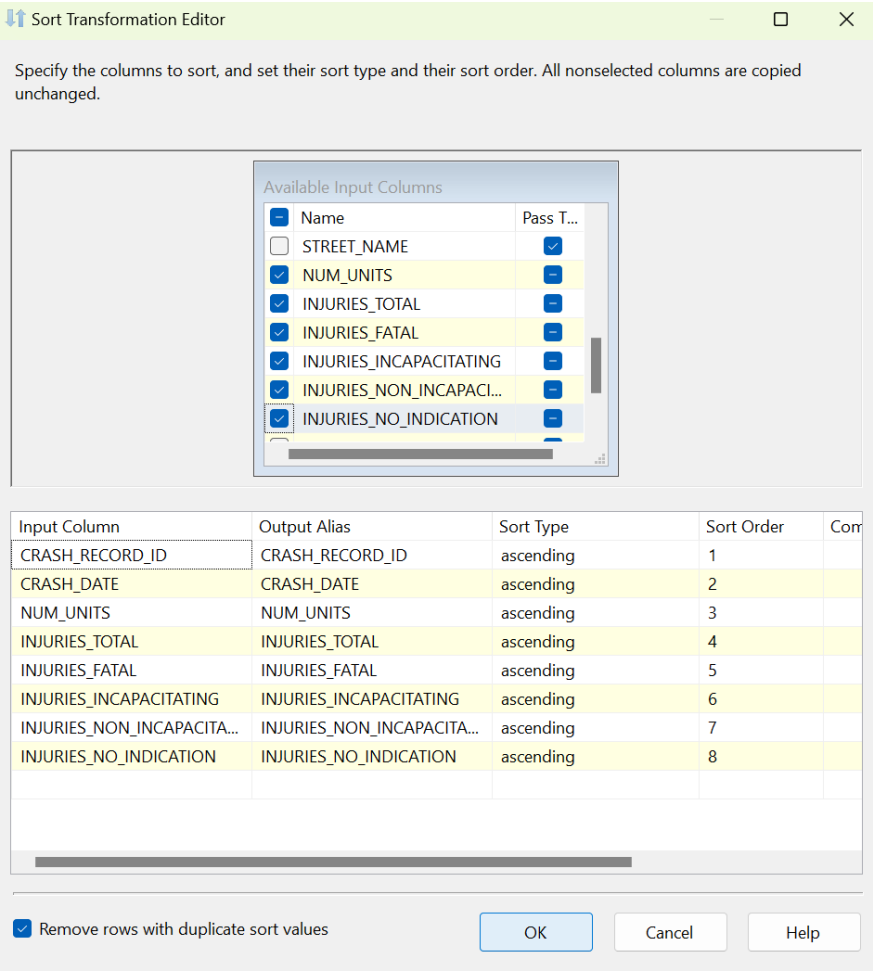
2.3.11 Tạo bảng Fact_Raw

Bước 1: Tạo một "Sort", đặt tên là "Sort Fact_Raw".

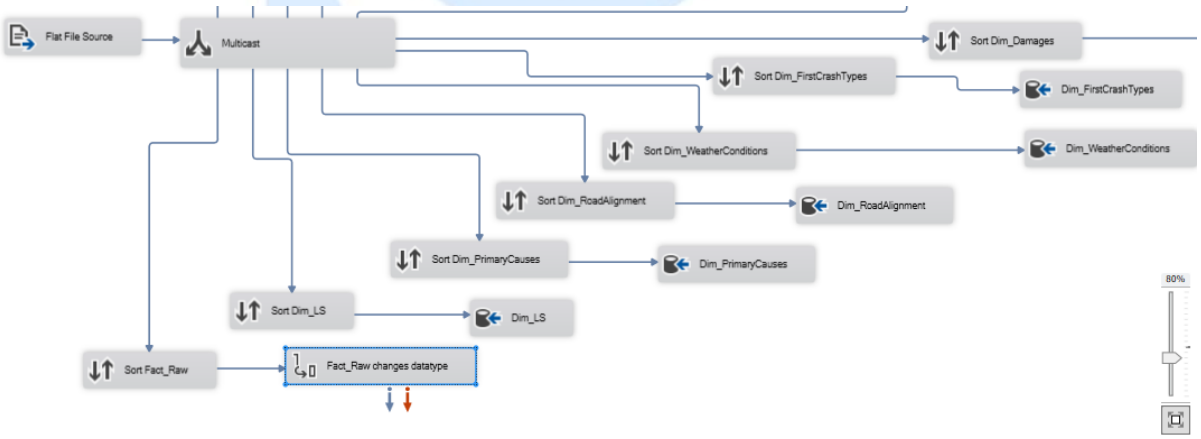


Bước 2: Nhấn chuột phải vào "Sort Fact_Raw", chọn "Edit".

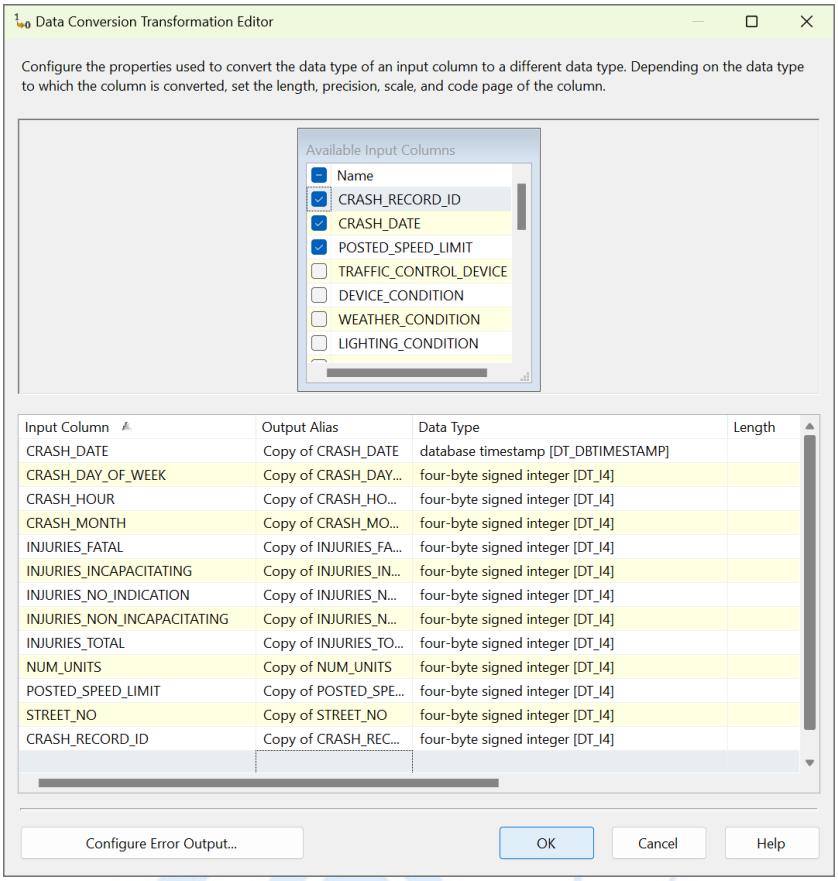
- Chọn các thuộc tính để làm dữ liệu cho "Sort Fact_Raw".
- Chọn vào "*Remove rows with duplicate sort values*" để xóa các dòng dữ liệu trùng nhau. Sau đó chọn "OK" để hoàn tất.



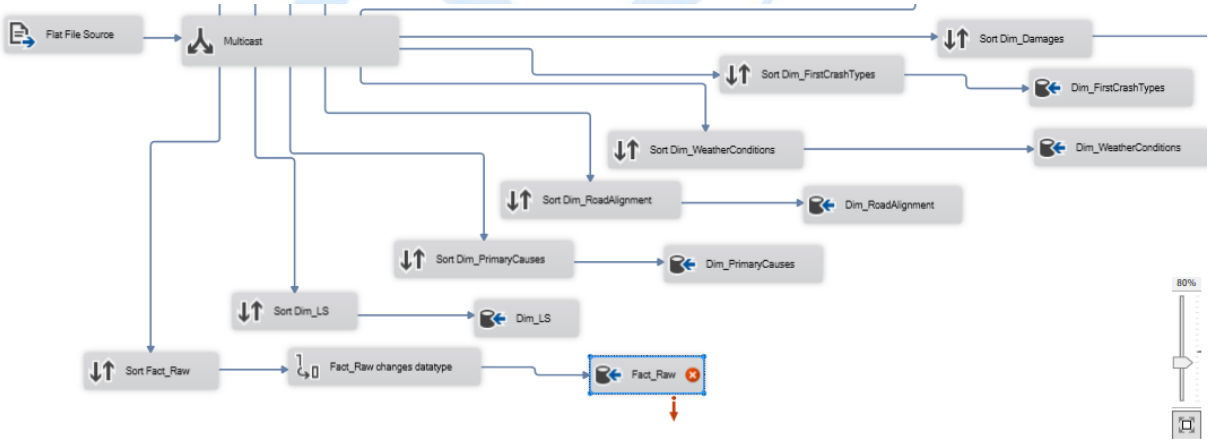
Bước 3: Tiến hành tạo một "Data Conversion" đặt tên là "Fact_Raw changes datatype" để đổi kiểu dữ liệu của các cột sang các kiểu dữ liệu phù hợp.



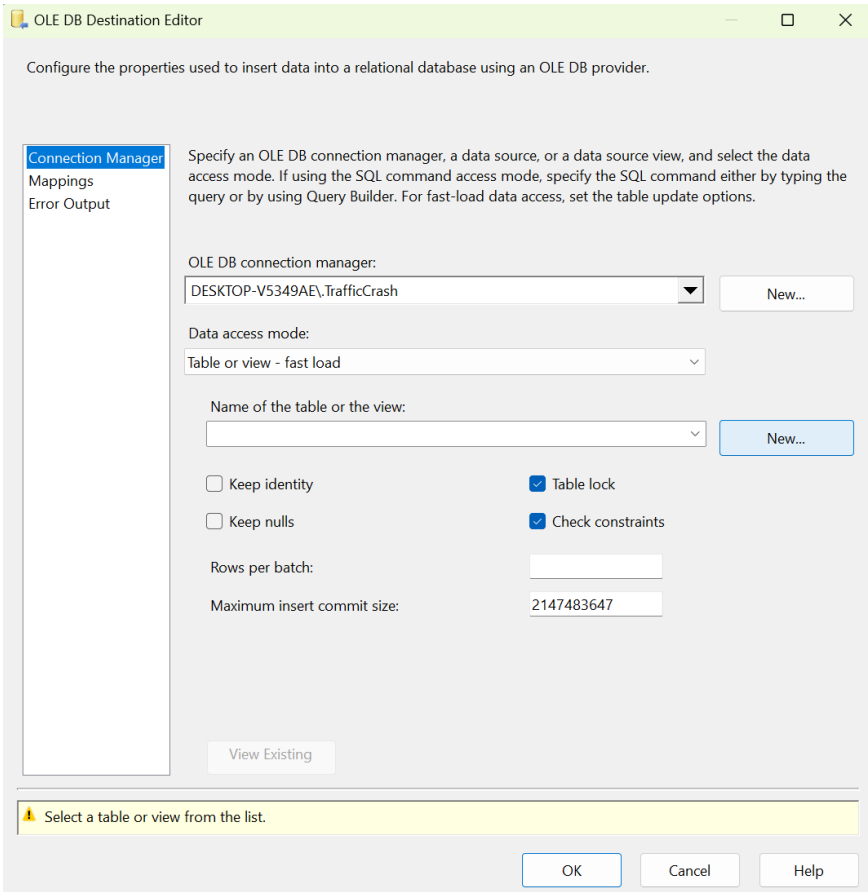
- Sau khi tiến hành đổi kiểu dữ liệu, nhấn "OK" để hoàn tất.



Bước 4: Tiến hành đổ dữ liệu bảng Fact_Raw vào TrafficCrash. Đầu tiên, tạo một OLE DB Destination, đặt tên là "Fact_Raw".



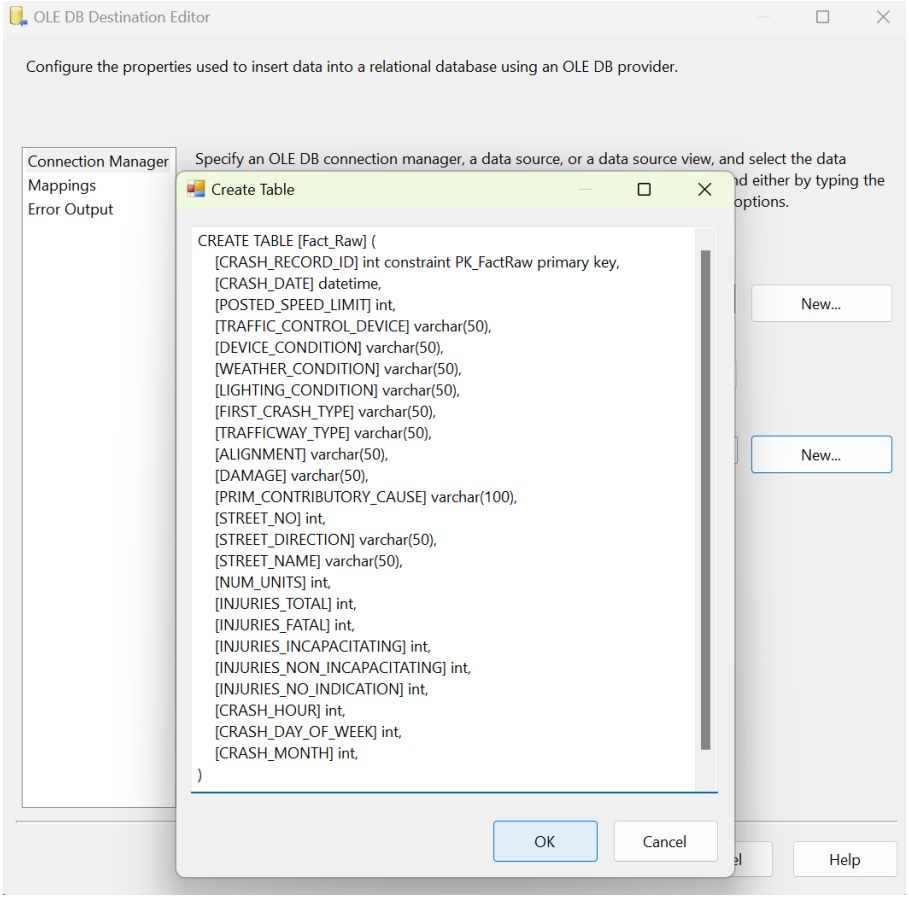
Bước 5: Nhấn chuột phải vào "Fact_Raw", chọn "Edit". Chọn "New" để tạo bảng Fact_Raw mới trong database có các cột là tất cả các cột từ dữ liệu gốc và chứa tất cả các dòng dữ liệu.



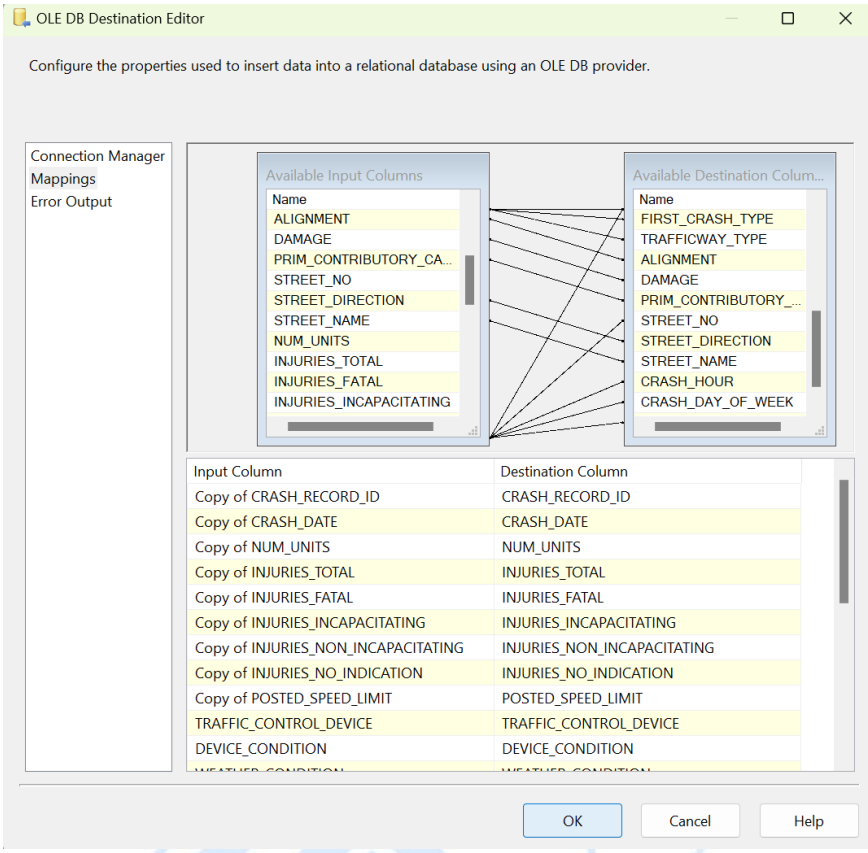
- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Fact_Raw] (
    [CRASH_RECORD_ID] int constraint PK_FactRaw primary key,
    [CRASH_DATE] datetime,
    [POSTED_SPEED_LIMIT] int,
    [TRAFFIC_CONTROL_DEVICE] varchar(50),
    [DEVICE_CONDITION] varchar(50),
    [WEATHER_CONDITION] varchar(50),
    [LIGHTING_CONDITION] varchar(50),
    [FIRST_CRASH_TYPE] varchar(50),
    [TRAFFICWAY_TYPE] varchar(50),
    [ALIGNMENT] varchar(50),
    [DAMAGE] varchar(50),
    [PRIM_CONTRIBUTORY_CAUSE] varchar(100),
    [STREET_NO] int,
    [STREET_DIRECTION] varchar(50),
    [STREET_NAME] varchar(50),
    [NUM_UNITS] int,
```

```
    [INJURIES_TOTAL] int,
    [INJURIES_FATAL] int,
    [INJURIES_INCAPACITATING] int,
    [INJURIES_NON_INCAPACITATING] int,
    [INJURIES_NO_INDICATION] int,
    [CRASH_HOUR] int,
    [CRASH_DAY_OF_WEEK] int,
    [CRASH_MONTH] int,
)
```



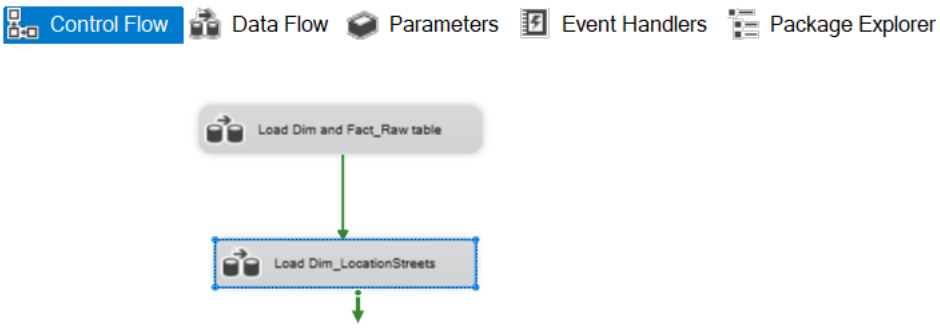
- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



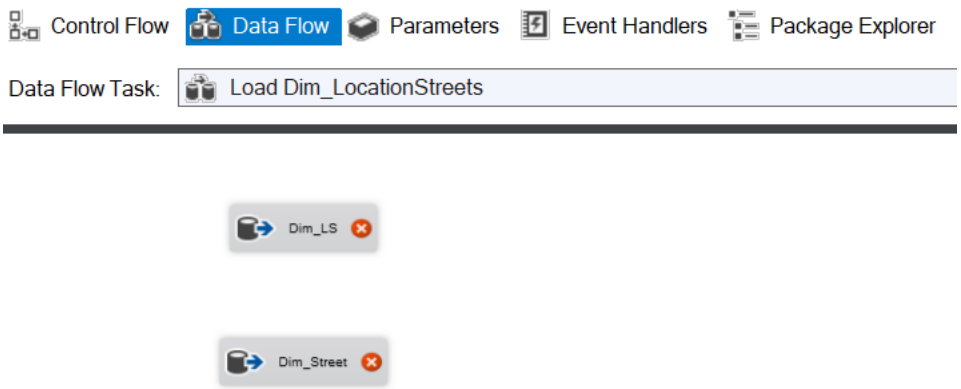
Sau đó, ta sẽ thực hiện quá trình lần lượt loại bỏ các cột dữ liệu trùng của bảng Fact với các Dimension, thực hiện thêm khóa ngoại vào bảng Fact nhằm thu gọn bảng Fact, tối ưu hóa quá trình phân tích dữ liệu.

2.3.12 Tạo bảng Dim_LocationStreets

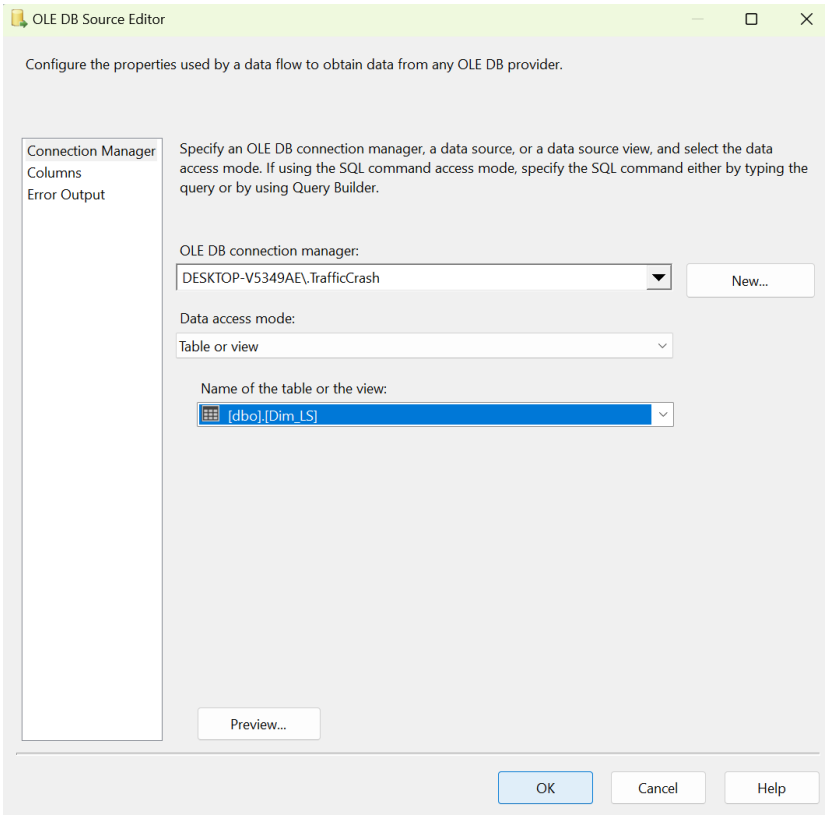
Bước 1: Ở tab Control Flow, tạo một Data Flow Task và đặt tên là "Load Dim_LocationStreets".



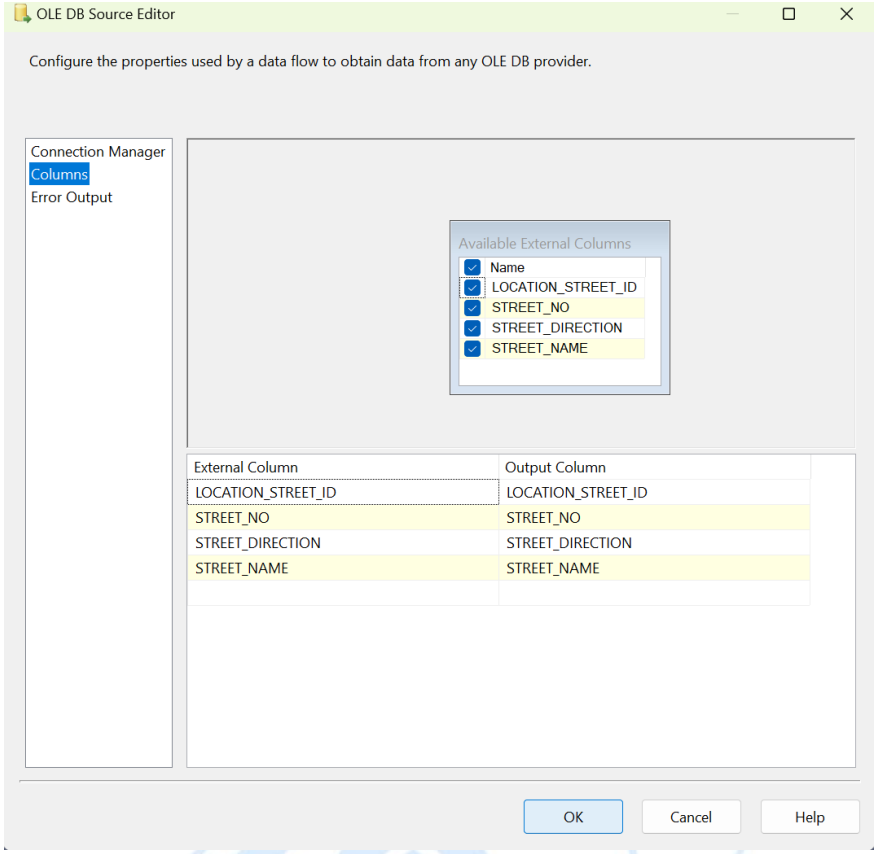
Bước 2: Nhấn chuột phải vào "Load Dim_LocationStreets" và chọn "Edit". Trong tab Data Flow ta tạo hai **OLE DB Source** và đặt tên là "Dim_LS và "Dim_Street".



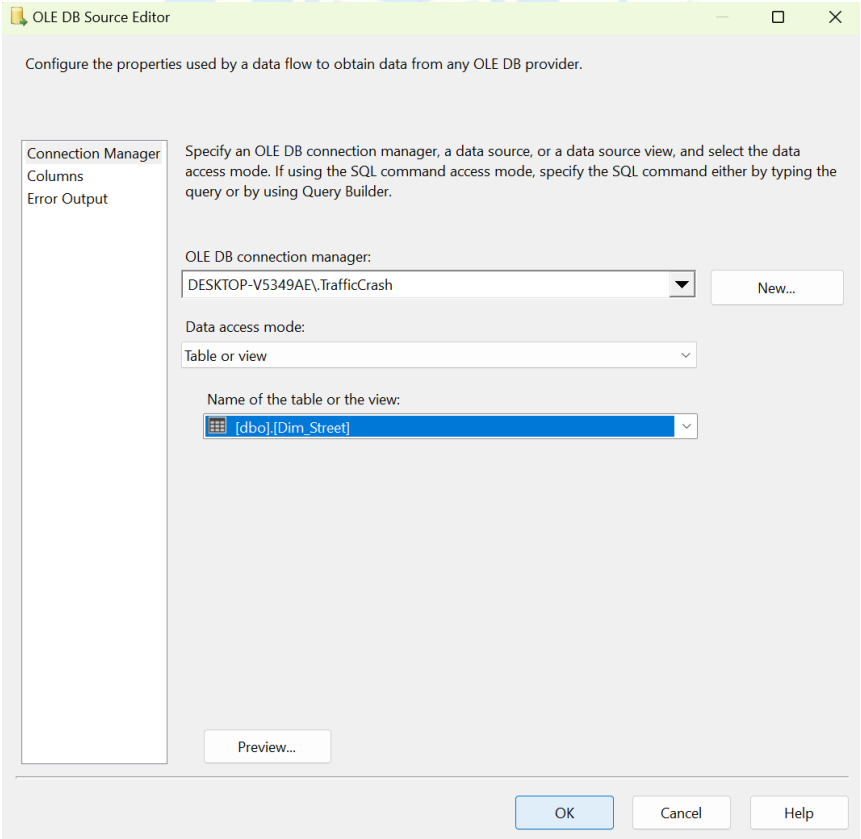
Bước 3: Nhấn chuột phải chọn "Edit", sau đó chọn bảng "Dim_LS" đã tạo trước đó làm data source cho bảng "Dim_LS" mới này.



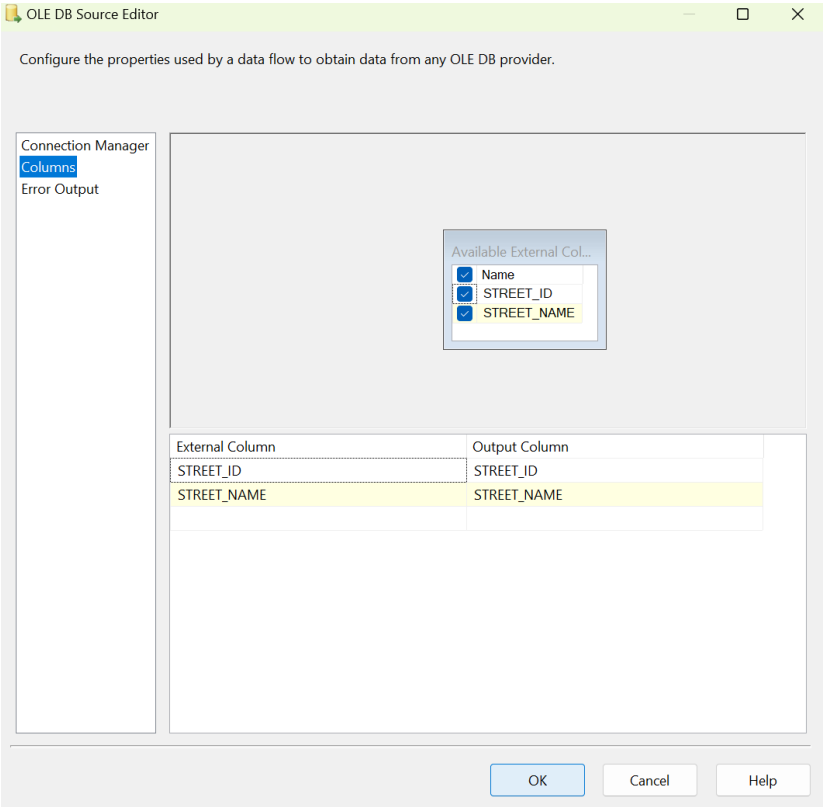
Bước 4: Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



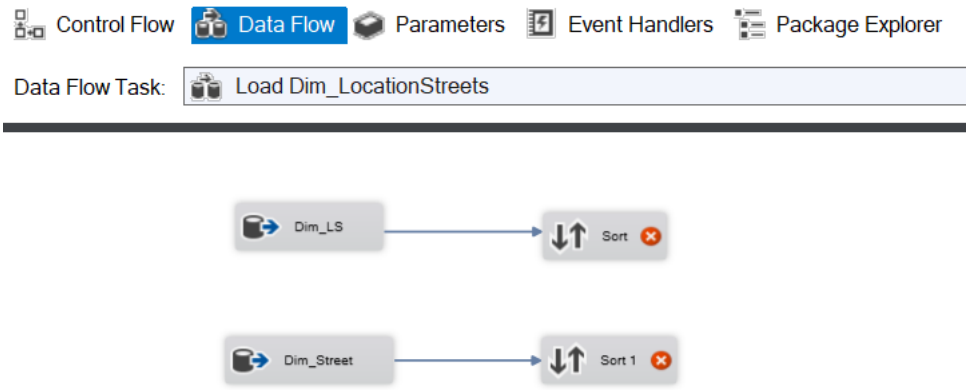
Bước 5: Tương tự thực hiện chọn bảng cho Dim_Street.



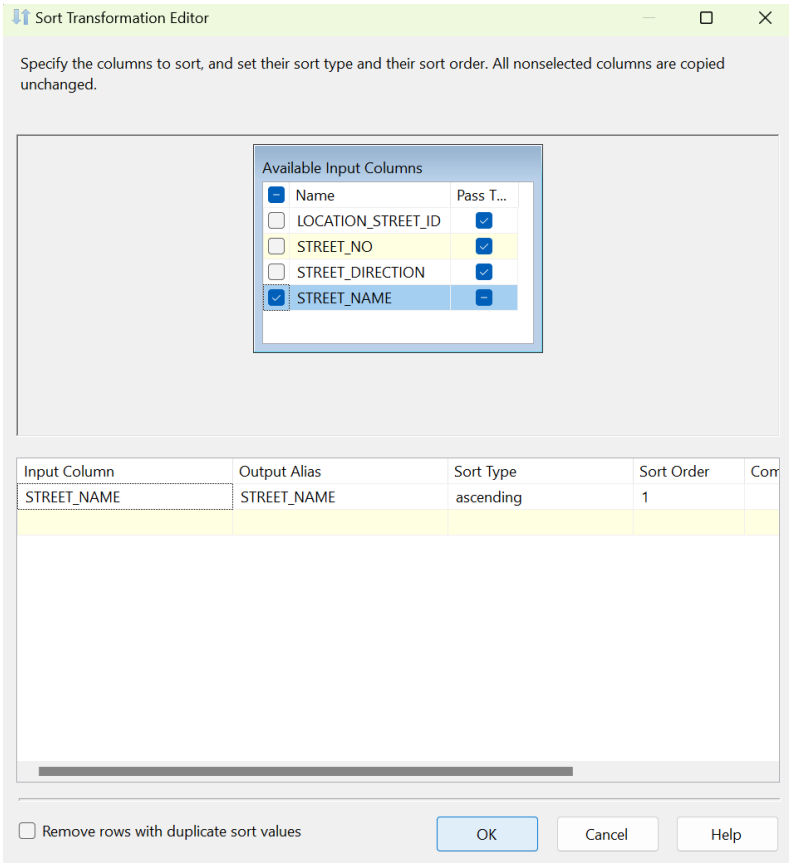
- Chọn mục “Columns” để xem xét các cột được ánh xạ. Sau đó nhấn “OK”.



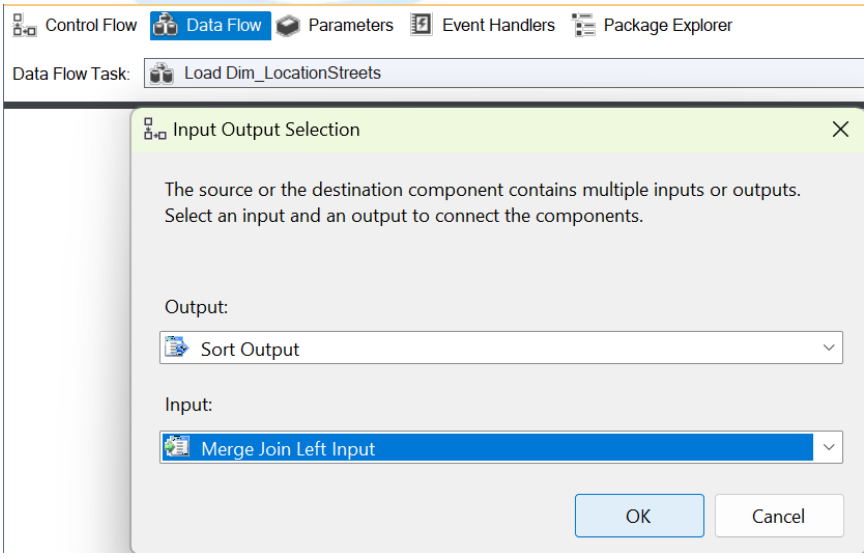
Bước 6: Tạo hai Sort đặt tên là “Sort” và “Sort1” tương ứng với mỗi “OLE DB Source” đã tạo. Thực hiện tạo các kết nối.



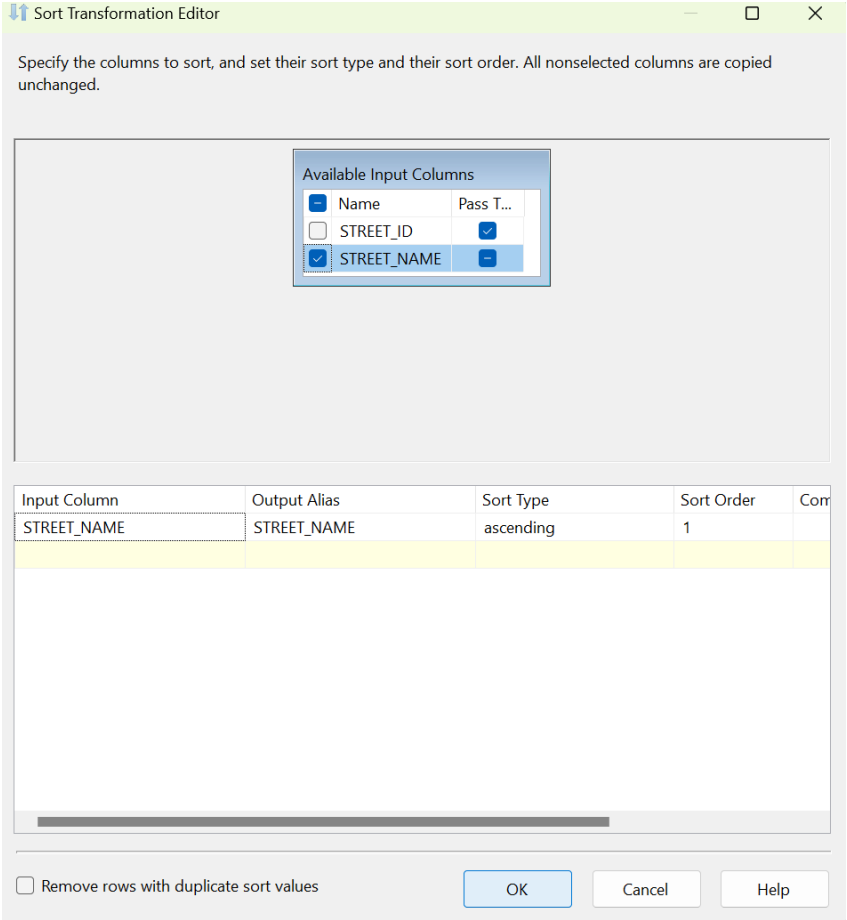
Bước 7: Ở “Sort”, nhấn chuột phải chọn “Edit” và chọn các cột theo thứ tự giống với bảng “Dim_Street” để chuẩn bị cho quá trình merge. Sau đó nhấn “OK”.



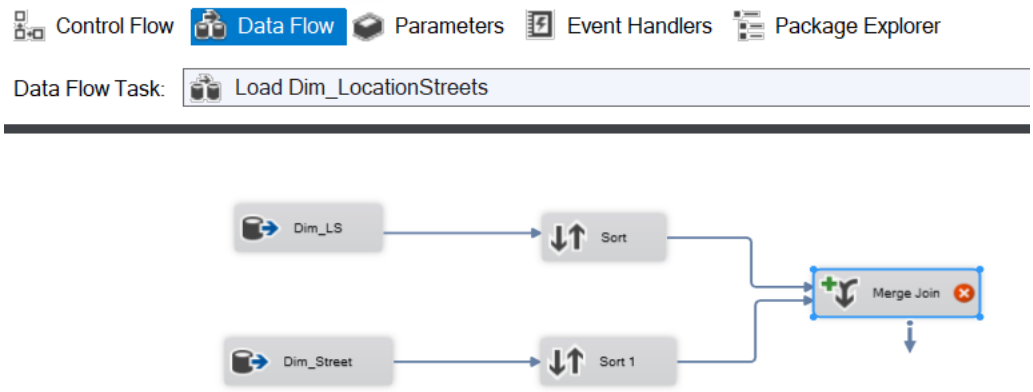
Bước 8: Tạo một Merge Join và nối với "Sort". Tiếp theo ta chọn "Merge Join Left Input" để giữ lại toàn bộ các dòng trong bảng "Dim_LS" bất kể có kết quả khi thực hiện phép kết trái với cột STREET_ID của bảng "Dim_Street" hay không.



Bước 9: Tương tự, ở "Sort1", nhấn chuột phải chọn "Edit" và chọn các cột theo thứ tự. Sau đó nhấn "OK".

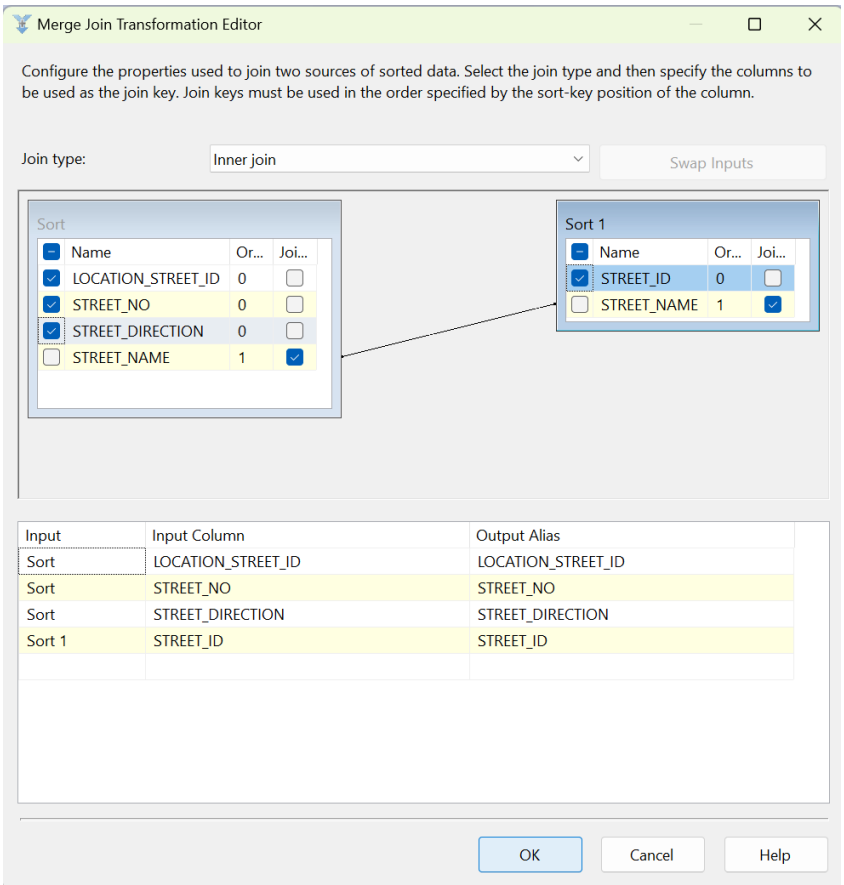


- Kết nối "Sort1" với "Merge Join".



Bước 10: Đầu tiên, nhấn chuột phải vào "Merge Join" và chọn "Edit", một hộp thoại Merge Join Transformation Editor xuất hiện.

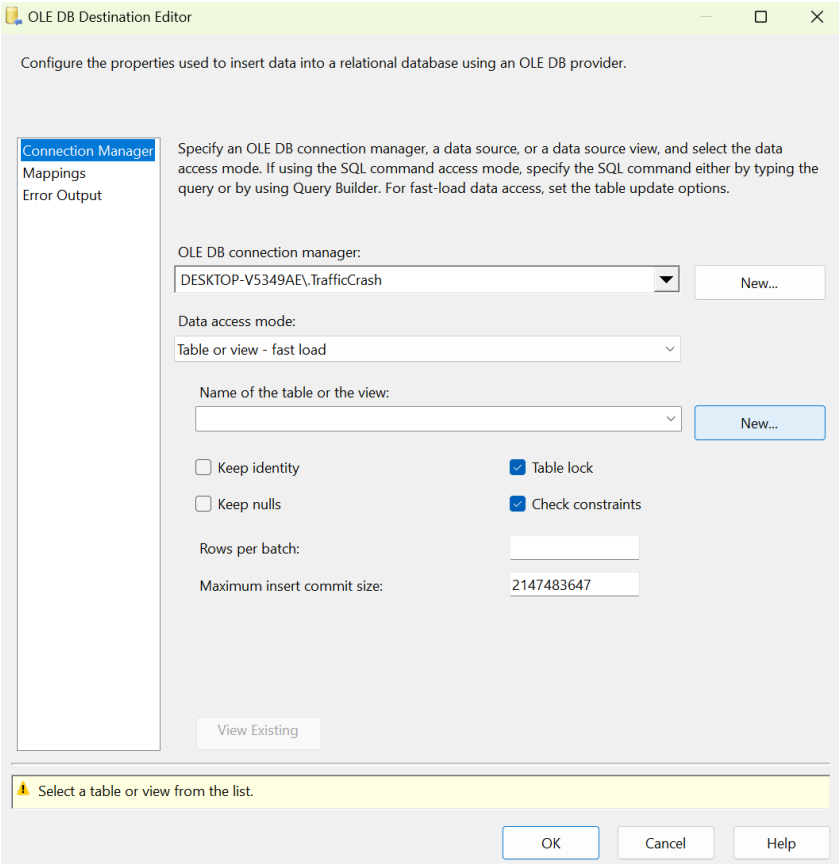
- Sau đó, ta chọn STREET_ID ở “Sort1” để merge vào “Dim_LS”.
- Tiếp theo, ta chọn tất cả các cột của “Sort” nhưng không lấy thuộc tính STREET_NAME.
- Nhấn “OK” để hoàn tất.



Kết quả sau khi merge là bảng Dim_LS không còn thuộc tính STREET_NAME và có thêm 1 thuộc tính mới là STREET_ID.

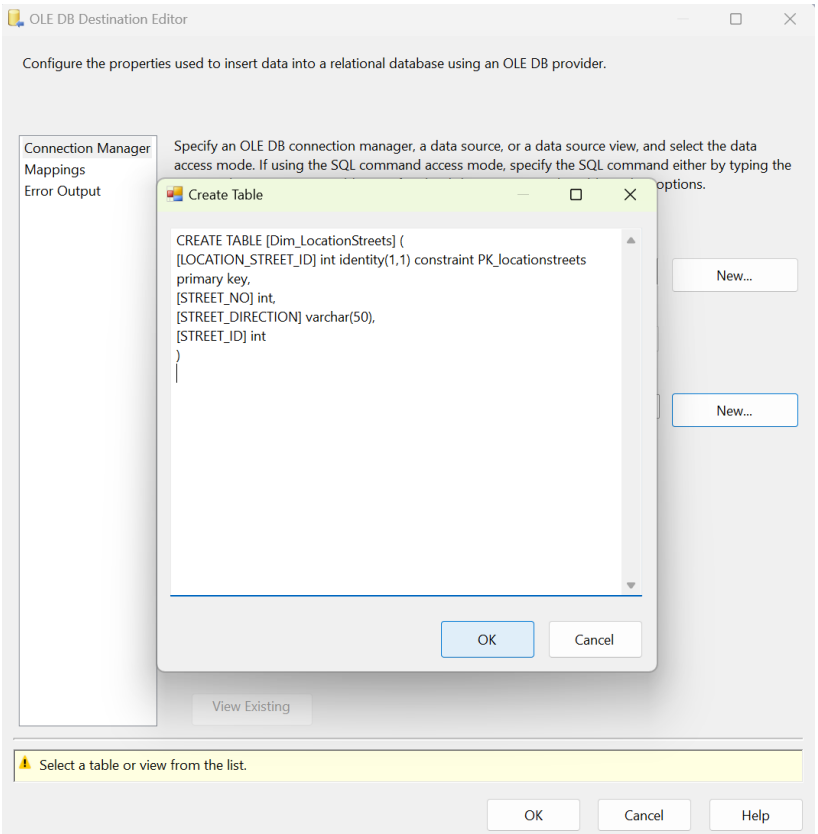
Bước 11: Tạo một OLE DB Destination đặt tên là “Dim_LocationStreets” để chứa tất cả những gì đã merge.

- Nhấn chuột phải vào “Dim_LocationStreets”, chọn “Edit”. Chọn “New” để tạo bảng.

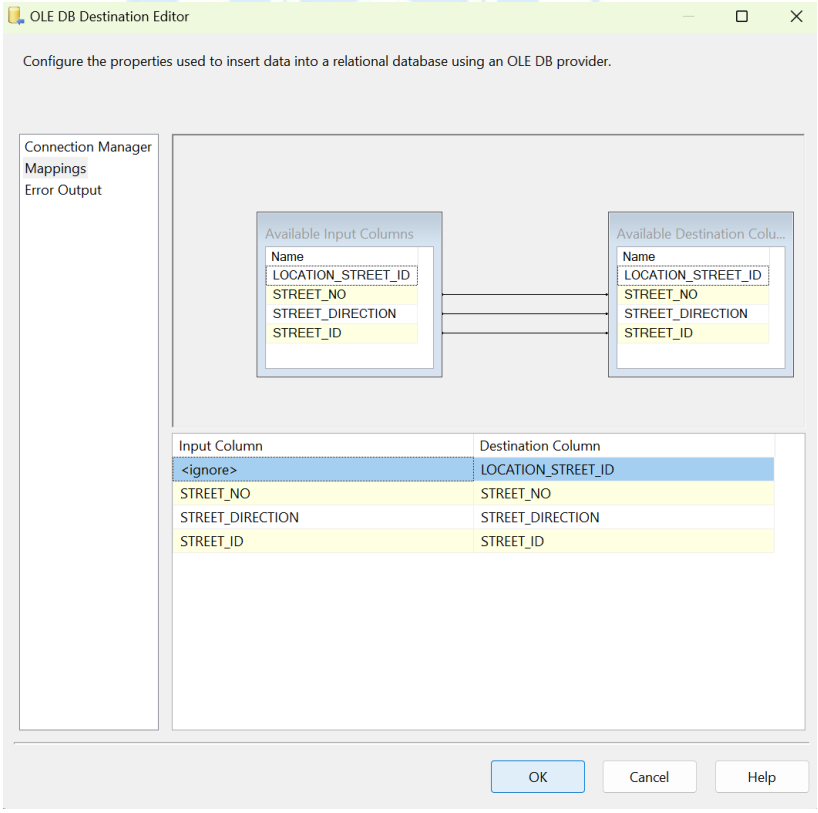


- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Dim_LocationStreets] (
[LOCATION_STREET_ID] int identity(1,1) constraint PK_locationstreets
primary key,
[STREET_NO] int,
[STREET_DIRECTION] varchar(50),
[STREET_ID] int
)
```



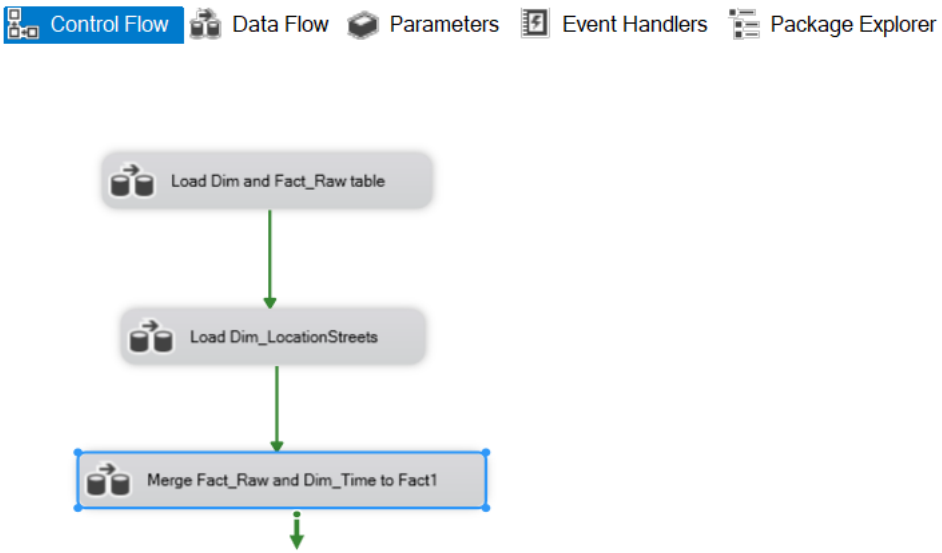
- Vào mục "Mappings" để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn "OK" để hoàn tất.



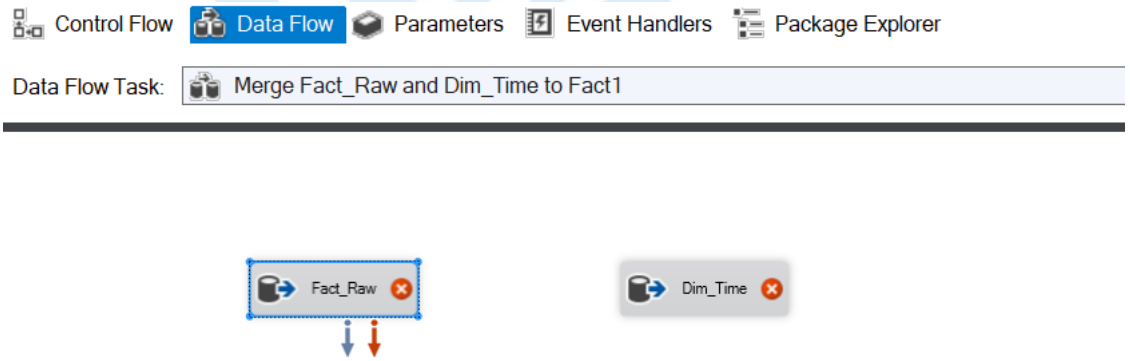
2.3.13 Tạo bảng Fact_TrafficCrashRecords

2.3.13.1 Merge Fact_Raw và Dim_Time vào Fact1

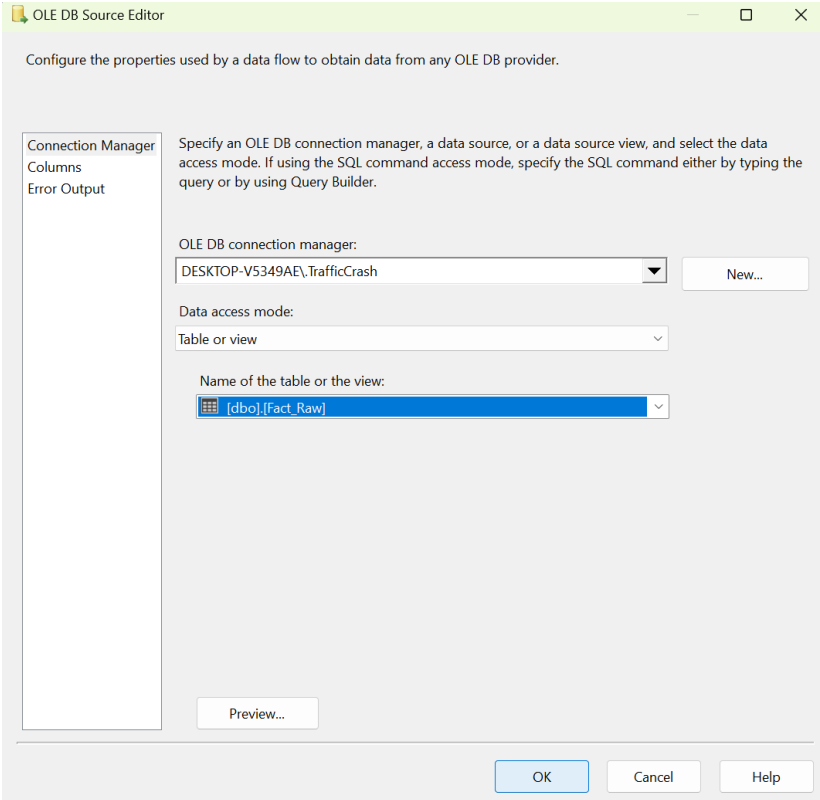
Bước 1: Ở tab Control Flow, tạo một Data Flow Task và đặt tên là "Merge Fact_Raw and Dim_Time to Fact1".



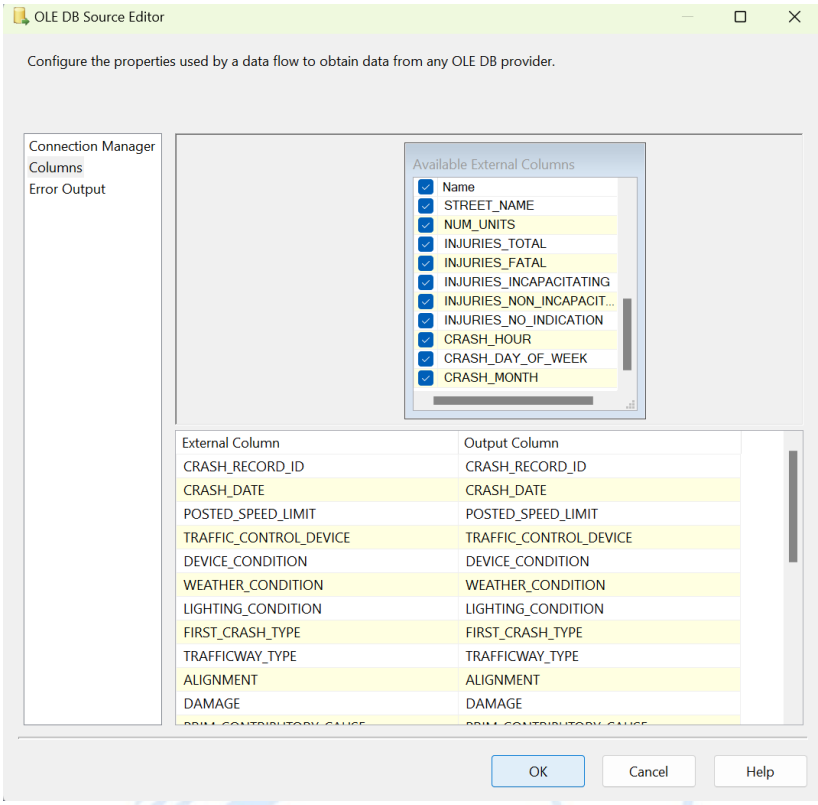
Bước 2: Nhấn chuột phải vào "Merge Fact_Raw and Dim_Time to Fact1" và chọn "Edit", trong tab Data Flow ta tạo hai **OLE DB Source** và đặt tên là "Fact_Raw" và "Dim_Time".



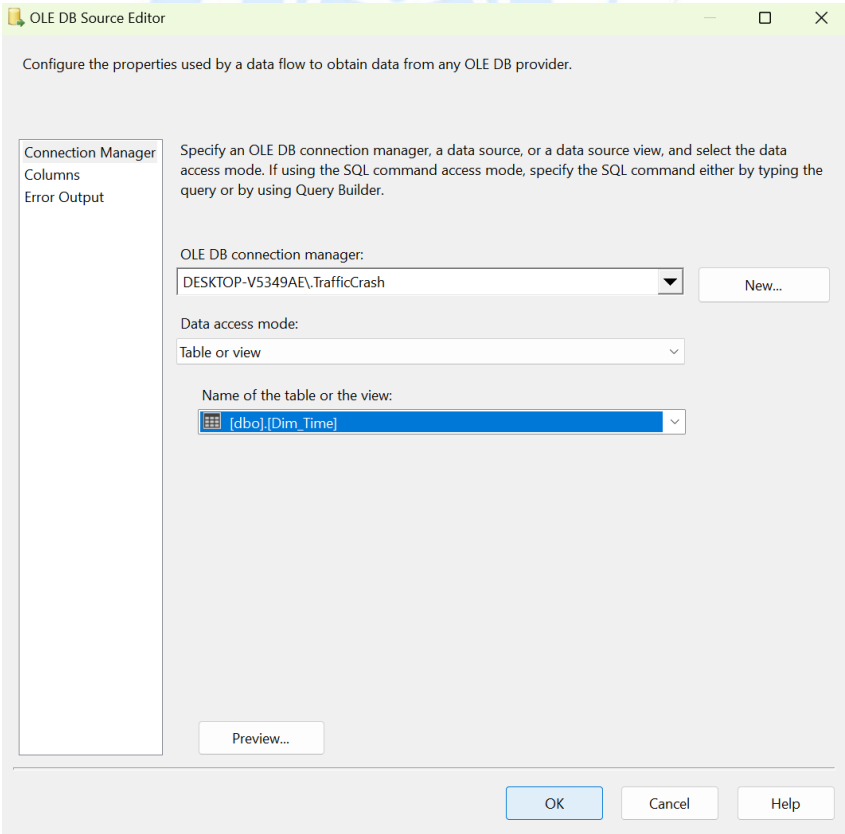
Bước 3: Nhấn chuột phải chọn "Edit", sau đó chọn bảng Fact_Raw đã tạo trước đó làm data source cho bảng Fact_Raw mới này.



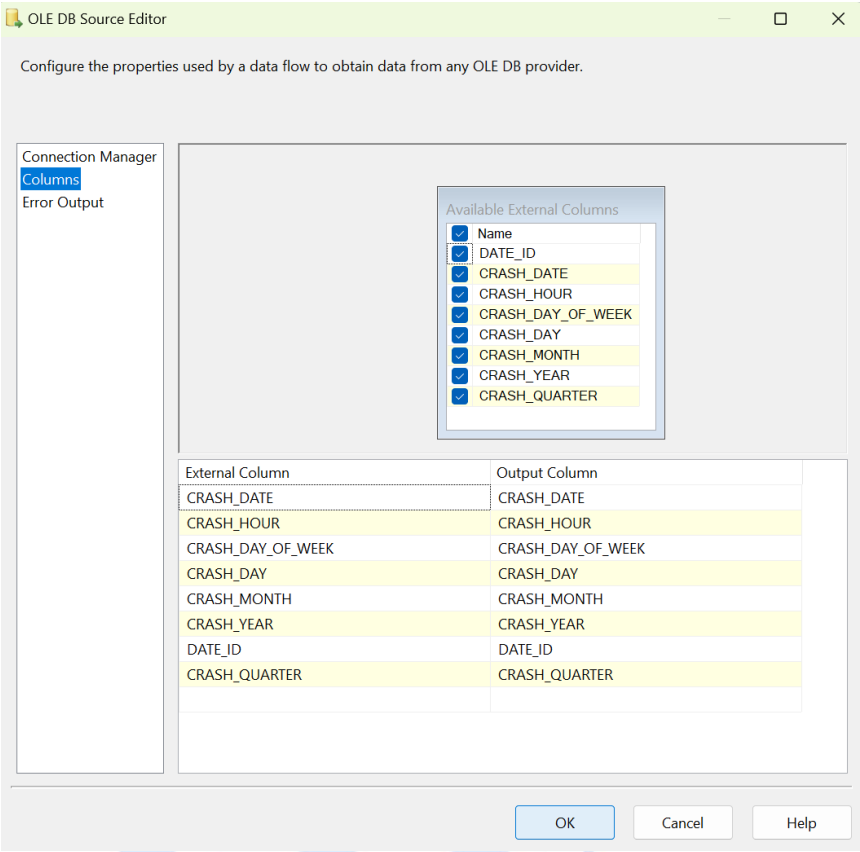
Bước 4: Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



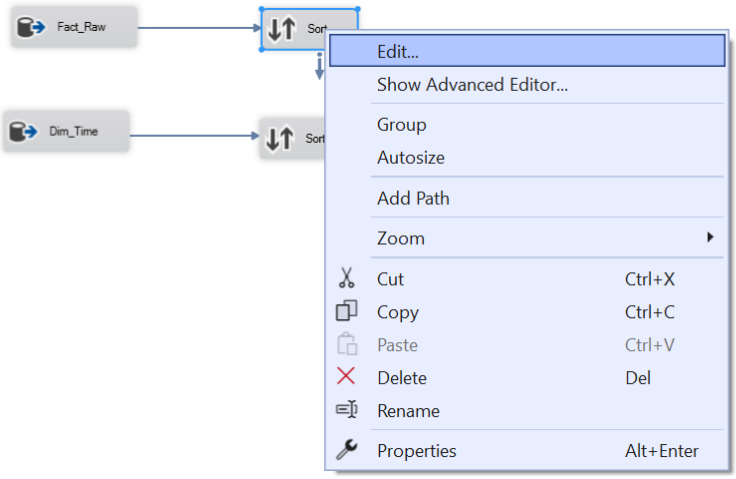
Bước 5: Tương tự thực hiện chọn bảng cho Dim_Time.



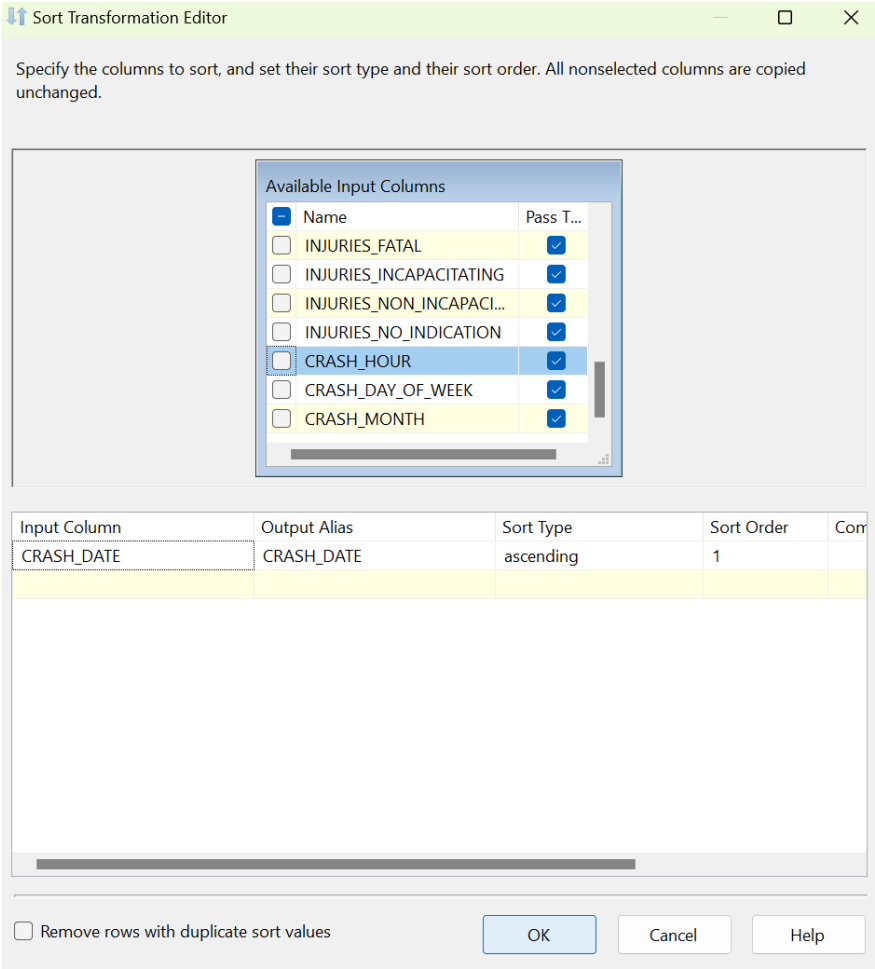
- Chọn mục “Columns” để xem xét các cột được ánh xạ. Sau đó nhấn “OK”.



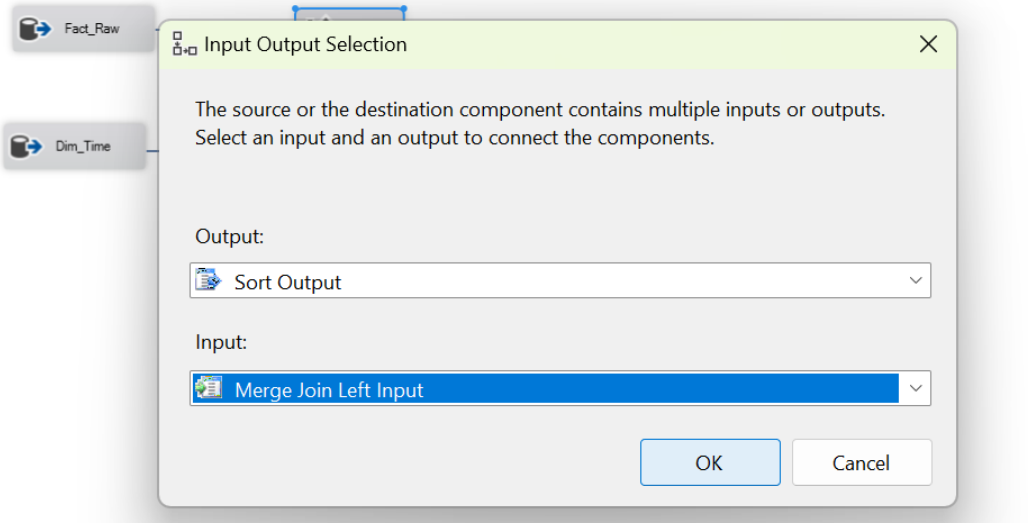
Bước 6: Tạo hai Sort đặt tên là “Sort” và “Sort1” tương ứng với mỗi “OLE DB Source” đã tạo. Thực hiện tạo các kết nối.



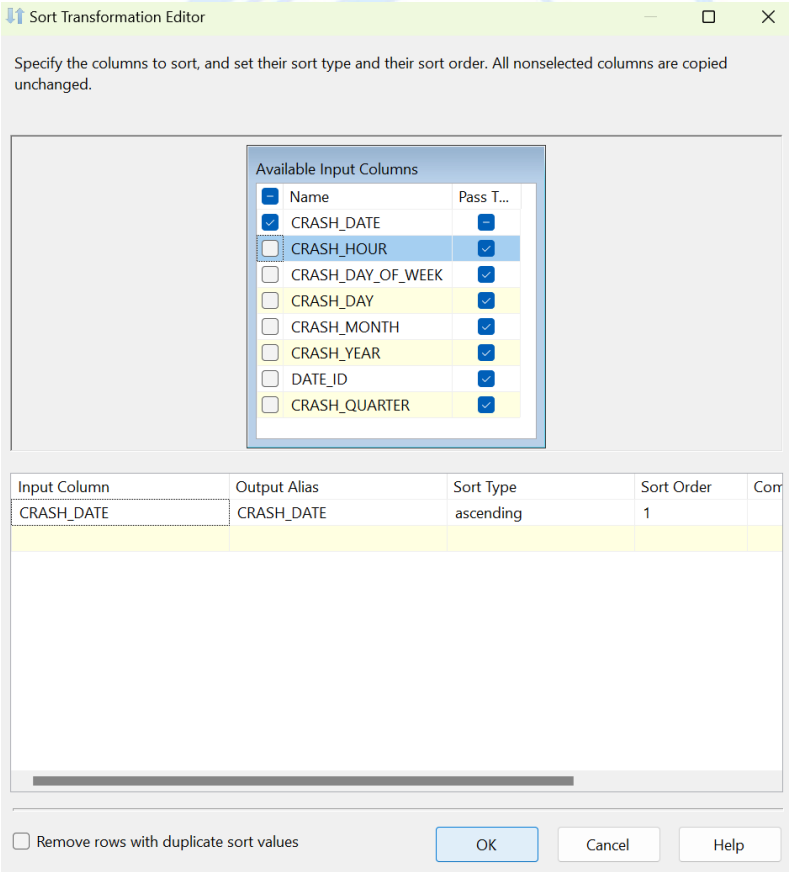
Bước 7: Ở “Sort”, nhấn chuột phải chọn “Edit” và chọn các cột theo thứ tự giống với bảng “Dim_Time” để chuẩn bị cho quá trình merge. Sau đó nhấn “OK”.



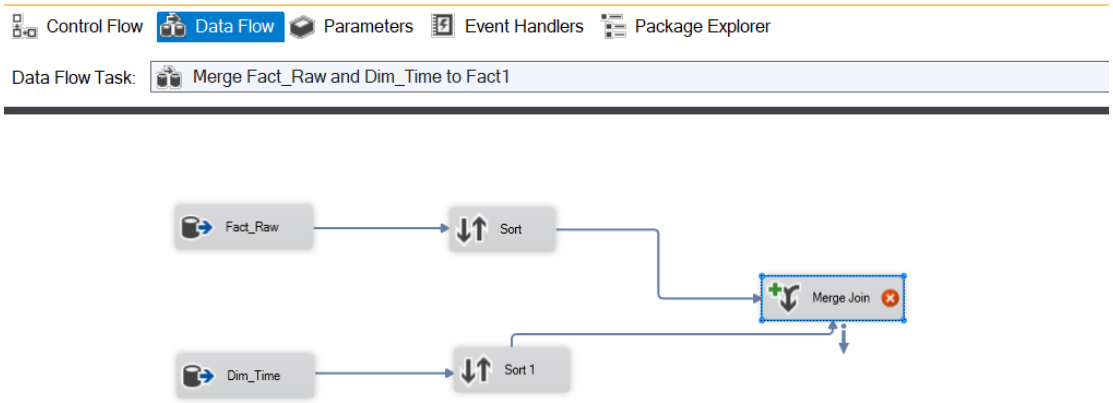
Bước 8: Tạo một Merge Join và nối với "Sort". Tiếp theo ta chọn "Merge Join Left Input" để giữ lại toàn bộ các dòng trong bảng "Fact_Raw" bất kể có kết quả khi thực hiện phép kết trái với cột DATE_ID của bảng "Dim_Time" hay không.



Bước 9: Tương tự, ở "Sort1", nhấn chuột phải chọn "Edit" và chọn các cột theo thứ tự. Sau đó nhấn "OK".

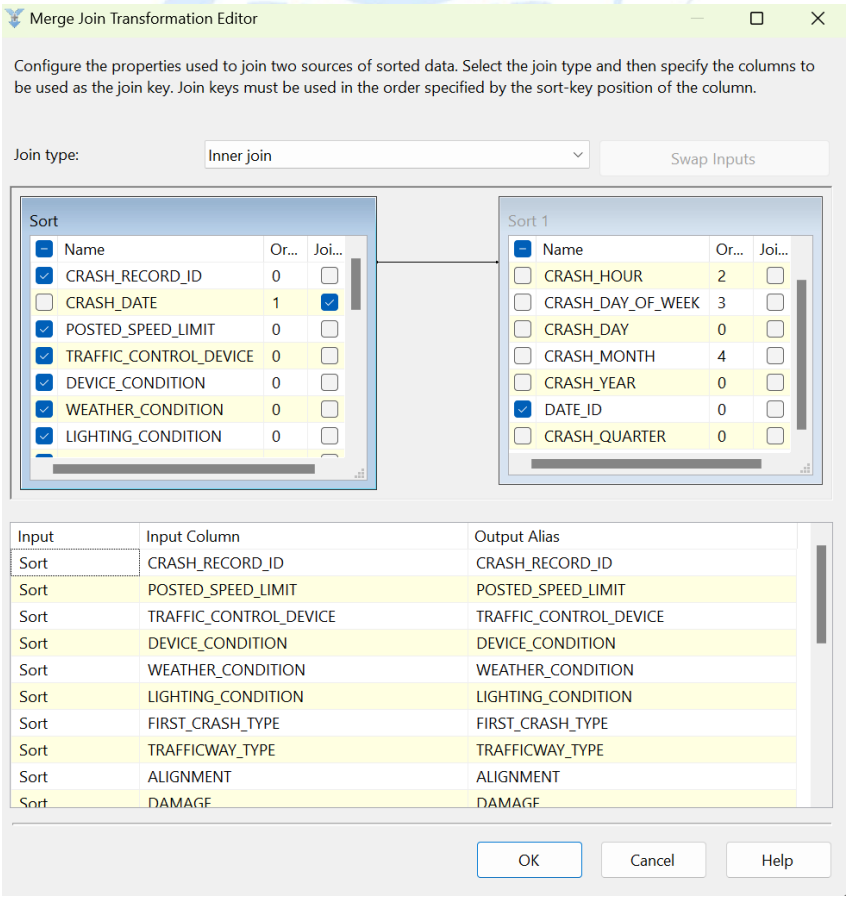


- Kết nối "Sort1" với "Merge Join".



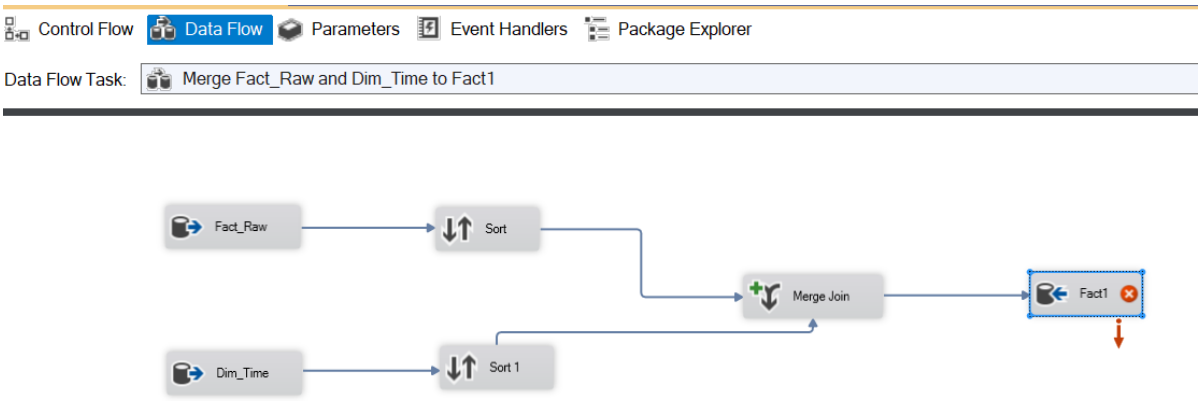
Bước 10: Đầu tiên, nhấn chuột phải vào "Merge Join" và chọn "Edit", một hộp thoại Merge Join Transformation Editor xuất hiện.

- Sau đó, ta chọn DATE_ID ở "Sort1" để merge vào "Fact_Raw".
- Tiếp theo, ta chọn tất cả các cột của "Sort" nhưng thuộc tính CRASH_DATE.
- Nhấn "OK" để hoàn tất.

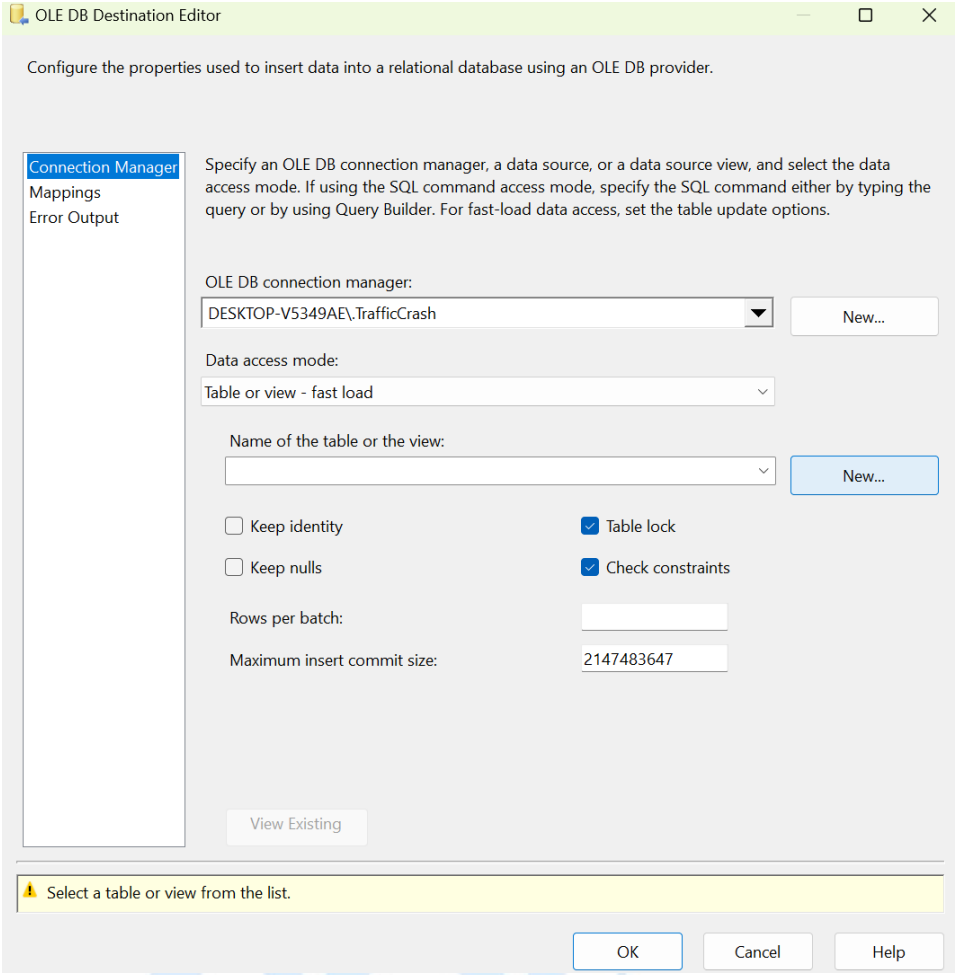


Kết quả sau khi merge là bảng Fact_Raw không còn thuộc tính CRASH_DATE và có thêm 1 thuộc tính mới là DATE_ID.

Bước 11: Tạo một OLE DB Destination đặt tên là "Fact1" để chứa tất cả những gì đã merge.



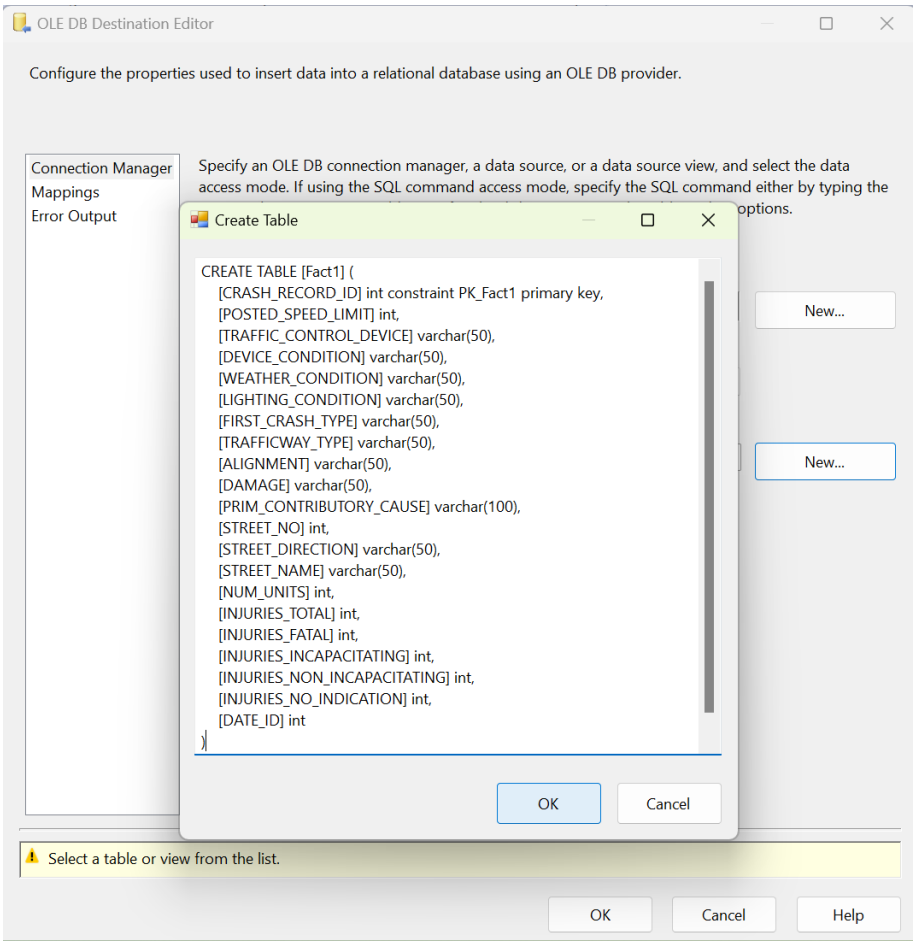
- Nhấn chuột phải vào "Fact1", chọn "Edit". Chọn "New" để tạo bảng.



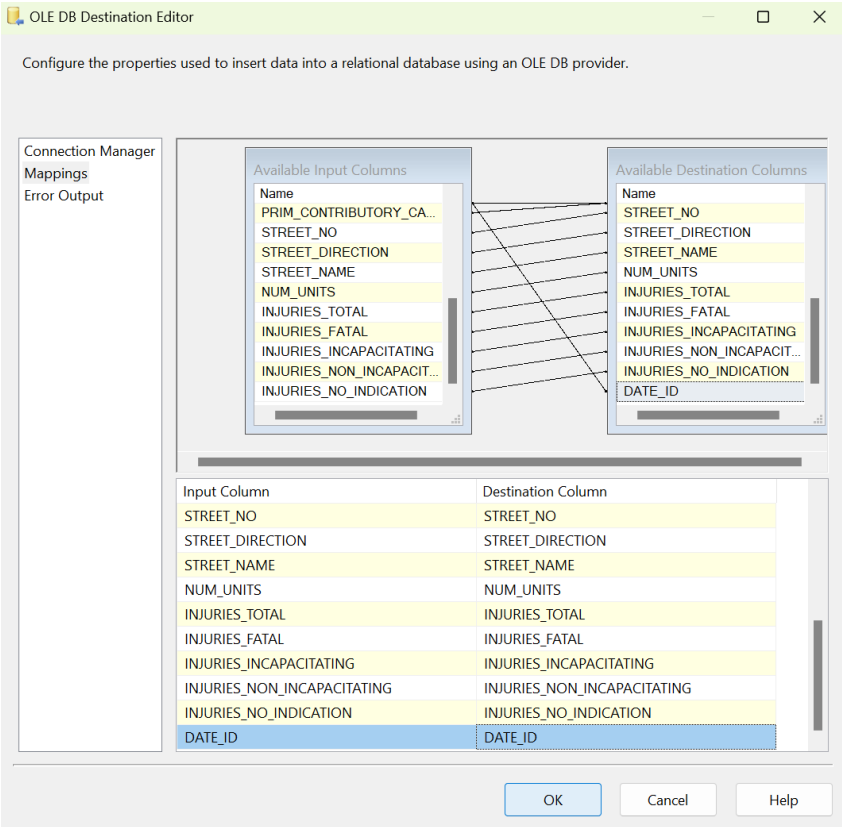
- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Fact1] (
[CRASH_RECORD_ID] int constraint PK_Fact1 primary key,
[POSTED_SPEED_LIMIT] int,
[TRAFFIC_CONTROL_DEVICE] varchar(50),
[DEVICE_CONDITION] varchar(50),
[WEATHER_CONDITION] varchar(50),
[LIGHTING_CONDITION] varchar(50),
[FIRST_CRASH_TYPE] varchar(50),
[TRAFFICWAY_TYPE] varchar(50),
[ALIGNMENT] varchar(50),
[DAMAGE] varchar(50),
[PRIM_CONTRIBUTORY_CAUSE] varchar(100),
[STREET_NO] int,
```

```
    [STREET_DIRECTION] varchar(50),
    [STREET_NAME] varchar(50),
    [NUM_UNITS] int,
    [INJURIES_TOTAL] int,
    [INJURIES_FATAL] int,
    [INJURIES_INCAPACITATING] int,
    [INJURIES_NON_INCAPACITATING] int,
    [INJURIES_NO_INDICATION] int,
    [DATE_ID] int
)
```



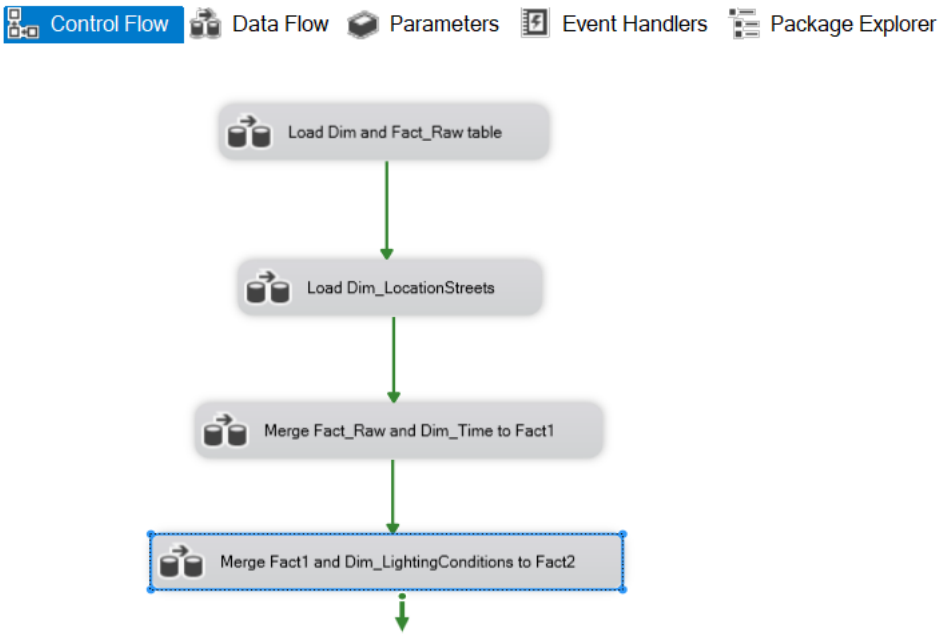
- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



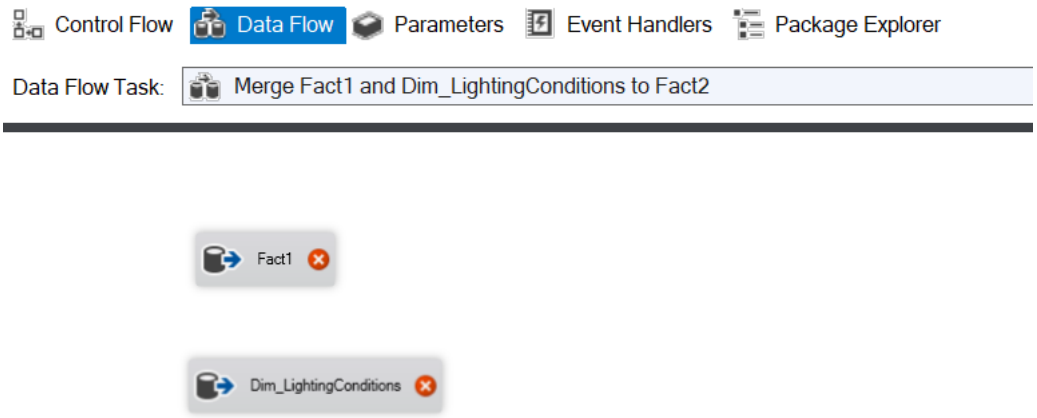
Tiếp tục quá trình merge bảng Fact1 với các Dimension còn lại để có 1 bảng Fact_TrafficCrashRecords hoàn chỉnh.

2.3.13.2 Merge Fact1 và Dim_LightingConditions vào Fact2.

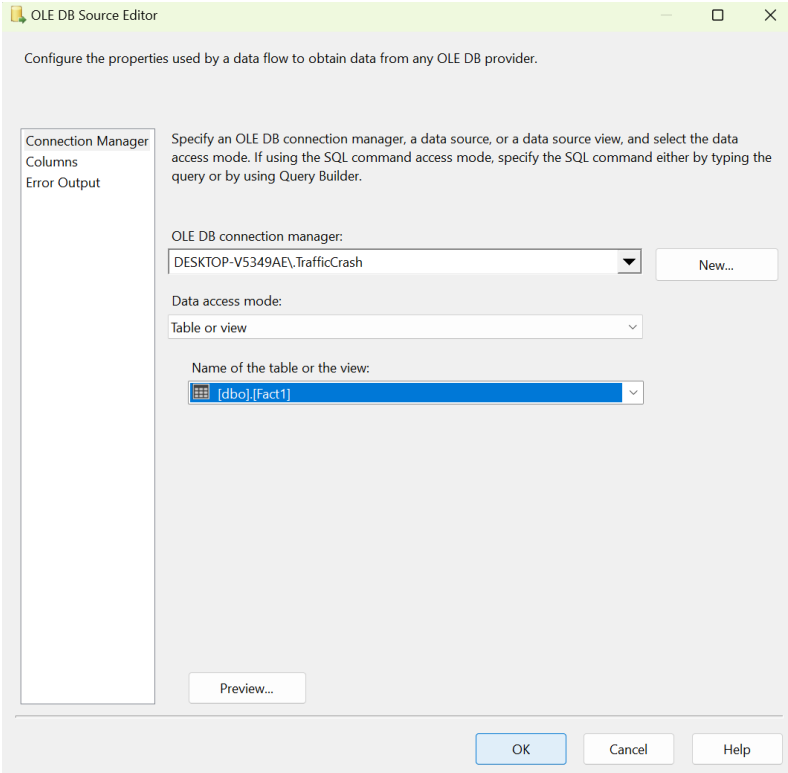
Bước 1: Ở tab Control Flow, tạo một Data Flow Task và đặt tên là “Merge Fact1 and Dim_LightingConditions to Fact2”.



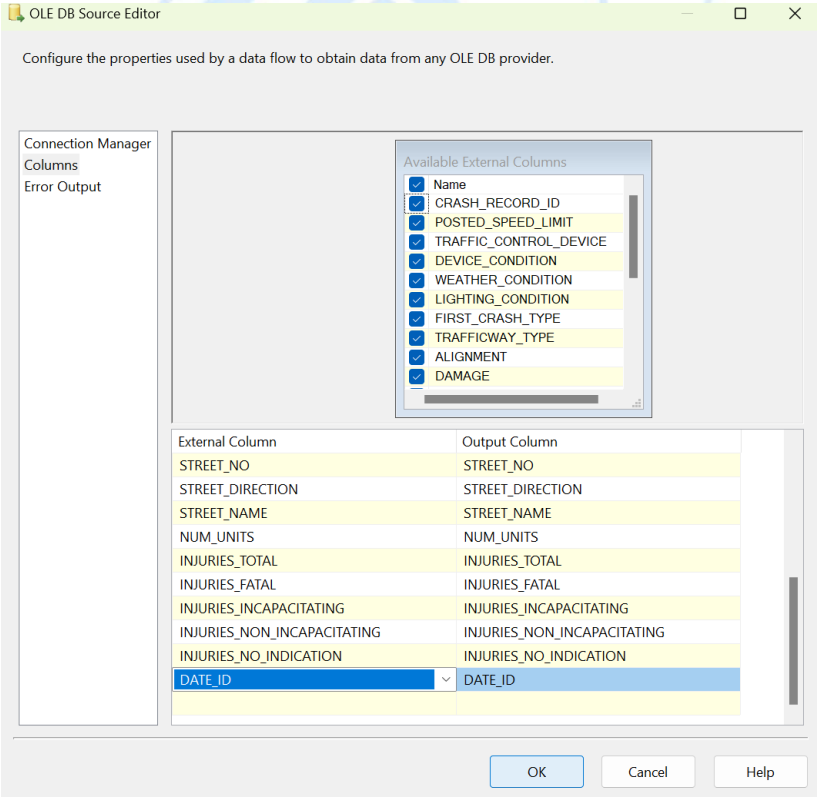
Bước 2: Nhấn chuột phải vào "Merge Fact1 and Dim_LightingConditions to Fact2" và chọn "Edit", trong tab Data Flow ta tạo hai **OLE DB Source** và đặt tên là "Fact1" và "Dim_LightingConditions".



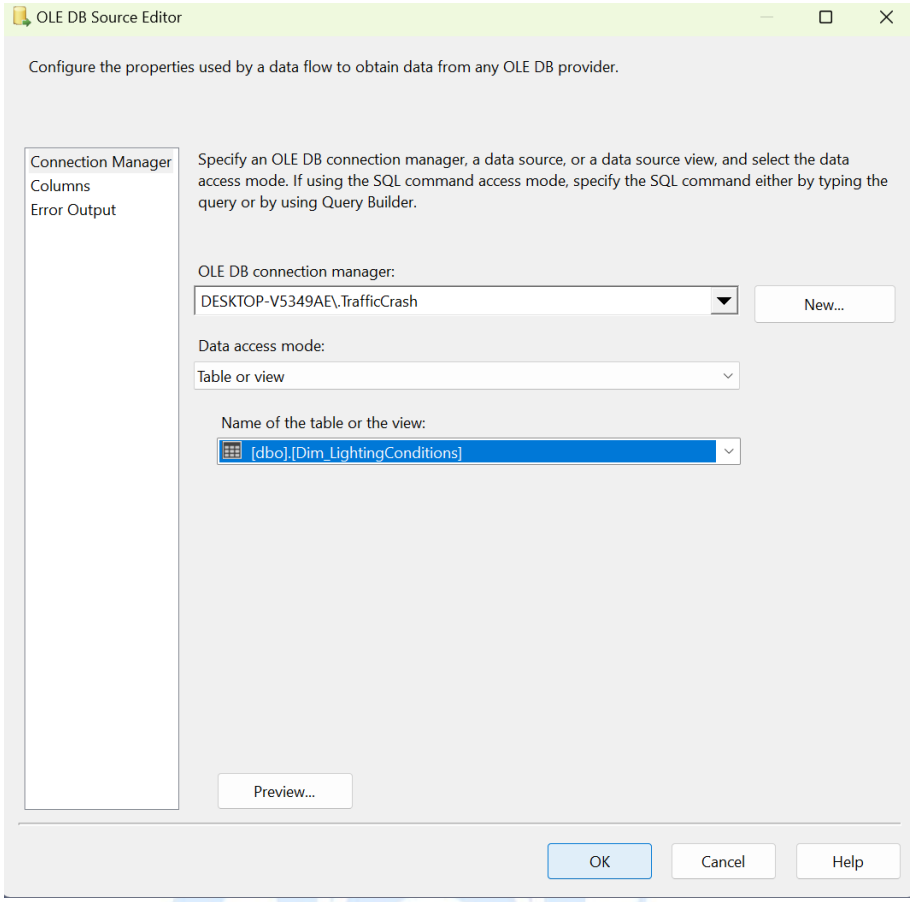
Bước 3: Nhấn chuột phải chọn "Edit", sau đó chọn bảng Fact1 đã tạo trước đó làm data source cho bảng Fact1 mới này.



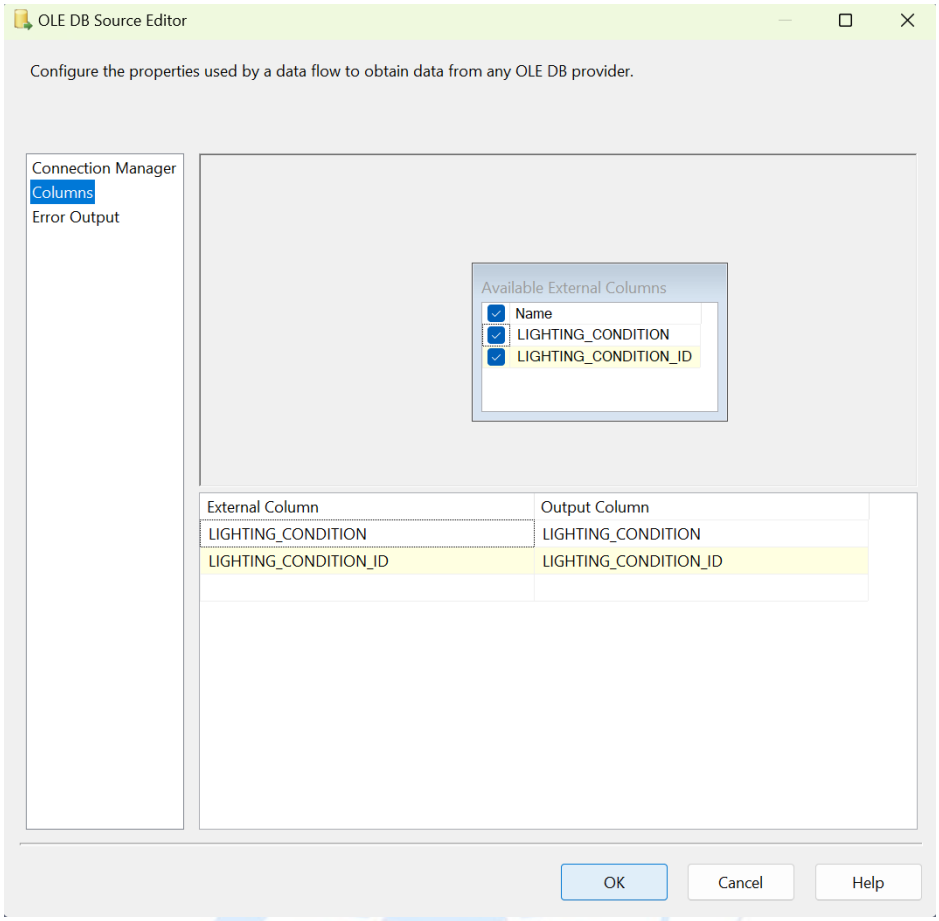
Bước 4: Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



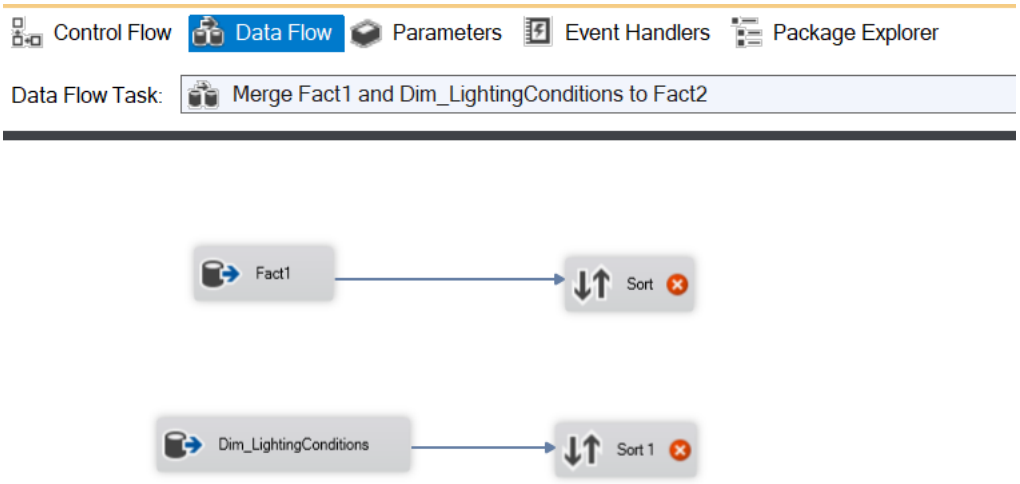
Bước 5: Tương tự thực hiện chọn bảng cho Dim_LightingConditions.



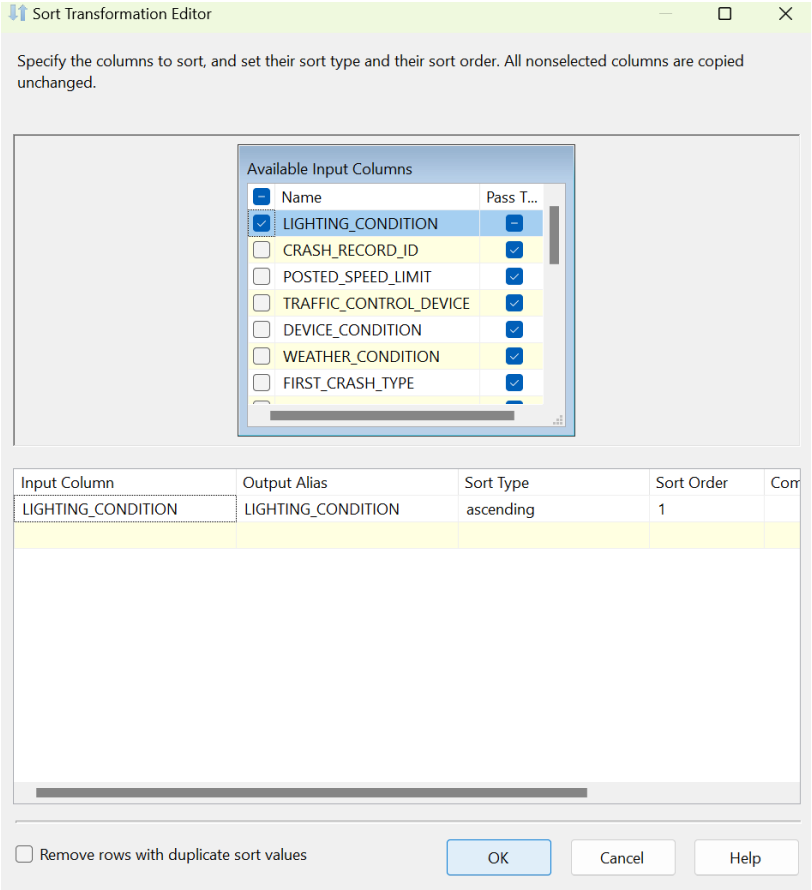
- Chọn mục “Columns” để xem xét các cột được ánh xạ. Sau đó nhấn “OK”.



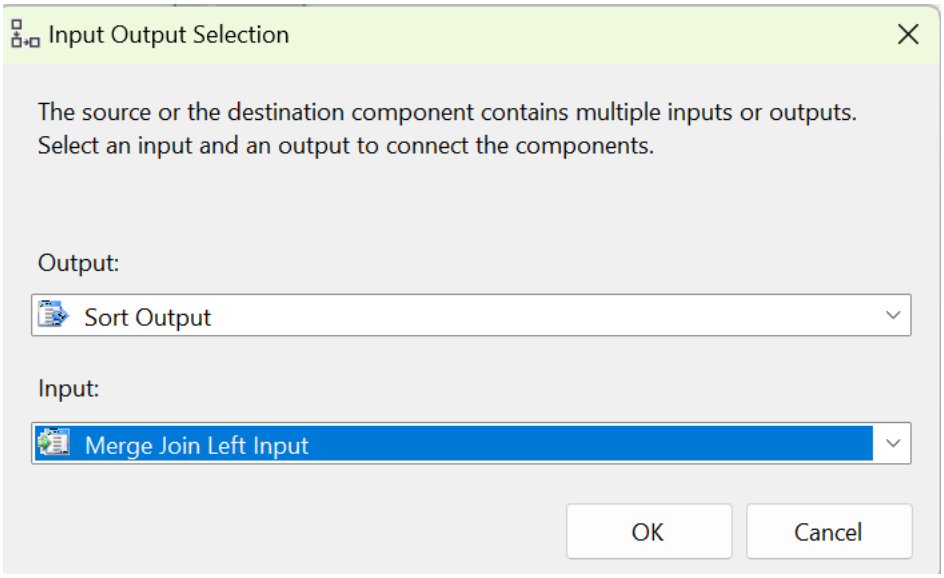
Bước 6: Tạo hai Sort đặt tên là "Sort" và "Sort1" tương ứng với mỗi "OLE DB Source" đã tạo. Thực hiện tạo các kết nối.



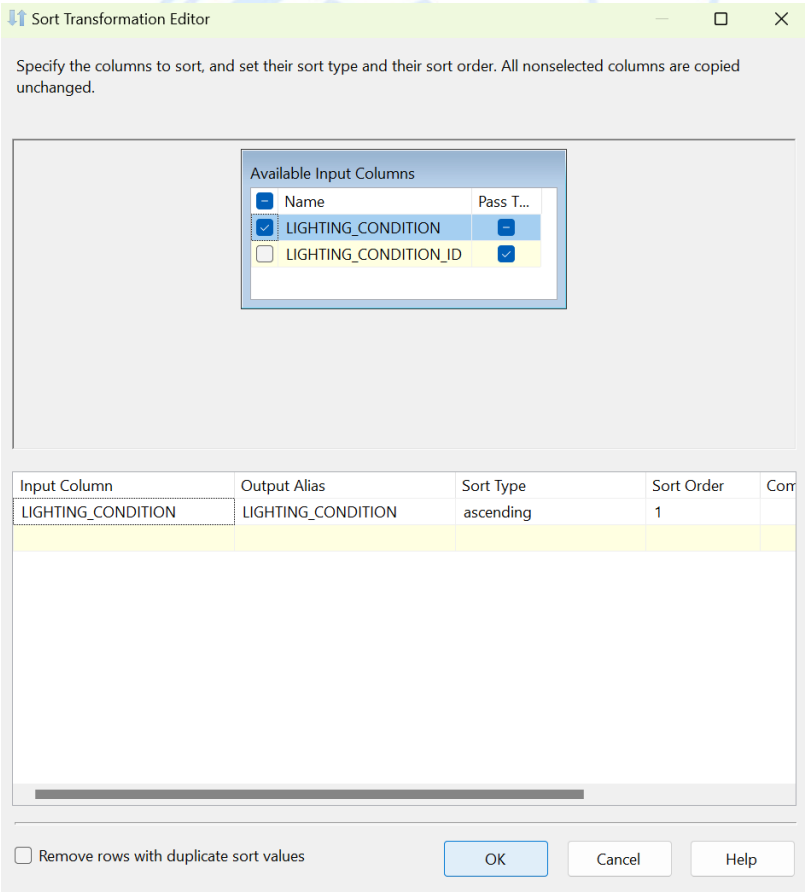
Bước 7: Ở "Sort", nhấn chuột phải chọn "Edit" và chọn cột để chuẩn bị cho quá trình merge. Sau đó nhấn "OK".



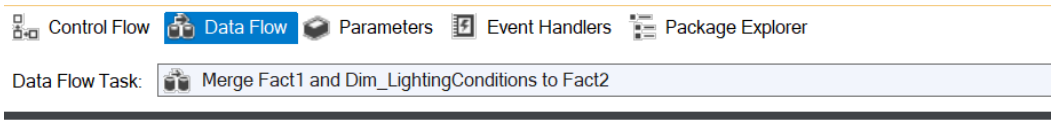
Bước 8: Tạo một Merge Join và nối với "Sort". Tiếp theo ta chọn "Merge Join Left Input" để giữ lại toàn bộ các dòng trong bảng "Fact1" bất kể có kết quả khi thực hiện phép kết trái với cột LIGHTING_CONDITION_ID của bảng "Dim_LightingConditions" hay không.

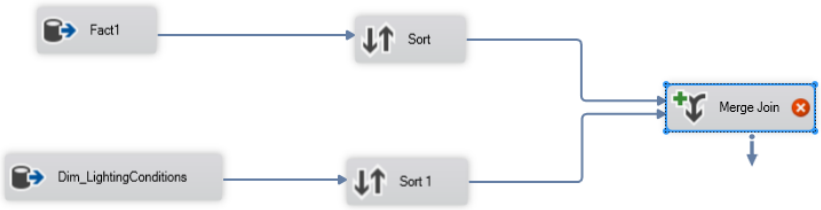


Bước 9: Tương tự, ở "Sort1", nhấn chuột phải chọn "Edit" và chọn cột. Sau đó nhấn "OK".



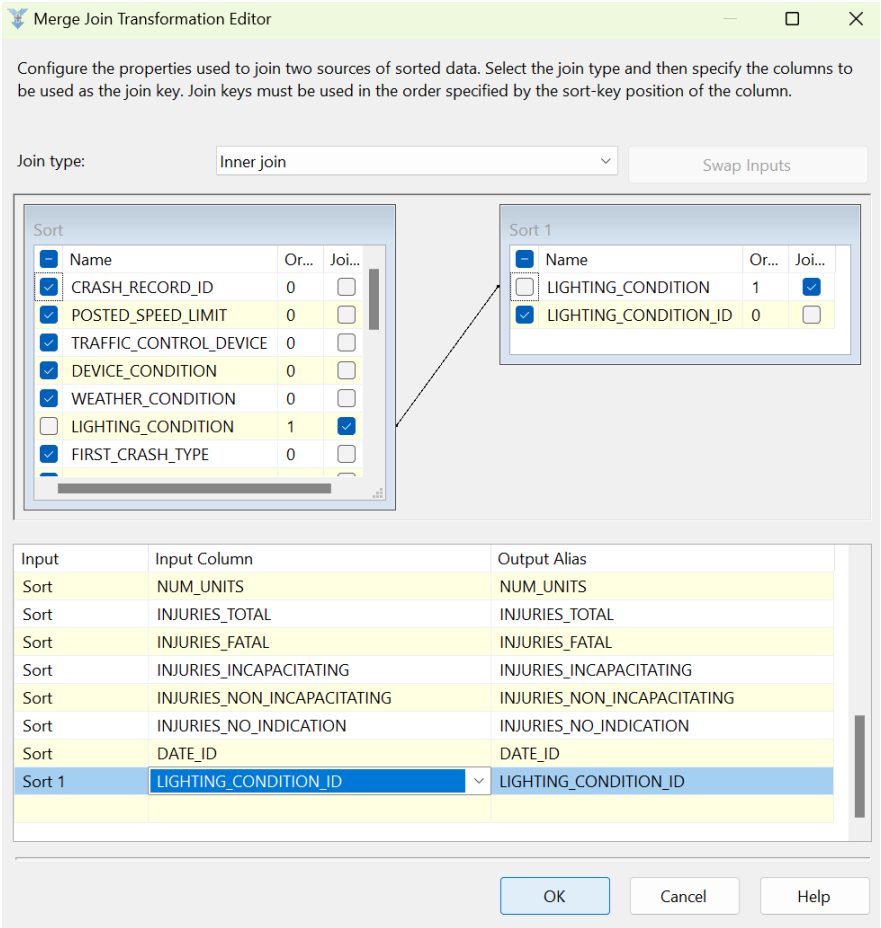
- Kết nối “Sort1” với Merge Join.





Bước 10: Đầu tiên, nhấn chuột phải vào “Merge Join” và chọn “Edit”, một hộp thoại Merge Join Transformation Editor xuất hiện.

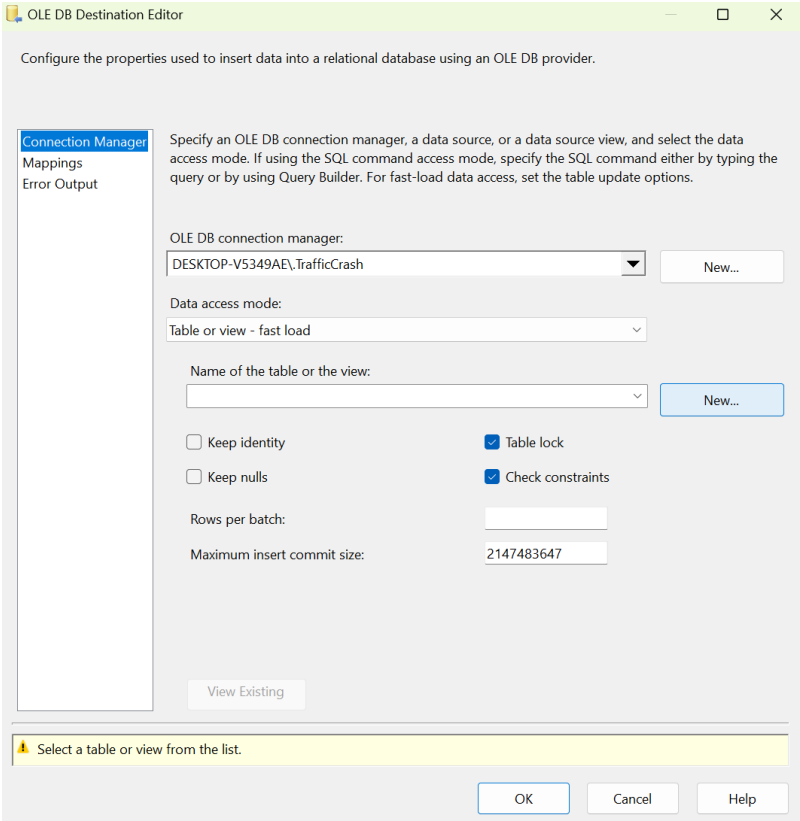
- Sau đó, ta chọn LIGHTING_CONDITION_ID ở “Sort1” để merge vào “Fact1”.
- Tiếp theo, ta chọn tất cả các cột của “Sort” nhưng không lấy thuộc tính LIGHTING_CONDITION.
- Nhấn “OK” để hoàn tất.



Kết quả sau khi merge là bảng Fact1 không còn thuộc tính LIGHTING_CONDITION và có thêm 1 thuộc tính mới là LIGHTING_CONDITION_ID.

Bước 11: Tạo một OLE DB Destination đặt tên là “Fact2” để chứa tất cả những gì đã merge.

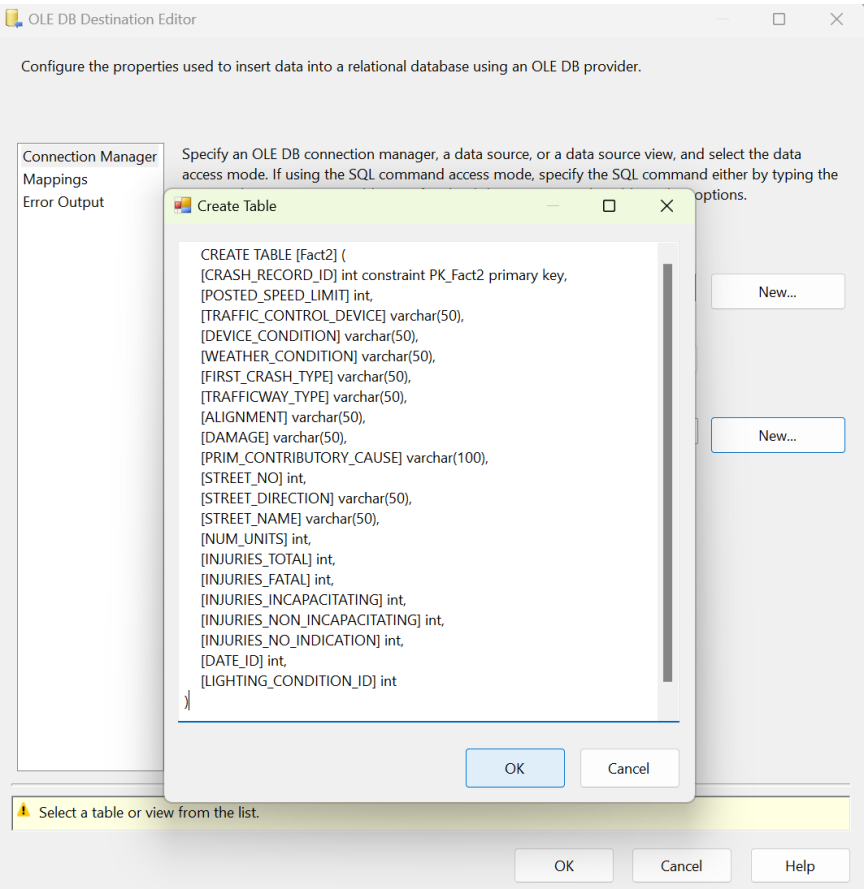
- Nhấn chuột phải vào “Fact2”, chọn “Edit”. Chọn “New” để tạo bảng.



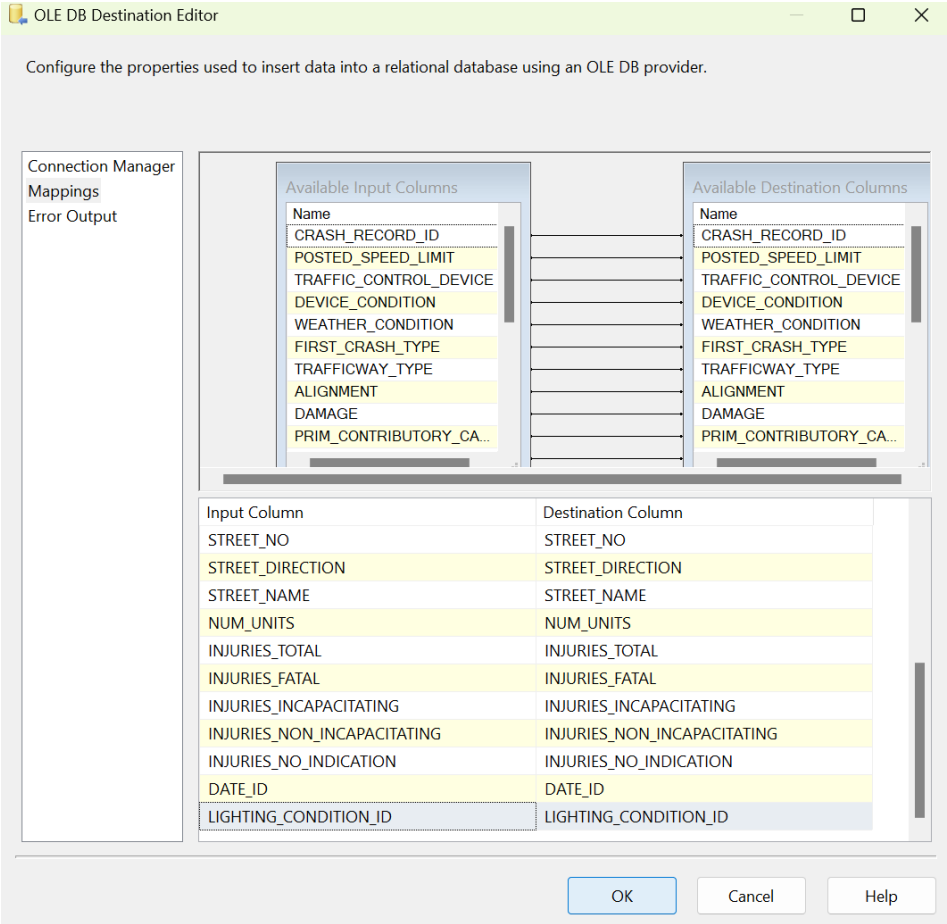
- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Fact2] (
[CRASH_RECORD_ID] int constraint PK_Fact2 primary key,
[POSTED_SPEED_LIMIT] int,
[TRAFFIC_CONTROL_DEVICE] varchar(50),
[DEVICE_CONDITION] varchar(50),
[WEATHER_CONDITION] varchar(50),
[FIRST_CRASH_TYPE] varchar(50),
[TRAFFICWAY_TYPE] varchar(50),
[ALIGNMENT] varchar(50),
[DAMAGE] varchar(50),
[PRIM_CONTRIBUTORY_CAUSE] varchar(100),
[STREET_NO] int,
[STREET_DIRECTION] varchar(50),
[STREET_NAME] varchar(50),
[NUM_UNITS] int,
[INJURIES_TOTAL] int,
[INJURIES_FATAL] int,
[INJURIES_INCAPACITATING] int,
[INJURIES_NON_INCAPACITATING] int,
```

```
[INJURIES_NO_INDICATION] int,
[DATE_ID] int,
[LIGHTING_CONDITION_ID] int
)
```



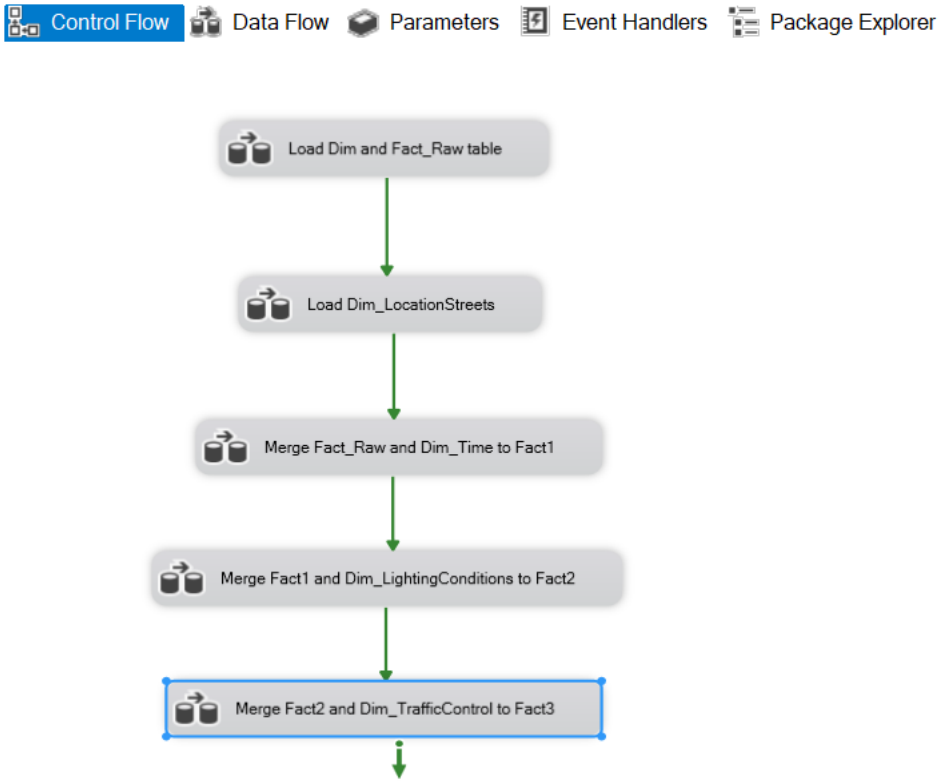
- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



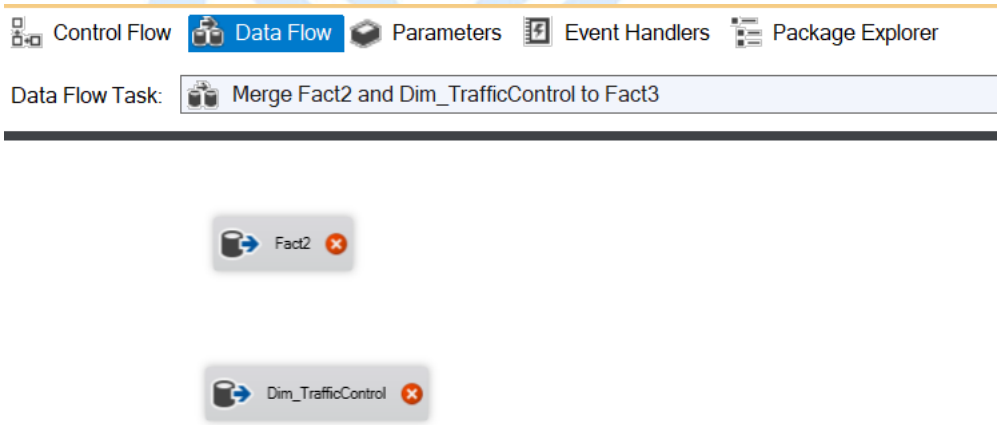
Tiếp tục quá trình merge bảng Fact2 với các Dimension còn lại để có 1 bảng Fact_TrafficCrashRecords hoàn chỉnh.

2.3.13.3 Merge Fact2 và Dim_TrafficControl vào Fact3.

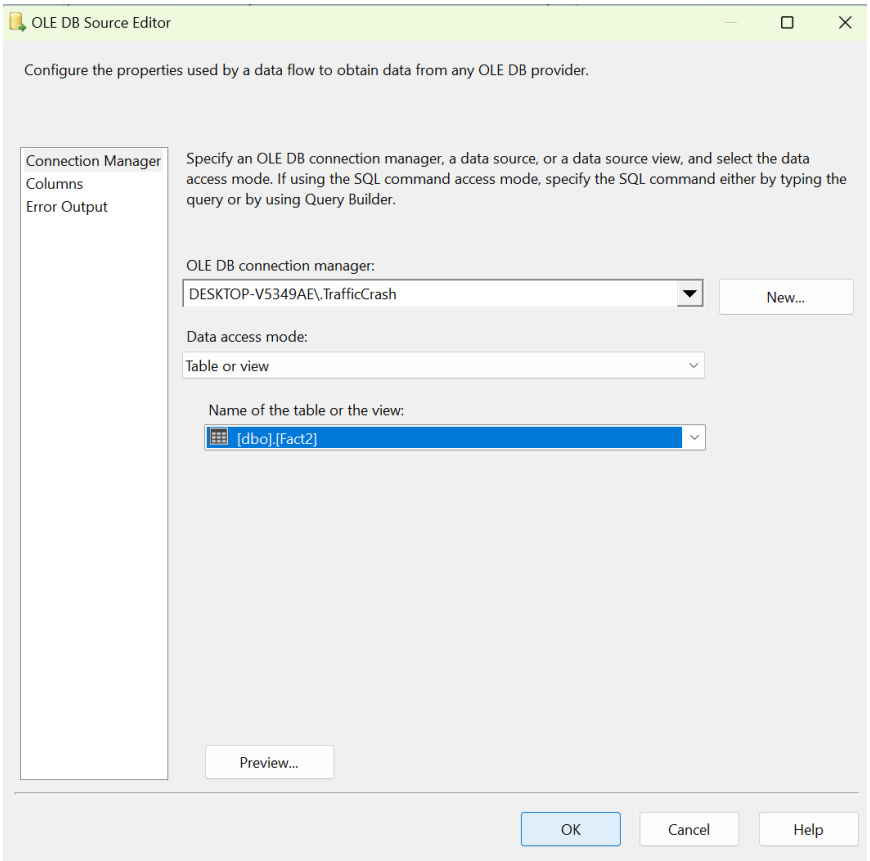
Bước 1: Ở tab Control Flow, tạo một Data Flow Task và đặt tên là “Merge Fact2 and Dim_TrafficControl to Fact3”.



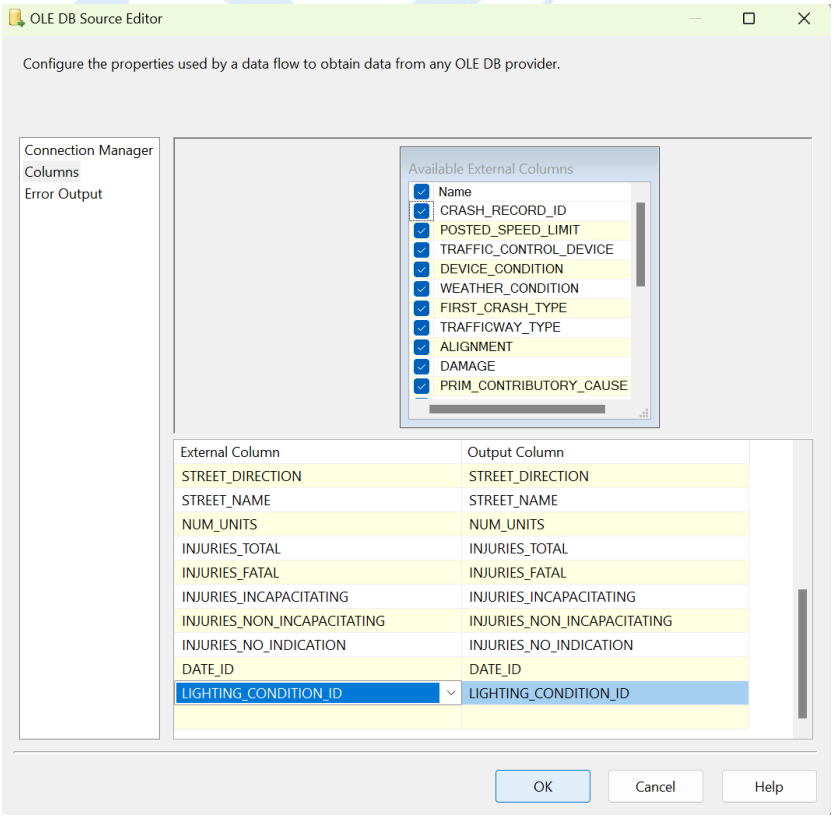
Bước 2: Nhấn chuột phải vào "Merge Fact2 and Dim_TrafficControl to Fact3" và chọn "Edit", trong tab Data Flow ta tạo hai **OLE DB Source** và đặt tên là "Fact2" và "Dim_TrafficControl".



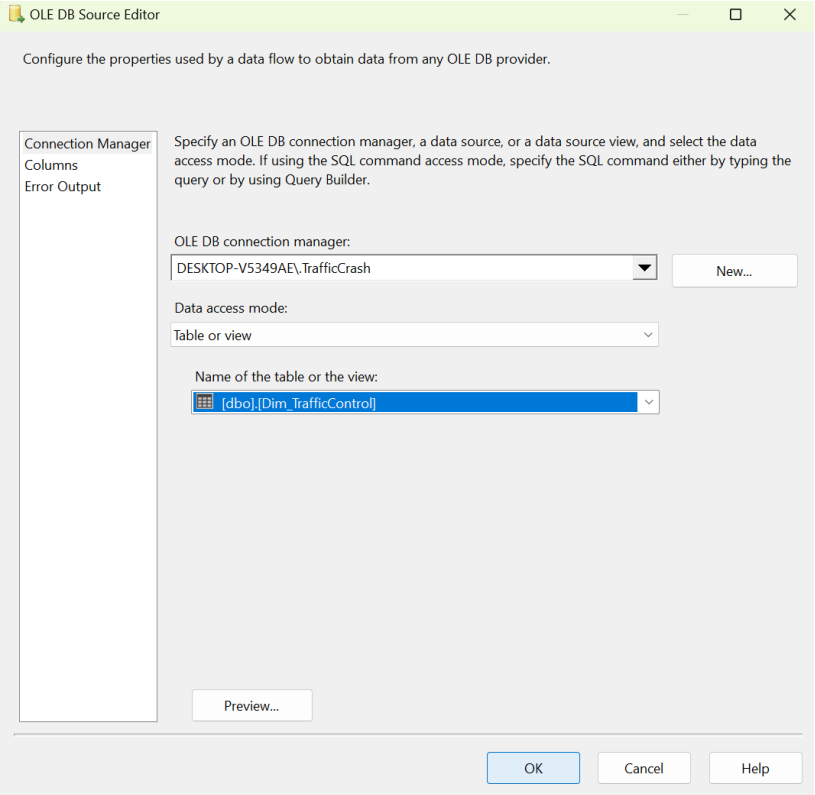
Bước 3: Nhấn chuột phải chọn "Edit", sau đó chọn bảng Fact2 đã tạo trước đó làm data source cho bảng Fact2 mới này.



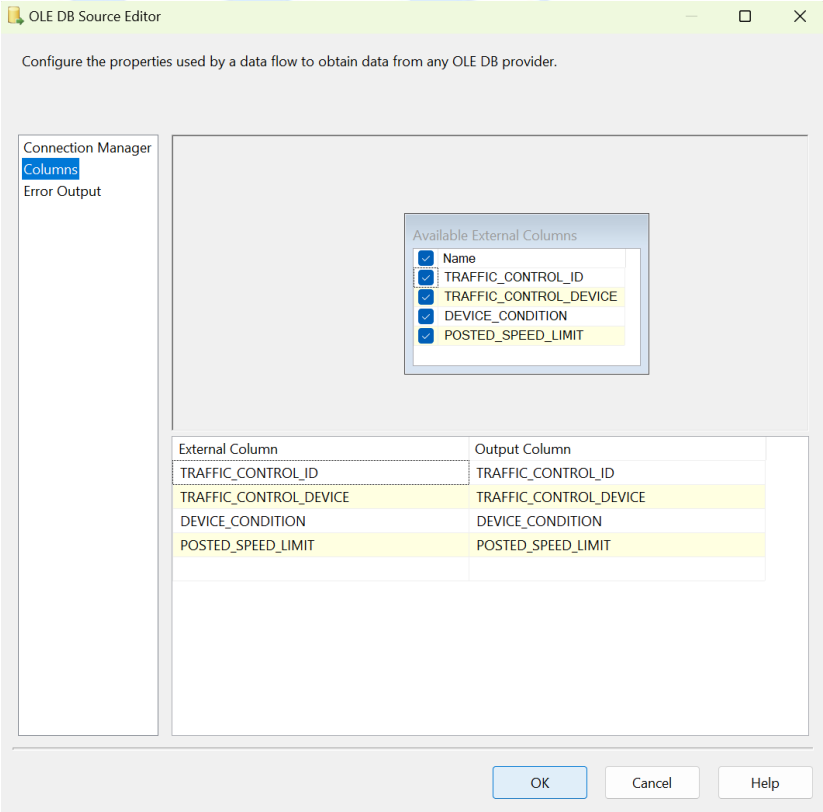
Bước 4: Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



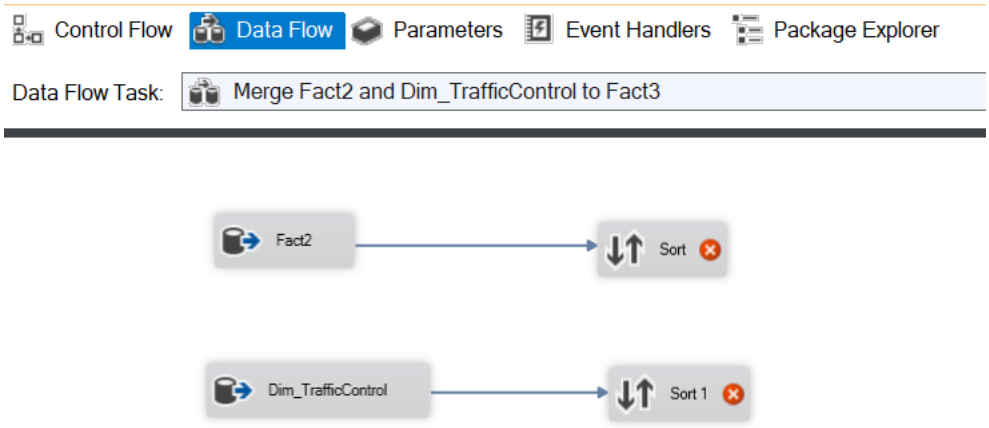
Bước 5: Tương tự thực hiện chọn bảng cho Dim_TrafficControl.



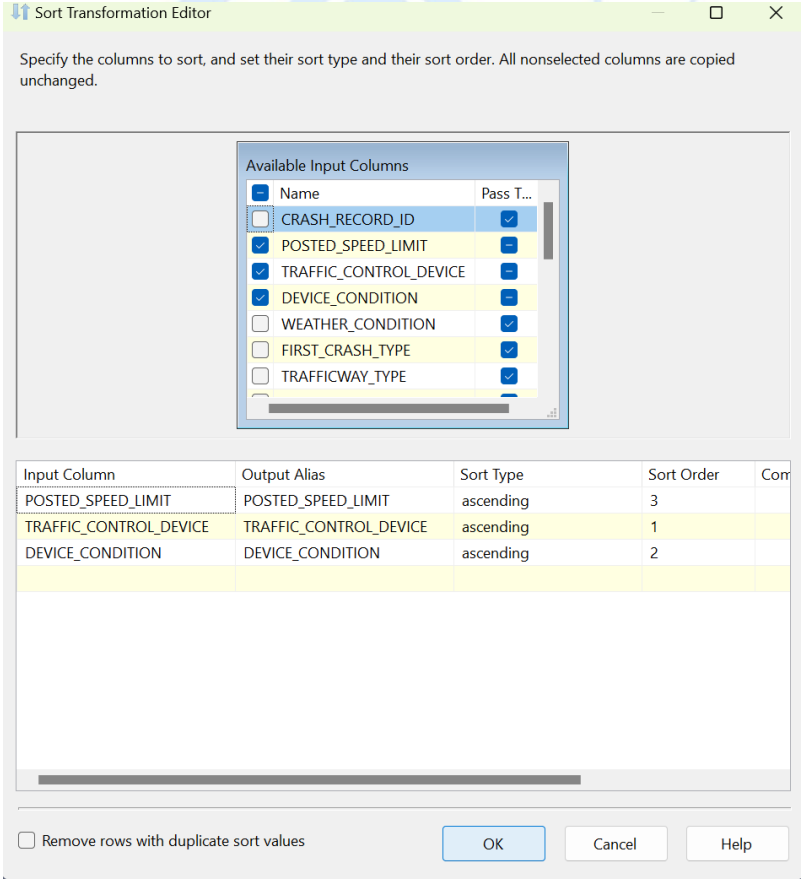
- Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



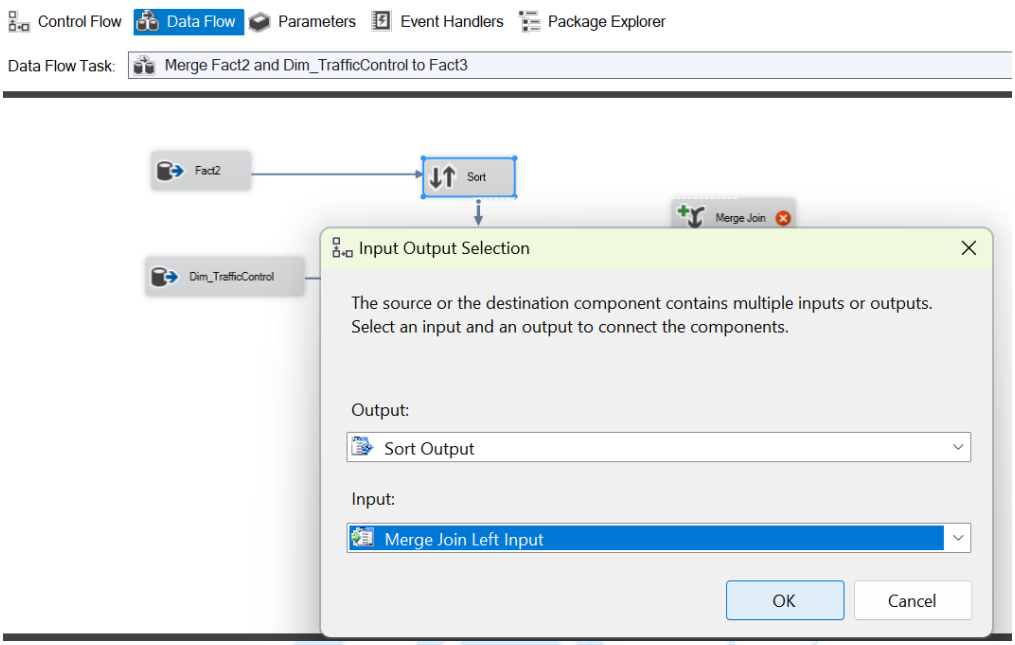
Bước 6: Tạo hai Sort đặt tên là "Sort" và "Sort1" tương ứng với mỗi "OLE DB Source" đã tạo. Thực hiện tạo các kết nối.



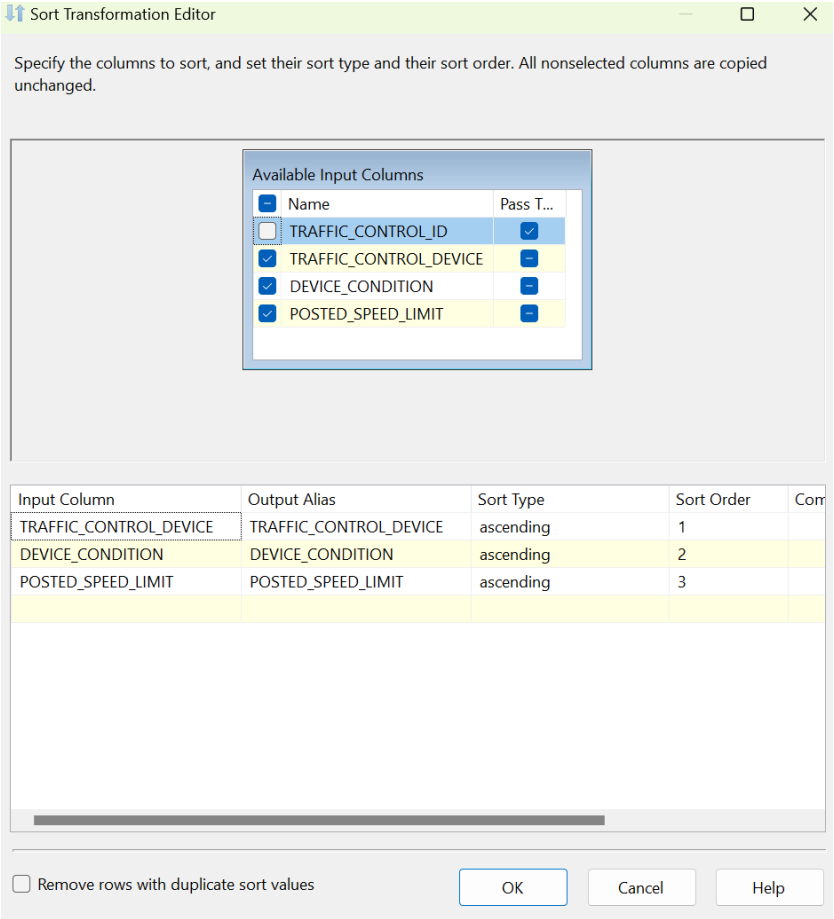
Bước 7: Ở "Sort", nhấn chuột phải chọn "Edit" và chọn cột để chuẩn bị cho quá trình merge. Sau đó nhấn "OK".



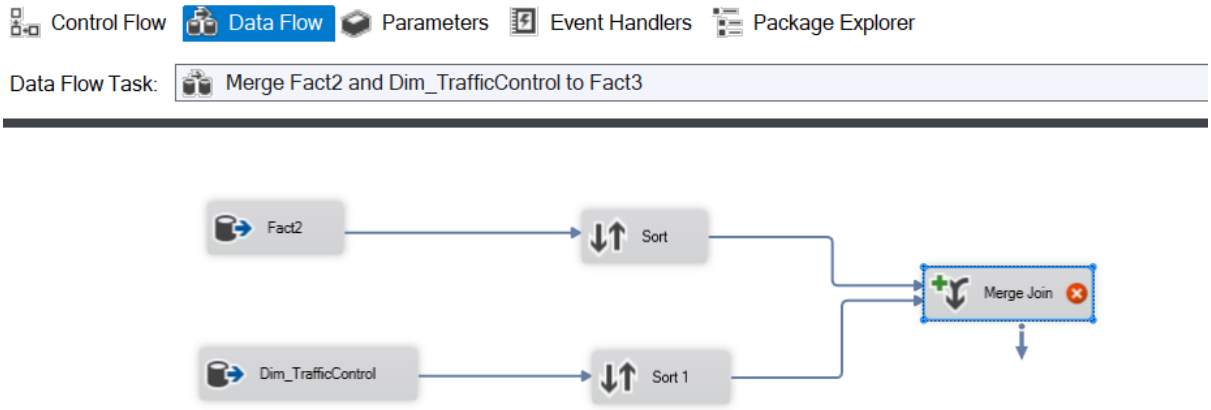
Bước 8: Tạo một Merge Join và nối với "Sort". Tiếp theo ta chọn "Merge Join Left Input" để giữ lại toàn bộ các dòng trong bảng "Fact2" bất kể có kết quả khi thực hiện phép kết trái với cột TRAFFIC_CONTROL_ID của bảng "Dim_TrafficControl" hay không.



Bước 9: Tương tự, ở "Sort1", nhấn chuột phải chọn "Edit" và chọn cột. Sau đó nhấn "OK".

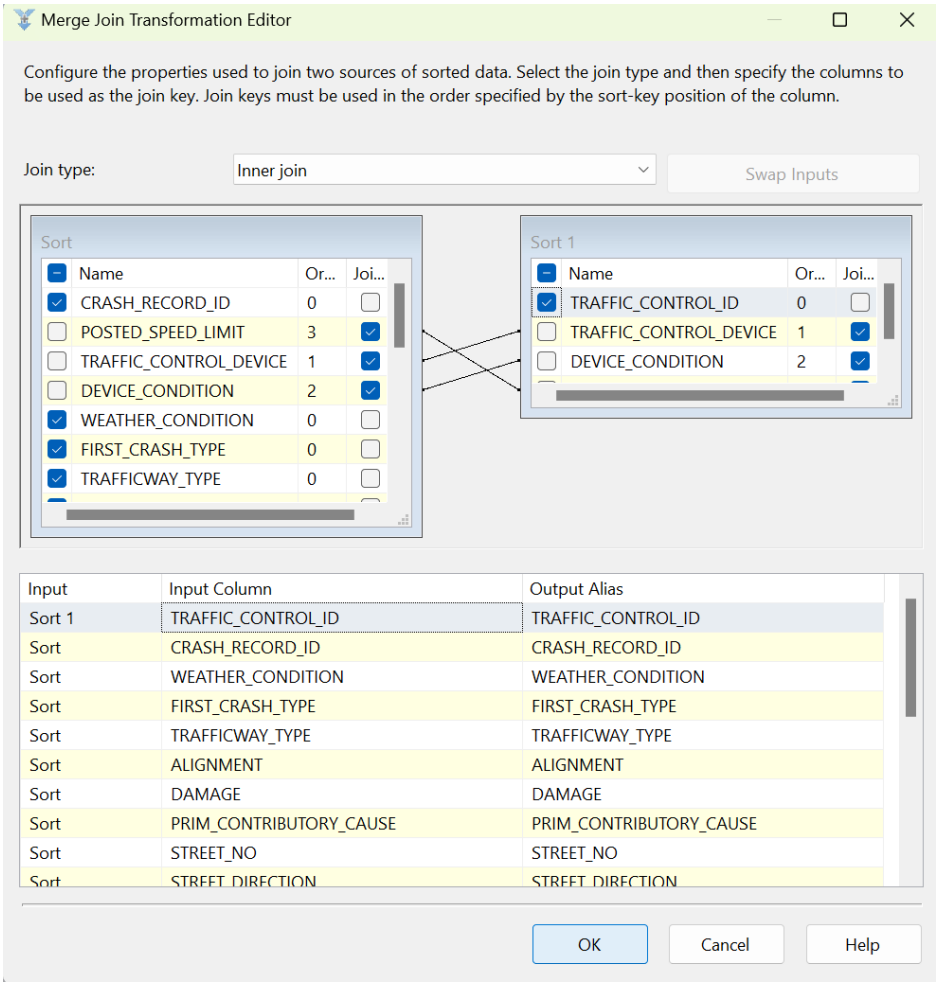


- Kết nối "Sort1" với "Merge Join".



Bước 10: Đầu tiên, nhấn chuột phải vào "Merge Join" và chọn "Edit", một hộp thoại Merge Join Transformation Editor xuất hiện.

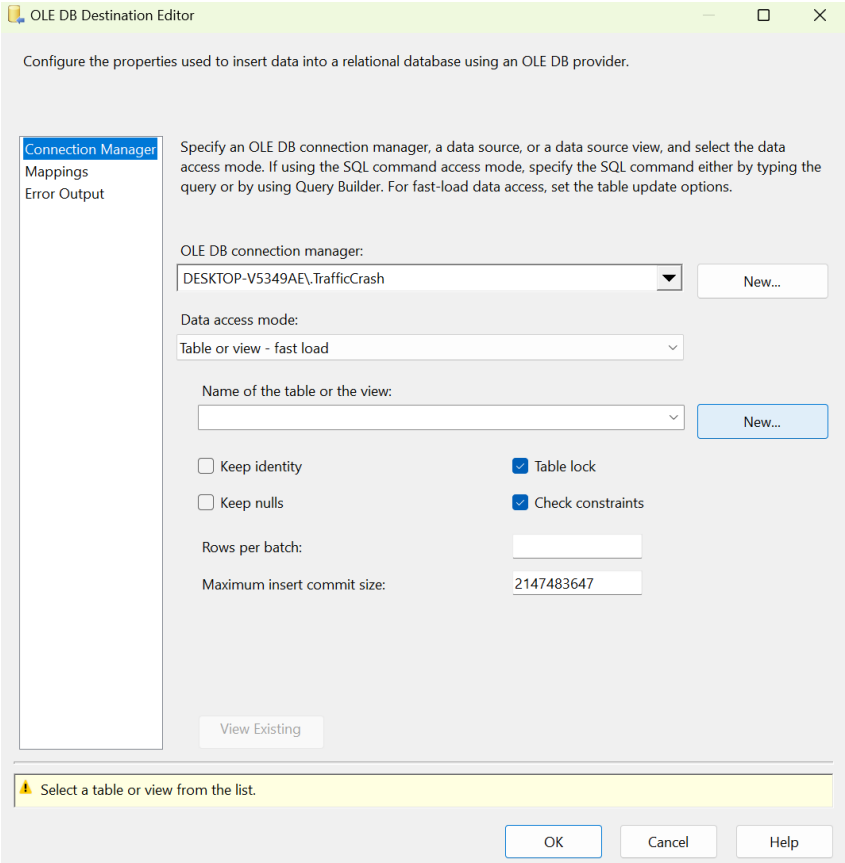
- Sau đó, ta chọn TRAFFIC_CONTROL_ID ở “Sort1” để merge vào “Fact2”.
- Tiếp theo, ta chọn tất cả các cột của “Sort” nhưng không lấy thuộc tính TRAFFIC_CONTROL_DEVICE, DEVICE_CONDITION, POSTED_SPEED_LIMIT.
- Nhấn “OK” để hoàn tất.



Kết quả sau khi merge là bảng Fact2 không còn các thuộc tính TRAFFIC_CONTROL_DEVICE, DEVICE_CONDITION, POSTED_SPEED_LIMIT và có thêm 1 thuộc tính mới là TRAFFIC_CONTROL_ID.

Bước 11: Tạo một OLE DB Destination đặt tên là “Fact3” để chứa tất cả những gì đã merge.

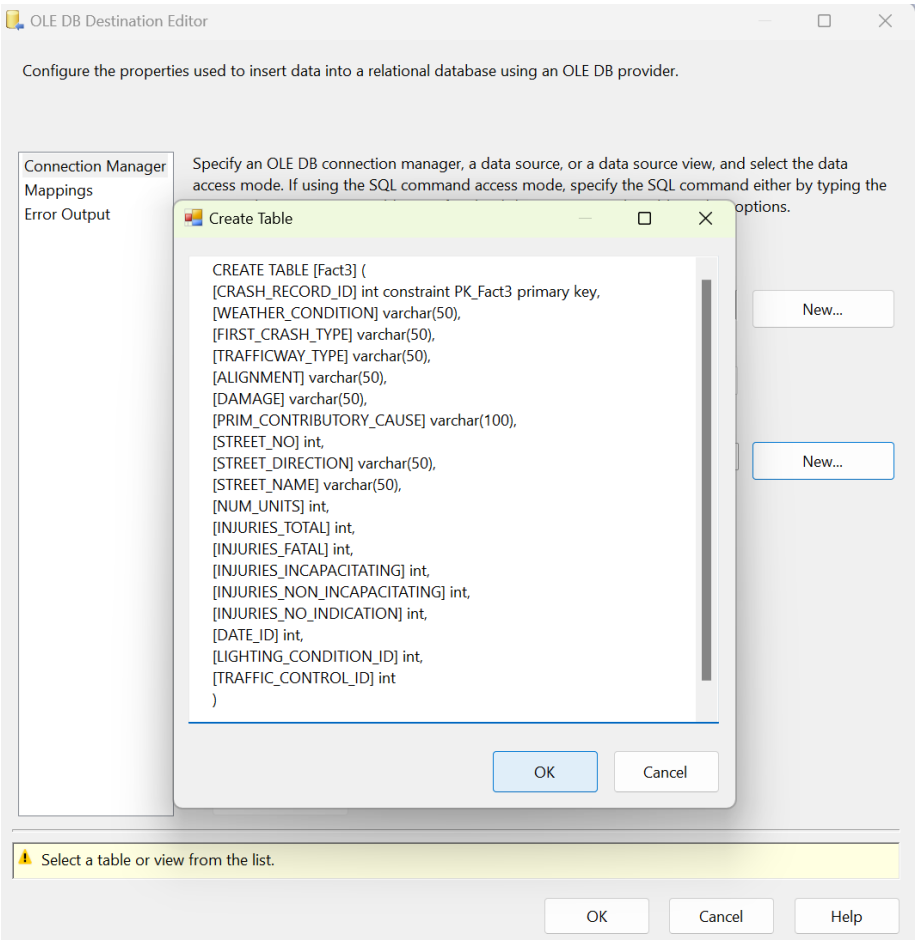
- Nhấn chuột phải vào “Fact3”, chọn “Edit”. Chọn “New” để tạo bảng.



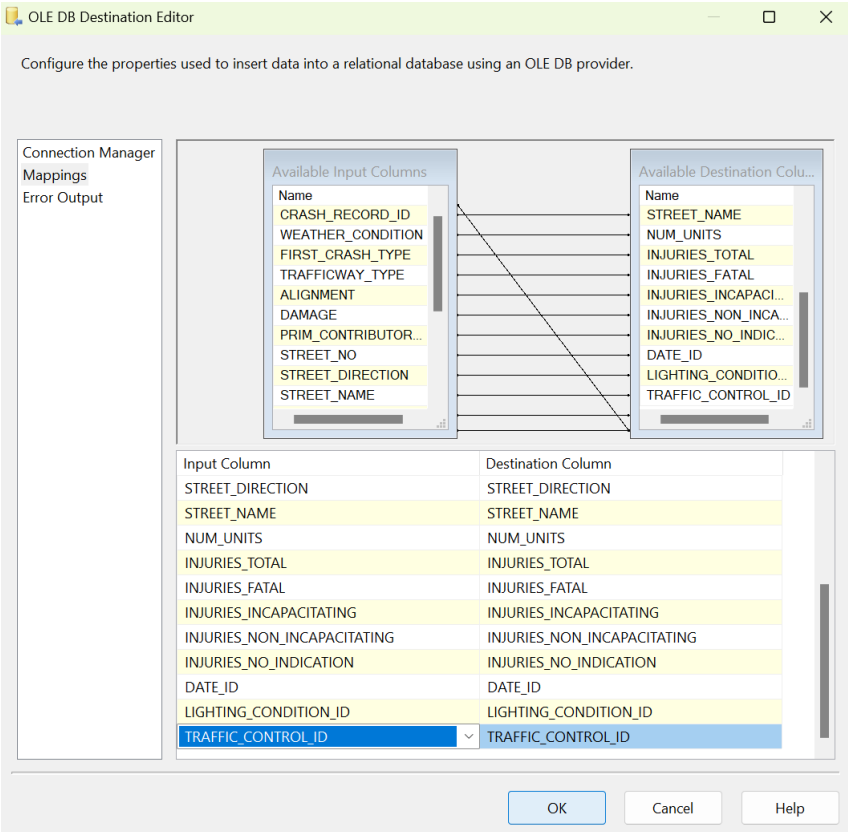
- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Fact3] (
[CRASH_RECORD_ID] int constraint PK_Fact3 primary key,
[WEATHER_CONDITION] varchar(50),
[FIRST_CRASH_TYPE] varchar(50),
[TRAFFICWAY_TYPE] varchar(50),
[ALIGNMENT] varchar(50),
[DAMAGE] varchar(50),
[PRIM_CONTRIBUTORY_CAUSE] varchar(100),
[STREET_NO] int,
[STREET_DIRECTION] varchar(50),
[STREET_NAME] varchar(50),
[NUM_UNITS] int,
[INJURIES_TOTAL] int,
[INJURIES_FATAL] int,
[INJURIES_INCAPACITATING] int,
[INJURIES_NON_INCAPACITATING] int,
[INJURIES_NO_INDICATION] int,
```

```
[DATE_ID] int,
[LIGHTING_CONDITION_ID] int,
[TRAFFIC_CONTROL_ID] int
)
```



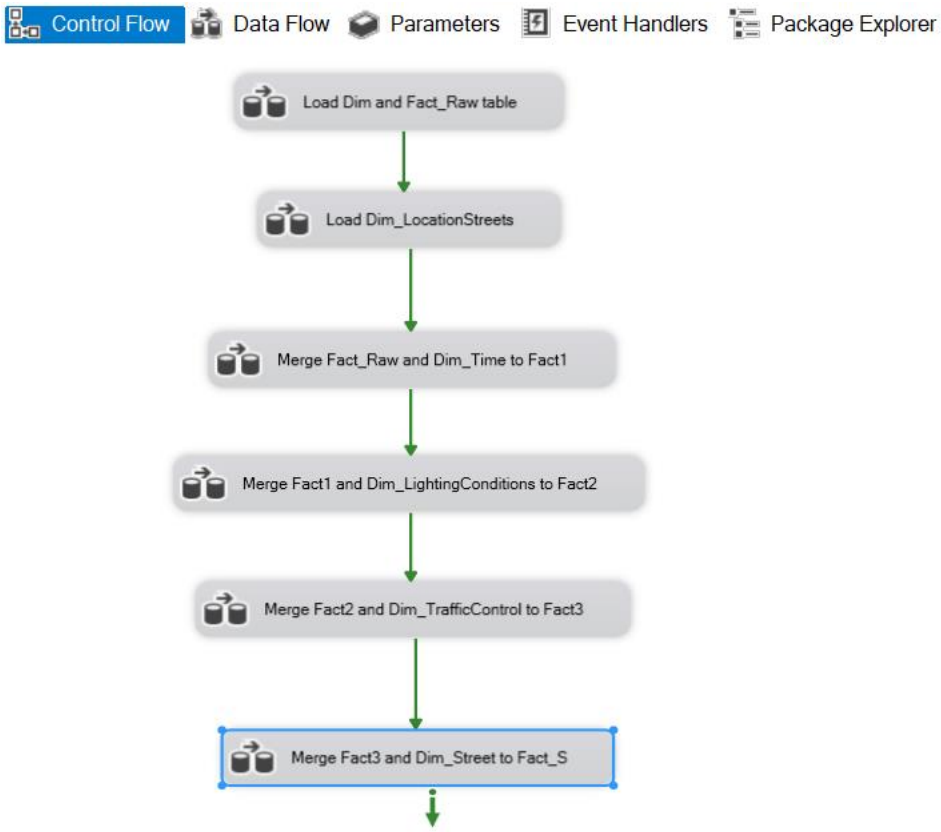
- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



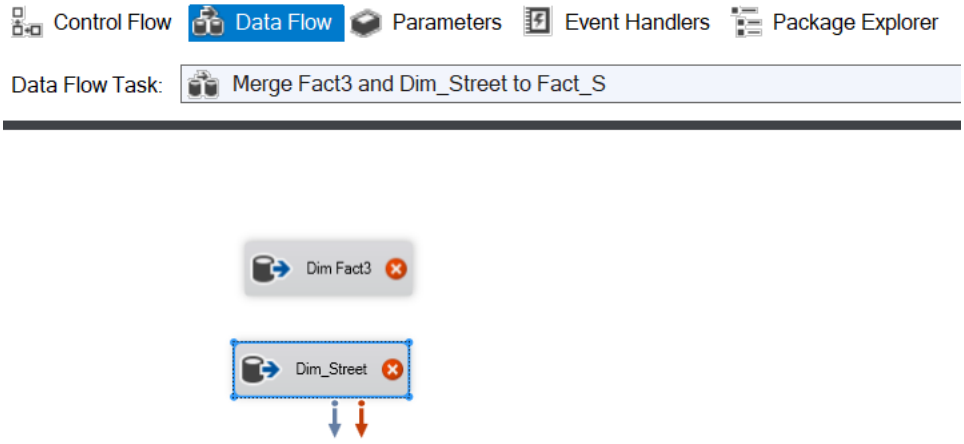
Tiếp tục quá trình merge bảng Fact3 với các Dimension còn lại để có 1 bảng Fact_TrafficCrashRecords hoàn chỉnh.

2.3.13.4 Merge Fact3 và Dim_Street vào Fact_S.

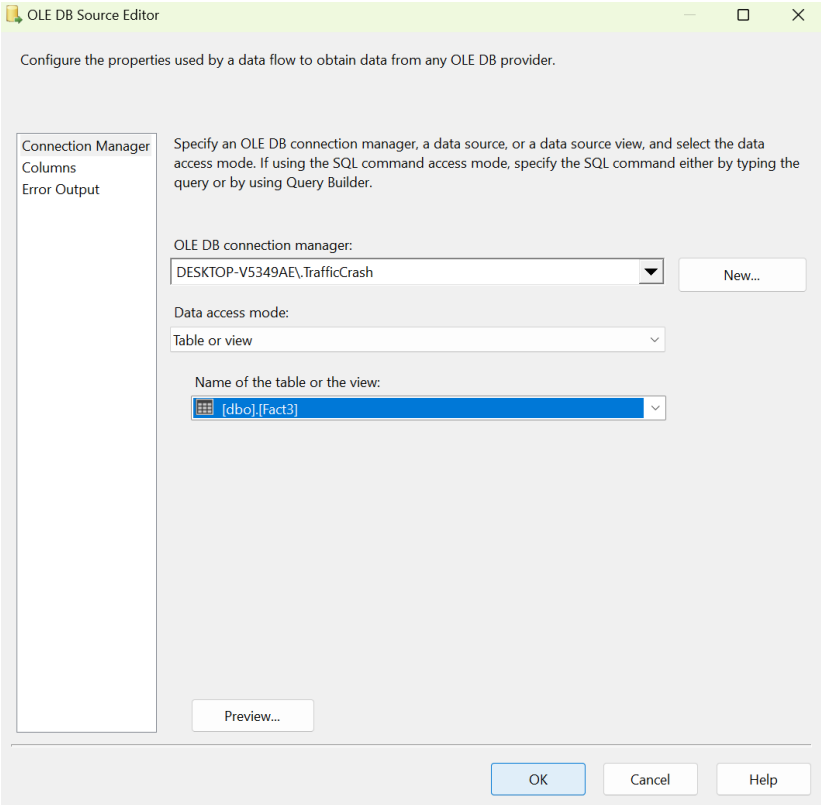
Bước 1: Ở tab Control Flow, tạo một Data Flow Task và đặt tên là "Merge Fact3 and Dim_Street to Fact_S".



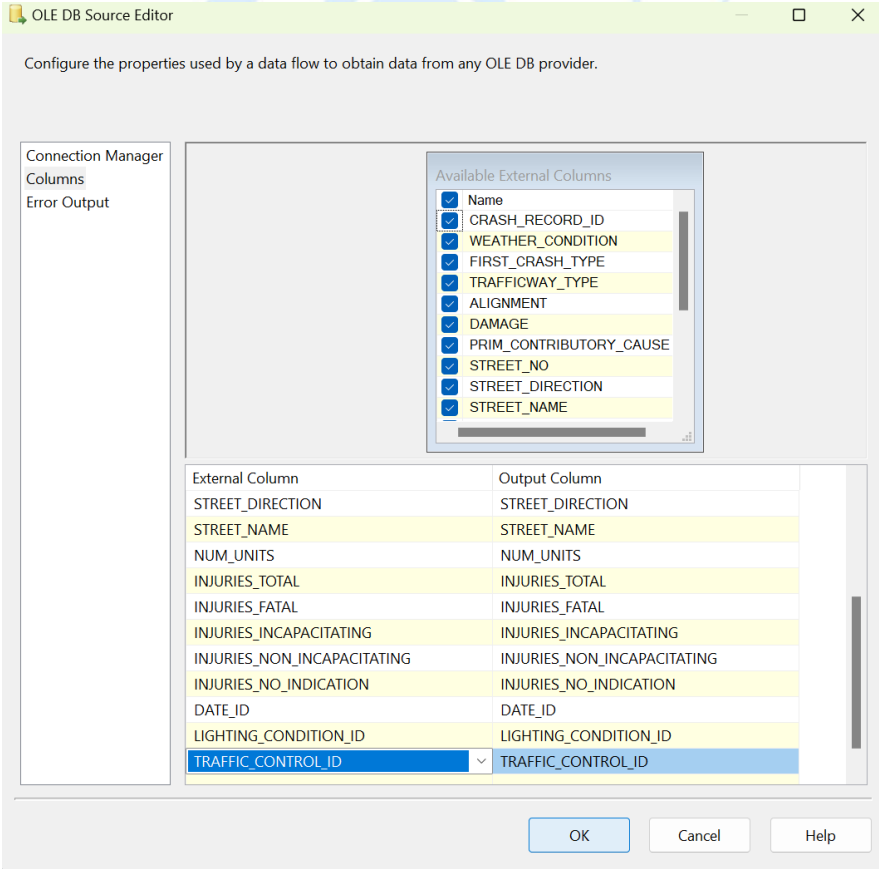
Bước 2: Nhấn chuột phải vào "Merge Fact3 and Dim_Street to Fact_S" và chọn "Edit", trong tab Data Flow ta tạo hai **OLE DB Source** và đặt tên là "Fact3" và "Dim_Street".



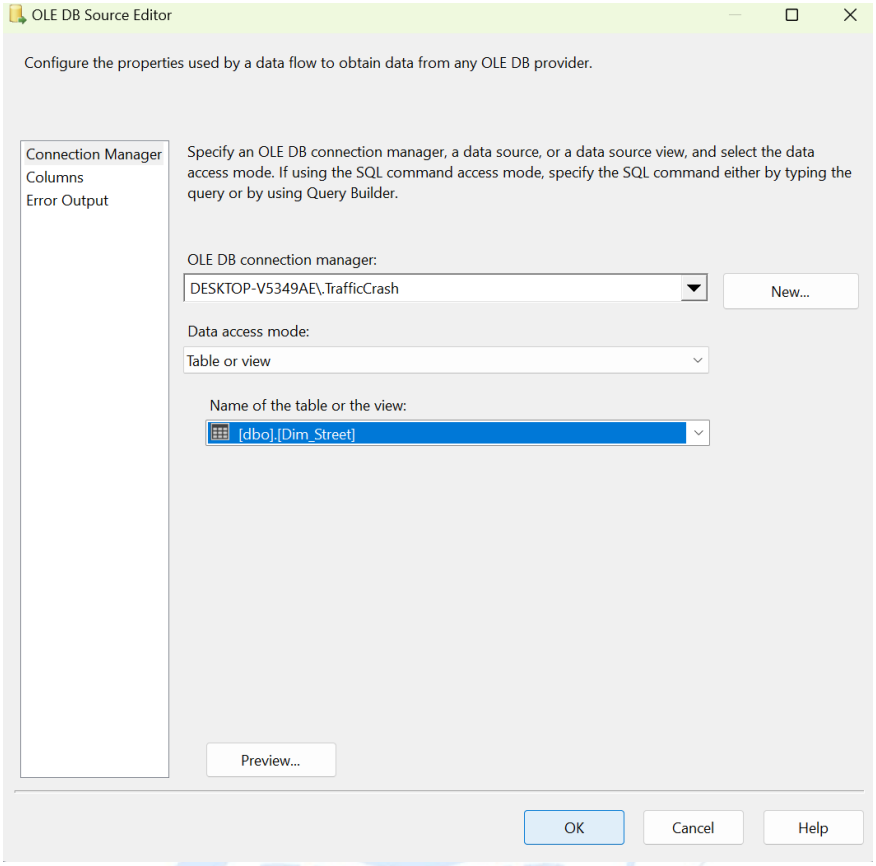
Bước 3: Nhấn chuột phải chọn "Edit", sau đó chọn bảng Fact3 đã tạo trước đó làm data source cho bảng Fact3 mới này.



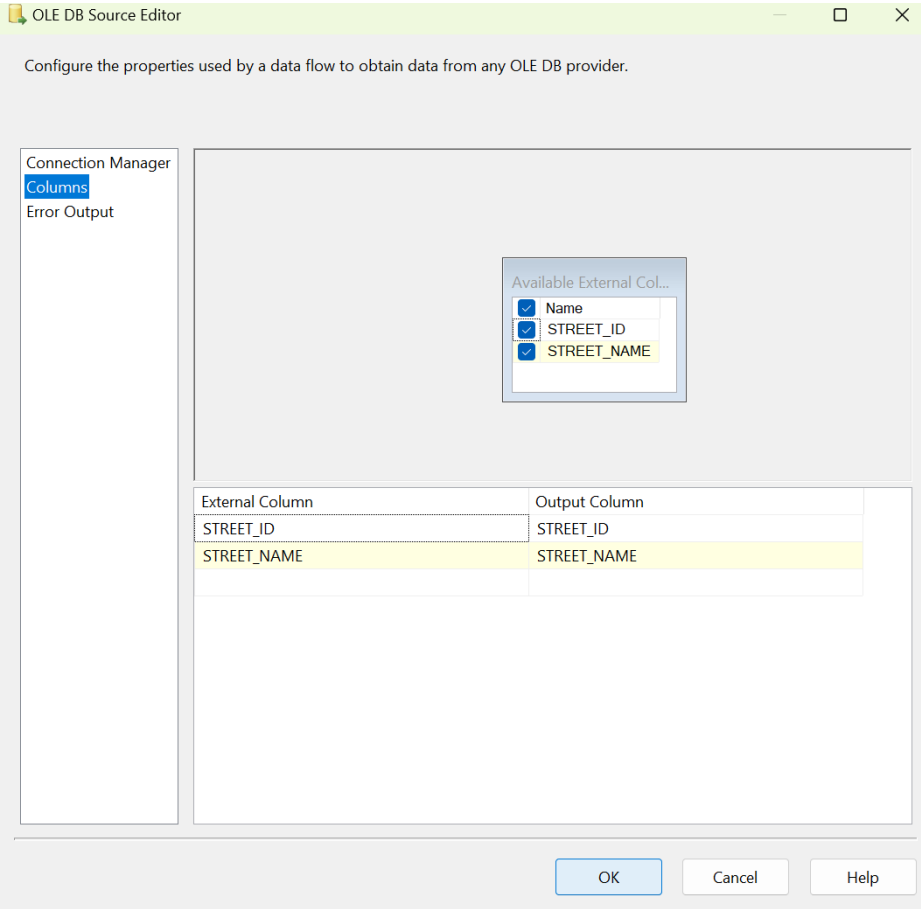
Bước 4: Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



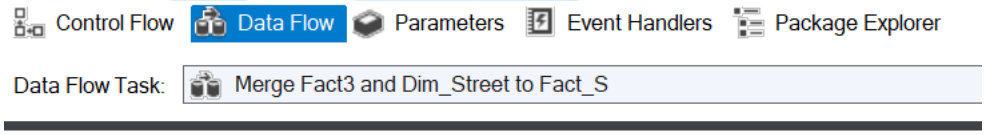
Bước 5: Tương tự thực hiện chọn bảng cho Dim_Street.



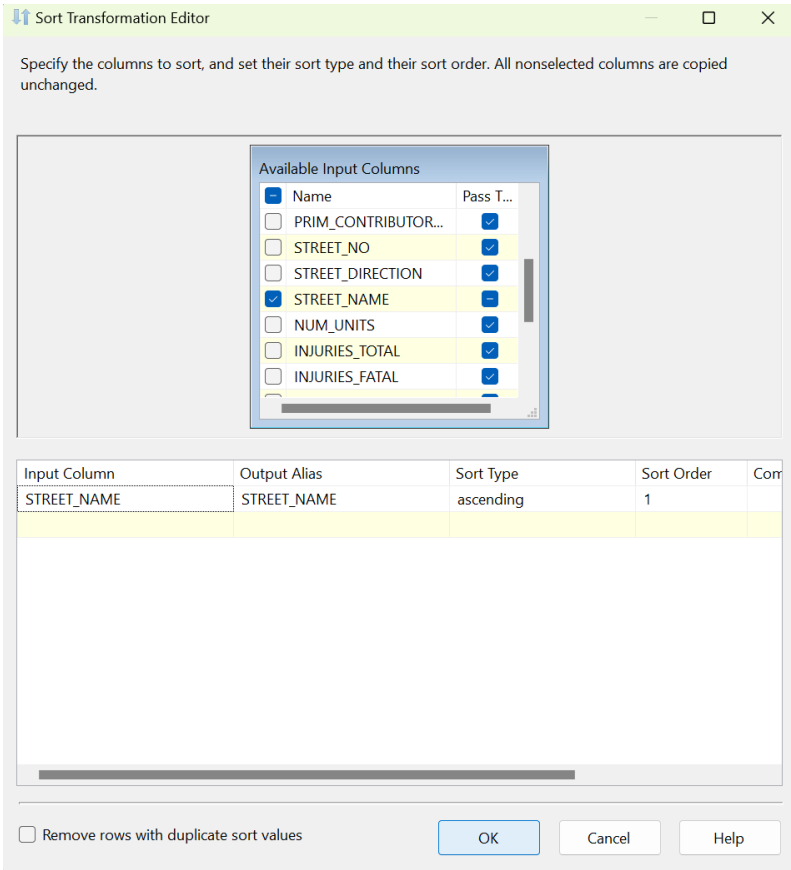
- Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



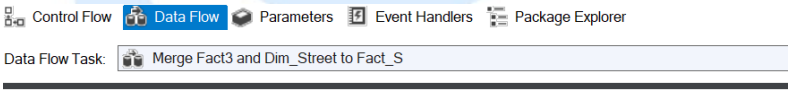
Bước 6: Tạo hai Sort đặt tên là "Sort" và "Sort1" tương ứng với mỗi "OLE DB Source" đã tạo. Thực hiện tạo các kết nối.

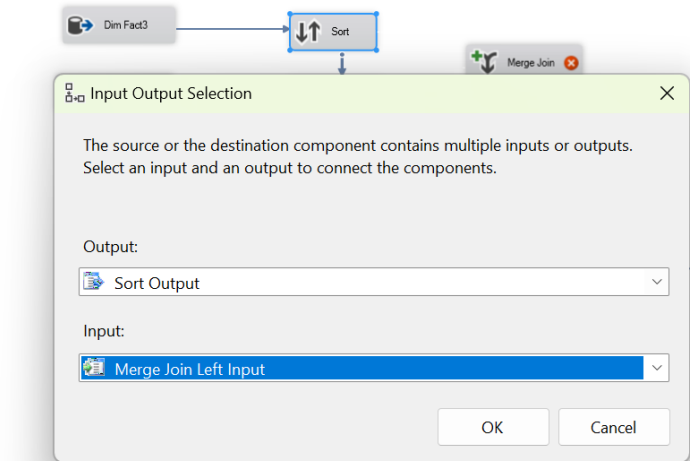


Bước 7: Ở "Sort", nhấn chuột phải chọn "Edit" và chọn cột để chuẩn bị cho quá trình merge. Sau đó nhấn "OK".

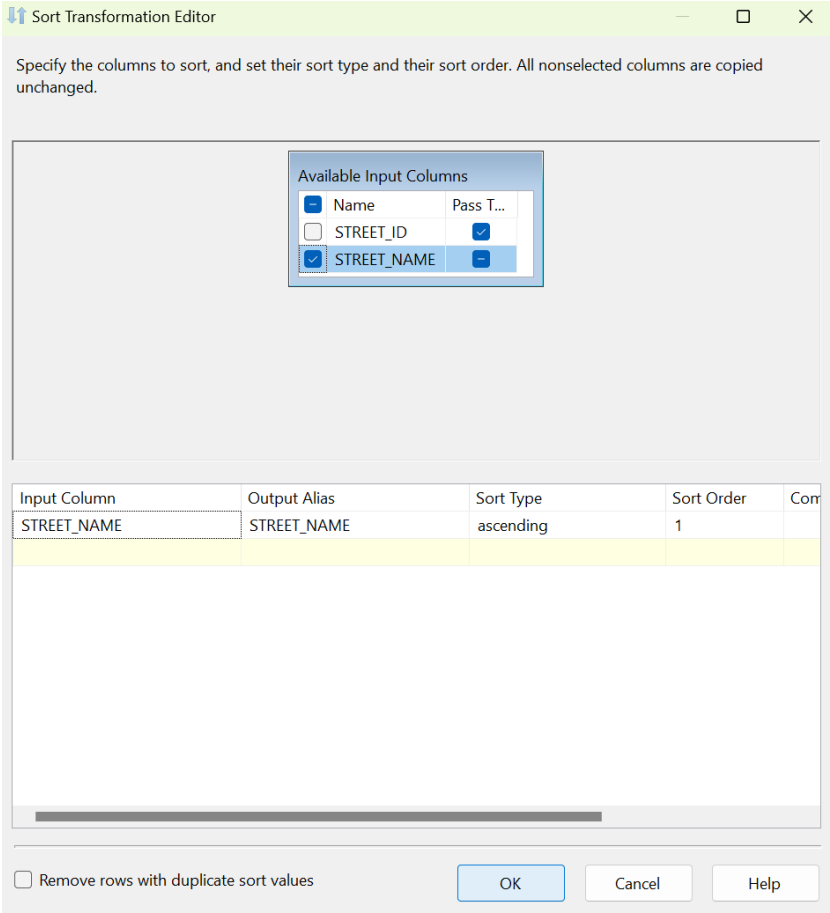


Bước 8: Tạo một Merge Join và nối với "Sort". Tiếp theo ta chọn "Merge Join Left Input" để giữ lại toàn bộ các dòng trong bảng "Fact3" bất kể có kết quả khi thực hiện phép kết trái với cột STREET_ID của bảng "Dim_Street" hay không.

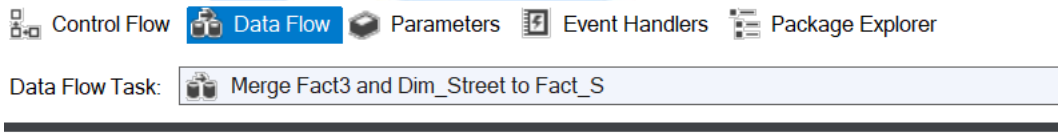


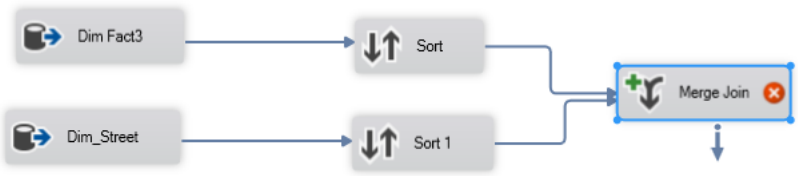


Bước 9: Tương tự, ở "Sort1", nhấn chuột phải chọn "Edit" và chọn cột. Sau đó nhấn "OK".



- Kết nối "Sort1" với "Merge Join".

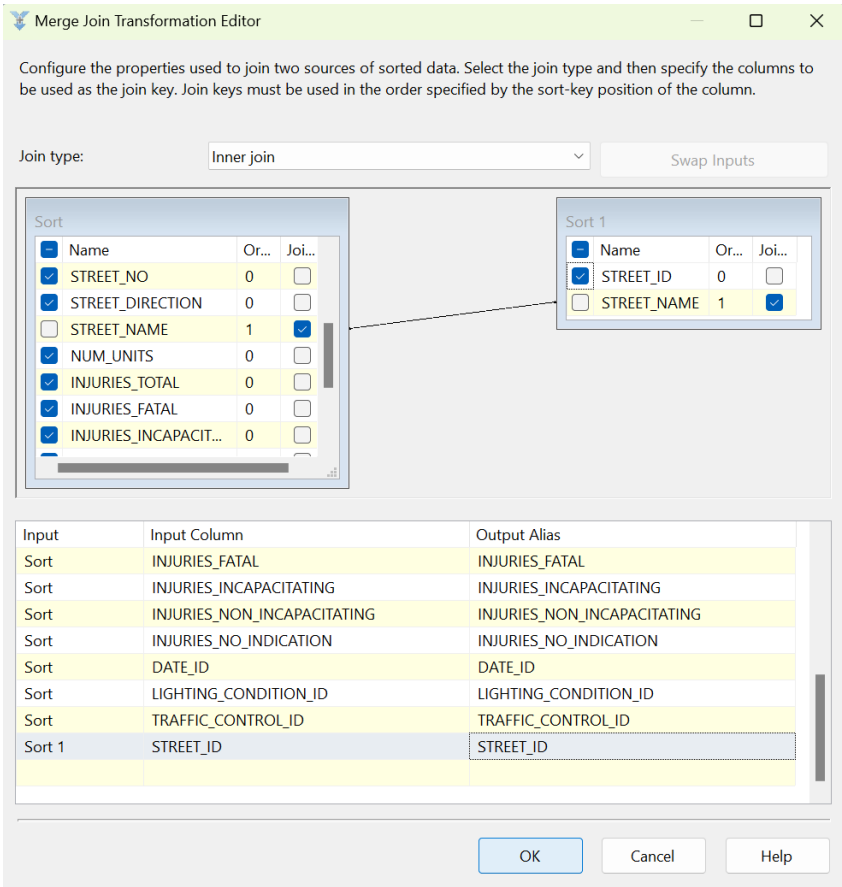




Bước 10: Đầu tiên, nhấn chuột phải vào "Merge Join" và chọn "Edit", một hộp thoại Merge Join Transformation Editor xuất hiện.

- Sau đó, ta chọn STREET_ID ở "Sort1" để merge vào "Fact3".

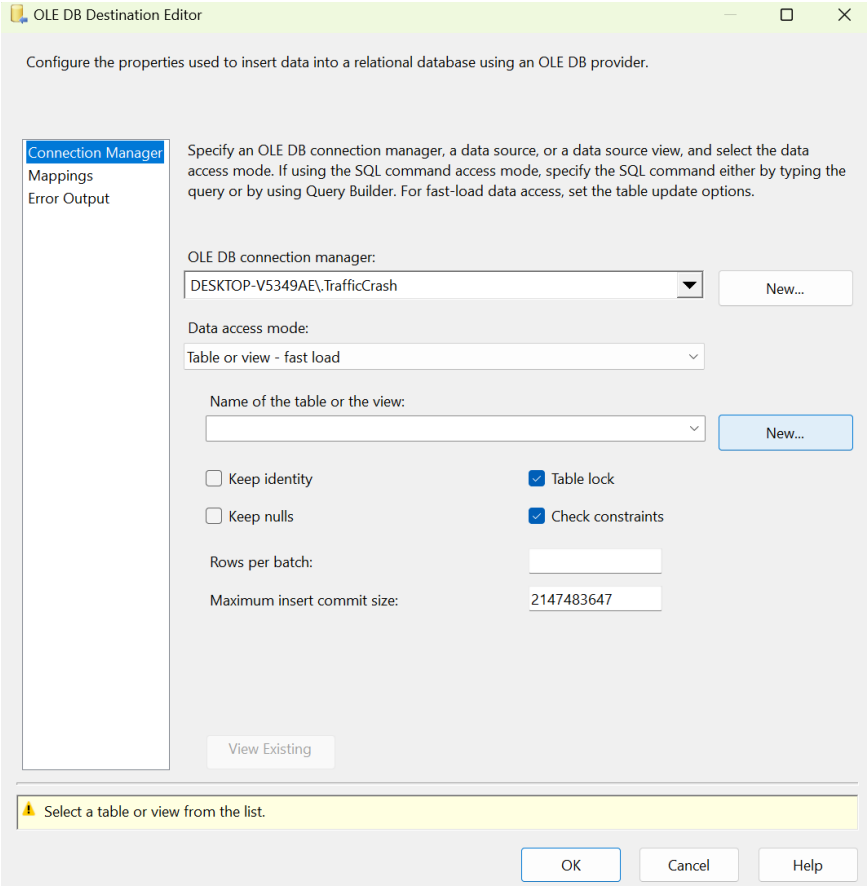
- Tiếp theo, ta chọn tất cả các cột của “Sort” nhưng không lấy thuộc tính STREET_NAME.
- Nhấn “OK” để hoàn tất.



Kết quả sau khi merge là bảng Fact3 không còn thuộc tính STREET_NAME và có thêm 1 thuộc tính mới là STREET_ID.

Bước 11: Tạo một OLE DB Destination đặt tên là “Fact_S” để chứa tất cả những gì đã merge.

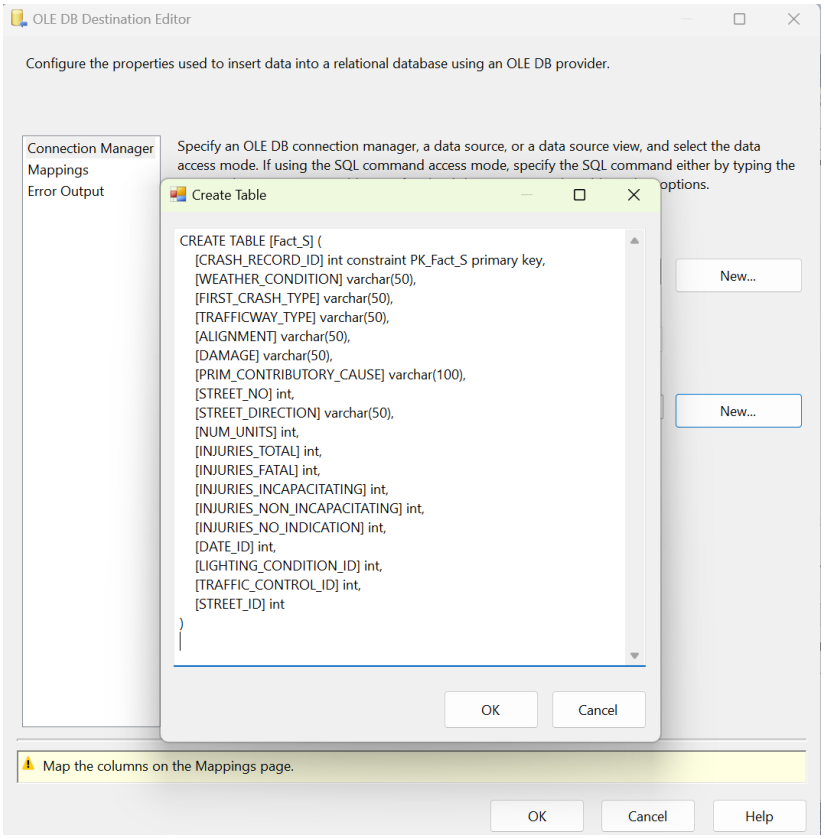
- Nhấn chuột phải vào “Fact_S”, chọn “Edit”. Chọn “New” để tạo bảng.



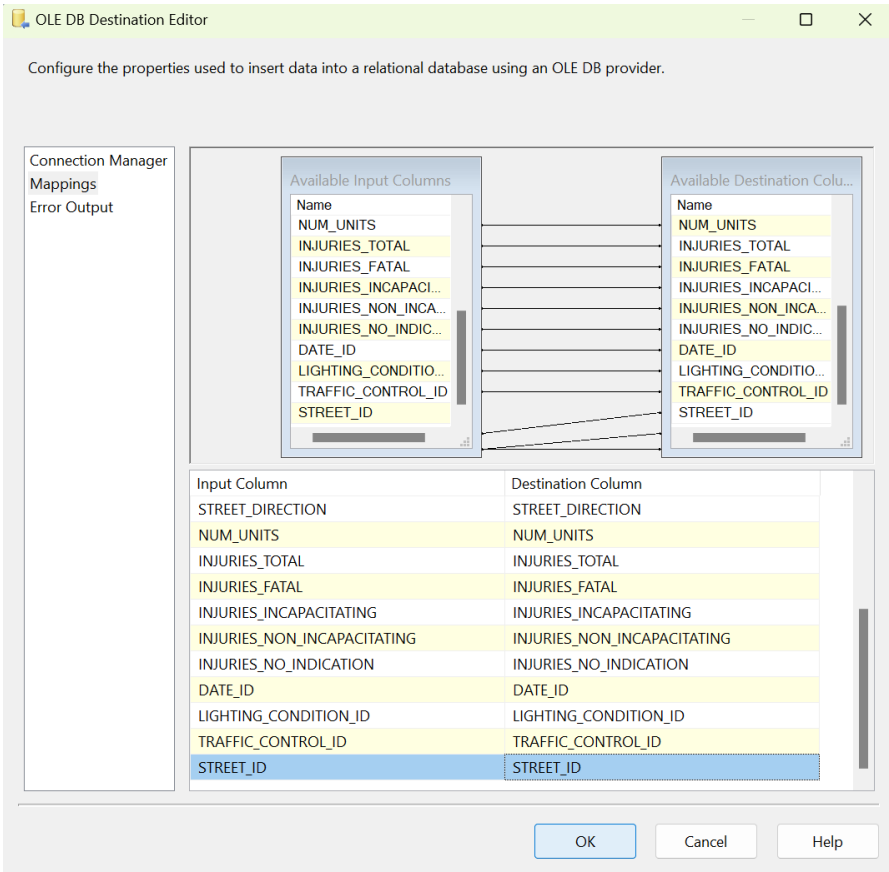
- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Fact_S] (
    [CRASH_RECORD_ID] int constraint PK_Fact_S primary key,
    [WEATHER_CONDITION] varchar(50),
    [FIRST_CRASH_TYPE] varchar(50),
    [TRAFFICWAY_TYPE] varchar(50),
    [ALIGNMENT] varchar(50),
    [DAMAGE] varchar(50),
    [PRIM_CONTRIBUTORY_CAUSE] varchar(100),
    [STREET_NO] int,
    [STREET_DIRECTION] varchar(50),
    [NUM_UNITS] int,
    [INJURIES_TOTAL] int,
    [INJURIES_FATAL] int,
```

```
    [INJURIES_INCAPACITATING] int,
    [INJURIES_NON_INCAPACITATING] int,
    [INJURIES_NO_INDICATION] int,
    [DATE_ID] int,
    [LIGHTING_CONDITION_ID] int,
    [TRAFFIC_CONTROL_ID] int,
    [STREET_ID] int
)
```



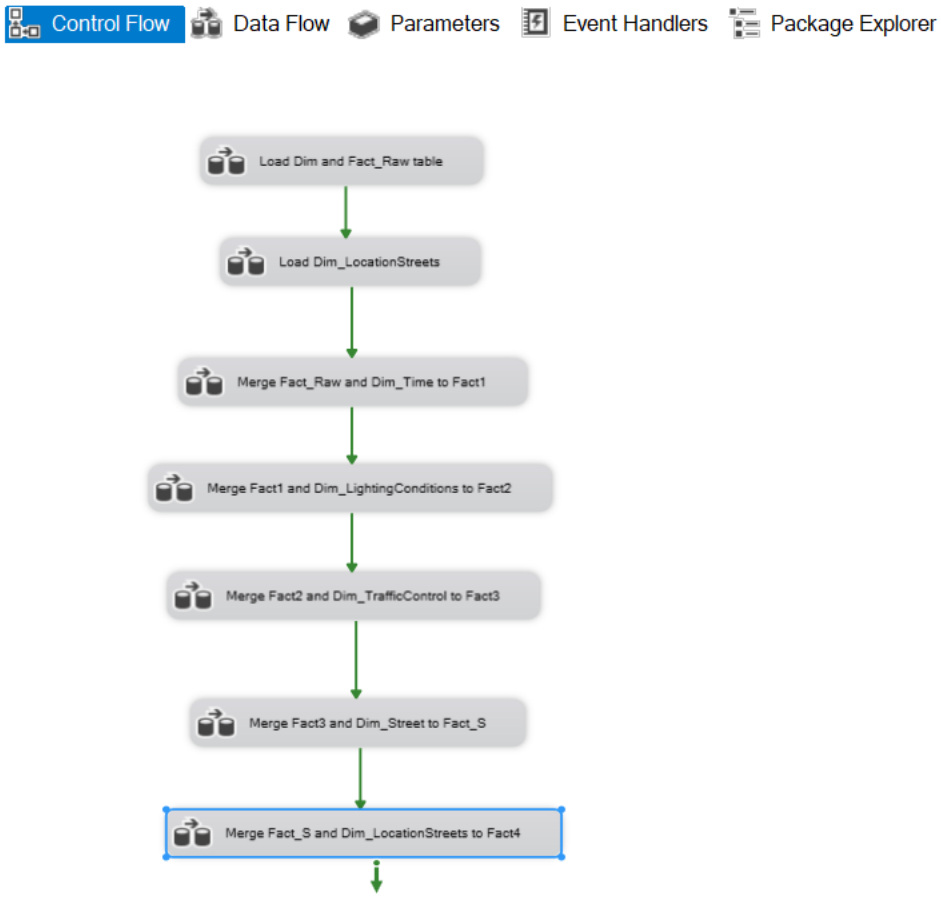
- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



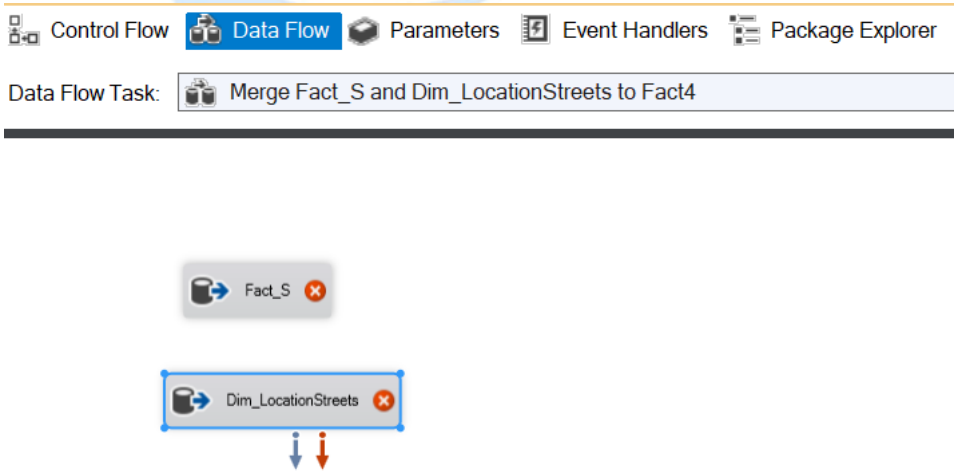
Tiếp tục quá trình merge bảng Fact_S với các Dimension còn lại để có 1 bảng Fact_TrafficCrashRecords hoàn chỉnh.

2.3.13.5 Merge Fact_S và Dim_LocationStreets vào Fact4.

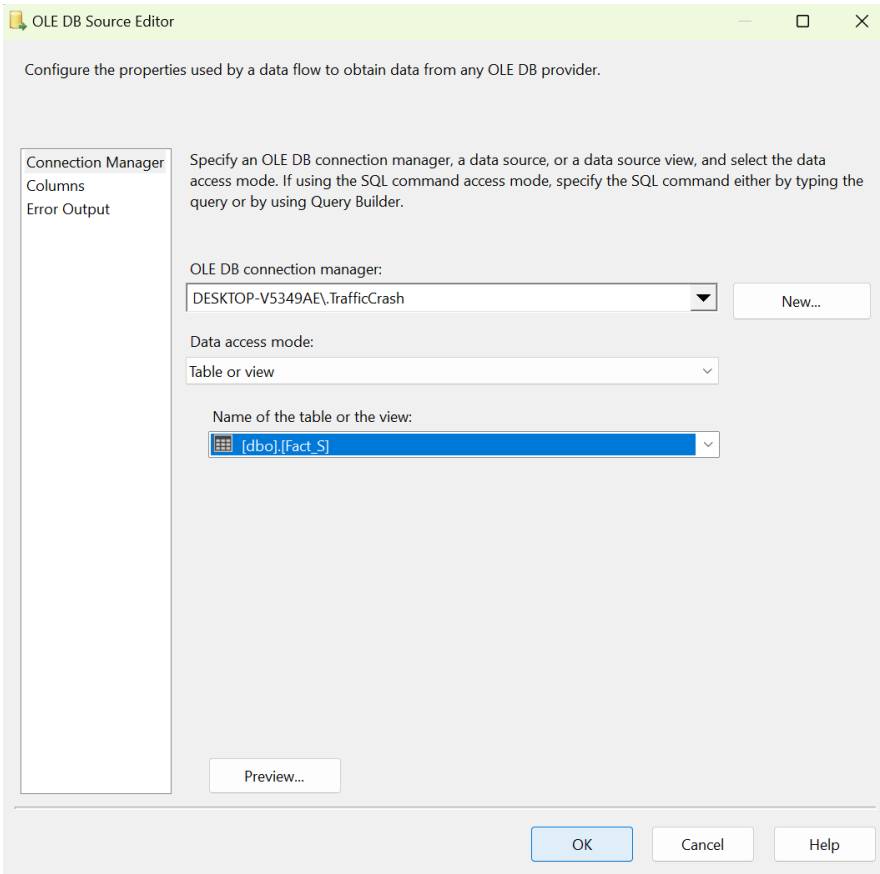
Bước 1: Ở tab Control Flow, tạo một Data Flow Task và đặt tên là “Merge Fact_S and Dim_LocationStreets to Fact4”.



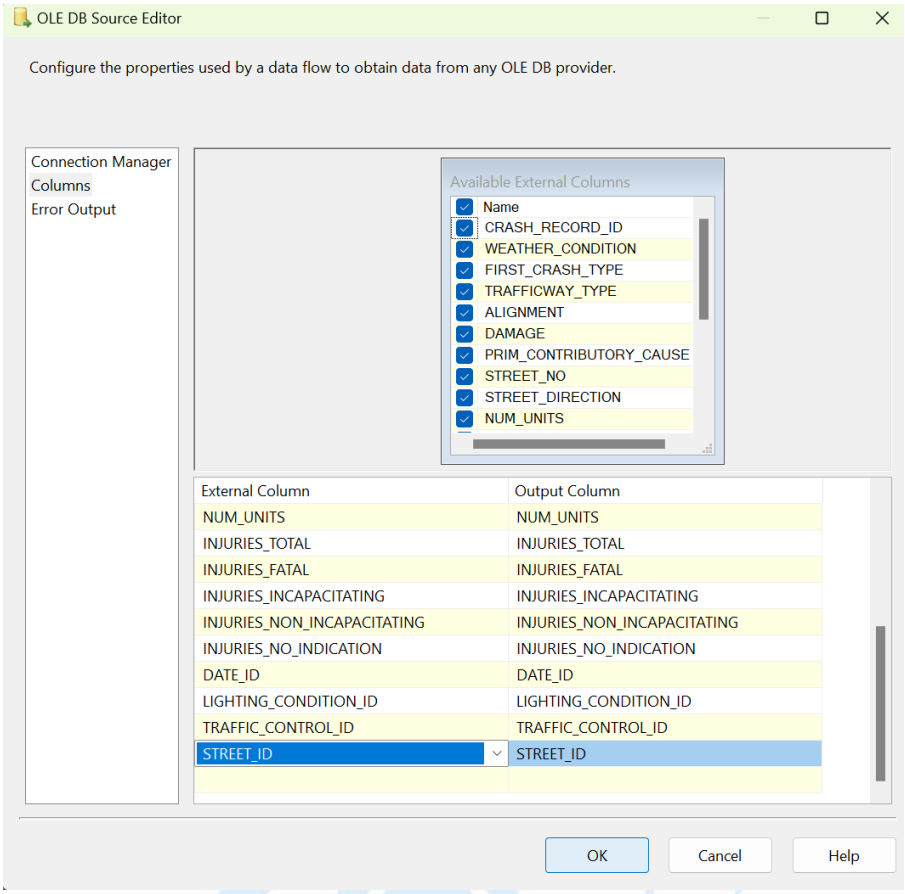
Bước 2: Nhấn chuột phải vào "Merge Fact_S and Dim_LocationStreets to Fact4" và chọn "Edit", trong tab Data Flow ta tạo hai **OLE DB Source** và đặt tên là "Fact_S" và "Dim_LocationStreets".



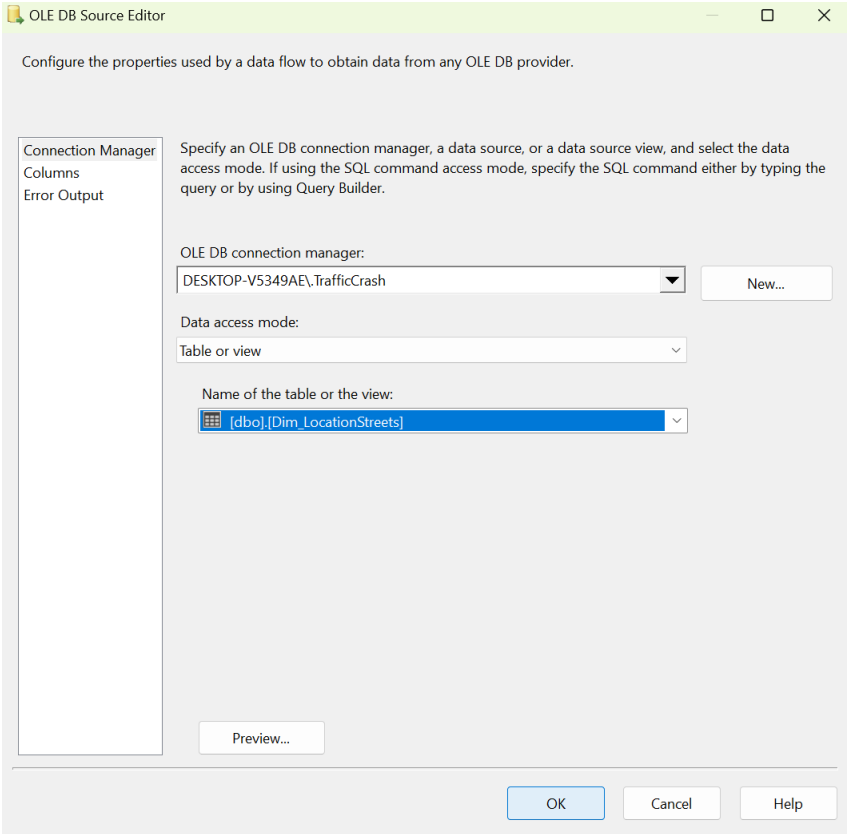
Bước 3: Nhấn chuột phải chọn "Edit", sau đó chọn bảng Fact_S đã tạo trước đó làm data source cho bảng Fact_S mới này.



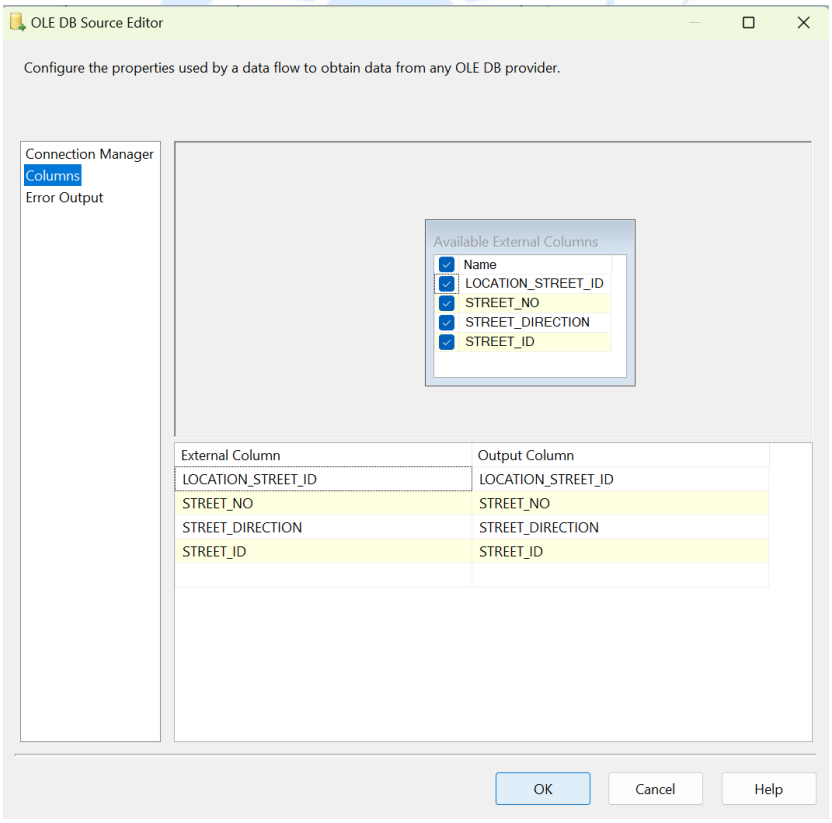
Bước 4: Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



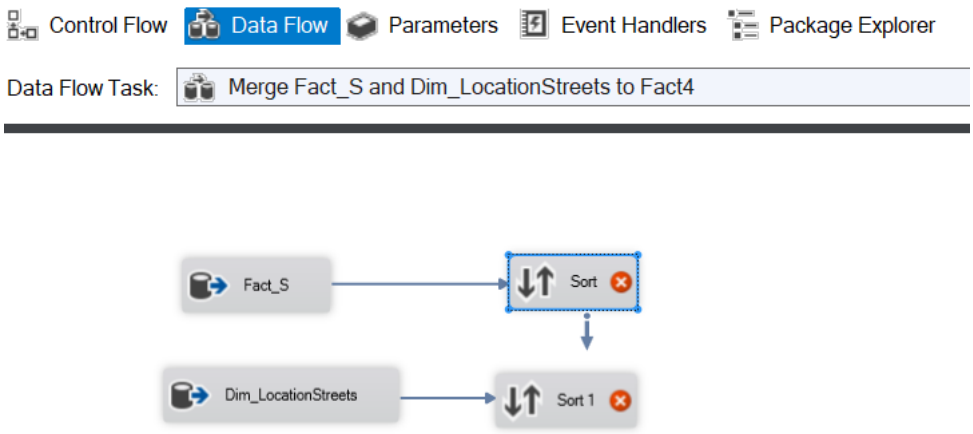
Bước 5: Tương tự thực hiện chọn bảng cho Dim_LocationStreets.



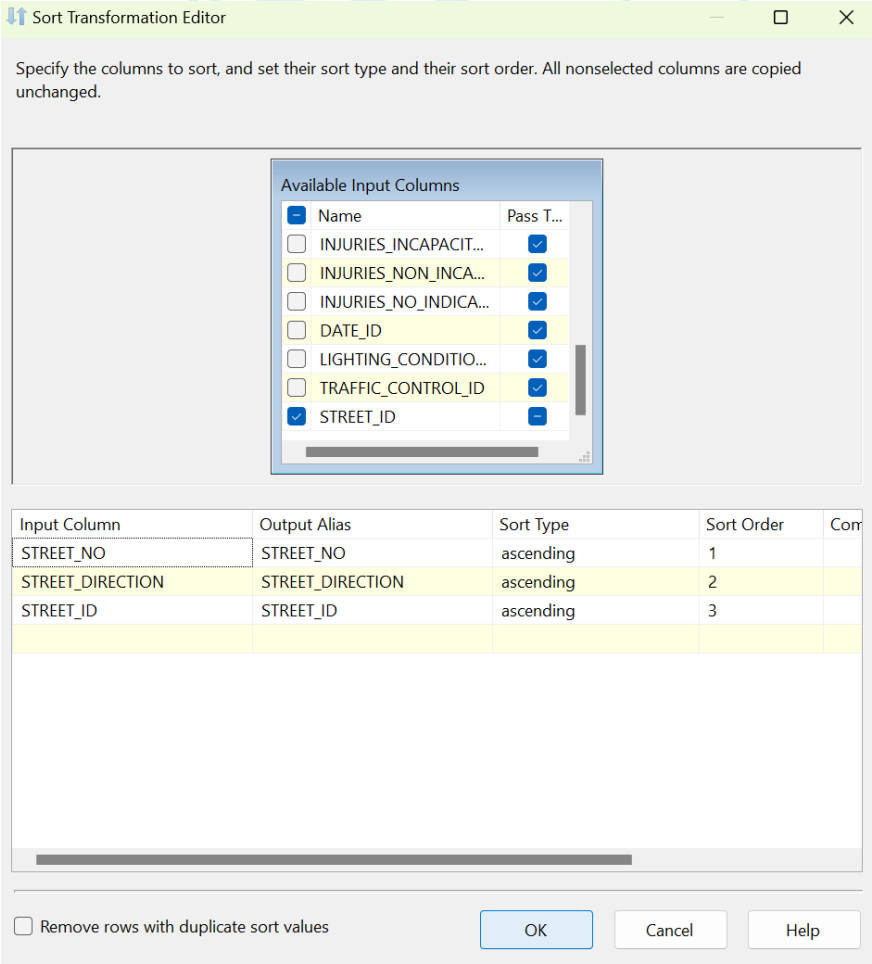
- Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



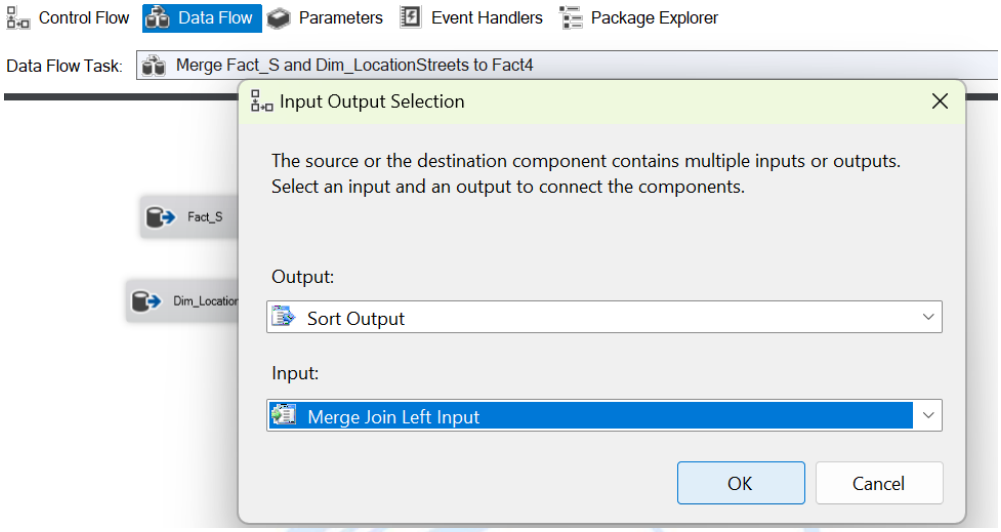
Bước 6: Tạo hai Sort đặt tên là "Sort" và "Sort1" tương ứng với mỗi "OLE DB Source" đã tạo. Thực hiện tạo các kết nối.



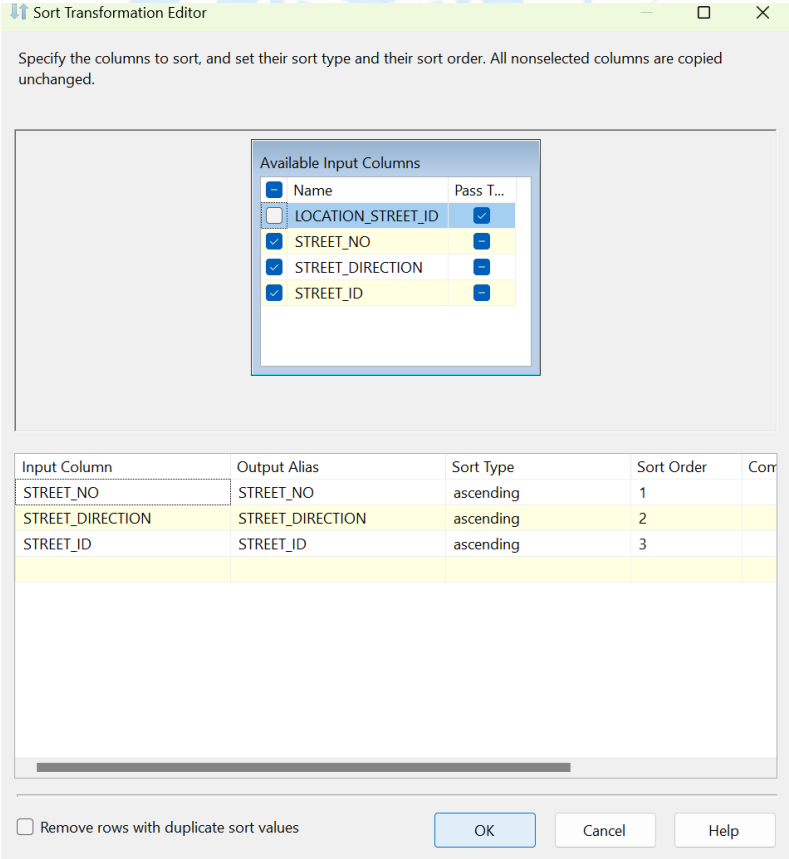
Bước 7: Ở "Sort", nhấn chuột phải chọn "Edit" và chọn cột để chuẩn bị cho quá trình merge. Sau đó nhấn "OK".



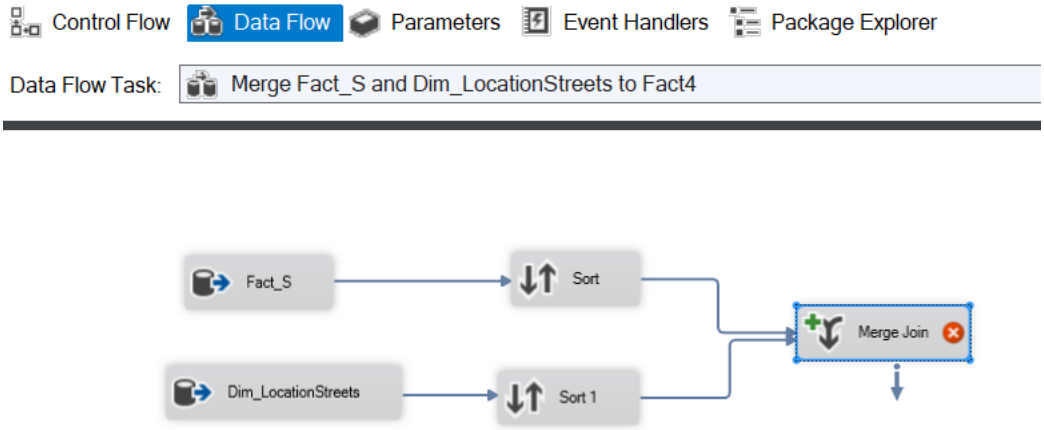
Bước 8: Tạo một Merge Join và nối với "Sort". Tiếp theo ta chọn "Merge Join Left Input" để giữ lại toàn bộ các dòng trong bảng "Fact2" bất kể có kết quả khi thực hiện phép kết trái với cột ID của bảng "Dim_TrafficControl" hay không.



Bước 9: Tương tự, ở "Sort1", nhấn chuột phải chọn "Edit" và chọn cột. Sau đó nhấn "OK".

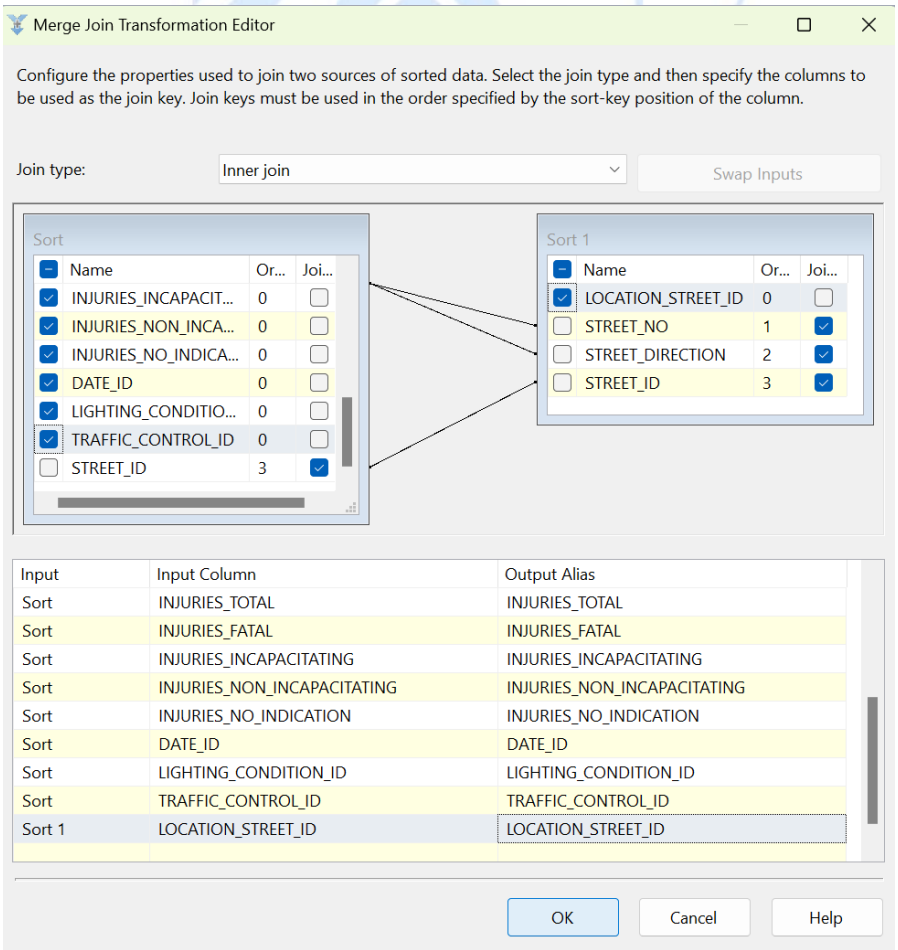


- Kết nối “Sort1” với “Merge Join”.



Bước 10: Đầu tiên, nhấn chuột phải vào “Merge Join” và chọn “Edit”, một hộp thoại Merge Join Transformation Editor xuất hiện.

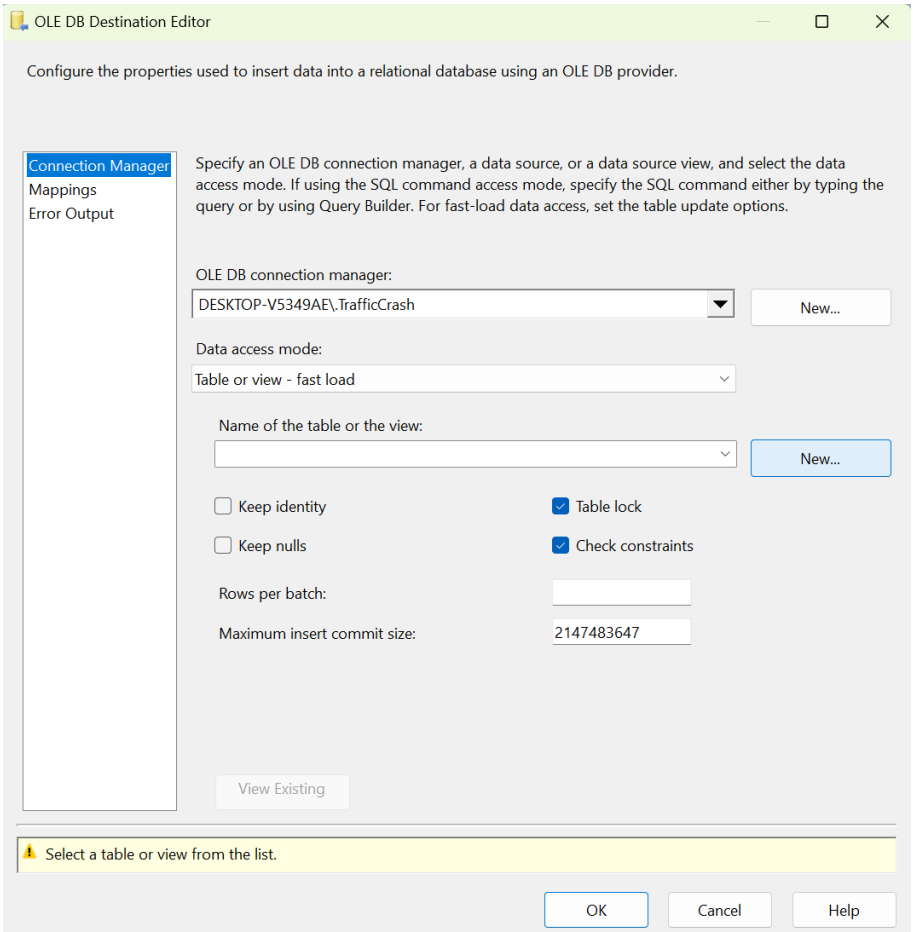
- Sau đó, ta chọn LOCATION_STREET_ID ở “Sort1” để merge vào “Fact_S”.
- Tiếp theo, ta chọn tất cả các cột của “Sort” nhưng không lấy thuộc tính STREET_NO, STREET_DIRECTION, STREET_ID.
- Nhấn “OK” để hoàn tất.



Kết quả sau khi merge là bảng Fact_S không còn các thuộc tính STREET_NO, STREET_DIRECTION, STREET_ID và có thêm 1 thuộc tính mới là LOCATION_STREET_ID.

Bước 11: Tạo một OLE DB Destination đặt tên là "Fact4" để chứa tất cả những gì đã merge.

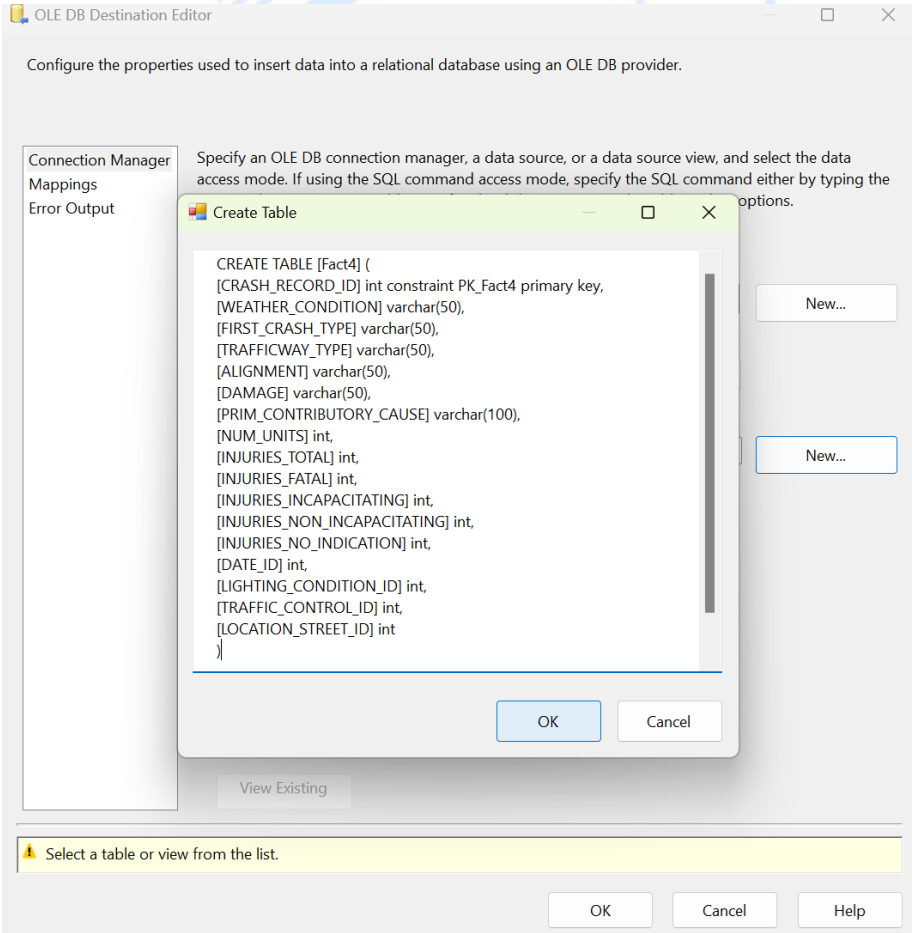
- Nhấn chuột phải vào "Fact4", chọn "Edit". Chọn "New" để tạo bảng.



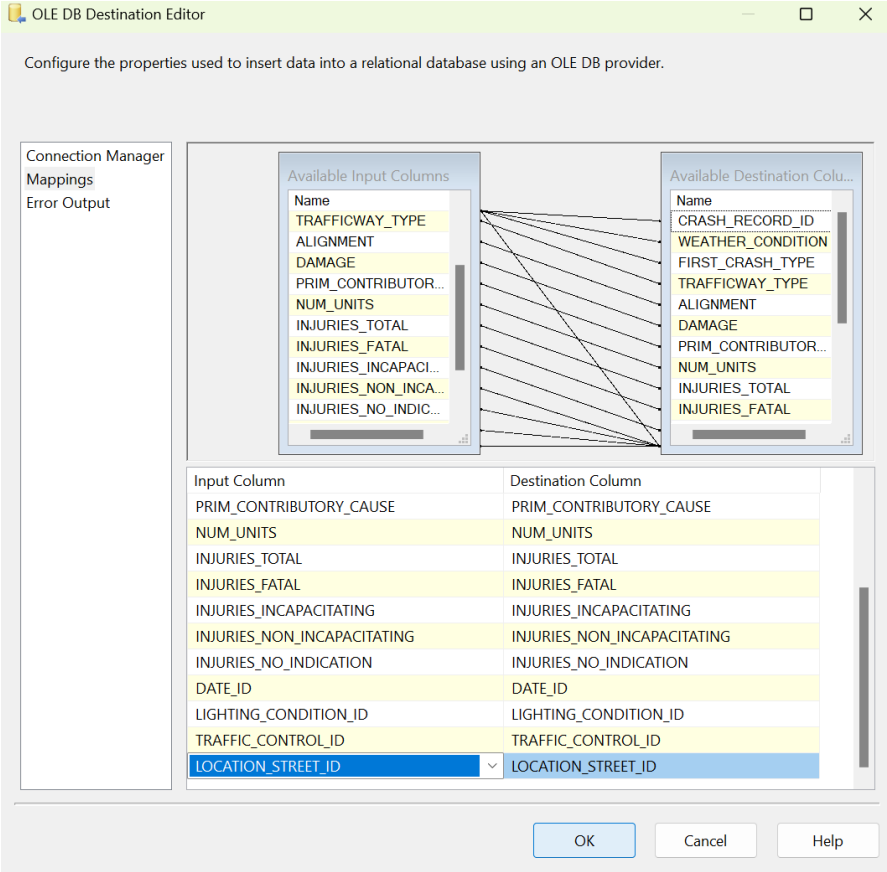
- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Fact4] (
[CRASH_RECORD_ID] int constraint PK_Fact4 primary key,
[WEATHER_CONDITION] varchar(50),
[FIRST_CRASH_TYPE] varchar(50),
[TRAFFICWAY_TYPE] varchar(50),
[ALIGNMENT] varchar(50),
[DAMAGE] varchar(50),
```

```
[PRIM_CONTRIBUTORY_CAUSE] varchar(100),
[NUM_UNITS] int,
[INJURIES_TOTAL] int,
[INJURIES_FATAL] int,
[INJURIES_INCAPACITATING] int,
[INJURIES_NON_INCAPACITATING] int,
[INJURIES_NO_INDICATION] int,
[DATE_ID] int,
[LIGHTING_CONDITION_ID] int,
[TRAFFIC_CONTROL_ID] int,
[LOCATION_STREET_ID] int
)
```



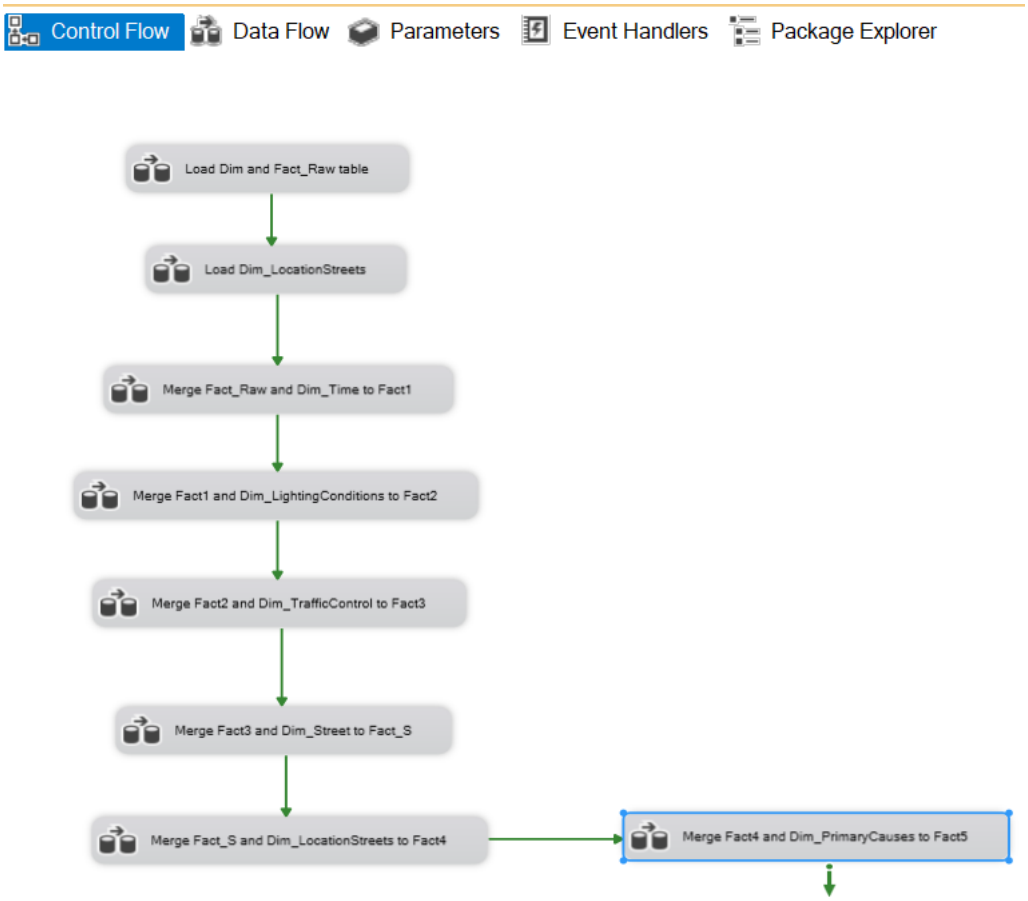
- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



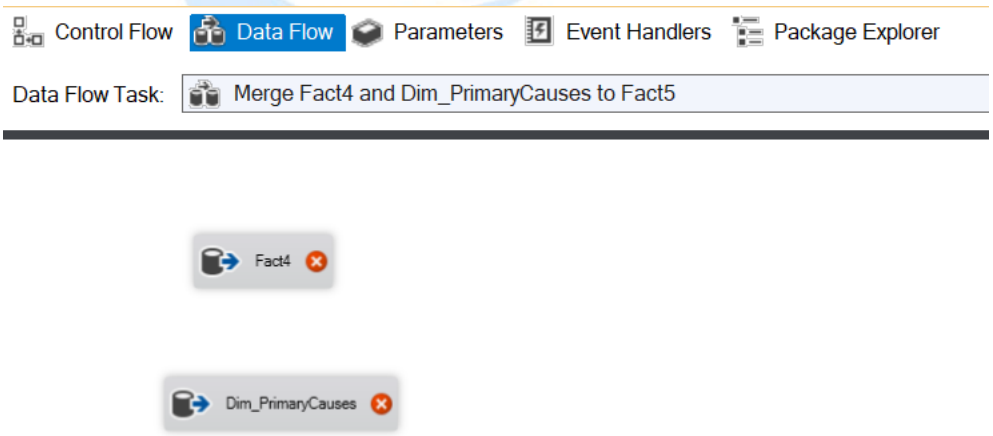
Tiếp tục quá trình merge bảng Fact4 với các Dimension còn lại để có 1 bảng Fact_TrafficCrashRecords hoàn chỉnh.

2.3.13.6 Merge Fact4 và Dim_PrimaryCauses vào Fact5.

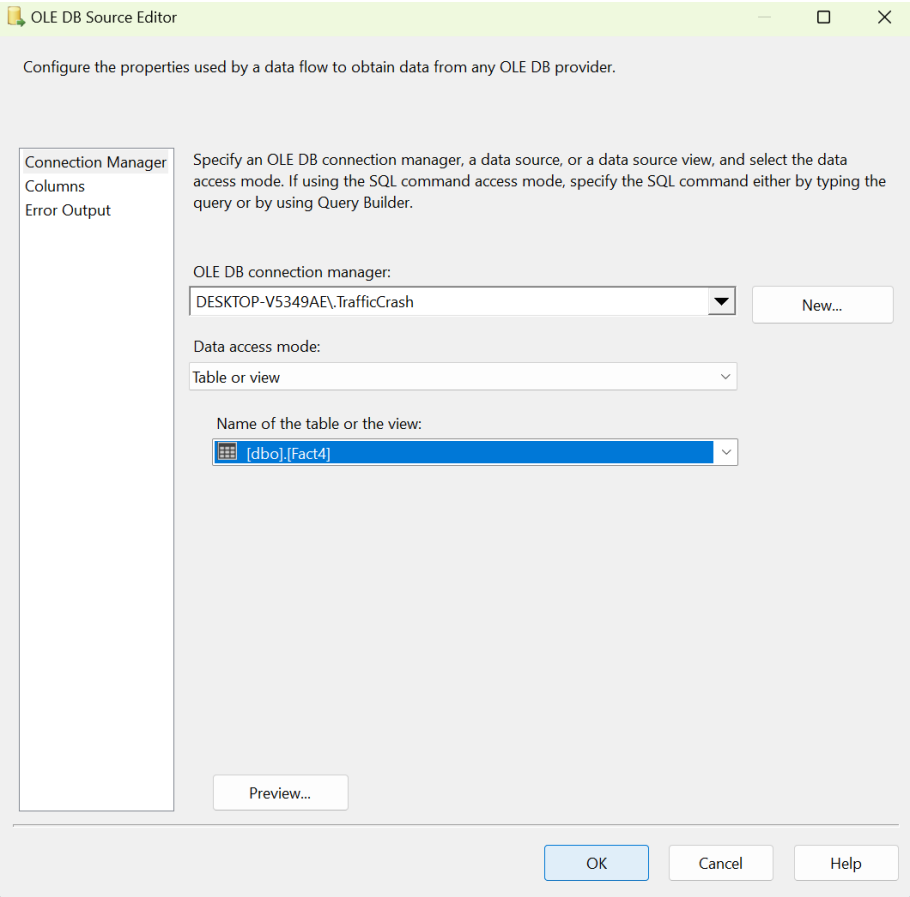
Bước 1: Ở tab Control Flow, tạo một Data Flow Task và đặt tên là “Merge Fact4 and Dim_PrimaryCauses to Fact5”.



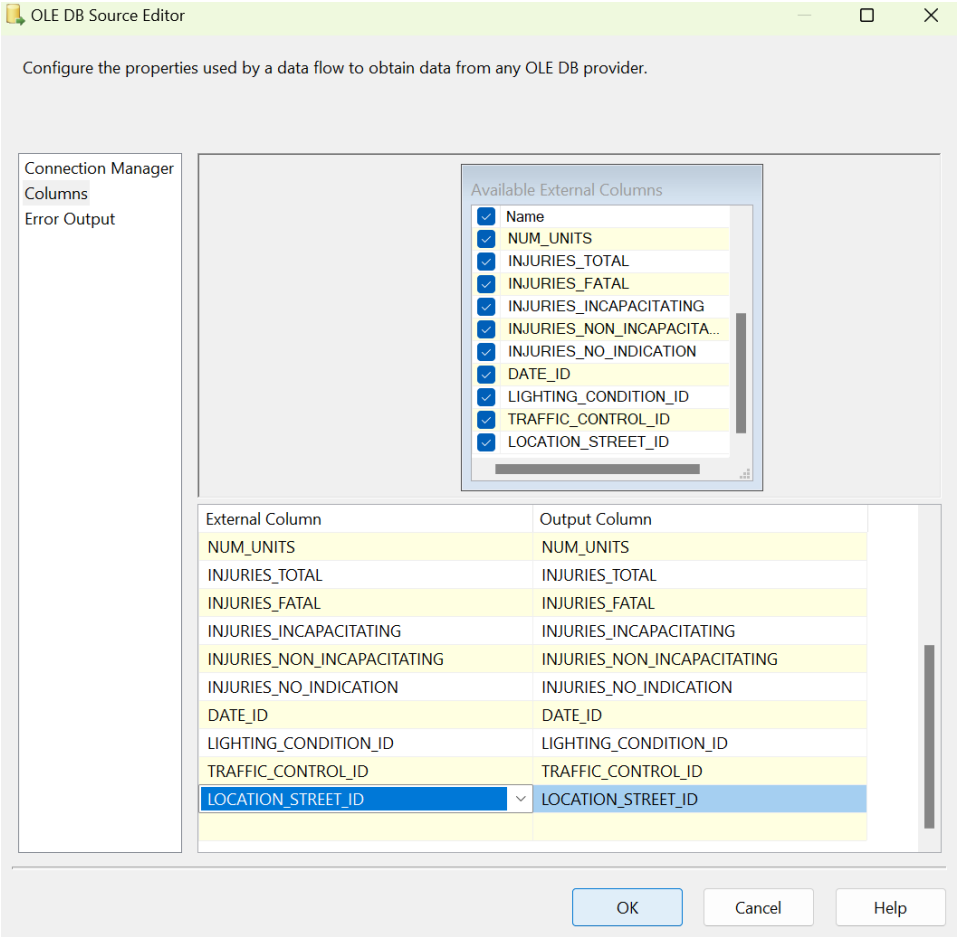
Bước 2: Nhấn chuột phải vào "Merge Fact4 and Dim_PrimaryCauses to Fact5" và chọn "Edit", trong tab Data Flow ta tạo hai **OLE DB Source** và đặt tên là "Fact4" và "Dim_PrimaryCauses".



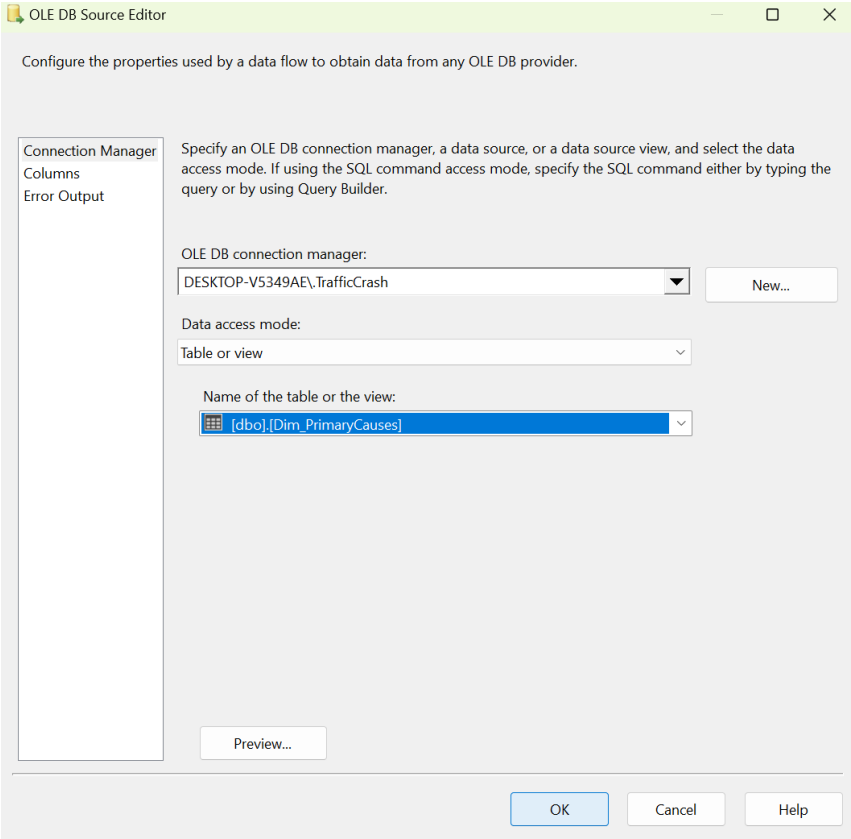
Bước 3: Nhấn chuột phải chọn "Edit", sau đó chọn bảng Fact4 đã tạo trước đó làm data source cho bảng Fact4 mới này.



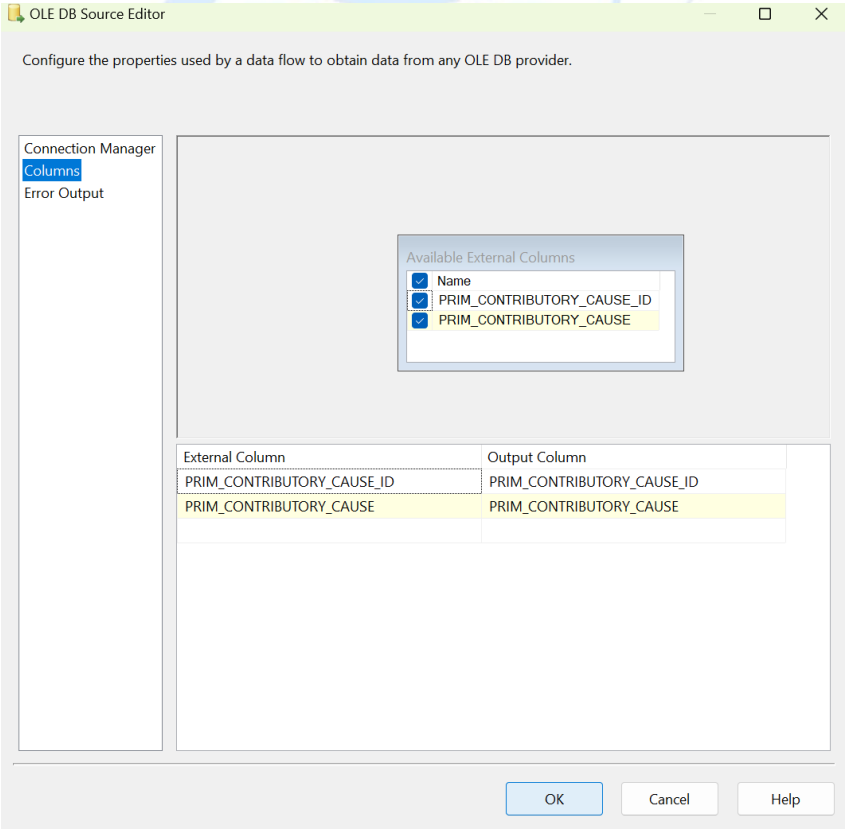
Bước 4: Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



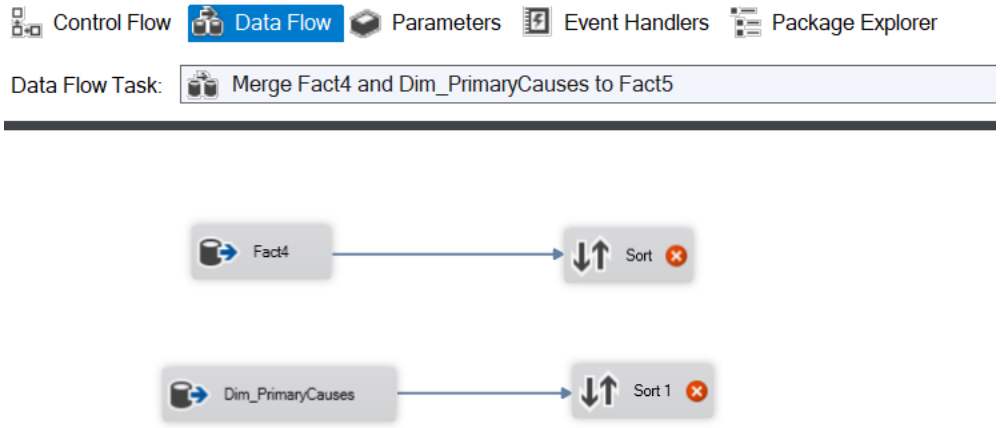
Bước 5: Tương tự thực hiện chọn bảng cho Dim_PrimaryCauses.



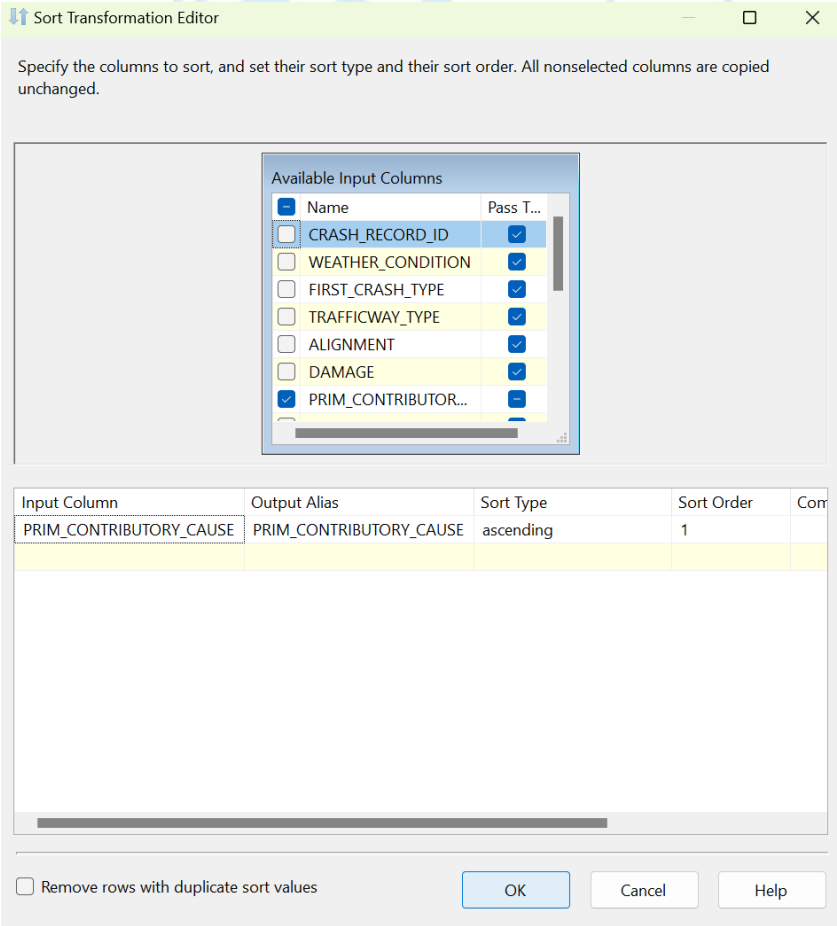
- Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



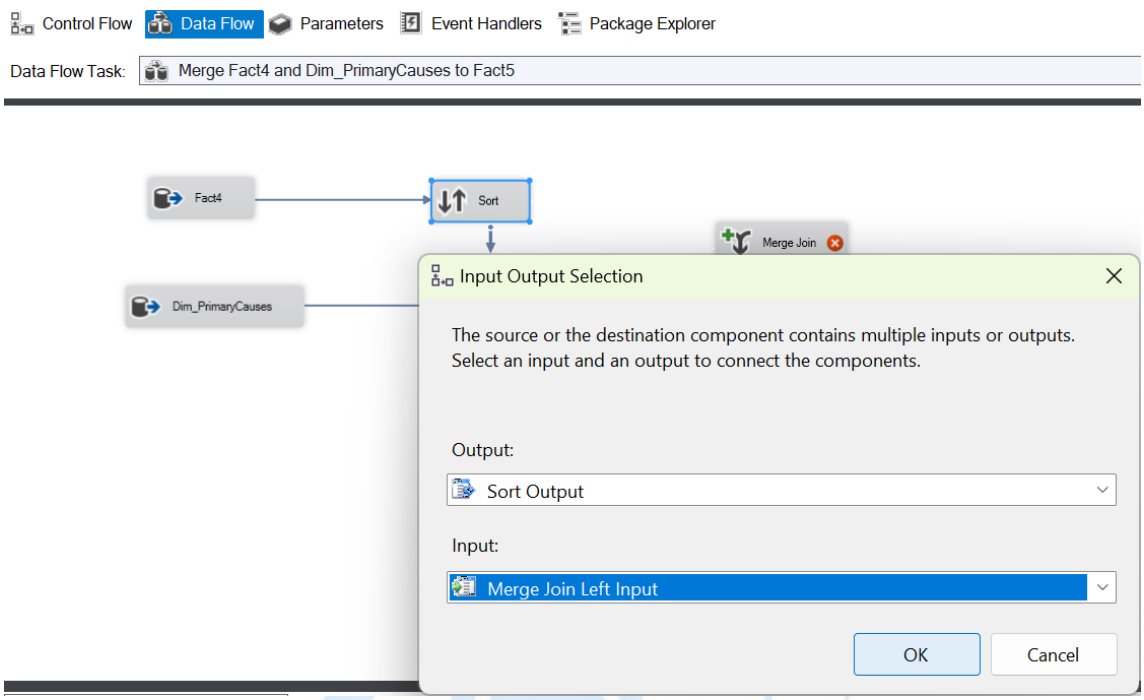
Bước 6: Tạo hai Sort đặt tên là "Sort" và "Sort1" tương ứng với mỗi "OLE DB Source" đã tạo. Thực hiện tạo các kết nối.



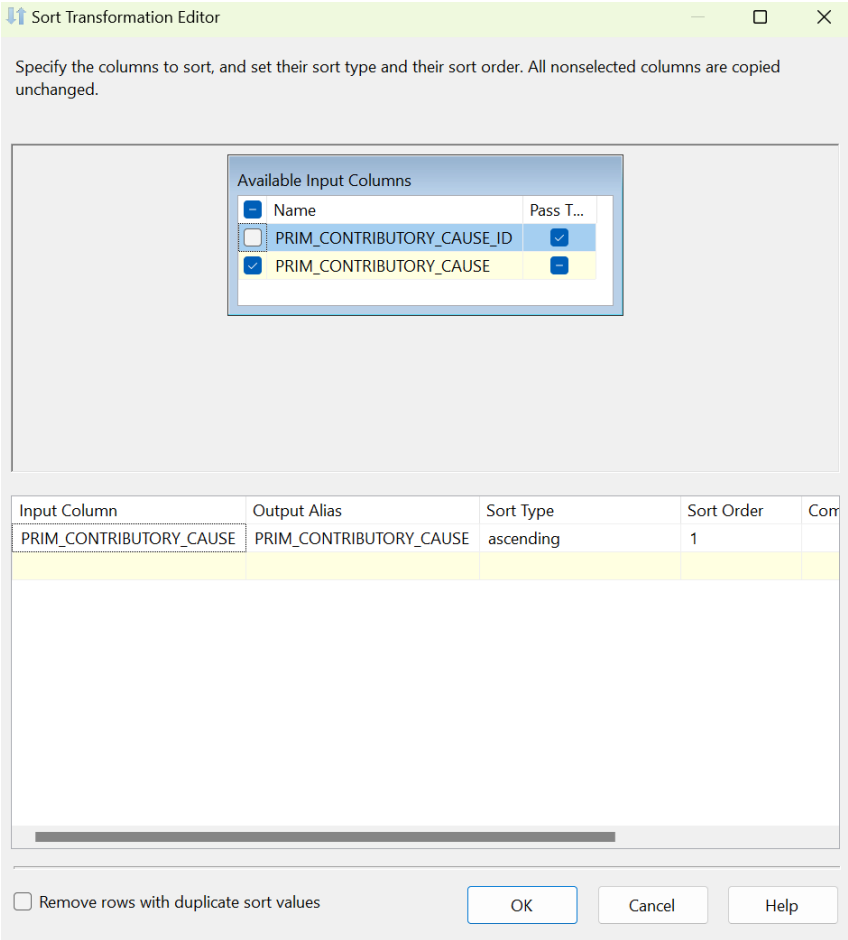
Bước 7: Ở "Sort", nhấn chuột phải chọn "Edit" và chọn cột để chuẩn bị cho quá trình merge. Sau đó nhấn "OK".



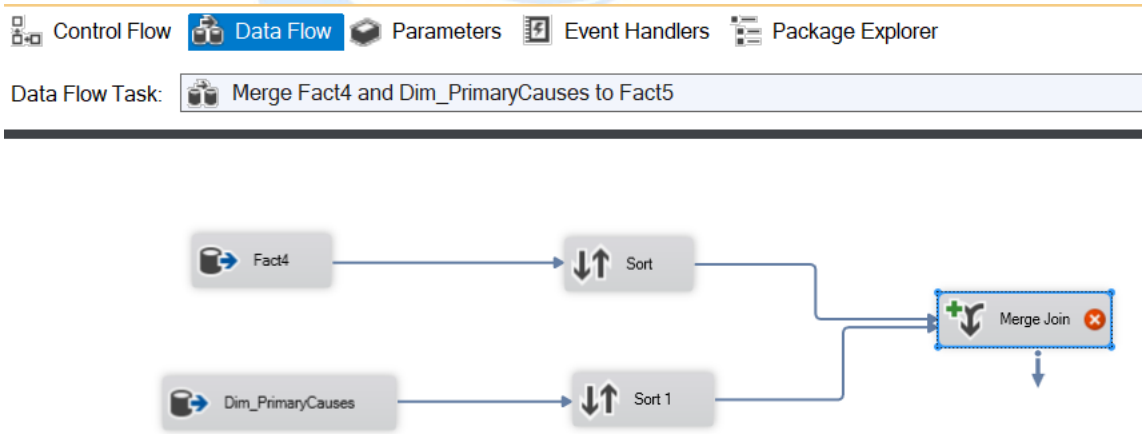
Bước 8: Tạo một Merge Join và nối với "Sort". Tiếp theo ta chọn "Merge Join Left Input" để giữ lại toàn bộ các dòng trong bảng "Fact4" bất kể có kết quả khi thực hiện phép kết trái với cột ID của bảng "Dim_PrimaryCauses" hay không.



Bước 9: Tương tự, ở "Sort1", nhấn chuột phải chọn "Edit" và chọn cột. Sau đó nhấn "OK".

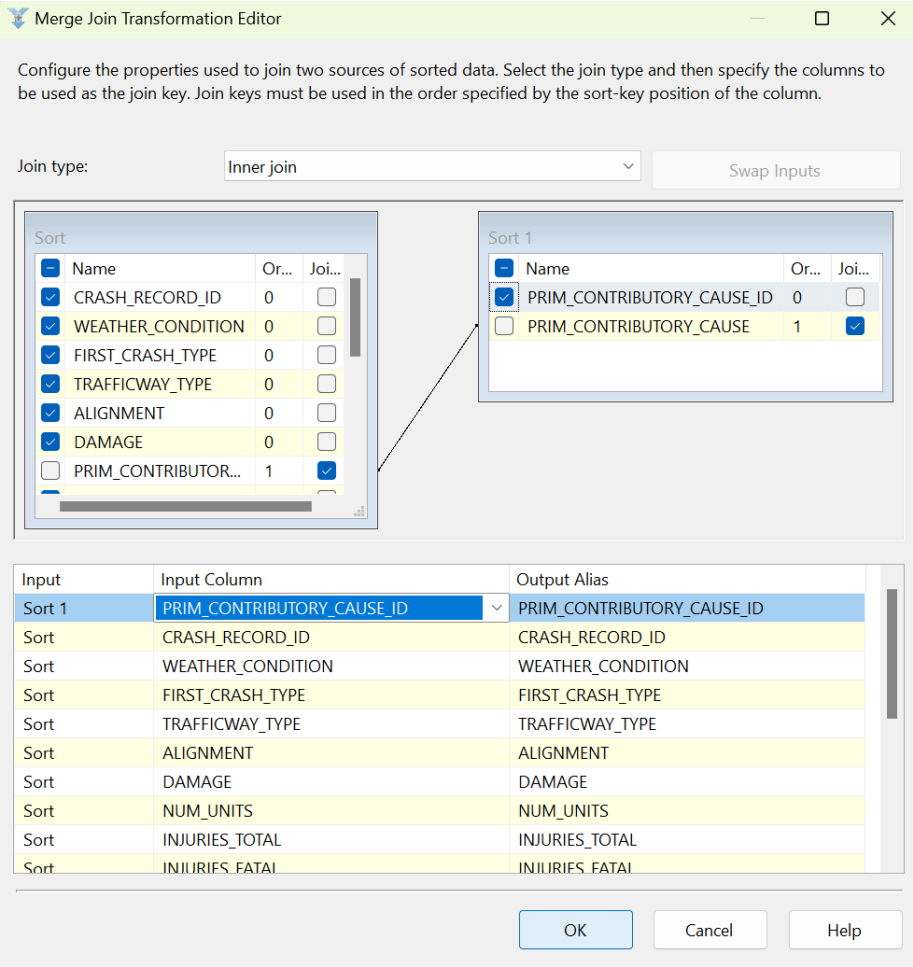


- Kết nối "Sort1" với Merge Join.



Bước 10: Đầu tiên, nhấn chuột phải vào "Merge Join" và chọn "Edit", một hộp thoại Merge Join Transformation Editor xuất hiện.

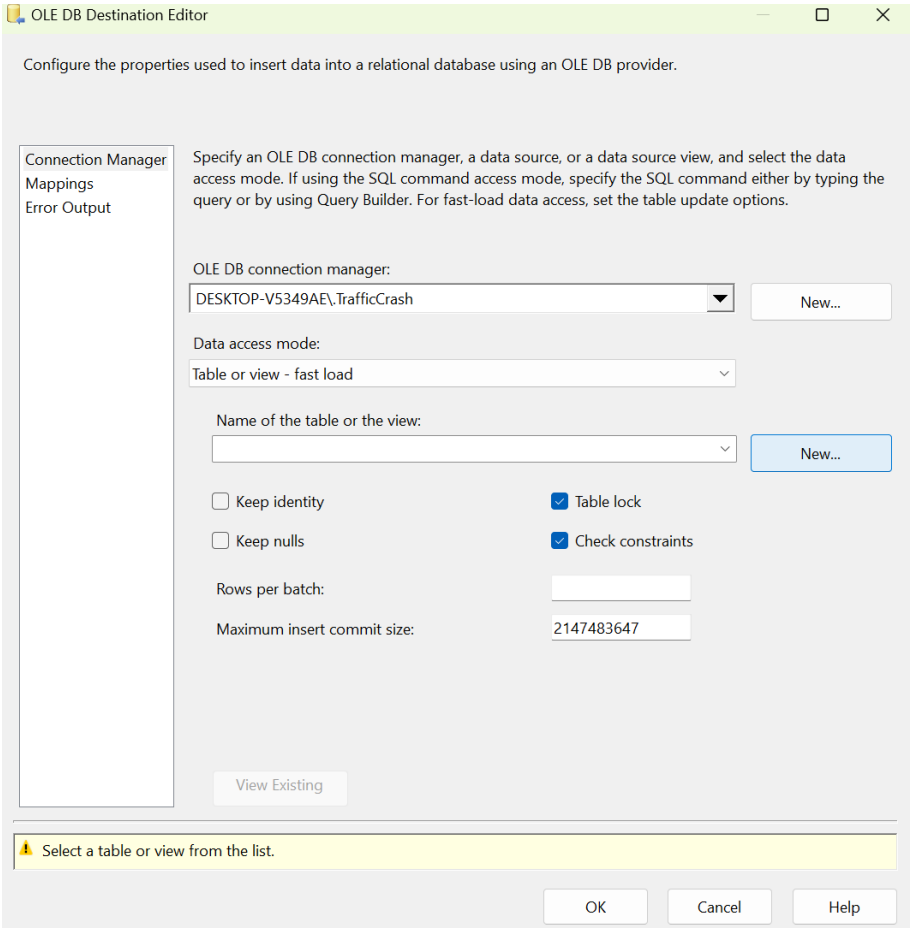
- Sau đó, ta chọn PRIM_CONTRIBUTORY_CAUSE_ID ở “Sort1” để merge vào “Fact4”.
- Tiếp theo, ta chọn tất cả các cột của “Sort” nhưng không lấy thuộc tính PRIM_CONTRIBUTORY_CAUSE.
- Nhấn “OK” để hoàn tất.



Kết quả sau khi merge là bảng Fact4 không còn các thuộc tính PRIM_CONTRIBUTORY_CAUSE và có thêm 1 thuộc tính mới là PRIM_CONTRIBUTORY_CAUSE_ID.

Bước 11: Tạo một OLE DB Destination đặt tên là “Fact5” để chứa tất cả những gì đã merge.

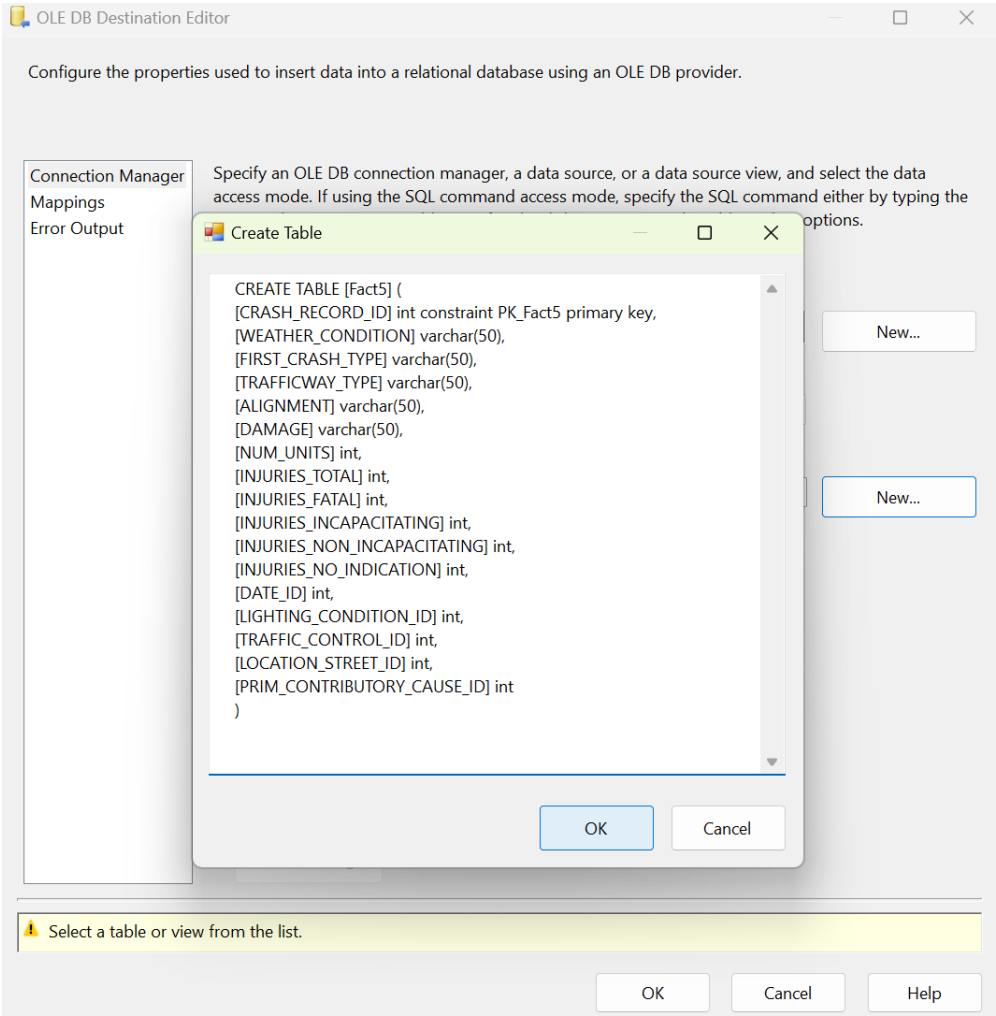
- Nhấn chuột phải vào “Fact5”, chọn “Edit”. Chọn “New” để tạo bảng.



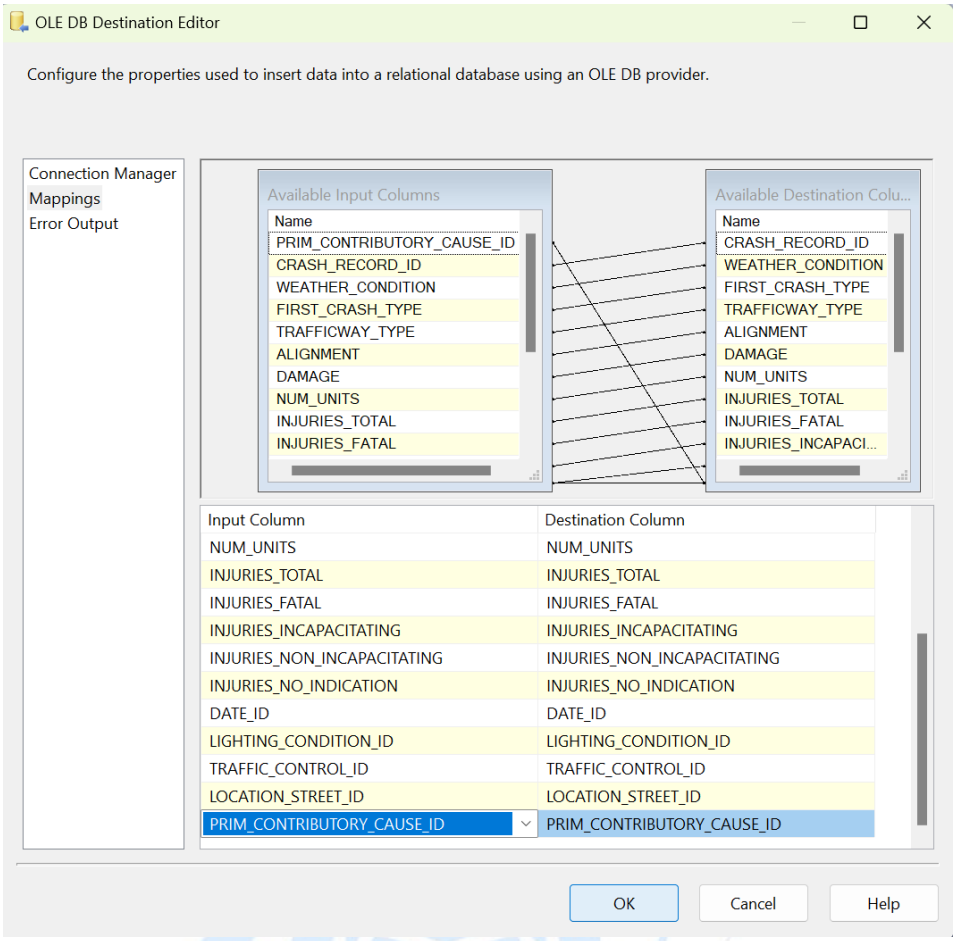
- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Fact5] (
[CRASH_RECORD_ID] int constraint PK_Fact5 primary key,
[WEATHER_CONDITION] varchar(50),
[FIRST_CRASH_TYPE] varchar(50),
[TRAFFICWAY_TYPE] varchar(50),
[ALIGNMENT] varchar(50),
[DAMAGE] varchar(50),
[NUM_UNITS] int,
[INJURIES_TOTAL] int,
[INJURIES_FATAL] int,
[INJURIES_INCAPACITATING] int,
```

```
[INJURIES_NON_INCAPACITATING] int,
[INJURIES_NO_INDICATION] int,
[DATE_ID] int,
[LIGHTING_CONDITION_ID] int,
[TRAFFIC_CONTROL_ID] int,
[LOCATION_STREET_ID] int,
[PRIM_CONTRIBUTORY_CAUSE_ID] int
)
```



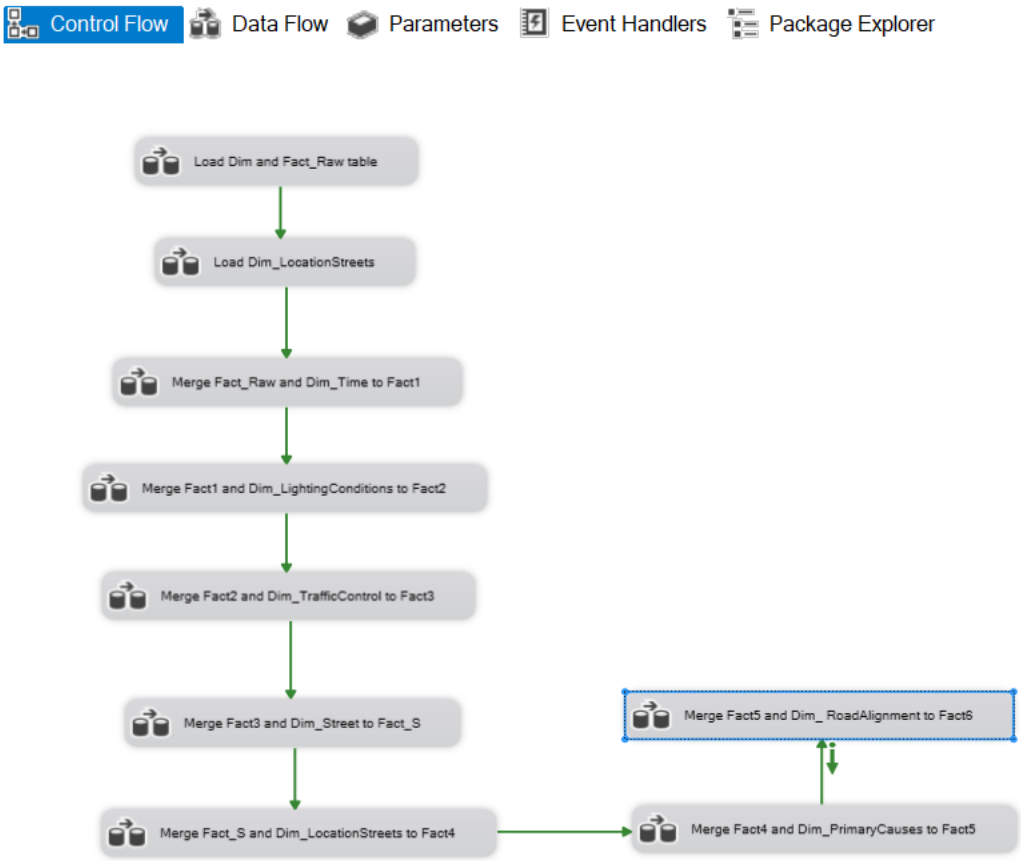
- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



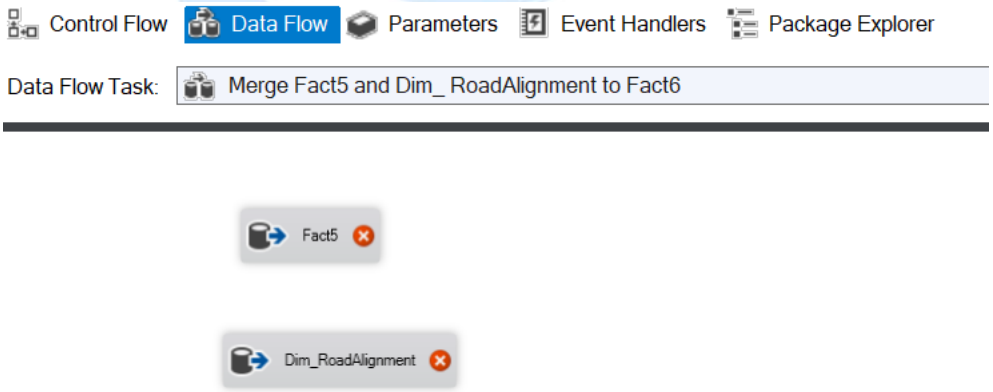
Tiếp tục quá trình merge bảng Fact5 với các Dimension còn lại để có 1 bảng Fact_TrafficCrashRecords hoàn chỉnh.

2.3.13.7 Merge Fact5 và Dim_RoadAlignment vào Fact6.

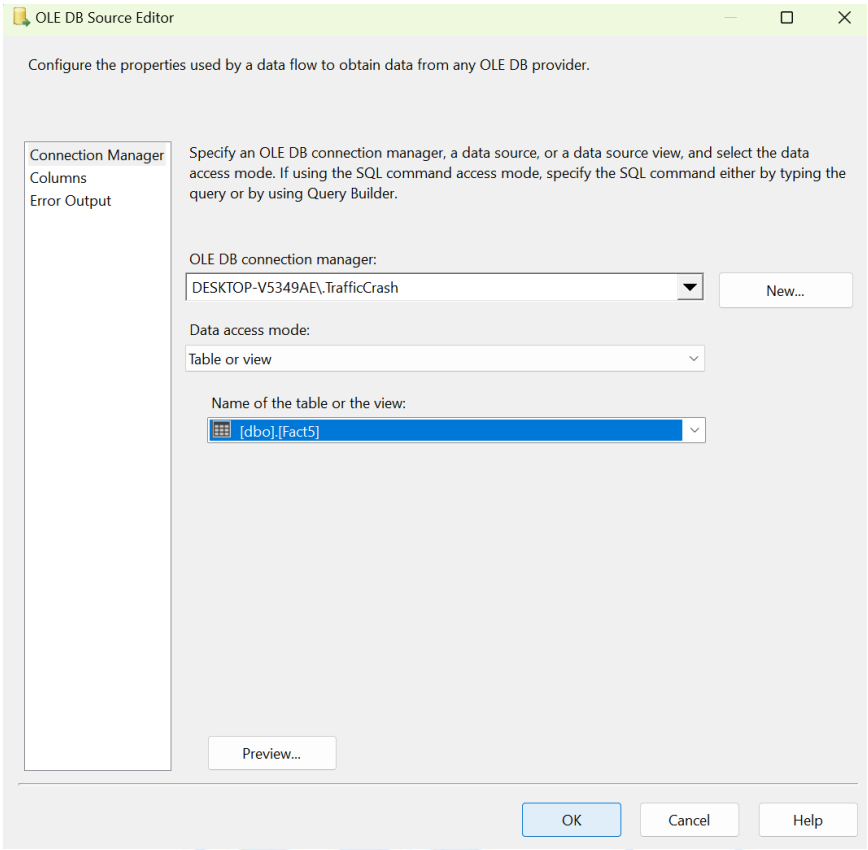
Bước 1: Ở tab Control Flow, tạo một Data Flow Task và đặt tên là “Merge Fact5 and Dim_ RoadAlignment to Fact6”.



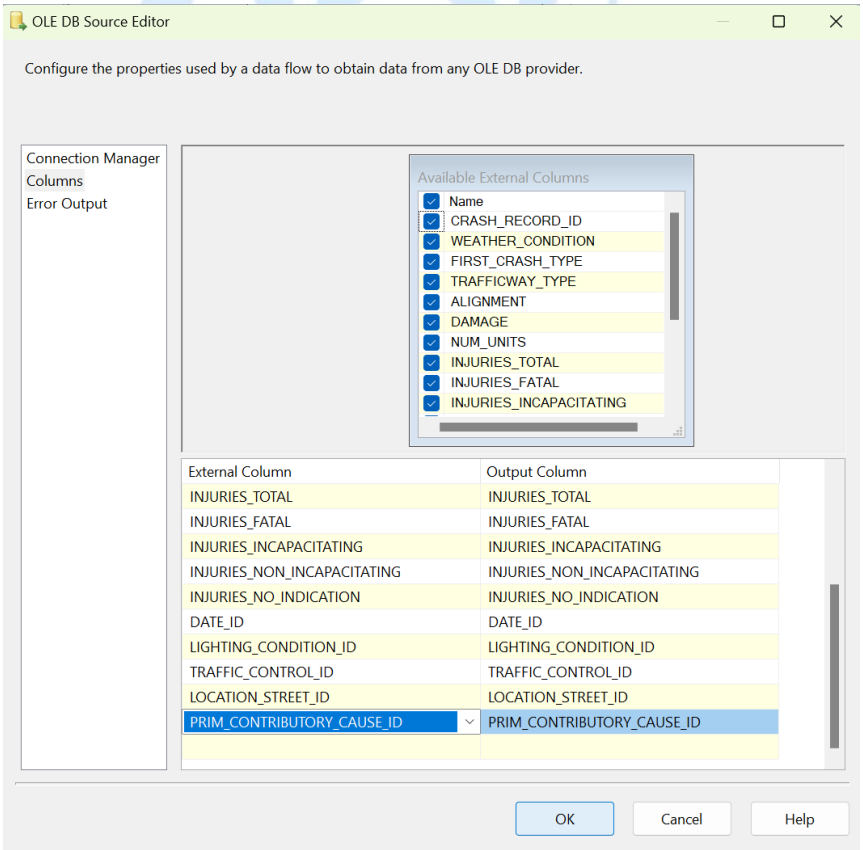
Bước 2: Nhấn chuột phải vào "Merge Fact5 and Dim_ RoadAlignment to Fact6" và chọn "Edit", trong tab Data Flow ta tạo hai **OLE DB Source** và đặt tên là "Fact5" và "Dim_RoadAlignment".



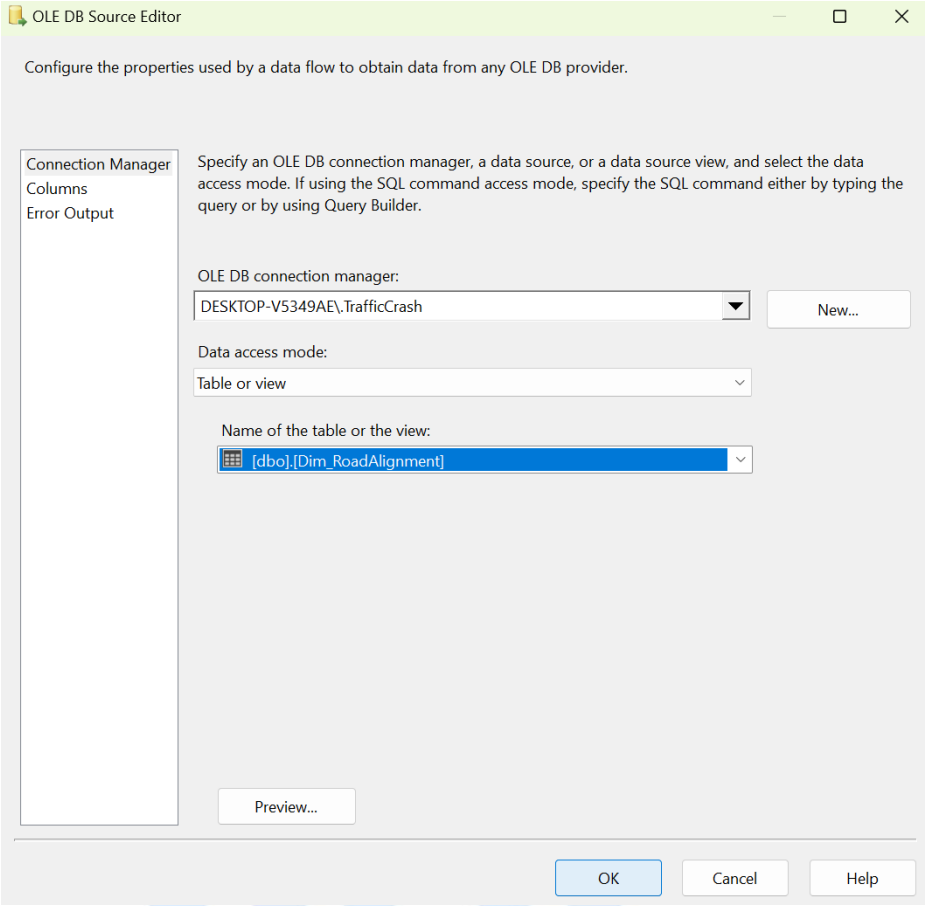
Bước 3: Nhấn chuột phải chọn "Edit", sau đó chọn bảng Fact5 đã tạo trước đó làm data source cho bảng Fact5 mới này.



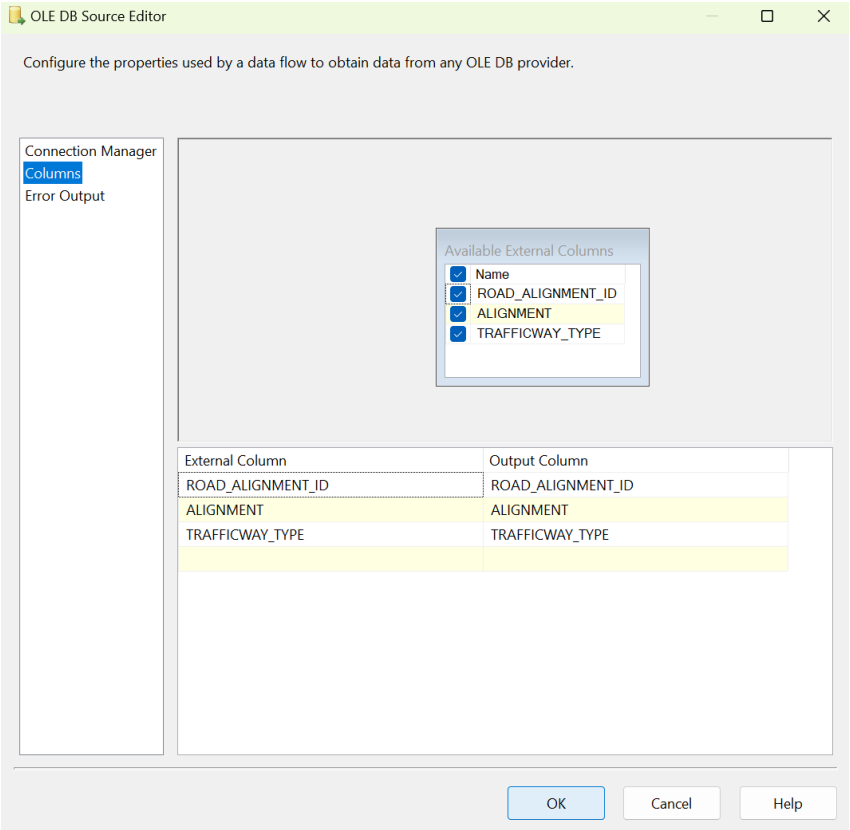
Bước 4: Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



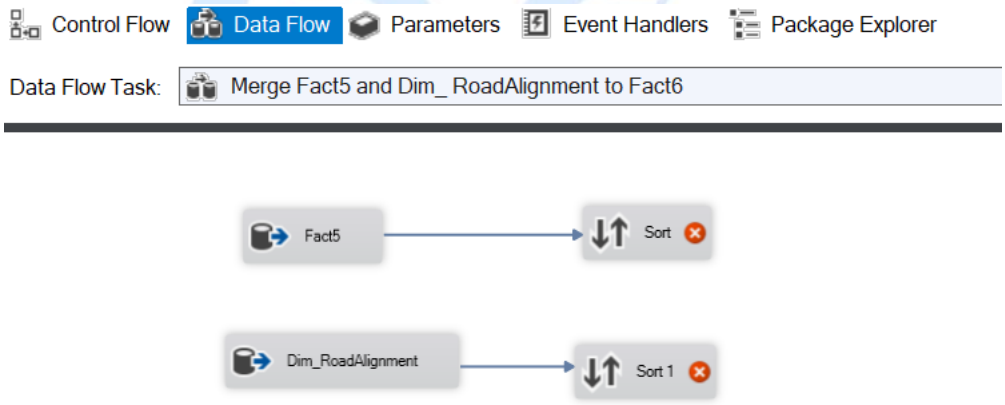
Bước 5: Tương tự thực hiện chọn bảng cho Dim_RoadAlignment.



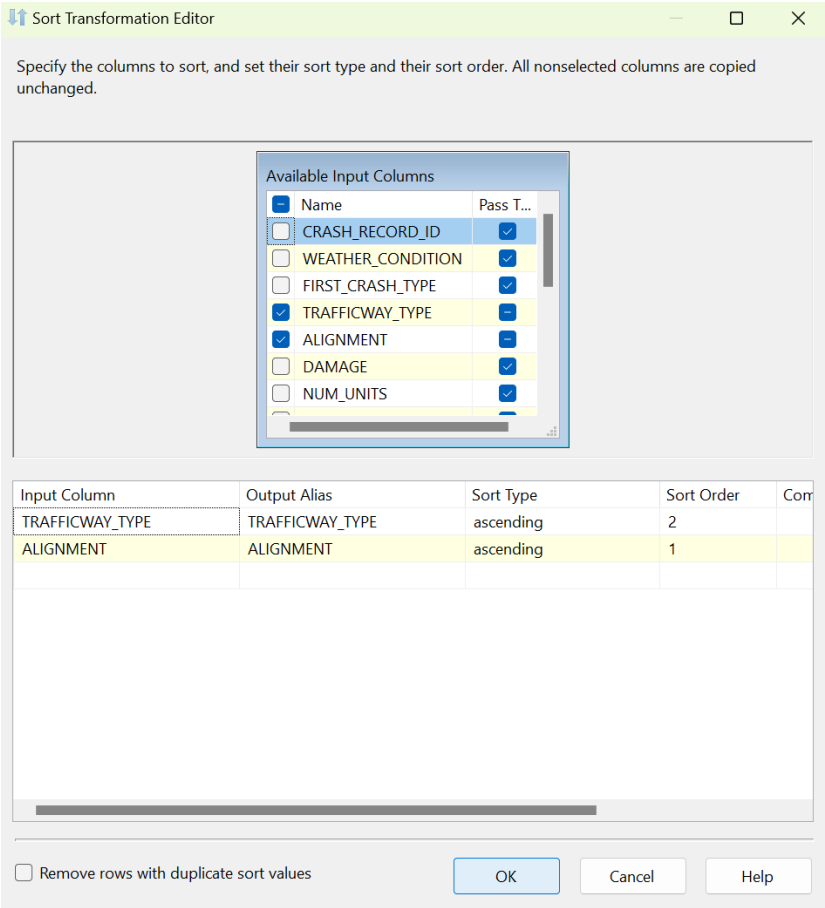
- Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



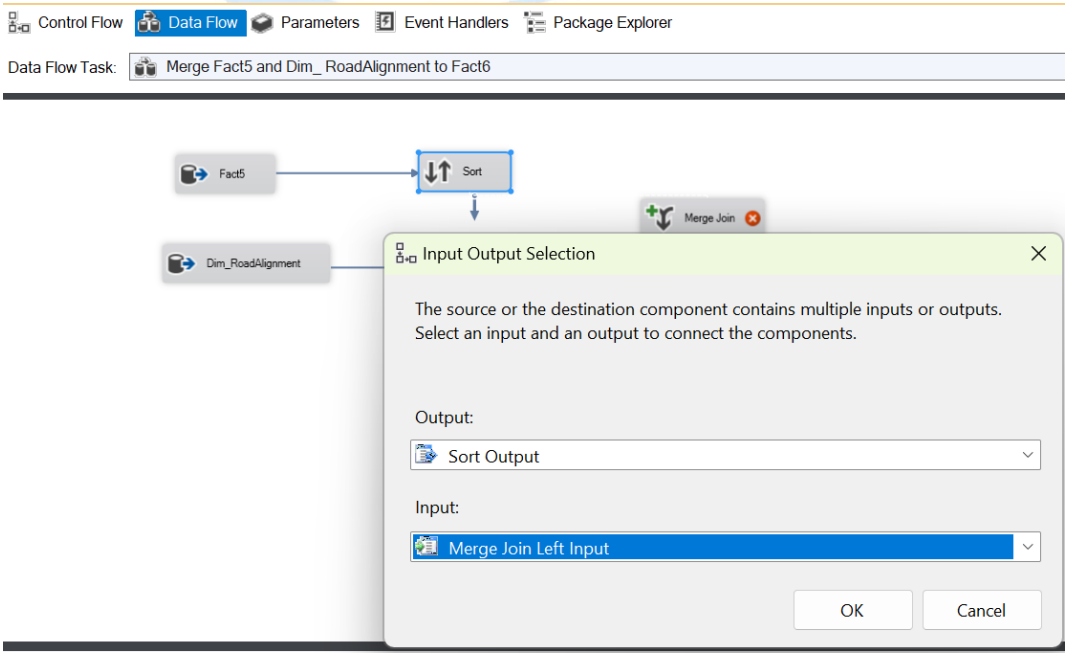
Bước 6: Tạo hai Sort đặt tên là "Sort" và "Sort1" tương ứng với mỗi "OLE DB Source" đã tạo. Thực hiện tạo các kết nối.



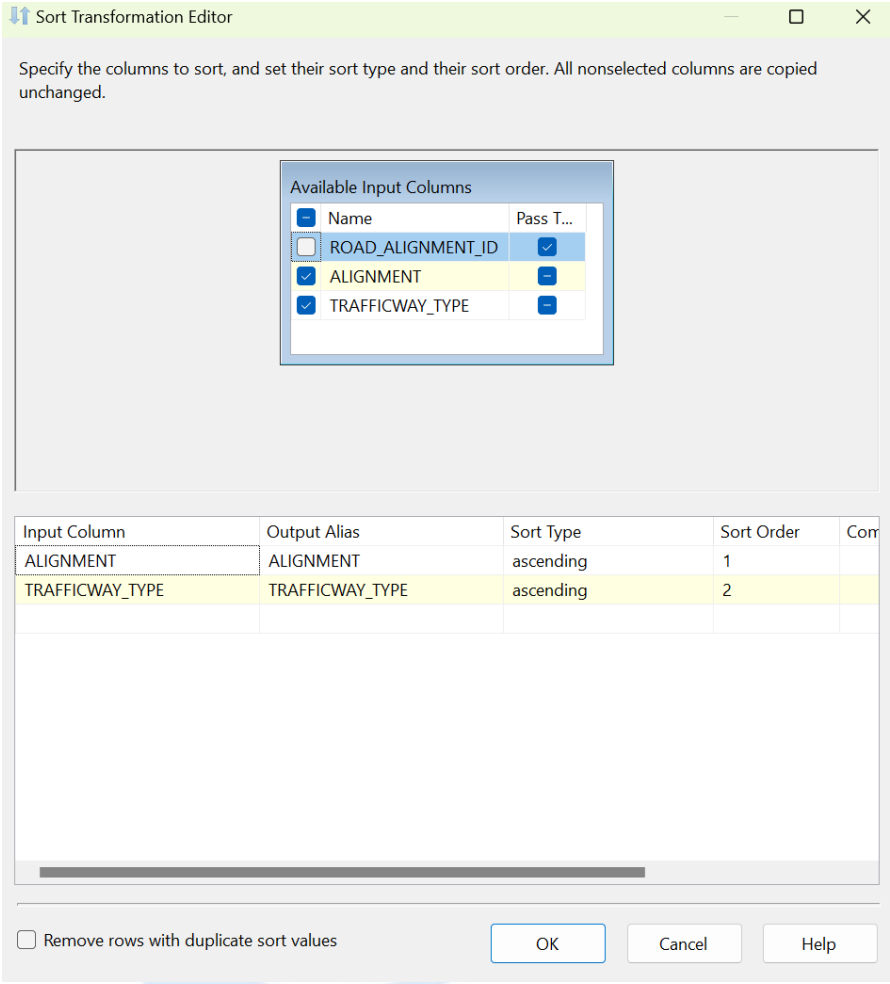
Bước 7: Ở "Sort", nhấn chuột phải chọn "Edit" và chọn cột để chuẩn bị cho quá trình merge. Sau đó nhấn "OK".



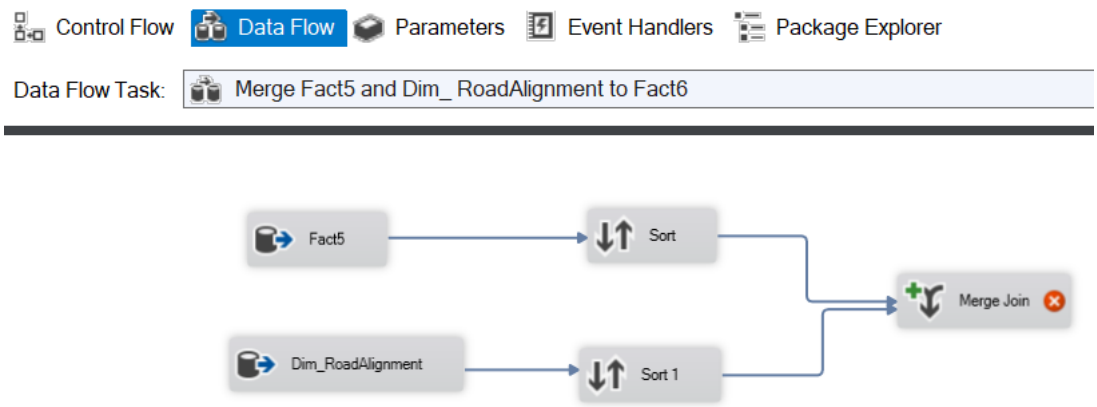
Bước 8: Tạo một Merge Join và nối với "Sort". Tiếp theo ta chọn "Merge Join Left Input" để giữ lại toàn bộ các dòng trong bảng "Fact5" bất kể có kết quả khi thực hiện phép kết trái với cột ID của bảng "Dim_RoadAlignment" hay không.



Bước 9: Tương tự, ở "Sort1", nhấn chuột phải chọn "Edit" và chọn cột. Sau đó nhấn "OK".

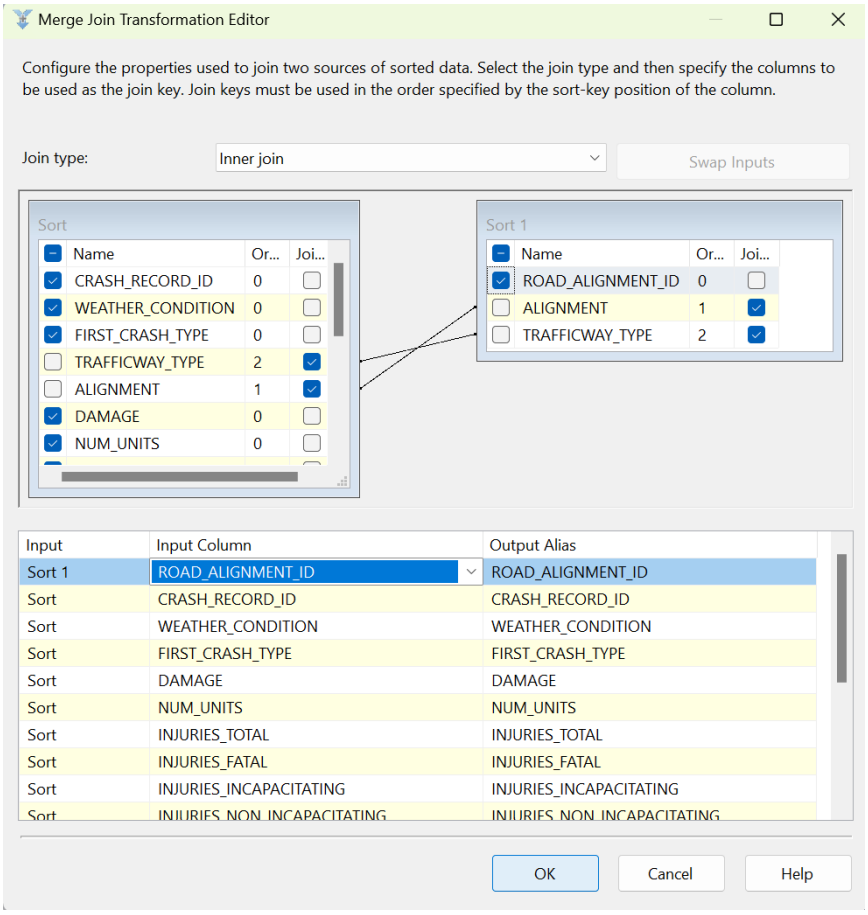


- Kết nối "Sort1" với Merge Join.



Bước 10: Đầu tiên, nhấn chuột phải vào "Merge Join" và chọn "Edit", một hộp thoại Merge Join Transformation Editor xuất hiện.

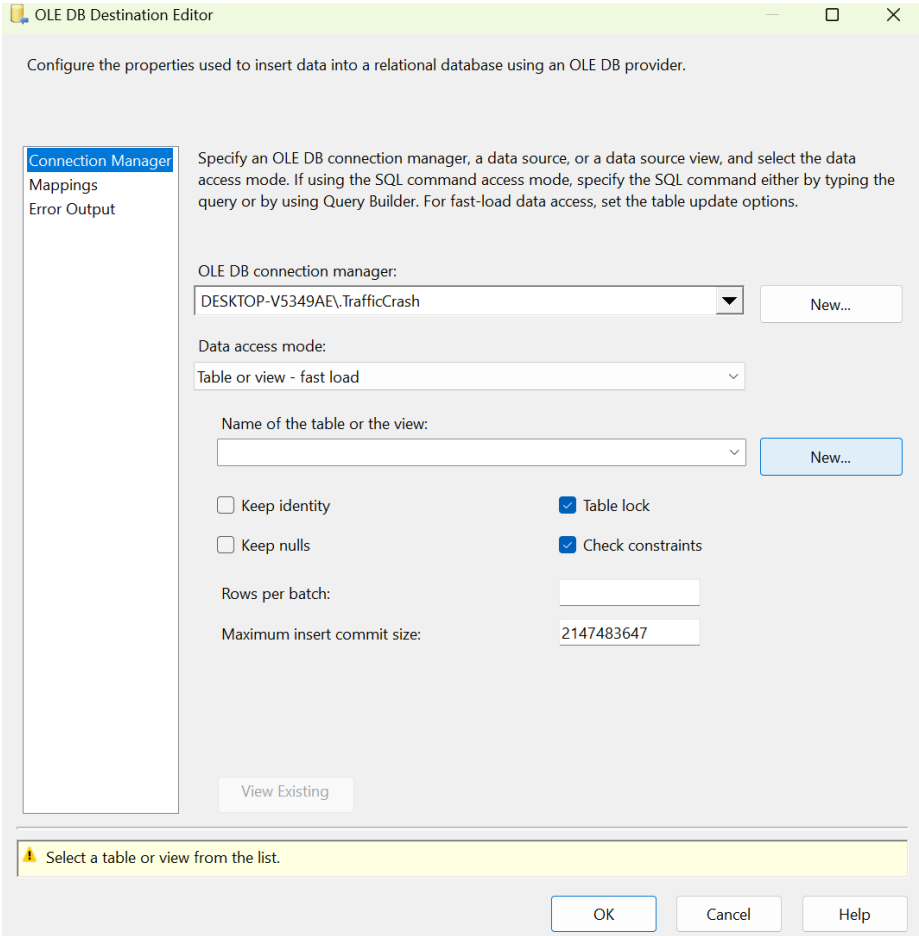
- Sau đó, ta chọn ROAD_ALIGNMENT_ID ở "Sort1" để merge vào "Fact5".
- Tiếp theo, ta chọn tất cả các cột của "Sort" nhưng không lấy thuộc tính ALIGNMENT và TRAFFICWAY_TYPE.
- Nhấn "OK" để hoàn tất.



Kết quả sau khi merge là bảng Fact5 không còn các thuộc tính ALIGNMENT và TRAFFICWAY_TYPE, có thêm 1 thuộc tính mới là ROAD_ALIGNMENT_ID.

Bước 11: Tạo một OLE DB Destination đặt tên là "Fact6" để chứa tất cả những gì đã merge.

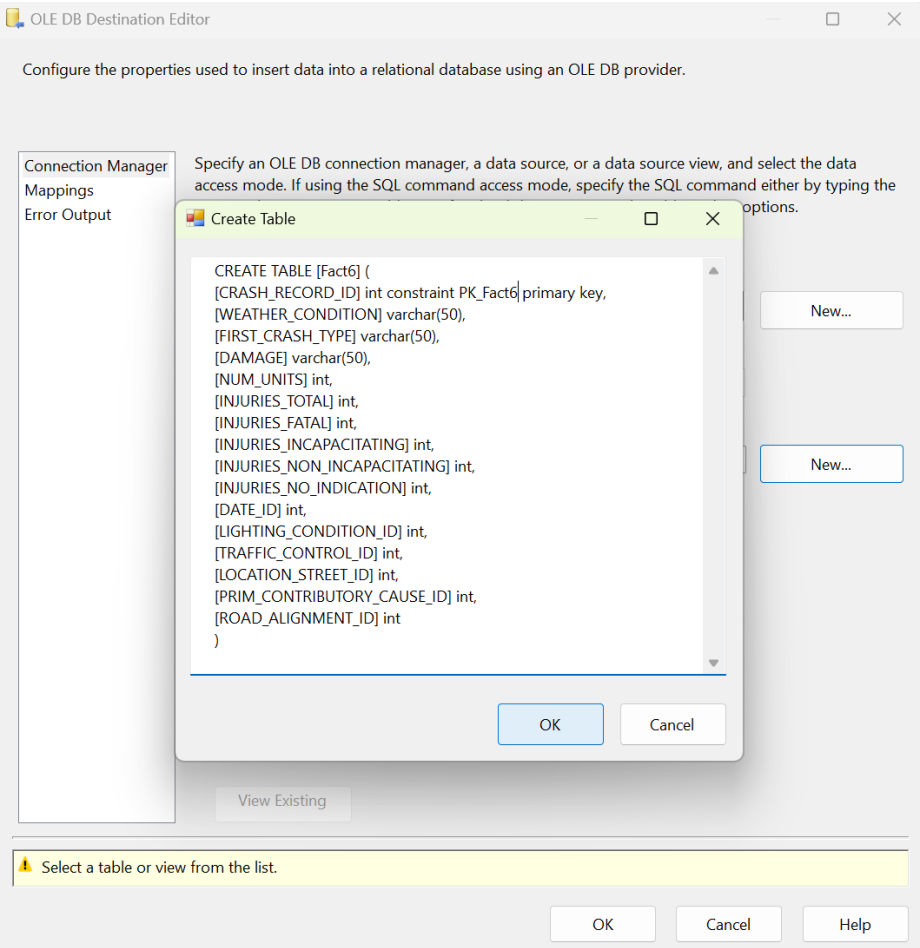
- Nhấn chuột phải vào "Fact6", chọn "Edit". Chọn "New" để tạo bảng.



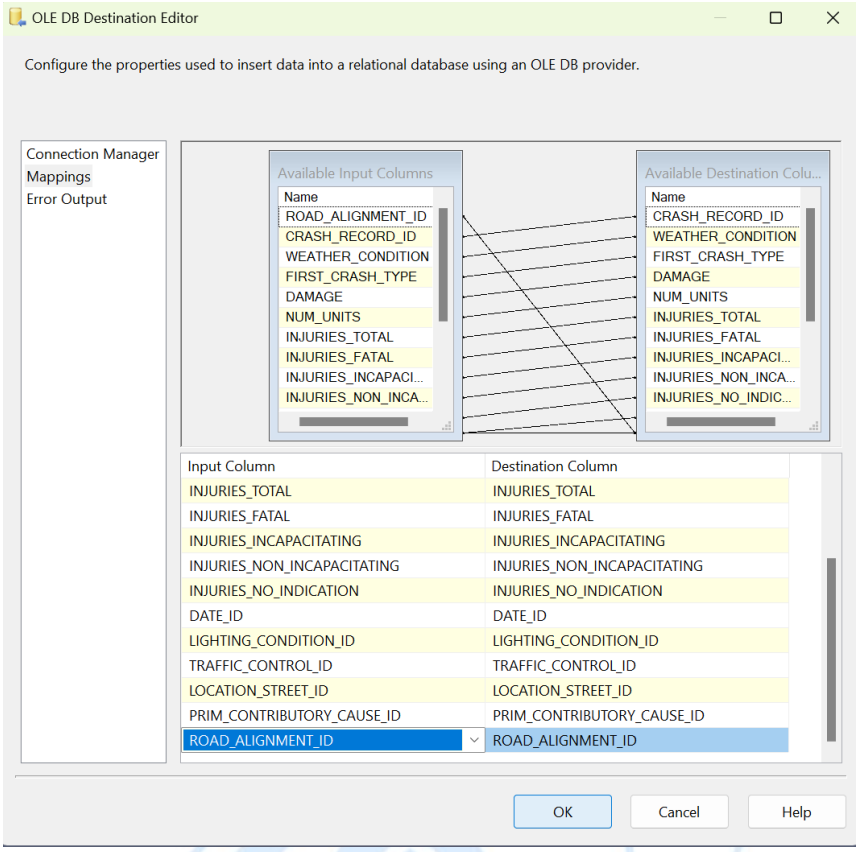
- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Fact6] (
[CRASH_RECORD_ID] int constraint PK_Fact6 primary key,
[WEATHER_CONDITION] varchar(50),
[FIRST_CRASH_TYPE] varchar(50),
[DAMAGE] varchar(50),
[NUM_UNITS] int,
[INJURIES_TOTAL] int,
[INJURIES_FATAL] int,
[INJURIES_INCAPACITATING] int,
[INJURIES_NON_INCAPACITATING] int,
[INJURIES_NO_INDICATION] int,
```

```
[DATE_ID] int,
[LIGHTING_CONDITION_ID] int,
[TRAFFIC_CONTROL_ID] int,
[LOCATION_STREET_ID] int,
[PRIM_CONTRIBUTORY_CAUSE_ID] int,
[ROAD_ALIGNMENT_ID] int
)
```



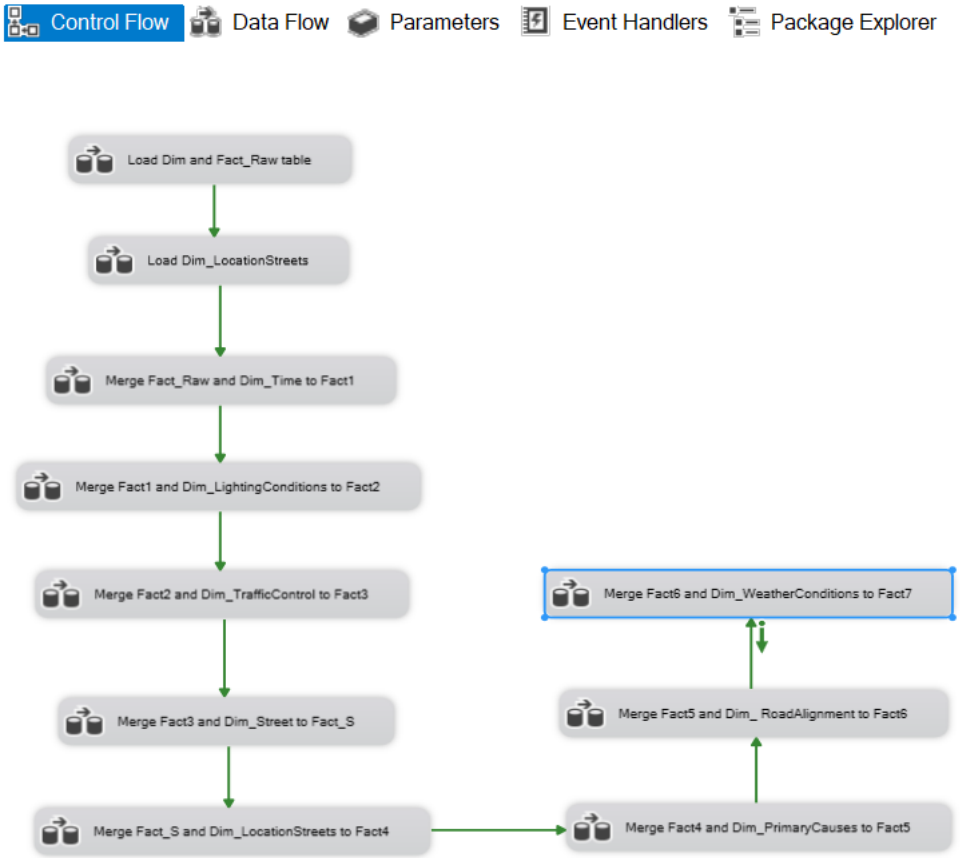
- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



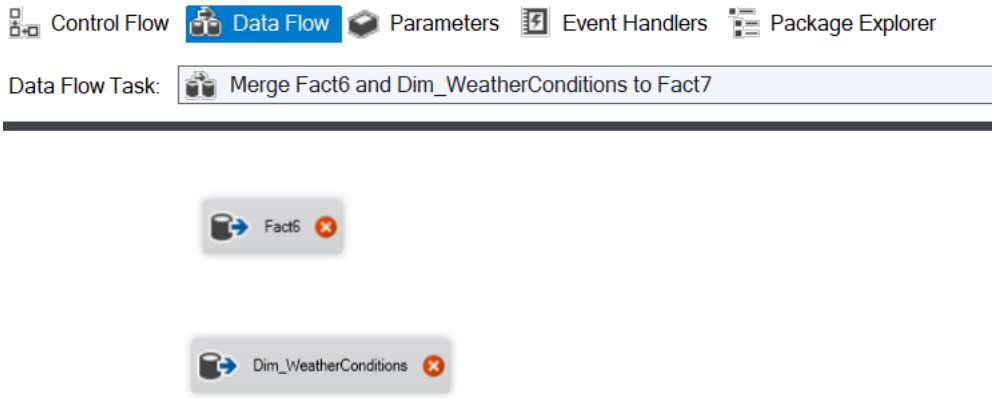
Tiếp tục quá trình merge bảng Fact6 với các Dimension còn lại để có 1 bảng Fact_TrafficCrashRecords hoàn chỉnh.

2.3.13.8 Merge Fact6 và Dim_WeatherConditions vào Fact7.

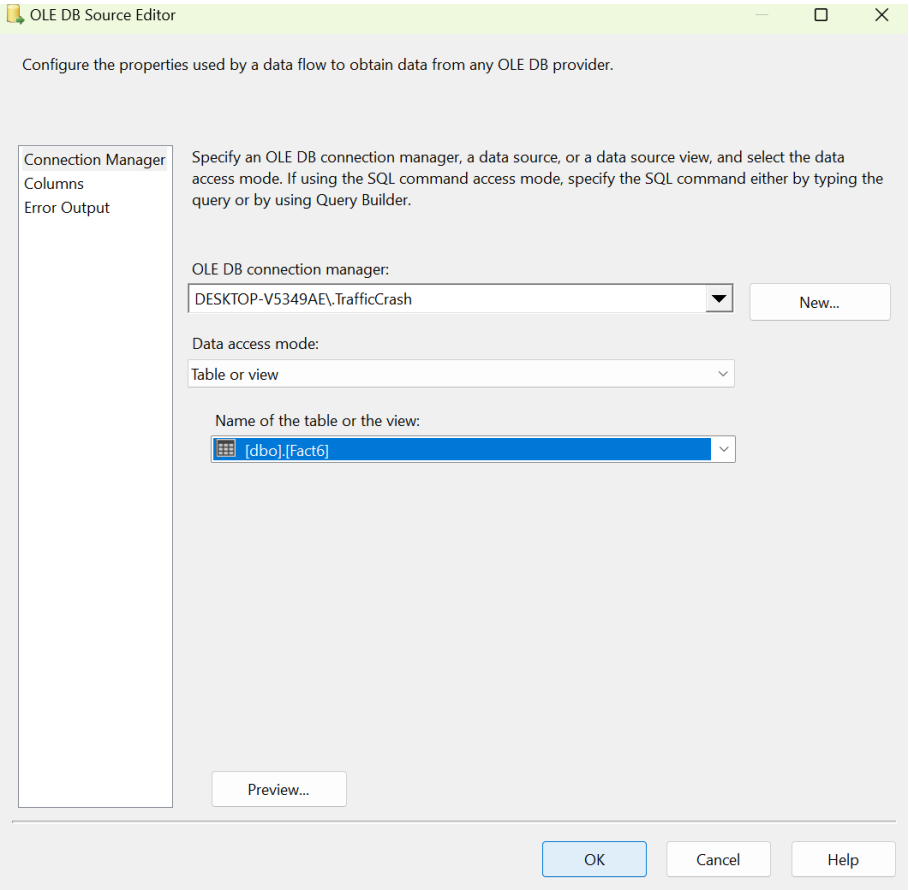
Bước 1: Ở tab Control Flow, tạo một Data Flow Task và đặt tên là “Merge Fact6 and Dim_WeatherConditions to Fact7”.



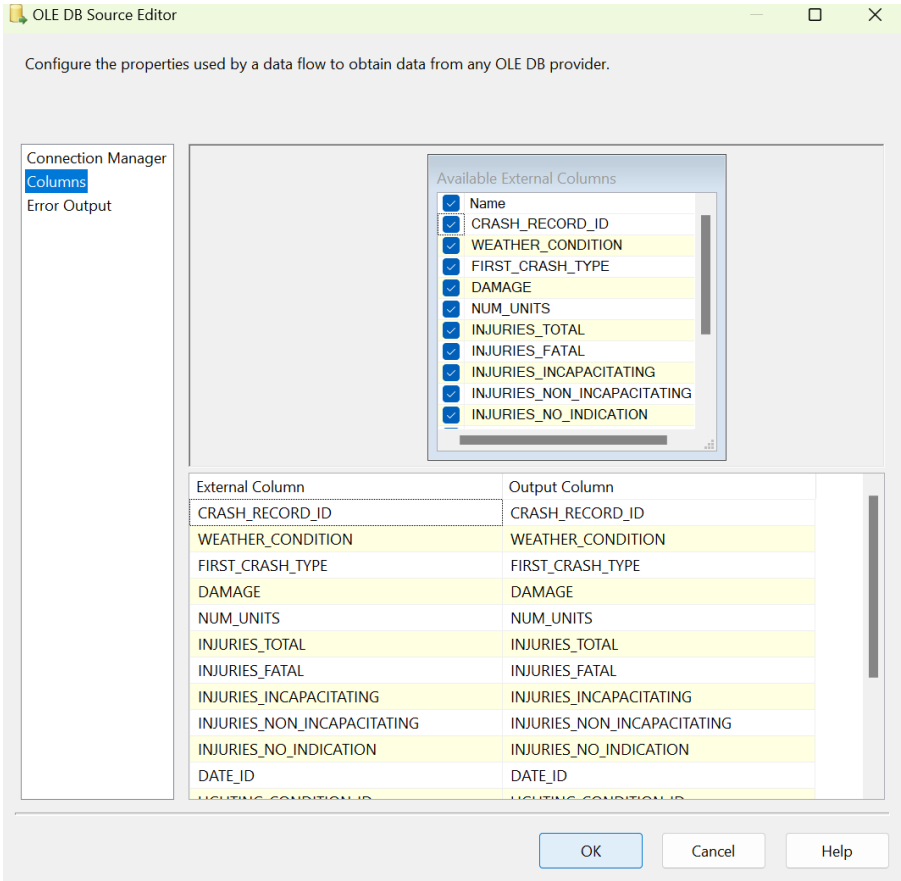
Bước 2: Nhấn chuột phải vào "Merge Fact6 and Dim_WeatherConditions to Fact7" và chọn "Edit", trong tab Data Flow ta tạo hai **OLE DB Source** và đặt tên là "Fact6" và "Dim_WeatherConditions".



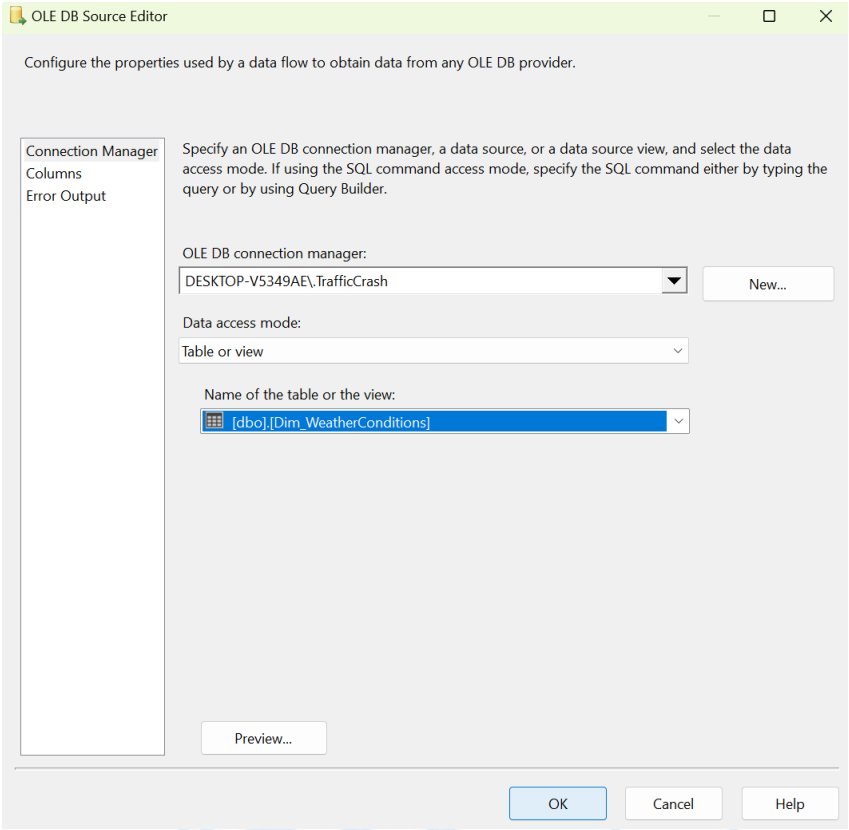
Bước 3: Nhấn chuột phải chọn "Edit", sau đó chọn bảng Fact6 đã tạo trước đó làm data source cho bảng Fact6 mới này.



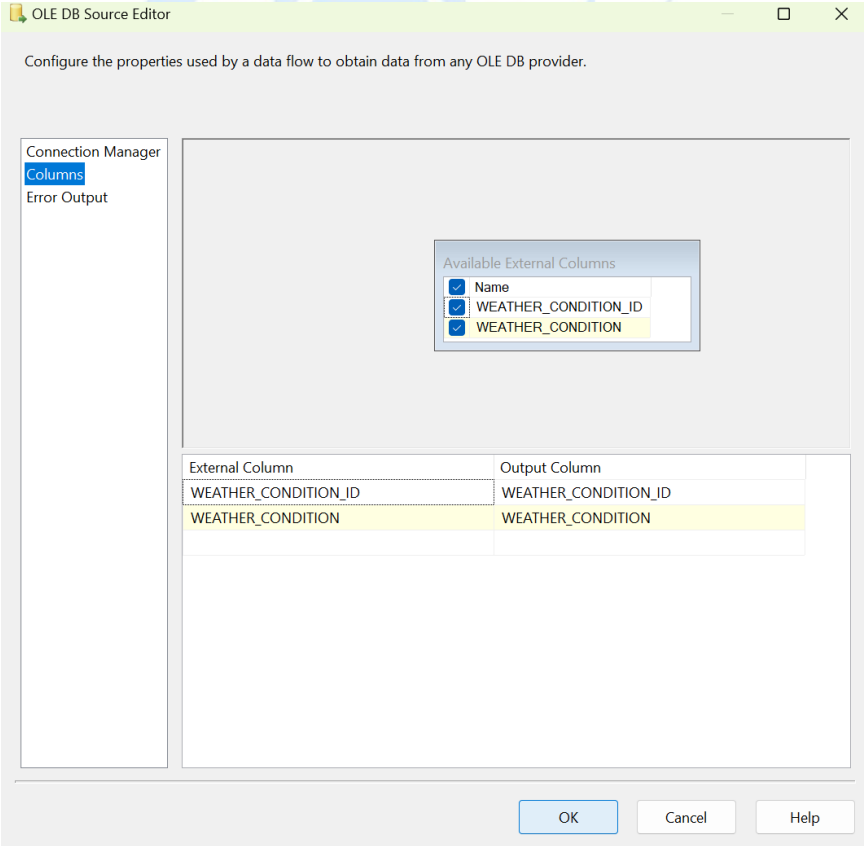
Bước 4: Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



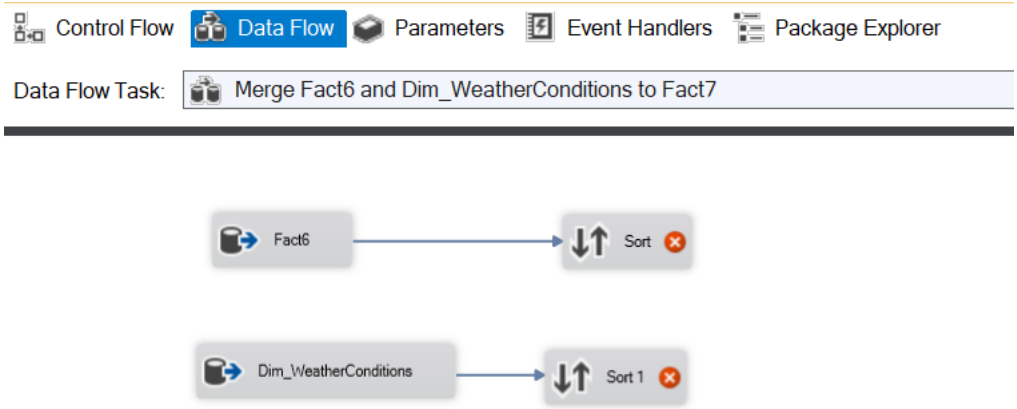
Bước 5: Tương tự thực hiện chọn bảng cho Dim_PrimaryCauses.



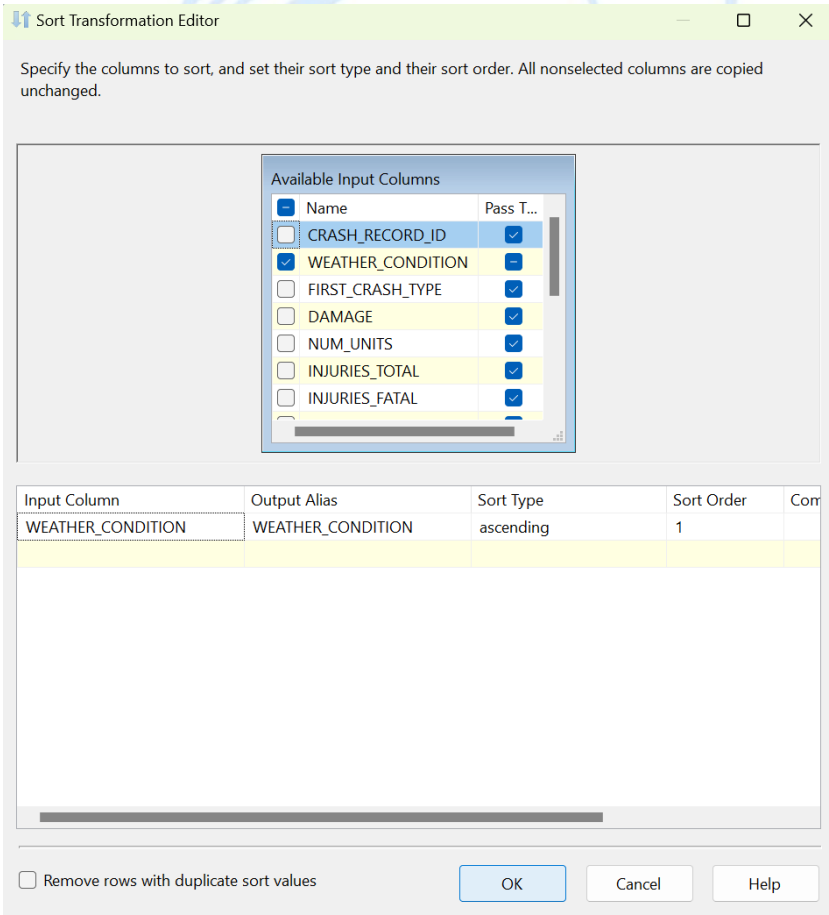
- Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



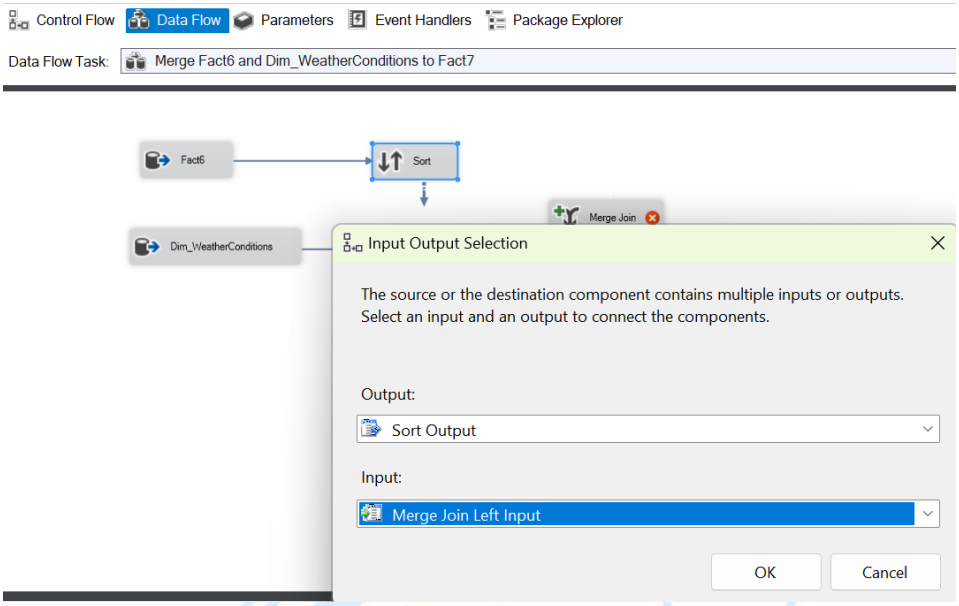
Bước 6: Tạo hai Sort đặt tên là "Sort" và "Sort1" tương ứng với mỗi "OLE DB Source" đã tạo. Thực hiện tạo các kết nối.



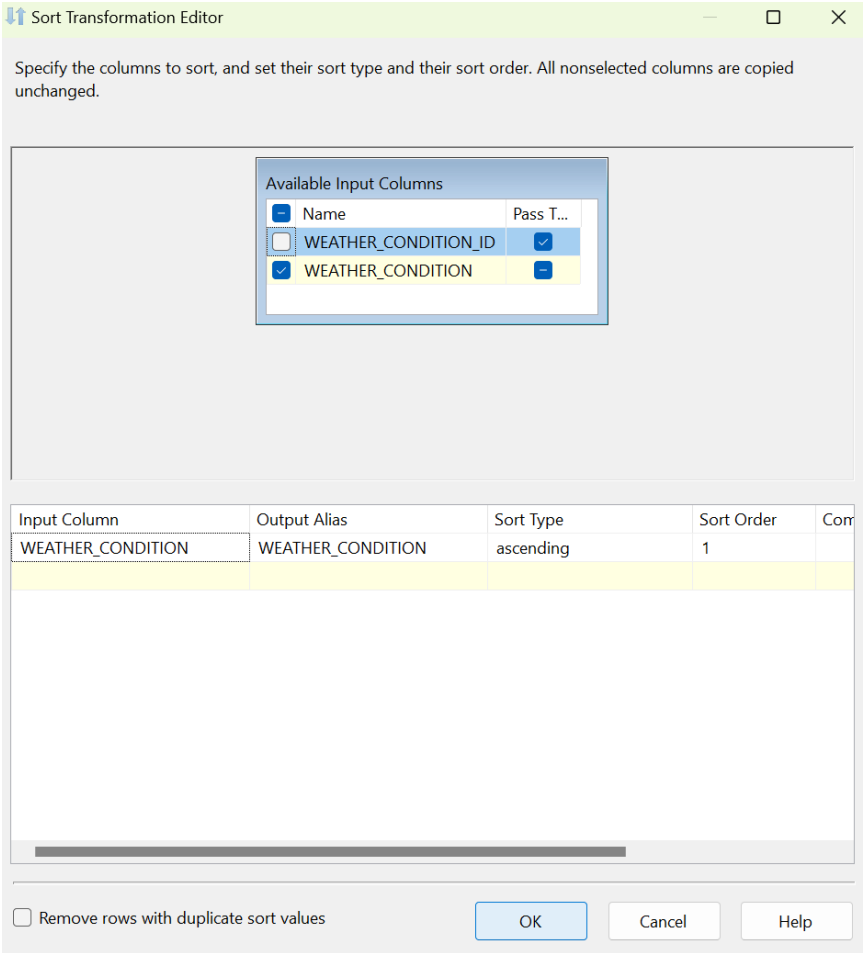
Bước 7: Ở "Sort", nhấn chuột phải chọn "Edit" và chọn cột để chuẩn bị cho quá trình merge. Sau đó nhấn "OK".



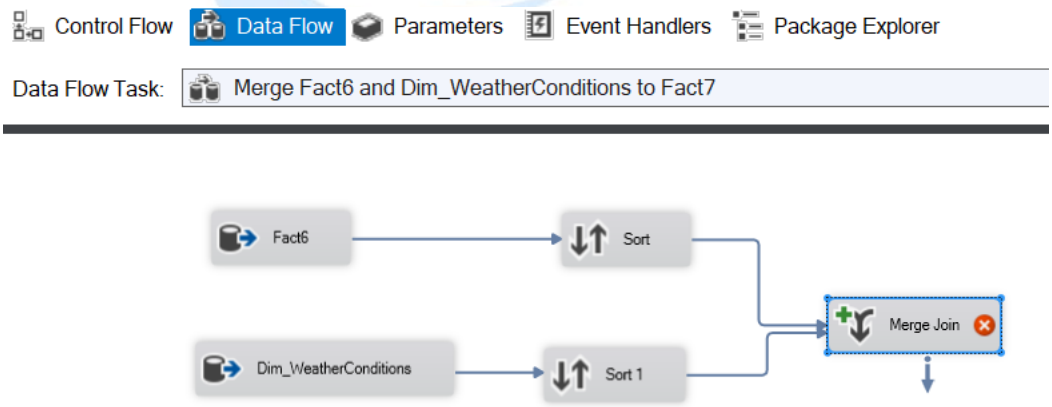
Bước 8: Tạo một Merge Join và nối với "Sort". Tiếp theo ta chọn "Merge Join Left Input" để giữ lại toàn bộ các dòng trong bảng "Fact6" bất kể có kết quả khi thực hiện phép kết trái với cột ID của bảng "Dim_PrimaryCauses" hay không.



Bước 9: Tương tự, ở "Sort1", nhấn chuột phải chọn "Edit" và chọn cột. Sau đó nhấn "OK".

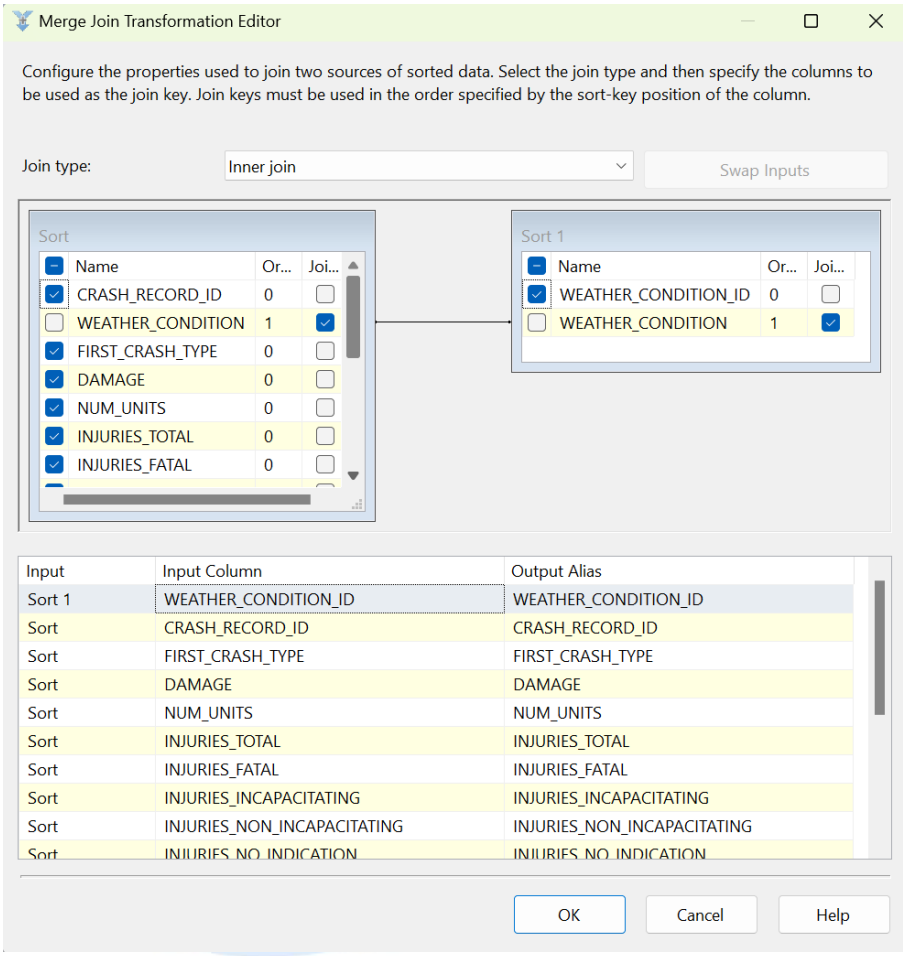


- Kết nối "Sort1" với Merge Join.



Bước 10: Đầu tiên, nhấn chuột phải vào "Merge Join" và chọn "Edit", một hộp thoại Merge Join Transformation Editor xuất hiện.

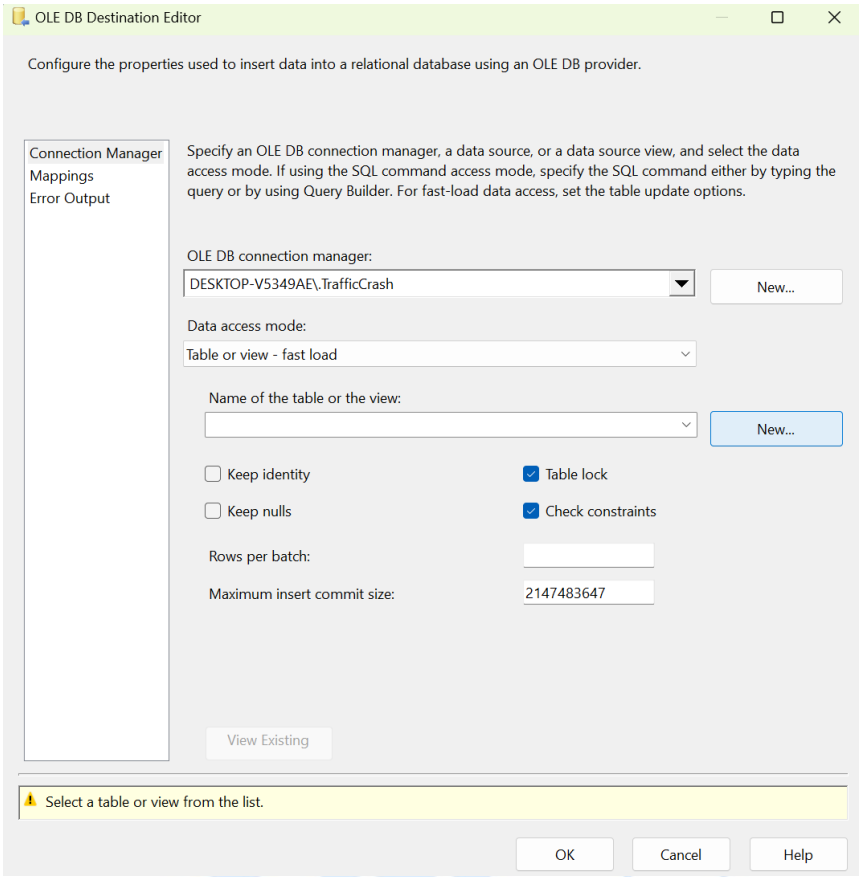
- Sau đó, ta chọn WEATHER_CONDITION_ID ở “Sort1” để merge vào “Fact6”.
- Tiếp theo, ta chọn tất cả các cột của “Sort” nhưng không lấy thuộc tính WEATHER_CONDITION.
- Nhấn “OK” để hoàn tất.



Kết quả sau khi merge là bảng Fact6 không còn thuộc tính WEATHER_CONDITION và có thêm 1 thuộc tính mới là WEATHER_CONDITION_ID.

Bước 11: Tạo một OLE DB Destination đặt tên là “Fact7” để chứa tất cả những gì đã merge.

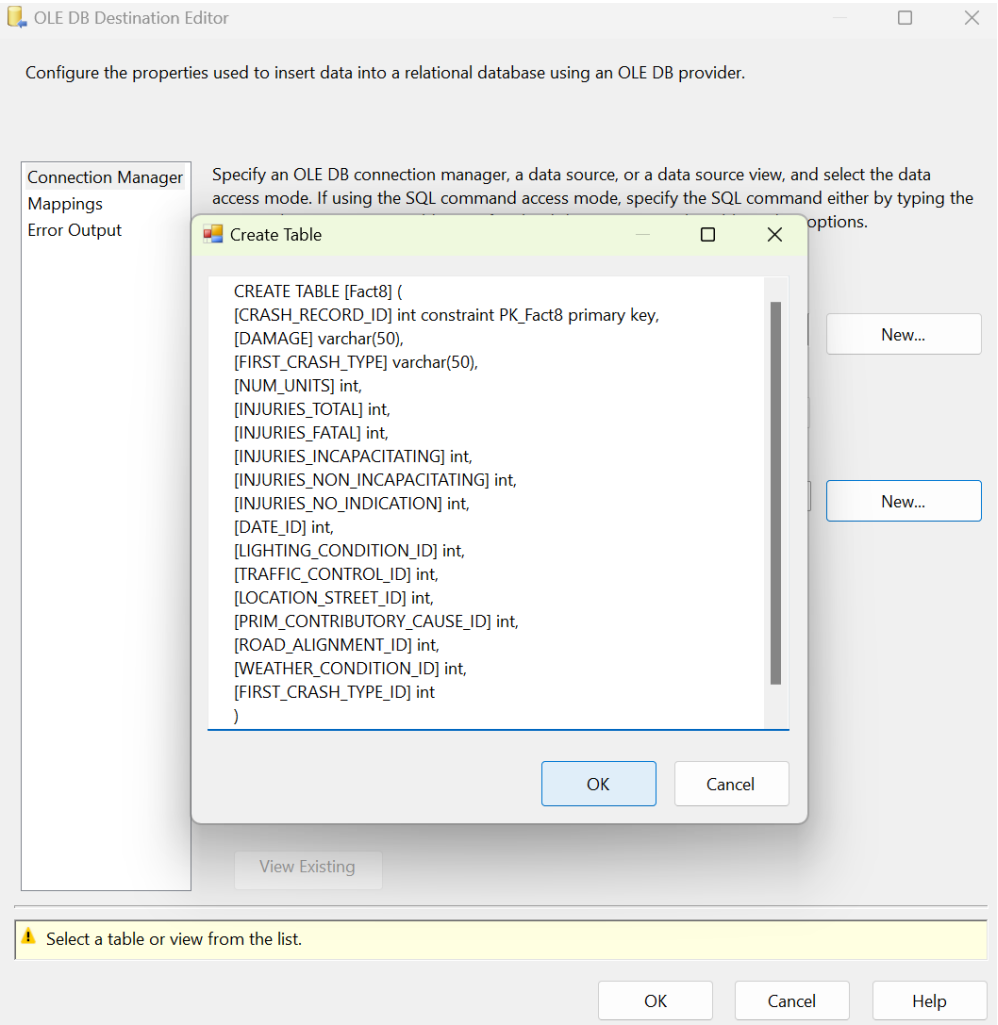
- Nhấn chuột phải vào “Fact7”, chọn “Edit”. Chọn “New” để tạo bảng.



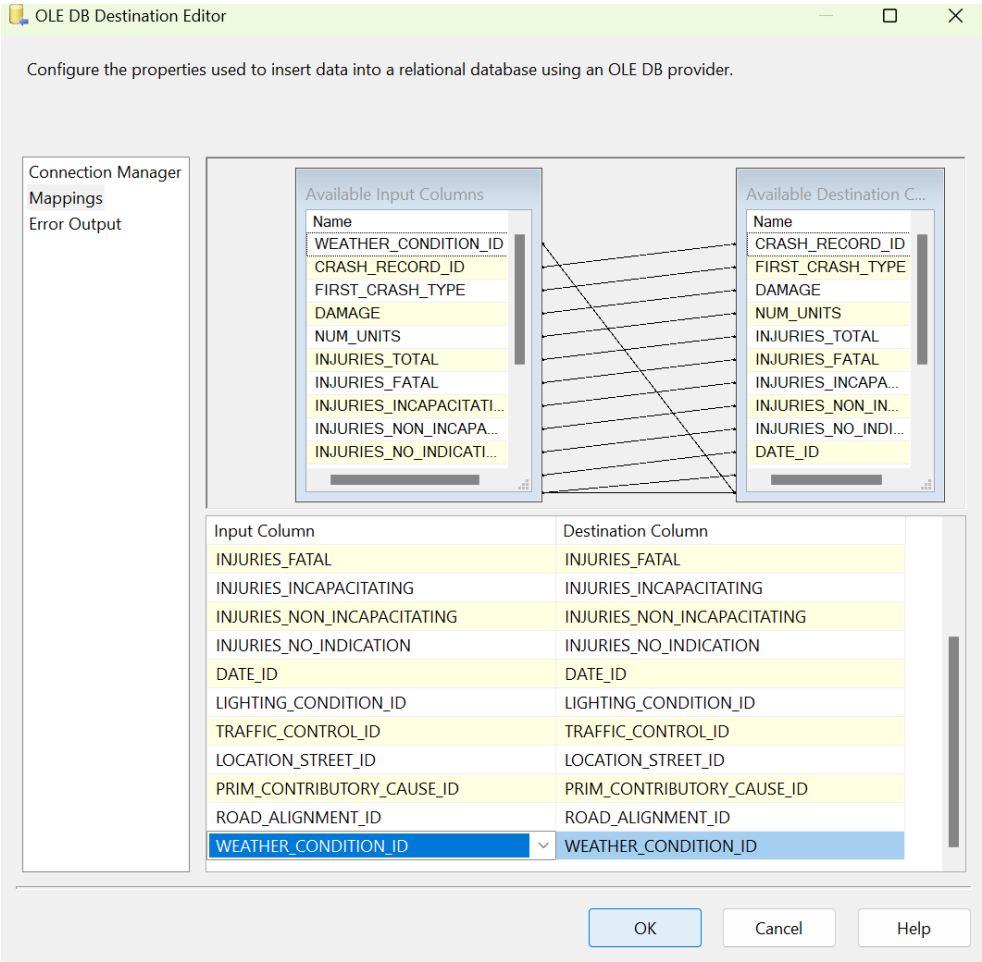
- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Fact7] (
[CRASH_RECORD_ID] int constraint PK_Fact7 primary key,
[FIRST_CRASH_TYPE] varchar(50),
[DAMAGE] varchar(50),
[NUM_UNITS] int,
[INJURIES_TOTAL] int,
[INJURIES_FATAL] int,
[INJURIES_INCAPACITATING] int,
[INJURIES_NON_INCAPACITATING] int,
[INJURIES_NO_INDICATION] int,
[DATE_ID] int,
[LIGHTING_CONDITION_ID] int,
```

```
[TRAFFIC_CONTROL_ID] int,
[LOCATION_STREET_ID] int,
[PRIM_CONTRIBUTORY_CAUSE_ID] int,
[ROAD_ALIGNMENT_ID] int,
[WEATHER_CONDITION_ID] int
)
```



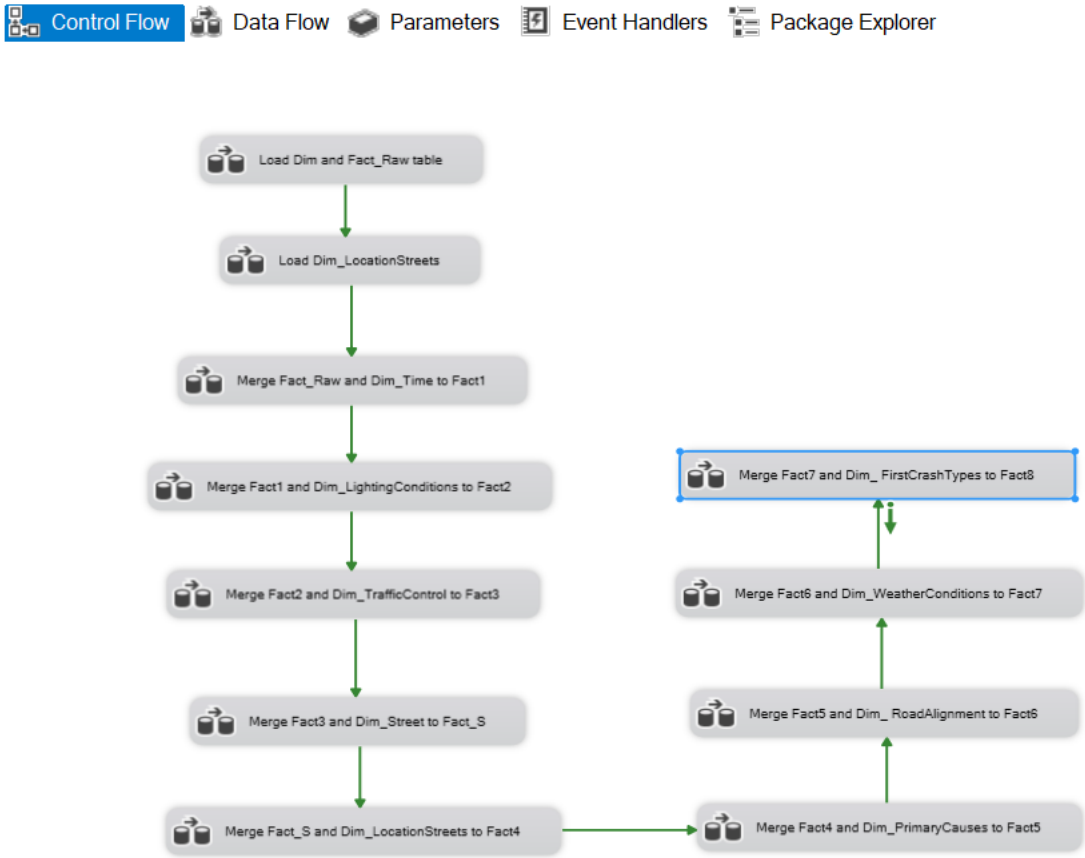
- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



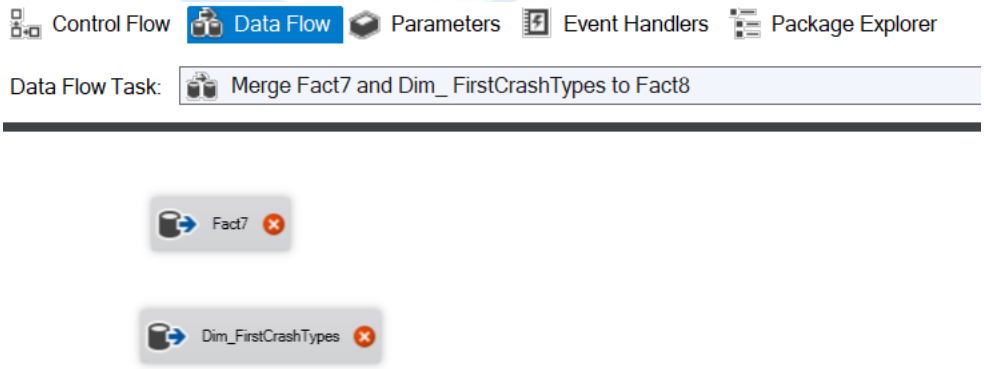
Tiếp tục quá trình merge bảng Fact7 với các Dimension còn lại để có 1 bảng Fact_TrafficCrashRecords hoàn chỉnh.

2.3.13.9 Merge Fact7 và Dim_FirstCrashTypes vào Fact8.

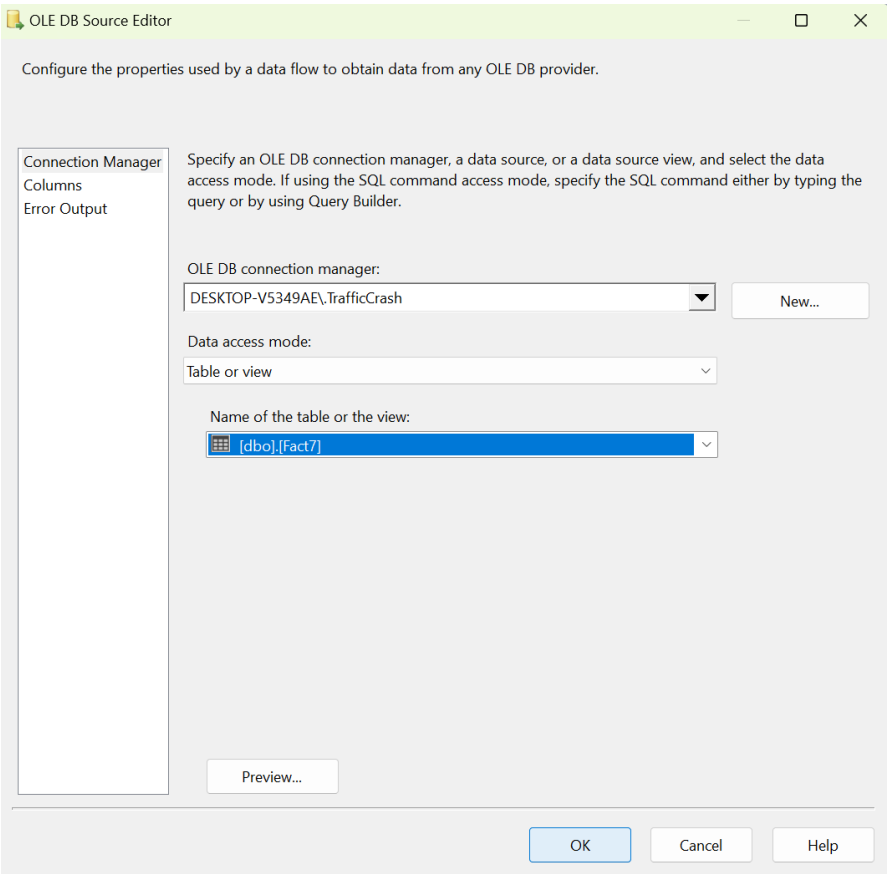
Bước 1: Ở tab Control Flow, tạo một Data Flow Task và đặt tên là “Merge Fact7 and Dim_ FirstCrashTypes to Fact8”.



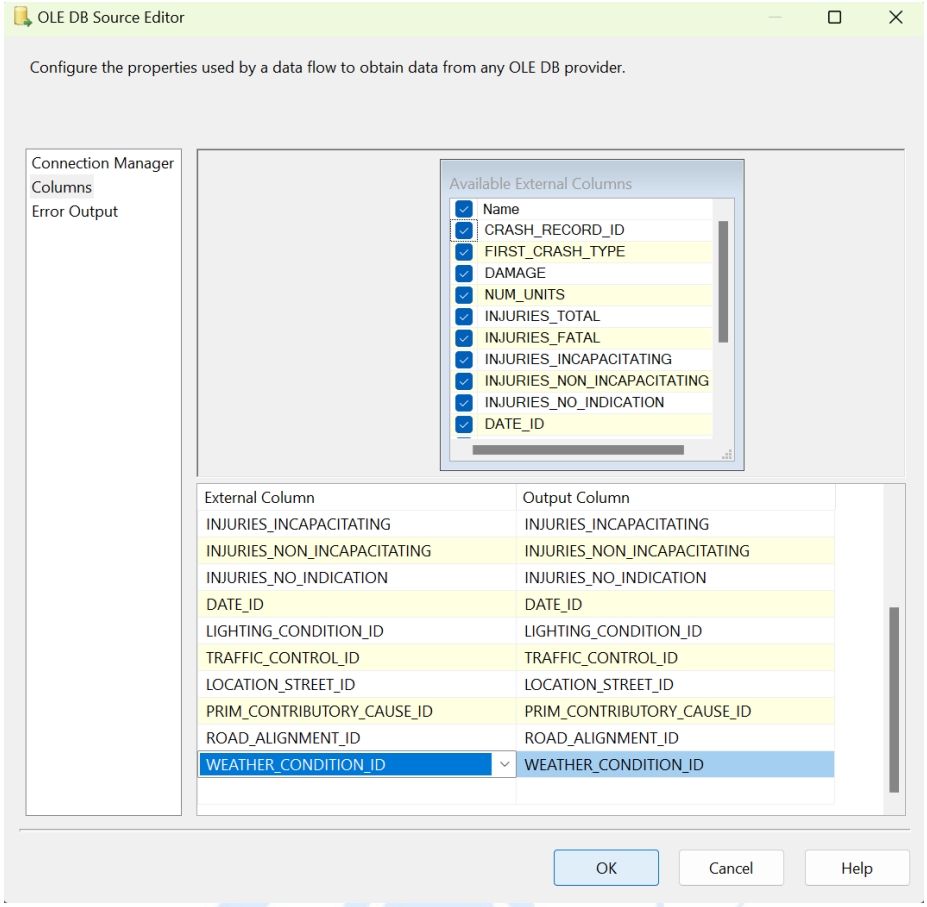
Bước 2: Nhấn chuột phải vào "Merge Fact7 and Dim_FirstCrashTypes to Fact8" và chọn "Edit", trong tab Data Flow ta tạo hai **OLE DB Source** và đặt tên là "Fact7" và "Dim_FirstCrashTypes".



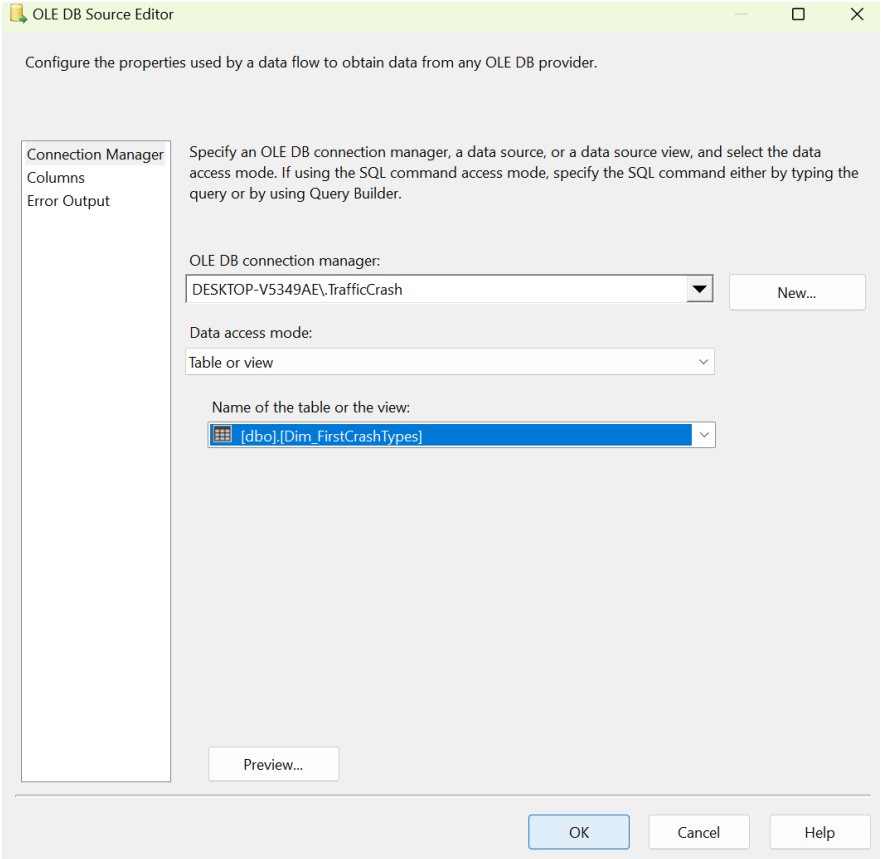
Bước 3: Nhấn chuột phải chọn "Edit", sau đó chọn bảng Fact7 đã tạo trước đó làm data source cho bảng Fact7 mới này.



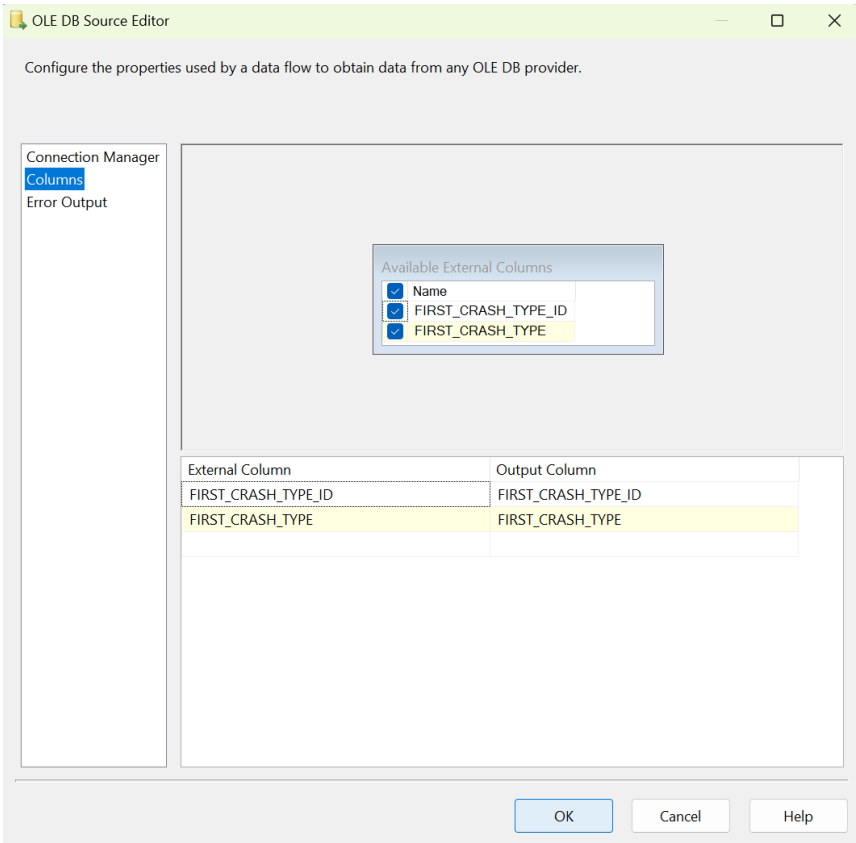
Bước 4: Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



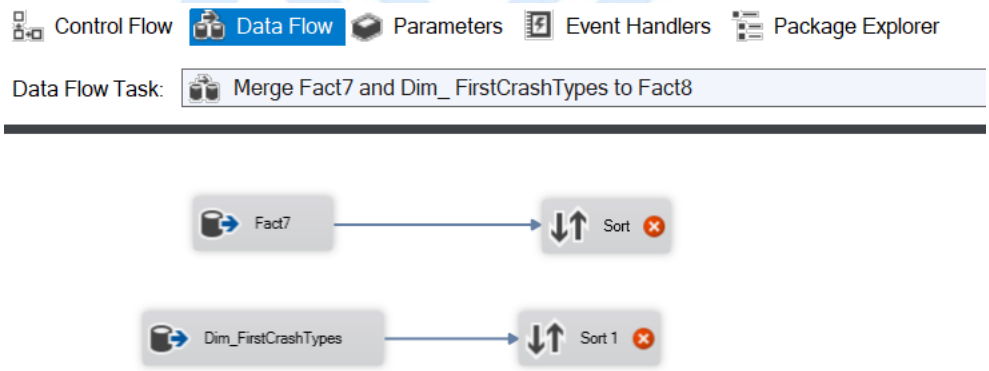
Bước 5: Tương tự thực hiện chọn bảng cho "Dim_FirstCrashTypes".



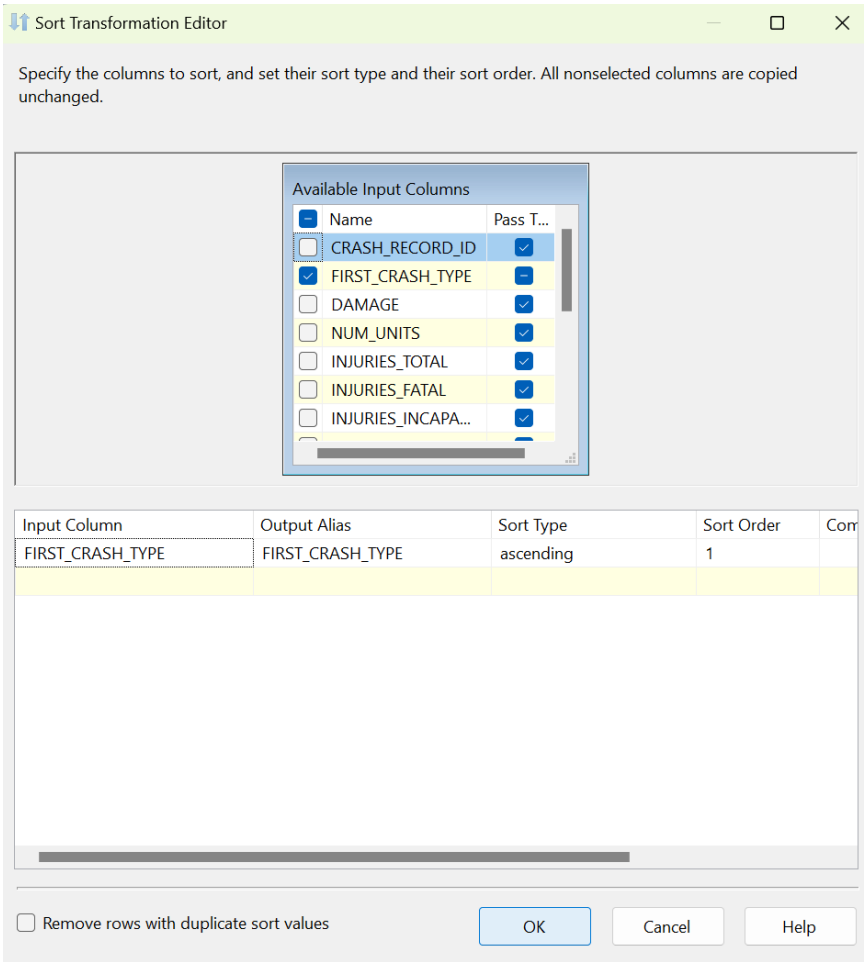
- Chọn mục “Columns” để xem xét các cột được ánh xạ. Sau đó nhấn “OK”.



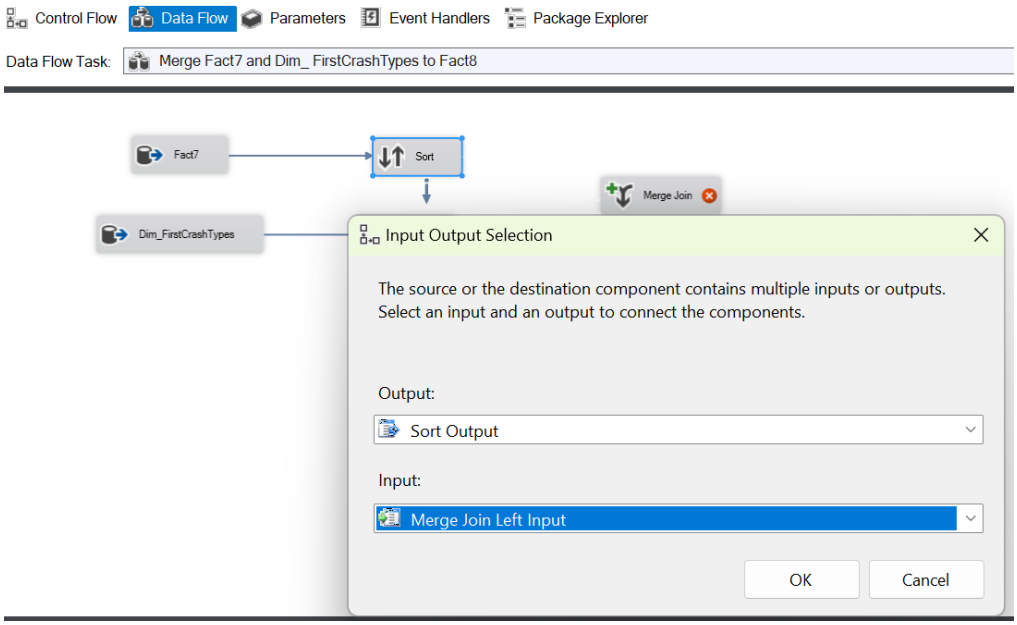
Bước 6: Tạo hai Sort đặt tên là "Sort" và "Sort1" tương ứng với mỗi "OLE DB Source" đã tạo. Thực hiện tạo các kết nối.



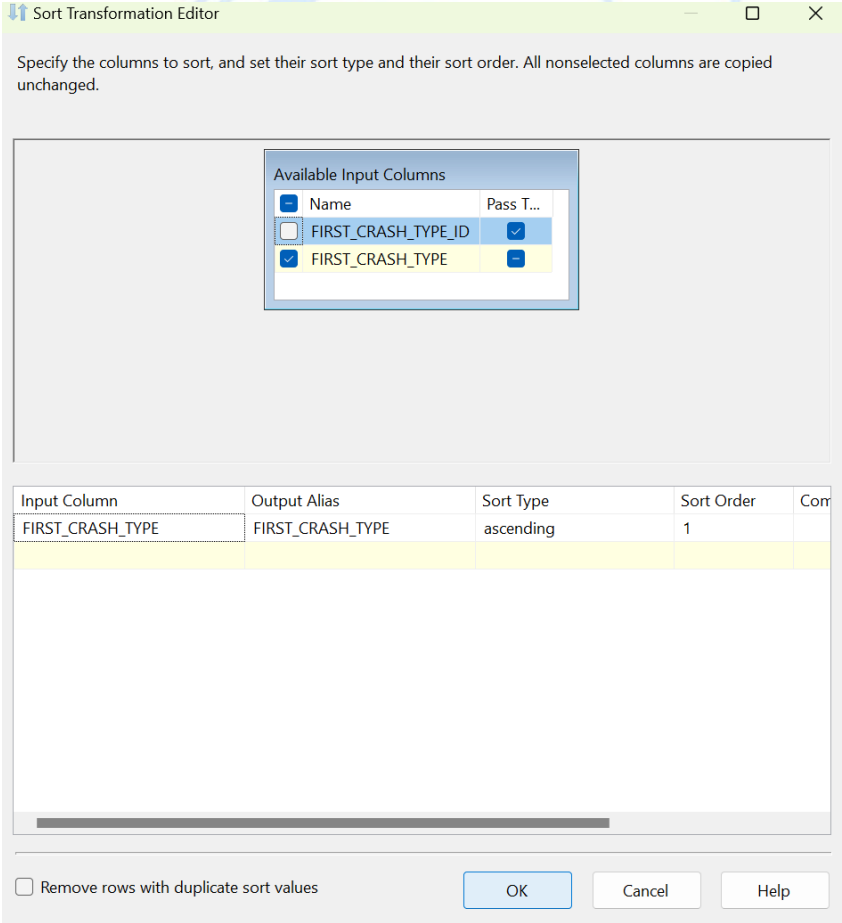
Bước 7: Ở "Sort", nhấn chuột phải chọn "Edit" và chọn cột để chuẩn bị cho quá trình merge. Sau đó nhấn "OK".



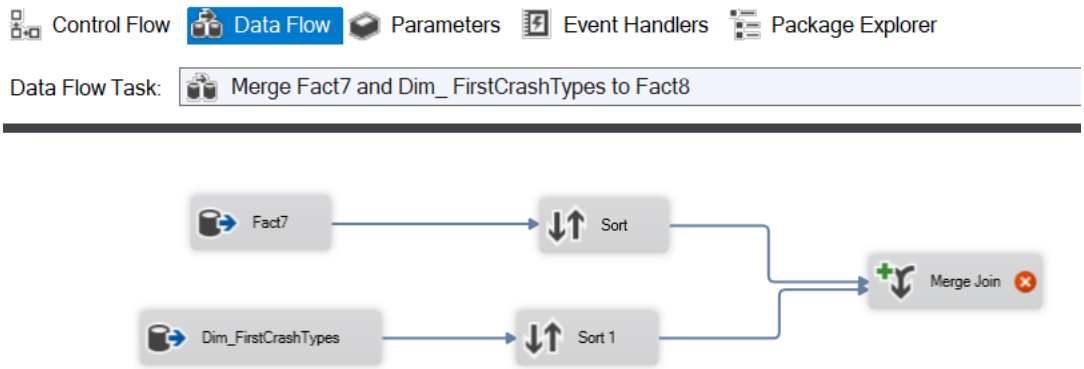
Bước 8: Tạo một Merge Join và nối với "Sort". Tiếp theo ta chọn "Merge Join Left Input" để giữ lại toàn bộ các dòng trong bảng "Fact7" bất kể có kết quả khi thực hiện phép kết trái với cột ID của bảng "Dim_FirstCrashTypes" hay không.



Bước 9: Tương tự, ở "Sort1", nhấn chuột phải chọn "Edit" và chọn cột. Sau đó nhấn "OK".

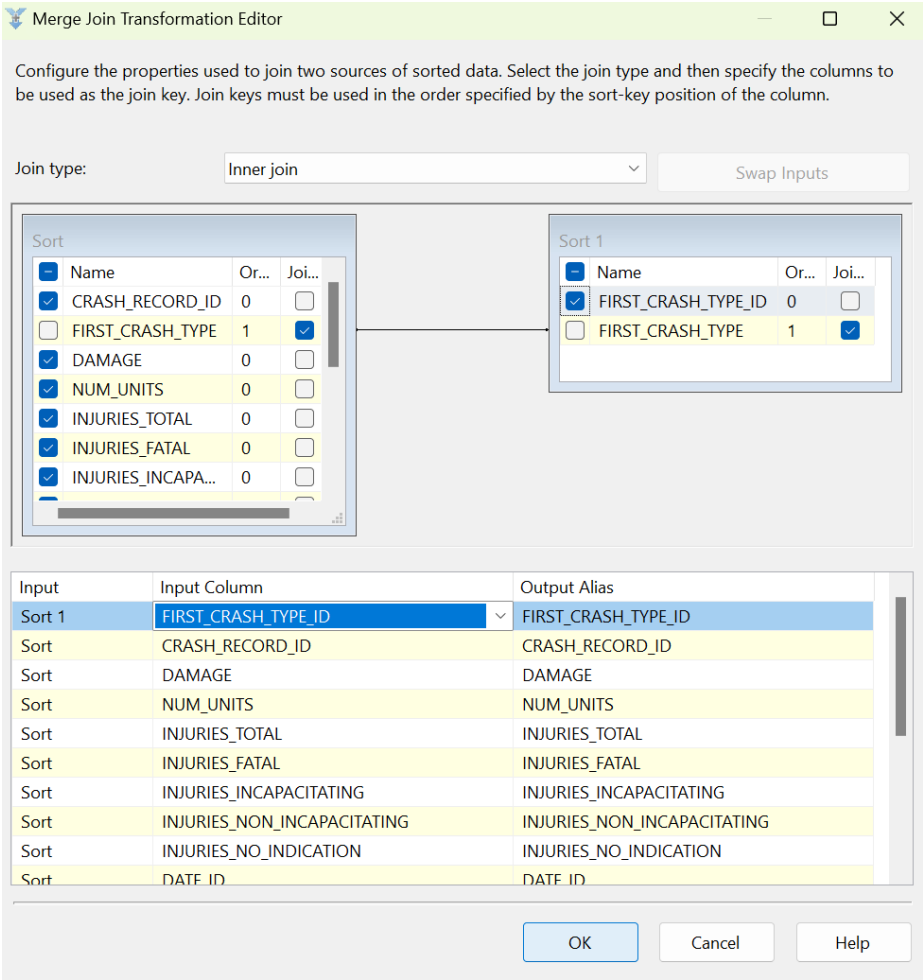


- Kết nối “Sort1” với “Merge Join”.



Bước 10: Đầu tiên, nhấn chuột phải vào “Merge Join” và chọn “Edit”, một hộp thoại Merge Join Transformation Editor xuất hiện.

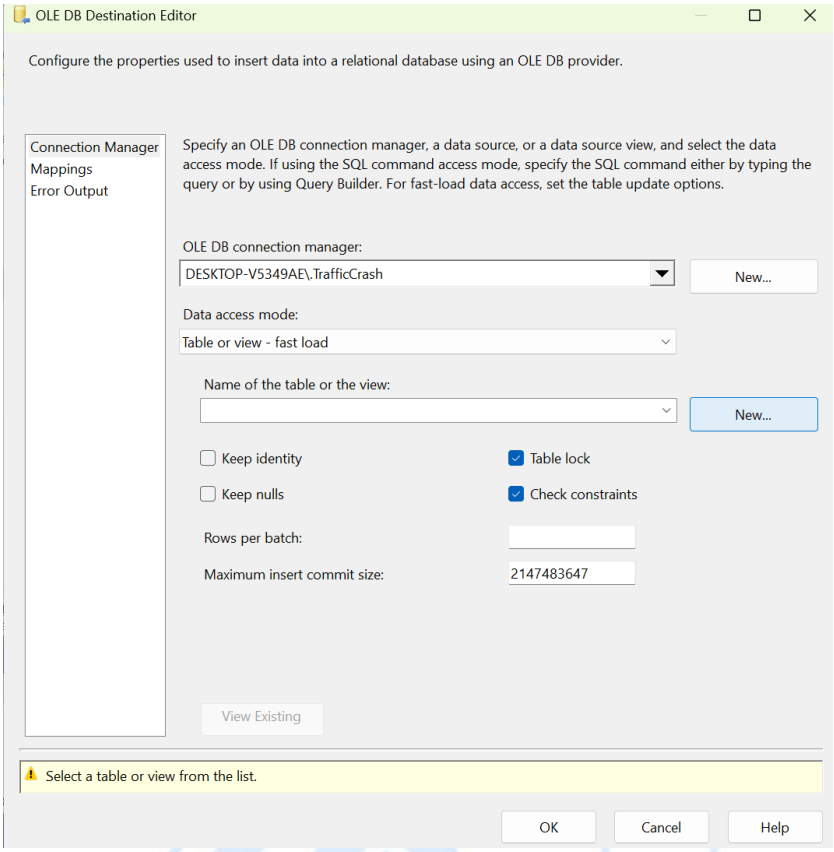
- Sau đó, ta chọn FIRST_CRASH_TYPE_ID ở “Sort1” để merge vào “Fact7”.
- Tiếp theo, ta chọn tất cả các cột của “Sort” nhưng không lấy thuộc tính FIRST_CRASH_TYPE.
- Nhấn “OK” để hoàn tất.



Kết quả sau khi merge là bảng Fact7 không còn thuộc tính FIRST_CRASH_TYPE và có thêm 1 thuộc tính mới là FIRST_CRASH_TYPE_ID.

Bước 11: Tạo một OLE DB Destination đặt tên là “Fact8” để chứa tất cả những gì đã merge.

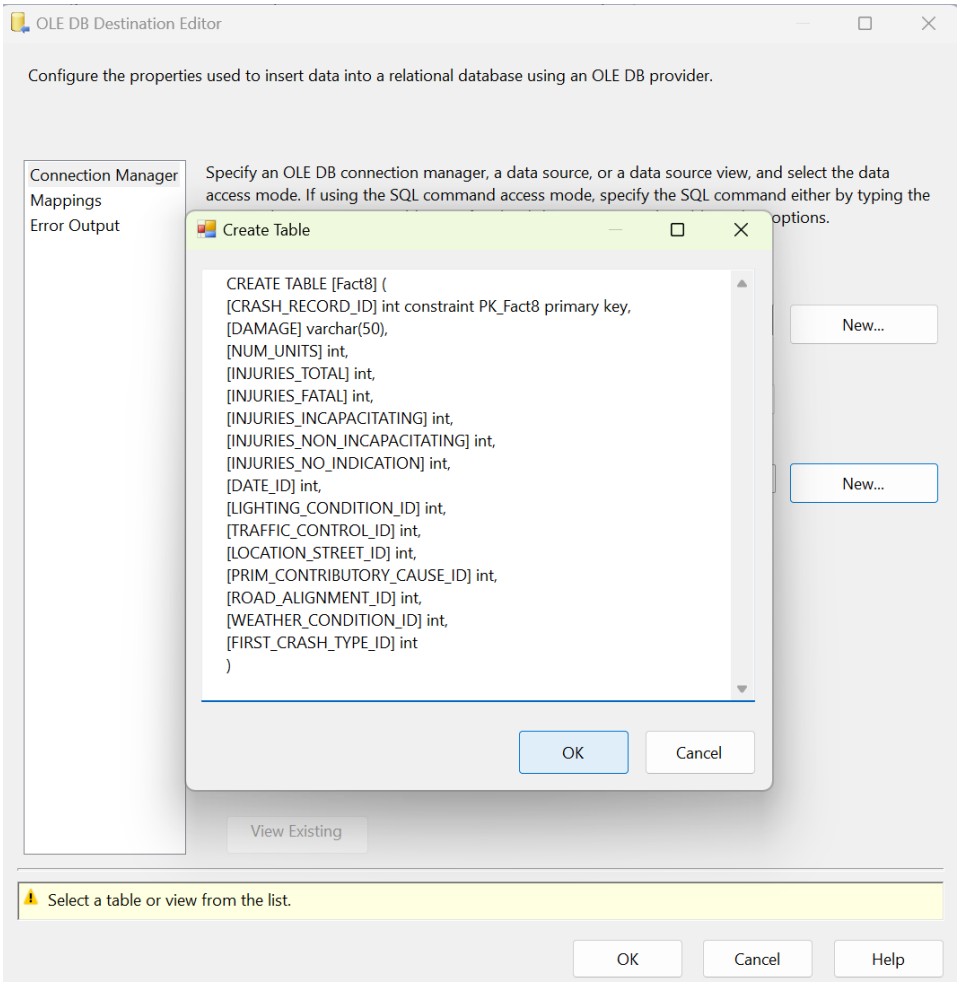
- Nhấn chuột phải vào “Fact8”, chọn “Edit”. Chọn “New” để tạo bảng.



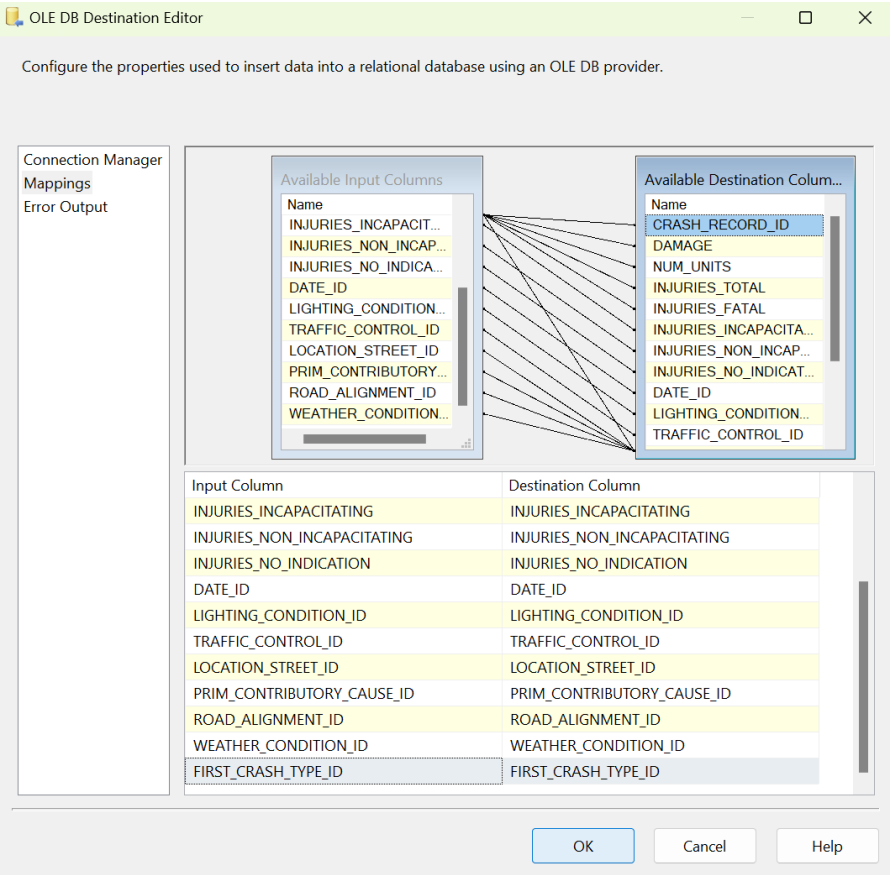
- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Fact8] (
[CRASH_RECORD_ID] int constraint PK_Fact8 primary key,
[DAMAGE] varchar(50),
[NUM_UNITS] int,
[INJURIES_TOTAL] int,
[INJURIES_FATAL] int,
[INJURIES_INCAPACITATING] int,
[INJURIES_NON_INCAPACITATING] int,
[INJURIES_NO_INDICATION] int,
[DATE_ID] int,
[LIGHTING_CONDITION_ID] int,
[TRAFFIC_CONTROL_ID] int,
[LOCATION_STREET_ID] int,
```

```
[PRIM_CONTRIBUTORY_CAUSE_ID] int,
[ROAD_ALIGNMENT_ID] int,
[WEATHER_CONDITION_ID] int,
[FIRST_CRASH_TYPE_ID] int
)
```



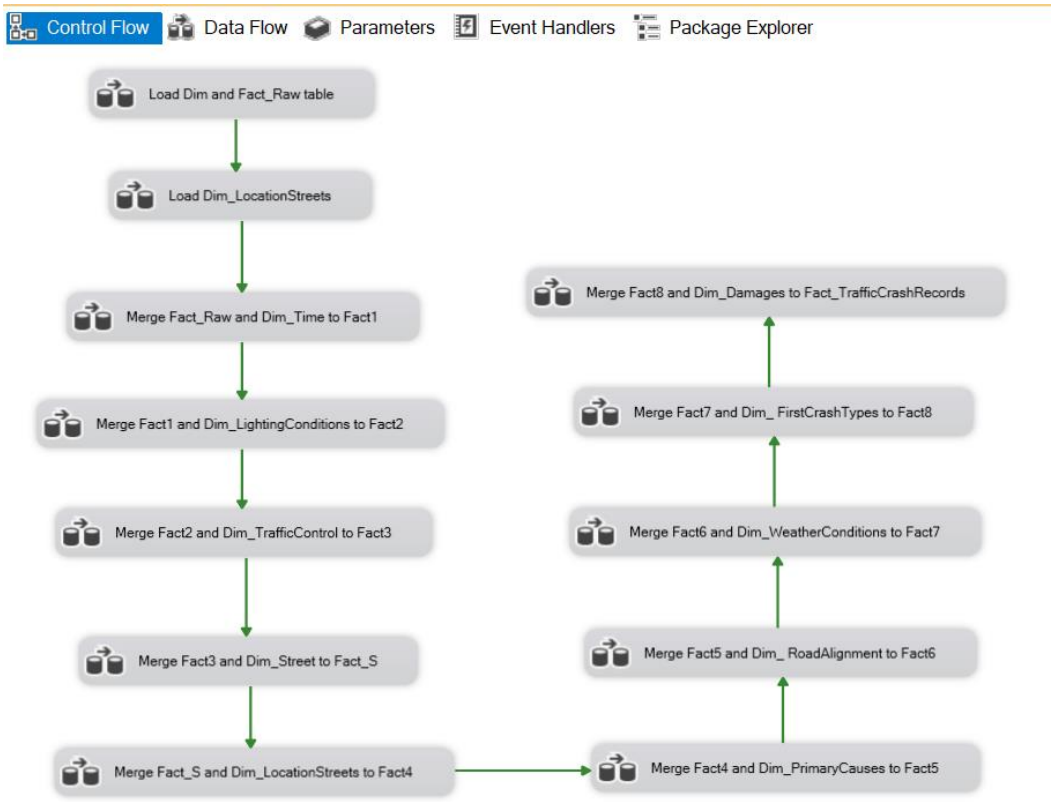
- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



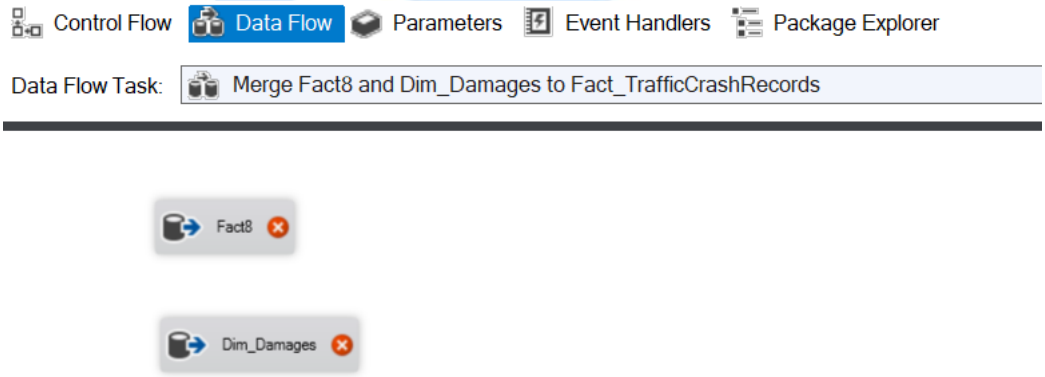
Tiếp tục quá trình merge bảng Fact8 với Dimension còn lại để có bảng Fact_TrafficCrashRecords hoàn chỉnh.

2.3.13.10 Merge Fact8 và Dim_Damages vào Fact_TrafficCrashRecords.

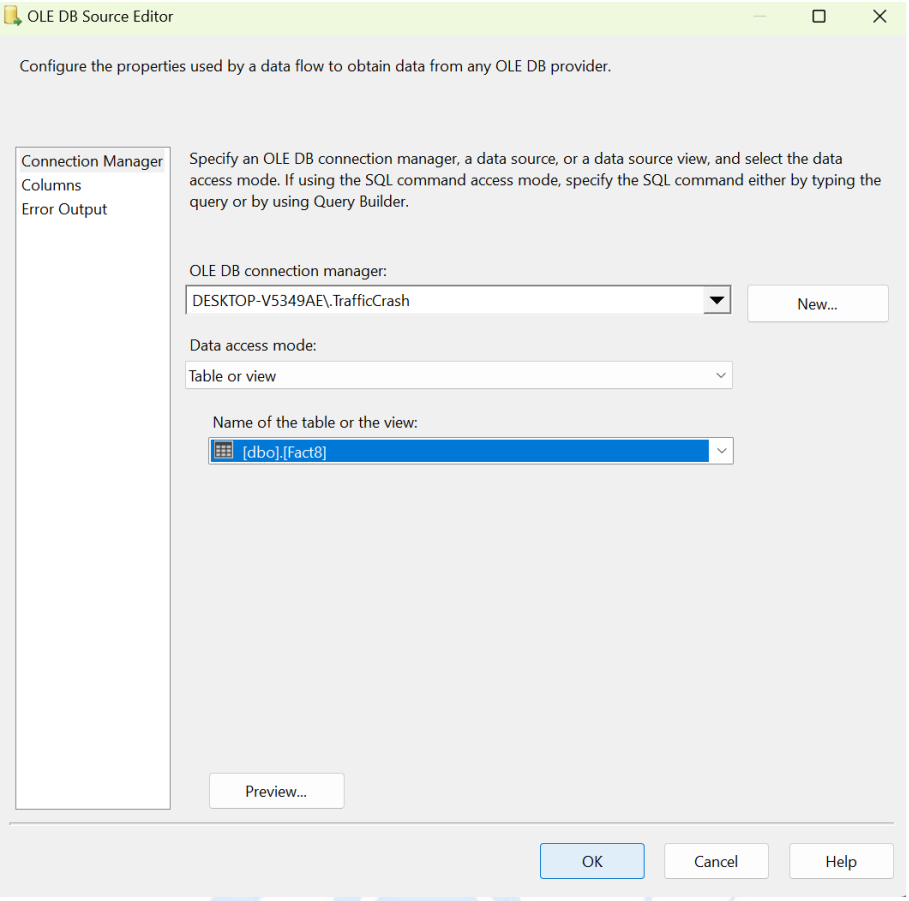
Bước 1: Ở tab Control Flow, tạo một Data Flow Task và đặt tên là “Merge Fact8 and Dim_Damages to Fact_TrafficCrashRecords”.



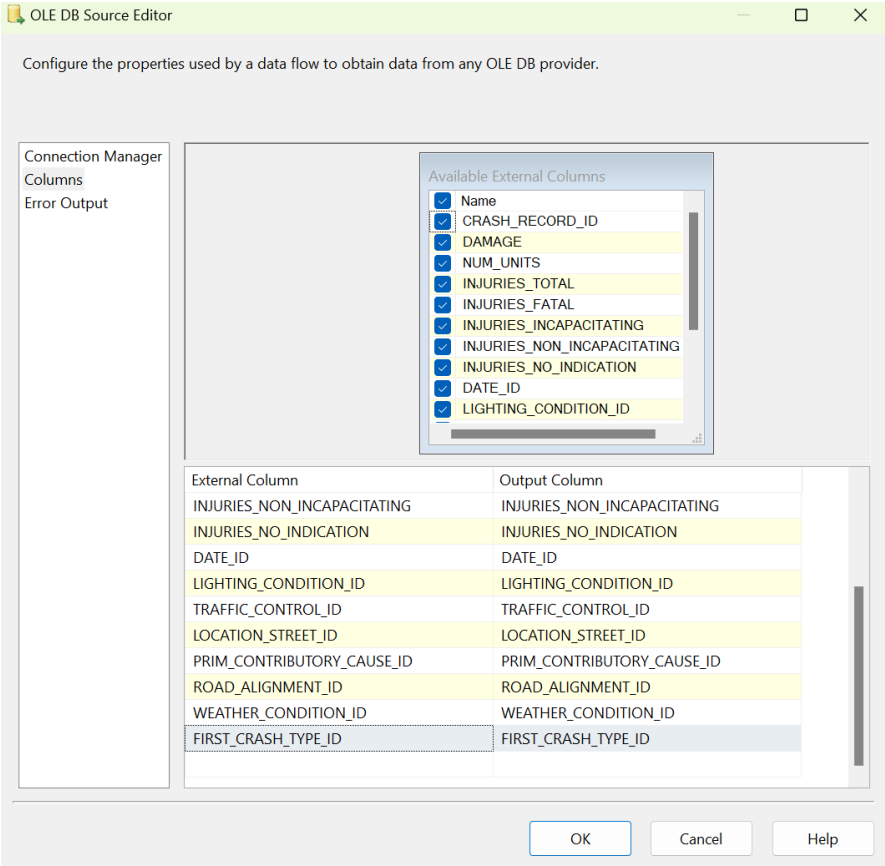
Bước 2: Nhấn chuột phải vào "Merge Fact8 and Dim_Damages to Fact_TrafficCrashRecords" và chọn "Edit", trong tab Data Flow ta tạo hai **OLE DB Source** và đặt tên là "Fact8" và "Dim_Damages".



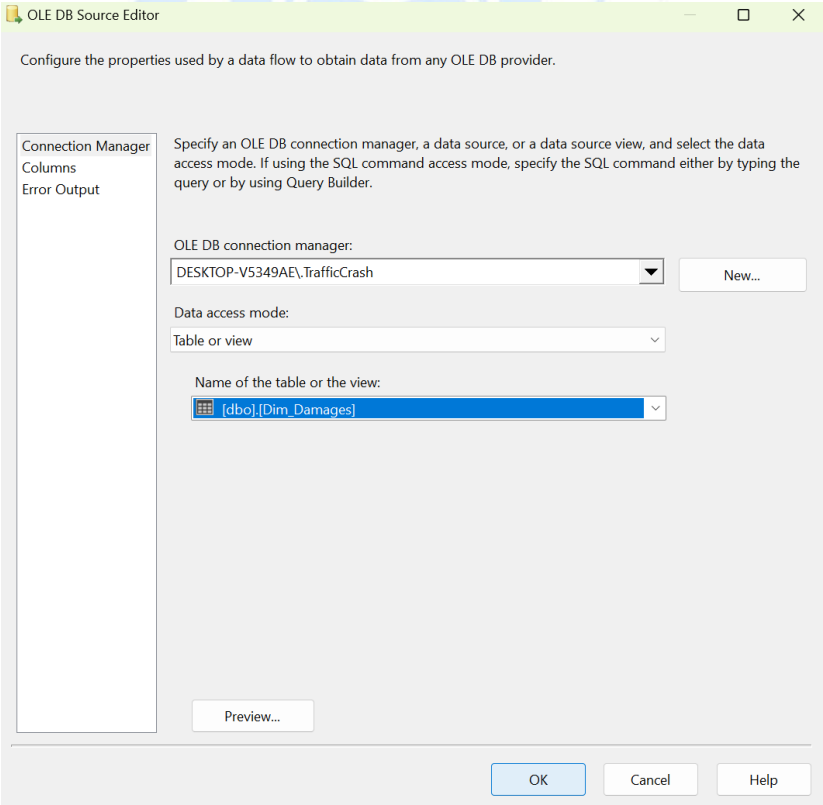
Bước 3: Nhấn chuột phải chọn "Edit", sau đó chọn bảng Fact8 đã tạo trước đó làm data source cho bảng Fact8 mới này.



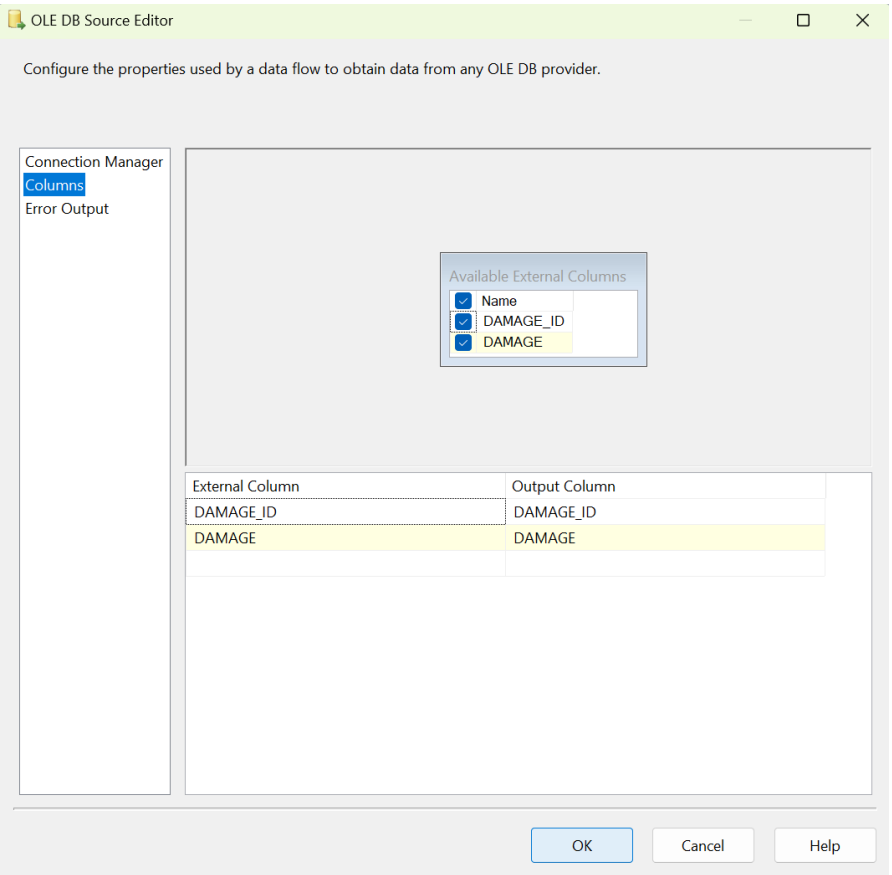
Bước 4: Chọn mục "Columns" để xem xét các cột được ánh xạ. Sau đó nhấn "OK".



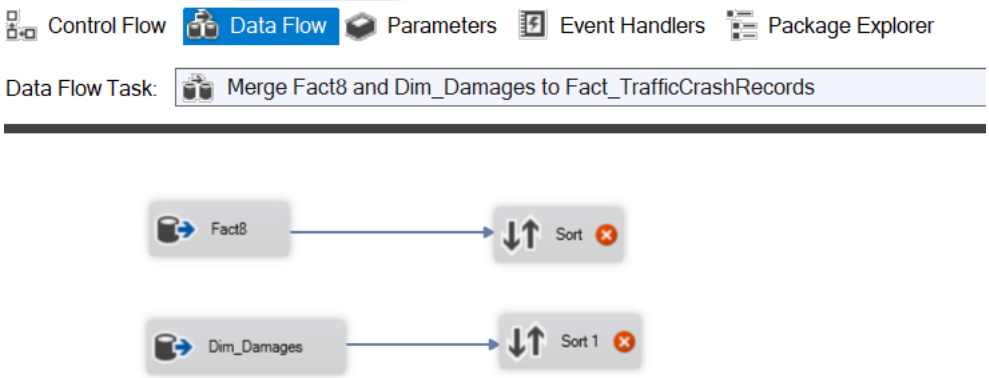
Bước 5: Tương tự thực hiện chọn bảng cho "Dim_Damages".



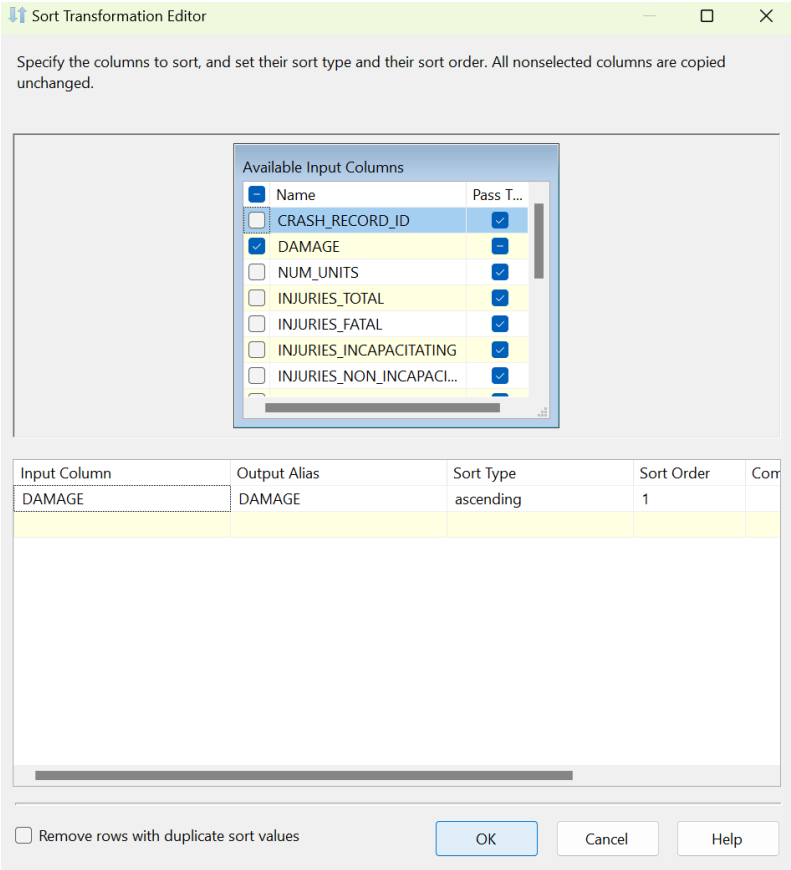
- Chọn mục “Columns” để xem xét các cột được ánh xạ. Sau đó nhấn “OK”.



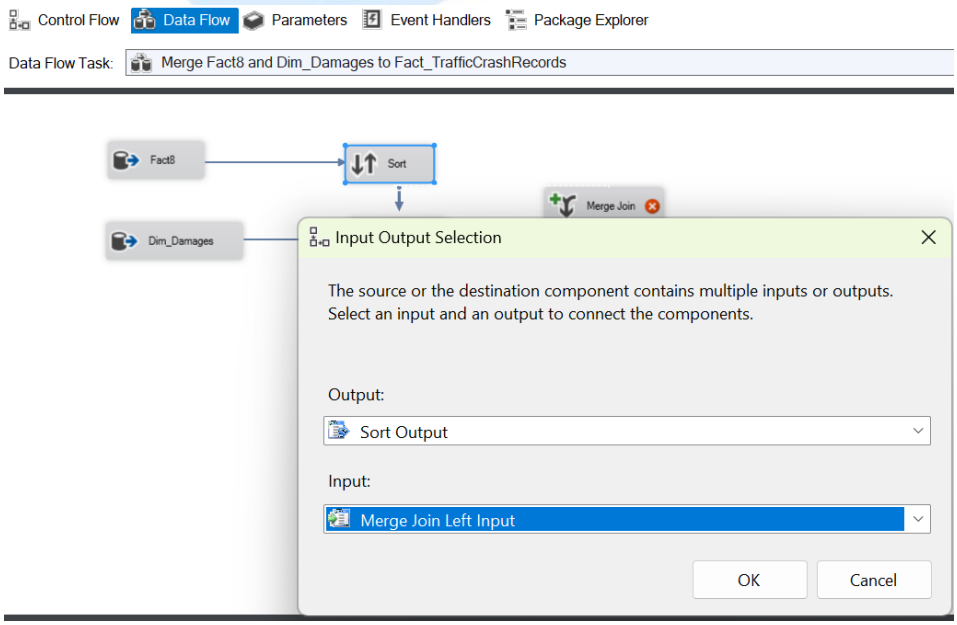
Bước 6: Tạo hai Sort đặt tên là “Sort” và “Sort1” tương ứng với mỗi “OLE DB Source” đã tạo. Thực hiện tạo các kết nối.



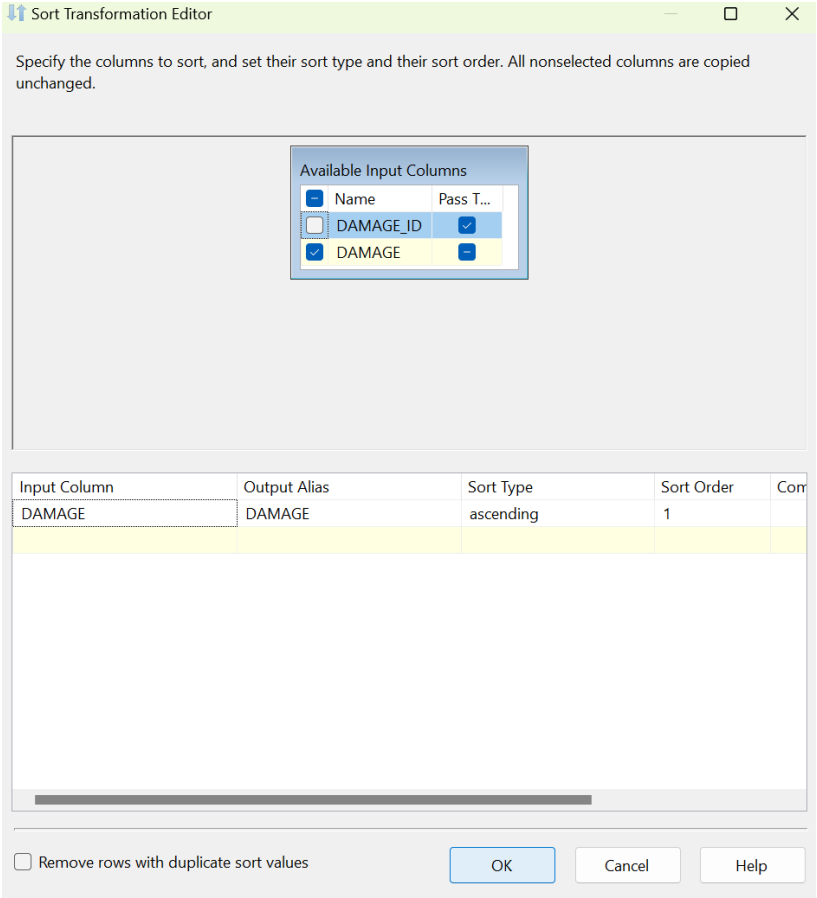
Bước 7: Ở “Sort”, nhấn chuột phải chọn “Edit” và chọn cột để chuẩn bị cho quá trình merge. Sau đó nhấn “OK”.



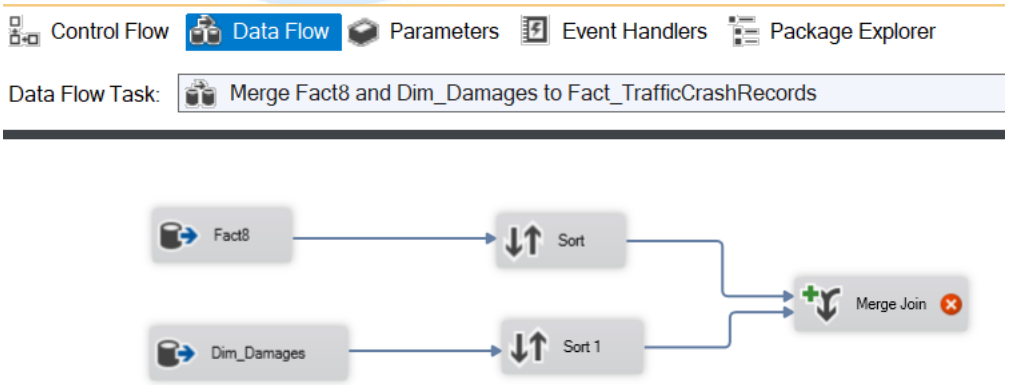
Bước 8: Tạo một Merge Join và nối với "Sort". Tiếp theo ta chọn "Merge Join Left Input" để giữ lại toàn bộ các dòng trong bảng "Fact8" bất kể có kết quả khi thực hiện phép kết trái với cột ID của bảng "Dim_Damages" hay không.



Bước 9: Tương tự, ở "Sort1", nhấn chuột phải chọn "Edit" và chọn cột. Sau đó nhấn "OK".



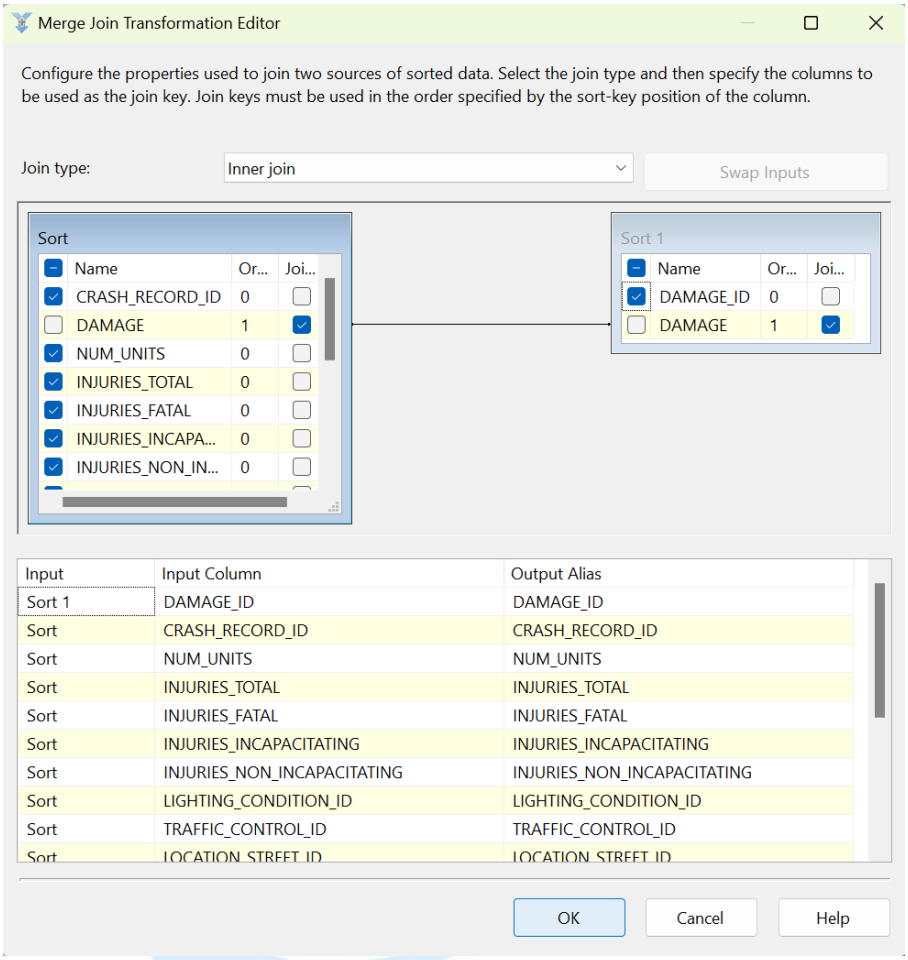
- Kết nối "Sort1" với "Merge Join".



Bước 10: Đầu tiên, nhấn chuột phải vào "Merge Join" và chọn "Edit", một hộp thoại Merge Join Transformation Editor xuất hiện.

- Sau đó, ta chọn DAMAGE_ID ở "Sort1" để merge vào "Fact8".

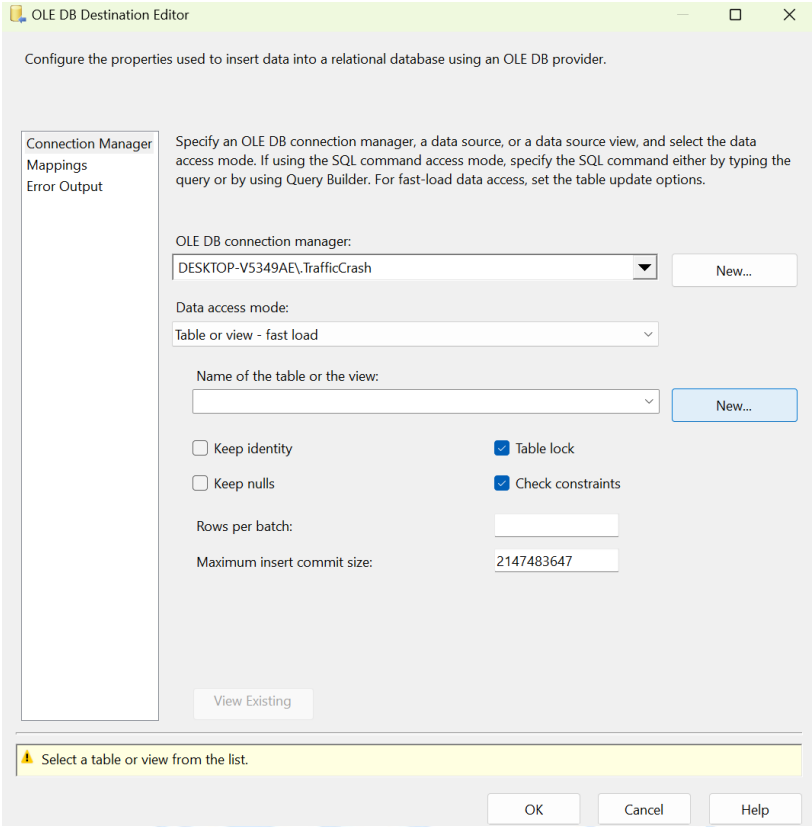
- Tiếp theo, ta chọn tất cả các cột của “Sort” nhưng không lấy thuộc tính DAMAGE.
- Nhấn “OK” để hoàn tất.



Kết quả sau khi merge là bảng Fact8 không còn thuộc tính DAMAGE và có thêm 1 thuộc tính mới là DAMAGE_ID.

Bước 11: Tạo một OLE DB Destination đặt tên là “Fact_TrafficCrashRecords” để chứa tất cả những gì đã merge.

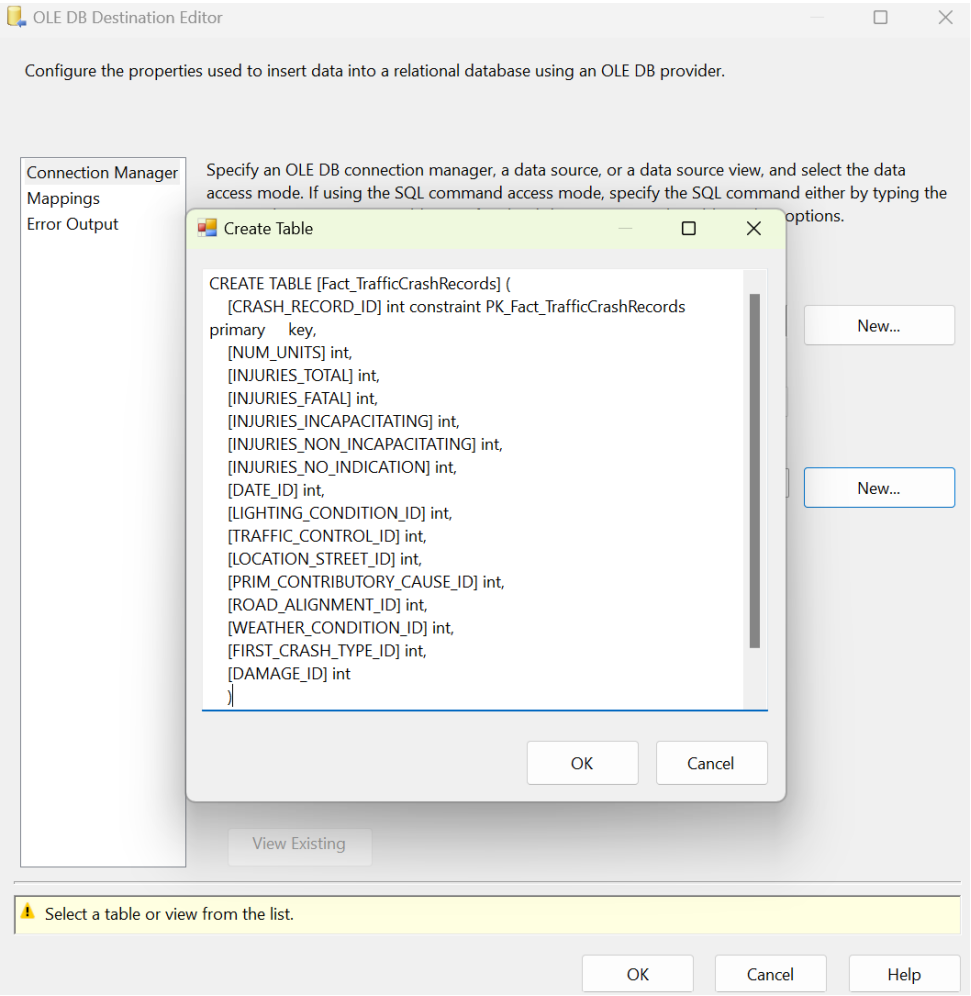
- Nhấn chuột phải vào “Fact_TrafficCrashRecords”, chọn “Edit”. Chọn “New” để tạo bảng.



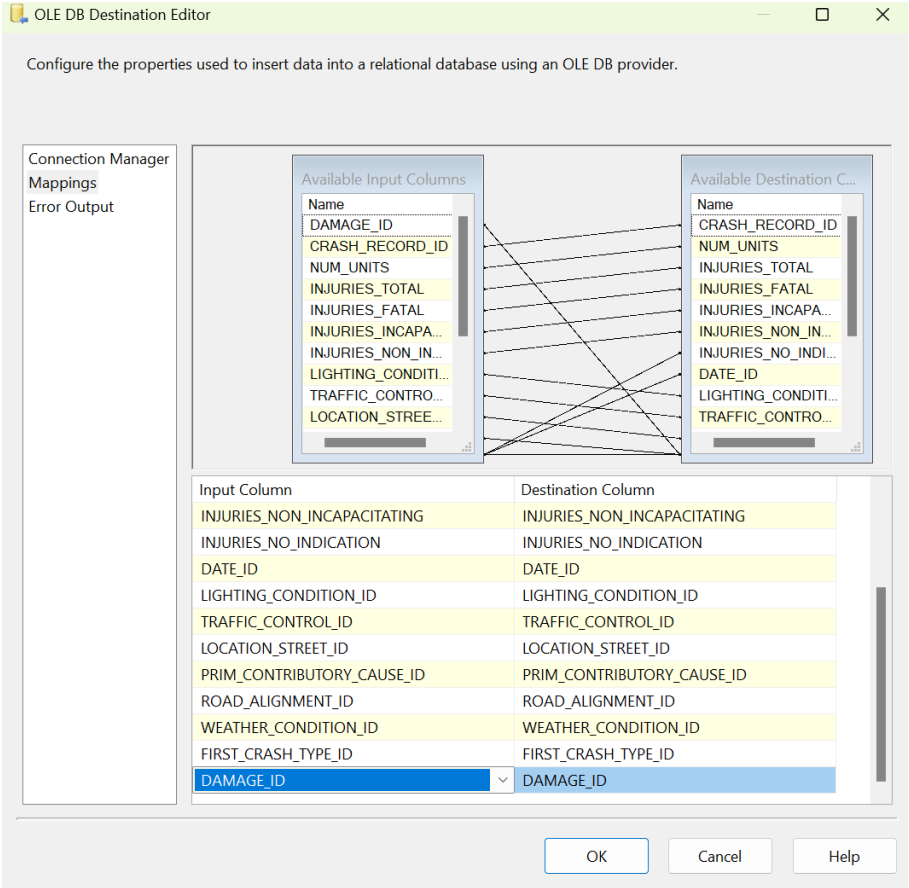
- Viết câu lệnh sql để tạo bảng. Sau đó nhấn "OK".

```
CREATE TABLE [Fact_TrafficCrashRecords] (
[CRASH_RECORD_ID] int constraint PK_Fact_TrafficCrashRecords primary key,
[NUM_UNITS] int,
[INJURIES_TOTAL] int,
[INJURIES_FATAL] int,
[INJURIES_INCAPACITATING] int,
[INJURIES_NON_INCAPACITATING] int,
[INJURIES_NO_INDICATION] int,
[DATE_ID] int,
[LIGHTING_CONDITION_ID] int,
[TRAFFIC_CONTROL_ID] int,
[LOCATION_STREET_ID] int,
```

```
[PRIM_CONTRIBUTORY_CAUSE_ID] int,
[ROAD_ALIGNMENT_ID] int,
[WEATHER_CONDITION_ID] int,
[FIRST_CRASH_TYPE_ID] int,
[DAMAGE_ID] int
)
```



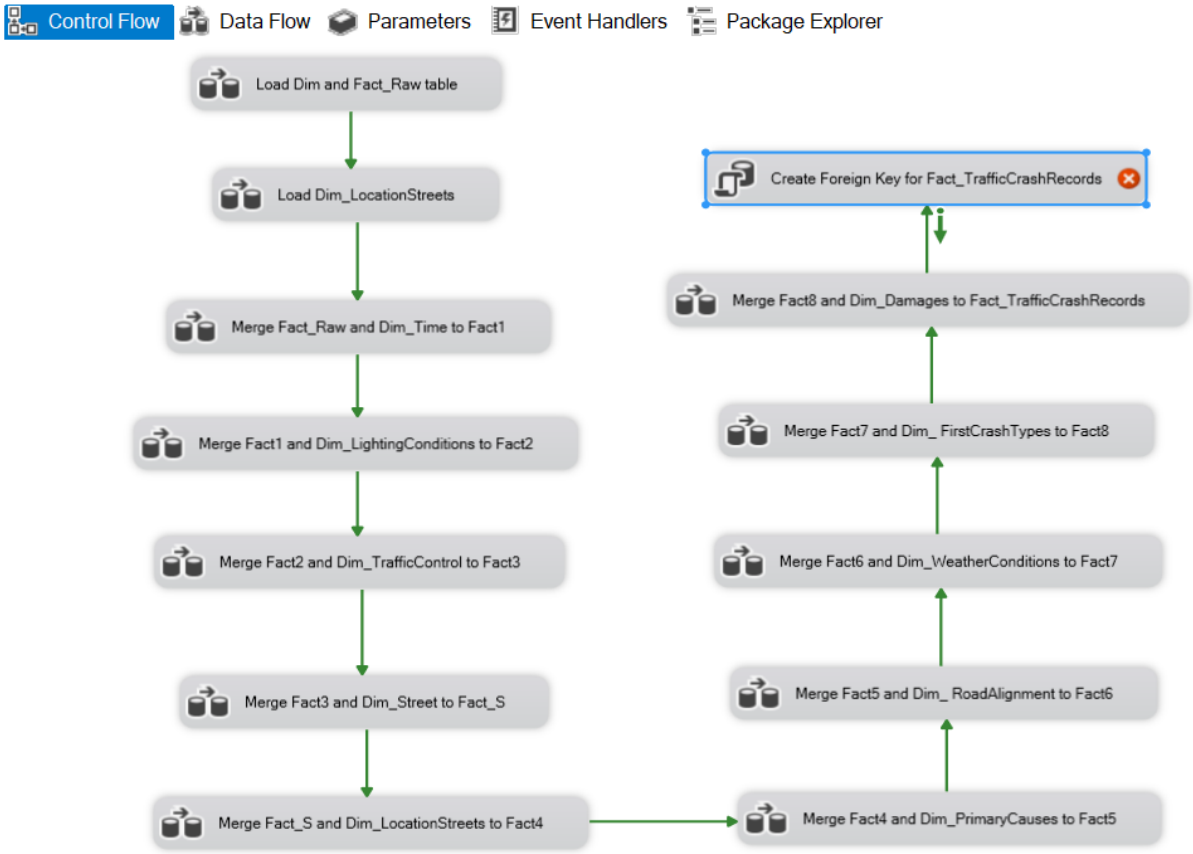
- Vào mục “Mappings” để kiểm tra việc ánh xạ các cột dữ liệu. Nhấn “OK” để hoàn tất.



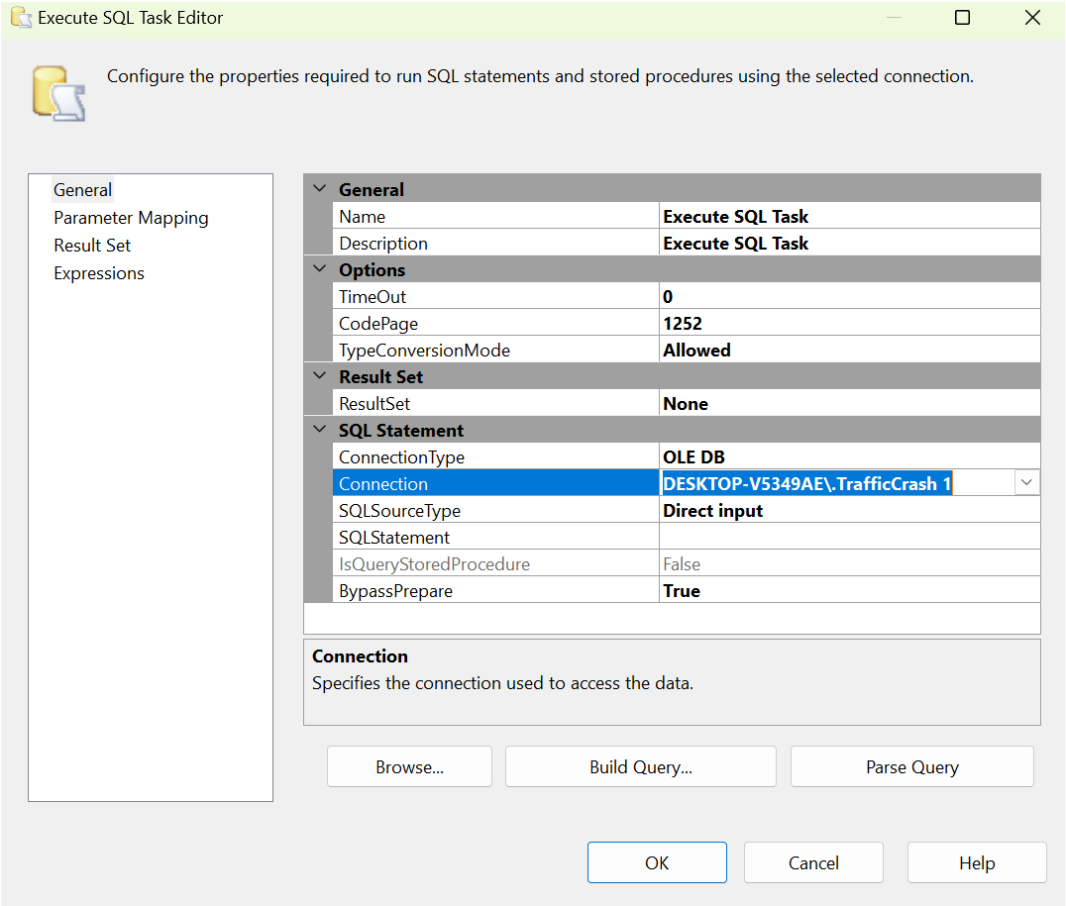
Tiếp theo, ta sẽ tiến hành tạo khóa ngoại từ bảng Fact đến các Dimension và chạy dự án SSIS.

2.3.14 Tạo khóa ngoại từ bảng Fact đến các Dimension.

Bước 1: Tạo một Execute SQL Task, đặt tên "Create Foreign Key for Fact_TrafficCrashRecords" để thực thi các câu lệnh SQL tạo các khóa ngoại từ các Dimension đến bảng Fact.



Bước 2: Nhấn chuột phải vào "Create Foreign Key for Fact_TrafficCrashRecords" và chọn "Edit". Ở ô Connection, chọn connection đã thiết lập đến data warehouse trong SQL Server.



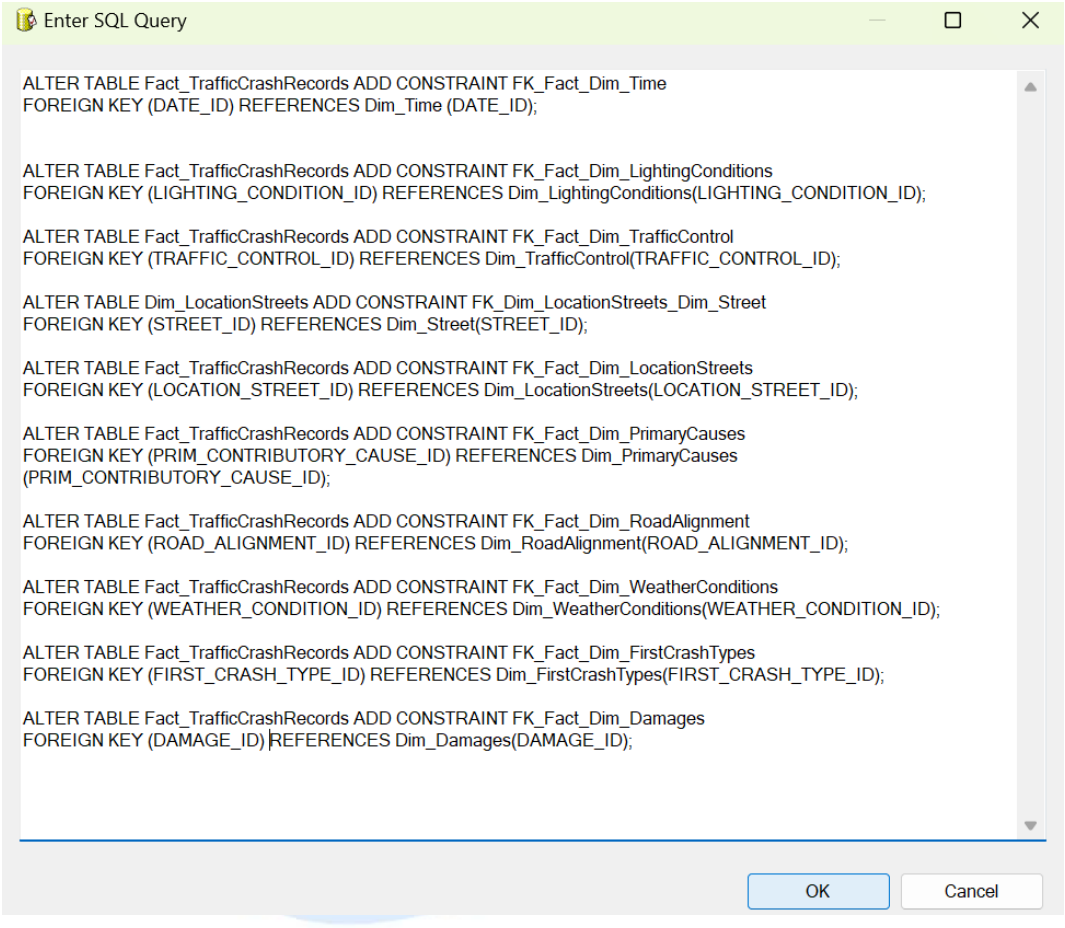
Bước 3: Ở ô SQLStatement, thêm các câu truy vấn SQL thực hiện tạo các khóa ngoại từ các Dimension đến bảng Fact.

- Câu lệnh tạo khóa ngoại:

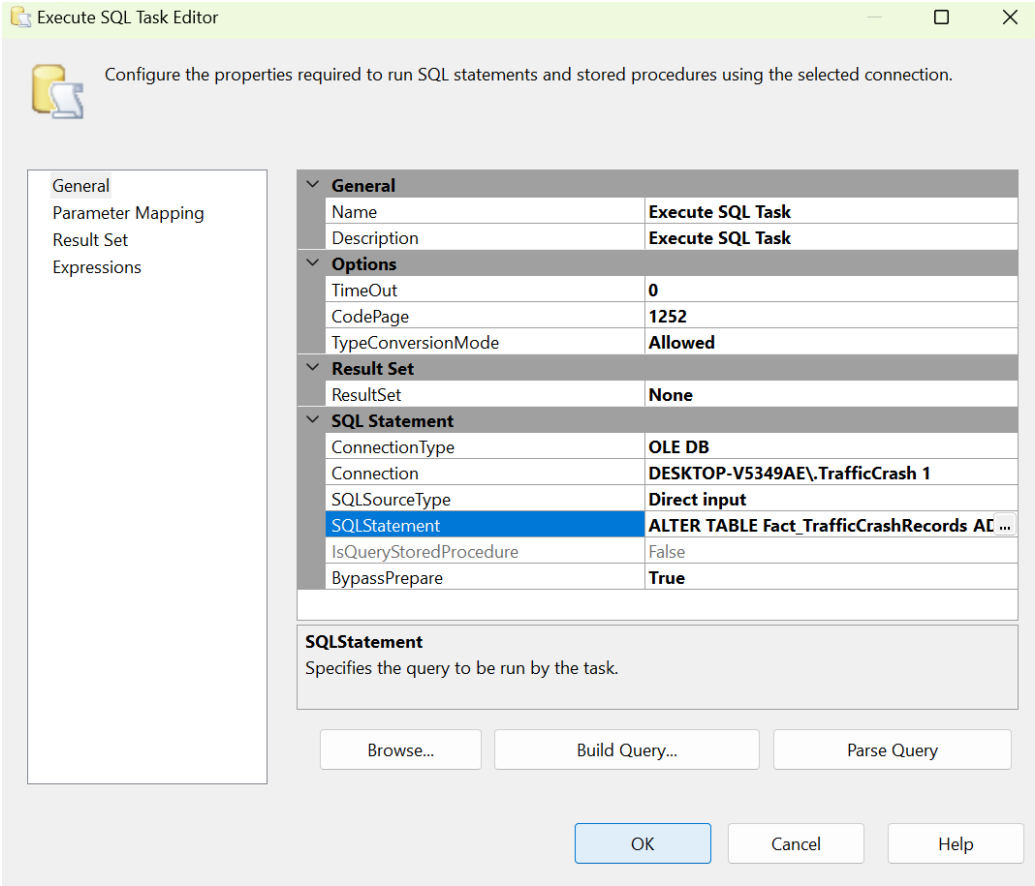
```
ALTER TABLE Fact_TrafficCrashRecords ADD CONSTRAINT
FK_Fact_Dim_Time
FOREIGN KEY (DATE_ID) REFERENCES Dim_Time (DATE_ID);

ALTER TABLE Fact_TrafficCrashRecords ADD CONSTRAINT
FK_Fact_Dim_LightingConditions
FOREIGN KEY (LIGHTING_CONDITION_ID) REFERENCES
Dim_LightingConditions(LIGHTING_CONDITION_ID);

ALTER TABLE Fact_TrafficCrashRecords ADD CONSTRAINT
FK_Fact_Dim_TrafficControl
```

```
FOREIGN KEY (TRAFFIC_CONTROL_ID) REFERENCES
Dim_TrafficControl(TRAFFIC_CONTROL_ID);


ALTER TABLE Dim_LocationStreets ADD CONSTRAINT
FK_Dim_LocationStreets_Dim_Street
FOREIGN KEY (STREET_ID) REFERENCES Dim_Street(STREET_ID);


ALTER TABLE Fact_TrafficCrashRecords ADD CONSTRAINT
FK_Fact_Dim_LocationStreets
FOREIGN KEY (LOCATION_STREET_ID) REFERENCES
Dim_LocationStreets(LOCATION_STREET_ID);


ALTER TABLE Fact_TrafficCrashRecords ADD CONSTRAINT
FK_Fact_Dim_PrimaryCauses
FOREIGN KEY (PRIM_CONTRIBUTORY_CAUSE_ID) REFERENCES
Dim_PrimaryCauses(PRIM_CONTRIBUTORY_CAUSE_ID);


ALTER TABLE Fact_TrafficCrashRecords ADD CONSTRAINT
FK_Fact_Dim_RoadAlignment
FOREIGN KEY (ROAD_ALIGNMENT_ID) REFERENCES
Dim_RoadAlignment(ROAD_ALIGNMENT_ID);


ALTER TABLE Fact_TrafficCrashRecords ADD CONSTRAINT
FK_Fact_Dim_WeatherConditions
FOREIGN KEY (WEATHER_CONDITION_ID) REFERENCES
Dim_WeatherConditions(WEATHER_CONDITION_ID);


ALTER TABLE Fact_TrafficCrashRecords ADD CONSTRAINT
FK_Fact_Dim_FirstCrashTypes
FOREIGN KEY (FIRST_CRASH_TYPE_ID) REFERENCES
Dim_FirstCrashTypes(FIRST_CRASH_TYPE_ID);
```

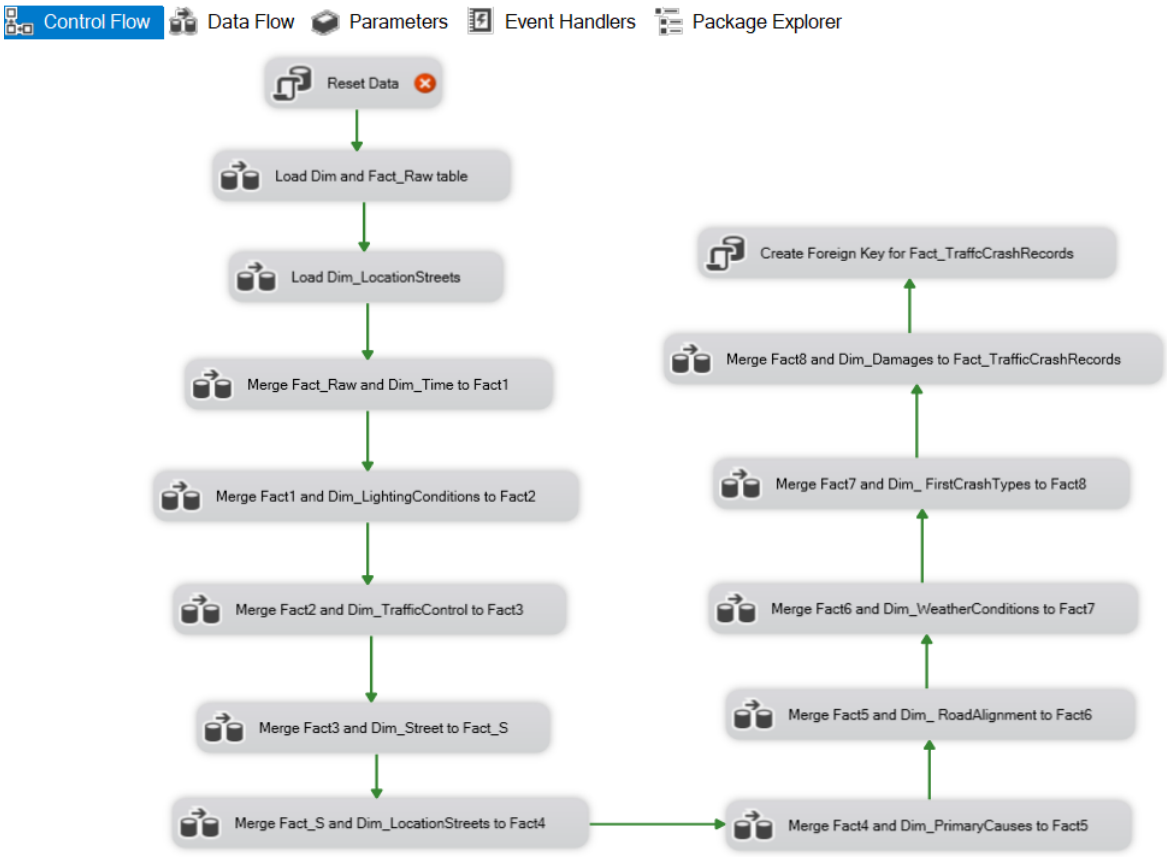
```
ALTER TABLE Fact_TrafficCrashRecords ADD CONSTRAINT
FK_Fact_Dim_Damages
FOREIGN KEY (DAMAGE_ID) REFERENCES
Dim_Damages(DAMAGE_ID);
```



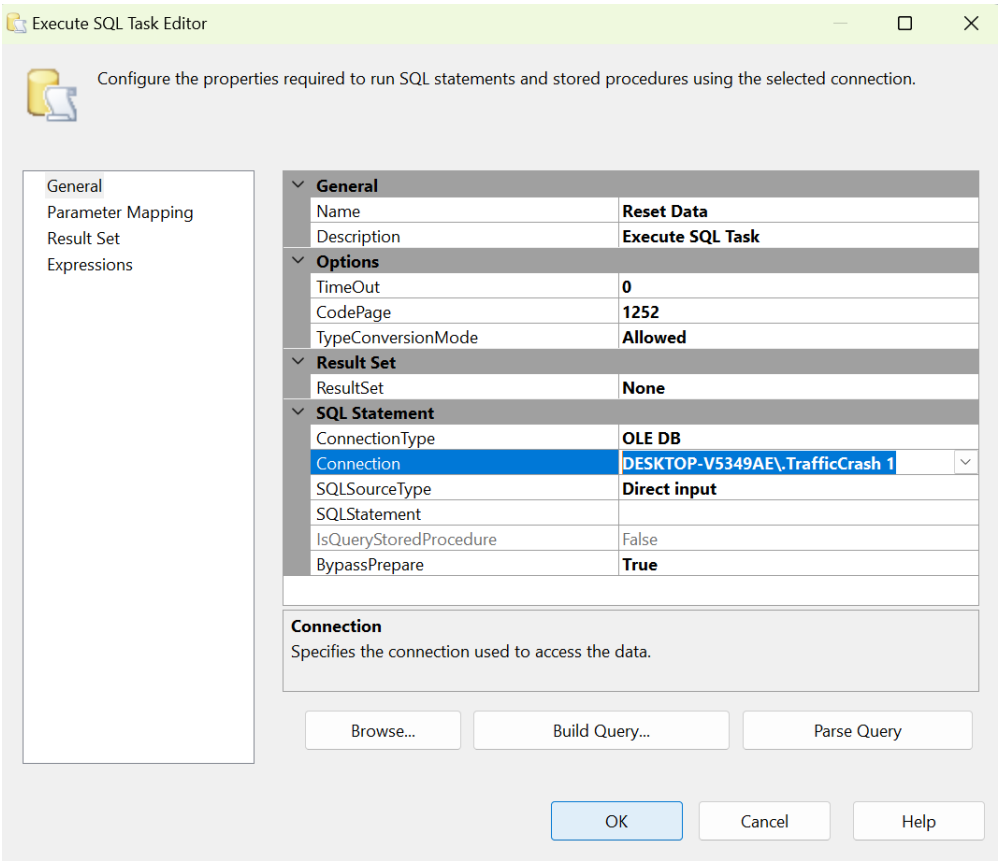
- Nhấn “OK” để hoàn tất quá trình.



Bước 4: Thêm vào một Execute SQL Task, đặt tên là "Reset Data", trước quá trình chia bảng Fact và các bảng Dim nhằm thực hiện nhiệm vụ đảm bảo xóa các khóa ngoại và đổ dữ liệu mới hoàn toàn (không bị trùng lặp dữ liệu cũ) mỗi khi chạy dự án.



Bước 5: Nhấn chuột phải vào "Reset Data" và chọn "Edit". Ở ô Connection, chọn connection đã thiết lập đến data warehouse trong SQL Server.

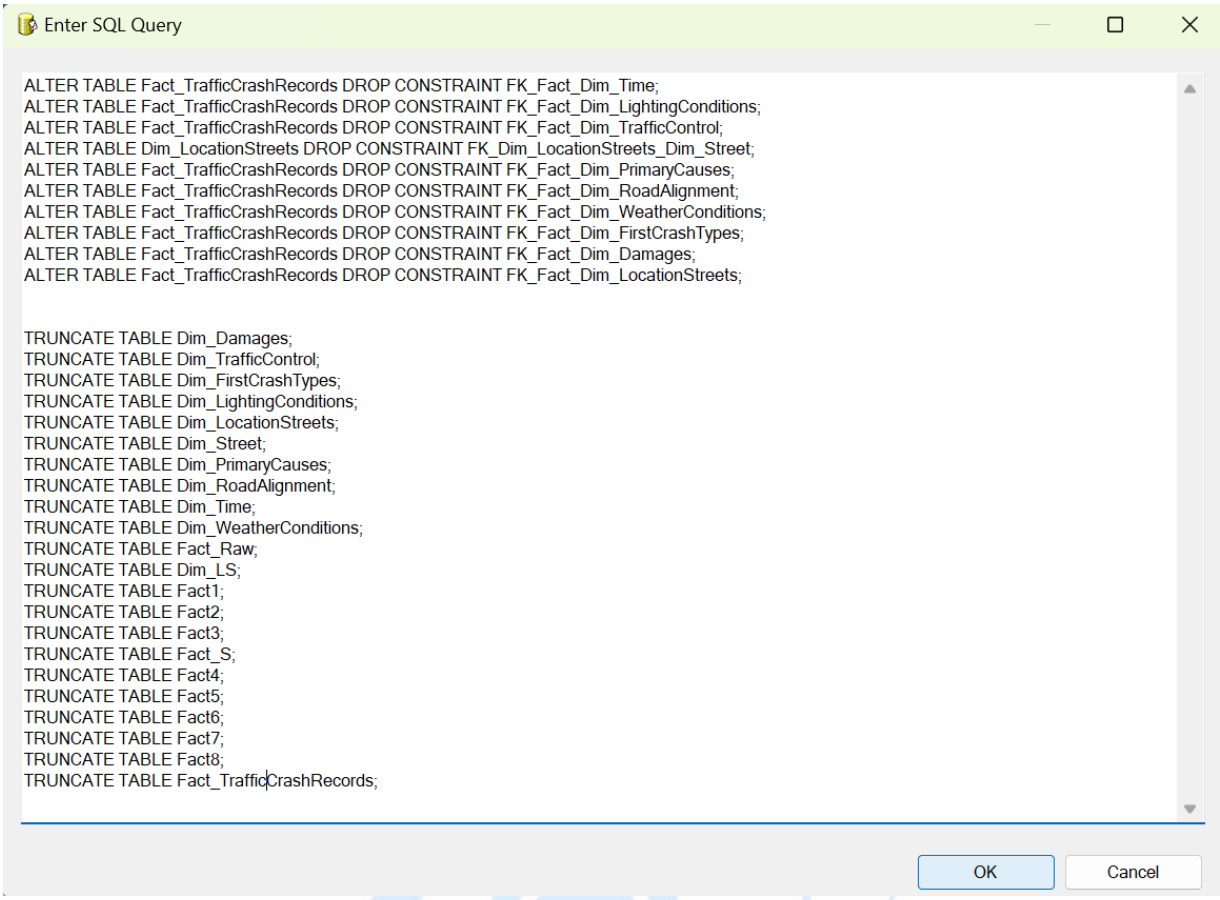


Bước 6: Ở ô SQLStatement, thêm các câu truy vấn SQL thực hiện xóa khóa ngoại và dữ liệu cũ trong các bảng mỗi khi chạy dự án. Nhấn OK để hoàn tất quá trình.

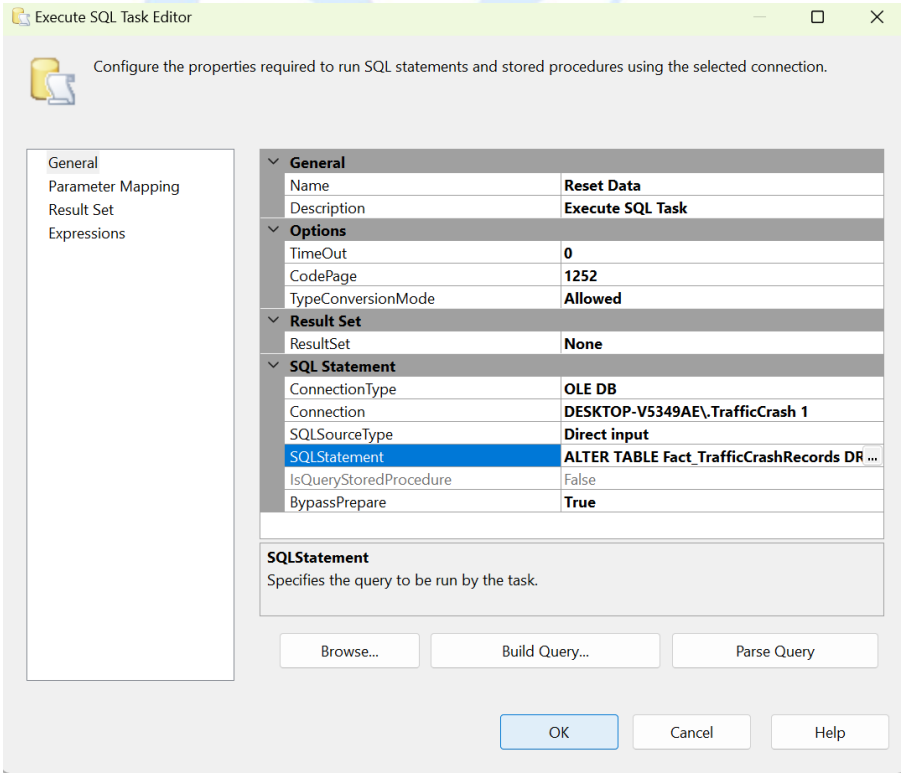
- Câu lệnh xóa khóa ngoại và dữ liệu:

```
ALTER TABLE Fact_TrafficCrashRecords DROP CONSTRAINT
FK_Fact_Dim_Time;
ALTER TABLE Fact_TrafficCrashRecords DROP CONSTRAINT
FK_Fact_Dim_LightingConditions;
ALTER TABLE Fact_TrafficCrashRecords DROP CONSTRAINT
FK_Fact_Dim_TrafficControl;
ALTER TABLE Dim_LocationStreets DROP CONSTRAINT
FK_Dim_LocationStreets_Dim_Street;
ALTER TABLE Fact_TrafficCrashRecords DROP CONSTRAINT
FK_Fact_Dim_PrimaryCauses;
ALTER TABLE Fact_TrafficCrashRecords DROP CONSTRAINT
FK_Fact_Dim_RoadAlignment;
ALTER TABLE Fact_TrafficCrashRecords DROP CONSTRAINT
FK_Fact_Dim_WeatherConditions;
ALTER TABLE Fact_TrafficCrashRecords DROP CONSTRAINT
FK_Fact_Dim_FirstCrashTypes;
```

```
ALTER TABLE Fact_TrafficCrashRecords DROP CONSTRAINT
FK_Fact_Dim_Damages;
ALTER TABLE Fact_TrafficCrashRecords DROP CONSTRAINT
FK_Fact_Dim_LocationStreets;


TRUNCATE TABLE Dim_Damages;
TRUNCATE TABLE Dim_TrafficControl;
TRUNCATE TABLE Dim_FirstCrashTypes;
TRUNCATE TABLE Dim_LightingConditions;
TRUNCATE TABLE Dim_LocationStreets;
TRUNCATE TABLE Dim_Street;
TRUNCATE TABLE Dim_PrimaryCauses;
TRUNCATE TABLE Dim_RoadAlignment;
TRUNCATE TABLE Dim_Time;
TRUNCATE TABLE Dim_WeatherConditions;
TRUNCATE TABLE Fact_Raw;
TRUNCATE TABLE Dim_LS;
TRUNCATE TABLE Fact1;
TRUNCATE TABLE Fact2;
TRUNCATE TABLE Fact3;
TRUNCATE TABLE Fact_S;
TRUNCATE TABLE Fact4;
TRUNCATE TABLE Fact5;
TRUNCATE TABLE Fact6;
TRUNCATE TABLE Fact7;
TRUNCATE TABLE Fact8;
TRUNCATE TABLE Fact_TrafficCrashRecords;
```



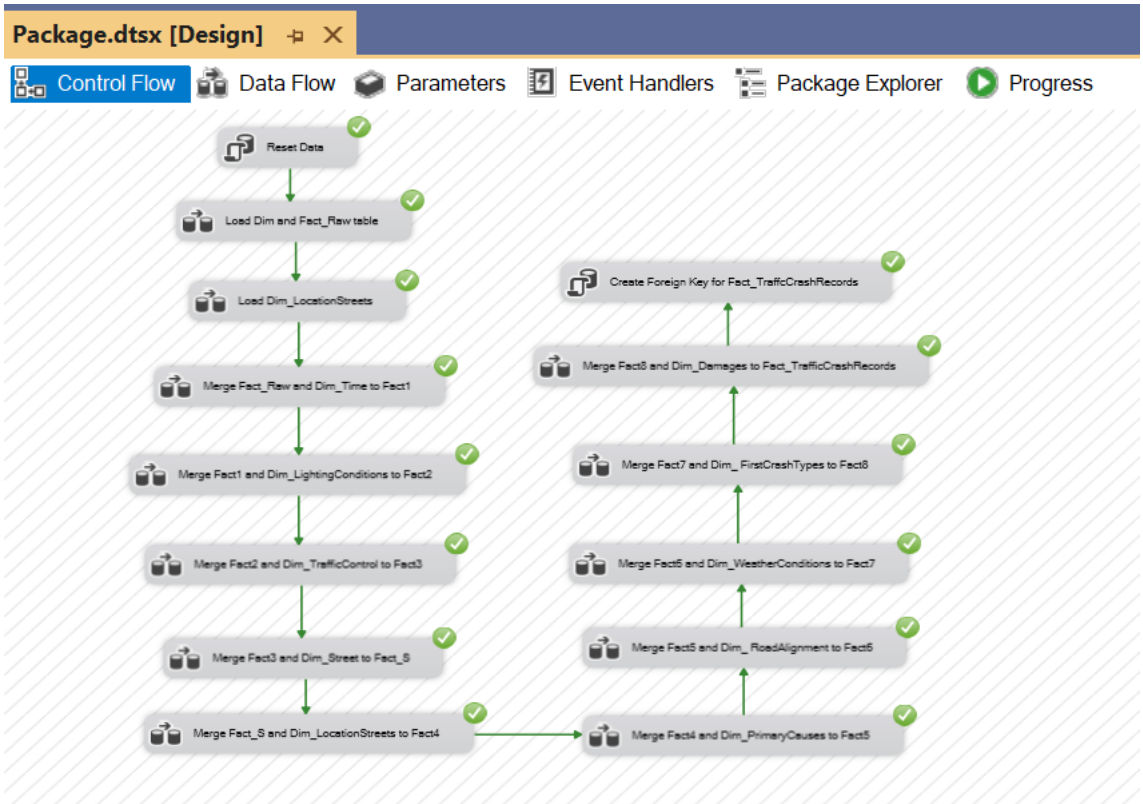
- Nhấn “OK” để hoàn tất quá trình.



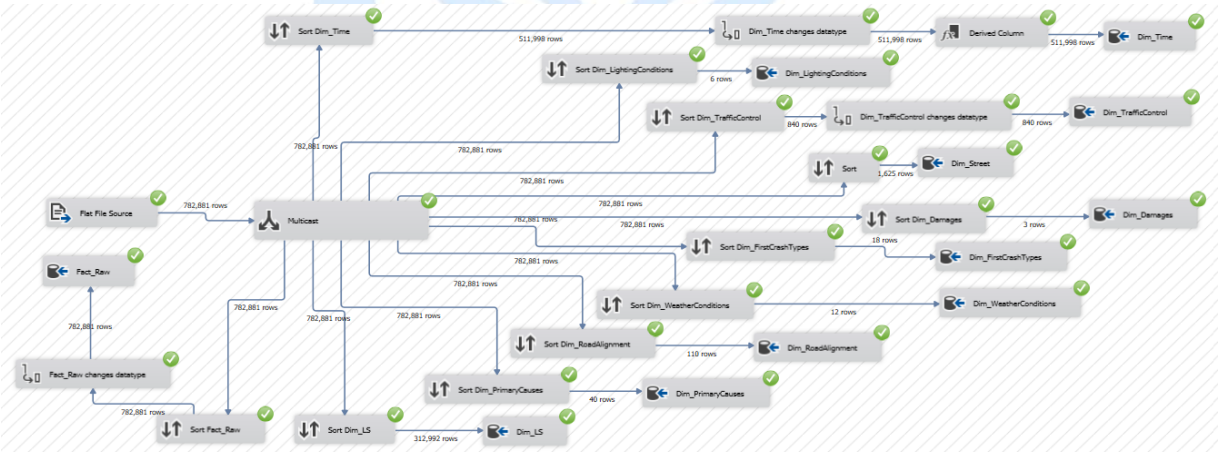
2.4 Chạy dự án SSIS

Nhấn nút **Start** trên thanh Menu để tiến hành chạy dự án.

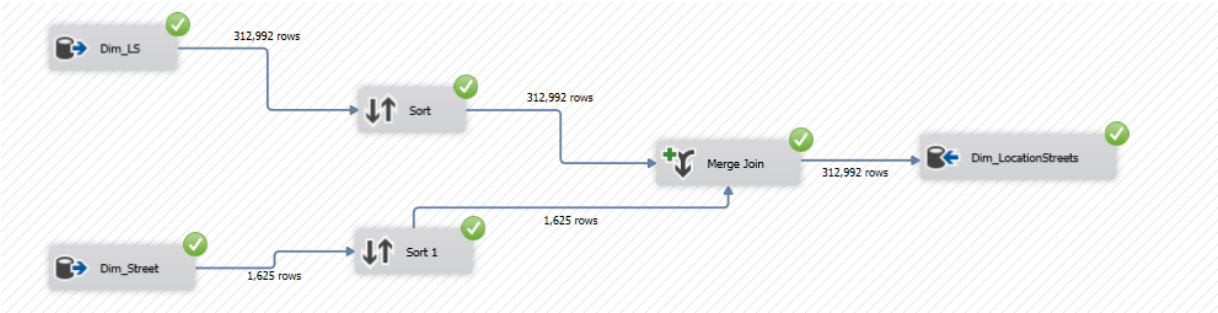
❖ Kết quả khi chạy dự án:



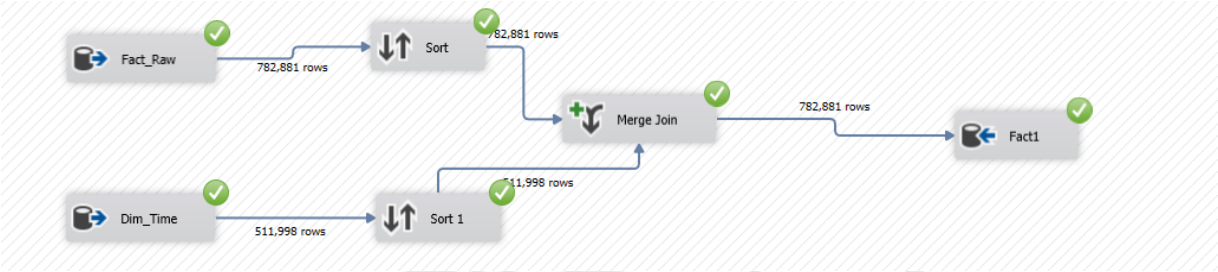
*Load Dim and Fact_Raw table.



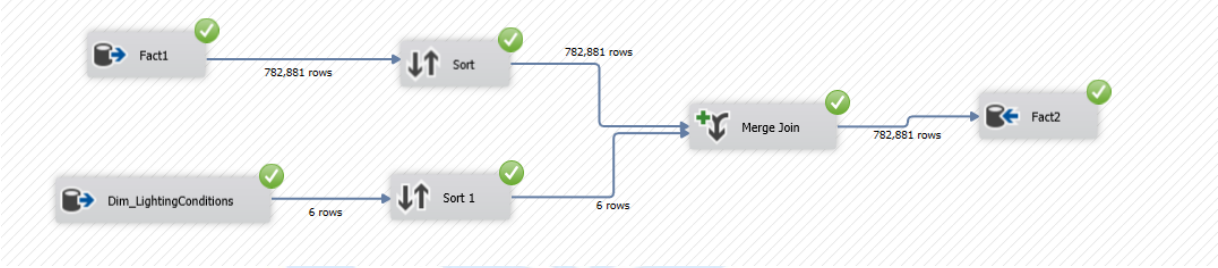
*Load Dim_LocationStreets.



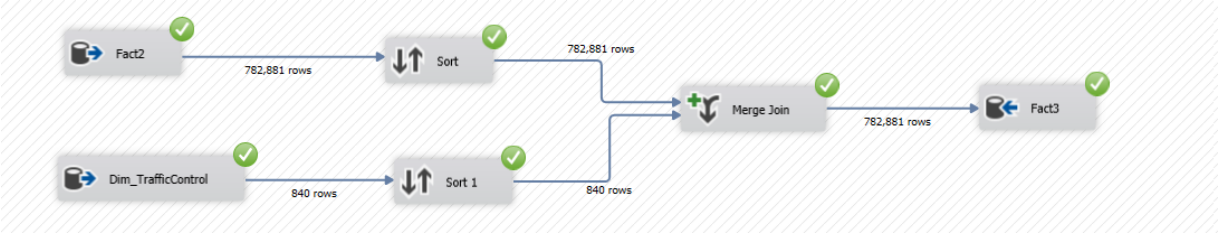
*Merge Fact_Raw and Dim_Time to Fact1.



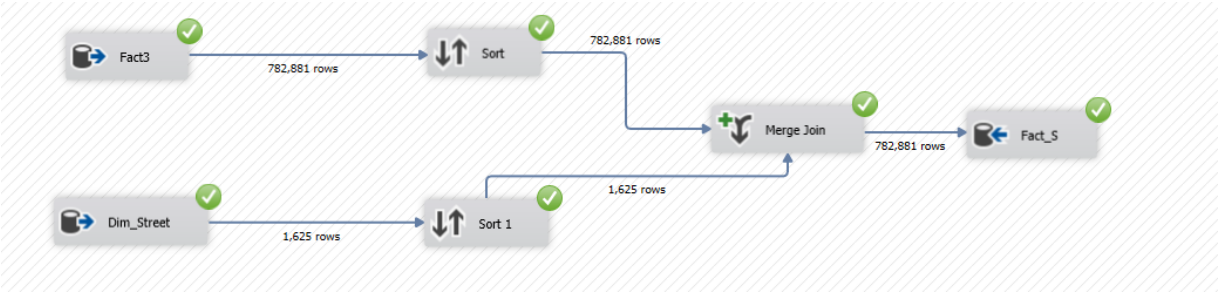
*Merge Fact1 and Dim_LightingConditions to Fact2.



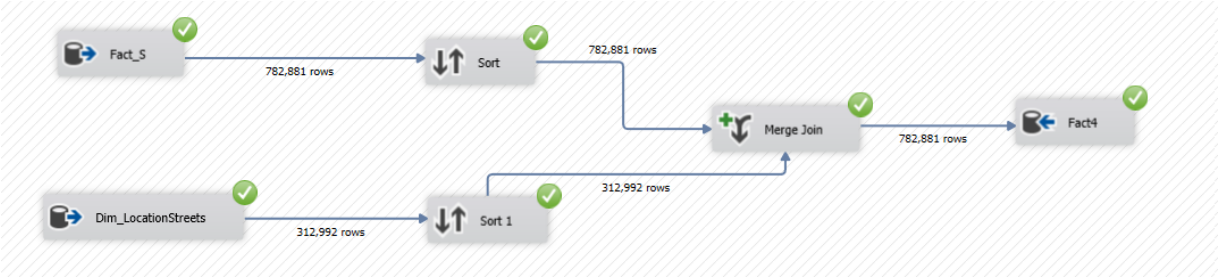
*Merge Fact2 and Dim_TrafficControl to Fact3.



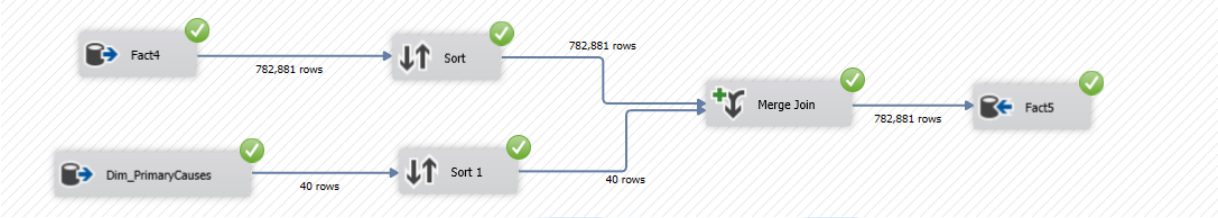
*Merge Fact3 and Dim_Street to Fact_S.



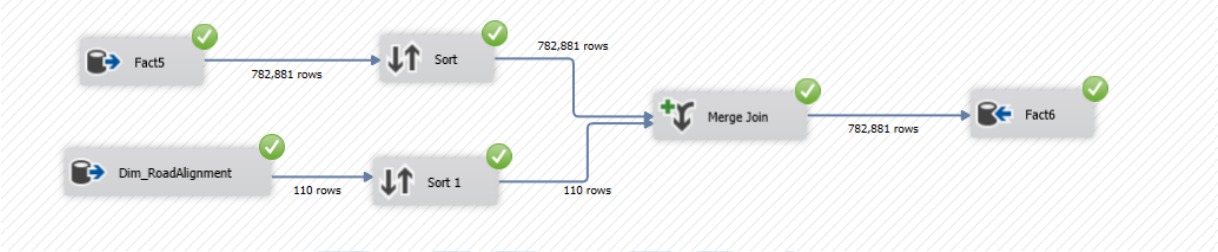
*Merge Fact_S and Dim_LocationStreets to Fact4.



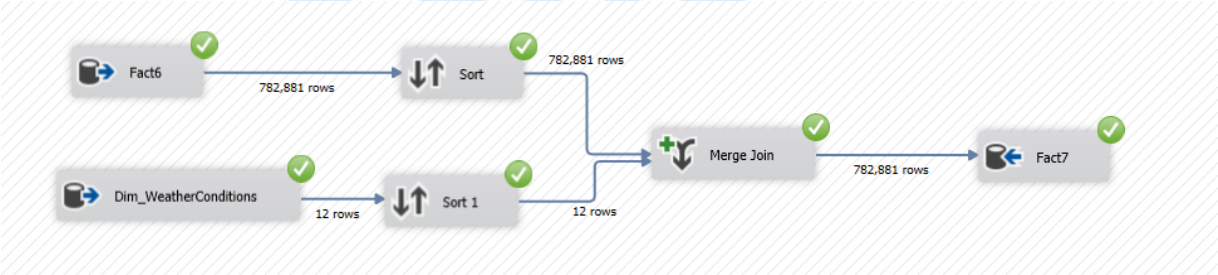
*Merge Fact4 and Dim_PrimaryCauses to Fact5.



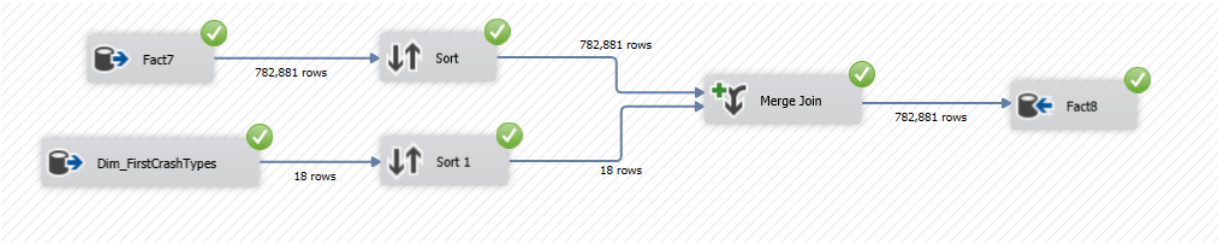
*Merge Fact5 and Dim_RoadAlignment to Fact6.



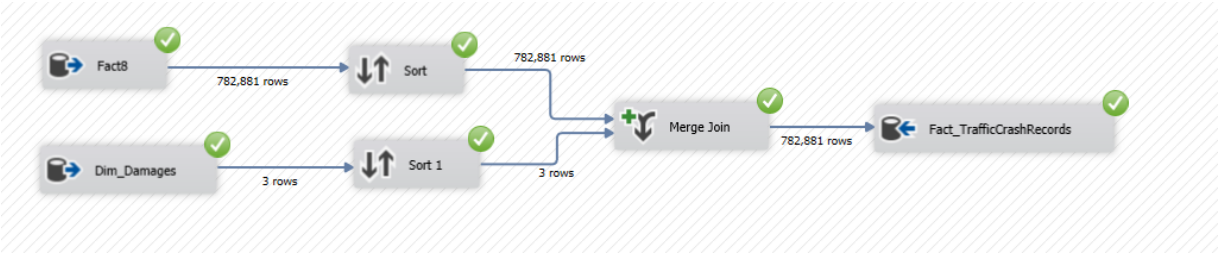
*Merge Fact6 and Dim_WeatherConditions to Fact7.



*Merge Fact7 and Dim_FirstCrashTypes to Fact8.



*Merge Fact8 and Dim_Damages to TrafficCrashRecords.



2.5 Kiểm tra dữ liệu các bảng

❖ Dữ liệu bảng Dim_Time.

	DATE_ID	CRASH_DATE	CRASH_HOUR	CRASH_DAY_OF_WEEK	CRASH_DAY	CRASH_MONTH	CRASH_QUARTER	CRASH_YEAR
1	1	2016-01-01 00:01:00.000	0	6	1	1	1	2016
2	2	2016-01-01 00:08:00.000	0	6	1	1	1	2016
3	3	2016-01-01 00:15:00.000	0	6	1	1	1	2016
4	4	2016-01-01 00:36:00.000	0	6	1	1	1	2016
5	5	2016-01-01 00:55:00.000	0	6	1	1	1	2016
6	6	2016-01-01 01:00:00.000	1	6	1	1	1	2016
7	7	2016-01-01 01:03:00.000	1	6	1	1	1	2016
8	8	2016-01-01 01:18:00.000	1	6	1	1	1	2016
9	9	2016-01-01 01:30:00.000	1	6	1	1	1	2016
10	10	2016-01-01 01:35:00.000	1	6	1	1	1	2016
11	11	2016-01-01 01:45:00.000	1	6	1	1	1	2016
12	12	2016-01-01 02:00:00.000	2	6	1	1	1	2016
13	13	2016-01-01 02:05:00.000	2	6	1	1	1	2016
14	14	2016-01-01 02:10:00.000	2	6	1	1	1	2016
15	15	2016-01-01 02:20:00.000	2	6	1	1	1	2016
16	16	2016-01-01 02:30:00.000	2	6	1	1	1	2016
17	17	2016-01-01 02:45:00.000	2	6	1	1	1	2016
18	18	2016-01-01 02:47:00.000	2	6	1	1	1	2016
19	19	2016-01-01 03:00:00.000	3	6	1	1	1	2016
20	20	2016-01-01 03:20:00.000	3	6	1	1	1	2016

Query executed successfully. | DESKTOP-V5349AE (16.0 RTM) | DESKTOP-V5349AE\DELL (79) | TrafficCrash | 00:00:01 | 511,998 rows

❖ Dữ liệu bảng Dim_LightingConditions.

	LIGHTING_CONDITION_ID	LIGHTING_CONDITION
1	1	DARKNESS
2	2	DARKNESS; LIGHTED ROAD
3	3	DAWN
4	4	DAYLIGHT
5	5	DUSK
6	6	UNKNOWN

Query executed successfully. | DESKTOP-V5349AE (16.0 RTM) | DESKTOP-V5349AE\DELL (79) | TrafficCrash | 00:00:00 | 6 rows

❖ Dữ liệu bảng Dim_TrafficControl.

	TRAFFIC_CONTROL_ID	TRAFFIC_CONTROL_DEVICE	DEVICE_CONDITION	POSTED_SPEED_LIMIT
1	1	BICYCLE CROSSING SIGN	FUNCTIONING PROPERLY	30
2	2	BICYCLE CROSSING SIGN	FUNCTIONING PROPERLY	35
3	3	BICYCLE CROSSING SIGN	NO CONTROLS	15
4	4	BICYCLE CROSSING SIGN	NO CONTROLS	25
5	5	BICYCLE CROSSING SIGN	NO CONTROLS	30
6	6	BICYCLE CROSSING SIGN	NOT FUNCTIONING	30
7	7	BICYCLE CROSSING SIGN	OTHER	30
8	8	BICYCLE CROSSING SIGN	UNKNOWN	25
9	9	BICYCLE CROSSING SIGN	UNKNOWN	30
10	10	DELINEATORS	FUNCTIONING IMPROPE...	30
11	11	DELINEATORS	FUNCTIONING IMPROPE...	40
12	12	DELINEATORS	FUNCTIONING PROPERLY	0
13	13	DELINEATORS	FUNCTIONING PROPERLY	10
14	14	DELINEATORS	FUNCTIONING PROPERLY	15
15	15	DELINEATORS	FUNCTIONING PROPERLY	25
16	16	DELINEATORS	FUNCTIONING PROPERLY	30
17	17	DELINEATORS	FUNCTIONING PROPERLY	35
18	18	DELINEATORS	FUNCTIONING PROPERLY	40
19	19	DELINEATORS	FUNCTIONING PROPERLY	45
20	20	DELINEATORS	FUNCTIONING PROPERLY	5

Query executed successfully. DESKTOP-V5349AE (16.0 RTM) DESKTOP-V5349AE\DELL (79) TrafficCrash 00:00:00 840 rows

❖ Dữ liệu bảng Dim_Street.

	STREET_ID	STREET_NAME
1	1	100TH DR
2	2	100TH PL
3	3	100TH ST
4	4	101ST PL
5	5	101ST ST
6	6	102ND PL
7	7	102ND ST
8	8	103RD PL
9	9	103RD ST
10	10	104TH PL
11	11	104TH ST
12	12	105TH PL
13	13	105TH ST
14	14	106TH PL
15	15	106TH ST
16	16	107TH PL
17	17	107TH ST
18	18	108TH PL
19	19	108TH ST
20	20	109TH PL

Query executed successfully. DESKTOP-V5349AE (16.0 RTM) DESKTOP-V5349AE\DELL (79) TrafficCrash 00:00:00 1,625 rows

❖ Dữ liệu bảng Dim_LocationStreets.

	LOCATION_STREET_ID	STREET_NO	STREET_DIRECTION	STREET_ID
1	1	3800	E	1
2	2	1101	E	2
3	3	1301	W	2
4	4	1450	W	2
5	5	1545	W	2
6	6	1649	W	2
7	7	19	E	2
8	8	27	E	2
9	9	30	E	2
10	10	326	W	2
11	11	35	E	2
12	12	1054	E	2
13	13	413	W	2
14	14	439	W	2
15	15	48	E	2
16	16	50	E	2
17	17	60	E	2
18	18	622	E	2
19	19	1114	E	2
20	20	661	E	2

Query executed successfully. | DESKTOP-V5349AE (16.0 RTM) | DESKTOP-V5349AE\DELL (79) | TrafficCrash | 00:00:01 | 312,992 rows

❖ Dữ liệu bảng Dim_PrimaryCauses.

	PRIM_CONTRIBUTORY_CAUSE_ID	PRIM_CONTRIBUTORY_CAUSE
1	1	ANIMAL
2	2	BICYCLE ADVANCING LEGALLY ON RED LIGHT
3	3	CELL PHONE USE OTHER THAN TEXTING
4	4	DISREGARDING OTHER TRAFFIC SIGNS
5	5	DISREGARDING ROAD MARKINGS
6	6	DISREGARDING STOP SIGN
7	7	DISREGARDING TRAFFIC SIGNALS
8	8	DISREGARDING YIELD SIGN
9	9	DISTRACTION - FROM INSIDE VEHICLE
10	10	DISTRACTION - FROM OUTSIDE VEHICLE
11	11	DISTRACTION - OTHER ELECTRONIC DEVICE (NAVIGATIO...
12	12	DRIVING ON WRONG SIDE/WRONG WAY
13	13	DRIVING SKILLS/KNOWLEDGE/EXPERIENCE
14	14	EQUIPMENT - VEHICLE CONDITION
15	15	EVASIVE ACTION DUE TO ANIMAL; OBJECT; NONMOTORIST
16	16	EXCEEDING AUTHORIZED SPEED LIMIT
17	17	EXCEEDING SAFE SPEED FOR CONDITIONS
18	18	FAILING TO REDUCE SPEED TO AVOID CRASH
19	19	FAILING TO YIELD RIGHT-OF-WAY
20	20	FOLLOWING TOO CLOSELY

Query executed successfully. | DESKTOP-V5349AE (16.0 RTM) | DESKTOP-V5349AE\DELL (66) | TrafficCrash | 00:00:00 | 40 rows

❖ Dữ liệu bảng Dim_RoadAlignment.

	ROAD_ALIGNMENT_ID	ALIGNMENT	TRAFFICWAY_TYPE
1	1	CURVE ON GRADE	ALLEY
2	2	CURVE ON GRADE	CENTER TURN LANE
3	3	CURVE ON GRADE	DIVIDED - W/MEDIAN (NOT RAISED)
4	4	CURVE ON GRADE	DIVIDED - W/MEDIAN BARRIER
5	5	CURVE ON GRADE	DRIVEWAY
6	6	CURVE ON GRADE	FIVE POINT, OR MORE
7	7	CURVE ON GRADE	FOUR WAY
8	8	CURVE ON GRADE	L-INTERSECTION
9	9	CURVE ON GRADE	NOT DIVIDED
10	10	CURVE ON GRADE	ONE-WAY
11	11	CURVE ON GRADE	OTHER
12	12	CURVE ON GRADE	PARKING LOT
13	13	CURVE ON GRADE	RAMP
14	14	CURVE ON GRADE	ROUNDABOUT
15	15	CURVE ON GRADE	T-INTERSECTION
16	16	CURVE ON GRADE	TRAFFIC ROUTE
17	17	CURVE ON GRADE	UNKNOWN
18	18	CURVE ON GRADE	UNKNOWN INTERSECTION TYPE
19	19	CURVE ON GRADE	Y-INTERSECTION
20	20	CURVE ON HILLCREST	ALLEY

Query executed successfully. | DESKTOP-V5349AE (16.0 RTM) | DESKTOP-V5349AE\DELL (79) | TrafficCrash | 00:00:00 | 110 rows

❖ Dữ liệu bảng Dim_WeatherConditions.

	WEATHER_CONDITION_ID	WEATHER_CONDITION
1	1	BLOWING SAND; SOIL; DIRT
2	2	BLOWING SNOW
3	3	CLEAR
4	4	CLOUDY/OVERCAST
5	5	FOG/SMOKE/HAZE
6	6	FREEZING RAIN/DRIZZLE
7	7	OTHER
8	8	RAIN
9	9	SEVERE CROSS WIND GATE
10	10	SLEET/HAIL
11	11	SNOW
12	12	UNKNOWN

Query executed successfully. | DESKTOP-V5349AE (16.0 RTM) | DESKTOP-V5349AE\DELL (79) | TrafficCrash | 00:00:00 | 12 rows

❖ Dữ liệu bảng Dim_FirstCrashTypes.

	FIRST_CRASH_TYPE_ID	FIRST_CRASH_TYPE
1	1	ANGLE
2	2	ANIMAL
3	3	FIXED OBJECT
4	4	HEAD ON
5	5	OTHER NONCOLLISION
6	6	OTHER OBJECT
7	7	OVERTURNED
8	8	PARKED MOTOR VEHICLE
9	9	PEDALCYCLIST
10	10	PEDESTRIAN
11	11	REAR END
12	12	REAR TO FRONT
13	13	REAR TO REAR
14	14	REAR TO SIDE
15	15	SIDESWIPE OPPOSITE DIRECTION
16	16	SIDESWIPE SAME DIRECTION
17	17	TRAIN
18	18	TURNING

Query executed successfully. | DESKTOP-V5349AE (16.0 RTM) | DESKTOP-V5349AE\DELL (79) | TrafficCrash | 00:00:00 | 18 rows

❖ Dữ liệu bảng Dim_Damages.

	DAMAGE_ID	DAMAGE
1	1	500 OR LESS
2	2	501 - 1.500
3	3	OVER 1.500

Query executed successfully. | DESKTOP-V5349AE (16.0 RTM) | DESKTOP-V5349AE\DELL (79) | TrafficCrash | 00:00:00 | 3 rows

❖ Dữ liệu bảng Fact_TrafficCrashRecords.

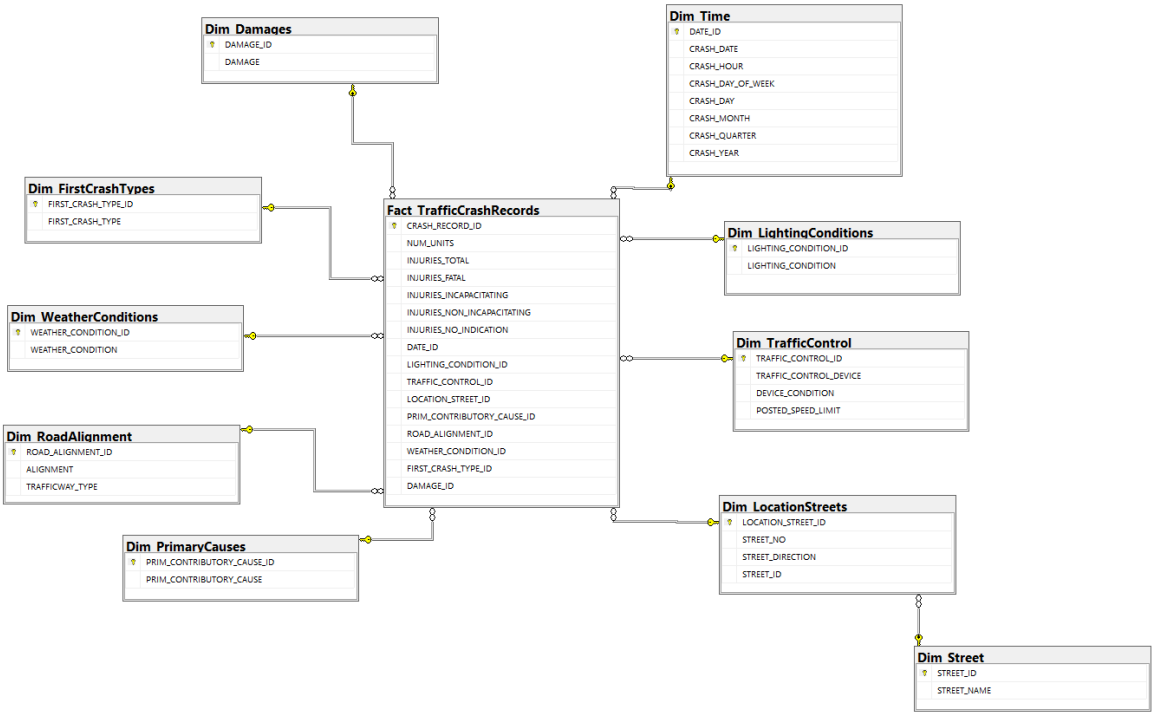
	CRASH_RECORD_ID	NUM_UNITS	INJURIES_TOTAL	INJURIES_FATAL	INJURIES_INCAPACITATING	INJURIES_NON_INCAPACITATING	INJURIES_NO_INDICATION	DATE_ID	LIGHTING_CONDITION_ID
1	1	1	0	0	0	0	1	1	2
2	2	2	0	0	0	0	2	2	2
3	3	2	0	0	0	0	2	3	2
4	4	3	1	0	0	0	1	4	2
5	5	2	0	0	0	0	3	5	2
6	6	2	0	0	0	0	2	6	1
7	7	2	0	0	0	0	1	6	6
8	8	2	0	0	0	0	1	6	6
9	9	2	1	0	0	1	2	7	2
10	10	3	1	0	0	0	0	8	2
11	11	2	0	0	0	0	2	9	2
12	12	2	0	0	0	0	4	10	2
13	13	2	0	0	0	0	3	11	2
14	14	2	0	0	0	0	1	12	6
15	15	1	0	0	0	0	1	12	6
16	16	2	0	0	0	0	4	12	1
17	17	3	0	0	0	0	5	13	2
18	18	4	1	0	0	1	0	14	2
19	19	2	1	0	0	0	0	15	2
20	20	3	0	0	0	0	1	16	2

Query executed successfully. DESKTOP-V5349AE (16.0 RTM) DESKTOP-V5349AE\DELL (79) TrafficCrash 00:00:03 782,881 rows

	TRAFFIC_CONTROL_ID	LOCATION_STREET_ID	PRIM_CONTRIBUTORY_CAUSE_ID	ROAD_ALIGNMENT_ID	WEATHER_CONDITION_ID	FIRST_CRASH_TYPE_ID	DAMAGE_ID
1	165	130517	27	64	3	6	1
2	165	203924	18	62	3	11	2
3	655	145398	23	64	3	11	1
4	785	281082	27	56	3	16	3
5	170	150993	18	62	3	11	3
6	655	94190	37	64	3	11	1
7	785	111945	37	66	12	8	3
8	165	226273	37	64	3	8	3
9	720	132425	37	62	3	11	3
10	165	119154	38	62	11	8	3
11	655	303724	20	62	3	11	1
12	161	121378	37	62	3	11	2
13	655	72955	27	56	3	16	3
14	165	195895	37	62	3	8	3
15	785	279159	27	65	12	5	1
16	156	137631	37	66	3	18	3
17	584	162640	6	56	3	1	3
18	165	132387	16	64	11	8	3
19	165	32396	23	62	3	8	3
20	165	31992	37	62	3	8	3

Query executed successfully. DESKTOP-V5349AE (16.0 RTM) DESKTOP-V5349AE\DELL (79) TrafficCrash 00:00:03 782,881 rows

2.6 Lược đồ cơ sở dữ liệu sau khi hoàn thành


Dim_Damages
DAMAGE_ID
DAMAGE
Dim_Time
DATE_ID
CRASH_DATE
CRASH_HOUR
CRASH_DAY_OF_WEEK
CRASH_DAY
CRASH_MONTH
CRASH_QUARTER
CRASH_YEAR
Dim_FirstCrashTypes
FIRST_CRASH_TYPE_ID
FIRST_CRASH_TYPE
Fact_TrafficCrashRecords
CRASH_RECORD_ID
NUM_UNITS
INJURIES_TOTAL
INJURIES_FATAL
INJURIES_INCAPACITATING
INJURIES_NON_INCAPACITATING
INJURIES_NO_INDICATION
DATE_ID
LIGHTING_CONDITION_ID
TRAFFIC_CONTROL_ID
LOCATION_STREET_ID
PRIM_CONTRIBUTORY_CAUSE_ID
ROAD_ALIGNMENT_ID
WEATHER_CONDITION_ID
FIRST_CRASH_TYPE_ID
DAMAGE_ID
Dim_LightingConditions
LIGHTING_CONDITION_ID
LIGHTING_CONDITION
Dim_WeatherConditions
WEATHER_CONDITION_ID
WEATHER_CONDITION
Dim_TrafficControl
TRAFFIC_CONTROL_ID
TRAFFIC_CONTROL_DEVICE
DEVICE_CONDITION
POSTED_SPEED_LIMIT
Dim_RoadAlignment
ROAD_ALIGNMENT_ID
ALIGNMENT
TRAFFICWAY_TYPE
Dim_LocationStreets
LOCATION_STREET_ID
STREET_NO
STREET_DIRECTION
STREET_ID
Dim_PrimaryCauses
PRIM_CONTRIBUTORY_CAUSE_ID
PRIM_CONTRIBUTORY_CAUSE
Dim_Street
STREET_ID
STREET_NAME

CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU TRỰC TUYẾN (SSAS)

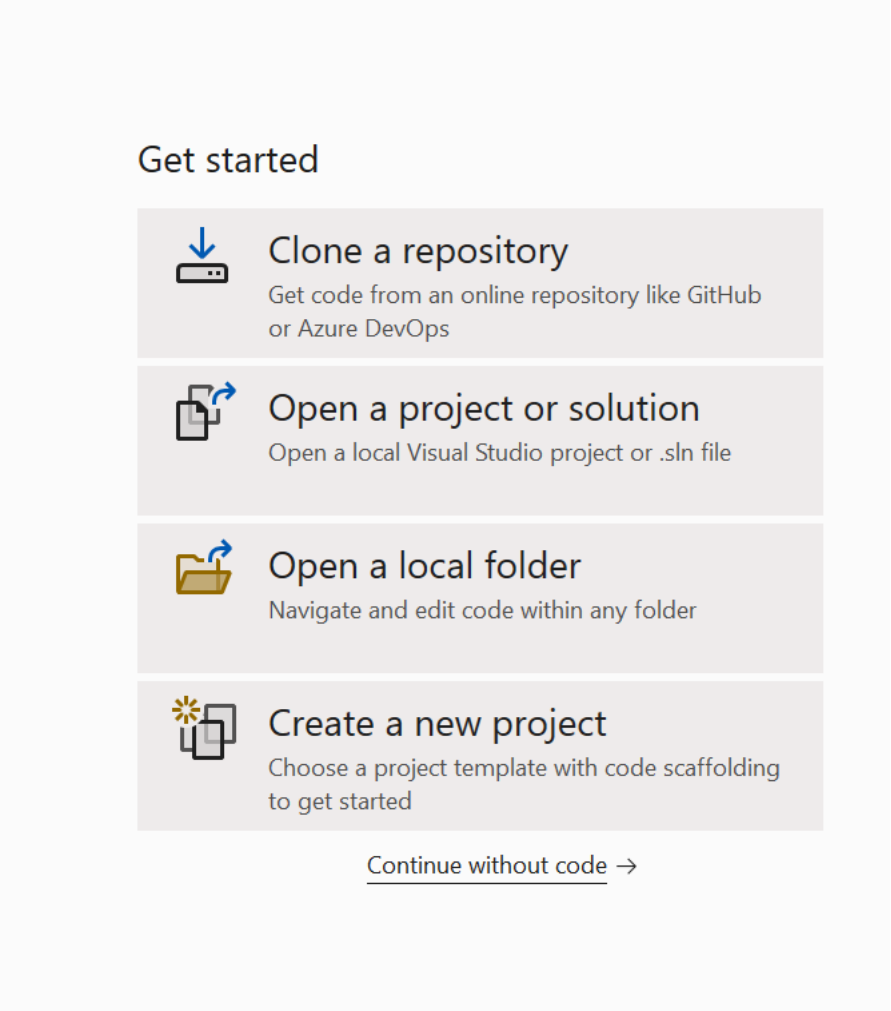
3.1 Chuẩn bị các công cụ

Để thực hiện được quá trình phân tích dữ liệu trực tuyến SSAS ta cần chuẩn bị và cài đặt các công cụ sau:

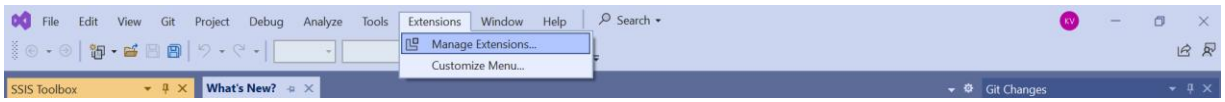
- Microsoft SQL Server có cài đặt Analysis Services.
- Microsoft Analysis Services Projects.

3.1.1 Cài đặt Microsoft Analysis Services Projects

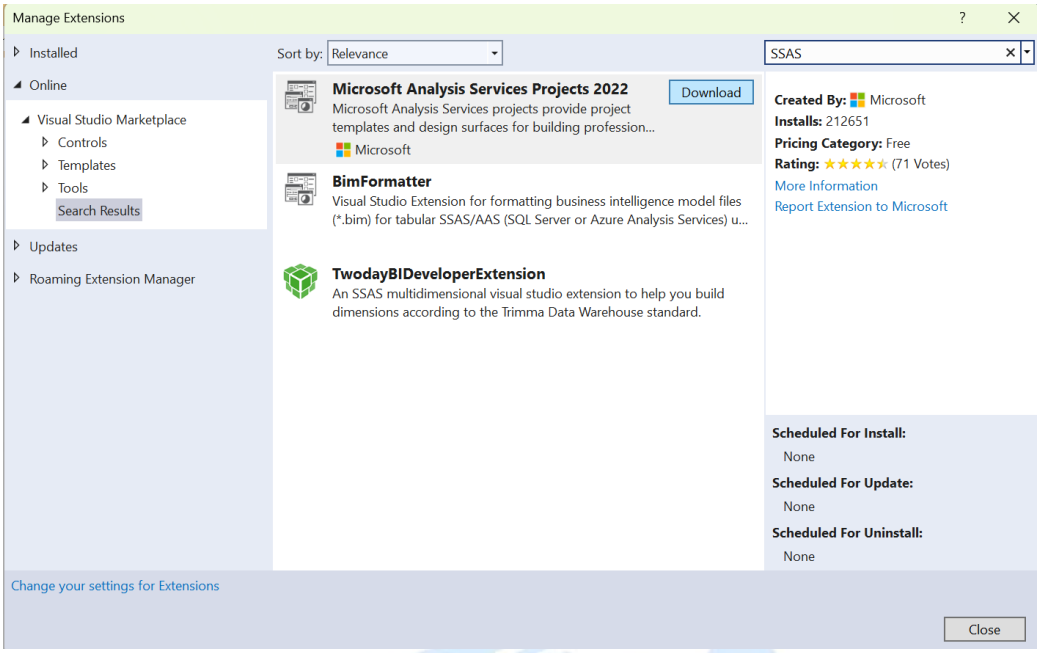
Bước 1: Mở **Visual Studio 2022** và chọn "Continue without code".



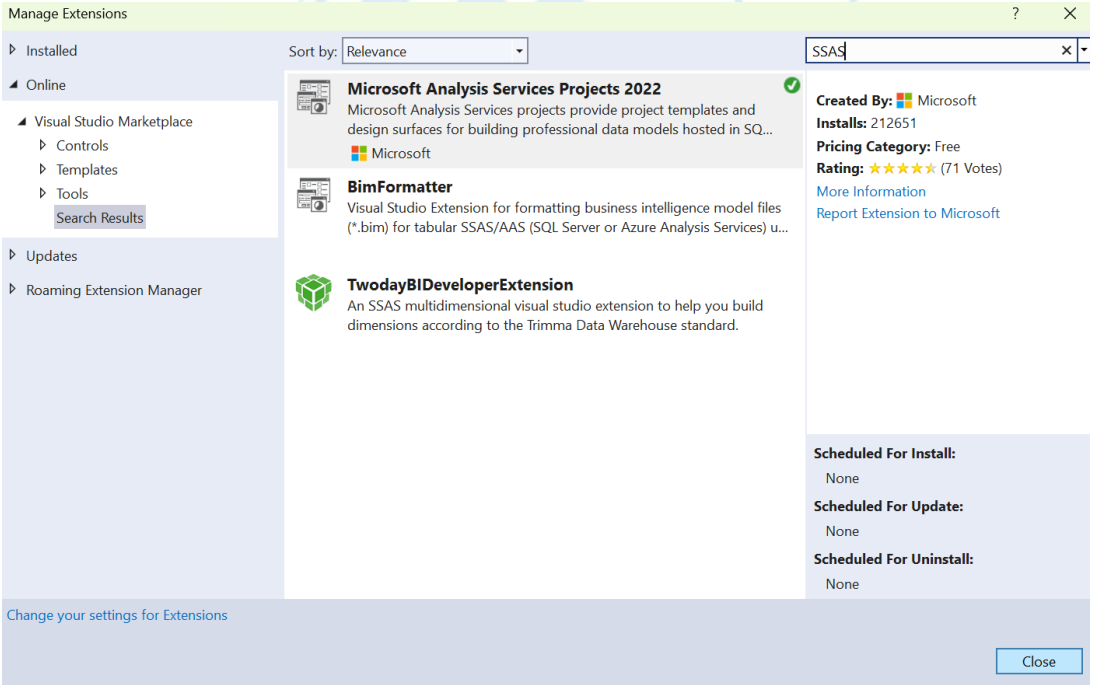
Bước 2: Tiếp theo, nhấn chọn "Extensions" > "Manage Extensions" ở giao diện chính.



Bước 3: Tiềm kiếm và nhấn “Download” để tải về công cụ **Microsoft Analysis Services Projects**.



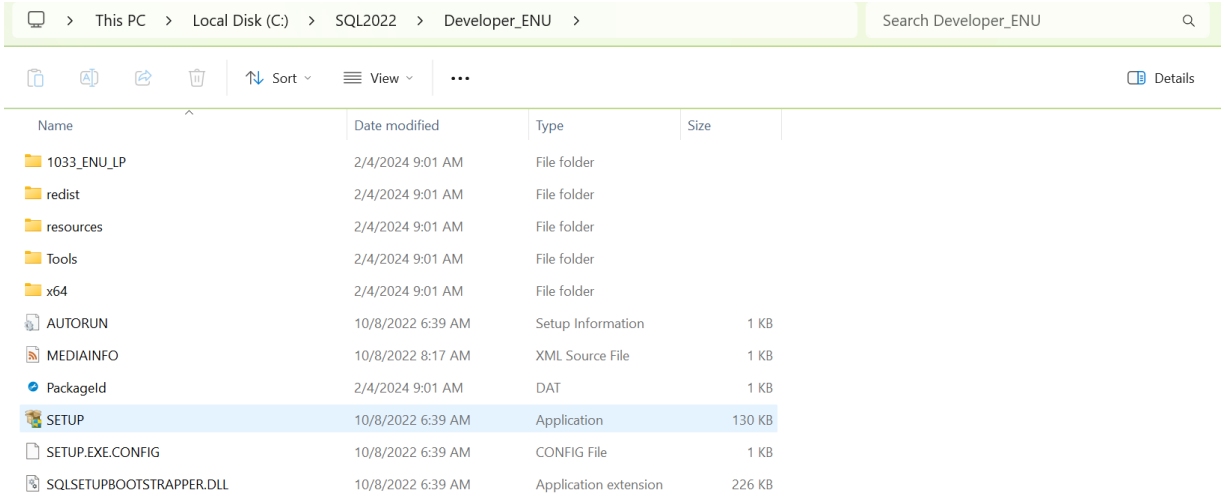
Bước 4: Sau đó, ta mở công cụ đã tải về và thực hiện các thao tác để hoàn tất quá trình cài đặt công cụ.



3.1.2 Cài đặt Analysis Services

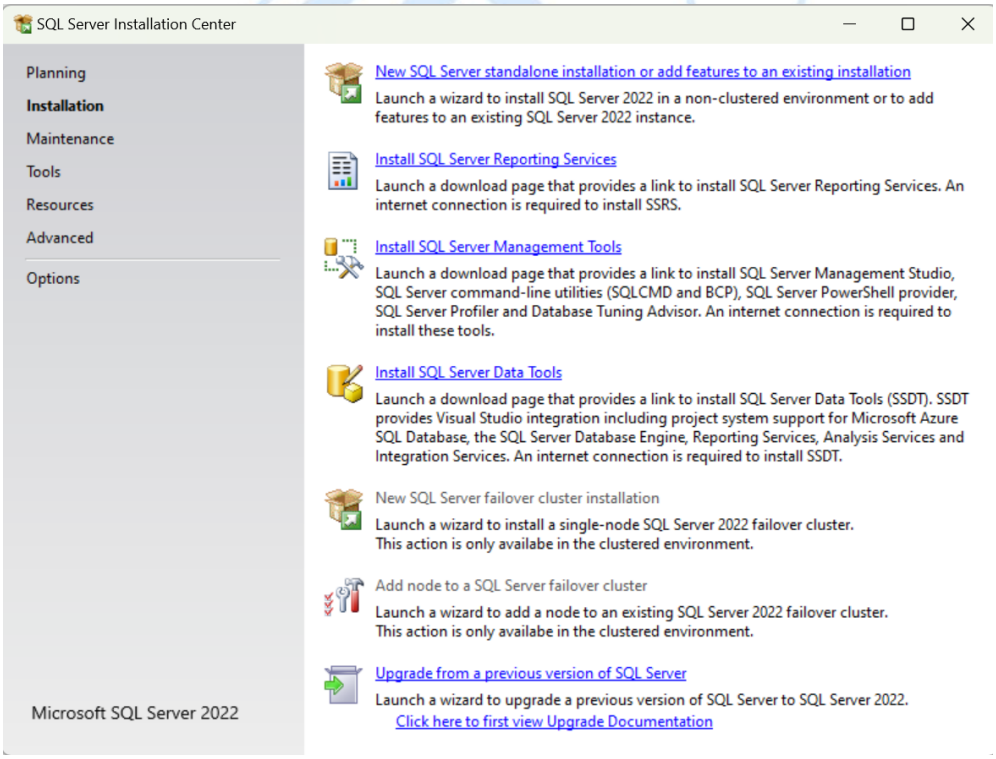
Bước 1: Mở **SQL Server Installation Center**. Cửa sổ này sẽ tự động được mở khi ta cài đặt **MS SQL Server**.

Nếu đã cài đặt MS SQL Server, ta vào thư mục của SQL Server (C:\SQL2022\Developer_ENU). Sau đó chọn **SETUP** để mở **SQL Server Installation Center**.

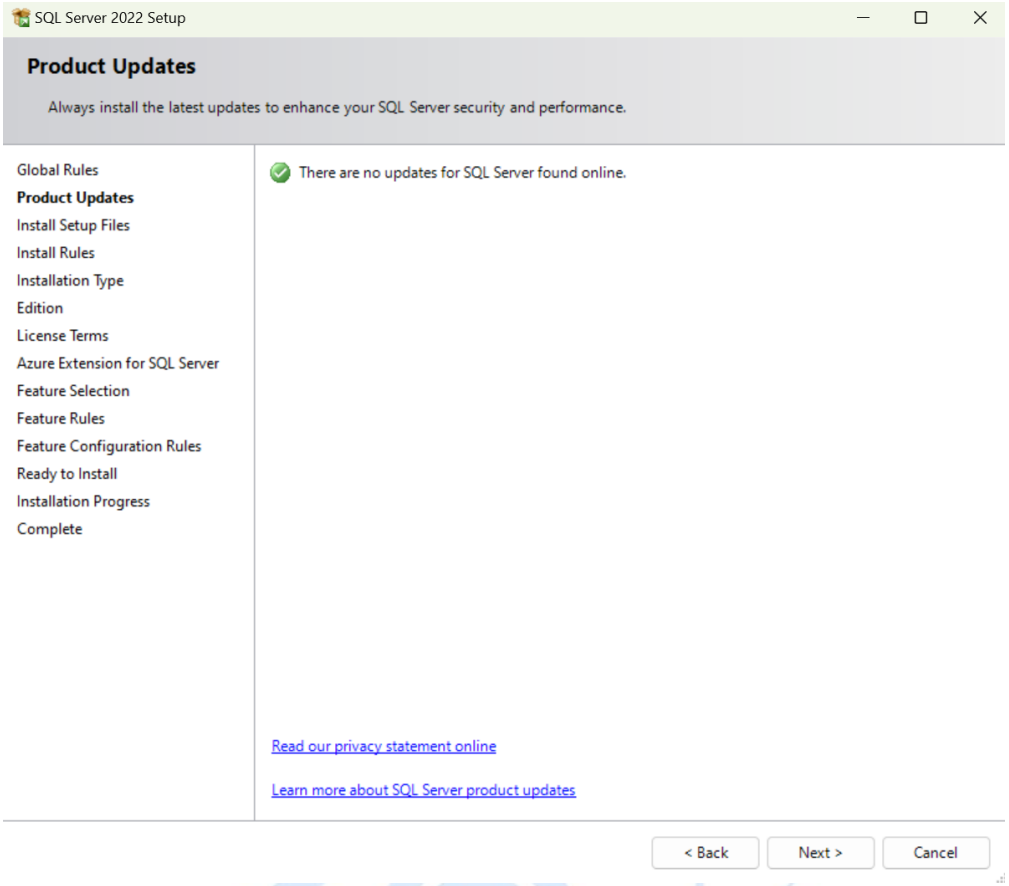


Bước 2: Chọn mục **Installation**, tại mục **Installation** nhấn chọn *New SQL Server standalone*

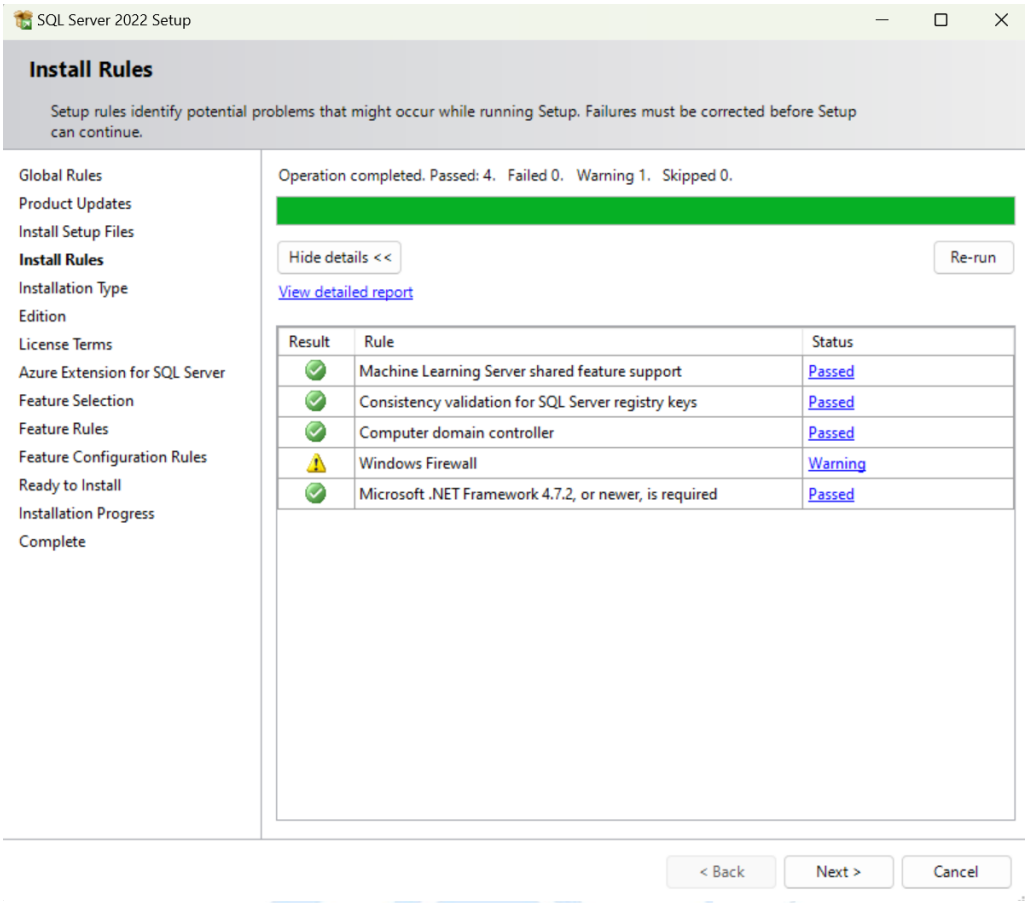
or add features to an existing installation.



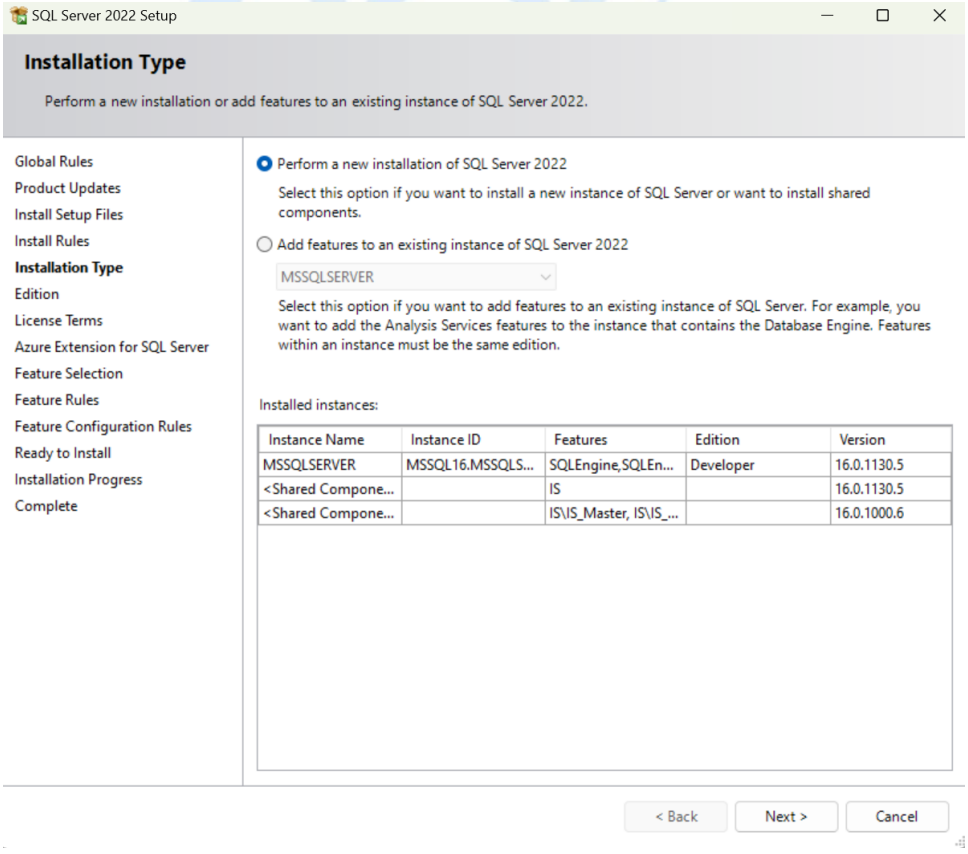
Bước 3: Hộp thoại mới xuất hiện, ta chọn "Next".



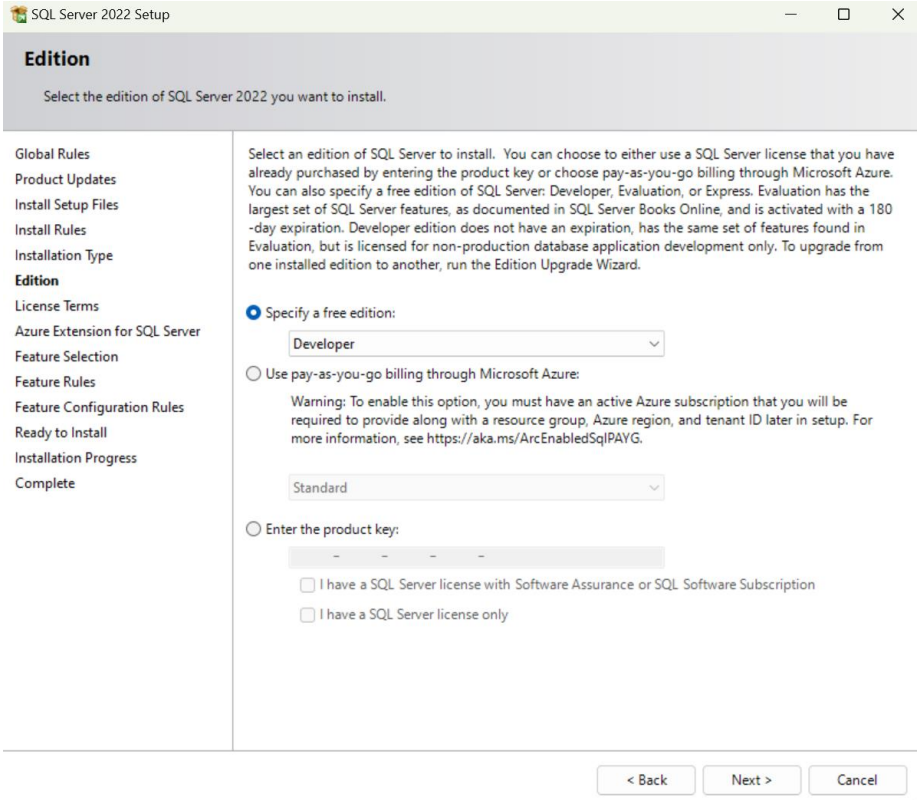
Bước 4: Sau đó tiếp tục chọn “Next”.



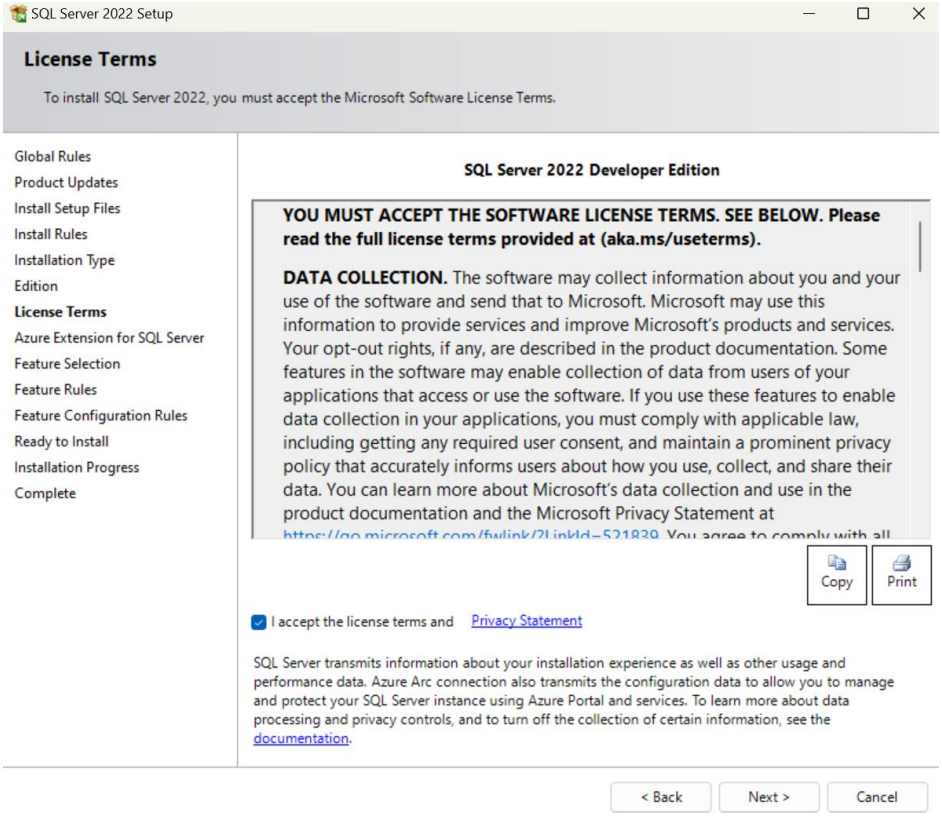
Bước 5: Tiếp tục chọn "Next".



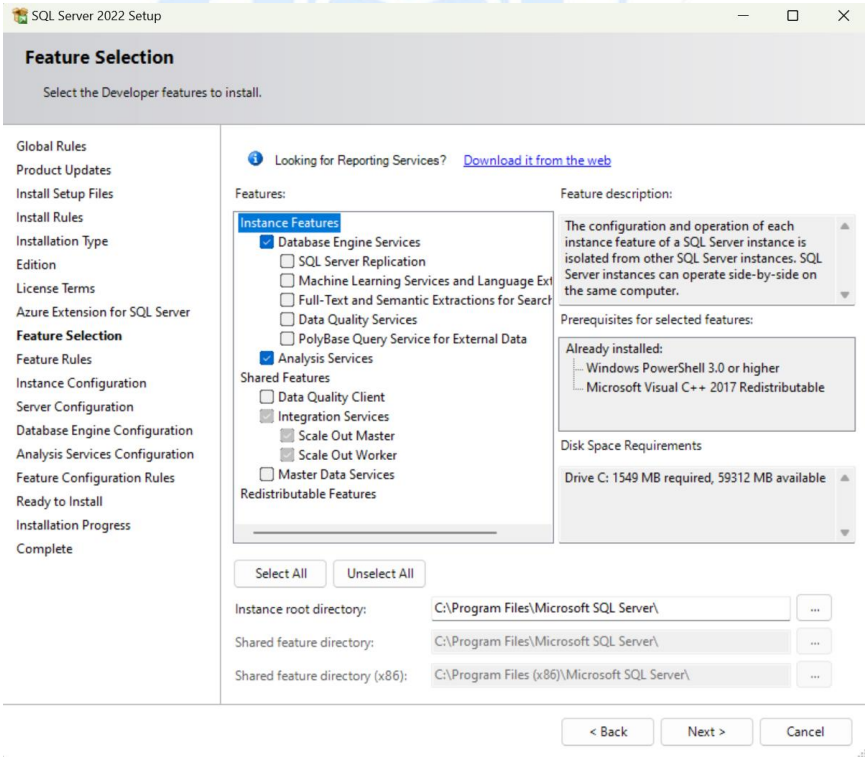
Bước 6: Tiếp tục chọn “Next”.



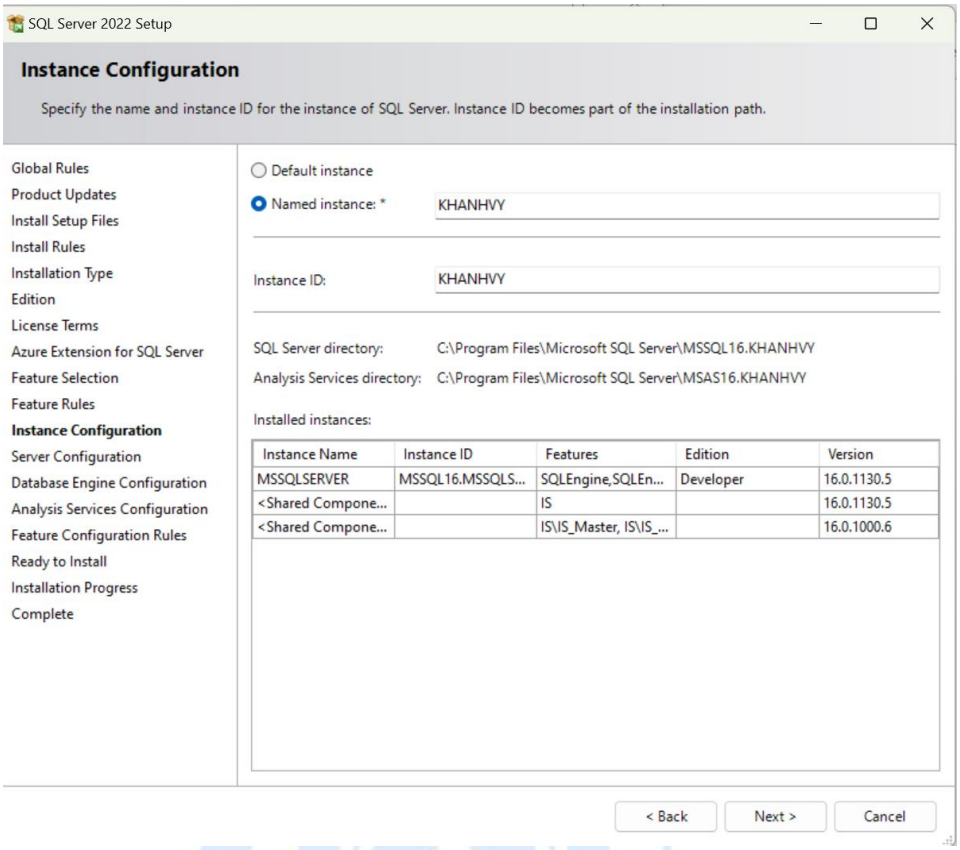
Bước 7: Chọn *I accept license terms and Privacy Statement*. Sau đó tiếp tục chọn “Next”.



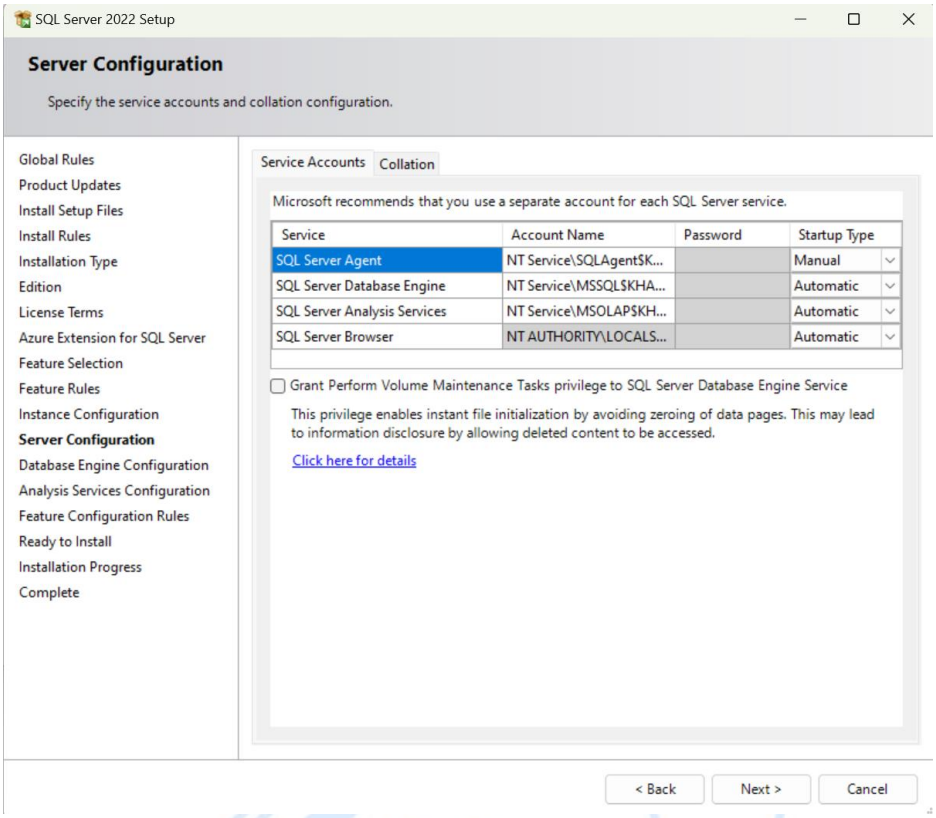
Bước 8: Ta chọn **Database Engine Services** và **Analysis Services**. Sau đó tiếp tục chọn "Next".



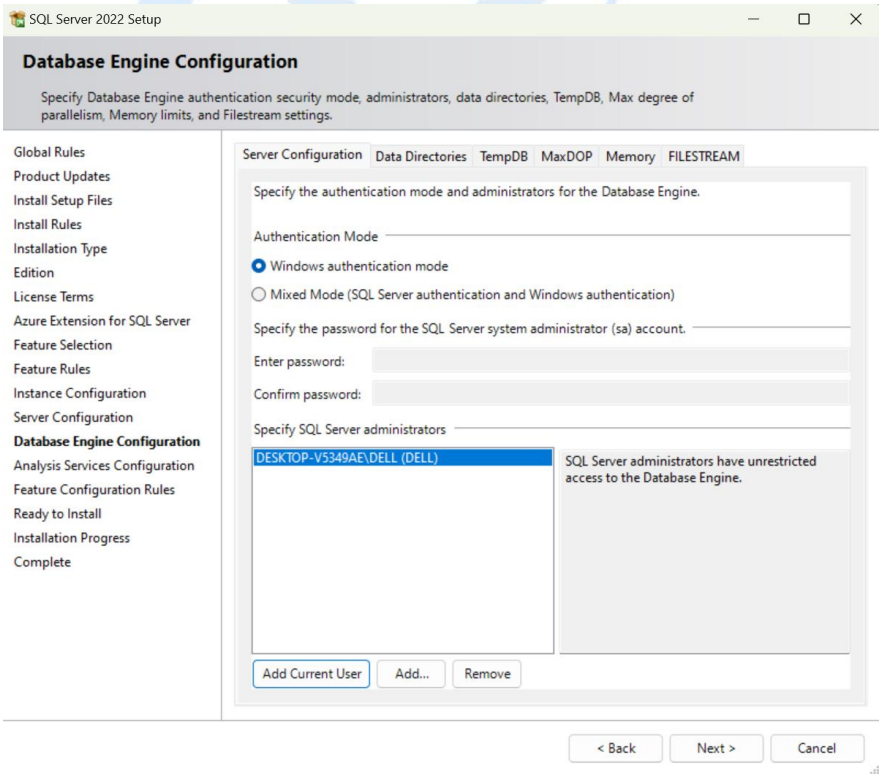
Bước 9: Chọn **Name instance**, sau đó ta đặt tên cho instance vừa tạo. Sau đó nhấn "Next".



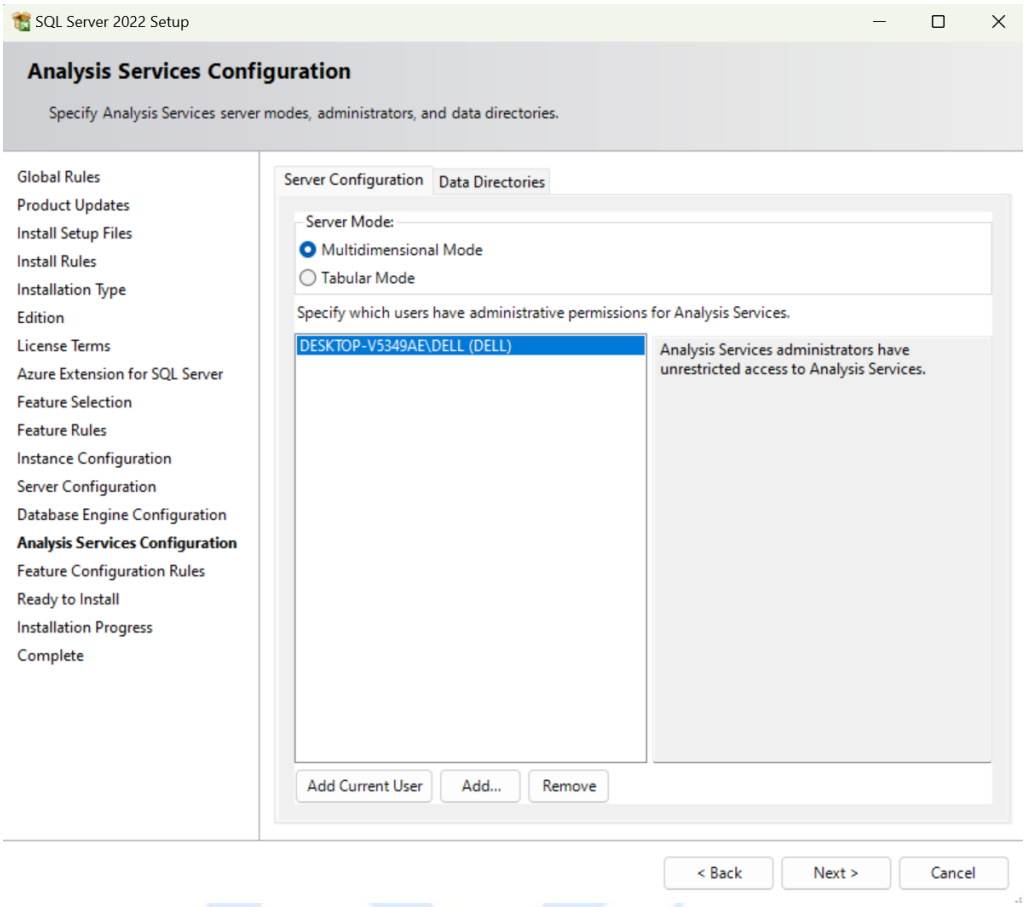
Bước 10: Tiếp tục chọn "Next".



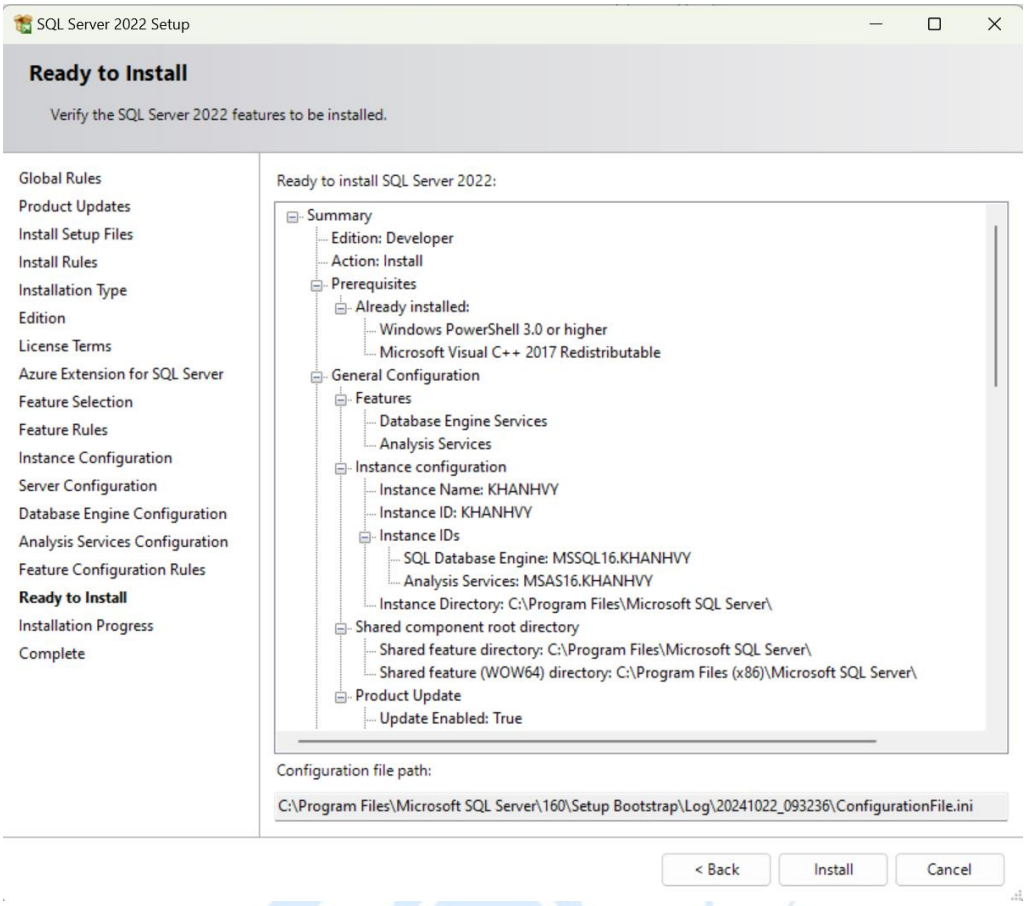
Bước 11: Chọn *Windowns authentication mode*. Tiếp theo, chọn **Add Current User**. Sau khi user xuất hiện, ta chọn vào tên user và chọn "Next".



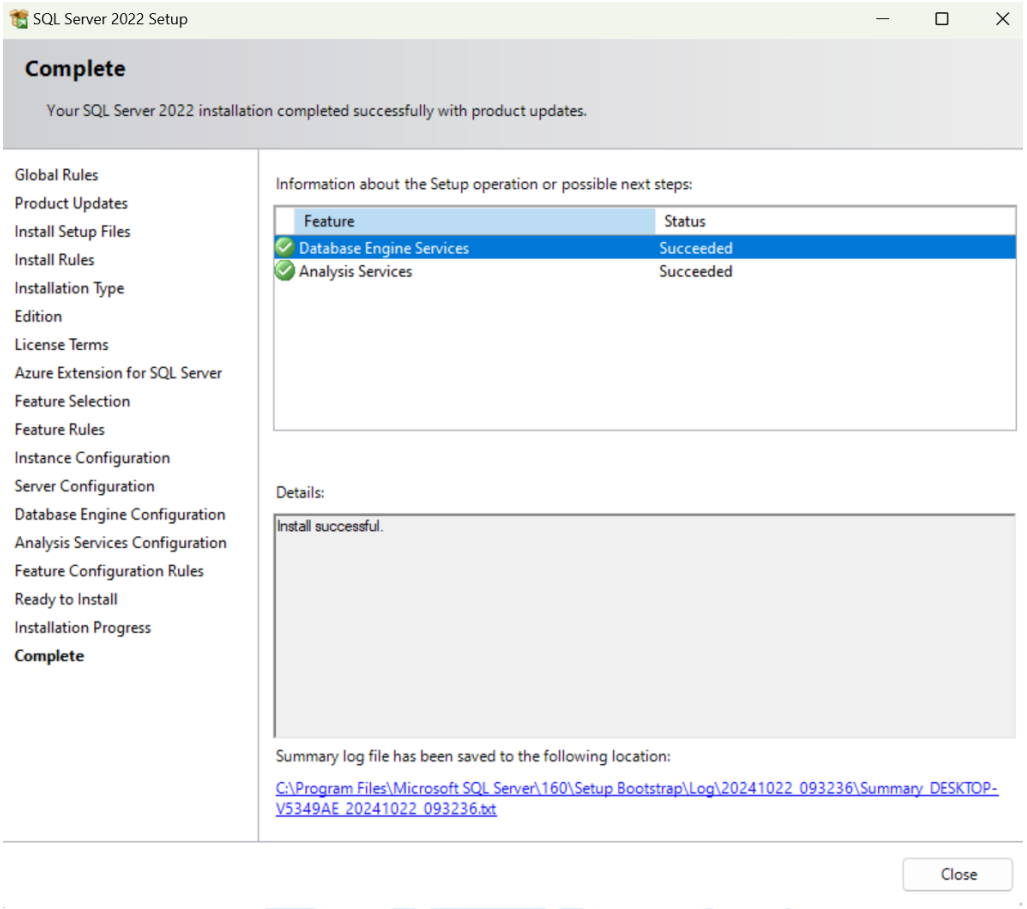
Bước 12: Chọn **Multidimensional Mode**. Tiếp theo, chọn **Add Current User**. Sau khi user xuất hiện, ta chọn vào tên user và chọn Next.



Bước 13: Chọn **Install** để tiến hành cài đặt.

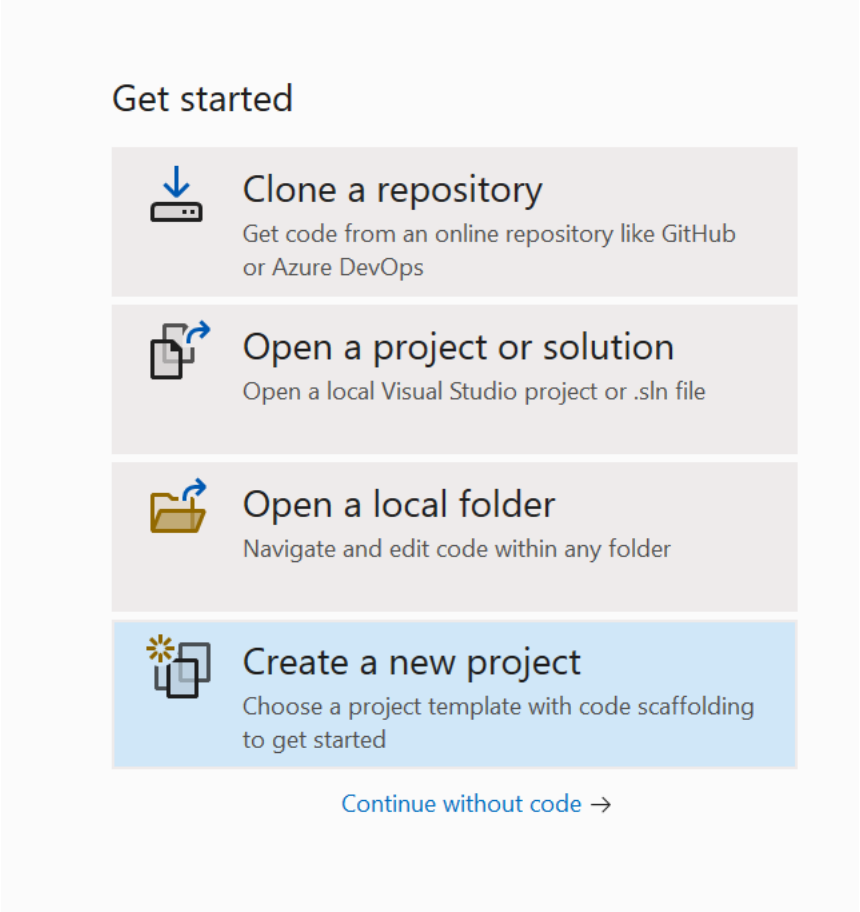


Bước 14: Nhấn “Close” để hoàn tất quá trình cài đặt.

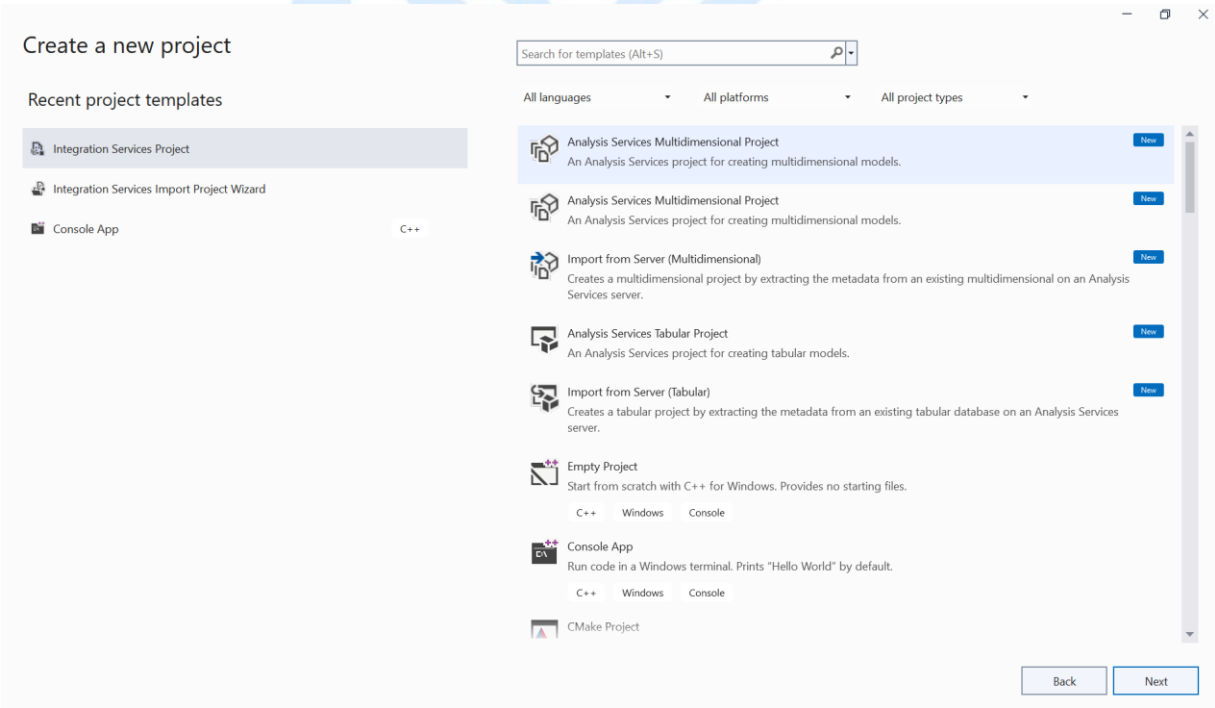


3.2 Tạo mới Project SSAS

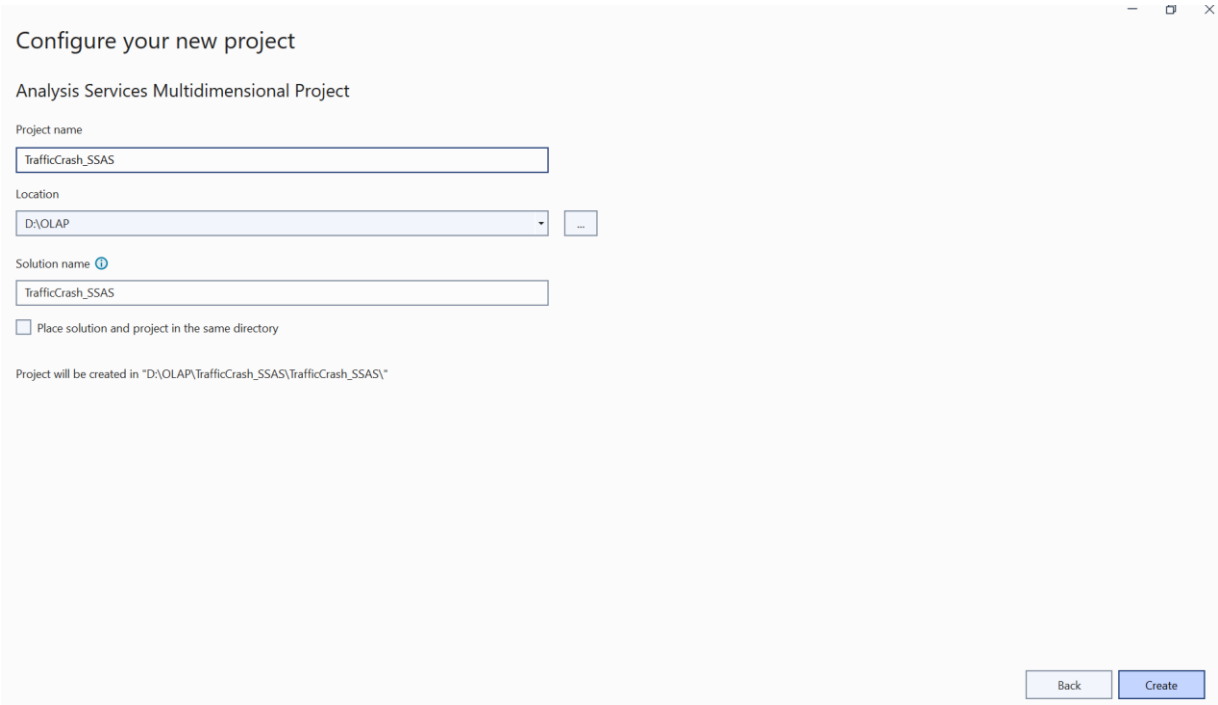
Bước 1: Mở **Visual Studio 2022** và chọn "Create a new project".



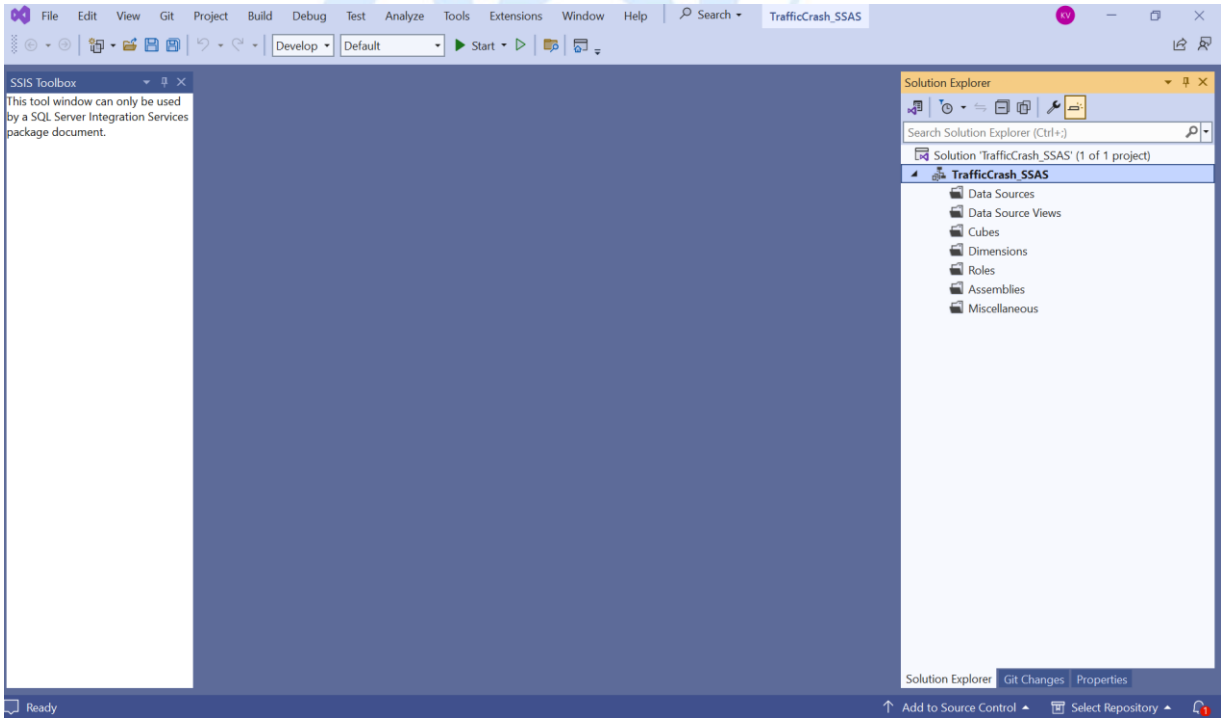
Bước 2: Tiếp theo, ta chọn **Analysis Services Multidimensional Project** và chọn "Next".



Bước 3: Nhập tên và thiết lập đường dẫn cho Project. Sau đó chọn "Create".

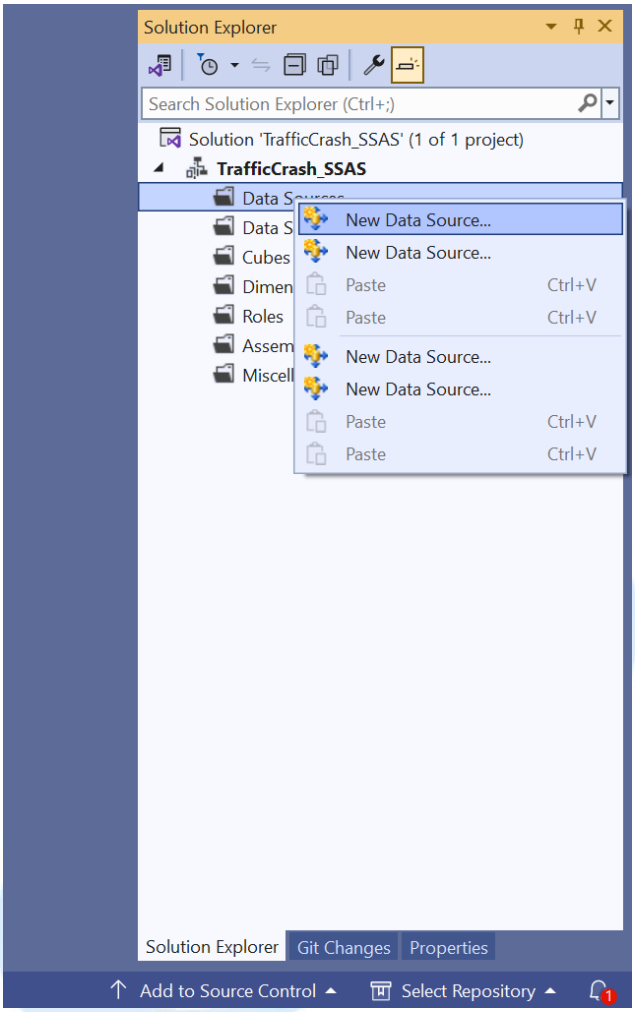


Bước 4: Giao diện quá trình SSAS hiện ra.

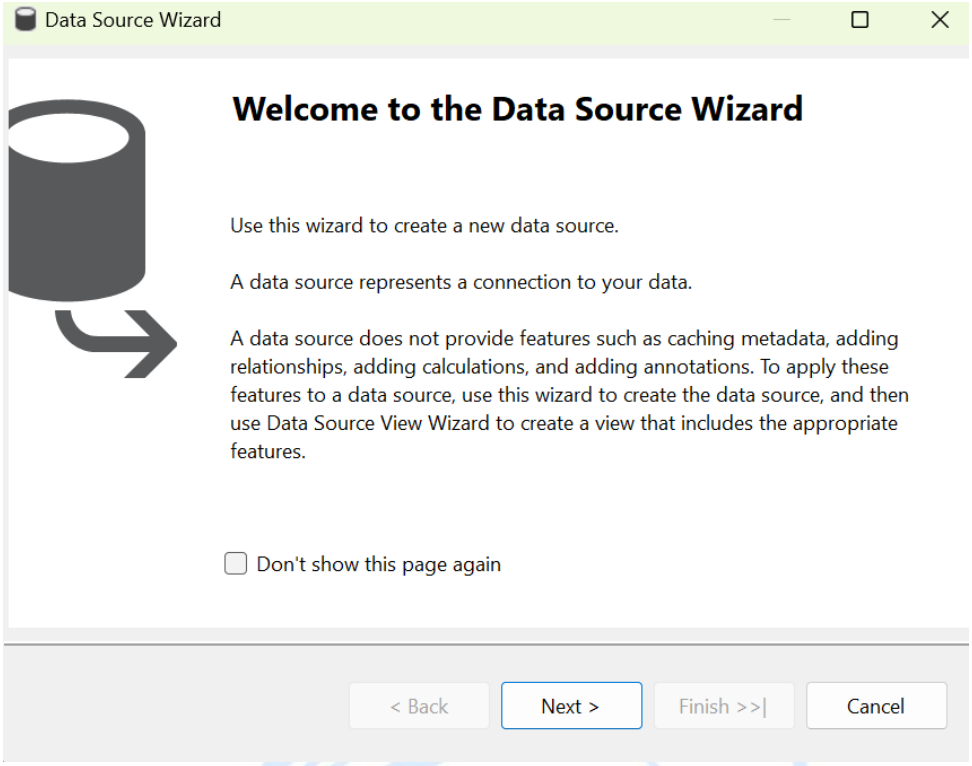


3.3 Xác định dữ liệu nguồn (Data Sources)

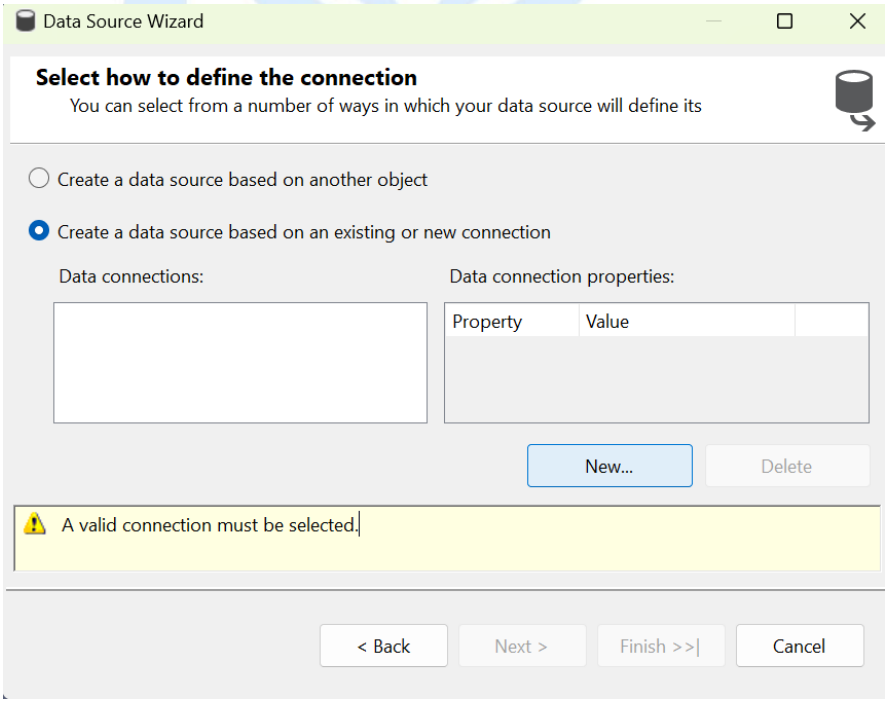
Bước 1: Đầu tiên ở phần **Solution Explorer** (góc phải màn hình) nhấn chuột phải vào **Data Sources** và chọn **New Data Source**.



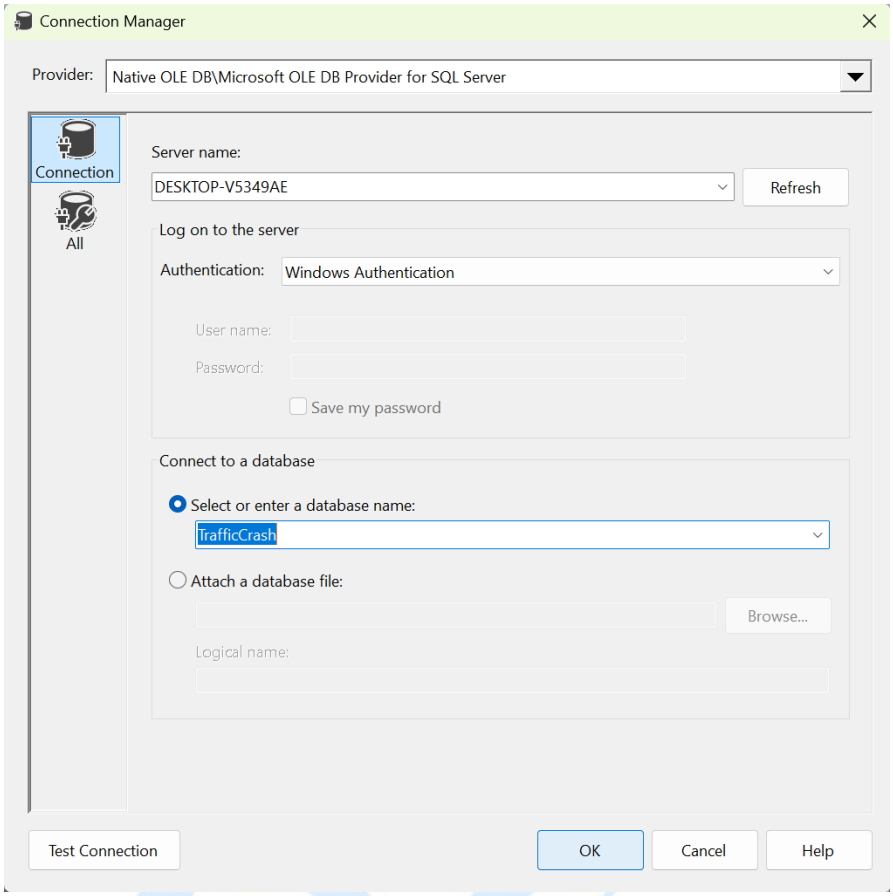
Bước 2: Hộp thoại **Data Source Wizard** xuất hiện, chọn "Next" để tiếp tục.



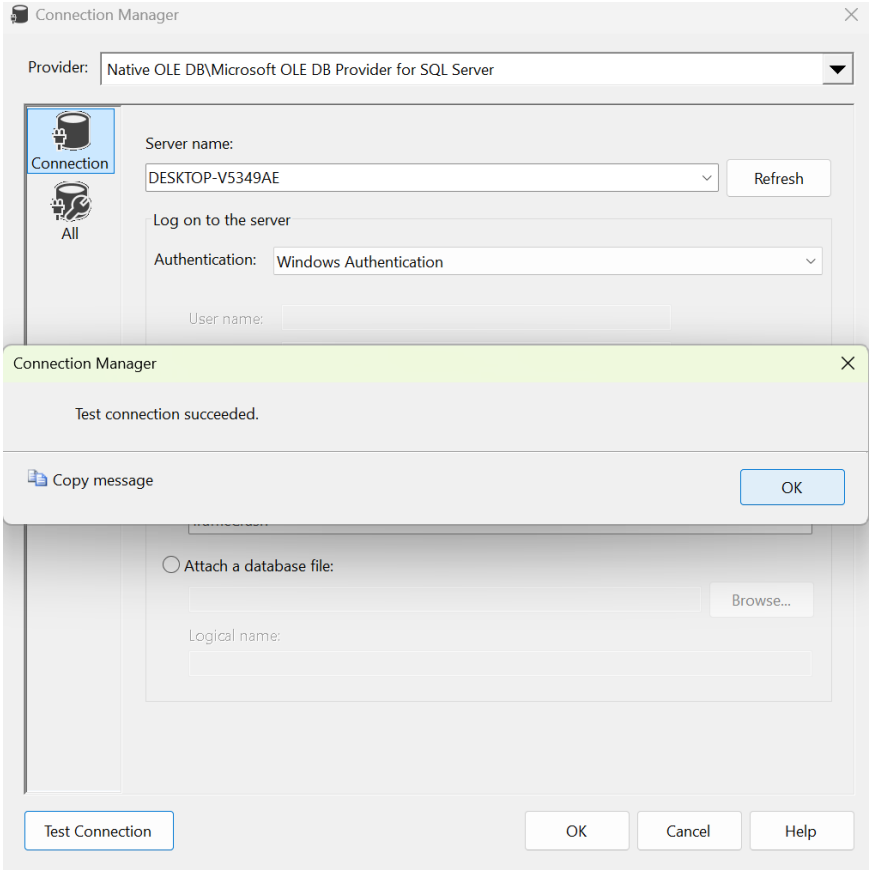
Bước 3: Ở trang **Select how to define the connection**, chọn "Create a data source based on an existing or new connection", chọn "New" để tạo mới một **Data connection**.



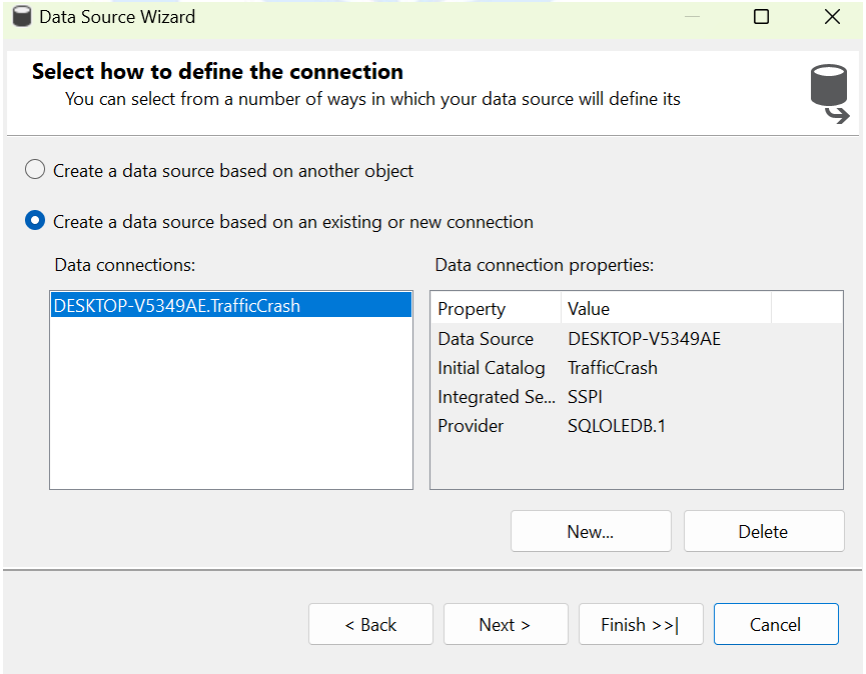
Bước 4: Hộp thoại **Connection Manager** xuất hiện, ta chọn **Server name** và cơ sở dữ liệu mà ta đã tạo ra từ quá trình SSIS.



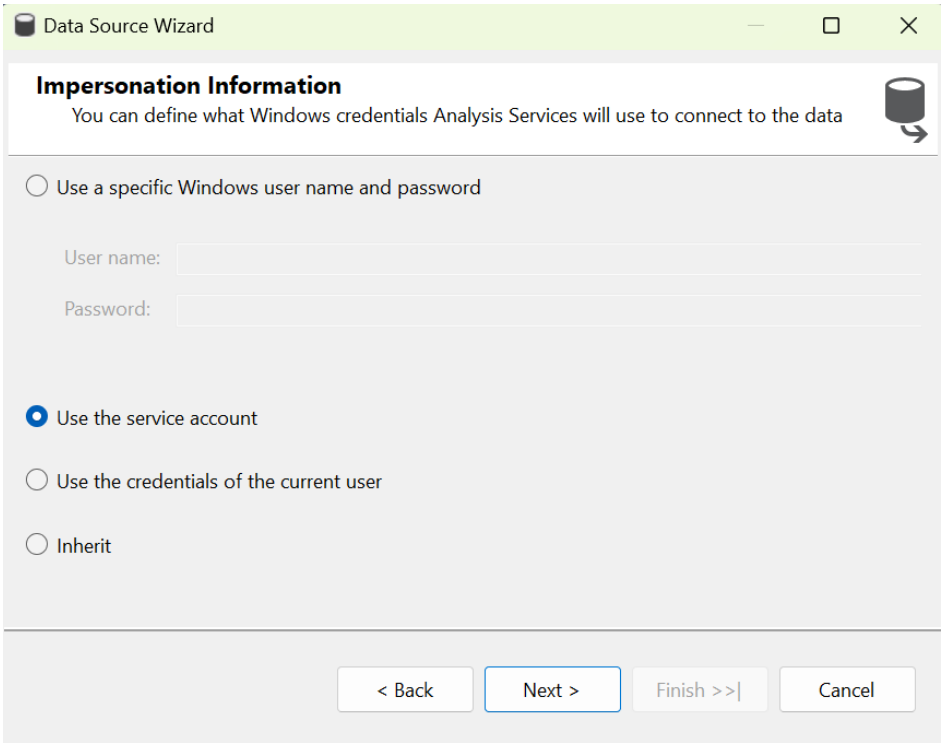
Nhấn chọn "Test" để kiểm tra. Sau đó nhấn "OK" để hoàn tất.



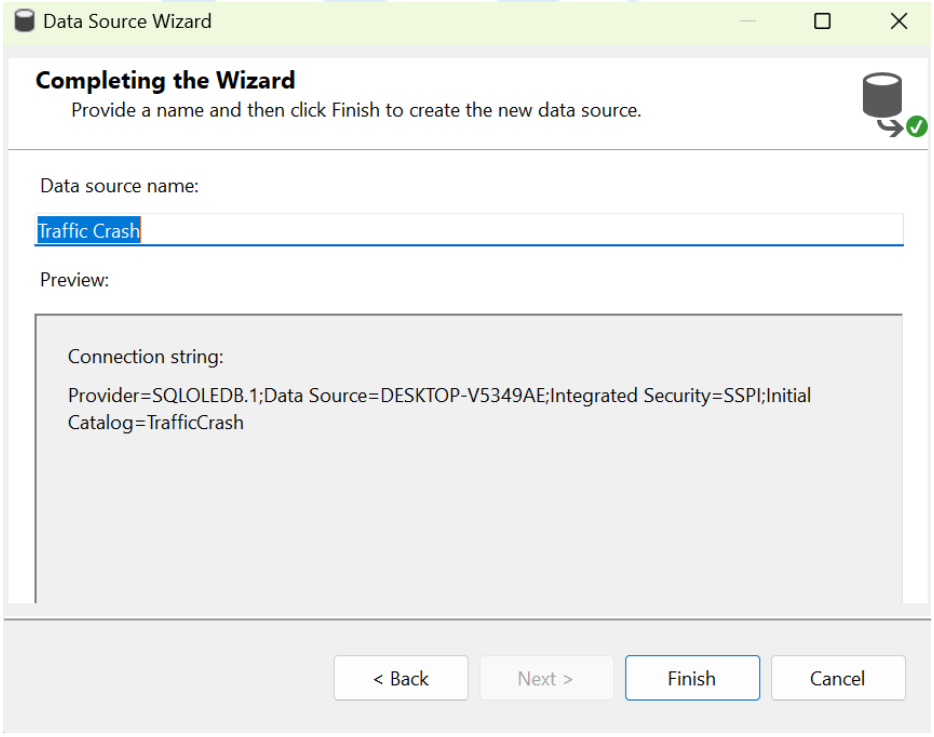
Bước 5: Chọn kết nối đã tạo từ quá trình SSIS rồi tiếp tục nhấn "Next".



Bước 6: Chọn **Use the service account** và nhấn "Next".

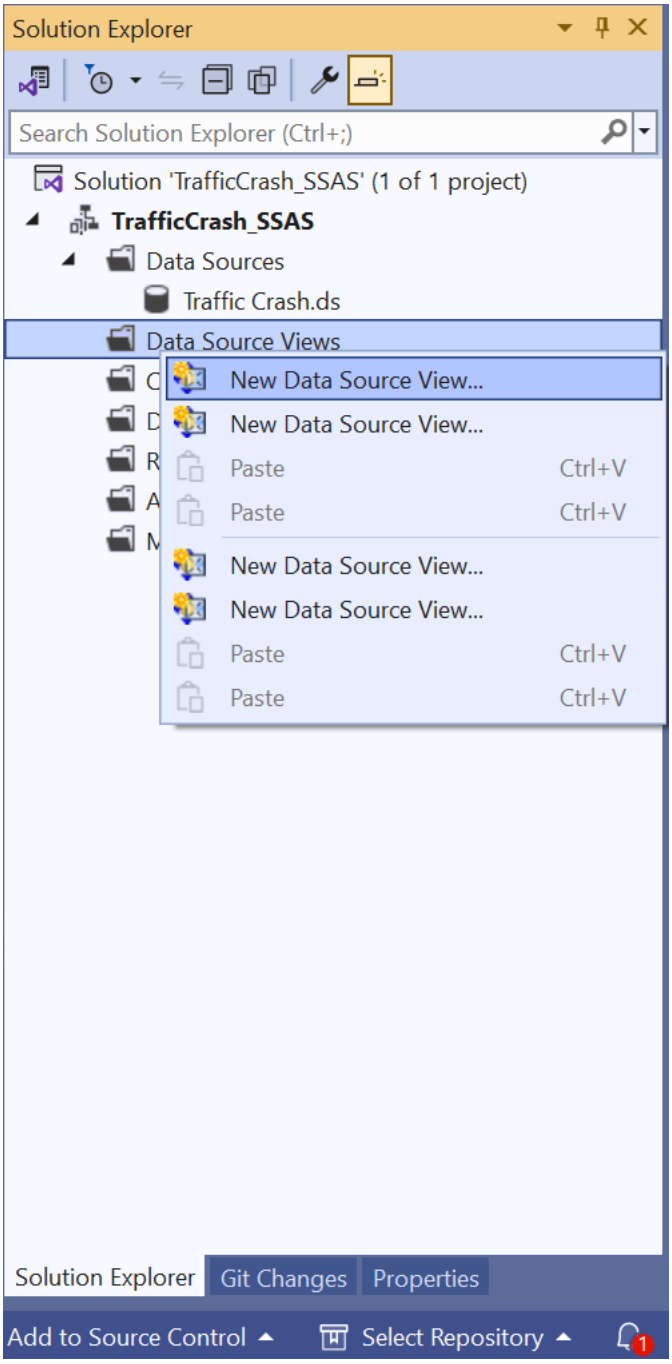


Bước 7: Nhấn "Finish" để hoàn tất.

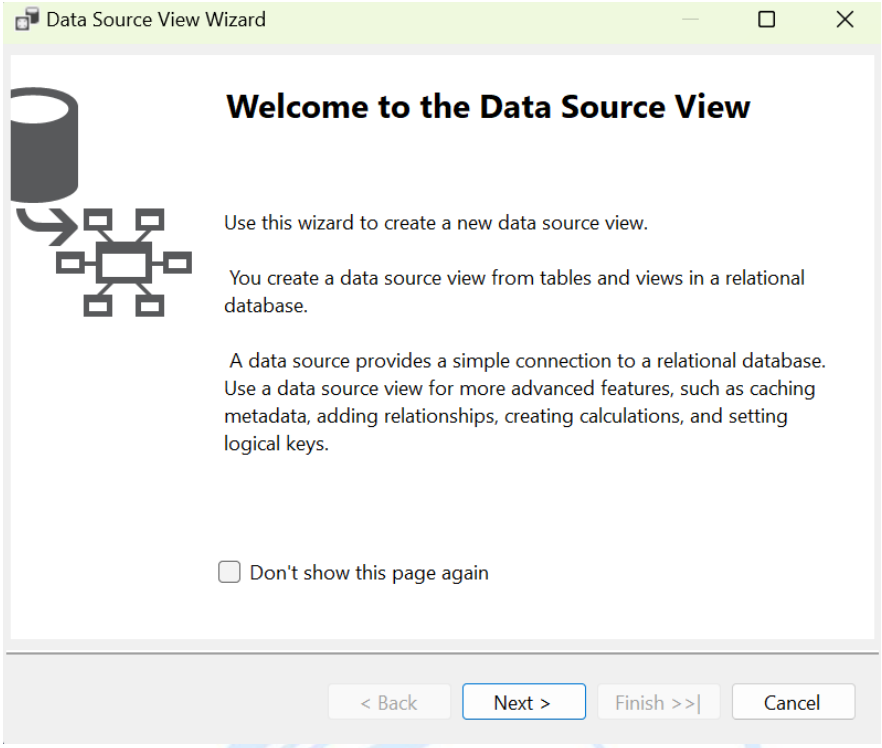


3.4 Xác định khung nhìn dữ liệu nguồn (Data Source Views)

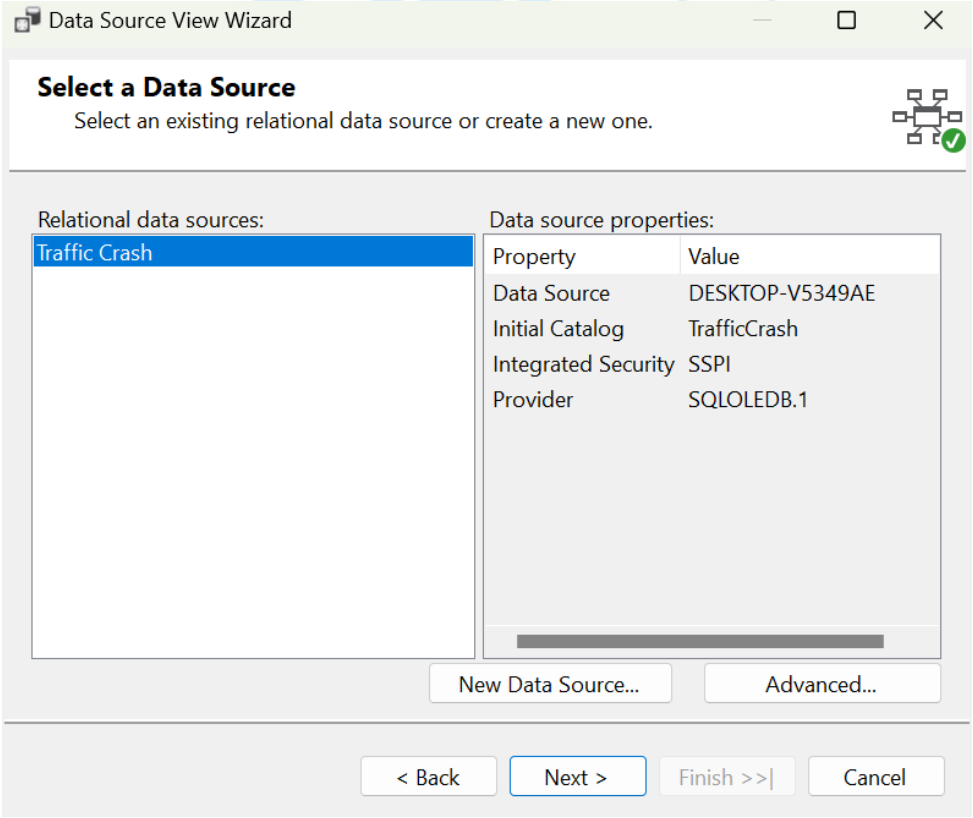
Bước 1: Ở phần **Solution Explorer** nhấn chuột phải vào **Data Source Views** và chọn **New Data Source Views**.



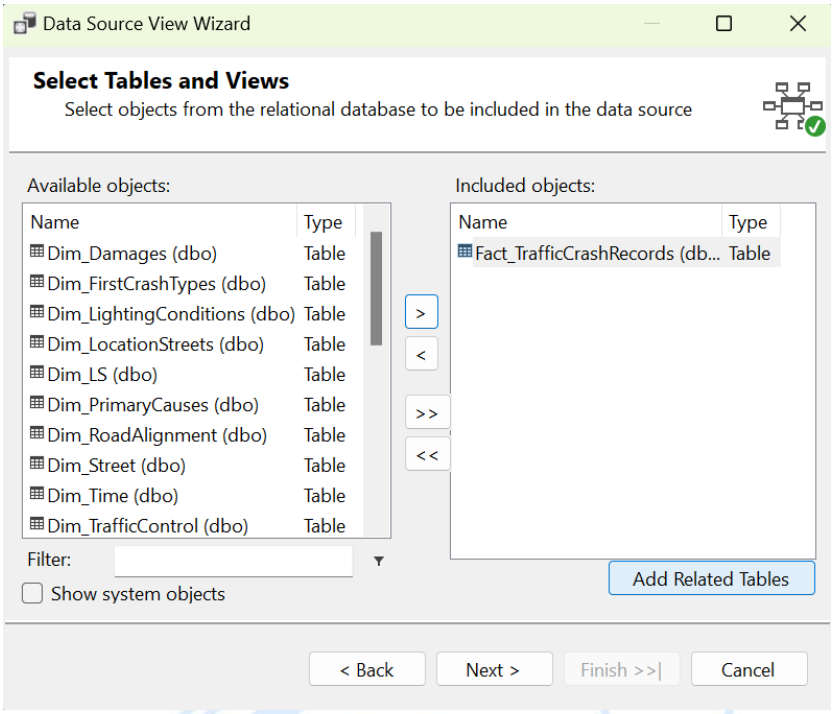
Bước 2: Hộp thoại **Data Source Wizard** xuất hiện, chọn "Next" để tiếp tục.



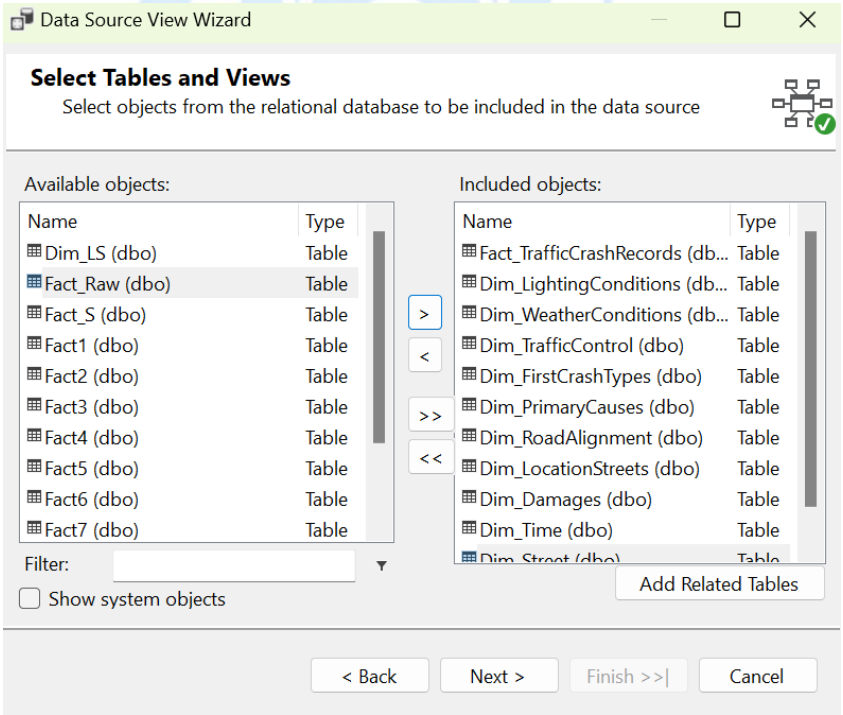
Bước 3: Chọn kho dữ liệu **Traffic Crash** và nhấn "Next".



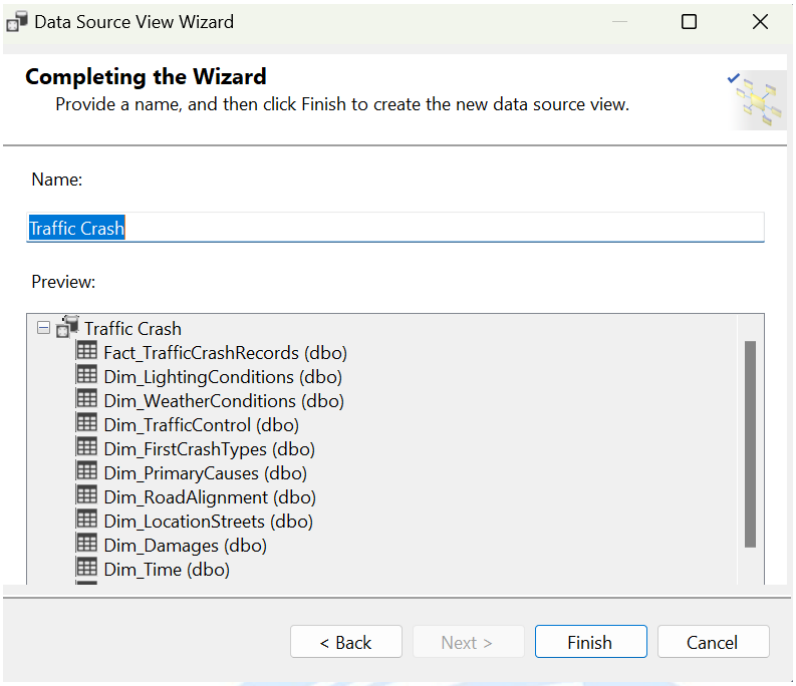
Bước 4: Chọn bảng **Fact_TrafficCrashRecords** và nhấn **Add Related Tables** để thêm các bảng **Dim** vào **Data source view**.



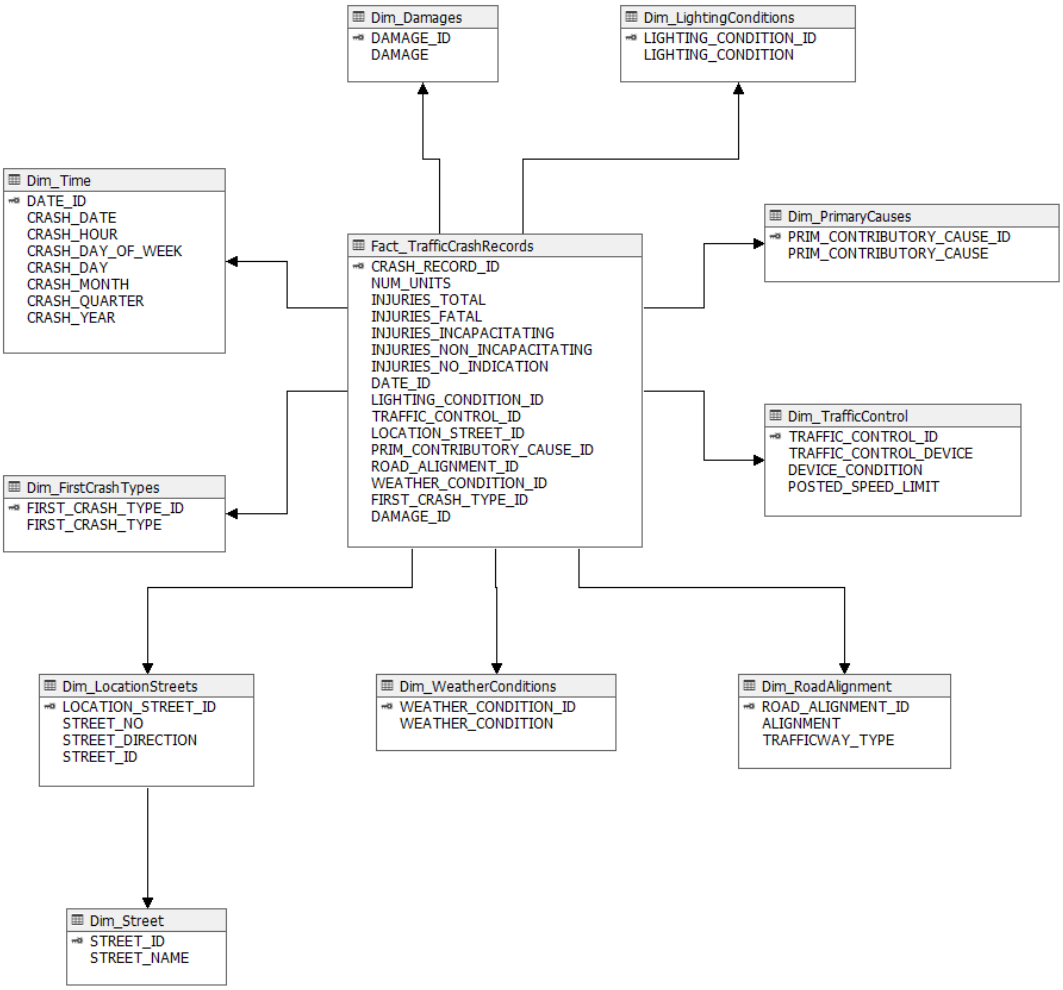
Bước 5: Nhấn "Next" để tiếp tục.



Bước 5: Chọn "Finish" để hoàn tất.



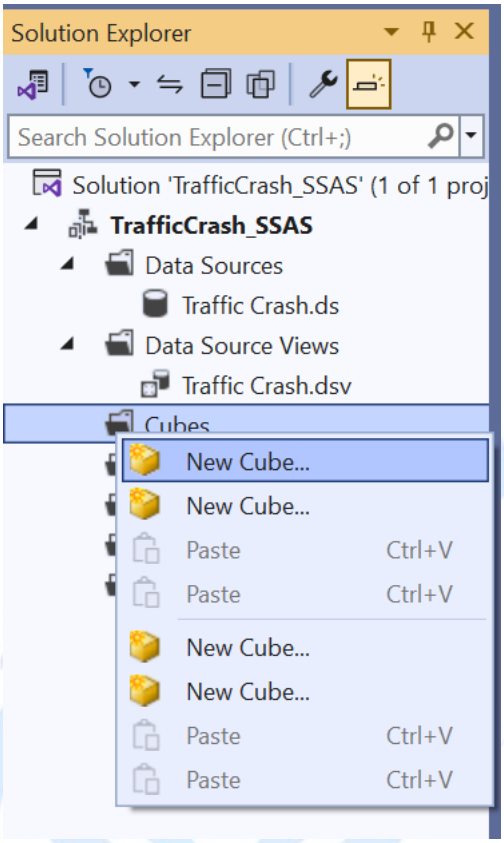
Sau khi kết thúc quá trình này, ta sẽ được Data Source View như hình sau:



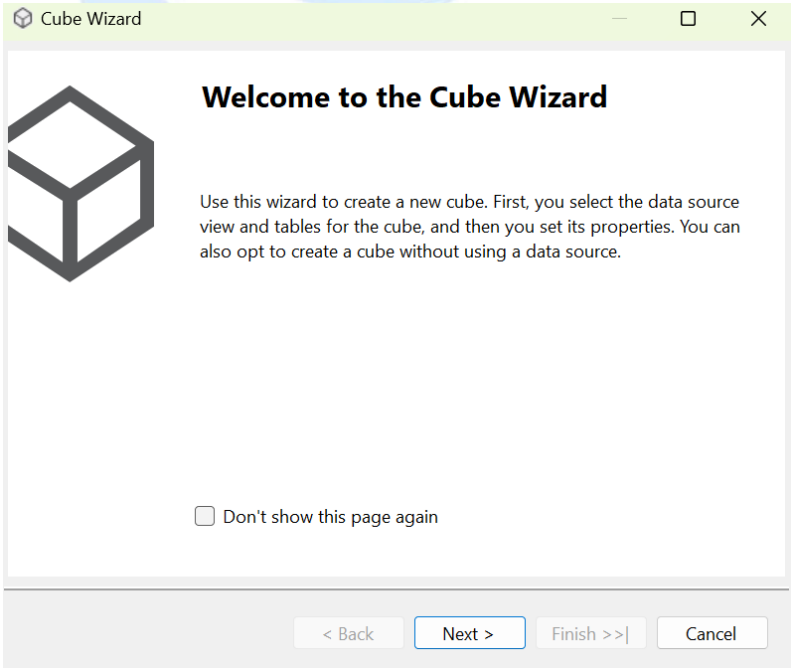
3.5 Xây dựng các khối (Cube) và deploy Cube

3.5.1 Tạo Cube và Dimension

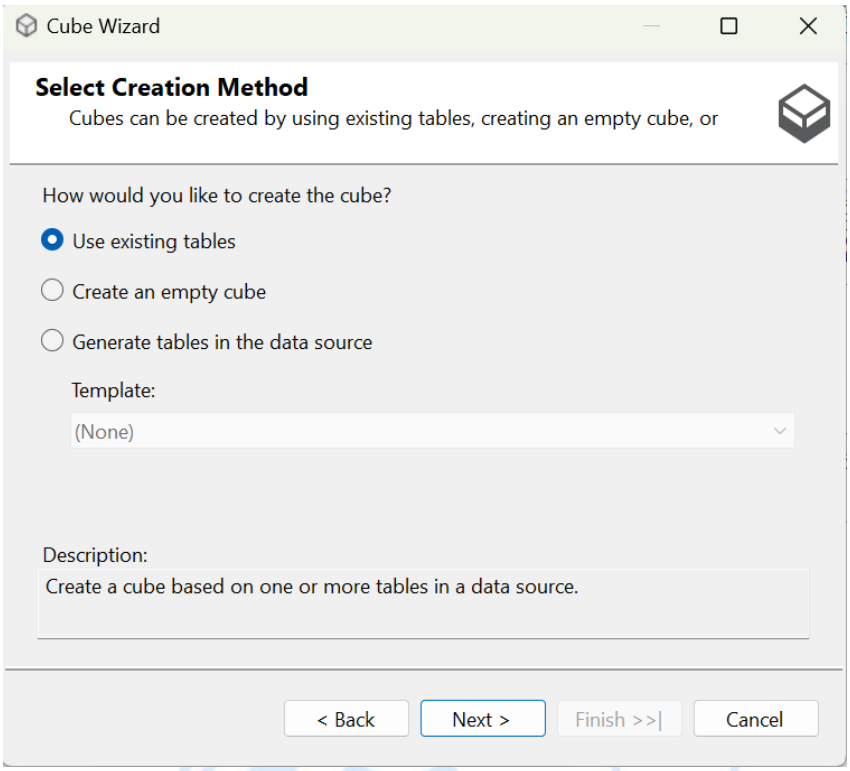
Bước 1: Tại **Solution Explorer**, nhấn phải chuột vào thư mục **Cubes** và chọn **New Cube**.



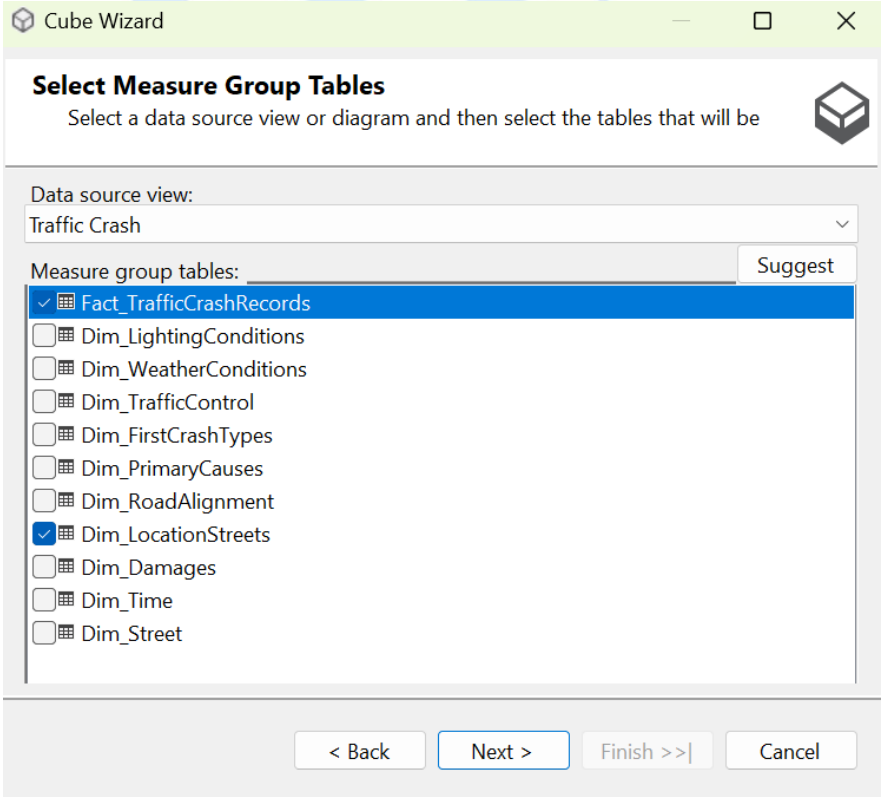
Bước 2: Hộp thoại **Cube Wizard** xuất hiện, chọn "Next" để tiếp tục.



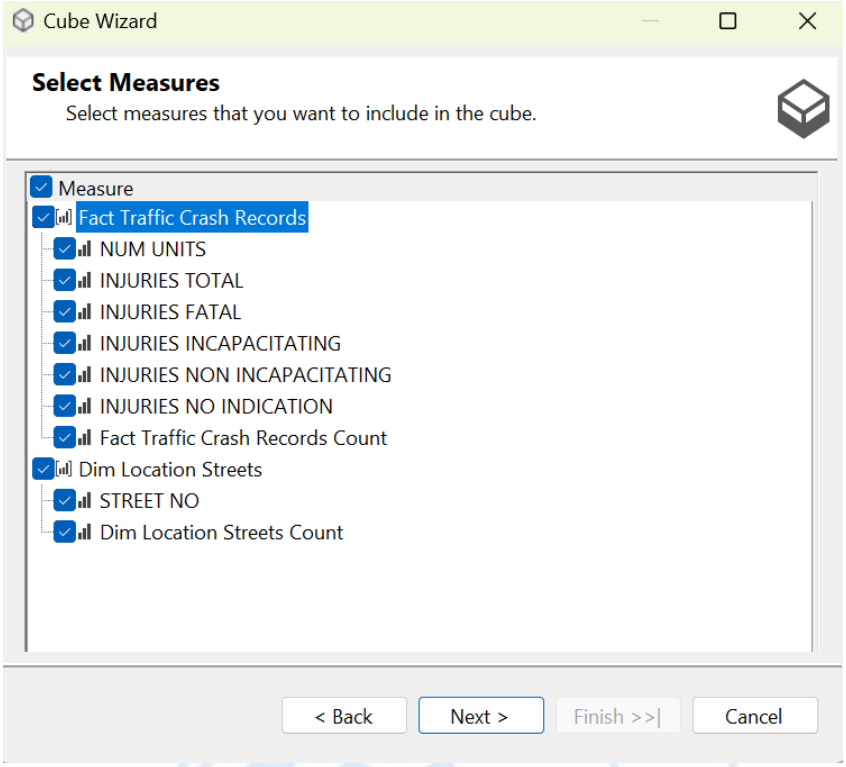
Bước 3: Chọn “Use existing tables”, sau đó chọn “Next” để tiếp tục.



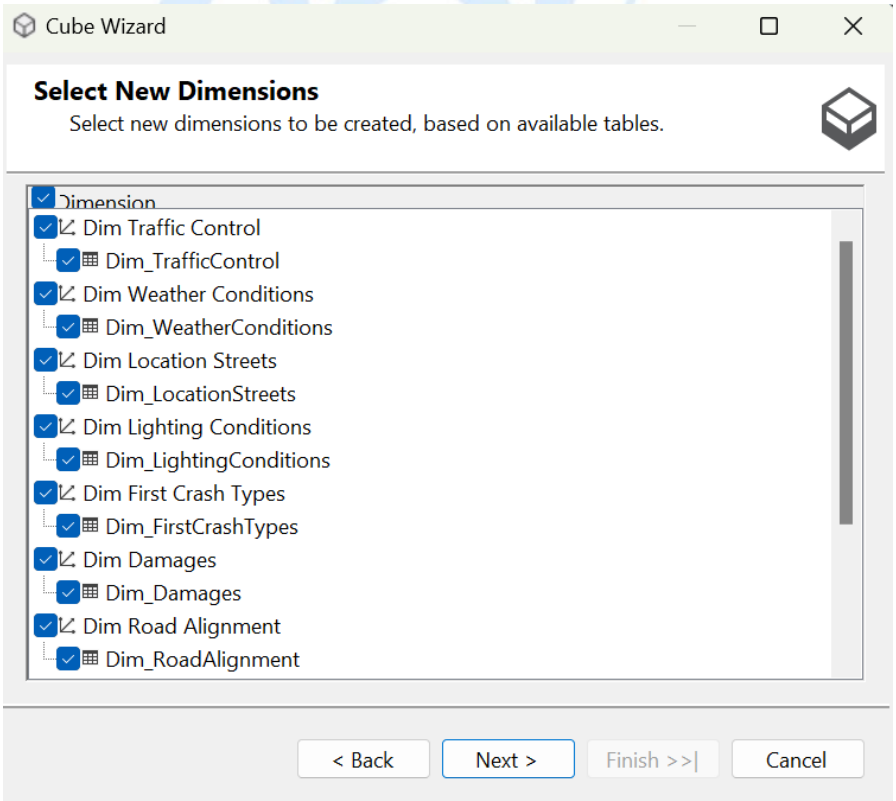
Bước 4: Tiếp tục chọn bảng **Fact_TrafficCrashRecords** và bảng **Dim_LocationStreets** để phân chia các measure group.và nhấn “Next”.



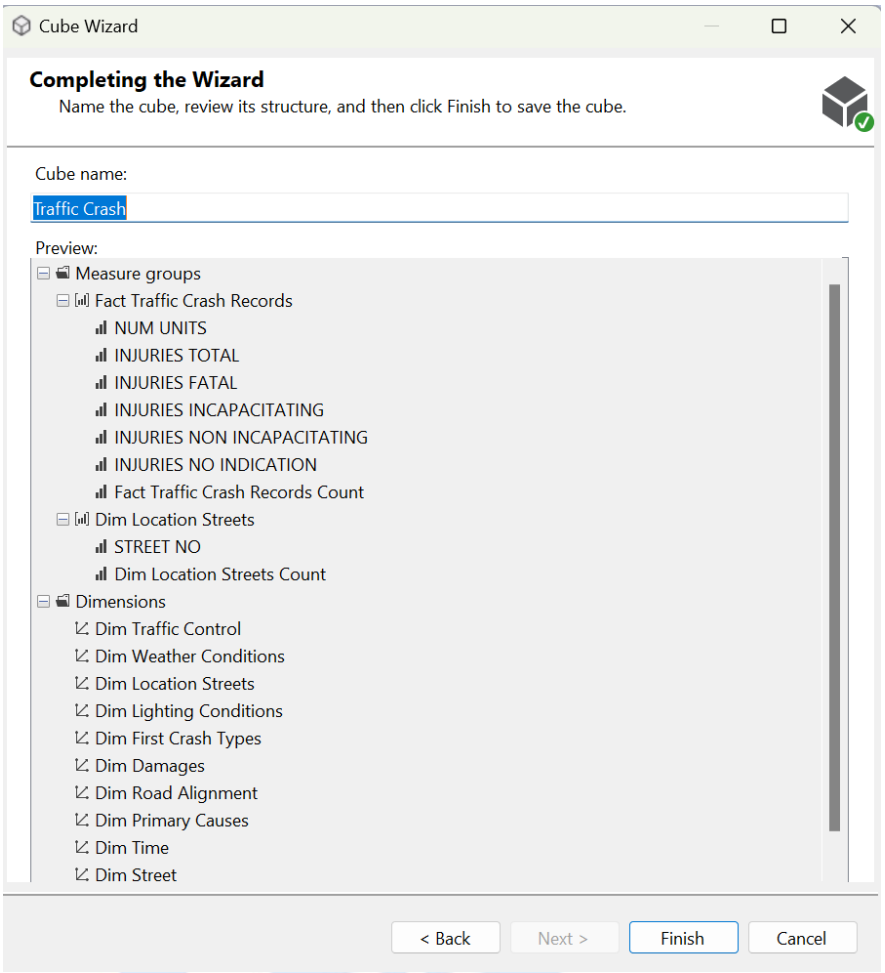
Bước 5: Chọn những độ đo đề xuất, sau đó chọn "Next" để tiếp tục.



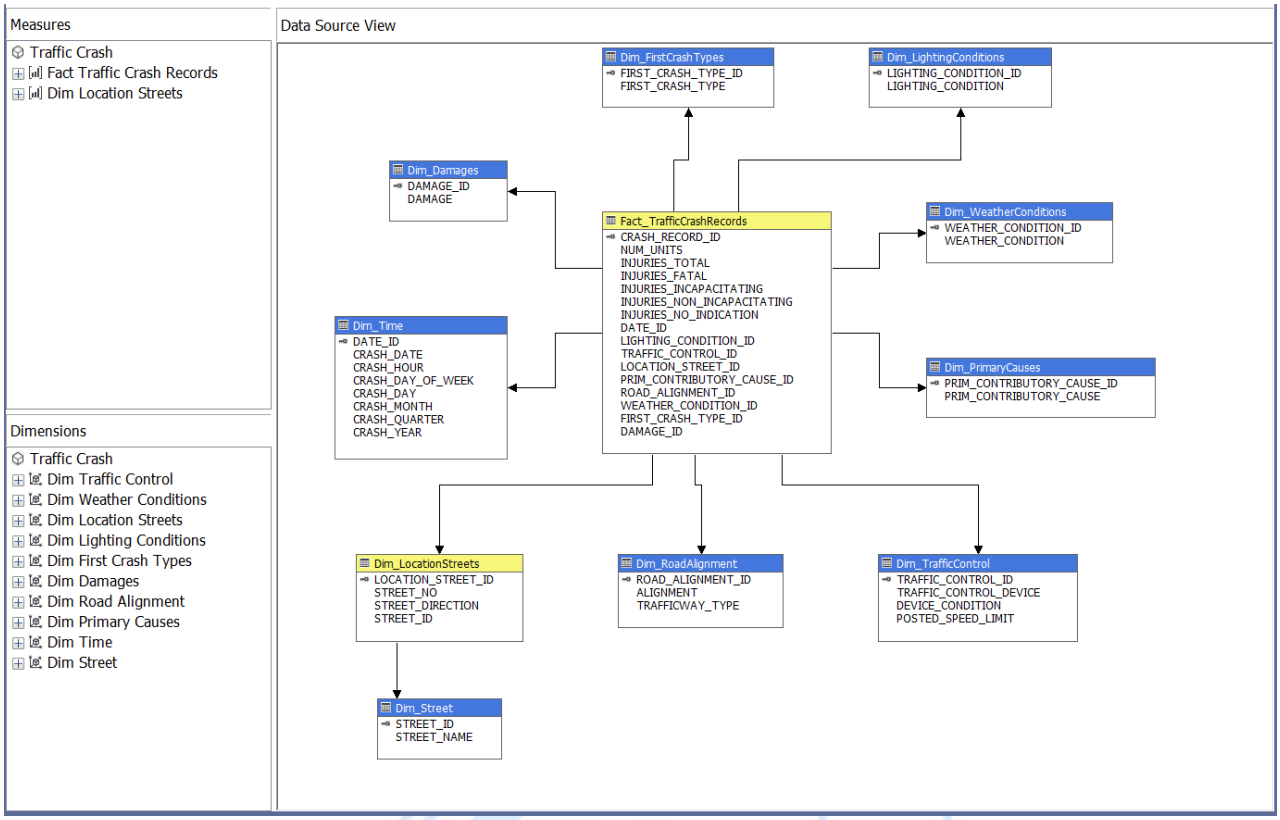
Bước 5: Để mặc định và nhấn "Next".



Bước 6: Xem lại các độ đo được tạo, các bảng Dimension và chọn "Finish" để hoàn tất.

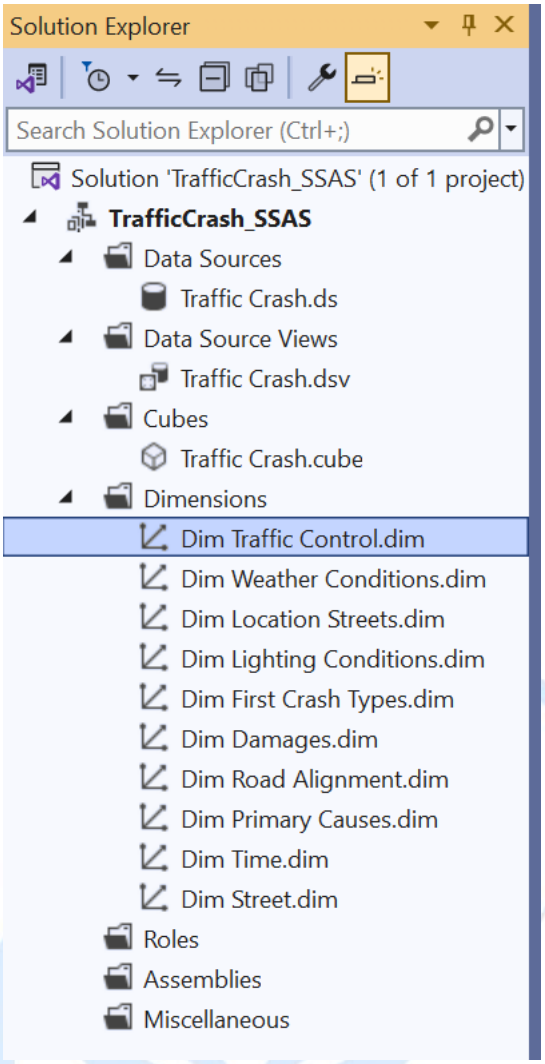


Kết quả sẽ được như hình.

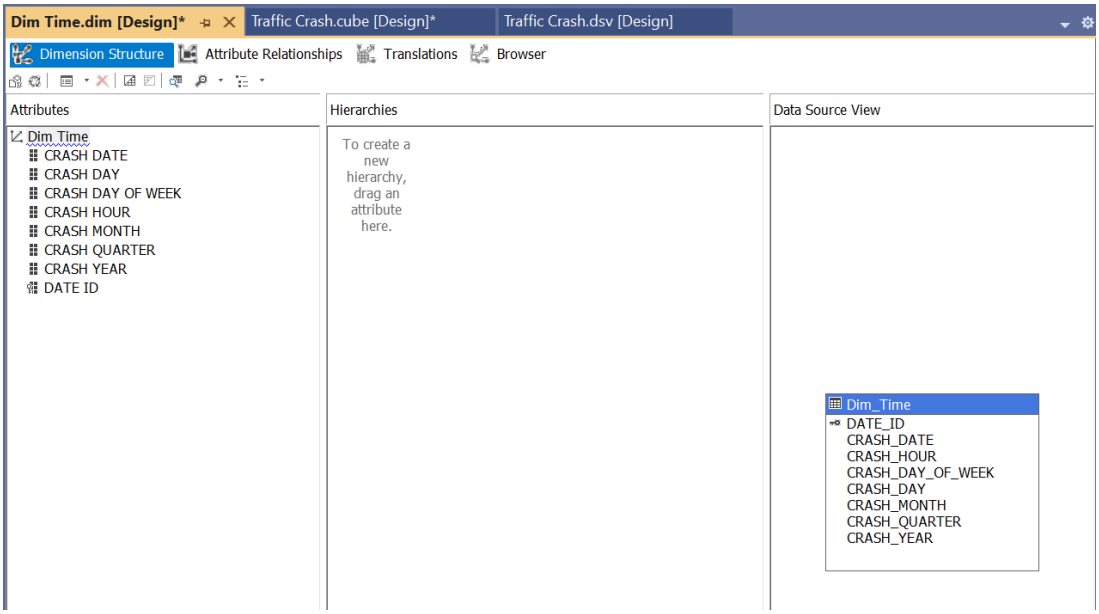


3.5.2 Thêm thuộc tính vào Dimension

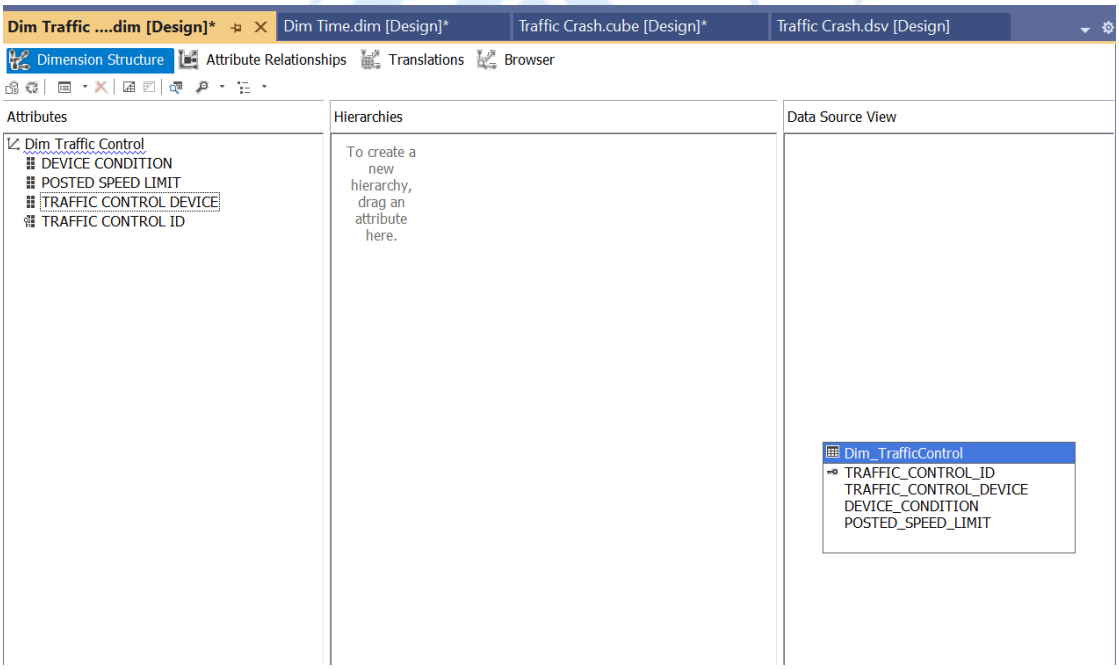
Ở **Solution Explorer**, nhấp đúp chuột trái vào từng bảng Dimension để tiến hành thêm các thuộc tính.



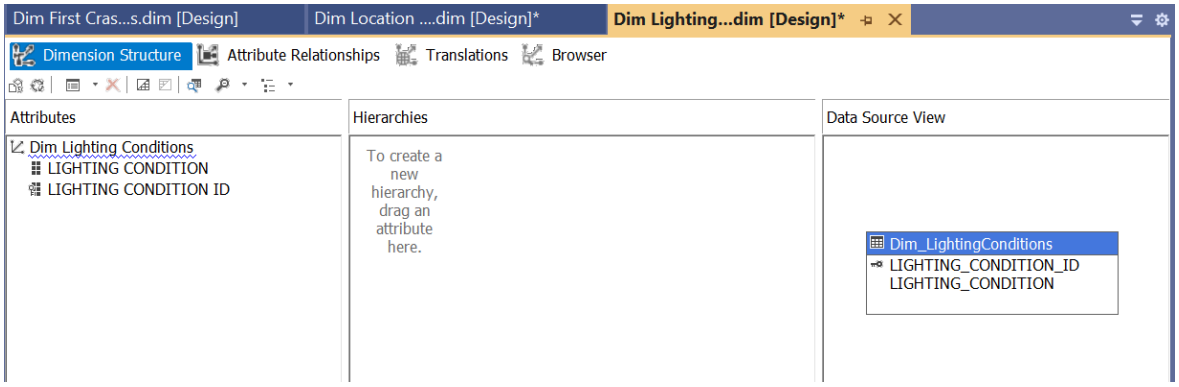
Ở bảng **Dim_Time,** kéo thả thuộc tính cần thiết ở cột Data Source View sang cột Attributes.



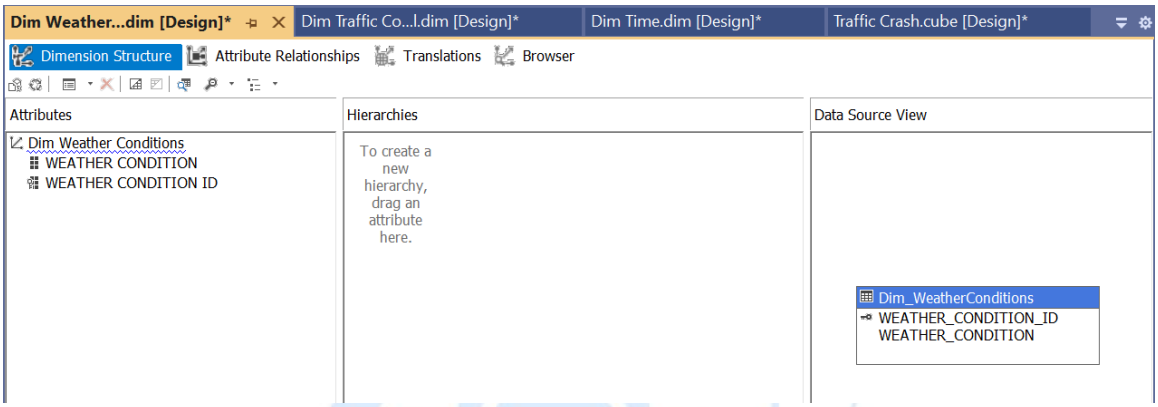
Ở bảng **Dim_TrafficControl,** kéo thả thuộc tính cần thiết ở cột Data Source View sang cột Attributes.



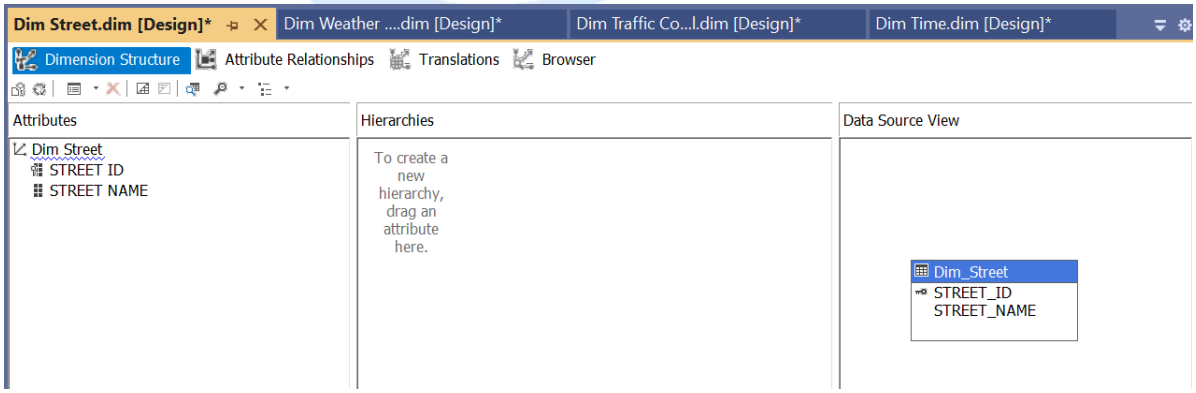
Ở bảng **Dim_LightingConditions,** kéo thả thuộc tính cần thiết ở cột Data Source View sang cột Attributes.



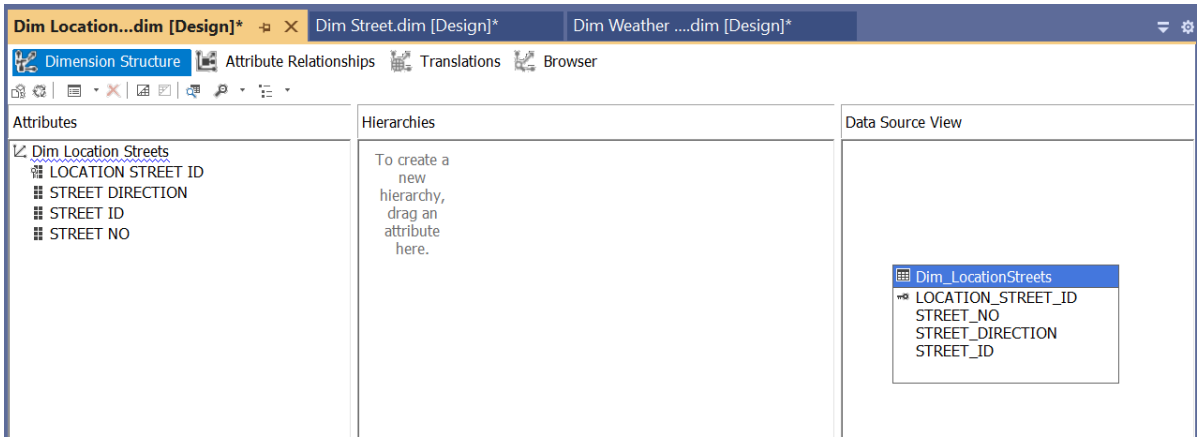
Ở bảng **Dim_WeatherConditions**, kéo thả thuộc tính cần thiết ở cột Data Source View sang cột Attributes.



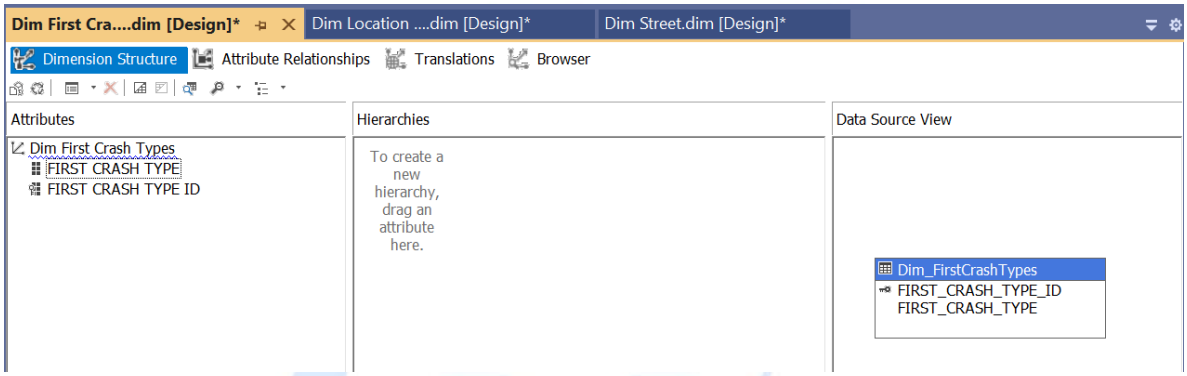
Ở bảng **Dim_Street**, kéo thả thuộc tính cần thiết ở cột Data Source View sang cột Attributes.



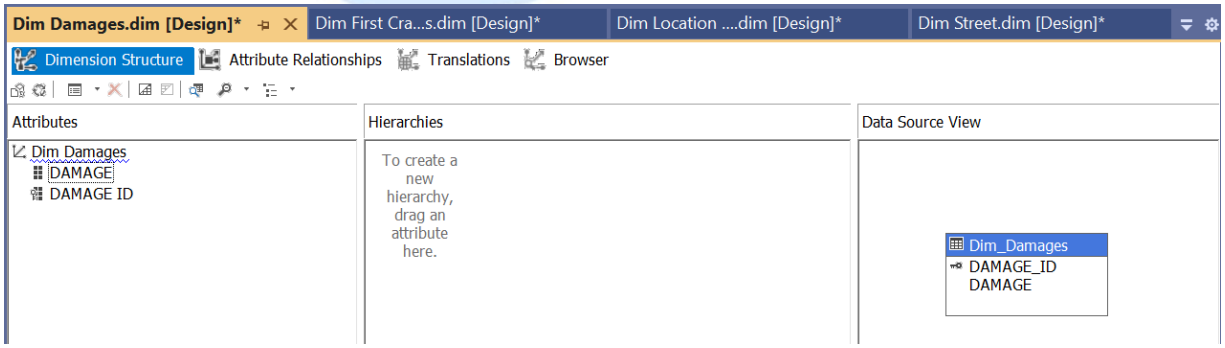
Ở bảng **Dim_LocationStreets**, kéo thả thuộc tính cần thiết ở cột Data Source View sang cột Attributes.



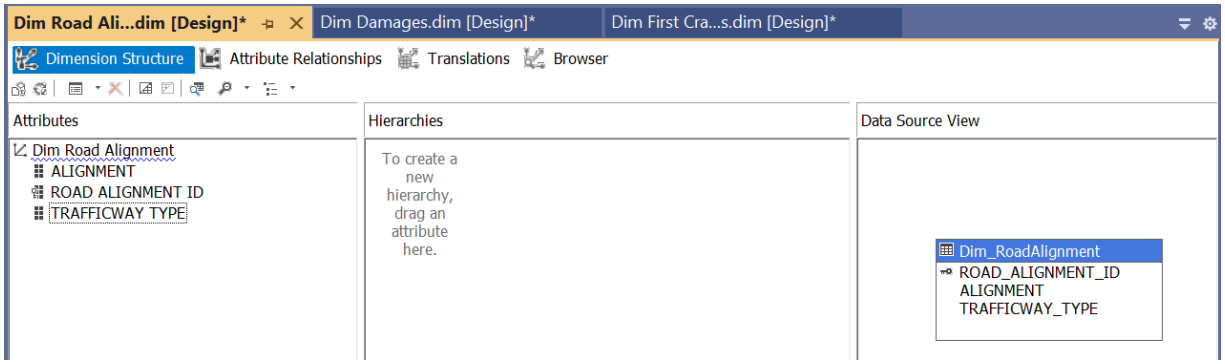
Ở bảng **Dim_FirstCrashTypes**, kéo thả thuộc tính cần thiết ở cột Data Source View sang cột Attributes.



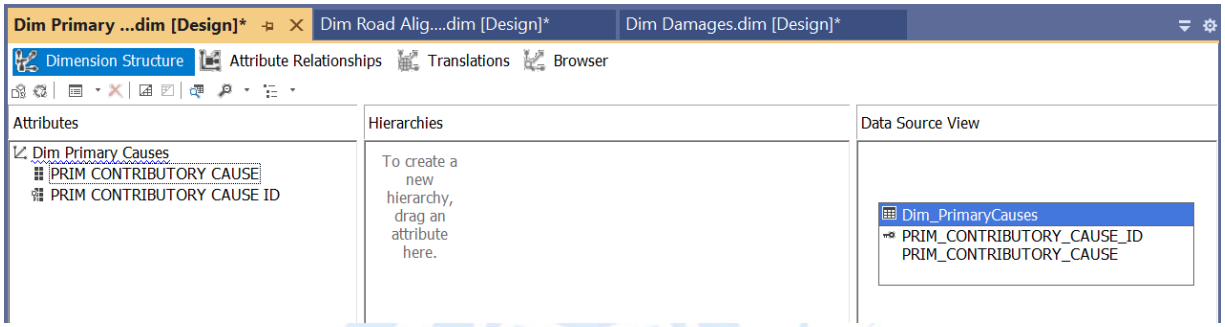
Ở bảng **Dim_Damages**, kéo thả thuộc tính cần thiết ở cột Data Source View sang cột Attributes.



Ở bảng **Dim_RoadAlignment**, kéo thả thuộc tính cần thiết ở cột Data Source View sang cột Attributes.



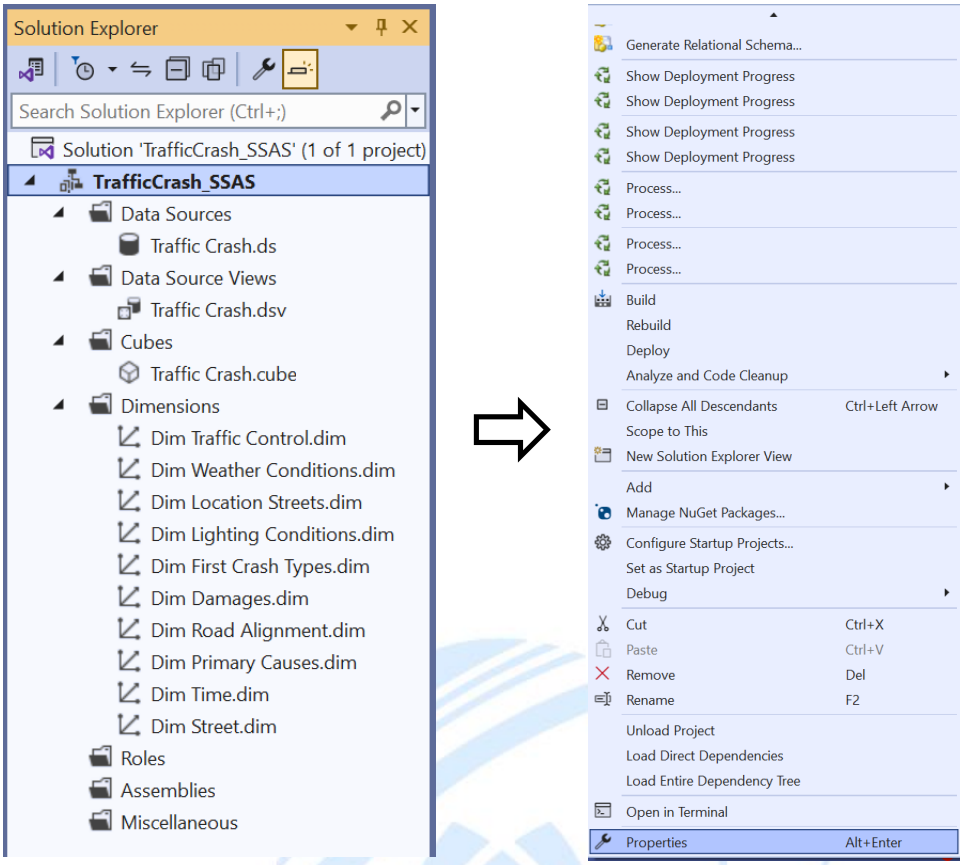
Ở bảng **Dim_PrimaryCauses**, kéo thả thuộc tính cần thiết ở cột Data Source View sang cột Attributes.



3.5.3 Deploy Analysis Services Project

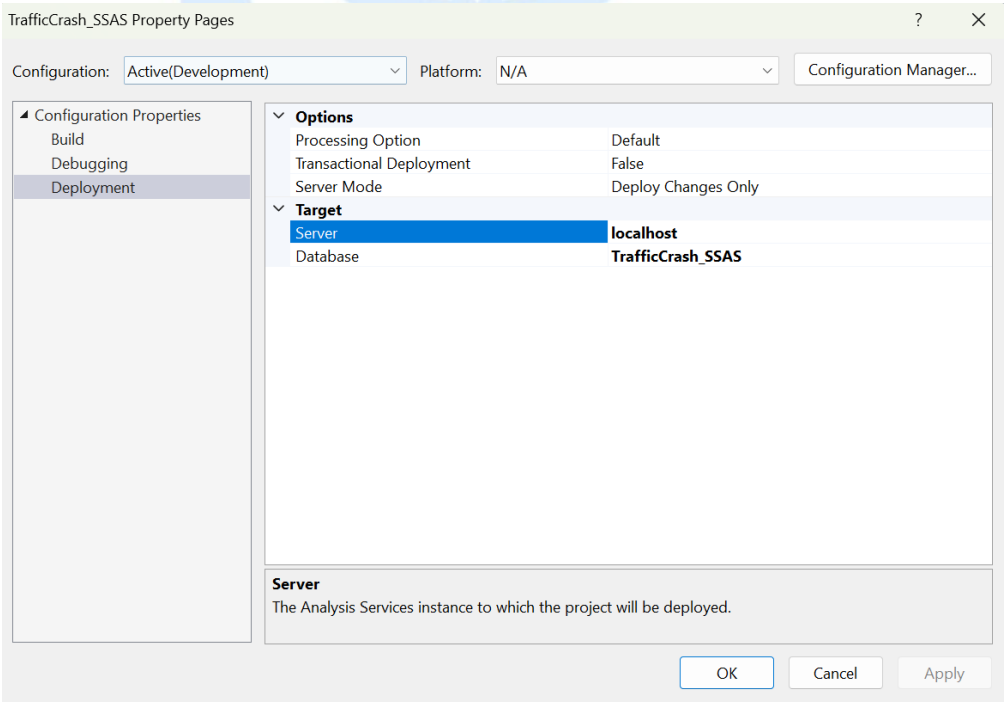
Bước 1: Tại **Solution Explorer**, nhấn chuột phải ở tên project và chọn **Properties** để chỉnh sửa kết nối đến Analysis Service của SQL Server.

Bước 1: Nhấp chuột phải vào Project đã tạo và chọn **Deploy**.



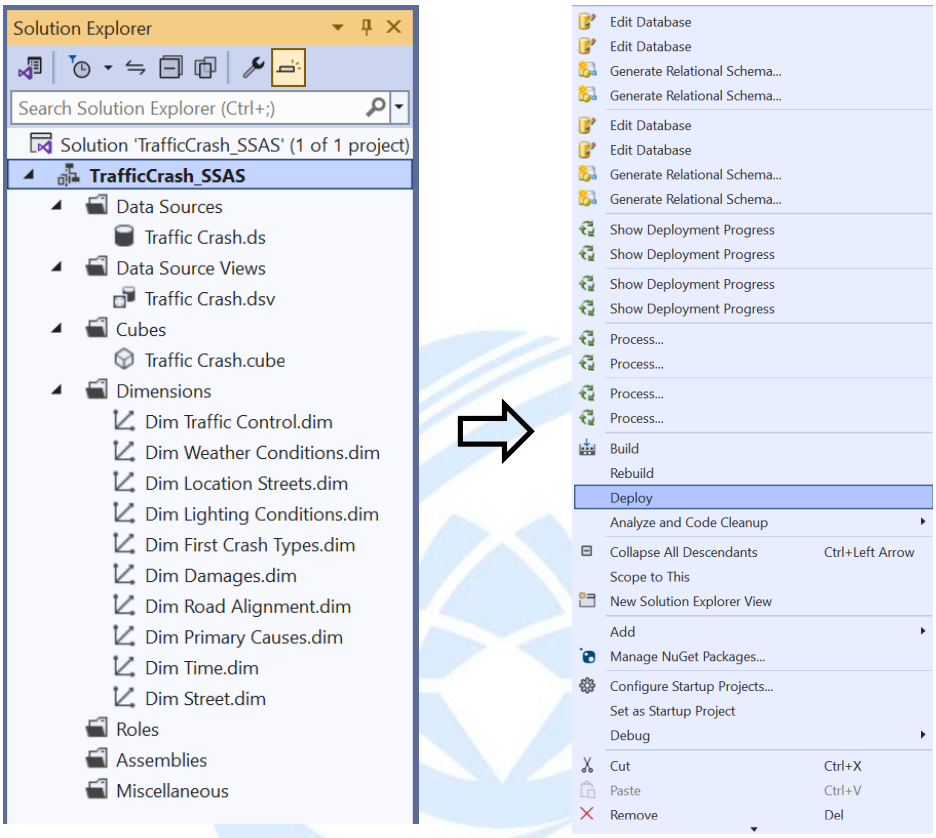
Bước 2: Ở cửa sổ vừa mở, đi đến **Deployment** và thực hiện đổi tên Server theo tên trong SQL Server.

- Server ban đầu là localhost:

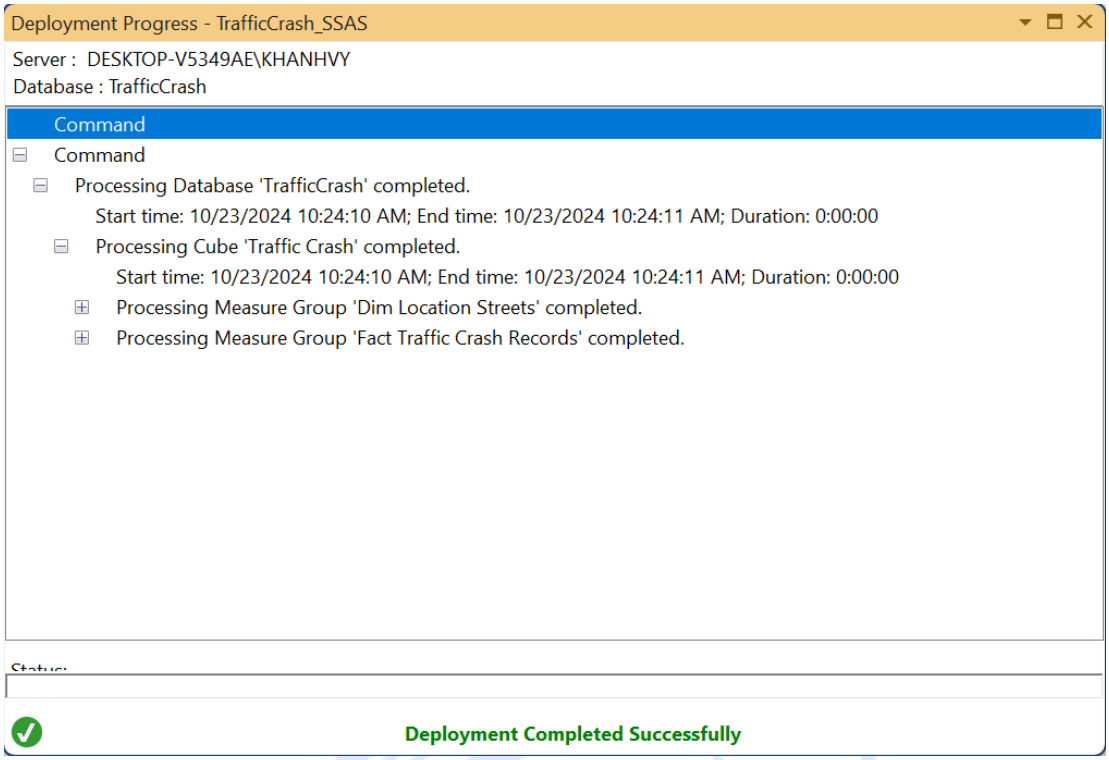


- Đặt lại tên cho Server. Sau đó nhấn “OK” để hoàn tất.

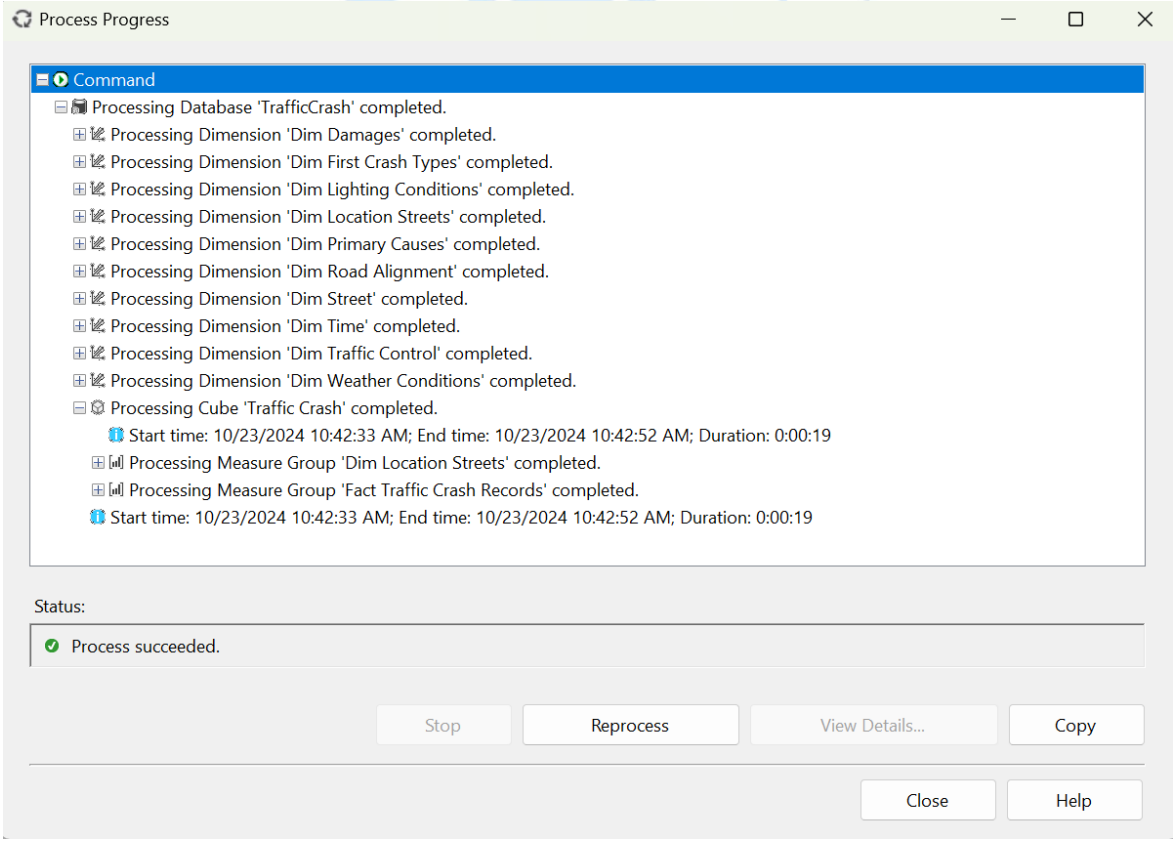
Bước 3: Nhấn chuột phải vào project đang hiện hành và chọn **Deploy**.



Khi deploy thành công sẽ có kết quả như hình.



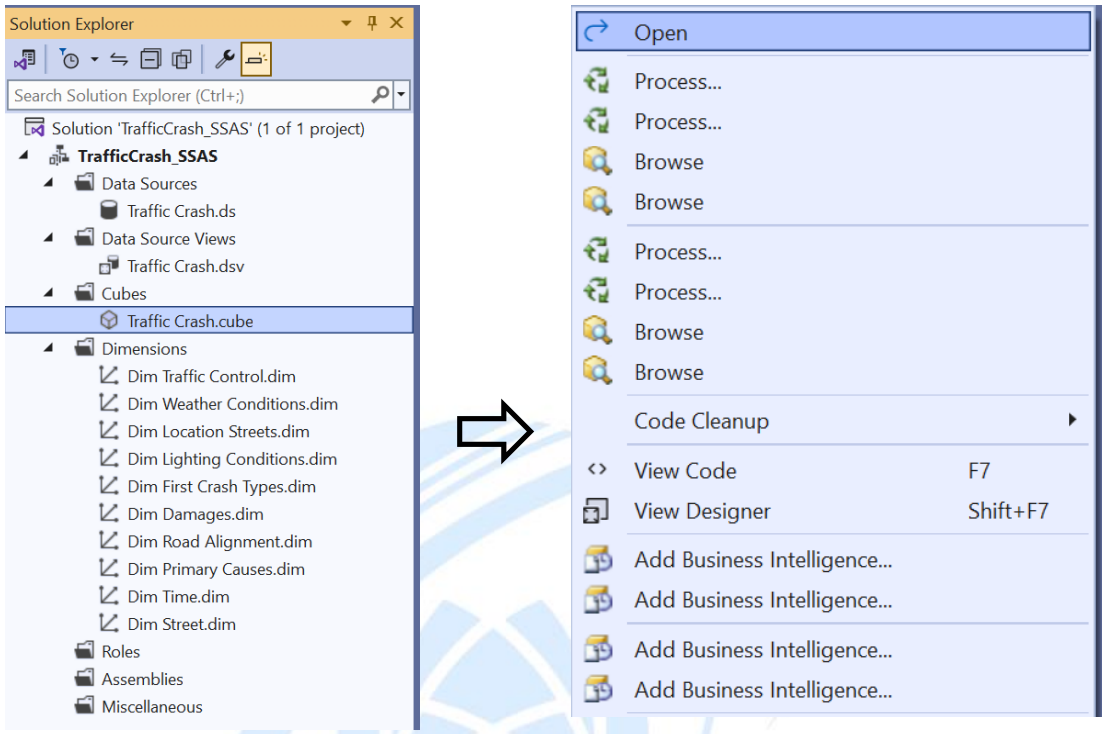
- Ta có thể vào Process Progress để kiểm tra.



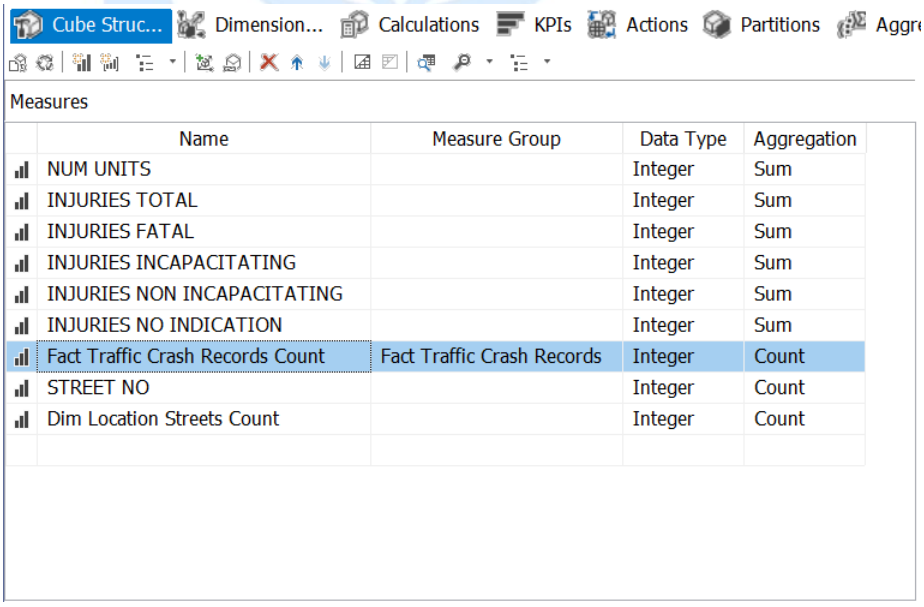
3.5.4 Tạo mới các Measures

3.5.4.1 Chỉnh sửa các Measure có sẵn

Bước 1: Nhấn chuột phải vào tên Cube đã tạo, chọn "Open" để mở khối Cube.



Bước 2: Nhấn chuột phải vào tên Cube, chọn **Show Measures Grid** để hiển thị chi tiết các độ đo.



Name	Measure Group	Data Type	Aggregation
NUM UNITS		Integer	Sum
INJURIES TOTAL		Integer	Sum
INJURIES FATAL		Integer	Sum
INJURIES INCAPACITATING		Integer	Sum
INJURIES NON INCAPACITATING		Integer	Sum
INJURIES NO INDICATION		Integer	Sum
Fact Traffic Crash Records Count	Fact Traffic Crash Records	Integer	Count
STREET NO		Integer	Count
Dim Location Streets Count		Integer	Count

Bước 3: Tiến hành đặt lại tên cho các **Measures** để dễ nhận biết.

Measures

Name	Measure Group	Data Type	Aggregation
Number of Units		Integer	Sum
Injuries Total		Integer	Sum
Injuries Fatal		Integer	Sum
Injuries Incapacitating		Integer	Sum
Injuries Non Incapacitating		Integer	Sum
Injuries No Indication		Integer	Sum
Number of Traffic Crash Records	Fact Traffic Crash Records	Integer	Count
Street No		Integer	Count
Number of Locations		Integer	Count

Bước 4: Tiến hành Deploy lại Project.

Deployment Progress - TrafficCrash_SSAS

Server : DESKTOP-V5349AE\KHANHVY
Database : TrafficCrash

Command

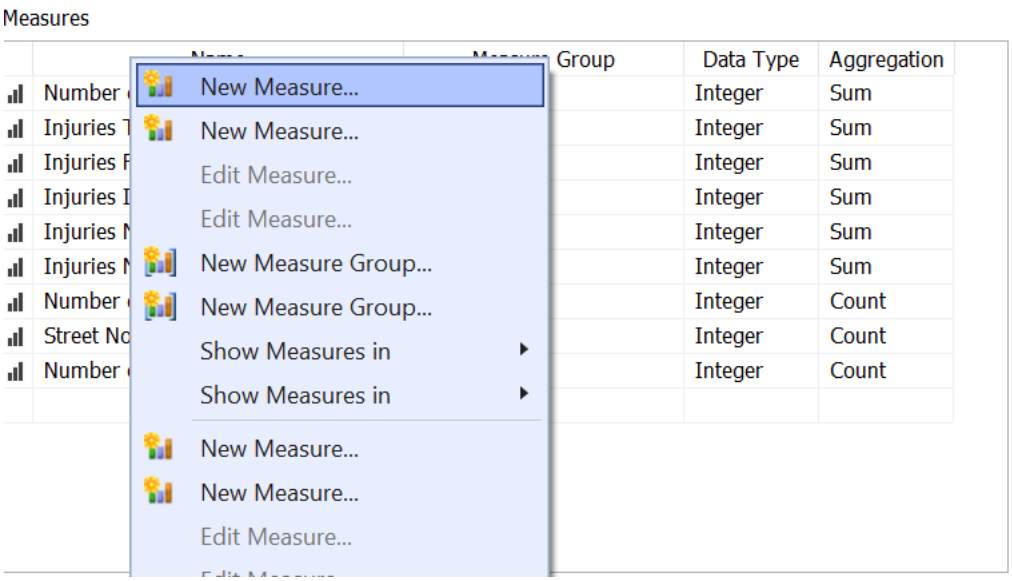
Command
Processing Database 'TrafficCrash' completed.
Start time: 10/23/2024 1:43:24 PM; End time: 10/23/2024 1:43:25 PM; Duration: 0:00:00
Processing Cube 'Traffic Crash' completed.
Start time: 10/23/2024 1:43:24 PM; End time: 10/23/2024 1:43:25 PM; Duration: 0:00:00
Processing Measure Group 'Dim Location Streets' completed.
Processing Measure Group 'Fact Traffic Crash Records' completed.

Status:

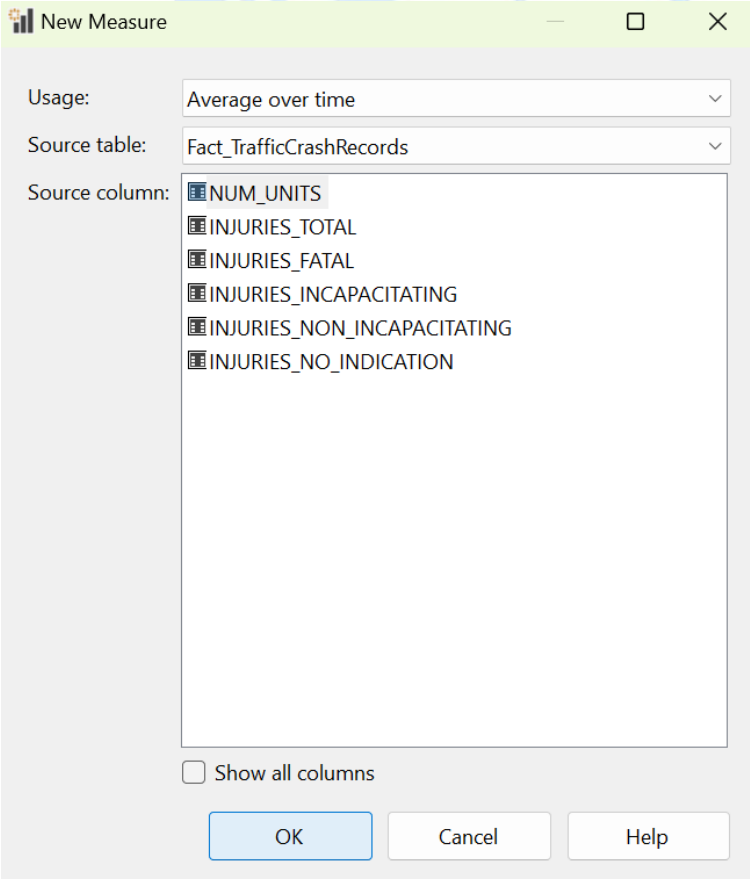
Deployment Completed Successfully

3.5.4.2 Tạo các Measure mới

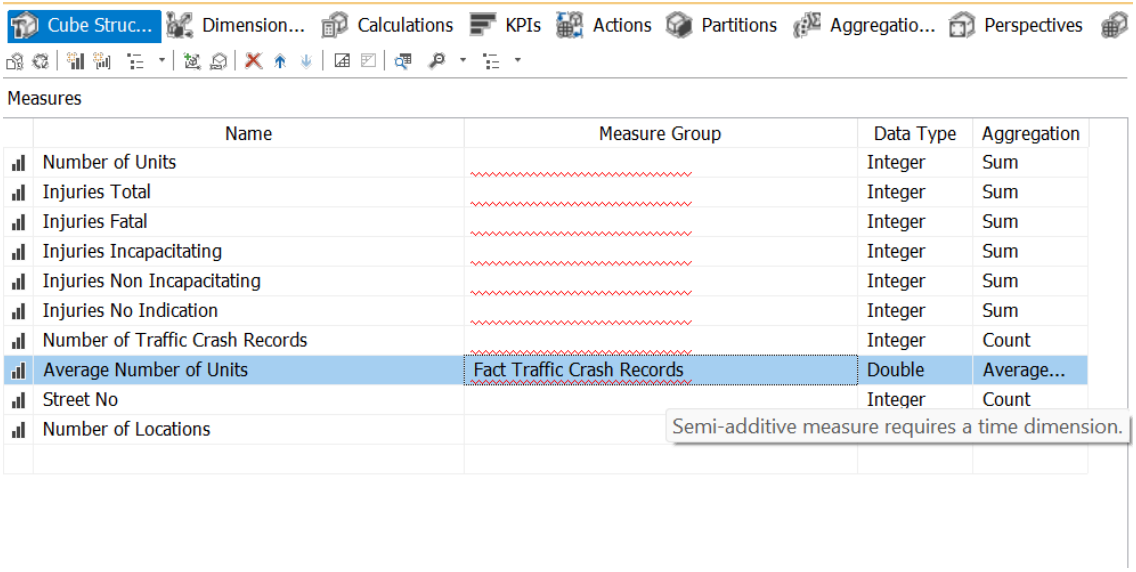
Bước 1: Chuột phải ở bảng **Measures** chọn **New Measure.**



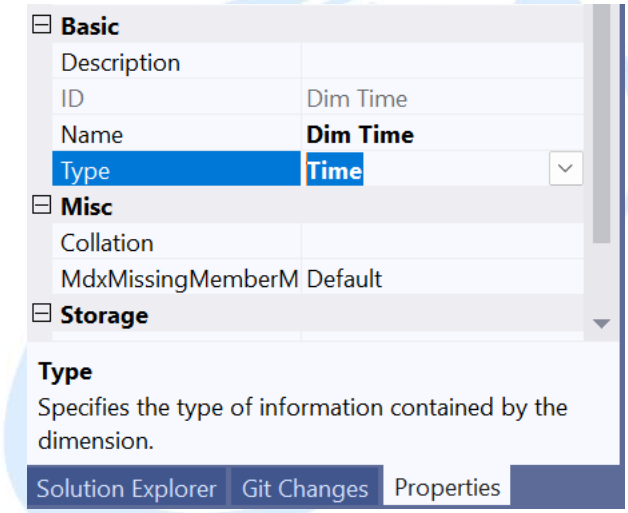
Bước 2: Chọn loại độ đo cần thêm và thông tin cần thiết. Sau đó nhấn "OK".



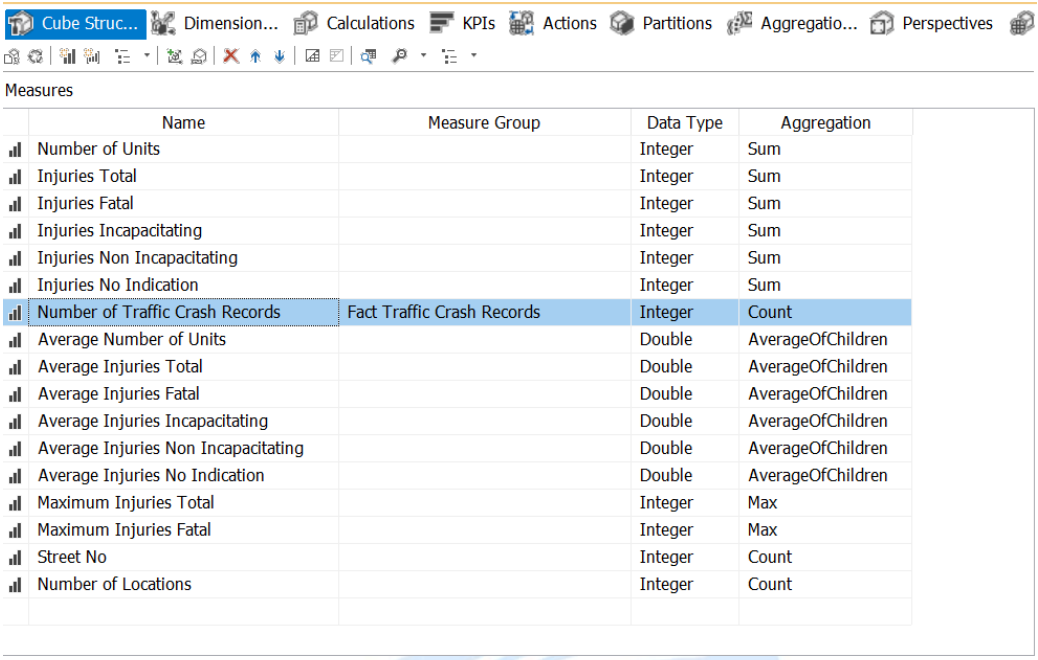
Bước 3: Đặt lại tên mới cho Measures, xuất hiện dòng "*Semi-additive measure requires a time dimesion*".



Nhấn chuột phải vào Dim_Time, chọn **Properties** và sửa **Type** thành **Time**.

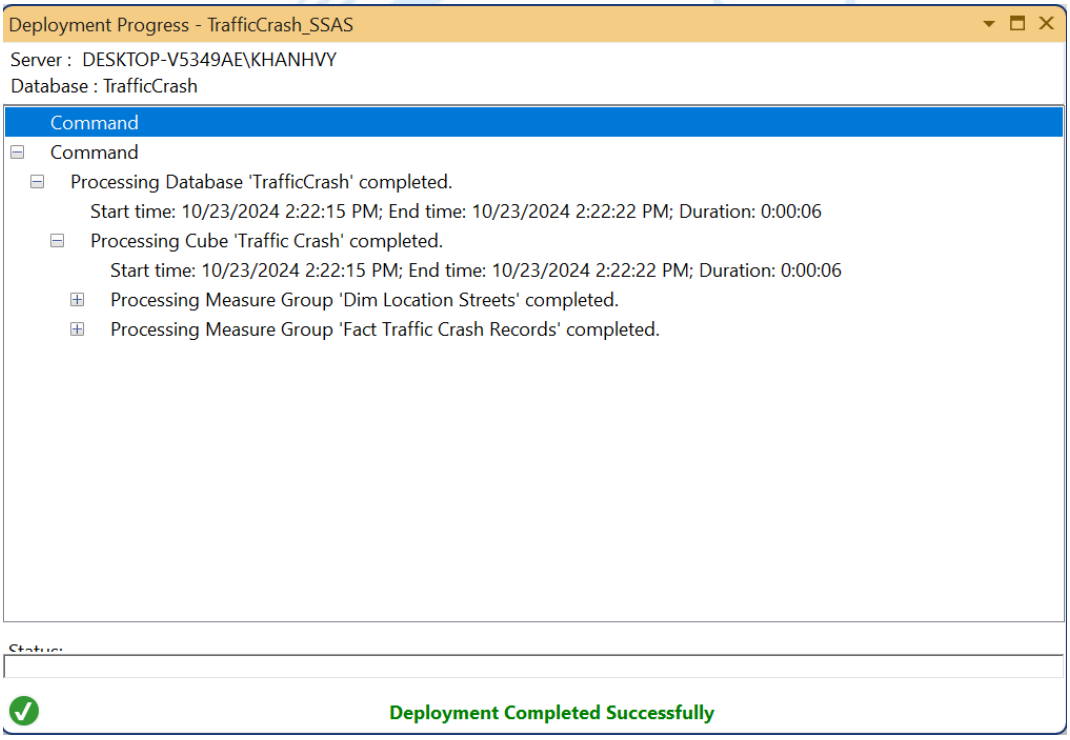


Bước 5: Tương tự, ta thực hiện tạo mới một số độ đo cần thiết.



Name	Measure Group	Data Type	Aggregation
Number of Units		Integer	Sum
Injuries Total		Integer	Sum
Injuries Fatal		Integer	Sum
Injuries Incapacitating		Integer	Sum
Injuries Non Incapacitating		Integer	Sum
Injuries No Indication		Integer	Sum
Number of Traffic Crash Records	Fact Traffic Crash Records	Integer	Count
Average Number of Units		Double	AverageOfChildren
Average Injuries Total		Double	AverageOfChildren
Average Injuries Fatal		Double	AverageOfChildren
Average Injuries Incapacitating		Double	AverageOfChildren
Average Injuries Non Incapacitating		Double	AverageOfChildren
Average Injuries No Indication		Double	AverageOfChildren
Maximum Injuries Total		Integer	Max
Maximum Injuries Fatal		Integer	Max
Street No		Integer	Count
Number of Locations		Integer	Count

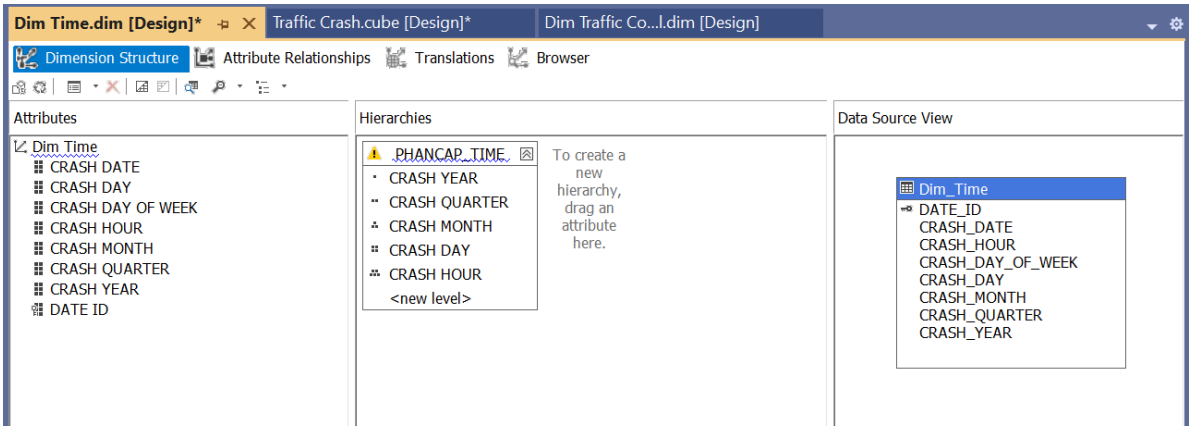
Bước 6: Tiến hành Deploy lại Project.



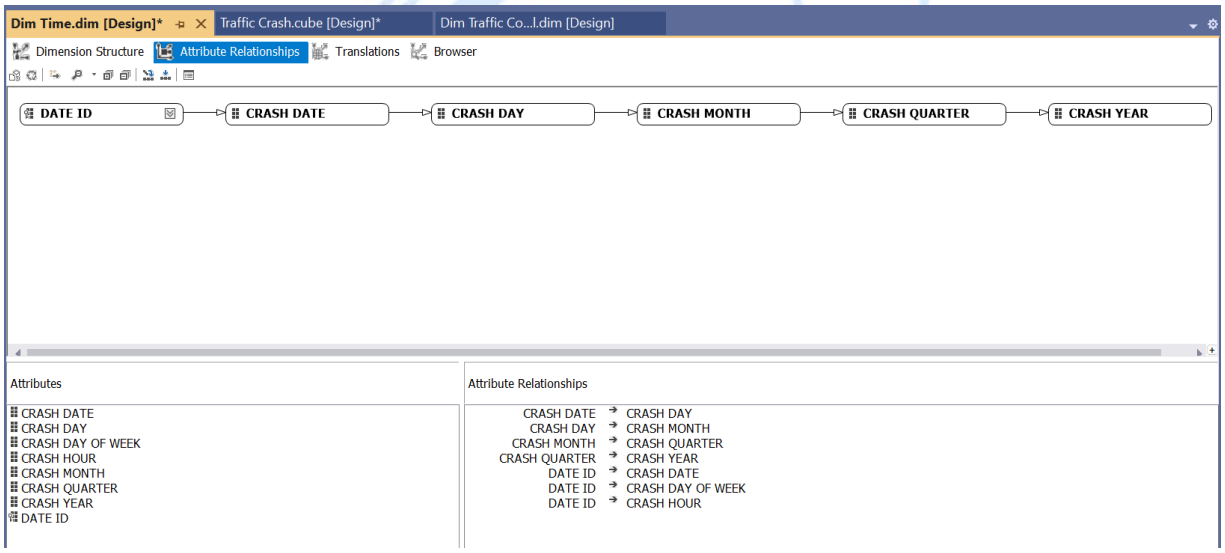
3.5.5 Tạo Hierachies và định nghĩa các Attribute Relationship

- Tạo Hierachy phân cấp theo CRASH_YEAR > CRASH_QUARTER > CRASH_MONTH > CRASH_DAY > CRASH_HOUR.

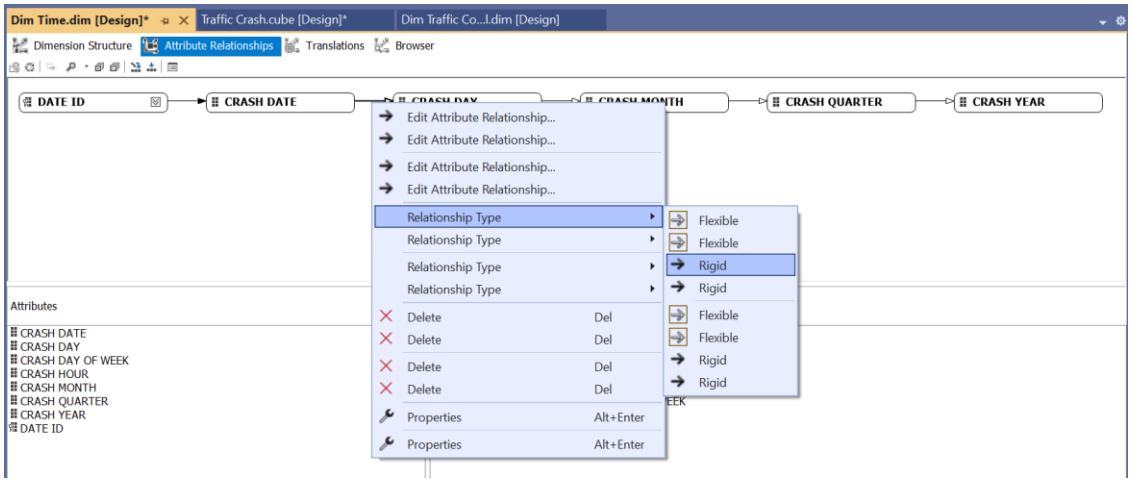
Bước 1: Kéo thả các thuộc tính CRASH_YEAR, CRASH_QUARTER, CRASH_MONTH, CRASH_DAY, CRASH_HOUR sang cột Hierachies. Với thứ tự phân cấp từ cao tới thấp nhất và đổi tên Hireachy thành PHANCAP_TIME.



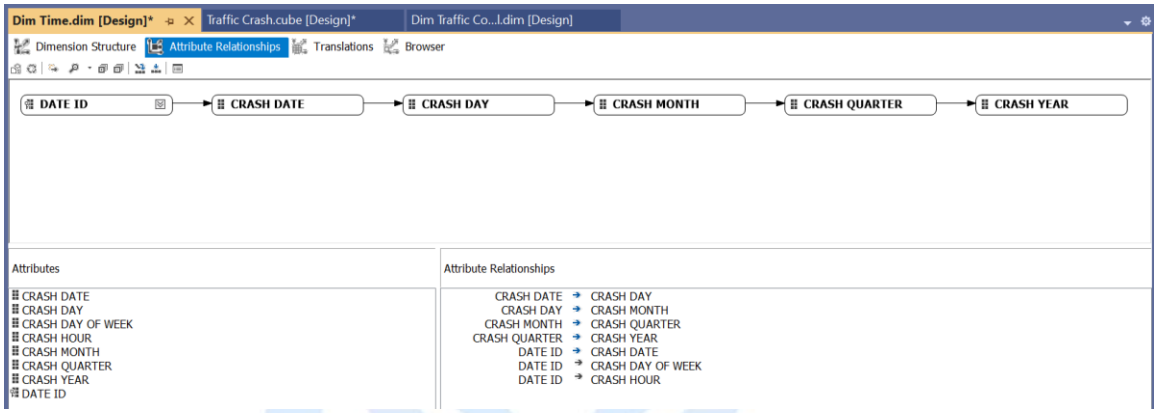
Bước 2: Chuyển sang tab **Attribute Relationships** để tiến hành định nghĩa Attribute Relationships. Tiến hành kéo thả phân cấp từ nhỏ đến lớn theo thứ tự từ trái sang phải.



Bước 3: Đổi các **Relationship Type** thành **Rigid**.

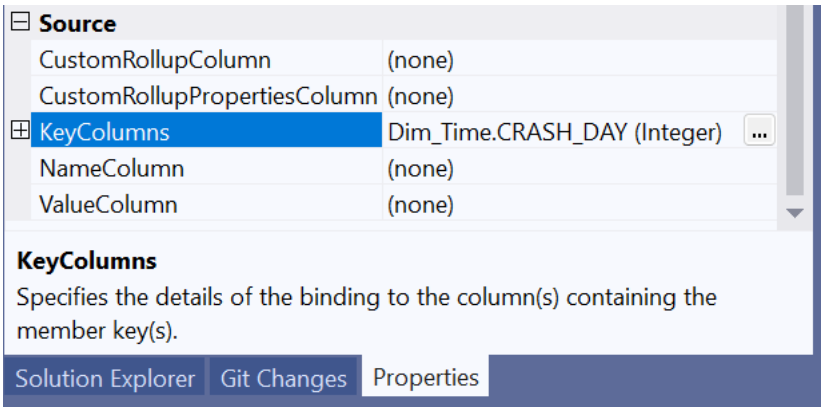


Mối quan hệ đã tạo:

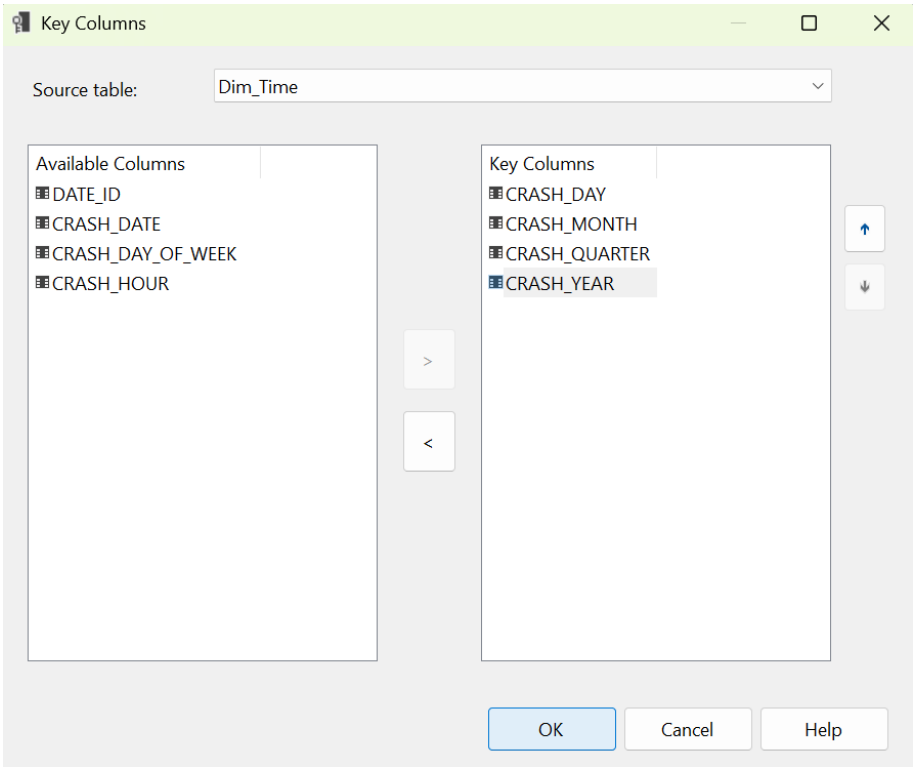


Bước 4: Chỉnh khóa cột (**Key Columns**) và tên cột (**Name Column**) của thuộc tính CRASH_DATE. Vì thuộc tính CRASH_DAY là thuộc tính cấp nhỏ nhất nên sẽ lấy khóa cột gồm chính nó và những thuộc tính cấp cao hơn.

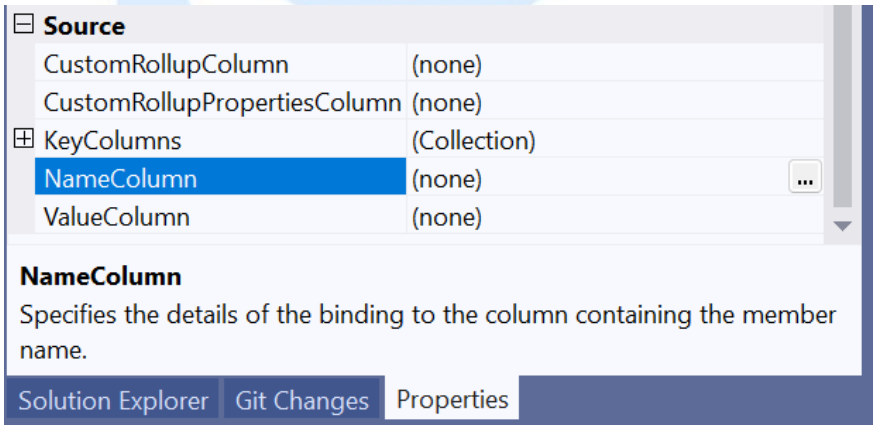
Tại cửa sổ **Properties** của thuộc tính CRASH_DAY, chọn **Key Columns**.



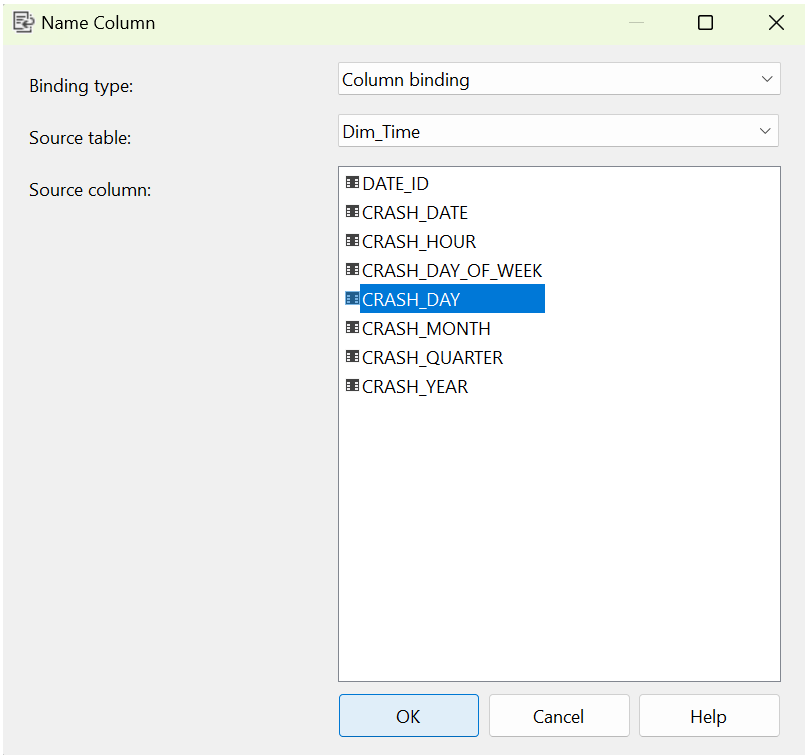
Thêm các thuộc tính vào **Key Columns**, sau đó "OK" để hoàn tất.



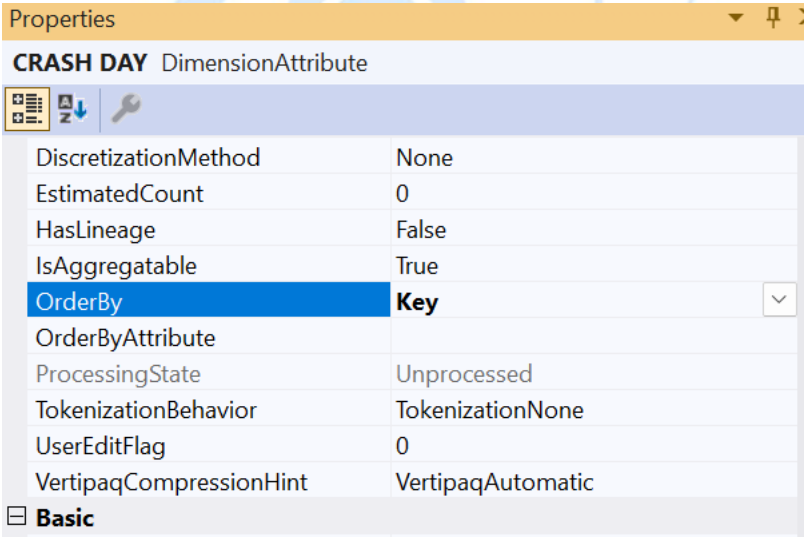
Bước 5: Tại cửa sổ **Properties**, chọn **NameColumn.**



Chọn tên thuộc tính sẽ hiển thị trên Hierachy là CRASH_DAY. Sau đó nhấn "OK".

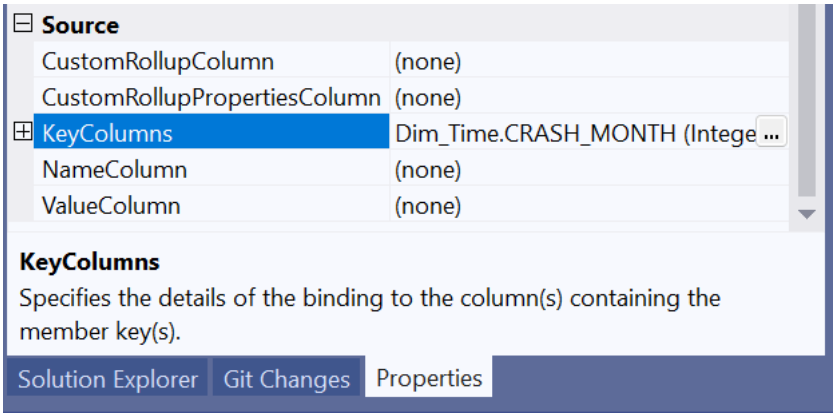


Bước 6: Chọn thuộc tính **OrderBy** là **Key** để CRASH_DAY được sắp xếp theo thứ tự tăng dần.

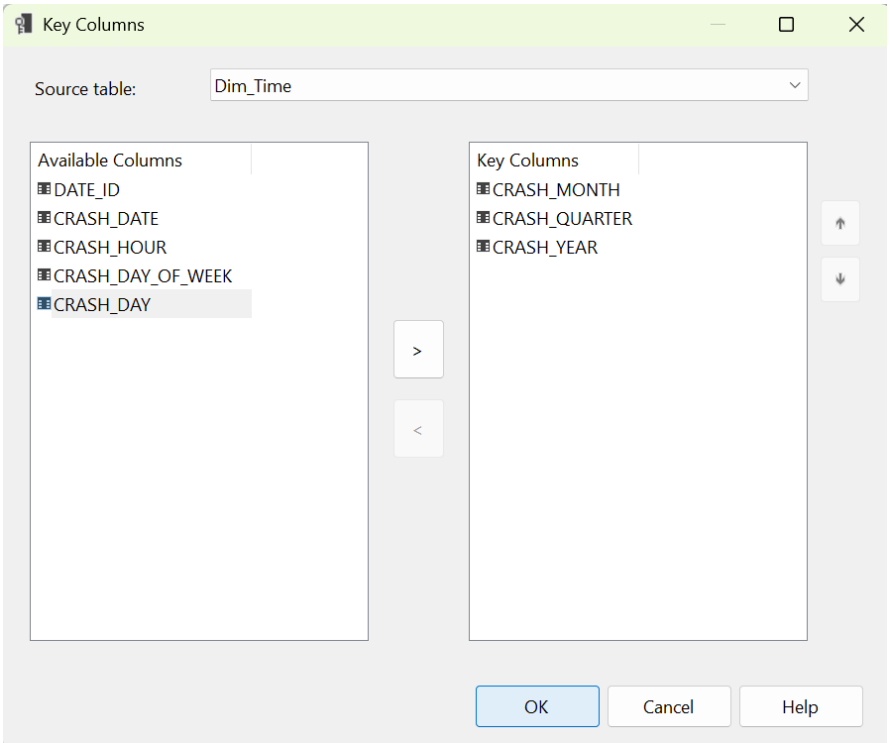


Bước 7: Chỉnh khóa cột (**KeyColumns**) và tên cột (**Name Column**) của thuộc tính CRASH_MONTH. Vì thuộc tính CRASH_MONTH là thuộc tính cấp nhỏ hơn CRASH_QUARTER, CRASH_YEAR nên sẽ lấy khóa cột gồm chính nó và những thuộc tính cấp cao hơn.

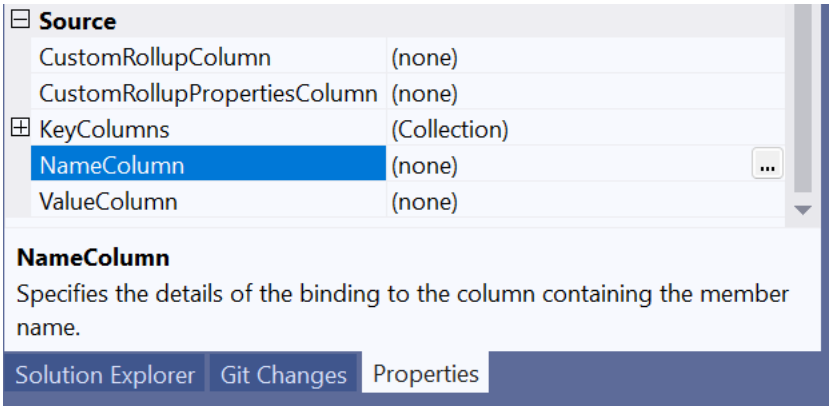
Tại cửa sổ **Properties** của thuộc tính CRASH_MONTH, chọn **Key Columns**.



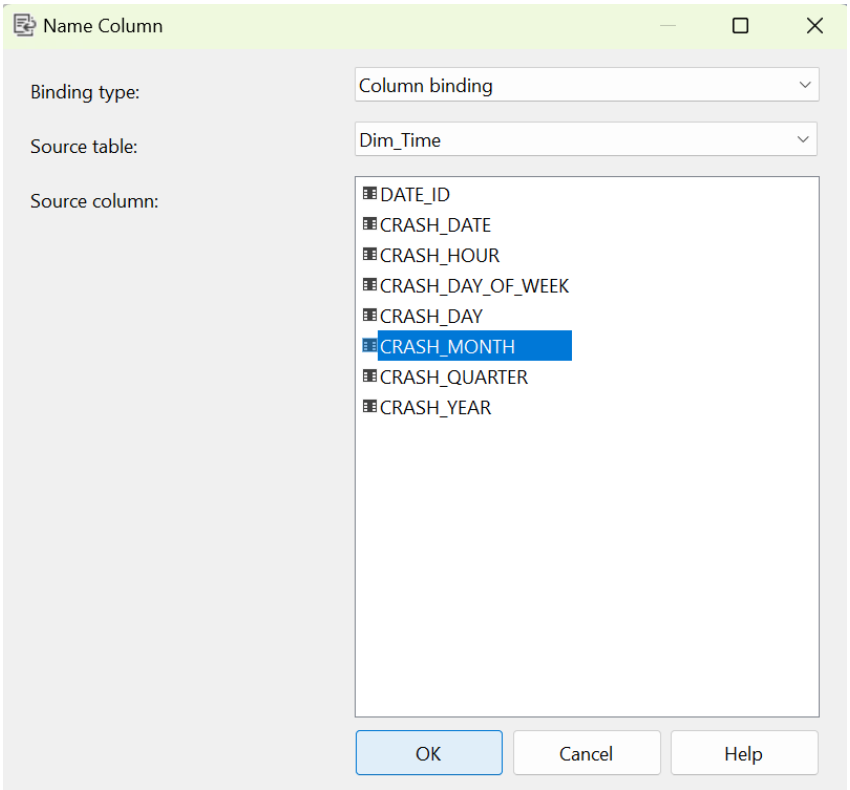
Thêm các thuộc tính vào **Key Columns**, sau đó "OK" để hoàn tất.



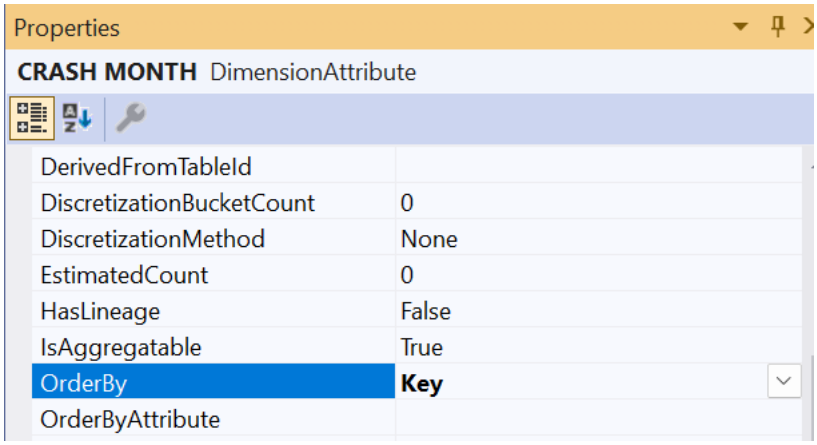
Bước 8: Tại cửa sổ **Properties** của thuộc tính CRASH_MONTH, ta chọn **Name Column**.



Chọn tên thuộc tính là CRASH_MONTH. Sau đó nhấn "OK".

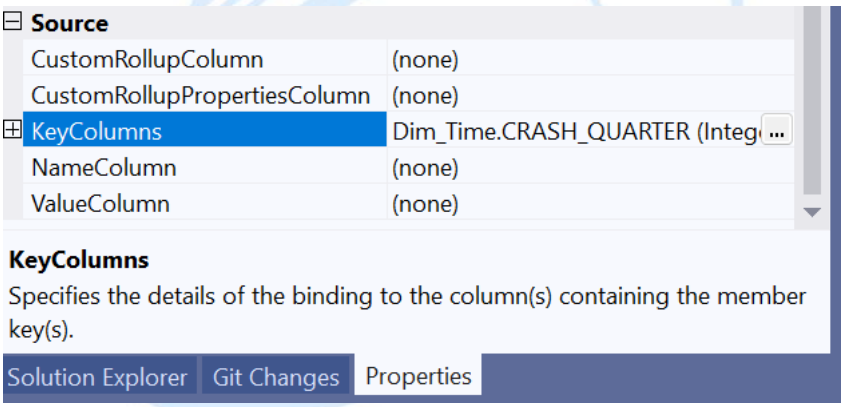


Bước 9: Chọn thuộc tính **OrderBy** là **Key** để CRASH_MONTH được sắp xếp theo thứ tự tăng dần.

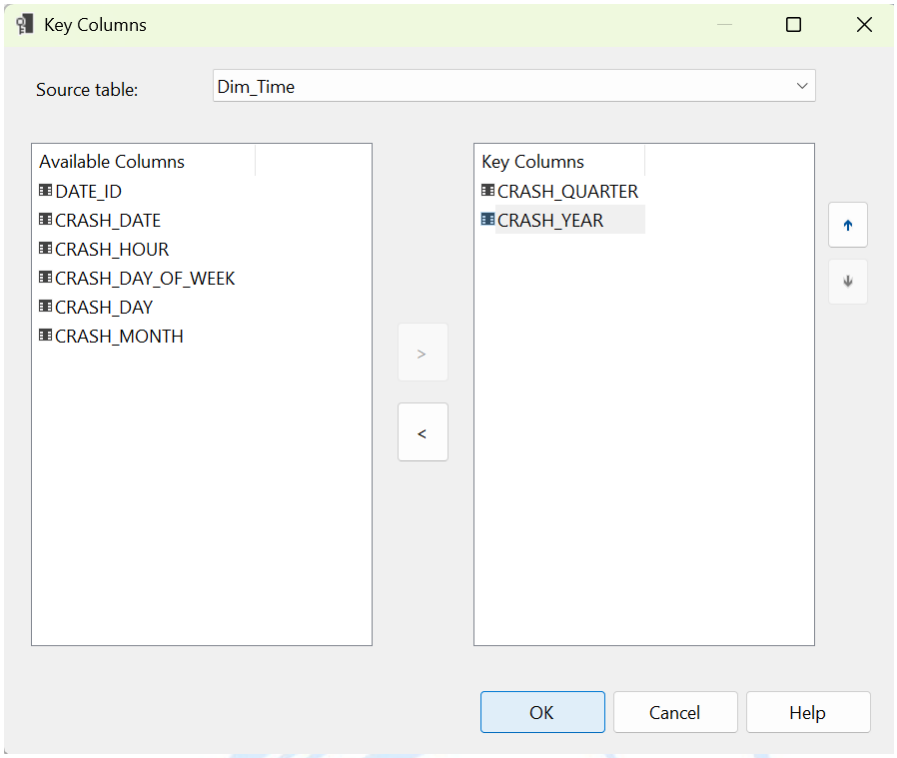


Bước 10: Chỉnh khóa cột (**KeyColumns**) và tên cột (**Name Column**) của thuộc tính CRASH_QUARTER. Vì thuộc tính CRASH_QUARTER là thuộc tính cấp nhỏ hơn CRASH_YEAR nên sẽ lấy khóa cột gồm chính nó và thuộc tính cấp cao hơn.

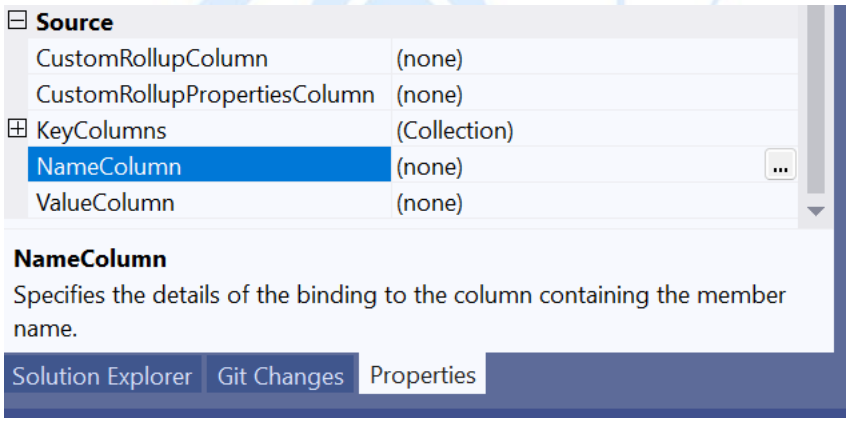
Tại cửa sổ **Properties** của thuộc tính CRASH_QUARTER, chọn **Key Columns**.



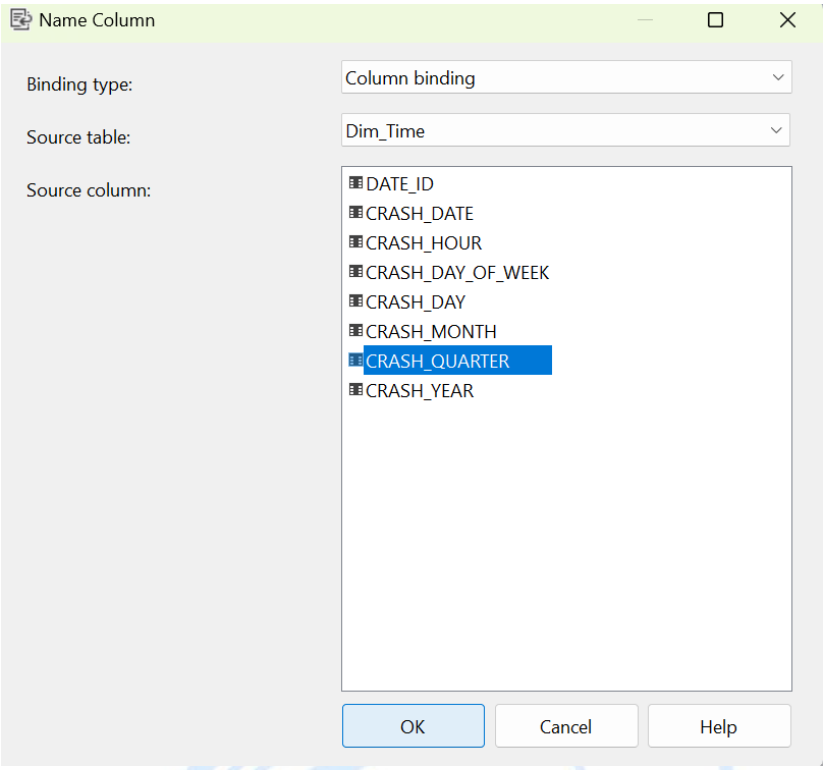
Thêm các thuộc tính vào **Key Columns**, sau đó "OK" để hoàn tất.



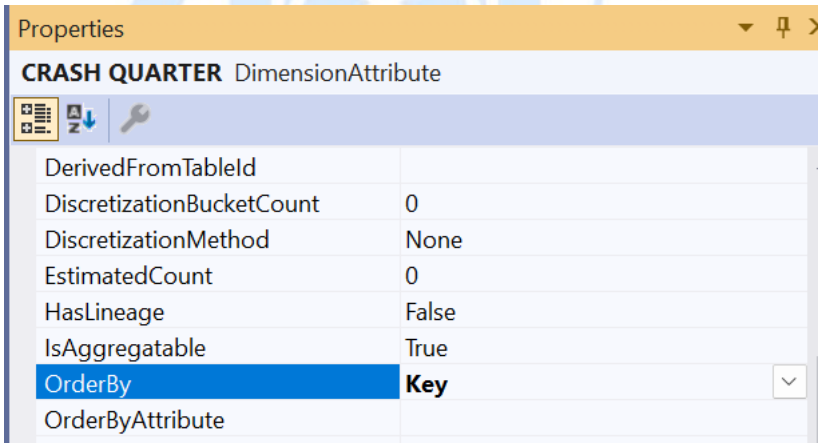
Bước 11: Tại cửa sổ **Properties** của thuộc tính CRASH_QUARTER, ta chọn **Name Column**.



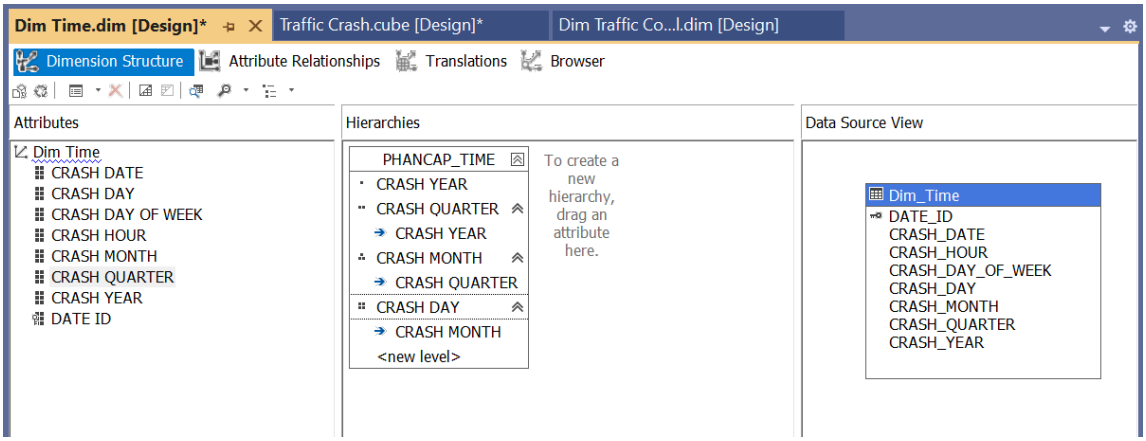
Chọn tên thuộc tính là CRASH_QUARTER. Sau đó nhấn “OK”.



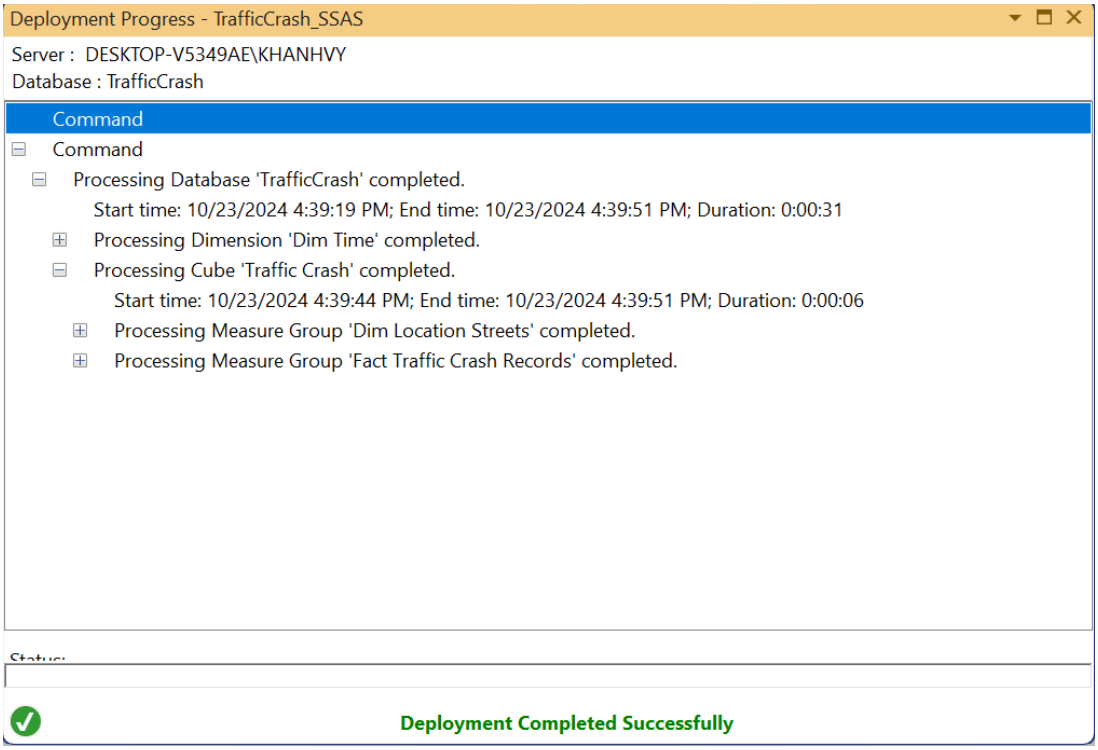
Bước 12: Chọn thuộc tính **OrderBy** là **Key** để CRASH_QUARTER được sắp xếp theo thứ tự tăng dần.



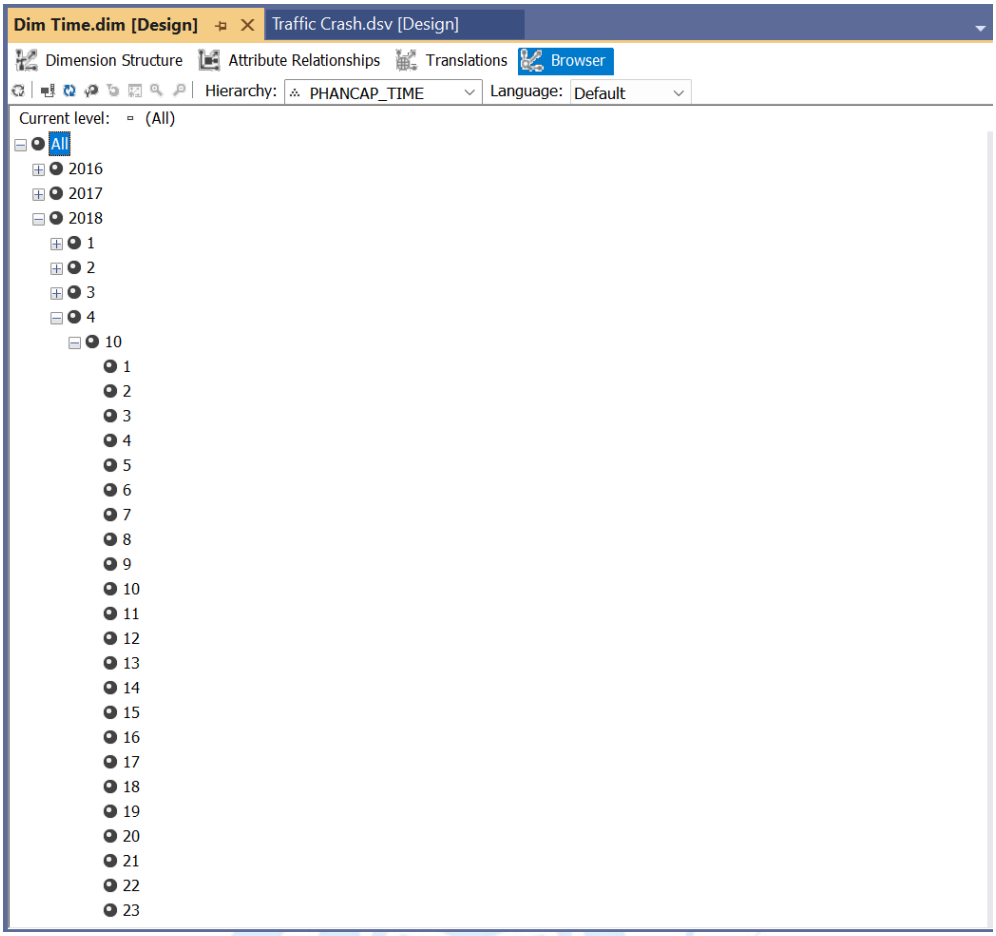
Ta thu được kết quả như hình dưới.



Bước 13: Thực hiện Deploy Project để đảm bảo không có lỗi xảy ra sau quá trình phân cấp.



Chuyển sang tab **Browser** để kiểm tra xem các thuộc tính đã phân cấp.


Dim Time.dim [Design]
Traffic Crash.dsv [Design]
Dimension Structure
Attribute Relationships
Translations
Browser
Hierarchy:
PHANCAP_TIME
Language:
Default
Current level:
(All)
All
2016
2017
2018
1
2
3
4
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

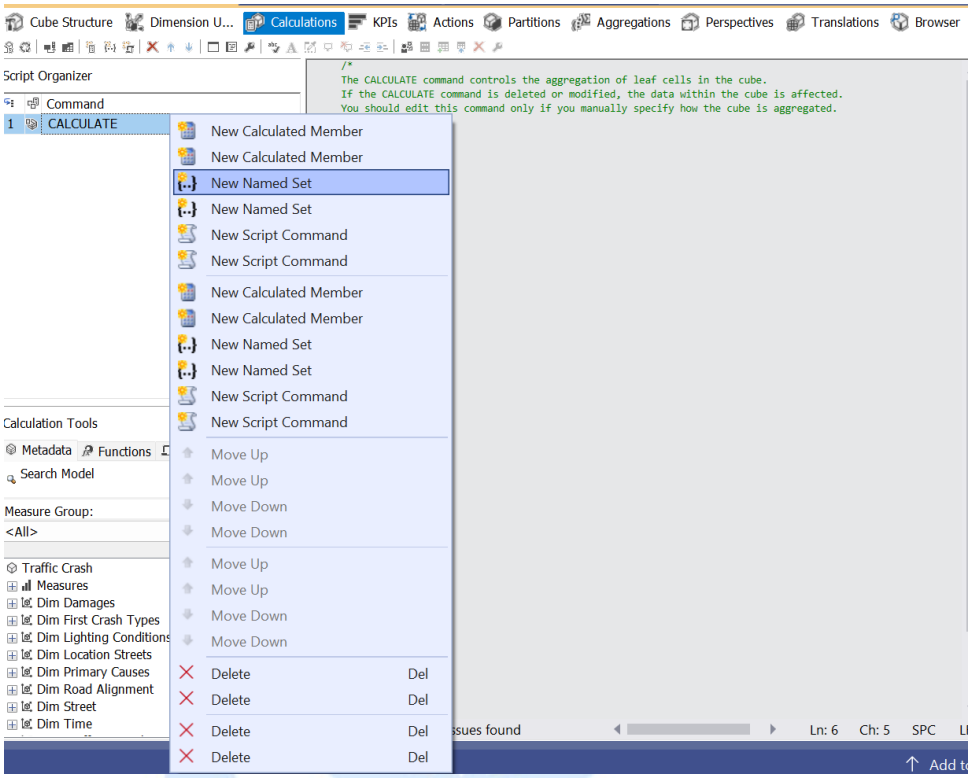
3.6 Thực hiện 15 câu truy vấn (MDX, Manual và Pivot Excel)

3.6.1 Câu truy vấn 1

Nội dung câu truy vấn: Liệt kê top 10 vụ tai nạn giao thông có số lượng tử vong cao nhất.

3.6.1.1 Manual

Bước 1: Vào mục **Calculations**, tạo mới một "New Named Set" đặt tên "C1".



Sau đó nhập câu truy vấn và tiến hành Deloy.

```
ORDER(
        TOPCOUNT(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM CONTRIBUTORY
CAUSE].MEMBERS,
            10,
            [Measures].[Injuries Fatal]
        ),
        [Measures].[Injuries Fatal],
        DESC
    )
```

Name:

[C1]

Expression

```
ORDER(
        TOPCOUNT(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM CONTRIBUTORY CAUSE].MEMBERS,
            10,
            [Measures].[Injuries Fatal]
        ),
        [Measures].[Injuries Fatal],
        DESC
    )
```

No issues found Ln: 11 Ch: 1 SPC CRLF

Additional Properties

Type: Dynamic

Display folder:

Bước 2: Vào mục **Browser**, kéo từ bảng Dim Primary Cause thuộc tính "PRIM CONTRIBUTORY CAUSE". Tại cột **Operator** chọn "In" và **Filter Expression** chọn vào Named Set "C1" vừa tạo. Sau đó kéo các thuộc tính để hiển thị kết quả câu truy vấn và nhấn "Click to execute the query". Ta được kết quả.

Dimension	Hierarchy	Operator	Filter Expression	Param...
Dim Primary Causes	PRIM CONTRIBU...	In	C1	
<Select dimension>				

PRIM CONTRIBUTORY CAUSE	Injuries Fatal
DISREGARDING TRAFFIC SIGNALS	77
DRIVING ON WRONG SIDE/WRONG WAY	26
EXCEEDING AUTHORIZED SPEED LIMIT	26
FAILING TO REDUCE SPEED TO AVOID CRASH	80
FAILING TO YIELD RIGHT-OF-WAY	77
NOT APPLICABLE	74
OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	45
PHYSICAL CONDITION OF DRIVER	79
UNABLE TO DETERMINE	323
UNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED)	21

3.6.1.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
SELECT
    {[Measures].[Injuries Fatal]} ON COLUMNS,
    ORDER(
        TOPCOUNT(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM CONTRIBUTORY
CAUSE].MEMBERS,
            10,
            [Measures].[Injuries Fatal]
        ),
        [Measures].[Injuries Fatal],
        DESC
    ) ON ROWS
     FROM [Traffic Crash]
```

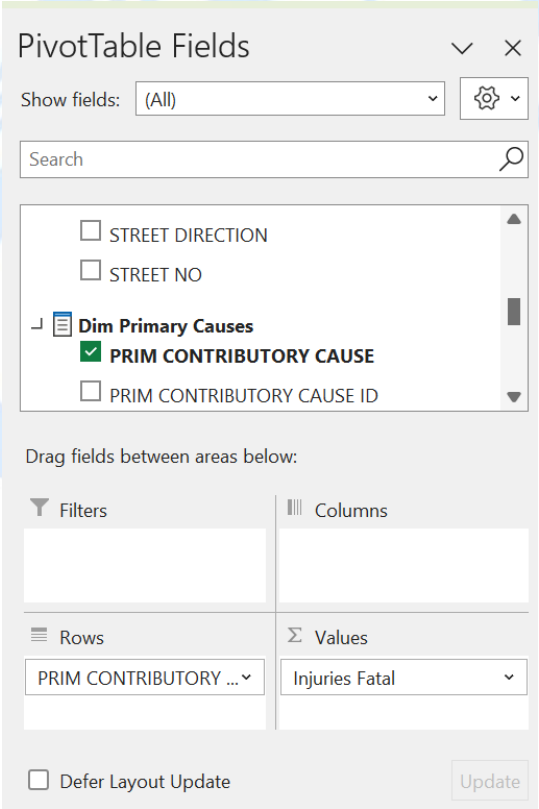
Kết quả câu truy vấn MDX:

Messages | Results

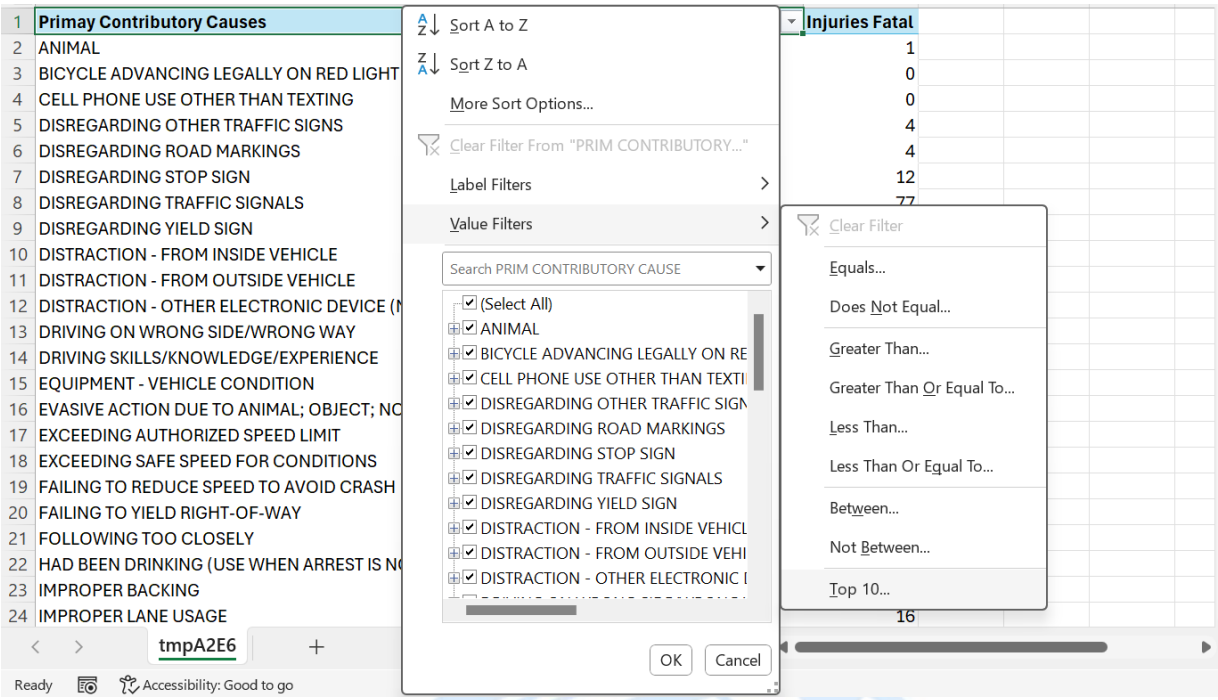
	Injuries Fatal
UNABLE TO DETERMINE	323
FAILING TO REDUCE SPEED TO AVOID CRASH	80
PHYSICAL CONDITION OF DRIVER	79
DISREGARDING TRAFFIC SIGNALS	77
FAILING TO YIELD RIGHT-OF-WAY	77
NOT APPLICABLE	74
OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	45
DRIVING ON WRONG SIDE/WRONG WAY	26
EXCEEDING AUTHORIZED SPEED LIMIT	26
UNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED)	21

3.6.1.3 Excel

Bước 1: Trong **PivotTable Fields**, tick chọn thuộc tính là "PRIM CONTRIBUTORY CAUSE" và "Injuries Fatal".



Bước 2: Nhấn vào biểu tượng cạnh Row Labels chọn **Value Filters**, sau đó chọn Top 10.



Kết quả:

Primay Contributory Causes	Injuries Fatal
DISREGARDING TRAFFIC SIGNALS	77
DRIVING ON WRONG SIDE/WRONG WAY	26
EXCEEDING AUTHORIZED SPEED LIMIT	26
FAILING TO REDUCE SPEED TO AVOID CRASH	80
FAILING TO YIELD RIGHT-OF-WAY	77
NOT APPLICABLE	74
OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	45
PHYSICAL CONDITION OF DRIVER	79
UNABLE TO DETERMINE	323
UNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED)	21
Grand Total	**828**

3.6.1.4 Nhận xét

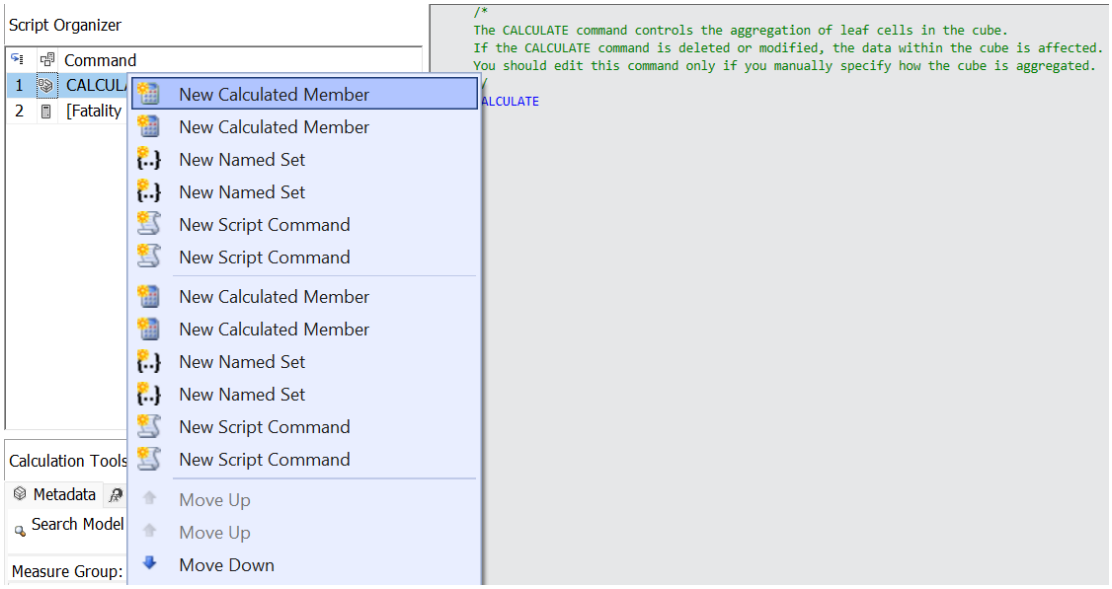
Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.2 Câu truy vấn 2

Nội dung câu truy vấn: Thống kê tỉ lệ tử vong và bị thương nặng của nạn nhân trong các nhân tố trực tiếp dẫn đến vụ tai nạn (FIRST_CRASH_TYPE).

3.6.2.1 Manual

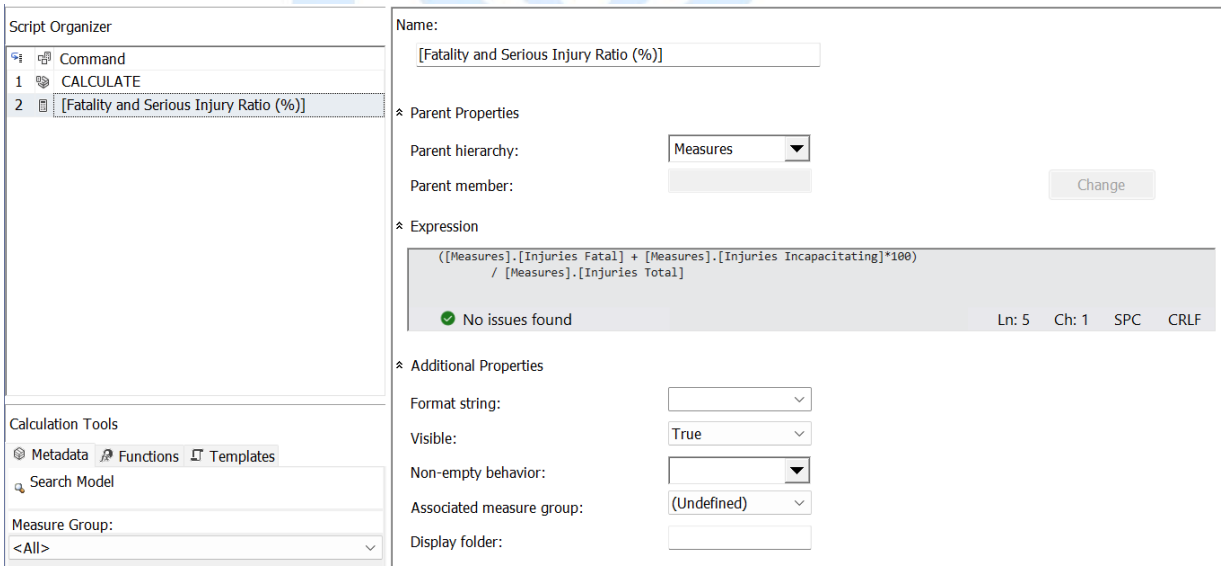
Bước 1: Vào mục **Calculations**, tạo mới một "New Calculated Member".



Sau đó nhập biểu thức và tiến hành Deloy.

Biểu thức:

```
([Measures].[Injuries Fatal] + [Measures].[Injuries Incapacitating]*100)
        / [Measures].[Injuries Total]
```



Bước 2: Vào mục **Browser**, kéo từ bảng Dim First Crash Types thuộc tính "FIRST CRASH TYPE" và biểu thức "Fatality and Serious Injury Ratio (%)" vừa tạo. Sau đó nhấn "Click to execute the query". Ta được kết quả.

Dimension	Hierarchy	Operator	Filter Expression	Paramet...
<Select dimension>				

FIRST CRASH TYPE	Fatality and Serious Injury Ratio (%)
ANGLE	9.14382693994218
ANIMAL	25.7948717948718
FIXED OBJECT	17.3561460987744
HEAD ON	14.5145811240721
OTHER NONCOLLISION	17.027972027972
OTHER OBJECT	17.9915611814346
OVERTURNED	19.6509090909091
PARKED MOTOR VEHICLE	13.0263090349076
PEDALCYCLIST	13.3604895104895
PEDESTRIAN	18.9091015785459
REAR END	6.12198283844682
REAR TO FRONT	8.00444444444445
REAR TO REAR	11.7647058823529
REAR TO SIDE	10.1149425287356
SIDESWIPE OPPOSITE DIRECTION	9.94249022892239
SIDESWIPE SAME DIRECTION	8.19076887394178
TRAIN	11.6153846153846
TURNING	8.51178181818182

3.6.2.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
WITH
MEMBER [Measures].[Fatality and Serious Injury Ratio (%)] AS
    IIF(
        [Measures].[Injuries Total] = 0,
        NULL,
        ([Measures].[Injuries Fatal] + [Measures].[Injuries Incapacitating]*100)
        / [Measures].[Injuries Total]
    )

SELECT
    {[Measures].[Fatality and Serious Injury Ratio (%)]} ON COLUMNS,
    [Dim First Crash Types].[FIRST CRASH TYPE].[FIRST CRASH TYPE].MEMBERS ON ROWS
FROM
         [Traffic Crash];
```

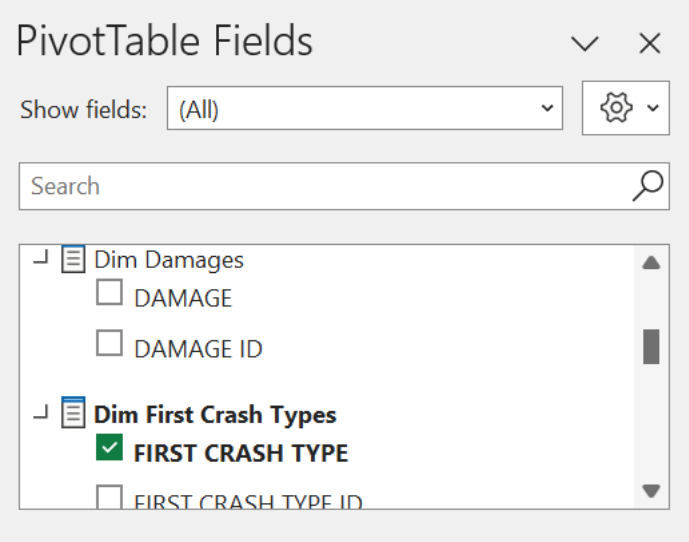
Kết quả câu truy vấn MDX:

Messages | Results

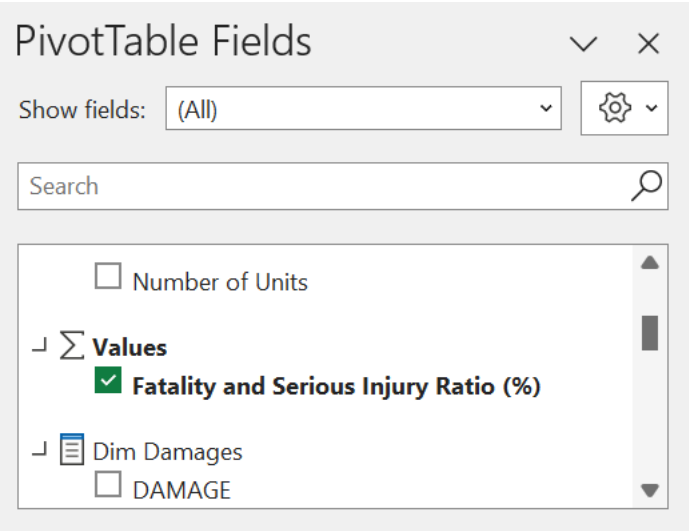
	Fatality and Serious Injury Ratio (%)
ANGLE	9.14382693994218
ANIMAL	25.7948717948718
FIXED OBJECT	17.3561460987744
HEAD ON	14.5145811240721
OTHER NONCOLLISION	17.027972027972
OTHER OBJECT	17.9915611814346
OVERTURNED	19.6509090909091
PARKED MOTOR VEHICLE	13.0263090349076
PEDALCYCLIST	13.3604895104895
PEDESTRIAN	18.9091015785459
REAR END	6.12198283844682
REAR TO FRONT	8.00444444444445
REAR TO REAR	11.7647058823529
REAR TO SIDE	10.1149425287356
SIDESWIPE OPPOSITE DIRECTION	9.94249022892239
SIDESWIPE SAME DIRECTION	8.19076887394178
TRAIN	11.6153846153846
TURNING	8.51178181818182

3.6.2.3 Excel

Bước 1: Trong **PivotTable Fields**, tick chọn thuộc tính là "FIRST CRASH TYPE".



Bước 2: Trong **Values** chọn biểu thức "Fatality and Serious Injury Ratio (%)" đã tạo.



Kết quả:

FIRST CRASH TYPE	Fatality and Serious Injury Ratio (%)
ANGLE	9.14382694
ANIMAL	25.79487179
FIXED OBJECT	17.3561461
HEAD ON	14.51458112
OTHER NONCOLLISION	17.02797203
OTHER OBJECT	17.99156118
OVERTURNED	19.65090909
PARKED MOTOR VEHICLE	13.02630903
PEDALCYCLIST	13.36048951
PEDESTRIAN	18.90910158
REAR END	6.121982838
REAR TO FRONT	8.004444444
REAR TO REAR	11.76470588
REAR TO SIDE	10.11494253
SIDESWIPE OPPOSITE DIRECTION	9.942490229
SIDESWIPE SAME DIRECTION	8.190768874
TRAIN	11.61538462
TURNING	8.511781818
Grand Total	**10.59080443**

3.6.2.4 Nhận xét

Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.3 Câu truy vấn 3

Nội dung câu truy vấn: Liệt kê các thông số về chấn thương do tai nạn giao thông từ 2016 đến 2023.

3.6.3.1 Manual

Vào mục **Browser**, kéo từ bảng Dim Time thuộc tính "CRASH YEAR" và các độ đo cần thiết. Sau đó nhấn "Click to execute the query". Ta được kết quả.

Dimension	Hierarchy	Operator	Filter Expression	Paramet...
<Select dimension>				

CRASH YEAR	Injuries Fatal	Injuries Incapacitating	Injuries Non Incapacitati...	Injuries No Indication	Number of Traffic Crash Records
2016	14	344	1614	92309	44239
2017	87	1466	6852	172054	83638
2018	127	2497	12664	242922	118729
2019	108	2376	12641	240098	117497
2020	146	2241	11465	179083	91857
2021	165	2378	13173	211331	108514
2022	150	2286	12735	211618	108107
2023	148	2225	13386	219276	110300

3.6.3.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
SELECT
    {[Measures].[Injuries Fatal],
      [Measures].[Injuries Incapacitating],
      [Measures].[Injuries Non Incapacitating],
      [Measures].[Injuries No Indication],
      [Measures].[Number of Traffic Crash Records]} ON COLUMNS,
    [Dim Time].[CRASH YEAR].&[2016]:[Dim Time].[CRASH YEAR].&[2023] ON ROWS
FROM [Traffic Crash];
```

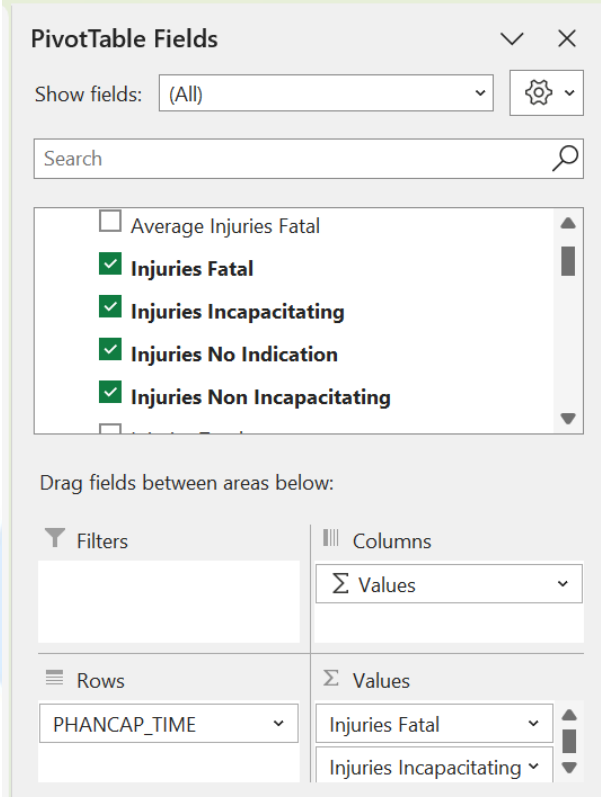
Kết quả câu truy vấn MDX:

Messages | Results

	Injuries Fatal	Injuries Incapacitating	Injuries Non Incapacitating	Injuries No Indication	Number of Traffic Crash Records
2016	14	344	1614	92309	44239
2017	87	1466	6852	172054	83638
2018	127	2497	12664	242922	118729
2019	108	2376	12641	240098	117497
2020	146	2241	11465	179083	91857
2021	165	2378	13173	211331	108514
2022	150	2286	12735	211618	108107
2023	148	2225	13386	219276	110300

3.6.3.3 Excel

Trong **PivotTable Fields**, tick chọn thuộc tính là "CRASH YEAR" cùng các độ đo cần thiết. Chuyển thuộc tính CRASH YEAR sang hiển thị ở dạng hàng.



Kết quả:

CRASH YEAR	Injuries Fatal	Injuries Incapacitating	Injuries Non Incapacitating	Injuries No Indication	Number of Traffic Crash Records
2016	14	344	1614	92309	44239
2017	87	1466	6852	172054	83638
2018	127	2497	12664	242922	118729
2019	108	2376	12641	240098	117497
2020	146	2241	11465	179083	91857
2021	165	2378	13173	211331	108514
2022	150	2286	12735	211618	108107
2023	148	2225	13386	219276	110300
Grand Total	**945**	**15813**	**84530**	**1568691**	**782881**

3.6.3.4 Nhận xét

Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.4 Câu truy vấn 4

Nội dung câu truy vấn: Trong 10 nguyên nhân gây va chạm giao thông phổ biến nhất ở mỗi năm, thống kê số lượng nạn nhân tử vong, bị thương nặng và bị thương nhẹ ở mỗi nguyên nhân từ năm 2021 đến năm 2023 khi điều kiện thời tiết là "CLEAR".

Ở đây ta sẽ không tính nguyên nhân là "NOT APPLICABLE" và "UNABLE TO DETERMINE" vì đây là hai nguyên nhân không xác định.

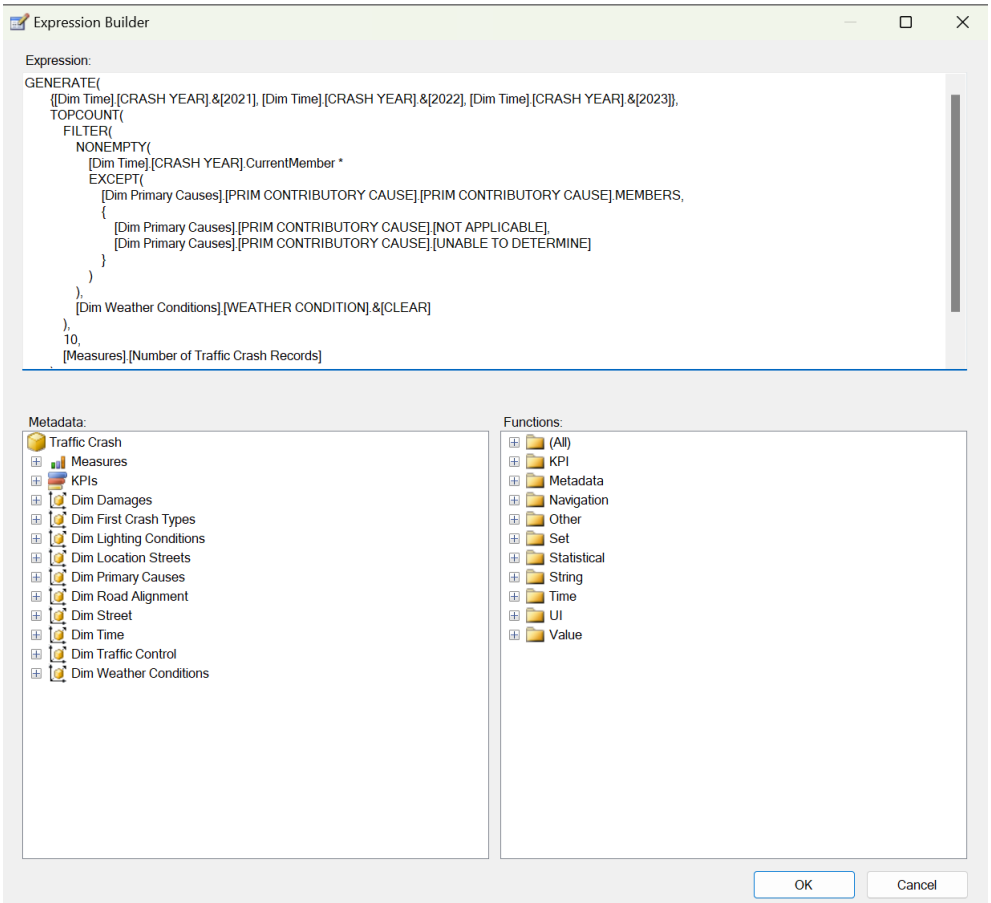
3.6.4.1 Manual

Bước 1: Vào mục **Browser**, chọn chọn thuộc tính PRIM CONTRIBUTORY CAUSE và trong **Operator** chọn **Custom**. Tiếp theo ta nhập biểu thức.

Biểu thức:

```
GENERATE(
  {[Dim Time].[CRASH YEAR].&[2021], [Dim Time].[CRASH YEAR].&[2022], [Dim Time].[CRASH YEAR].&[2023]},
      TOPCOUNT(
        FILTER(
          NONEMPTY(
            [Dim Time].[CRASH YEAR].CurrentMember *
            EXCEPT(
              [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM CONTRIBUTORY CAUSE].MEMBERS,
              {
                [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[NOT APPLICABLE],
                [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[UNABLE TO DETERMINE]
              }
            )
          ),
          [Dim Weather Conditions].[WEATHER CONDITION].&[CLEAR]
        ),
        10,
        [Measures].[Number of Traffic Crash Records]
      )
  )
```

Sau đó nhấn “OK”.



Bước 2: Tiếp theo, ta kéo từ bảng Dim Primary Causes thuộc tính “PRIM CONTRIBUTORY CAUSE” và “CRASH YEAR” từ bảng Dim Time, cùng các độ đo cần thiết. Sau đó nhấn “Click to execute the query”. Ta được kết quả.

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim Primary Causes	PRIM CONTRIBUTORY C...	Custom	GENERATE({[Dim Time].[CRASH YEAR].&[2021], [Dim Time].[...	
<Select dimension>				

PRIM CONTRIBUTORY CAUSE	CRASH YEAR	Injuries Fatal	Injuries Incapacitating	Injuries Non Incapacita...
DRIVING SKILLS/KNOWLEDGE/EXPERIENCE	2021	1	53	312
DRIVING SKILLS/KNOWLEDGE/EXPERIENCE	2022	0	52	307
DRIVING SKILLS/KNOWLEDGE/EXPERIENCE	2023	2	63	377
FAILING TO REDUCE SPEED TO AVOID CRASH	2021	19	192	1068
FAILING TO REDUCE SPEED TO AVOID CRASH	2022	13	156	845
FAILING TO REDUCE SPEED TO AVOID CRASH	2023	12	148	924
FAILING TO YIELD RIGHT-OF-WAY	2021	11	409	2374
FAILING TO YIELD RIGHT-OF-WAY	2022	14	427	2502
FAILING TO YIELD RIGHT-OF-WAY	2023	11	386	2753
FOLLOWING TOO CLOSELY	2021	1	98	674
FOLLOWING TOO CLOSELY	2022	1	89	690
FOLLOWING TOO CLOSELY	2023	1	79	797
IMPROPER BACKING	2021	1	12	56
IMPROPER BACKING	2022	0	17	78
IMPROPER BACKING	2023	0	8	50
IMPROPER LANE USAGE	2021	2	37	250
IMPROPER LANE USAGE	2022	2	41	282
IMPROPER LANE USAGE	2023	6	43	272
IMPROPER OVERTAKING/PASSING	2021	3	52	338
IMPROPER OVERTAKING/PASSING	2022	4	59	303
IMPROPER OVERTAKING/PASSING	2023	2	70	444
IMPROPER TURNING/NO SIGNAL	2021	1	60	445
IMPROPER TURNING/NO SIGNAL	2022	3	62	446
IMPROPER TURNING/NO SIGNAL	2023	0	88	535
OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR...	2023	7	56	294

3.6.4.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
SELECT
    {[Measures].[Injuries Fatal],  [Measures].[Injuries Incapacitating],
[Measures].[Injuries Non Incapacitating]} ON COLUMNS,

    NON EMPTY
    GENERATE(
        {[Dim Time].[CRASH YEAR].&[2021], [Dim Time].[CRASH YEAR].&[2022],
[Dim Time].[CRASH YEAR].&[2023]},
        TOPCOUNT(
            FILTER(
                NONEMPTY(
                    [Dim Time].[CRASH YEAR].CurrentMember *
                    EXCEPT(
                        [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM
CONTRIBUTORY CAUSE].MEMBERS,
                        {
                            [Dim Primary Causes].[PRIM CONTRIBUTORY
CAUSE].[NOT APPLICABLE],
                            [Dim Primary Causes].[PRIM CONTRIBUTORY
CAUSE].[UNABLE TO DETERMINE]
                        }
                    )
                ),
                [Dim Weather Conditions].[WEATHER CONDITION].&[CLEAR]
            ),
```

```
            10,
            [Measures].[Number of Traffic Crash Records]
        )
    ) ON ROWS
     FROM [Traffic Crash];
```

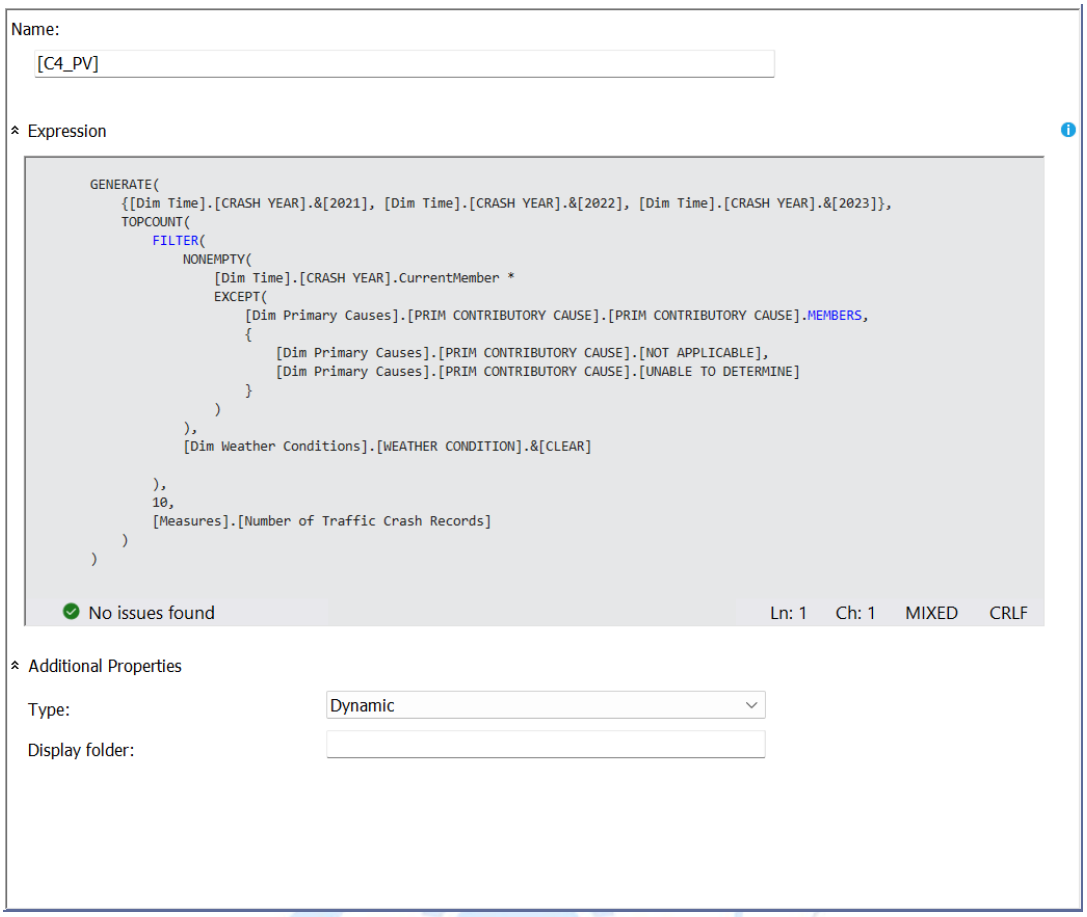
Kết quả câu truy vấn MDX:

Messages | Results

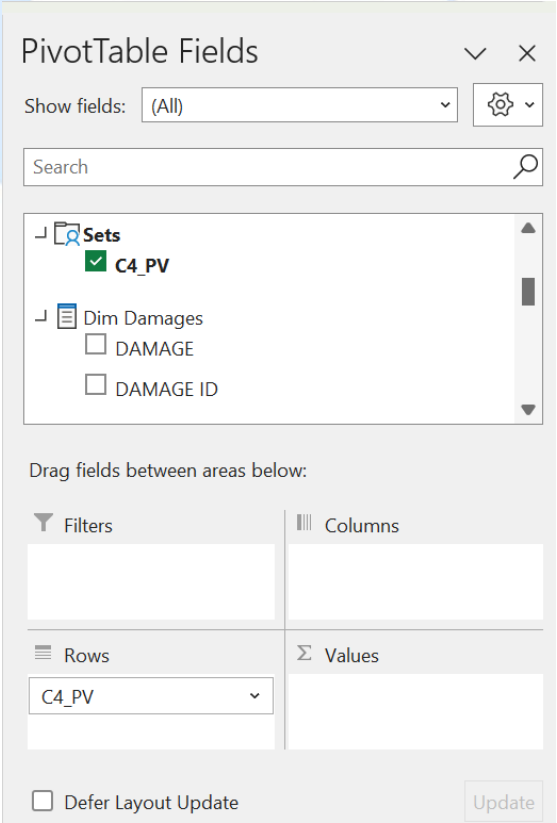
		Injuries Fatal	Injuries Incapacitating	Injuries Non Incapacitating
2021	FAILING TO YIELD RIGHT-OF-WAY	11	409	2374
2021	FOLLOWING TOO CLOSELY	1	98	674
2021	IMPROPER OVERTAKING/PASSING	3	52	338
2021	FAILING TO REDUCE SPEED TO AVOID CRASH	19	192	1068
2021	IMPROPER BACKING	1	12	56
2021	DRIVING SKILLS/KNOWLEDGE/EXPERIENCE	1	53	312
2021	IMPROPER LANE USAGE	2	37	250
2021	IMPROPER TURNING/NO SIGNAL	1	60	445
2021	DISREGARDING TRAFFIC SIGNALS	19	165	1049
2021	WEATHER	0	29	163
2022	FAILING TO YIELD RIGHT-OF-WAY	14	427	2502
2022	FOLLOWING TOO CLOSELY	1	89	690
2022	IMPROPER OVERTAKING/PASSING	4	59	303
2022	FAILING TO REDUCE SPEED TO AVOID CRASH	13	156	845
2022	DRIVING SKILLS/KNOWLEDGE/EXPERIENCE	0	52	307
2022	IMPROPER BACKING	0	17	78
2022	IMPROPER TURNING/NO SIGNAL	3	62	446
2022	IMPROPER LANE USAGE	2	41	282
2022	DISREGARDING TRAFFIC SIGNALS	10	186	1003
2022	WEATHER	0	23	243
2023	FAILING TO YIELD RIGHT-OF-WAY	11	386	2753
2023	FOLLOWING TOO CLOSELY	1	79	797
2023	IMPROPER OVERTAKING/PASSING	2	70	444
2023	DRIVING SKILLS/KNOWLEDGE/EXPERIENCE	2	63	377
2023	FAILING TO REDUCE SPEED TO AVOID CRASH	12	148	924
2023	IMPROPER TURNING/NO SIGNAL	0	88	535
2023	IMPROPER BACKING	0	8	50
2023	IMPROPER LANE USAGE	6	43	272
2023	DISREGARDING TRAFFIC SIGNALS	9	152	965
2023	OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	7	56	294

3.6.4.3 Excel

Bước 1: Tạo nameset “C4_PV” cho PivotTable Fields.



Bước 2: Trong **Sets**, ta chọn Named Set đã tạo là "C4_PV".



Kết quả:

PRIMARY CONTRIBUTORY CAUSE	Injuries Fatal	Injuries Incapacitating	Injuries Non Incapacitating
2021			
DISREGARDING TRAFFIC SIGNALS	19	165	1049
DRIVING SKILLS/KNOWLEDGE/EXPERIENCE	1	53	312
FAILING TO REDUCE SPEED TO AVOID CRASH	19	192	1068
FAILING TO YIELD RIGHT-OF-WAY	11	409	2374
FOLLOWING TOO CLOSELY	1	98	674
IMPROPER BACKING	1	12	56
IMPROPER LANE USAGE	2	37	250
IMPROPER OVERTAKING/PASSING	3	52	338
IMPROPER TURNING/NO SIGNAL	1	60	445
WEATHER	0	29	163
2022			
DISREGARDING TRAFFIC SIGNALS	10	186	1003
DRIVING SKILLS/KNOWLEDGE/EXPERIENCE	0	52	307
FAILING TO REDUCE SPEED TO AVOID CRASH	13	156	845
FAILING TO YIELD RIGHT-OF-WAY	14	427	2502
FOLLOWING TOO CLOSELY	1	89	690
IMPROPER BACKING	0	17	78
IMPROPER LANE USAGE	2	41	282
IMPROPER OVERTAKING/PASSING	4	59	303
IMPROPER TURNING/NO SIGNAL	3	62	446
WEATHER	0	23	243
2023			
DISREGARDING TRAFFIC SIGNALS	9	152	965
DRIVING SKILLS/KNOWLEDGE/EXPERIENCE	2	63	377
FAILING TO REDUCE SPEED TO AVOID CRASH	12	148	924
FAILING TO YIELD RIGHT-OF-WAY	11	386	2753
FOLLOWING TOO CLOSELY	1	79	797
IMPROPER BACKING	0	8	50
IMPROPER LANE USAGE	6	43	272
IMPROPER OVERTAKING/PASSING	2	70	444
IMPROPER TURNING/NO SIGNAL	0	88	535
OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	7	56	294

3.6.4.4 Nhận xét

Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.5 Câu truy vấn 5

Nội dung câu truy vấn: Liệt kê top 10 con đường có nhiều vị trí xảy ra tai nạn giao thông nhất, cùng với hướng đường và số địa điểm đã xảy ra va chạm giao thông ở mỗi hướng đường của con đường đó.

3.6.5.1 Manual

Bước 1: Vào mục **Calculations**, tạo mới một "New Named Set".

Sau đó nhập câu truy vấn và tiến hành Deloy.

Câu truy vấn:

```
TOPCOUNT(
        [Dim Street].[STREET NAME].[STREET NAME].MEMBERS,
        10,
        [Measures].[Number of Locations]
    )
```

Name:

[C5]

Expression

```
TOPCOUNT(
        [Dim Street].[STREET NAME].[STREET NAME].MEMBERS,
        10,
        [Measures].[Number of Locations]
    )
```

No issues found Ln: 7 Ch: 1 SPC CRLF

Additional Properties

Type: Dynamic

Display folder:

Bước 2: Vào mục **Browser**, kéo từ bảng Dim Street thuộc tính "STREET NAME". Tại cột **Operator** chọn "In" và **Filter Expression** chọn vào Named Set "C5" vừa tạo. Sau đó kéo các thuộc tính để hiển thị kết quả câu truy vấn và nhấn "Click to execute the query". Ta được kết quả.

Dimension	Hierarchy	Operator	Filter Expression	Param...
Dim Street	STREET NAME	In	C5	
<Select dimension>				

STREET NAME	STREET DIRECTION	Number of Locations
ASHLAND AVE	N	1828
ASHLAND AVE	S	2608
CALIFORNIA AVE	N	1330
CALIFORNIA AVE	S	1495
CENTRAL AVE	N	1928
CENTRAL AVE	S	573
CICERO AVE	N	2020
CICERO AVE	S	1545
CLARK ST	N	2222
CLARK ST	S	415
DAMEN AVE	N	1320
DAMEN AVE	S	1426
HALSTED ST	N	1105
HALSTED ST	S	2743
KEDZIE AVE	N	1467
KEDZIE AVE	S	2454
PULASKI RD	N	1975
PULASKI RD	S	2545
WESTERN AVE	N	2490
WESTERN AVE	S	2907

3.6.5.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
SELECT
    {
        [Measures].[Number of Locations]
    } ON COLUMNS,

    NONEMPTY(
        CROSSJOIN(
            TOPCOUNT(
                [Dim Street].[STREET NAME].[STREET NAME].MEMBERS,
                10,
                [Measures].[Number of Locations]
            ),
            [Dim Location Streets].[STREET DIRECTION].[STREET DIRECTION].MEMBERS
        ),
        [Measures].[Number of Locations]
    ) ON ROWS

FROM [Traffic Crash];
```

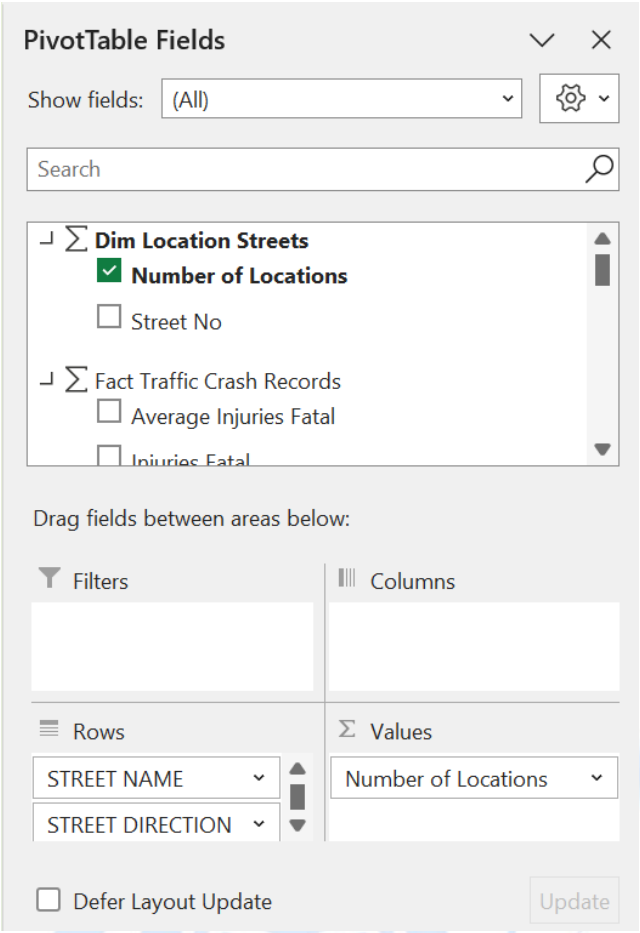
Kết quả câu truy vấn MDX:

Messages | Results

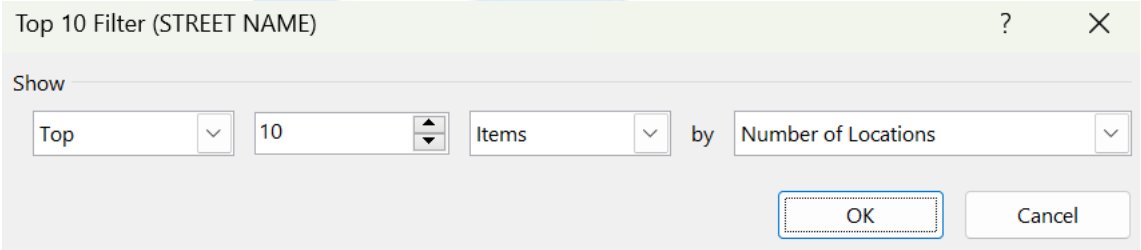
		Number of Locations
WESTERN AVE	N	2490
WESTERN AVE	S	2907
PULASKI RD	N	1975
PULASKI RD	S	2545
ASHLAND AVE	N	1828
ASHLAND AVE	S	2608
KEDZIE AVE	N	1467
KEDZIE AVE	S	2454
HALSTED ST	N	1105
HALSTED ST	S	2743
CICERO AVE	N	2020
CICERO AVE	S	1545
CALIFORNIA AVE	N	1330
CALIFORNIA AVE	S	1495
DAMEN AVE	N	1320
DAMEN AVE	S	1426
CLARK ST	N	2222
CLARK ST	S	415
CENTRAL AVE	N	1928
CENTRAL AVE	S	573

3.6.5.3 Excel

Bước 1: Trong **PivotTable Fields**, chọn thuộc tính là “STREET DIRECTION”, “STREET NAME” và “Number of Locations”.



Bước 2: Nhấn vào biểu tượng cạnh Row Labels chọn **Value Filters**, sau đó chọn Top 10.



Kết quả:

Row Labels	Number of Locations
⊟ASHLAND AVE	
N	1828
S	2608
⊟CALIFORNIA AVE	
N	1330
S	1495
⊟CENTRAL AVE	
N	1928
S	573
⊟CICERO AVE	
N	2020
S	1545
⊟CLARK ST	
N	2222
S	415
⊟DAMEN AVE	
N	1320
S	1426
⊟HALSTED ST	
N	1105
S	2743
⊟KEDZIE AVE	
N	1467
S	2454
⊟PULASKI RD	
N	1975
S	2545
⊟WESTERN AVE	
N	2490
S	2907
Grand Total	**36396**

3.6.5.4 Nhận xét

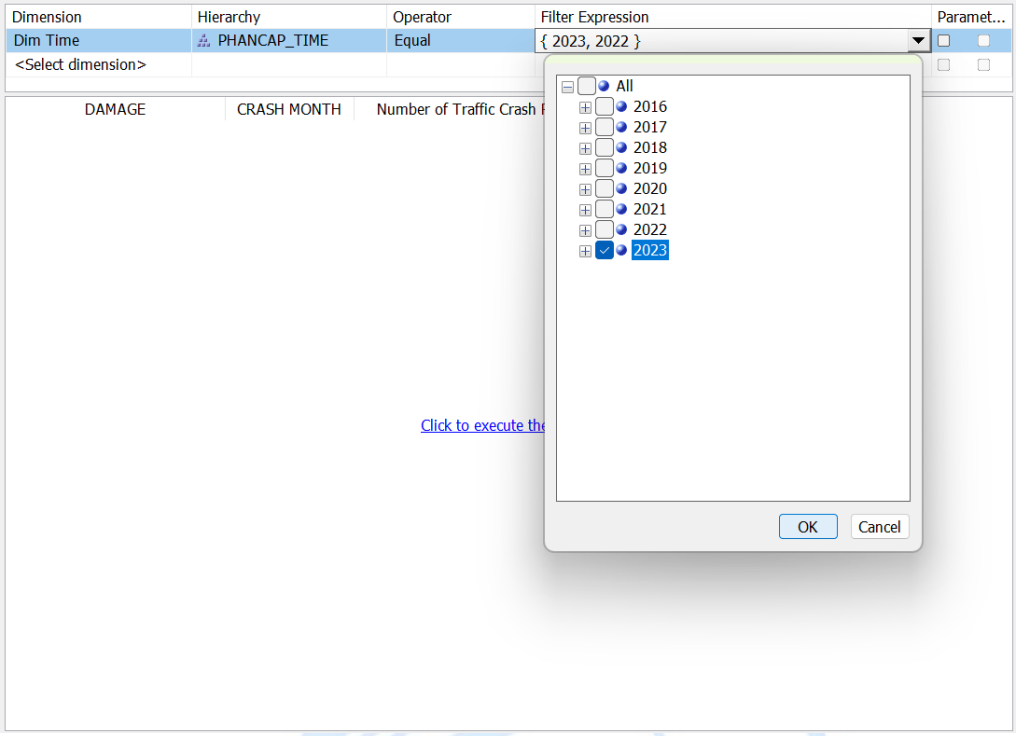
Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.6 Câu truy vấn 6

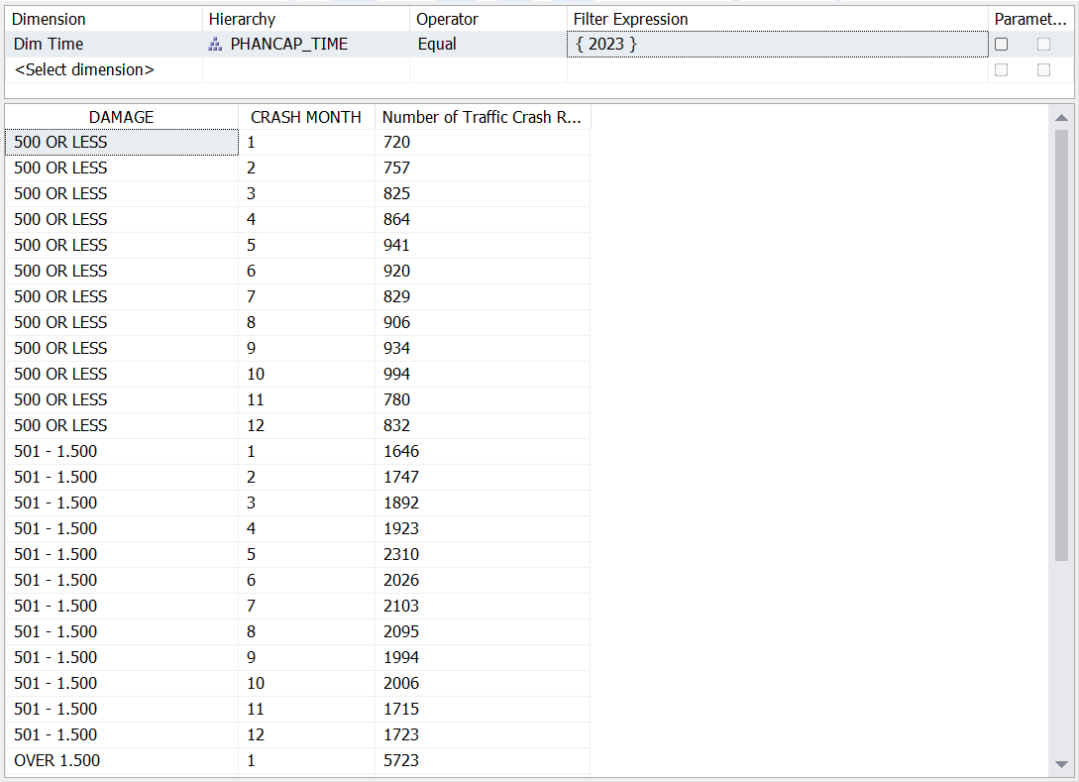
Nội dung câu truy vấn: Thống kê số lượng các vụ va chạm giao thông cho từng mức độ thiệt hại theo từng tháng trong năm 2023.

3.6.6.1 Manual

Bước 1: Chọn **Hierachy** PHANCAP_TIME ở Dim Time. Ở **Operator** chọn "Equal". Tiếp theo ở **Filter Expression** chọn năm 2023.



Bước 2: Kéo các thuộc tính để hiển thị kết quả câu truy vấn và nhấn "Click to execute the query". Ta được kết quả.



Dimension | Hierarchy | Operator | Filter Expression | Paramet...
Dim Time | PHANCAP_TIME | Equal | { 2023 }
<Select dimension>

DAMAGE	CRASH MONTH	Number of Traffic Crash R...
500 OR LESS	1	720
500 OR LESS	2	757
500 OR LESS	3	825
500 OR LESS	4	864
500 OR LESS	5	941
500 OR LESS	6	920
500 OR LESS	7	829
500 OR LESS	8	906
500 OR LESS	9	934
500 OR LESS	10	994
500 OR LESS	11	780
500 OR LESS	12	832
501 - 1.500	1	1646
501 - 1.500	2	1747
501 - 1.500	3	1892
501 - 1.500	4	1923
501 - 1.500	5	2310
501 - 1.500	6	2026
501 - 1.500	7	2103
501 - 1.500	8	2095
501 - 1.500	9	1994
501 - 1.500	10	2006
501 - 1.500	11	1715
501 - 1.500	12	1723
OVER 1.500	1	5723

3.6.6.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
WITH
    MEMBER [Measures].[Accident Count Less 500] AS
        SUM(
            FILTER(
                [Dim Damages].[DAMAGE].[DAMAGE].MEMBERS,
                [Dim Damages].[DAMAGE].CURRENTMEMBER IS [Dim Damages].[DAMAGE].[500 OR LESS]
            ),
            [Measures].[Number of Traffic Crash Records]
        )

    MEMBER [Measures].[Accident Count 501 1500] AS
        SUM(
            FILTER(
                [Dim Damages].[DAMAGE].[DAMAGE].MEMBERS,
                [Dim Damages].[DAMAGE].CURRENTMEMBER IS [Dim Damages].[DAMAGE].[501 - 1.500]
            ),
            [Measures].[Number of Traffic Crash Records]
        )

    MEMBER [Measures].[Accident Count Over 1500] AS
        SUM(
            FILTER(
                [Dim Damages].[DAMAGE].[DAMAGE].MEMBERS,
                [Dim Damages].[DAMAGE].CURRENTMEMBER IS [Dim Damages].[DAMAGE].[OVER 1.500]
            ),
            [Measures].[Number of Traffic Crash Records]
        )

SELECT
    {
        [Measures].[Accident Count Less 500],
        [Measures].[Accident Count 501 1500],
        [Measures].[Accident Count Over 1500]
    } ON COLUMNS,

    [Dim Time].[CRASH MONTH].[CRASH MONTH].MEMBERS ON ROWS

FROM
    [Traffic Crash]
WHERE
    (
        [Dim Time].[CRASH YEAR].[CRASH YEAR].&[2023]
    );
```

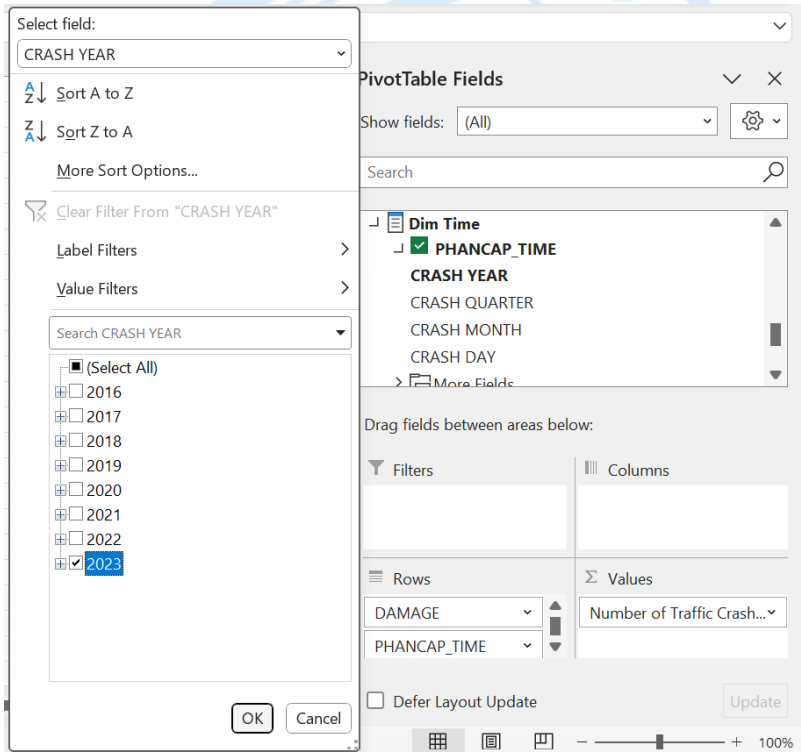
Kết quả câu truy vấn MDX:

Messages | Results

	Accident Count Less 500	Accident Count 501 1500	Accident Count Over 1500
1	720	1646	5723
2	757	1747	5572
3	825	1892	6157
4	864	1923	6200
5	941	2310	6872
6	920	2026	6619
7	829	2103	6567
8	906	2095	6749
9	934	1994	6650
10	994	2006	7063
11	780	1715	6236
12	832	1723	6410

3.6.6.3 Excel

Bước 1: Trong **PivotTable Fields**, ta chọn thuộc tính là “DAMAGE” và “Number of Traffic Crash Records” . Sau đó chọn năm 2023 trong “PHANCAP_TIME”.



Kết quả:

Row Labels	Number of Traffic Crash Records
500 OR LESS	
⊟2023	
⊟1	
⊞1	720
⊞2	757
⊞3	825
⊟2	
⊞4	864
⊞5	941
⊞6	920
⊟3	
⊞7	829
⊞8	906
⊞9	934
⊟4	
⊞10	994
⊞11	780
⊞12	832
501 - 1.500	
⊟2023	
⊟1	
⊞1	1646
⊞2	1747
⊞3	1892
⊟2	
⊞4	1923
⊞5	2310
⊞6	2026
⊟3	
⊞7	2103
⊞8	2095
⊞9	1994
⊟4	
⊞10	2006

3.6.6.4 Nhận xét

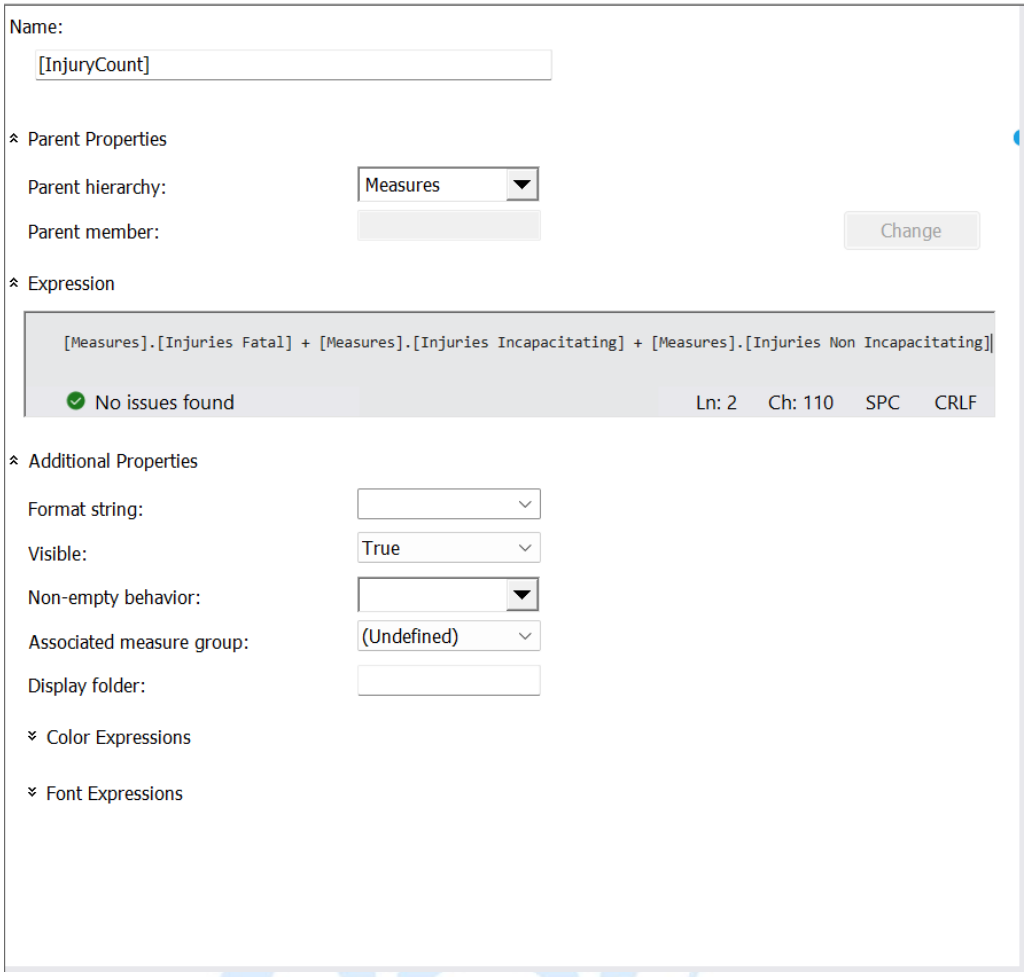
Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.7 Câu truy vấn 7

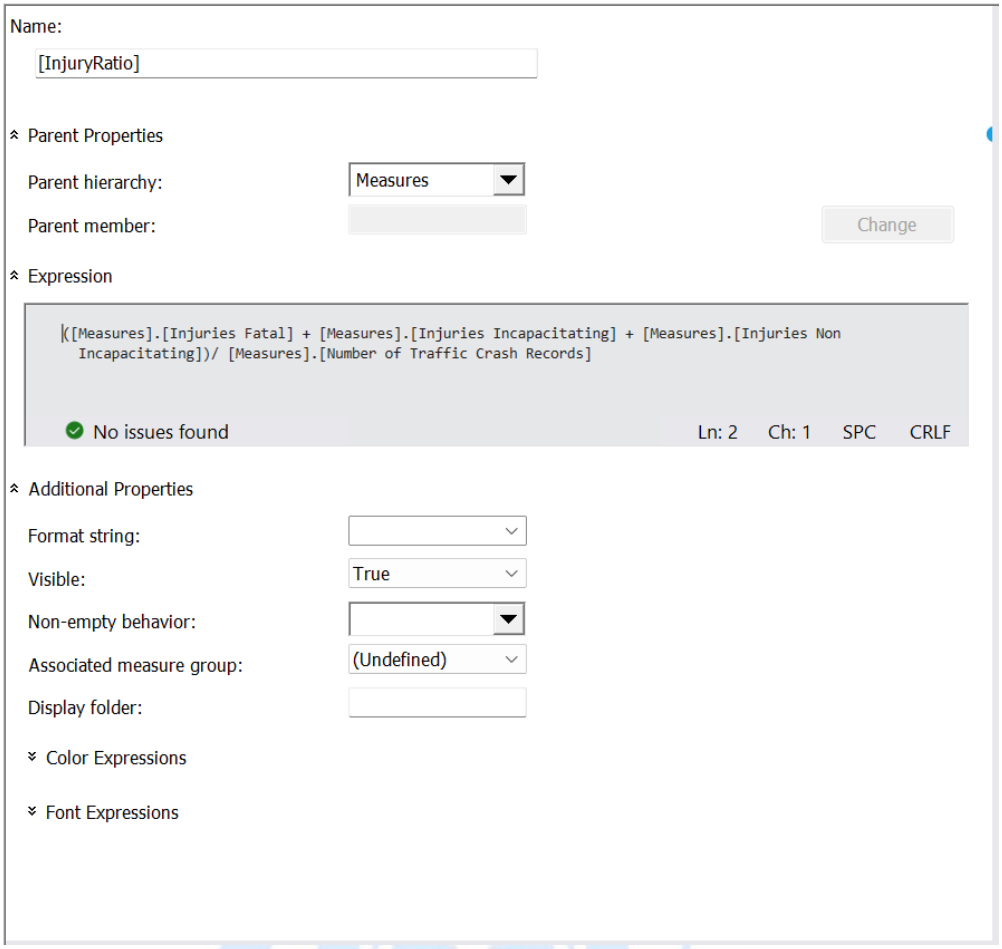
Nội dung câu truy vấn: Thống kê số nạn nhân bị thương ở mỗi vụ tai nạn [(số nạn nhân tử vong + bị thương nặng + bị thương nhẹ) / số vụ tai nạn] xảy ra ở loại cấu trúc đường cong (ALIGNMENT có kí tự "CURVE") khi trời tối (LIGHTING CONDITION có kí tự "DARK" trong tên) qua từng năm (từ 2016 đến 2023).

3.6.7.1 Manual

Bước 1: Vào mục **Calculations**, tạo mới hai Calculated Member đặt tên là "InjuryCount" để tính tổng số lượng các nạn nhân bị thương nhẹ, bị thương nặng và tử vong.



Calculated Member "InjuryRatio" để tính tỉ lệ bị thương (tử vong + bị thương nặng + bị thương nhẹ) ở mỗi vụ tai nạn (nạn nhân/vụ).



Bước 2: Tạo hai Named Set cần thiết cho câu truy vấn. Named Set đầu tiên đặt tên là "C7" để lọc ra các kiến trúc đường cong (có chứa kí tự "CURVE").

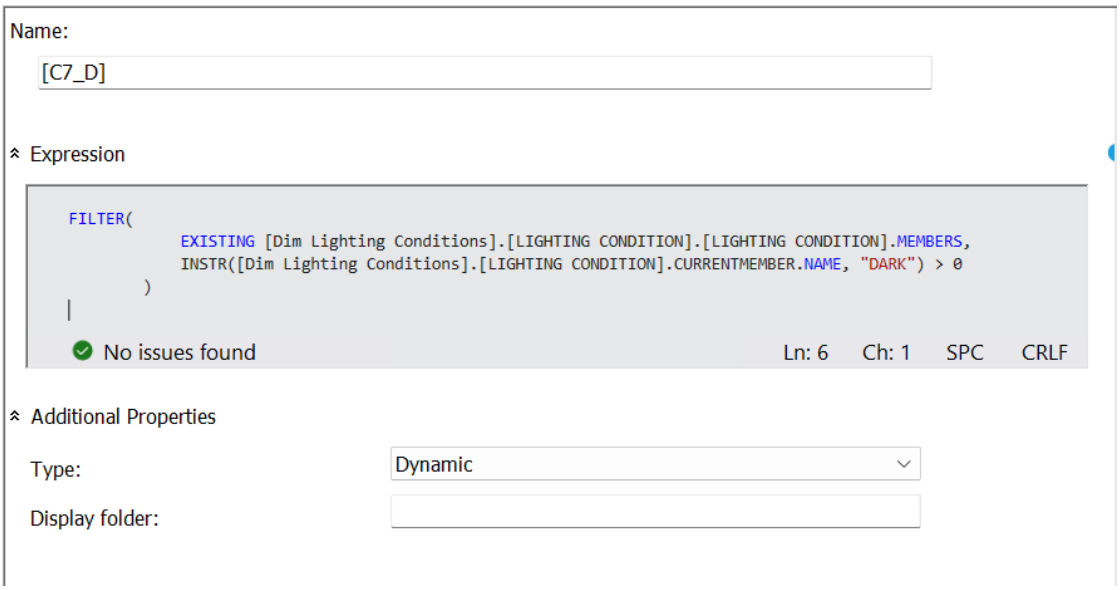
Câu truy vấn:

```
FILTER(
            EXISTING [Dim Road Alignment].[ALIGNMENT].[ALIGNMENT].MEMBERS,
            INSTR([Dim Road Alignment].[ALIGNMENT].CURRENTMEMBER.NAME, "CURVE") > 0
        )
```

Named Set thứ hai đặt tên là “C7_D” để lọc ra điều kiện ánh sáng khi trời tối (có kí tự “DARK”).

Câu truy vấn:

```
FILTER(
            EXISTING [Dim Lighting Conditions].[LIGHTING CONDITION].[LIGHTING
CONDITION].MEMBERS,
            INSTR([Dim Lighting Conditions].[LIGHTING CONDITION].CURRENTMEMBER.NAME,
"DARK") > 0
        )
```



Bước 3: Vào mục **Browser**, kéo từ bảng Dim Road Alignment thuộc tính "ALIGNMENT". Tại cột **Operator** chọn "In" và **Filter Expression** chọn vào Named Set "C7" vừa tạo. Tương tự, kéo từ bảng Dim Lighting Conditions thuộc tính "LIGHTING CONDITION". Tại cột **Operator** chọn "In" và **Filter Expression** chọn vào Named Set "C7_D" .

Sau đó kéo các thuộc tính cần thiết để hiển thị kết quả câu truy vấn và nhấn "Click to execute the query". Ta được kết quả.

Dimension	Hierarchy	Operator	Filter Expression	Param...
Dim Road Alignment	ALIGNMENT	In	C7	
Dim Lighting Condi...	LIGHTING CONDI...	In	C7_D	
<Select dimension>				

CRASH YEAR	InjuryCount	Number of Traffic Cr...	InjuryRatio
2016	15	124	0.120967741935484
2017	50	284	0.176056338028169
2018	94	420	0.223809523809524
2019	97	371	0.261455525606469
2020	100	420	0.238095238095238
2021	123	450	0.273333333333333
2022	104	363	0.286501377410468
2023	110	375	0.293333333333333

3.6.7.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
WITH
    MEMBER [Measures].[InjuryCount] AS
        ([Measures].[Injuries Fatal] + [Measures].[Injuries Incapacitating] +
[Measures].[Injuries Non Incapacitating])

    MEMBER [Measures].[InjuryRatio] AS
```

```
        IIF(
            [Measures].[Number of Traffic Crash Records] > 0,
            [Measures].[InjuryCount] / [Measures].[Number of Traffic Crash Records],
            NULL
        )

SELECT
    {
        [Measures].[InjuryCount],
        [Measures].[Number of Traffic Crash Records],
        [Measures].[InjuryRatio]
    } ON COLUMNS,

    [Dim Time].[CRASH YEAR].[CRASH YEAR].MEMBERS ON ROWS

FROM
    [Traffic Crash]

WHERE
    (
      FILTER(
        EXISTING [Dim Road Alignment].[ALIGNMENT].[ALIGNMENT].MEMBERS,
        INSTR([Dim Road Alignment].[ALIGNMENT].CURRENTMEMBER.NAME, "CURVE") > 0
        ),
      FILTER(
        EXISTING [Dim Lighting Conditions].[LIGHTING CONDITION].[LIGHTING
CONDITION].MEMBERS,
        INSTR([Dim Lighting Conditions].[LIGHTING CONDITION].CURRENTMEMBER.NAME, "DARK") > 0
        )
    );
```

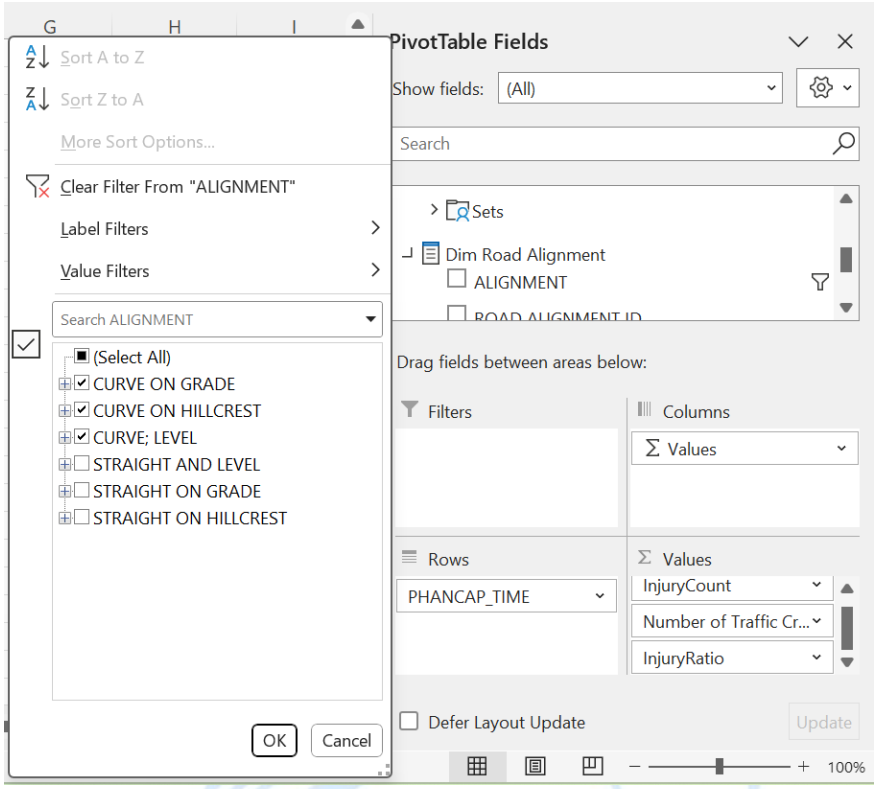
Kết quả câu truy vấn MDX:

Messages | Results

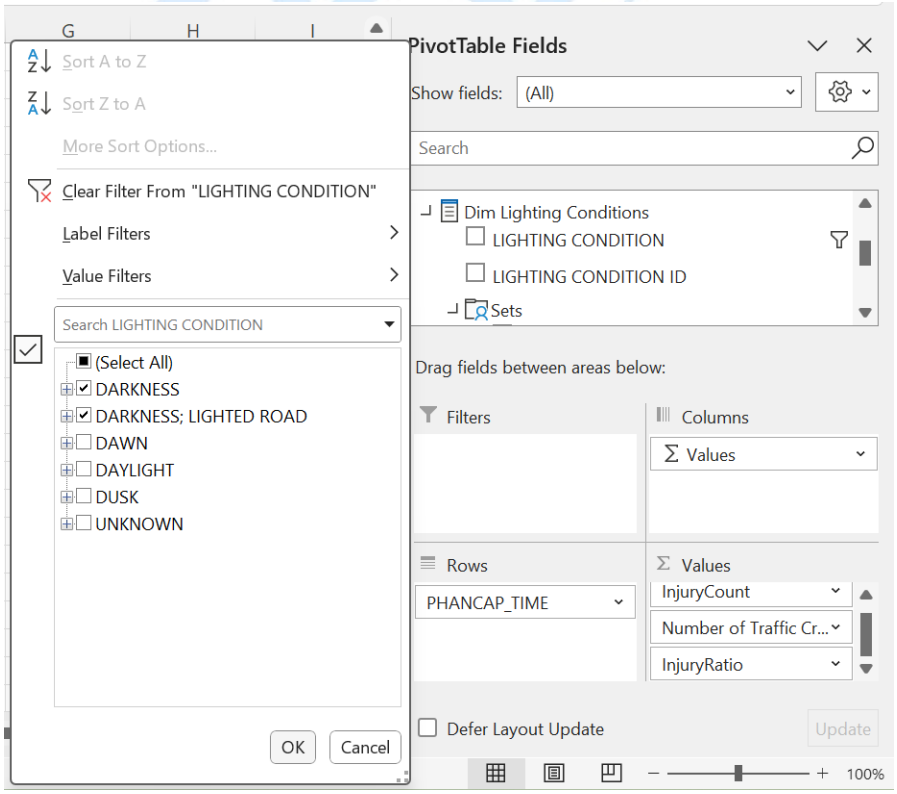
	InjuryCount	Number of Traffic Crash Records	InjuryRatio
2016	15	124	0.120967741935484
2017	50	284	0.176056338028169
2018	94	420	0.223809523809524
2019	97	371	0.261455525606469
2020	100	420	0.238095238095238
2021	123	450	0.273333333333333
2022	104	363	0.286501377410468
2023	110	375	0.293333333333333

3.6.7.3 Excel

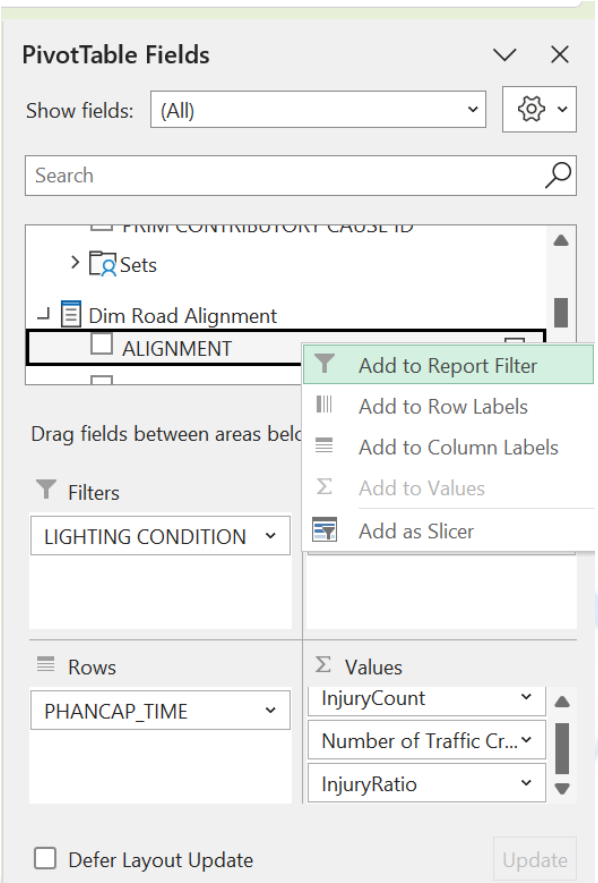
Bước 1: Trong **PivotTable Fields**, chọn thuộc tính là “ALIGNMENT” và chọn các giá trị có kí tự “CURVE”.



Bước 2: Trong **PivotTable Fields**, chọn thuộc tính là "LIGHTING CONDITION" và chọn các giá trị có kí tự "DARK".



Bước 3: Thêm hai giá trị "ALIGNMENT" và "LIGHTING CONDITION" đã lọc và **Filters**. Chọn các thuộc tính độ đo và Calculated Member cần hiển thị.



Ta được kết quả.

LIGHTING CONDITION	(Multiple Items)		
ALIGNMENT	(Multiple Items)		
Row Labels	**InjuryCount**	**Number of Traffic Crash Records**	**InjuryRatio**
2016	15	124	0.120967742
2017	50	284	0.176056338
2018	94	420	0.223809524
2019	97	371	0.261455526
2020	100	420	0.238095238
2021	123	450	0.273333333
2022	104	363	0.286501377
2023	110	375	0.293333333
Grand Total	**693**	**2807**	**0.246882793**

3.6.7.4 Nhận xét

Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.8 Câu truy vấn 8

Nội dung câu truy vấn: Thống kê số lượng phương tiện va chạm nhiều nhất trong một vụ tai nạn giao thông ở các nguyên nhân.

3.6.8.1 Manual

Vào mục **Browser**, kéo từ bảng Dim Primary Cause thuộc tính "PRIM CONTRIBUTORY CAUSE" và các thuộc tính độ đo để hiển thị kết quả câu truy vấn và nhấn "Click to execute the query". Ta được kết quả.

Dimension	Hierarchy	Operator	Filter Expression	Parame...
Dim Primary Causes	PRIM CONTRIBUTO...	Equal	{ All }	
<Select dimension>				

PRIM CONTRIBUTORY CAUSE	Max Number of Units
ANIMAL	7
BICYCLE ADVANCING LEGALLY ON RED LIGHT	3
CELL PHONE USE OTHER THAN TEXTING	10
DISREGARDING OTHER TRAFFIC SIGNS	6
DISREGARDING ROAD MARKINGS	8
DISREGARDING STOP SIGN	10
DISREGARDING TRAFFIC SIGNALS	8
DISREGARDING YIELD SIGN	4
DISTRACTION - FROM INSIDE VEHICLE	8
DISTRACTION - FROM OUTSIDE VEHICLE	8
DISTRACTION - OTHER ELECTRONIC DEVICE (NAVIGATION DEVICE; DVD PLAYER; ETC.)	6
DRIVING ON WRONG SIDE/WRONG WAY	10
DRIVING SKILLS/KNOWLEDGE/EXPERIENCE	7
EQUIPMENT - VEHICLE CONDITION	10
EVASIVE ACTION DUE TO ANIMAL; OBJECT; NONMOTORIST	7
EXCEEDING AUTHORIZED SPEED LIMIT	18
EXCEEDING SAFE SPEED FOR CONDITIONS	6
FAILING TO REDUCE SPEED TO AVOID CRASH	10
FAILING TO YIELD RIGHT-OF-WAY	10
FOLLOWING TOO CLOSELY	9
HAD BEEN DRINKING (USE WHEN ARREST IS NOT MADE)	15
IMPROPER BACKING	7
IMPROPER LANE USAGE	9
IMPROPER OVERTAKING/PASSING	9
IMPROPER TURNING/NO SIGNAL	8

3.6.8.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
SELECT
    {
        [Measures].[Max Number of Units]
    } ON COLUMNS,

    NONEMPTY(
        GENERATE(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM CONTRIBUTORY
CAUSE].MEMBERS,
            {[Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].CURRENTMEMBER}
        )
    ) ON ROWS
```

```
FROM
    [Traffic Crash];
```

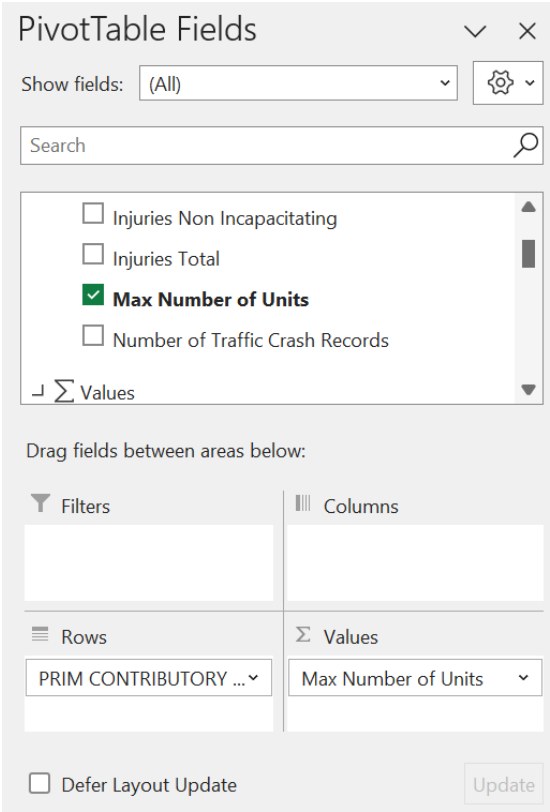
Kết quả câu truy vấn MDX:

Messages | Results

	Max Number of Units
ANIMAL	7
BICYCLE ADVANCING LEGALLY ON RED LIGHT	3
CELL PHONE USE OTHER THAN TEXTING	10
DISREGARDING OTHER TRAFFIC SIGNS	6
DISREGARDING ROAD MARKINGS	8
DISREGARDING STOP SIGN	10
DISREGARDING TRAFFIC SIGNALS	8
DISREGARDING YIELD SIGN	4
DISTRACTION - FROM INSIDE VEHICLE	8
DISTRACTION - FROM OUTSIDE VEHICLE	8
DISTRACTION - OTHER ELECTRONIC DEVICE (NAVIGATION DEVICE; DVD PLAYER; ETC.)	6
DRIVING ON WRONG SIDE/WRONG WAY	10
DRIVING SKILLS/KNOWLEDGE/EXPERIENCE	7
EQUIPMENT - VEHICLE CONDITION	10
EVASIVE ACTION DUE TO ANIMAL; OBJECT; NONMOTORIST	7
EXCEEDING AUTHORIZED SPEED LIMIT	18
EXCEEDING SAFE SPEED FOR CONDITIONS	6
FAILING TO REDUCE SPEED TO AVOID CRASH	10
FAILING TO YIELD RIGHT-OF-WAY	10
FOLLOWING TOO CLOSELY	9
HAD BEEN DRINKING (USE WHEN ARREST IS NOT MADE)	15
IMPROPER BACKING	7
IMPROPER LANE USAGE	9
IMPROPER OVERTAKING/PASSING	9
IMPROPER TURNING/NO SIGNAL	8
MOTORCYCLE ADVANCING LEGALLY ON RED LIGHT	4
NOT APPLICABLE	10
OBSTRUCTED CROSSWALKS	3
OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	18
PASSING STOPPED SCHOOL BUS	3

3.6.8.3 Excel

Trong **PivotTable Fields**, ta chọn thuộc tính là “PRIM CONTRIBUTORY CAUSE” và “Max Number of Units”.


PivotTable Fields
Show fields:
(All)
Search
Injuries Non Incapacitating
Injuries Total
Max Number of Units
Number of Traffic Crash Records
Values
Drag fields between areas below:
Filters
Columns
Rows
Values
PRIM CONTRIBUTORY ...
Max Number of Units
Defer Layout Update
Update

Kết quả:

PRIMARY CONTRIBUTORY CAUSE	Max Number of Units
ANIMAL	7
BICYCLE ADVANCING LEGALLY ON RED LIGHT	3
CELL PHONE USE OTHER THAN TEXTING	10
DISREGARDING OTHER TRAFFIC SIGNS	6
DISREGARDING ROAD MARKINGS	8
DISREGARDING STOP SIGN	10
DISREGARDING TRAFFIC SIGNALS	8
DISREGARDING YIELD SIGN	4
DISTRACTION - FROM INSIDE VEHICLE	8
DISTRACTION - FROM OUTSIDE VEHICLE	8
DISTRACTION - OTHER ELECTRONIC DEVICE (NAVIGATION DEVICE; DVD PLAYER; ETC.)	6
DRIVING ON WRONG SIDE/WRONG WAY	10
DRIVING SKILLS/KNOWLEDGE/EXPERIENCE	7
EQUIPMENT - VEHICLE CONDITION	10
EVASIVE ACTION DUE TO ANIMAL; OBJECT; NONMOTORIST	7
EXCEEDING AUTHORIZED SPEED LIMIT	18
EXCEEDING SAFE SPEED FOR CONDITIONS	6
FAILING TO REDUCE SPEED TO AVOID CRASH	10
FAILING TO YIELD RIGHT-OF-WAY	10
FOLLOWING TOO CLOSELY	9
HAD BEEN DRINKING (USE WHEN ARREST IS NOT MADE)	15
IMPROPER BACKING	7
IMPROPER LANE USAGE	9
IMPROPER OVERTAKING/PASSING	9
IMPROPER TURNING/NO SIGNAL	8
MOTORCYCLE ADVANCING LEGALLY ON RED LIGHT	4
NOT APPLICABLE	10
OBSTRUCTED CROSSWALKS	3
OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	18
PASSING STOPPED SCHOOL BUS	3
PHYSICAL CONDITION OF DRIVER	12
RELATED TO BUS STOP	4
ROAD CONSTRUCTION/MAINTENANCE	5
ROAD ENGINEERING/SURFACE/MARKING DEFECTS	5
TEXTING	9
TURNING RIGHT ON RED	6
UNABLE TO DETERMINE	18
UNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED)	18
VISION OBSCURED (SIGNS; TREE LIMBS; BUILDINGS; ETC.)	6
WEATHER	12
Grand Total	**18**

3.6.8.4 Nhận xét

Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.9 Câu truy vấn 9

Nội dung câu truy vấn: Thống kê tổng số nạn nhân, số vụ tai nạn theo từng thiết bị điều khiển (TRAFFIC CONTROL DEVICE) và từng nguyên nhân (PRIM CONTRIBUTORY CAUSE bắt đầu bằng ký tự 'F').

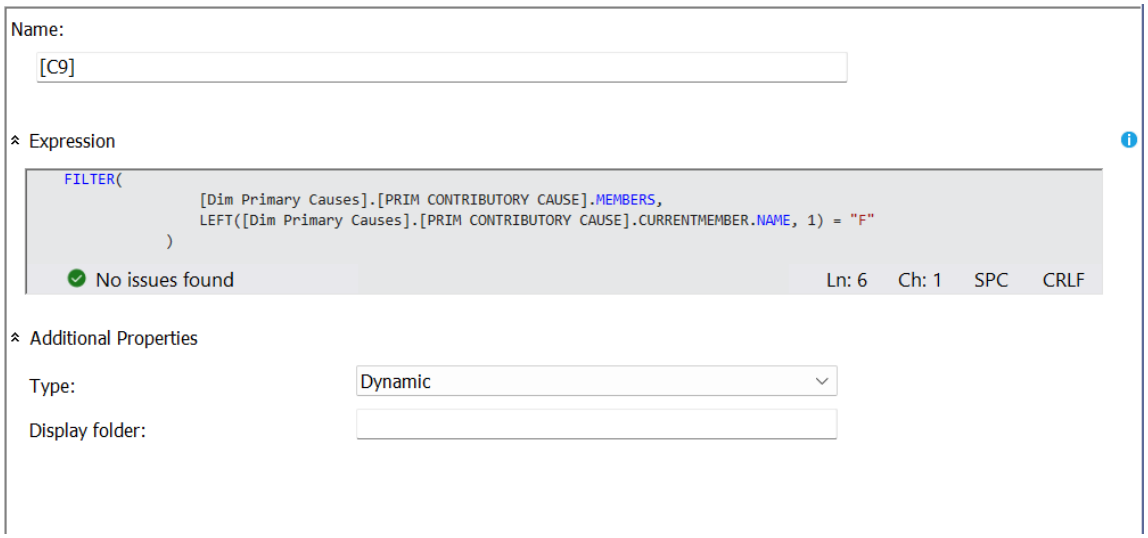
3.6.9.1 Manual

Bước 1: Vào mục Calculations, tạo mới một “New Named Set” đặt tên là “C9”.

Sau đó nhập câu truy vấn và tiến hành Deloy.

Câu truy vấn:

```
            FILTER(
                [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].MEMBERS,
                LEFT([Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].CURRENTMEMBER.NAME,
1) = "F"
            )
```



Bước 2: Vào mục **Browser** chọn từ Dim Primary Cause thuộc tính "PRIM CONTRIBUTORY CAUSE". Tại cột **Operator** chọn "In" và **Filter Expression** chọn vào Named Set "C9" vừa tạo. Sau đó kéo các thuộc tính và độ đo để hiển thị kết quả câu truy vấn và nhấn "Click to execute the query". Ta được kết quả.

Dimension	Hierarchy	Operator	Filter Expression	Parame...
Dim Primary Causes	PRIM CONTRIBUTOR...	In	C9	
<Select dimension>				

TRAFFIC CONTROL DE...	PRIM CONTRIBUTORY CAUSE	Injuries Total	Number of Traffic Crash Records
BICYCLE CROSSING SI...	FAILING TO REDUCE SPEED TO AVOID CRASH	0	1
BICYCLE CROSSING SI...	FAILING TO YIELD RIGHT-OF-WAY	6	8
DELINEATORS	FAILING TO REDUCE SPEED TO AVOID CRASH	2	9
DELINEATORS	FAILING TO YIELD RIGHT-OF-WAY	9	32
DELINEATORS	FOLLOWING TOO CLOSELY	0	15
FLASHING CONTROL ...	FAILING TO REDUCE SPEED TO AVOID CRASH	19	25
FLASHING CONTROL ...	FAILING TO YIELD RIGHT-OF-WAY	15	34
FLASHING CONTROL ...	FOLLOWING TOO CLOSELY	9	31
LANE USE MARKING	FAILING TO REDUCE SPEED TO AVOID CRASH	21	48
LANE USE MARKING	FAILING TO YIELD RIGHT-OF-WAY	40	129
LANE USE MARKING	FOLLOWING TOO CLOSELY	23	96
NO CONTROLS	FAILING TO REDUCE SPEED TO AVOID CRASH	4785	16023
NO CONTROLS	FAILING TO YIELD RIGHT-OF-WAY	8558	40118
NO CONTROLS	FOLLOWING TOO CLOSELY	5290	32085
NO PASSING	FAILING TO REDUCE SPEED TO AVOID CRASH	0	2
NO PASSING	FAILING TO YIELD RIGHT-OF-WAY	1	1
OTHER	FAILING TO REDUCE SPEED TO AVOID CRASH	70	185
OTHER	FAILING TO YIELD RIGHT-OF-WAY	158	448
OTHER	FOLLOWING TOO CLOSELY	65	498
OTHER RAILROAD CR...	FAILING TO REDUCE SPEED TO AVOID CRASH	1	9
OTHER RAILROAD CR...	FAILING TO YIELD RIGHT-OF-WAY	1	13
OTHER RAILROAD CR...	FOLLOWING TOO CLOSELY	10	42
OTHER REG. SIGN	FAILING TO REDUCE SPEED TO AVOID CRASH	21	39
OTHER REG. SIGN	FAILING TO YIELD RIGHT-OF-WAY	32	145
OTHER REG. SIGN	FOLLOWING TOO CLOSELY	10	63

3.6.9.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
SELECT
    {
        [Measures].[Injuries Total],
        [Measures].[Number of Traffic Crash Records]
    } ON COLUMNS,

    NONEMPTY(
      CROSSJOIN(
          [Dim Traffic Control].[TRAFFIC CONTROL DEVICE].MEMBERS,
          FILTER(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].MEMBERS,
            LEFT([Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].CURRENTMEMBER.NAME, 1) = "F"
            )
        )
    ) ON ROWS

FROM
    [Traffic Crash];
```

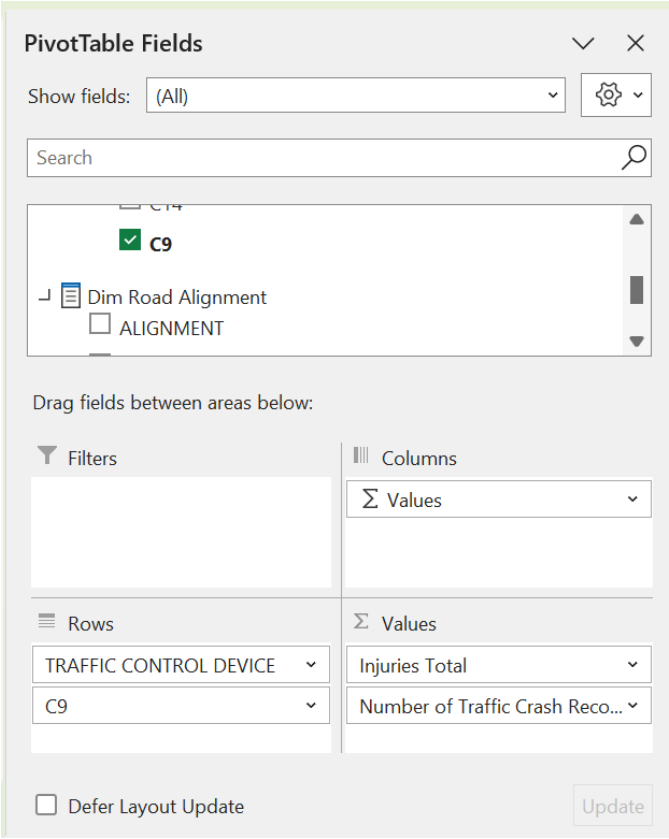
Kết quả câu truy vấn MDX:

Messages | Results

		Injuries Total	Number of Traffic Crash Records
BICYCLE CROSSING SIGN	FAILING TO REDUCE SPEED TO AVOID CRA...	0	1
BICYCLE CROSSING SIGN	FAILING TO YIELD RIGHT-OF-WAY	6	8
DELINEATORS	FAILING TO REDUCE SPEED TO AVOID CRA...	2	9
DELINEATORS	FAILING TO YIELD RIGHT-OF-WAY	9	32
DELINEATORS	FOLLOWING TOO CLOSELY	0	15
FLASHING CONTROL SIGNAL	FAILING TO REDUCE SPEED TO AVOID CRA...	19	25
FLASHING CONTROL SIGNAL	FAILING TO YIELD RIGHT-OF-WAY	15	34
FLASHING CONTROL SIGNAL	FOLLOWING TOO CLOSELY	9	31
LANE USE MARKING	FAILING TO REDUCE SPEED TO AVOID CRA...	21	48
LANE USE MARKING	FAILING TO YIELD RIGHT-OF-WAY	40	129
LANE USE MARKING	FOLLOWING TOO CLOSELY	23	96
NO CONTROLS	FAILING TO REDUCE SPEED TO AVOID CRA...	4785	16023
NO CONTROLS	FAILING TO YIELD RIGHT-OF-WAY	8558	40118
NO CONTROLS	FOLLOWING TOO CLOSELY	5290	32085
NO PASSING	FAILING TO REDUCE SPEED TO AVOID CRA...	0	2
NO PASSING	FAILING TO YIELD RIGHT-OF-WAY	1	1
OTHER	FAILING TO REDUCE SPEED TO AVOID CRA...	70	185
OTHER	FAILING TO YIELD RIGHT-OF-WAY	158	448
OTHER	FOLLOWING TOO CLOSELY	65	498
OTHER RAILROAD CROSSING	FAILING TO REDUCE SPEED TO AVOID CRA...	1	9
OTHER RAILROAD CROSSING	FAILING TO YIELD RIGHT-OF-WAY	1	13
OTHER RAILROAD CROSSING	FOLLOWING TOO CLOSELY	10	42
OTHER REG. SIGN	FAILING TO REDUCE SPEED TO AVOID CRA...	21	39
OTHER REG. SIGN	FAILING TO YIELD RIGHT-OF-WAY	32	145
OTHER REG. SIGN	FOLLOWING TOO CLOSELY	10	63
OTHER WARNING SIGN	FAILING TO REDUCE SPEED TO AVOID CRA...	16	38
OTHER WARNING SIGN	FAILING TO YIELD RIGHT-OF-WAY	12	55
OTHER WARNING SIGN	FOLLOWING TOO CLOSELY	6	51
PEDESTRIAN CROSSING SIGN	FAILING TO REDUCE SPEED TO AVOID CRA...	19	41
PEDESTRIAN CROSSING SIGN	FAILING TO YIELD RIGHT-OF-WAY	133	139

3.6.9.3 Excel

Trong **PivotTable Fields**, ta chọn thuộc tính là "TRAFFIC CONTROL DEVICE" và Named Set "C9" đã tạo cùng với độ đo "Injuries Total" và "Number of Traffic Crash Records".



Kết quả:

Row Labels	Injuries Total	Number of Traffic Crash Records
BICYCLE CROSSING SIGN		
FAILING TO REDUCE SPEED TO AVOID CRASH	0	1
FAILING TO YIELD RIGHT-OF-WAY	6	8
DELINEATORS		
FAILING TO REDUCE SPEED TO AVOID CRASH	2	9
FAILING TO YIELD RIGHT-OF-WAY	9	32
FOLLOWING TOO CLOSELY	0	15
FLASHING CONTROL SIGNAL		
FAILING TO REDUCE SPEED TO AVOID CRASH	19	25
FAILING TO YIELD RIGHT-OF-WAY	15	34
FOLLOWING TOO CLOSELY	9	31
LANE USE MARKING		
FAILING TO REDUCE SPEED TO AVOID CRASH	21	48
FAILING TO YIELD RIGHT-OF-WAY	40	129
FOLLOWING TOO CLOSELY	23	96
NO CONTROLS		
FAILING TO REDUCE SPEED TO AVOID CRASH	4785	16023
FAILING TO YIELD RIGHT-OF-WAY	8558	40118
FOLLOWING TOO CLOSELY	5290	32085
NO PASSING		
FAILING TO REDUCE SPEED TO AVOID CRASH	0	2
FAILING TO YIELD RIGHT-OF-WAY	1	1
OTHER		
FAILING TO REDUCE SPEED TO AVOID CRASH	70	185
FAILING TO YIELD RIGHT-OF-WAY	158	448
FOLLOWING TOO CLOSELY	65	498
OTHER RAILROAD CROSSING		
FAILING TO REDUCE SPEED TO AVOID CRASH	1	9
FAILING TO YIELD RIGHT-OF-WAY	1	13
FOLLOWING TOO CLOSELY	10	42
OTHER REG. SIGN		
FAILING TO REDUCE SPEED TO AVOID CRASH	21	39
FAILING TO YIELD RIGHT-OF-WAY	32	145
FOLLOWING TOO CLOSELY	10	63

3.6.9.4 Nhận xét

Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.10 Câu truy vấn 10

Nội dung câu truy vấn: Thống kê tổng số lượng vụ tai nạn và số lượng tử vong ở mỗi vụ tai nạn qua từng tháng và quý trong năm (từ 2016 đến 2023) do nguyên nhân 'ANIMAL'.

3.6.10.1 Manual

Vào mục **Browser**, kéo từ bảng Dim Primary Cause thuộc tính "PRIM CONTRIBUTORY CAUSE". Tại cột **Operator** chọn "Equal" và **Filter Expression** chọn vào "ANIMAL". Tiếp theo ta kéo từ bảng Dim Time Hierarchy "PHANCAP_TIME" đã tạo và ở **Filter Expression** ta chọn các năm từ 2016 đến 2023.Sau đó kéo các thuộc tính để hiển thị kết quả câu truy vấn và nhấn "Click to execute the query".

Ta được kết quả:

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim Primary Causes	PRIM CONTRIBUTORY ...	Equal	{ ANIMAL }	
Dim Time	PHANCAP_TIME	Equal	{ 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023 }	
<Select dimension>				

CRASH YEAR	CRASH QUARTER	CRASH MONTH	Number of Traffic Crash R...	Injuries Total	Injuries Fatal
2016	1	1	2	0	0
2016	1	2	1	0	0
2016	1	3	4	0	0
2016	2	4	3	0	0
2016	2	5	1	0	0
2016	2	6	1	0	0
2016	3	7	1	0	0
2016	3	8	1	0	0
2016	3	9	1	1	0
2016	4	10	4	0	0
2016	4	11	3	1	0
2016	4	12	3	2	0
2017	1	1	4	1	0
2017	1	2	3	1	0
2017	1	3	4	0	0
2017	2	4	2	2	0
2017	2	5	4	0	0
2017	2	6	8	0	0
2017	3	7	4	0	0
2017	3	8	6	0	0
2017	3	9	11	2	0
2017	4	10	6	0	0
2017	4	11	12	4	0
2017	4	12	7	1	0
2018	1	1	6	0	0

3.6.10.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
SELECT
    NON EMPTY {
        [Measures].[Number of Traffic Crash Records],
        [Measures].[Injuries Total],
              [Measures].[Injuries Fatal]
    } ON COLUMNS,

    NON EMPTY {
        DrillDownLevel(
            DrillDownLevel(
                DrillDownLevel(
                    [Dim Time].[PHANCAP_TIME]
                )
            )
        )
    } ON ROWS
FROM
    [Traffic Crash]
WHERE {
       [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].&[ANIMAL]
       };
```

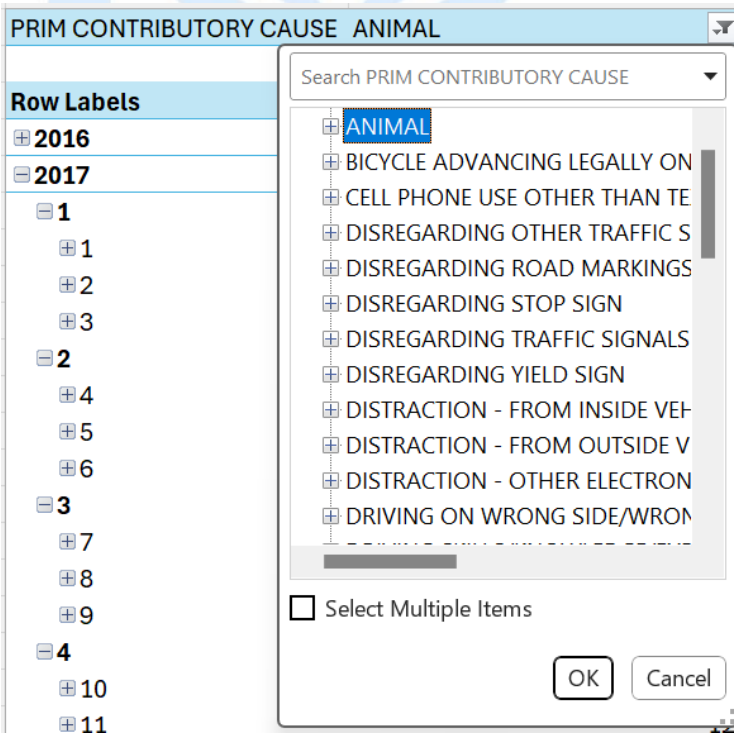
Kết quả câu truy vấn MDX:

Messages | Results

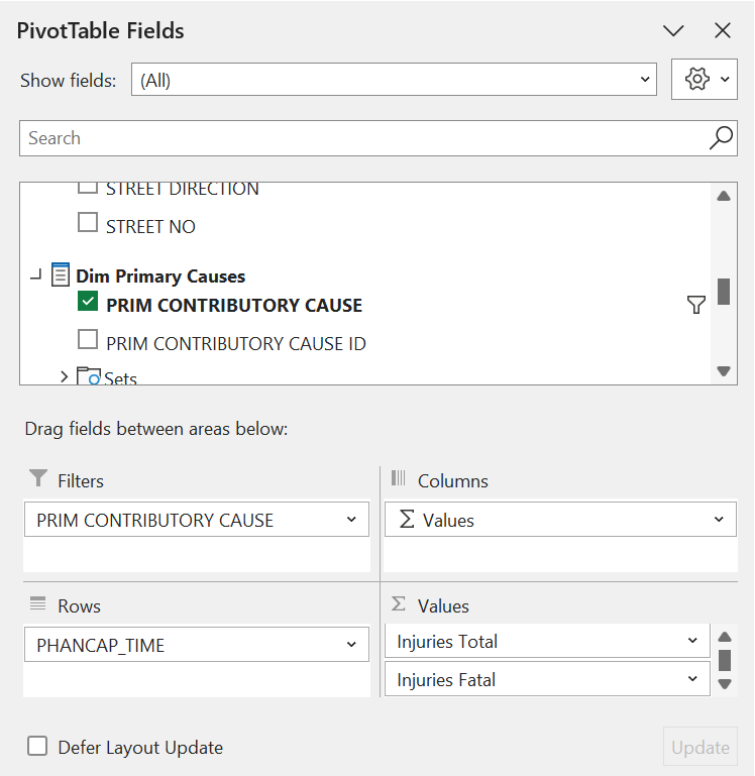
	Number of Traffic Crash Records	Injuries Total	Injuries Fatal
All	668	96	1
2016	25	4	0
1	7	0	0
1	2	0	0
2	1	0	0
3	4	0	0
2	5	0	0
4	3	0	0
5	1	0	0
6	1	0	0
3	3	1	0
7	1	0	0
8	1	0	0
9	1	1	0
4	10	3	0
10	4	0	0
11	3	1	0
12	3	2	0
2017	71	11	0
1	11	2	0
1	4	1	0
2	3	1	0
3	4	0	0
2	14	2	0
4	2	2	0
5	4	0	0
6	8	0	0
3	21	2	0
7	4	0	0
8	6	0	0

3.6.10.3 Excel

Bước 1: Trong **PivotTable Fields**, ta chọn **Filters** là thuộc tính là "PRIM CONTRIBUTORY CAUSE" có giá trị là "ANIMAL".



Bước 2: Nhấn vào biểu tượng cạnh Row Labels chọn **Value Filters**.



Kết quả:

PRIM CONTRIBUTORY CAUSE	ANIMAL		
Row Labels	**Number of Traffic Crash Records**	**Injuries Total**	**Injuries Fatal**
2016	**25**	**4**	**0**
2017			
1			
1	4	1	0
2	3	1	0
3	4	0	0
2			
4	2	2	0
5	4	0	0
6	8	0	0
3			
7	4	0	0
8	6	0	0
9	11	2	0
4			
10	6	0	0
11	12	4	0
12	7	1	0
2018			
1	**26**	**4**	**0**
2	**29**	**8**	**0**

3.6.10.4 Nhận xét

Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.11 Câu truy vấn 11

Nội dung câu truy vấn: Thống kê các nguyên nhân gây tai nạn ở mỗi loại đường (TRAFFICWAY TYPE) với số lượng tử vong lớn hơn hoặc bằng 6.

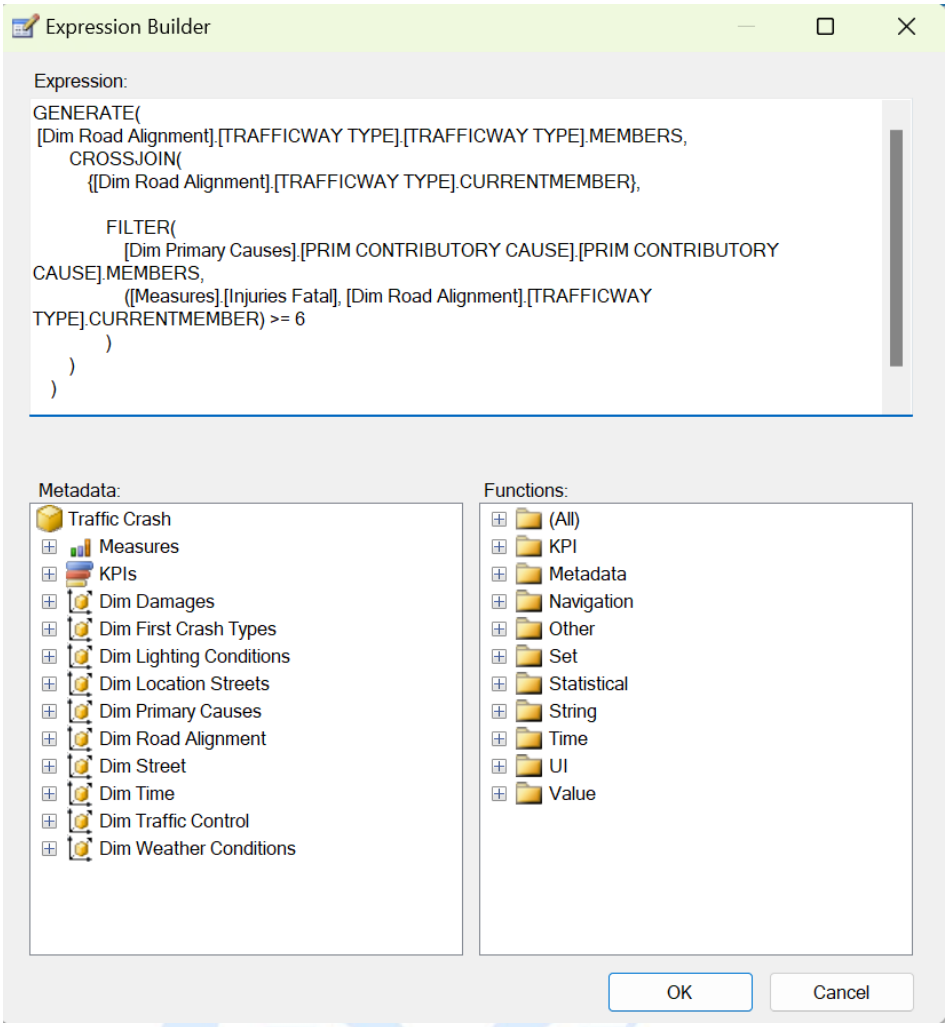
3.6.11.1 Manual

Bước 1: Vào mục **Browser**, chọn chọn thuộc tính "TRAFFICWAY TYPE" của Dim Road Alignment. Trong **Operator** chọn **Custom**. Tiếp theo ta nhập biểu thức.

Biểu thức:

```
GENERATE(
 [Dim Road Alignment].[TRAFFICWAY TYPE].[TRAFFICWAY TYPE].MEMBERS,
      CROSSJOIN(
         {[Dim Road Alignment].[TRAFFICWAY TYPE].CURRENTMEMBER},


            FILTER(
               [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].
[PRIM CONTRIBUTORY CAUSE].MEMBERS,
               ([Measures].[Injuries Fatal], [Dim Road Alignment].[TRAFFICWAY TYPE].
CURRENTMEMBER) >= 6
            )
      )
   )
```

Sau đó nhấn "OK".



Bước 2: Tiếp theo, ta kéo từ bảng Dim Road Alignment thuộc tính "TRAFFICWAY TYPE" và từ bảng Dim Primary Causes thuộc tính "PRIM CONTRIBUTORY CAUSE", cùng các độ đo cần thiết. Sau đó nhấn "Click to execute the query".

Ta được kết quả.

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim Road Alignment	TRAFFICWAY TYPE	Custom	GENERATE([Dim Road Alignment].[TRAFFICWAY TYPE].[TRAFFI...	
<Select dimension>				

TRAFFICWAY TYPE	PRIM CONTRIBUTORY CAUSE	Injuries Fatal
ALLEY	UNABLE TO DETERMINE	6
DIVIDED - W/MEDIAN (NOT RAISED)	DISREGARDING TRAFFIC SIGNALS	17
DIVIDED - W/MEDIAN (NOT RAISED)	EXCEEDING AUTHORIZED SPEED LIMIT	8
DIVIDED - W/MEDIAN (NOT RAISED)	FAILING TO REDUCE SPEED TO AVOID CRASH	16
DIVIDED - W/MEDIAN (NOT RAISED)	FAILING TO YIELD RIGHT-OF-WAY	23
DIVIDED - W/MEDIAN (NOT RAISED)	IMPROPER OVERTAKING/PASSING	6
DIVIDED - W/MEDIAN (NOT RAISED)	NOT APPLICABLE	17
DIVIDED - W/MEDIAN (NOT RAISED)	OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	10
DIVIDED - W/MEDIAN (NOT RAISED)	PHYSICAL CONDITION OF DRIVER	8
DIVIDED - W/MEDIAN (NOT RAISED)	UNABLE TO DETERMINE	52
DIVIDED - W/MEDIAN BARRIER	DRIVING ON WRONG SIDE/WRONG WAY	6
DIVIDED - W/MEDIAN BARRIER	FAILING TO REDUCE SPEED TO AVOID CRASH	14
DIVIDED - W/MEDIAN BARRIER	NOT APPLICABLE	8
DIVIDED - W/MEDIAN BARRIER	OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	9
DIVIDED - W/MEDIAN BARRIER	UNABLE TO DETERMINE	55
FOUR WAY	DISREGARDING STOP SIGN	6
FOUR WAY	DISREGARDING TRAFFIC SIGNALS	31
FOUR WAY	FAILING TO YIELD RIGHT-OF-WAY	18
FOUR WAY	UNABLE TO DETERMINE	22
NOT DIVIDED	DISREGARDING TRAFFIC SIGNALS	17
NOT DIVIDED	DISTRACTION - FROM OUTSIDE VEHICLE	8
NOT DIVIDED	DRIVING ON WRONG SIDE/WRONG WAY	14
NOT DIVIDED	EXCEEDING AUTHORIZED SPEED LIMIT	10
NOT DIVIDED	FAILING TO REDUCE SPEED TO AVOID CRASH	27
NOT DIVIDED	FAILING TO YIELD RIGHT-OF-WAY	22

3.6.11.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
WITH
SET [TopCausesByTrafficWay] AS
    GENERATE(
        [Dim Road Alignment].[TRAFFICWAY TYPE].[TRAFFICWAY TYPE].MEMBERS,
        CROSSJOIN(
            {[Dim Road Alignment].[TRAFFICWAY TYPE].CURRENTMEMBER},

                FILTER(
                    [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].
[PRIM CONTRIBUTORY CAUSE].MEMBERS,
                    ([Measures].[Injuries Fatal], [Dim Road Alignment].
[TRAFFICWAY TYPE].CURRENTMEMBER) >= 6
                )
        )
    )
SELECT
    NON EMPTY {
        [Measures].[Number of Traffic Crash Records],
        [Measures].[Injuries Fatal]
    } ON COLUMNS,

    NON EMPTY {
        [TopCausesByTrafficWay]
    } ON ROWS

FROM
    [Traffic Crash];
```

Kết quả câu truy vấn MDX:

Messages | Results

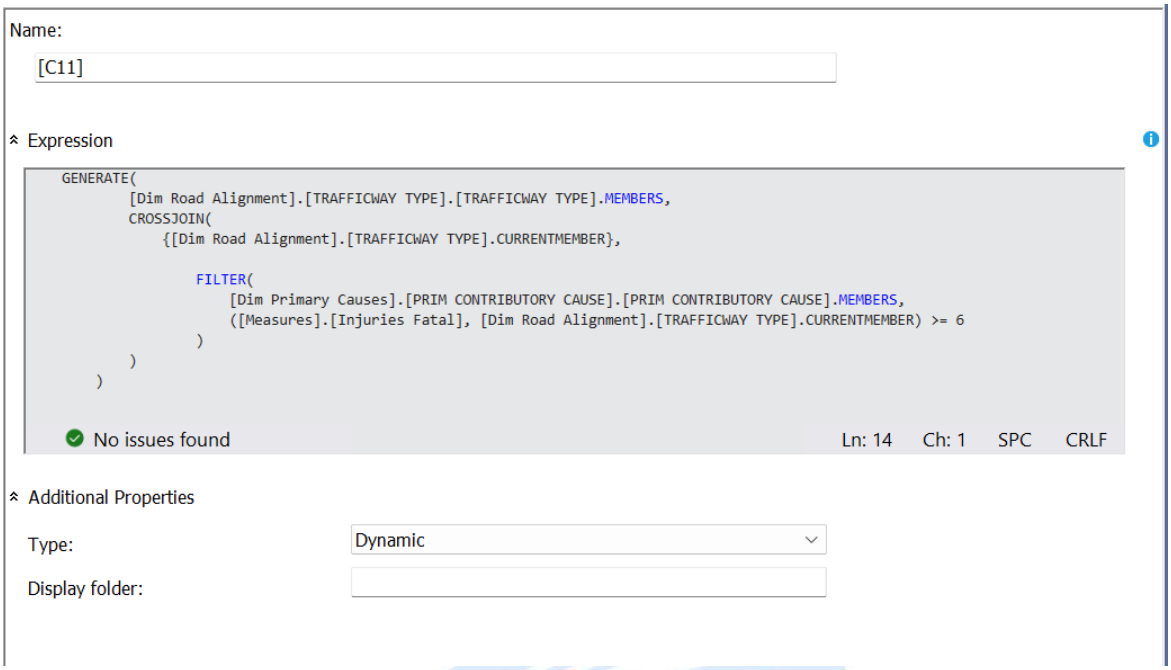
		Number of Traffic Crash Records	Injuries Fatal
ALLEY	UNABLE TO DETERMINE	5680	6
DIVIDED - W/MEDIAN (NOT RAISED)	DISREGARDING TRAFFIC SIGNALS	2370	17
DIVIDED - W/MEDIAN (NOT RAISED)	EXCEEDING AUTHORIZED SPEED LIMIT	325	8
DIVIDED - W/MEDIAN (NOT RAISED)	FAILING TO REDUCE SPEED TO AVOID CRASH	6634	16
DIVIDED - W/MEDIAN (NOT RAISED)	FAILING TO YIELD RIGHT-OF-WAY	15578	23
DIVIDED - W/MEDIAN (NOT RAISED)	IMPROPER OVERTAKING/PASSING	7391	6
DIVIDED - W/MEDIAN (NOT RAISED)	NOT APPLICABLE	4277	17
DIVIDED - W/MEDIAN (NOT RAISED)	OPERATING VEHICLE IN ERRATIC; RECKLESS; ...	1249	10
DIVIDED - W/MEDIAN (NOT RAISED)	PHYSICAL CONDITION OF DRIVER	865	8
DIVIDED - W/MEDIAN (NOT RAISED)	UNABLE TO DETERMINE	43610	52
DIVIDED - W/MEDIAN BARRIER	DRIVING ON WRONG SIDE/WRONG WAY	206	6
DIVIDED - W/MEDIAN BARRIER	FAILING TO REDUCE SPEED TO AVOID CRASH	3103	14
DIVIDED - W/MEDIAN BARRIER	NOT APPLICABLE	1856	8
DIVIDED - W/MEDIAN BARRIER	OPERATING VEHICLE IN ERRATIC; RECKLESS; ...	683	9
DIVIDED - W/MEDIAN BARRIER	UNABLE TO DETERMINE	13614	55
FOUR WAY	DISREGARDING STOP SIGN	2364	6
FOUR WAY	DISREGARDING TRAFFIC SIGNALS	5137	31
FOUR WAY	FAILING TO YIELD RIGHT-OF-WAY	10107	18
FOUR WAY	UNABLE TO DETERMINE	13573	22
NOT DIVIDED	DISREGARDING TRAFFIC SIGNALS	4640	17
NOT DIVIDED	DISTRACTION - FROM OUTSIDE VEHICLE	1358	8
NOT DIVIDED	DRIVING ON WRONG SIDE/WRONG WAY	1502	14
NOT DIVIDED	EXCEEDING AUTHORIZED SPEED LIMIT	911	10
NOT DIVIDED	FAILING TO REDUCE SPEED TO AVOID CRASH	14438	27
NOT DIVIDED	FAILING TO YIELD RIGHT-OF-WAY	37940	22
NOT DIVIDED	IMPROPER OVERTAKING/PASSING	18612	6
NOT DIVIDED	NOT APPLICABLE	19442	26
NOT DIVIDED	OPERATING VEHICLE IN ERRATIC; RECKLESS; ...	3851	16
NOT DIVIDED	PHYSICAL CONDITION OF DRIVER	2144	31
NOT DIVIDED	UNABLE TO DETERMINE	132198	134

3.6.11.3 Excel

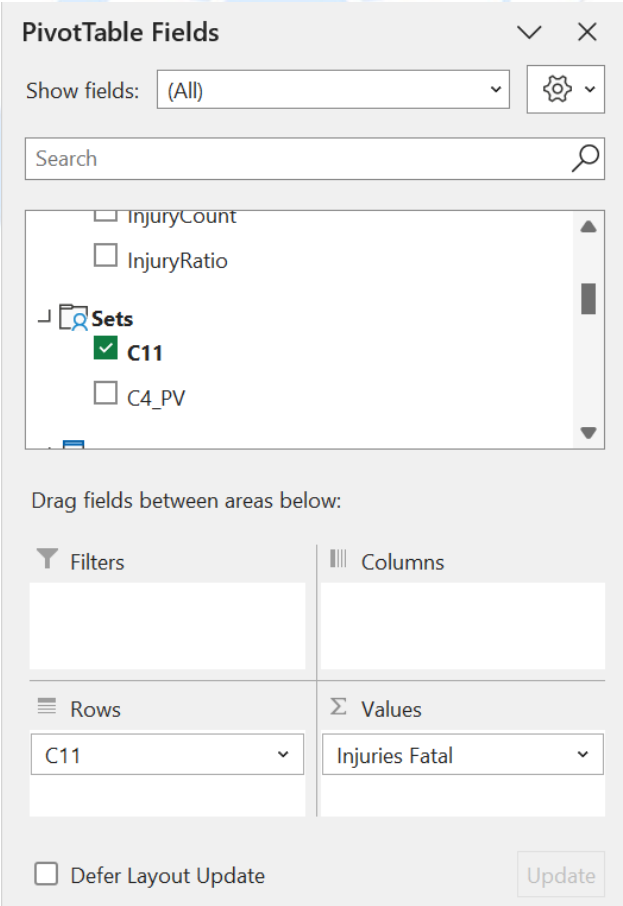
Bước 1: Tạo Named Set đặt tên là "C11" cho **PivotTable Fields**.

Câu lệnh:

```
GENERATE(
        [Dim Road Alignment].[TRAFFICWAY TYPE].[TRAFFICWAY TYPE].MEMBERS,
        CROSSJOIN(
            {[Dim Road Alignment].[TRAFFICWAY TYPE].CURRENTMEMBER},

                FILTER(
                    [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM
CONTRIBUTORY CAUSE].MEMBERS,
                    ([Measures].[Injuries Fatal], [Dim Road
Alignment].[TRAFFICWAY TYPE].CURRENTMEMBER) >= 6
                )
        )
    )
```



Bước 2: Trong **Sets**, ta chọn Named Set "C11" đã tạo. Tiếp theo, ta chọn độ đo cần thiết.



Kết quả:

Row Labels	Injuries Fatal
ALLEY	
UNABLE TO DETERMINE	6
DIVIDED - W/MEDIAN (NOT RAISED)	
DISREGARDING TRAFFIC SIGNALS	17
EXCEEDING AUTHORIZED SPEED LIMIT	8
FAILING TO REDUCE SPEED TO AVOID CRASH	16
FAILING TO YIELD RIGHT-OF-WAY	23
IMPROPER OVERTAKING/PASSING	6
NOT APPLICABLE	17
OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	10
PHYSICAL CONDITION OF DRIVER	8
UNABLE TO DETERMINE	52
DIVIDED - W/MEDIAN BARRIER	
DRIVING ON WRONG SIDE/WRONG WAY	6
FAILING TO REDUCE SPEED TO AVOID CRASH	14
NOT APPLICABLE	8
OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	9
UNABLE TO DETERMINE	55
FOUR WAY	
DISREGARDING STOP SIGN	6
DISREGARDING TRAFFIC SIGNALS	31
FAILING TO YIELD RIGHT-OF-WAY	18
UNABLE TO DETERMINE	22
NOT DIVIDED	
DISREGARDING TRAFFIC SIGNALS	17
DISTRACTION - FROM OUTSIDE VEHICLE	8
DRIVING ON WRONG SIDE/WRONG WAY	14
EXCEEDING AUTHORIZED SPEED LIMIT	10
FAILING TO REDUCE SPEED TO AVOID CRASH	27
FAILING TO YIELD RIGHT-OF-WAY	22
IMPROPER OVERTAKING/PASSING	6
NOT APPLICABLE	26
OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	16
PHYSICAL CONDITION OF DRIVER	31
UNABLE TO DETERMINE	134
UNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED)	11
ONE-WAY	
FAILING TO REDUCE SPEED TO AVOID CRASH	6
NOT APPLICABLE	15
PHYSICAL CONDITION OF DRIVER	24
UNABLE TO DETERMINE	29
OTHER	
UNABLE TO DETERMINE	10

3.6.11.4 Nhận xét

Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.12 Câu truy vấn 12

Nội dung câu truy vấn: Liệt kê các loại đường giao thông (TRAFFICWAY TYPE) là top 5 loại đường xảy ra nhiều vụ tai nạn nhất trong cả hai năm 2022 và 2023.

3.6.12.1 Manual

Bước 1: Vào mục **Calculations**, tạo mới một "New Named Set" đặt tên "C12".

Sau đó nhập câu truy vấn và tiến hành Deloy.

Nội dung câu lệnh:

```
INTERSECT(
    TOPCOUNT(
[Dim Road Alignment].[TRAFFICWAY TYPE].[TRAFFICWAY TYPE].MEMBERS,
        5,
        ([Measures].[Number of Traffic Crash Records],
[Dim Time].[PHANCAP_Time].&[2022])
    ),

    TOPCOUNT(
        [Dim Road Alignment].[TRAFFICWAY TYPE].[TRAFFICWAY TYPE].MEMBERS,
        5,
        ([Measures].[Number of Traffic Crash Records],
[Dim Time].[PHANCAP_Time].&[2023])
    )
)
```

Bước 2: Vào mục **Browser**, kéo từ bảng "Dim Road Alignment" thuộc tính "TRAFFICWAY TYPE". Tại cột **Operator** chọn "In" và **Filter Expression** chọn vào Named Set "C12" vừa tạo.

Bước 3: Tiếp theo, kéo từ bảng "Dim Time" hierarchy "PHANCAP_Time" đã tạo. Tại cột **Operator** chọn "Equal" và **Filter Expression** chọn năm 2022 và 2023.

Sau đó kéo các thuộc tính để hiển thị kết quả câu truy vấn và nhấn "Click to execute the query". Ta được kết quả.

Dimension	Hierarchy	Operator	Filter Expression	Parame...
Dim Road Alignment	TRAFFICWAY TYPE	In	C12	
Dim Time	PHANCAP_TIME	Equal	{ 2022, 2023 }	
<Select dimension>				

TRAFFICWAY TYPE	Number of Traffic Crash...	Injuries Total	Injuries Fatal
DIVIDED - W/MEDIAN (NOT RAISED)	29041	6306	51
FOUR WAY	24843	10662	35
NOT DIVIDED	90987	16858	123
ONE-WAY	25703	2676	20
PARKING LOT	13909	740	2

3.6.12.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
WITH
SET [Top5CrashTrafficWays2022] AS
    TOPCOUNT(
        [Dim Road Alignment].[TRAFFICWAY TYPE].[TRAFFICWAY TYPE].MEMBERS,
        5,
        ([Measures].[Number of Traffic Crash Records],
[Dim Time].[PHANCAP_Time].&[2022])
    )

SET [Top5CrashTrafficWays2023] AS
    TOPCOUNT(
        [Dim Road Alignment].[TRAFFICWAY TYPE].[TRAFFICWAY TYPE].MEMBERS,
        5,
        ([Measures].[Number of Traffic Crash Records],
[Dim Time].[PHANCAP_Time].&[2023])
    )

SET [IntersectionTopTrafficWays] AS
    INTERSECT(
        [Top5CrashTrafficWays2022],
        [Top5CrashTrafficWays2023]
```

```
    )
```

```
-- Calculate totals for 2022 and 2023 for the intersection TRAFFICWAY TYPE
MEMBER [Measures].[Total Number of Traffic Crash Records] AS
    AGGREGATE(
        {[Dim Time].[PHANCAP_Time].&[2022], [Dim Time].[PHANCAP_Time].&[2023]},
        [Measures].[Number of Traffic Crash Records]
    )

MEMBER [Measures].[Total Injuries Fatal] AS
    AGGREGATE(
        {[Dim Time].[PHANCAP_Time].&[2022], [Dim Time].[PHANCAP_Time].&[2023]},
        [Measures].[Injuries Fatal]
    )

MEMBER [Measures].[Total Injuries] AS
    AGGREGATE(
        {[Dim Time].[PHANCAP_Time].&[2022], [Dim Time].[PHANCAP_Time].&[2023]},
        [Measures].[Injuries Total]
    )

SELECT
    NON EMPTY {
        [Measures].[Total Number of Traffic Crash Records],
              [Measures].[Total Injuries],
        [Measures].[Total Injuries Fatal]

    } ON COLUMNS,

    NON EMPTY {
        [IntersectionTopTrafficWays]
    } ON ROWS

FROM
    [Traffic Crash];
```

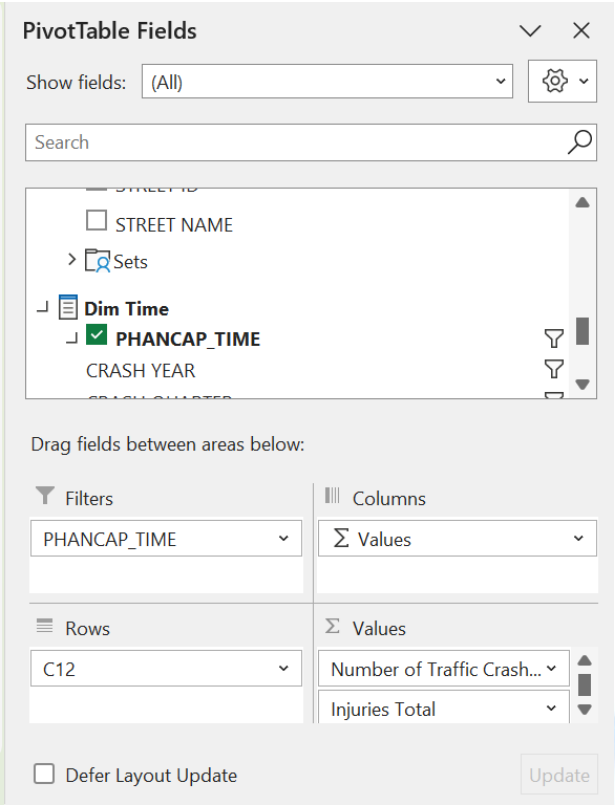
Kết quả câu truy vấn MDX:

Messages | Results

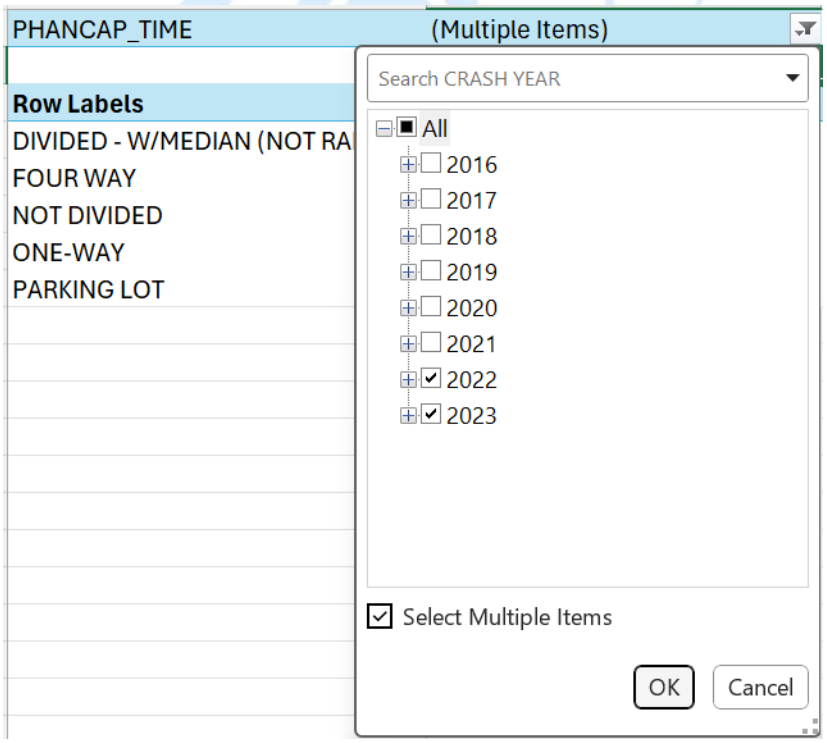
	Total Number of Traffic Crash Records	Total Injuries	Total Injuries Fatal
NOT DIVIDED	90987	16858	123
DIVIDED - W/MEDIAN (NOT RAISED)	29041	6306	51
ONE-WAY	25703	2676	20
FOUR WAY	24843	10662	35
PARKING LOT	13909	740	2

3.6.12.3 Excel

Bước 1: Trong **PivotTable Fields**, ta chọn thuộc tính là “TRAFFICWAY TYPE” và các thuộc tính độ đo cần thiết. Chọn “PHANCAP_Time” vào khu vực **Filters**.



Bước 2: Sau đó, chọn năm là 2022 và 2023.



Kết quả:

PHANCAP_TIME	(Multiple Items)		
Row Labels	**Number of Traffic Crash Records**	**Injuries Total**	**Injuries Fatal**
DIVIDED - W/MEDIAN (NOT RAISED)	29041	6306	51
FOUR WAY	24843	10662	35
NOT DIVIDED	90987	16858	123
ONE-WAY	25703	2676	20
PARKING LOT	13909	740	2

3.6.12.4 Nhận xét

Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.13 Câu truy vấn 13

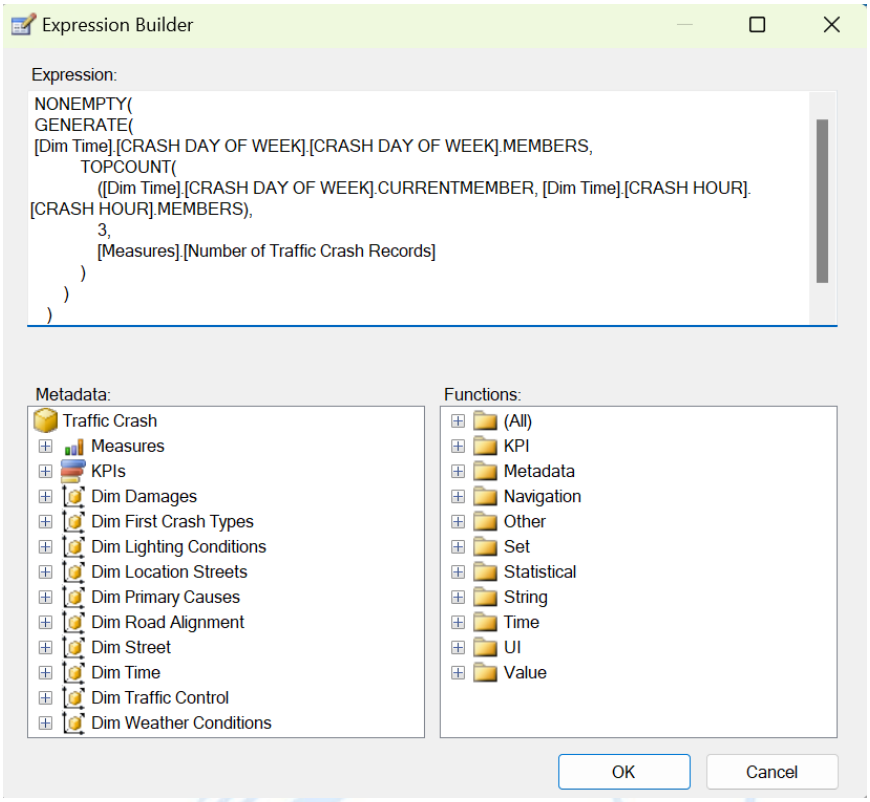
Nội dung câu truy vấn: Thống kê top 3 khung giờ xảy ra nhiều vụ tai nạn nhất của các ngày trong tuần.

3.6.13.1 Manual

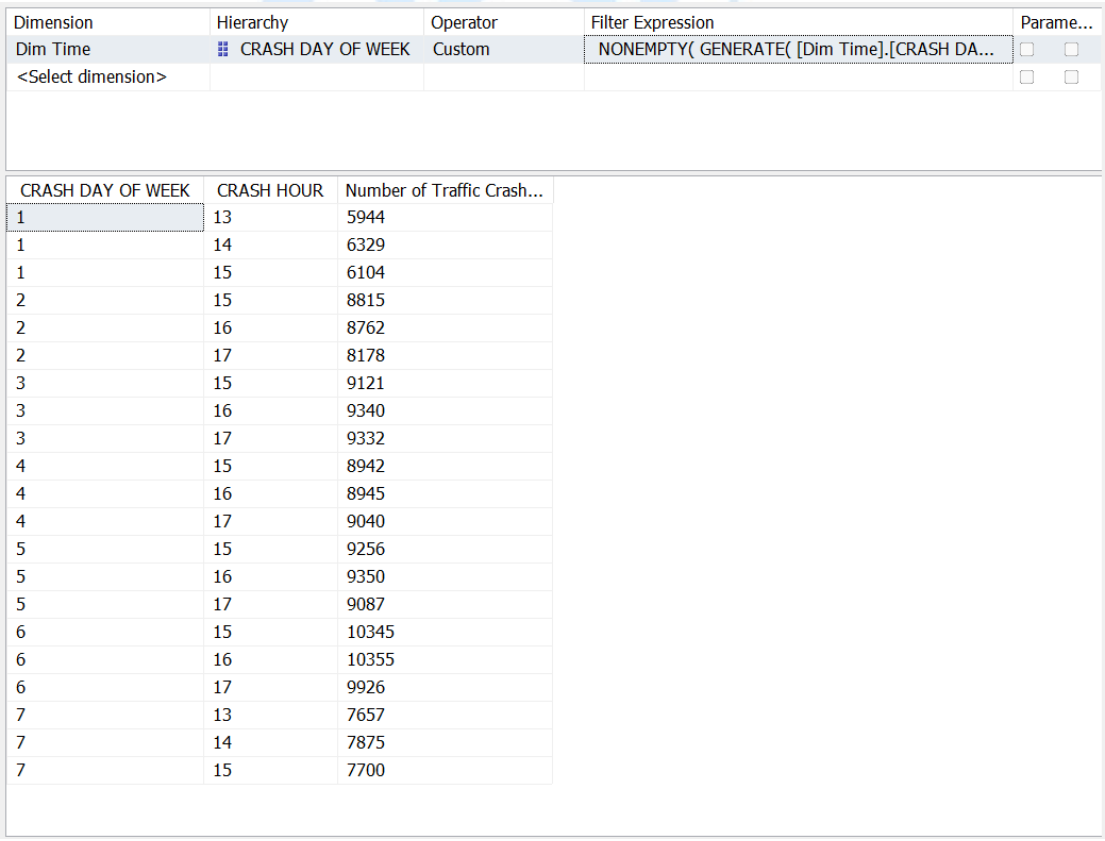
Bước 1: Vào mục **Browser**, kéo từ bảng "Dim Time" thuộc tính "CRASH DAY OF WEEK". Tại cột **Operator** chọn "Custom". Ở **Filter Expression** nhập vào câu lệnh truy vấn.

Nội dung câu lệnh:

```
NONEMPTY(
 GENERATE(
 [Dim Time].[CRASH DAY OF WEEK].[CRASH DAY OF WEEK].MEMBERS,
        TOPCOUNT(
            ([Dim Time].[CRASH DAY OF WEEK].CURRENTMEMBER, [Dim Time].[CRASH
HOUR].[CRASH HOUR].MEMBERS),
            3,
            [Measures].[Number of Traffic Crash Records]
        )
    )
  )
```



Bước 2: Sau đó, kéo các thuộc tính để hiển thị kết quả câu truy vấn và nhấn "Click to execute the query". Ta được kết quả.



Dimension	Hierarchy	Operator	Filter Expression	Parame...
Dim Time	CRASH DAY OF WEEK	Custom	NONEMPTY(GENERATE([Dim Time].[CRASH DA...	
<Select dimension>				

CRASH DAY OF WEEK	CRASH HOUR	Number of Traffic Crash...
1	13	5944
1	14	6329
1	15	6104
2	15	8815
2	16	8762
2	17	8178
3	15	9121
3	16	9340
3	17	9332
4	15	8942
4	16	8945
4	17	9040
5	15	9256
5	16	9350
5	17	9087
6	15	10345
6	16	10355
6	17	9926
7	13	7657
7	14	7875
7	15	7700

3.6.13.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
SELECT
    {[Measures].[Number of Traffic Crash Records]} ON COLUMNS,

    NONEMPTY(
        GENERATE(
            [Dim Time].[CRASH DAY OF WEEK].[CRASH DAY OF WEEK].MEMBERS,
            TOPCOUNT(
                ([Dim Time].[CRASH DAY OF WEEK].CURRENTMEMBER, [Dim Time].[CRASH
HOUR].[CRASH HOUR].MEMBERS),
                3,
                [Measures].[Number of Traffic Crash Records]
            )
        )
    ) ON ROWS

FROM
    [Traffic Crash];
```

Kết quả câu truy vấn MDX:

Messages | Results

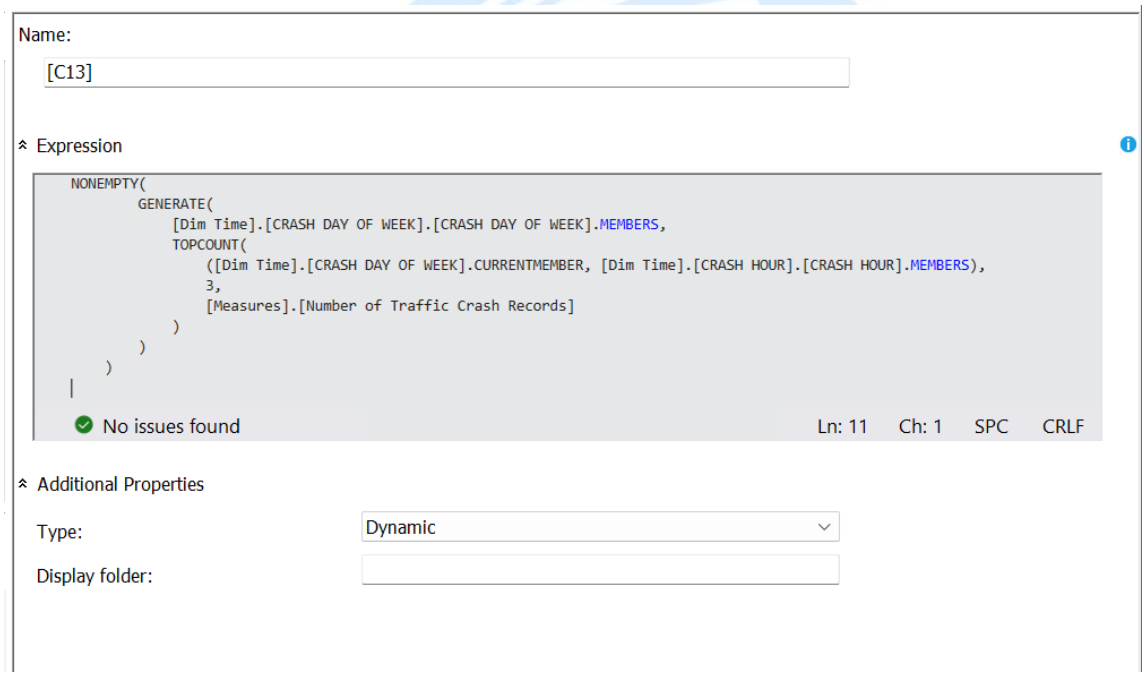
		Number of Traffic Crash Records
1	14	6329
1	15	6104
1	13	5944
2	15	8815
2	16	8762
2	17	8178
3	16	9340
3	17	9332
3	15	9121
4	17	9040
4	16	8945
4	15	8942
5	16	9350
5	15	9256
5	17	9087
6	16	10355
6	15	10345
6	17	9926
7	14	7875
7	15	7700
7	13	7657

3.6.13.3 Excel

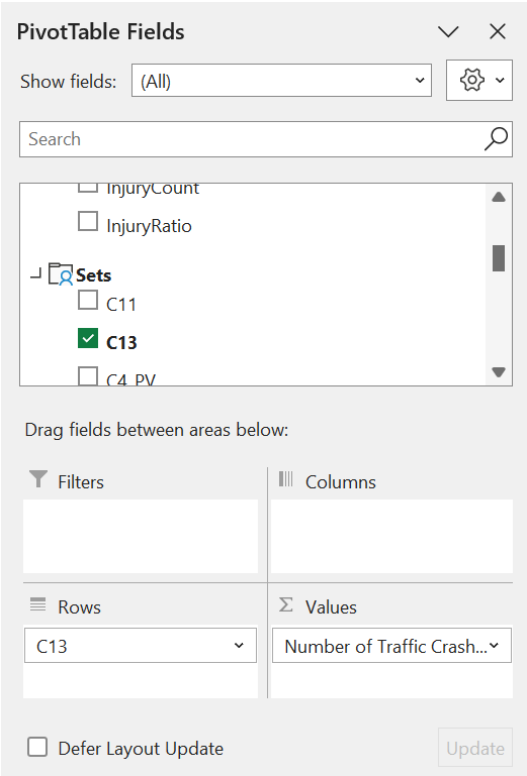
Bước 1: Vào mục **Calculations**, tạo mới một "New Named Set" đặt tên "C13. Sau đó nhập câu truy vấn và tiến hành Deloy.

Nội dung câu lệnh:

```
NONEMPTY(
        GENERATE(
            [Dim Time].[CRASH DAY OF WEEK].[CRASH DAY OF WEEK].MEMBERS,
            TOPCOUNT(
                ([Dim Time].[CRASH DAY OF WEEK].CURRENTMEMBER, [Dim Time].[CRASH
HOUR].[CRASH HOUR].MEMBERS),
                3,
                [Measures].[Number of Traffic Crash Records]
            )
        )
    )
```



Bước 2: Trong **PivotTable Fields**, ta chọn thuộc tính là Named Set "C13" và thuộc tính độ đo cần thiết.



Kết quả:

Row Labels	Number of Traffic Crash Records
1	
13	5944
14	6329
15	6104
2	
15	8815
16	8762
17	8178
3	
15	9121
16	9340
17	9332
4	
15	8942
16	8945
17	9040
5	
15	9256
16	9350
17	9087
6	
15	10345
16	10355
17	9926
7	
13	7657
14	7875
15	7700

3.6.13.4 Nhận xét

Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.14 Câu truy vấn 14

Nội dung câu truy vấn: Thống kê Top 5 nguyên nhân phổ biến nhất xuất hiện ở một trong các tập hợp: ({Nguyên nhân gây tử vong trung bình cao nhất}, {Nguyên nhân gây thiệt hại về tài sản cao nhất (OVER 1.500)}, {Nguyên nhân gây ra số lượng tử vong cao nhất khi tốc độ không vượt quá 40}) trong năm 2020.

3.6.14.1 Manual

Bước 1: Vào mục **Calculations**, tạo mới một "New Named Set" đặt tên "C14".

Sau đó nhập câu truy vấn và tiến hành Deloy.

Nội dung câu lệnh:

```
UNION(
  UNION(TOPCOUNT(
        FILTER(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].
[PRIM CONTRIBUTORY CAUSE].MEMBERS,
            [Measures].[Average Injuries Fatal] > 0
        ),
        5,
        [Measures].[Average Injuries Fatal]
    ),
        TOPCOUNT(
        FILTER(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].
[PRIM CONTRIBUTORY CAUSE].MEMBERS,
            ([Dim Damages].[DAMAGE].&[OVER 1.500])
        ),
        5,
        [Measures].[Number of Traffic Crash Records]
    )
   ),
        TOPCOUNT(
        FILTER(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].
[PRIM CONTRIBUTORY CAUSE].MEMBERS,
           ([Measures].[Injuries Fatal] > 0 AND [Dim Traffic Control].
[POSTED SPEED LIMIT] <= 40)
        ),
        5,
        [Measures].[Injuries Fatal]
    )
  )
```

Name:

[C14]

Expression

```
UNION(
        UNION(TOPCOUNT(
        FILTER(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM CONTRIBUTORY CAUSE].MEMBERS,
            [Measures].[Average Injuries Fatal] > 0
        ),
        5,
        [Measures].[Average Injuries Fatal]
    ), TOPCOUNT(
        FILTER(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM CONTRIBUTORY CAUSE].MEMBERS,
            ([Dim Damages].[DAMAGE].&[OVER 1.500])
        ),
        5,
        [Measures].[Number of Traffic Crash Records]
    )
),
        TOPCOUNT(
        FILTER(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM CONTRIBUTORY CAUSE].MEMBERS,
           ([Measures].[Injuries Fatal] > 0 AND [Dim Traffic Control].[POSTED SPEED LIMIT] <= 40)
        ),
        5,
        [Measures].[Injuries Fatal]
    )
    )
```

No issues found Ln: 27 Ch: 1 SPC CRLF

Additional Properties

Type: Dynamic

Display folder:

Bước 2: Vào mục **Browser**, kéo từ bảng "Dim Primary Causes" thuộc tính "PRIM CONTRIBUTORY CAUSE". Tại cột **Operator** chọn "In" và **Filter Expression** chọn vào Named Set "C14" vừa tạo.

Bước 3: Tiếp theo, kéo từ bảng "Dim Time" hierarchy "PHANCAP_Time" đã tạo. Tại cột **Operator** chọn "Equal" và **Filter Expression** chọn năm 2020.

Sau đó kéo các thuộc tính để hiển thị kết quả câu truy vấn và nhấn "Click to execute the query". Ta được kết quả.

Dimension	Hierarchy	Operator	Filter Expression	Parameters
Dim Primary Causes	PRIM CONTRIBUTORY ...	In	C14	
Dim Time	PHANCAP_TIME	Equal	{ 2020 }	
<Select dimension>				

PRIM CONTRIBUTORY CAUSE	Number of Traffic Crash Records	Injuries Fatal	Average Injuries Fatal
DISREGARDING TRAFFIC SIGNALS	2334	16	0.00690548122572292
DRIVING ON WRONG SIDE/WRONG WAY	575	1	0.00173913043478261
FAILING TO REDUCE SPEED TO AVOID CRASH	4807	11	0.00233001482736708
FAILING TO YIELD RIGHT-OF-WAY	9708	16	0.00173498156582086
FOLLOWING TOO CLOSELY	8211	0	0
IMPROPER OVERTAKING/PASSING	3992	4	0.00103092783505155
NOT APPLICABLE	4785	10	0.00217155266015201
PHYSICAL CONDITION OF DRIVER	653	15	0.0229709035222052
UNABLE TO DETERMINE	36179	55	0.00190568587367035
UNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED)	637	2	0.00313971742543171

3.6.14.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
WITH
-- Tập hợp top 5 nguyên nhân có trung bình số lượng tử vong cao nhất
SET [Top5FatalAverageCauses] AS
    TOPCOUNT(
        FILTER(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM CONTRIBUTORY
CAUSE].MEMBERS,
            [Measures].[Average Injuries Fatal] > 0
        ),
        5,
        [Measures].[Average Injuries Fatal]
    )
-- Tập hợp top 5 nguyên nhân gây thiệt hại tài sản “OVER 1.500” nhiều nhất
SET [Top5ModerateDamageCauses] AS
    TOPCOUNT(
        FILTER(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM CONTRIBUTORY
CAUSE].MEMBERS,
            ([Dim Damages].[DAMAGE].&[OVER 1.500])
        ),
        5,
        [Measures].[Number of Traffic Crash Records]
    )

-- Tập hợp top 5 nguyên nhân gây ra số lượng tử vong nhiều nhất khi tốc độ cho phép
không vượt quá 40
SET [Top5FatalInjuryCausesUnder40Speed] AS
    TOPCOUNT(
```

```
        FILTER(
            [Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM CONTRIBUTORY
CAUSE].MEMBERS,
           ([Measures].[Injuries Fatal] > 0 AND [Dim Traffic Control].[POSTED SPEED
LIMIT] <= 40)
        ),
        5,
        [Measures].[Injuries Fatal]
    )

-- Kết hợp ba tập hợp trên bằng UNION
SET [UnionTopCrashCauses] AS
    UNION(
        UNION([Top5FatalAverageCauses], [Top5ModerateDamageCauses]),
        [Top5FatalInjuryCausesUnder40Speed]
    )

SELECT
    {[Measures].[Number of Traffic Crash Records], [Measures].[Injuries Fatal],
[Average Injuries Fatal]} ON COLUMNS,
    [UnionTopCrashCauses] ON ROWS

FROM
    [Traffic Crash]
WHERE
 ([Dim Time].[CRASH YEAR].&[2020]);
```

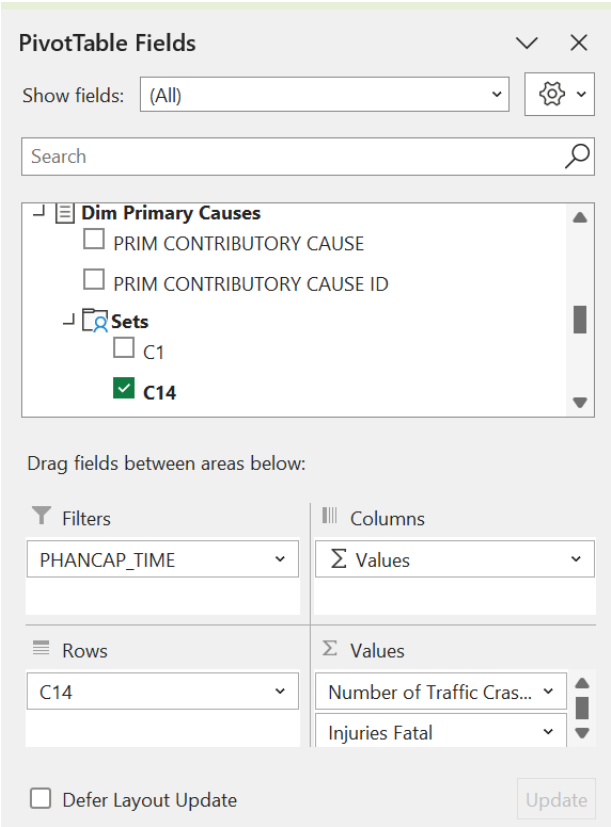
Kết quả câu truy vấn MDX:

Messages | Results

	Number of Traffic Crash Records	Injuries Fatal	Average Injuries Fatal
PHYSICAL CONDITION OF DRIVER	653	15	2.29709035222052E-02
DISREGARDING TRAFFIC SIGNALS	2334	16	6.90548122572292E-03
EVASIVE ACTION DUE TO ANIMAL; OBJECT; NONMOTORIST	194	1	5.15463917525773E-03
OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	1393	6	4.32900432900433E-03
DISREGARDING OTHER TRAFFIC SIGNS	233	1	4.29184549356223E-03
UNABLE TO DETERMINE	36179	55	1.90568587367035E-03
FAILING TO YIELD RIGHT-OF-WAY	9708	16	1.73498156582086E-03
FOLLOWING TOO CLOSELY	8211	0	0
FAILING TO REDUCE SPEED TO AVOID CRASH	4807	11	2.33001482736708E-03
NOT APPLICABLE	4785	10	2.17155266015201E-03

3.6.14.3 Excel

Trong **PivotTable Fields**, ta chọn Named Set đã tạo là “C14” và các thuộc tính độ đo cần thiết. Chọn “PHANCAP_Time” vào khu vực **Filters** và chọn năm là 2020.



Kết quả:

PHANCAP_TIME	2020		
Row Labels	**Number of Traffic Crash Records**	**Injuries Fatal**	**Average Injuries Fatal**
DISREGARDING OTHER TRAFFIC SIGNS	233	1	0.004291845
DISREGARDING TRAFFIC SIGNALS	2334	16	0.006905481
EVASIVE ACTION DUE TO ANIMAL; OBJECT; NONMOTORIST	194	1	0.005154639
FAILING TO REDUCE SPEED TO AVOID CRASH	4807	11	0.002330015
FAILING TO YIELD RIGHT-OF-WAY	9708	16	0.001734982
FOLLOWING TOO CLOSELY	8211	0	0
NOT APPLICABLE	4785	10	0.002171553
OPERATING VEHICLE IN ERRATIC; RECKLESS; CARELESS; NEGLIGENT OR AGGRESSIVE MANNER	1393	6	0.004329004
PHYSICAL CONDITION OF DRIVER	653	15	0.022970904
UNABLE TO DETERMINE	36179	55	0.001905686

3.6.14.4 Nhận xét

Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

3.6.15 Câu truy vấn 15

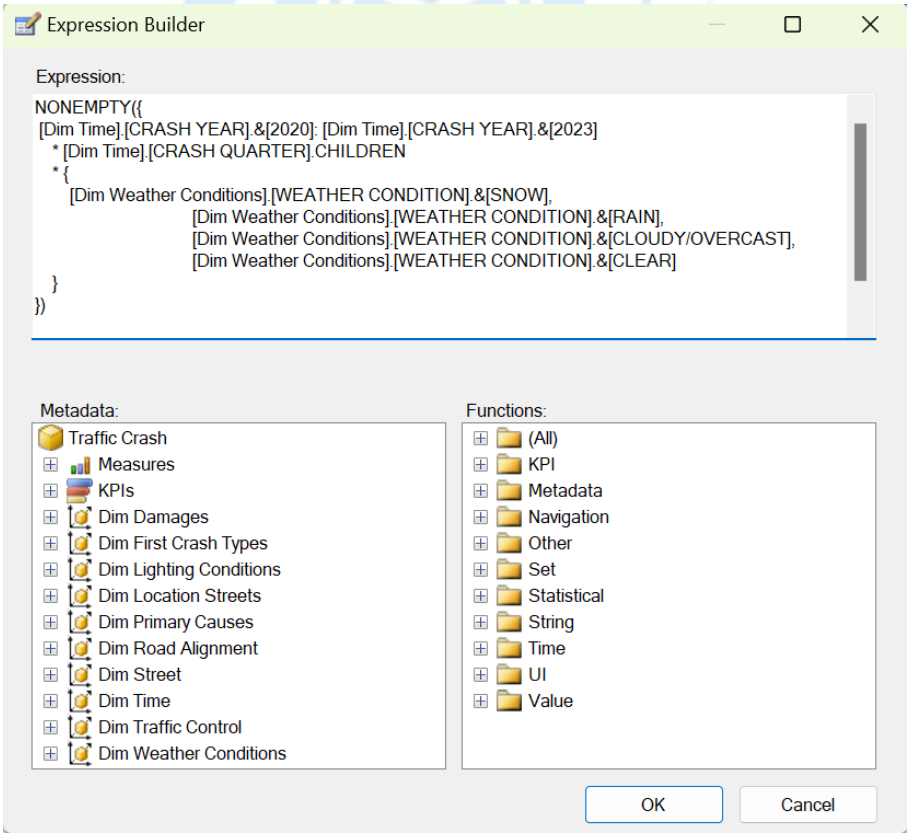
Nội dung câu truy vấn: Thống kê số vụ tai nạn và số lượng tử vong trong các điều kiện thời tiết ('SNOW', 'RAIN', 'CLOUDY/OVERCAST', 'CLEAR') ở từng quý từ năm 2020 đến 2023.

3.6.15.1 Manual

Bước 1: Vào mục **Browser**, kéo từ bảng "Dim Weather Conditions" thuộc tính "WEATHER CONDITION". Tại cột **Operator** chọn "Custom". Ở **Filter Expression** nhập vào câu lệnh truy vấn.

Nội dung câu lệnh:

```
NONEMPTY({
 [Dim Time].[CRASH YEAR].&[2020]: [Dim Time].[CRASH YEAR].&[2023]
   * [Dim Time].[CRASH QUARTER].CHILDREN
   * {
      [Dim Weather Conditions].[WEATHER CONDITION].&[SNOW],
      [Dim Weather Conditions].[WEATHER CONDITION].&[RAIN],
      [Dim Weather Conditions].[WEATHER CONDITION].&[CLOUDY/OVERCAST],
      [Dim Weather Conditions].[WEATHER CONDITION].&[CLEAR]
   }
})
```



Bước 2: Tiếp theo ta kéo các thuộc tính để hiển thị kết quả câu truy vấn và nhấn "Click to execute the query". Ta được kết quả.

Dimension	Hierarchy	Operator	Filter Expression	Paramet...
Dim Weather Conditions	WEATHER CONDITION	Custom	NONEMPTY({ [Dim Time].[CRASH YEAR].&[2020]: [D...	
<Select dimension>				

CRASH YEAR	CRASH QUARTER	WEATHER CONDITION	Number of Traffic Crash R...	Injuries Fatal
2020	1	CLEAR	17701	15
2020	1	CLOUDY/OVERCAST	1028	0
2020	1	RAIN	1678	3
2020	1	SNOW	2439	2
2020	2	CLEAR	15386	34
2020	2	CLOUDY/OVERCAST	383	1
2020	2	RAIN	1883	3
2020	2	SNOW	81	0
2020	3	CLEAR	23292	45
2020	3	CLOUDY/OVERCAST	433	1
2020	3	RAIN	1549	3
2020	3	SNOW	9	0
2020	4	CLEAR	17652	35
2020	4	CLOUDY/OVERCAST	656	0
2020	4	RAIN	2481	3
2020	4	SNOW	510	0
2021	1	CLEAR	16035	17
2021	1	CLOUDY/OVERCAST	724	1
2021	1	RAIN	654	2
2021	1	SNOW	4059	2
2021	2	CLEAR	23532	32
2021	2	CLOUDY/OVERCAST	794	0
2021	2	RAIN	2298	6
2021	2	SNOW	27	0
2021	3	CLEAR	26352	49

3.6.15.2 MDX

Viết câu truy vấn và chạy thực thi kết quả trong SSMS.

Câu truy vấn MDX:

```
SELECT {[Measures].[Number of Traffic Crash Records],
        [Measures].[Injuries Fatal]} ON COLUMNS,
NONEMPTY({
    [Dim Time].[CRASH YEAR].&[2020]: [Dim Time].[CRASH YEAR].&[2023]
    * [Dim Time].[CRASH QUARTER].CHILDREN
    * {
        [Dim Weather Conditions].[WEATHER CONDITION].&[SNOW],
              [Dim Weather Conditions].[WEATHER CONDITION].&[RAIN],
              [Dim Weather Conditions].[WEATHER CONDITION].&[CLOUDY/OVERCAST],
              [Dim Weather Conditions].[WEATHER CONDITION].&[CLEAR]
    }
}) ON ROWS
FROM [Traffic Crash];
```

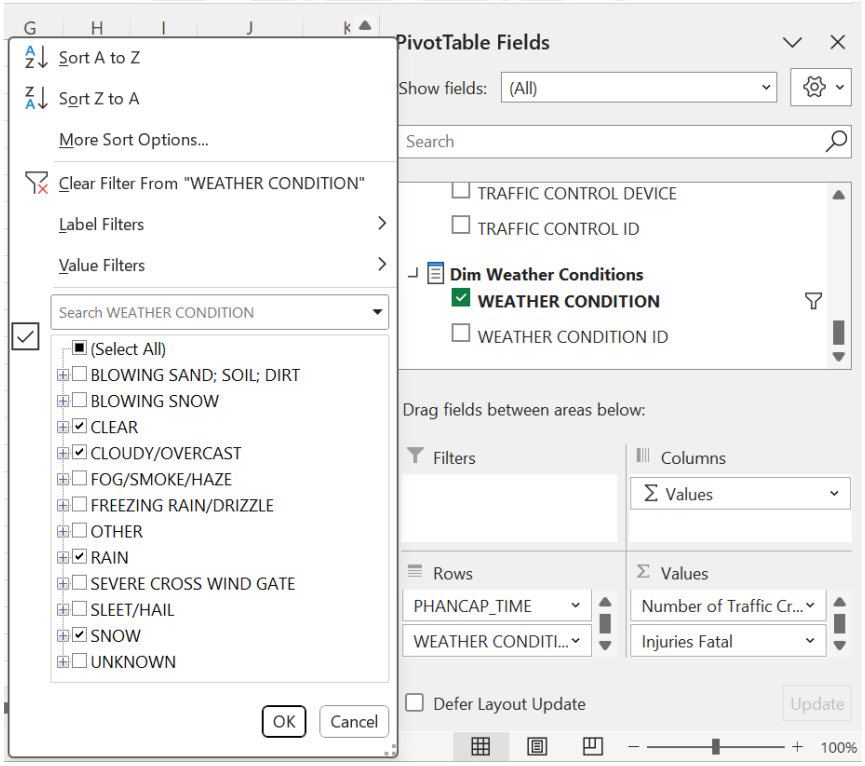
Kết quả câu truy vấn MDX:

Messages | Results

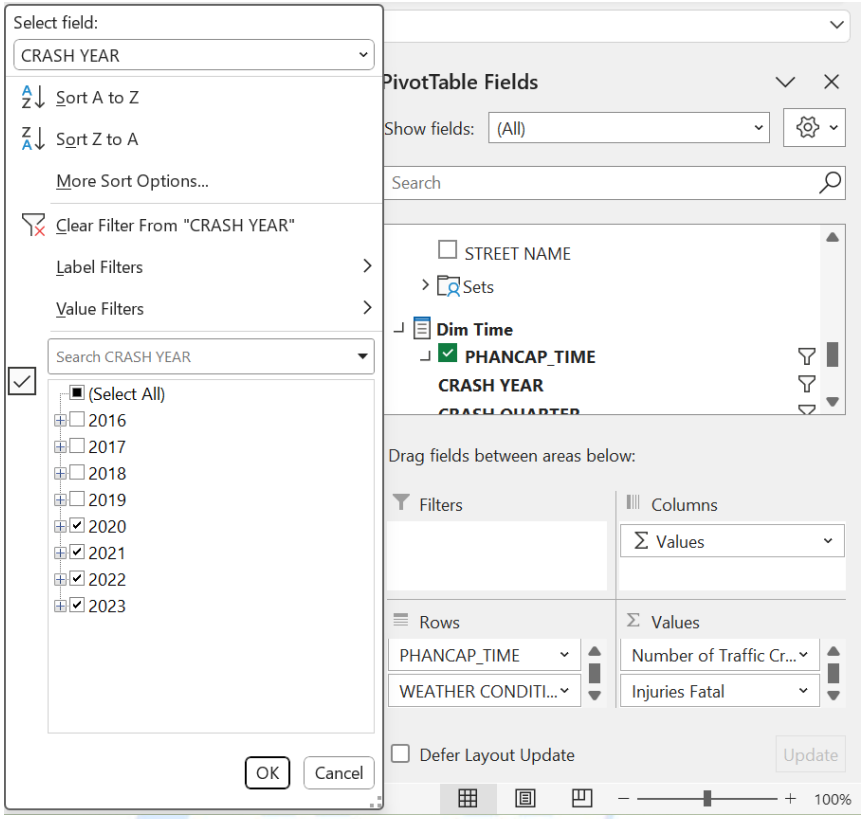
			Number of Traffic Crash Records	Injuries Fatal
2020	1	SNOW	2439	2
2020	1	RAIN	1678	3
2020	1	CLOUDY/OVERCAST	1028	0
2020	1	CLEAR	17701	15
2020	2	SNOW	81	0
2020	2	RAIN	1883	3
2020	2	CLOUDY/OVERCAST	383	1
2020	2	CLEAR	15386	34
2020	3	SNOW	9	0
2020	3	RAIN	1549	3
2020	3	CLOUDY/OVERCAST	433	1
2020	3	CLEAR	23292	45
2020	4	SNOW	510	0
2020	4	RAIN	2481	3
2020	4	CLOUDY/OVERCAST	656	0
2020	4	CLEAR	17652	35
2021	1	SNOW	4059	2
2021	1	RAIN	654	2
2021	1	CLOUDY/OVERCAST	724	1
2021	1	CLEAR	16035	17
2021	2	SNOW	27	0
2021	2	RAIN	2298	6
2021	2	CLOUDY/OVERCAST	794	0
2021	2	CLEAR	23532	32
2021	3	SNOW	3	0
2021	3	RAIN	1306	2
2021	3	CLOUDY/OVERCAST	576	1
2021	3	CLEAR	26352	49
2021	4	SNOW	223	1
2021	4	RAIN	3540	5

3.6.15.3 Excel

Bước 1: Trong **PivotTable Fields**, ở "WEATHER CONDITTION" của bảng "Dim Weather Conditions", ta chọn các giá trị ('SNOW', 'RAIN', 'CLOUDY/OVERCAST', 'CLEAR') để thống kê.



Bước 2: Chọn "PHANCAP_Time" và ở mục **Select field,** chọn các năm từ 2020 đến 2023.



Bước 3: Chọn các thuộc tính độ đo cần thiết.

Kết quả:

Row Labels	Number of Traffic Crash Records	Injuries Fatal
2020		
1		
CLEAR	17701	15
CLOUDY/OVERCAST	1028	0
RAIN	1678	3
SNOW	2439	2
2		
CLEAR	15386	34
CLOUDY/OVERCAST	383	1
RAIN	1883	3
SNOW	81	0
3		
CLEAR	23292	45
CLOUDY/OVERCAST	433	1
RAIN	1549	3
SNOW	9	0
4		
CLEAR	17652	35
CLOUDY/OVERCAST	656	0
RAIN	2481	3
SNOW	510	0
2021		
1		
CLEAR	16035	17
CLOUDY/OVERCAST	724	1
RAIN	654	2
SNOW	4059	2
2		
CLEAR	23532	32
CLOUDY/OVERCAST	794	0
RAIN	2298	6
SNOW	27	0
3		
CLEAR	26352	49
CLOUDY/OVERCAST	576	1
RAIN	1306	2
SNOW	3	0
4		
CLEAR	20600	41
CLOUDY/OVERCAST	1080	2
RAIN	3540	5
SNOW	223	1
2022		

3.6.15.4 Nhận xét

Nhận xét: Kết quả khi thực hiện câu truy vấn bằng thao tác Manual (trên Cube), MDX (trên SSMS) và Pivot Table (trên Excel) giống nhau.

CHƯƠNG 4: QUÁ TRÌNH TẠO BÁO CÁO DỮ LIỆU(SSRS)

4.1 Tạo project SSRS trên Power BI

4.1.1 Chuẩn bị công cụ Power BI

Bước 1: Truy cập link để tiến hành download công cụ Power BI.

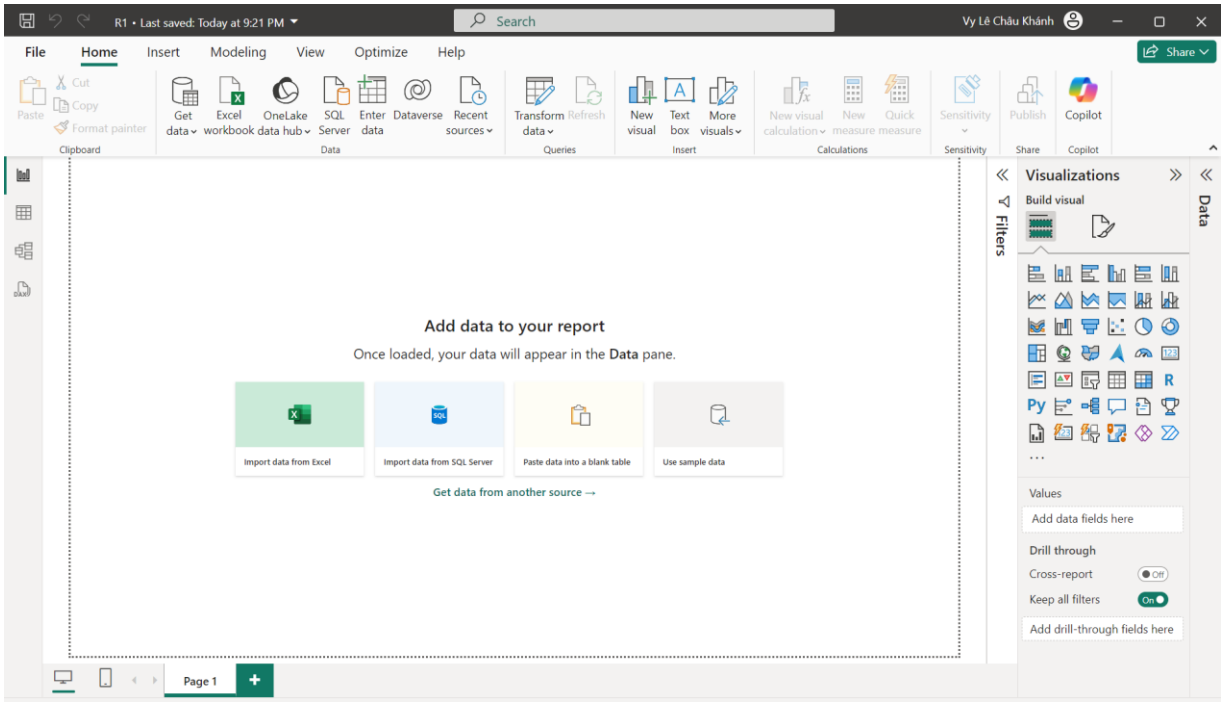
https://www.microsoft.com/vi-vn/download/details.aspx?id=58494

Bước 2: Chọn file *PBIDesktopSetup_x64.exe* và chọn Download để tiến hành cài đặt Power BI về máy.

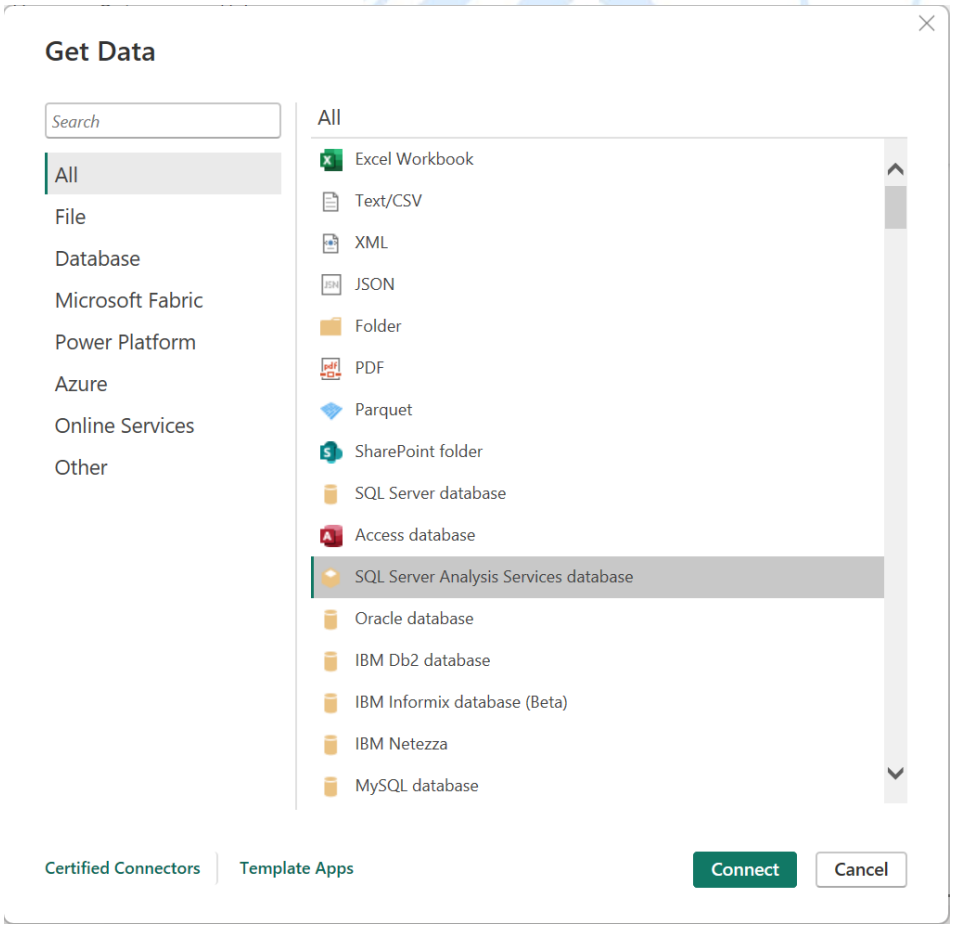
Bước 3: Thực hiện một số thao tác cài đặt và tiến hành đăng nhập Power BI bằng tài khoản Microsoft của trường (@ms.uit.edu.vn).

4.1.2 Thực hiện tạo báo cáo trên Power BI

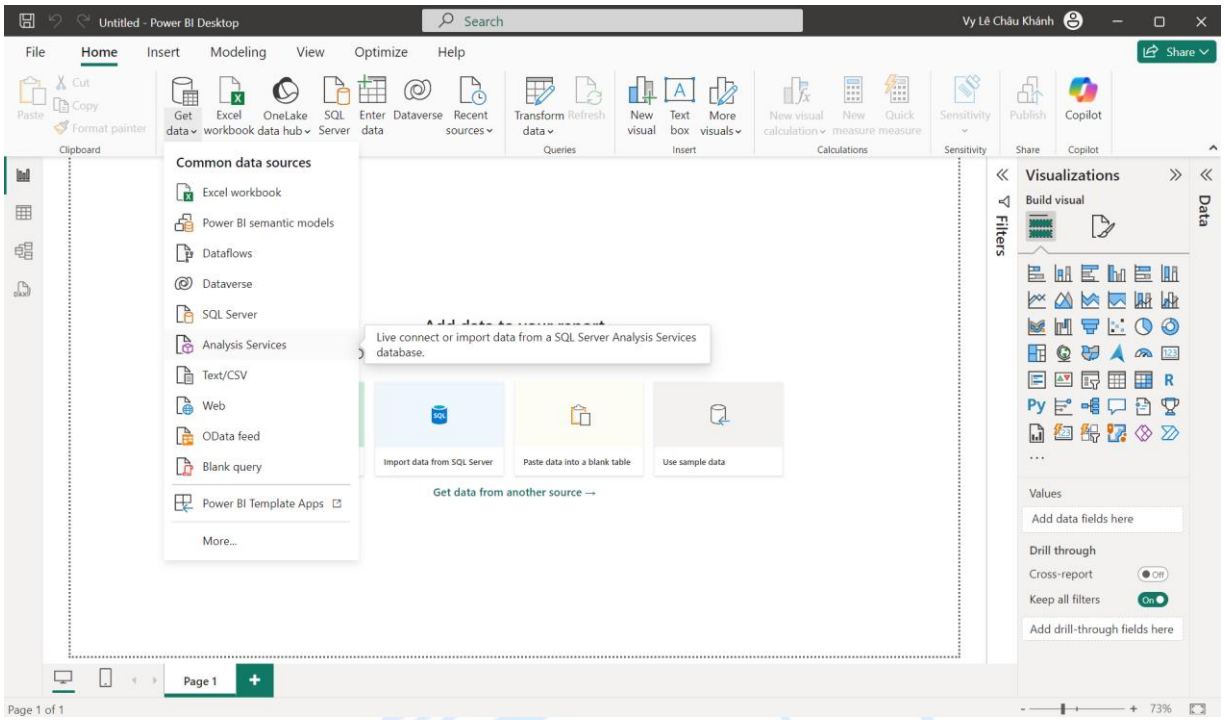
Bước 1.1: Mở Power BI, tạo mới một project và lưu vào thư mục phù hợp. Để tạo dataset, nhấn *Get data from another source*.



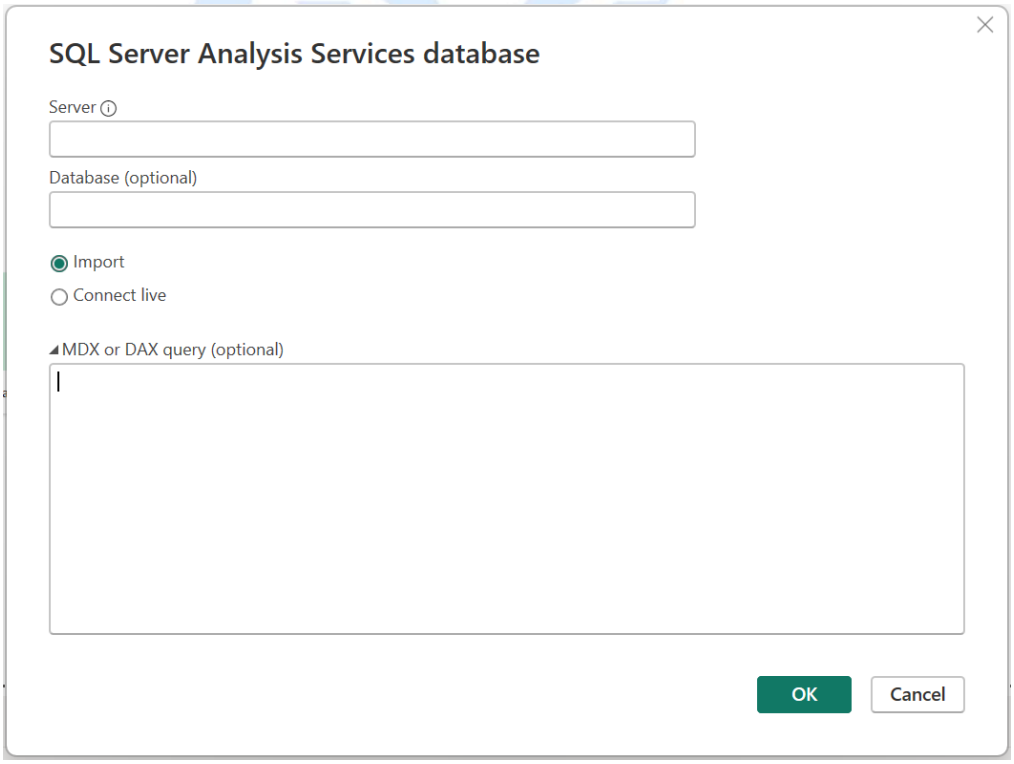
Ta tìm từ khóa trong thanh tìm kiếm và chọn SQL Server Analysis Services database.



Bước 1.2: Hoặc ta có thể lấy dữ liệu khi chọn **Get data** > **Analysis Services**.



Bước 2: Nhập Server name và Database, sau đó chọn **Import**, bên dưới là MDX or DAX query, nhập câu truy vấn MDX phù hợp cho report. Nhấn "OK" để nhận lấy dữ liệu.



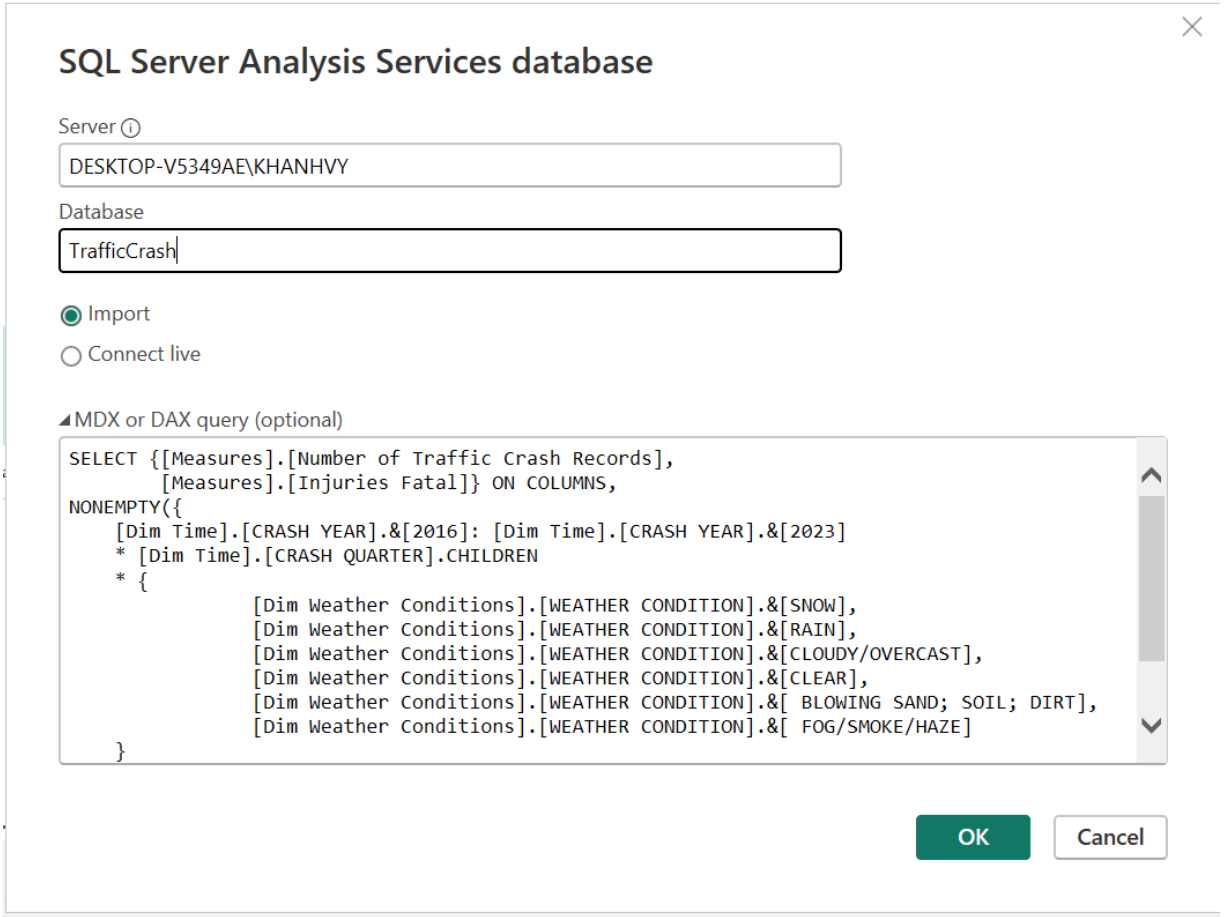
4.1.2.1 Report 1

Nội dung: Thống kê tổng số vụ tai nạn, tổng số nạn nhân và trung bình số lượng tử vong của từng quý từ năm 2016 đến 2023 trong các điều kiện thời tiết (‘SNOW’, ‘RAIN’, ‘CLOUDY/OVERCAST’, ‘CLEAR’, ‘BLOWING SAND; SOIL; DIRT’, ‘FOG/SMOKE/HAZE’).

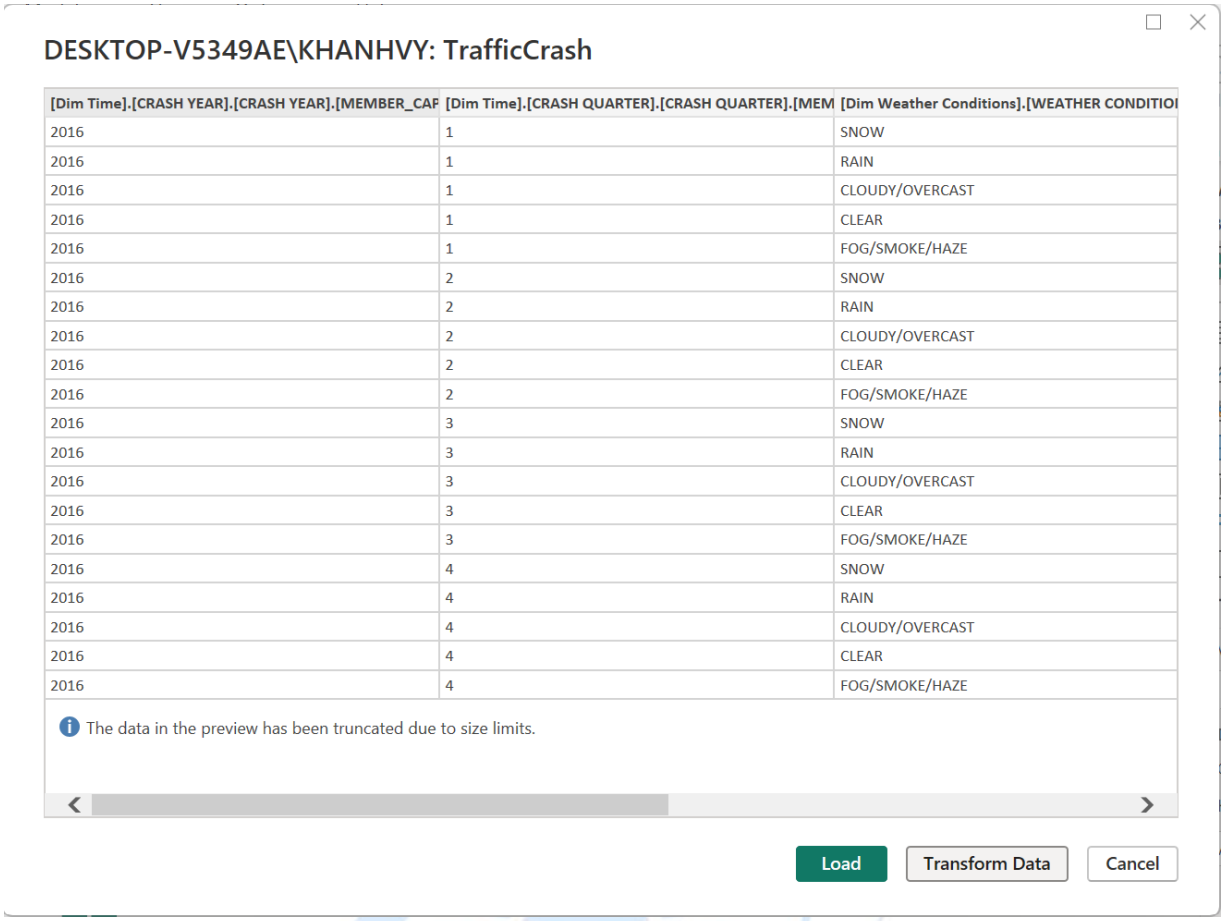
Bước 1: Nhập Server name Database và câu truy vấn MDX. Sau đó nhấn “OK” để xác nhận.

Câu lệnh:

```
SELECT {[Measures].[Number of Traffic Crash Records],
[Measures].[Average Injuries Fatal],
    [Measures].[Injuries Fatal], [Measures].[Injuries Total]} ON COLUMNS,
NONEMPTY({
  [Dim Time].[CRASH YEAR].&[2016]: [Dim Time].[CRASH YEAR].&[2023]
  * [Dim Time].[CRASH QUARTER].CHILDREN
  * {
                [Dim Weather Conditions].[WEATHER CONDITION].&[SNOW],
                [Dim Weather Conditions].[WEATHER CONDITION].&[RAIN],
                [Dim Weather Conditions].[WEATHER CONDITION].&[CLOUDY/OVERCAST],
                [Dim Weather Conditions].[WEATHER CONDITION].&[CLEAR],
                [Dim Weather Conditions].[WEATHER CONDITION].&[BLOWING SAND; SOIL; DIRT],
                [Dim Weather Conditions].[WEATHER CONDITION].&[FOG/SMOKE/HAZE]
  }
}) ON ROWS
FROM [Traffic Crash];
```



Bước 2: Ta tiến hành chọn "Transform data" để biến đổi dữ liệu.



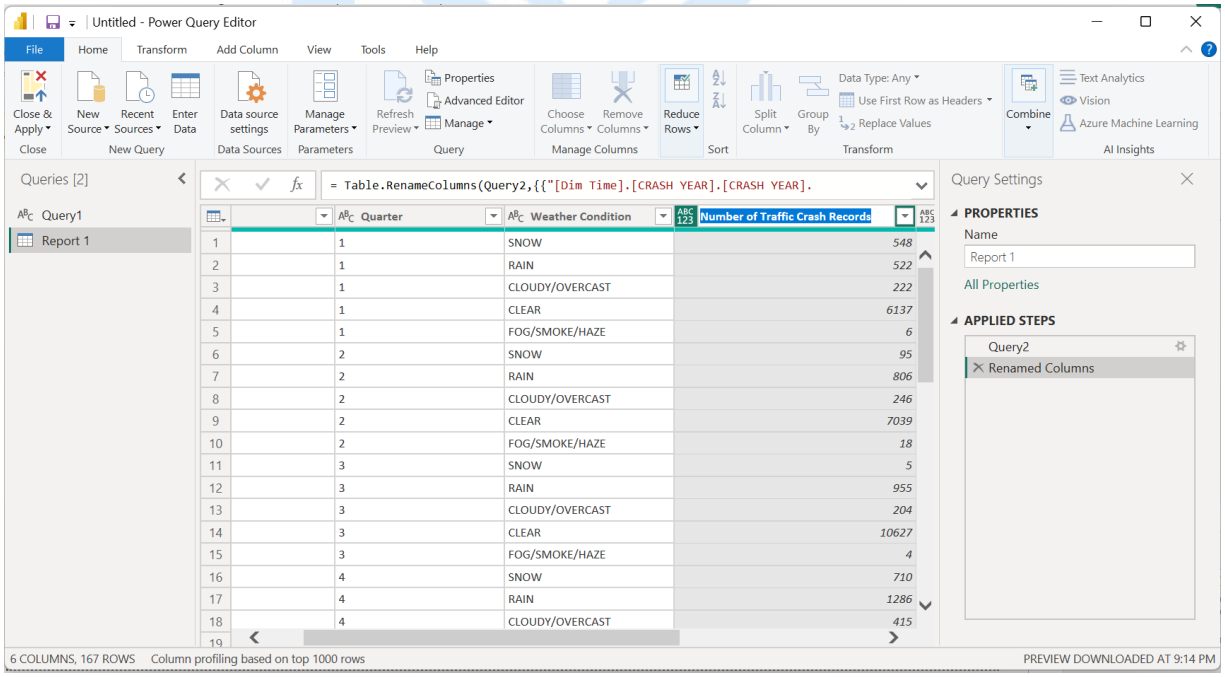
DESKTOP-V5349AE\KHANHVY: TrafficCrash

[Dim Time].[CRASH YEAR].[CRASH YEAR].[MEMBER_CAF	[Dim Time].[CRASH QUARTER].[CRASH QUARTER].[MEM	[Dim Weather Conditions].[WEATHER CONDITIOI
2016	1	SNOW
2016	1	RAIN
2016	1	CLOUDY/OVERCAST
2016	1	CLEAR
2016	1	FOG/SMOKE/HAZE
2016	2	SNOW
2016	2	RAIN
2016	2	CLOUDY/OVERCAST
2016	2	CLEAR
2016	2	FOG/SMOKE/HAZE
2016	3	SNOW
2016	3	RAIN
2016	3	CLOUDY/OVERCAST
2016	3	CLEAR
2016	3	FOG/SMOKE/HAZE
2016	4	SNOW
2016	4	RAIN
2016	4	CLOUDY/OVERCAST
2016	4	CLEAR
2016	4	FOG/SMOKE/HAZE

The data in the preview has been truncated due to size limits.

Load | Transform Data | Cancel

Bước 3: Tiến hành đổi tên cột dữ liệu cho phù hợp.



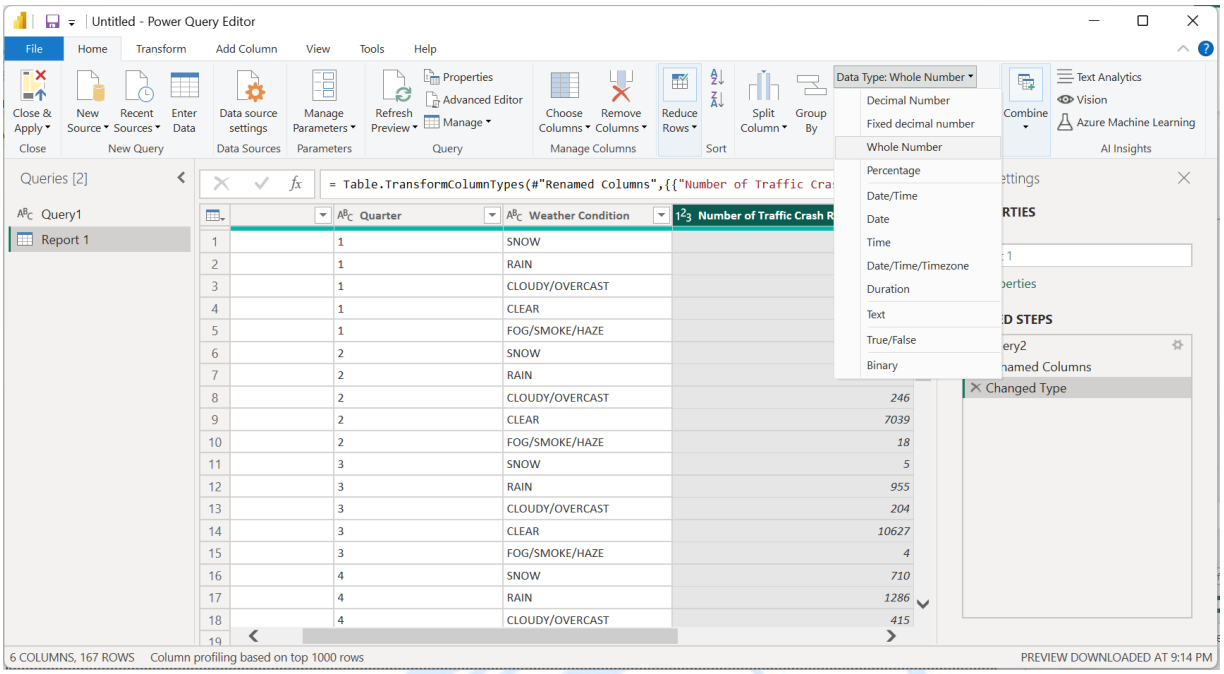
Untitled - Power Query Editor

= Table.RenameColumns(Query2,{{"[Dim Time].[CRASH YEAR].[CRASH YEAR].

	Quarter	Weather Condition	Number of Traffic Crash Records
1	1	SNOW	548
2	1	RAIN	522
3	1	CLOUDY/OVERCAST	222
4	1	CLEAR	6137
5	1	FOG/SMOKE/HAZE	6
6	2	SNOW	95
7	2	RAIN	806
8	2	CLOUDY/OVERCAST	246
9	2	CLEAR	7039
10	2	FOG/SMOKE/HAZE	18
11	3	SNOW	5
12	3	RAIN	955
13	3	CLOUDY/OVERCAST	204
14	3	CLEAR	10627
15	3	FOG/SMOKE/HAZE	4
16	4	SNOW	710
17	4	RAIN	1286
18	4	CLOUDY/OVERCAST	415

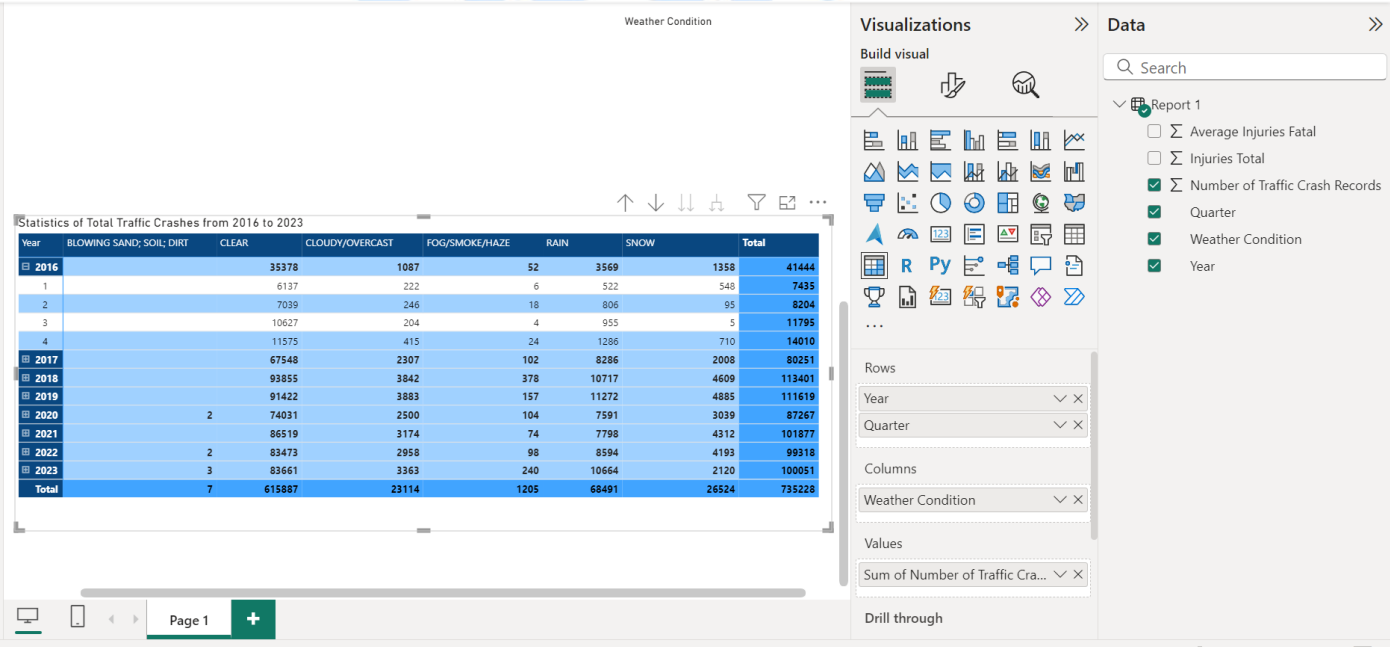
6 COLUMNS, 167 ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 9:14 PM

Bước 4: Đổi lại kiểu dữ liệu nếu chưa phù hợp.



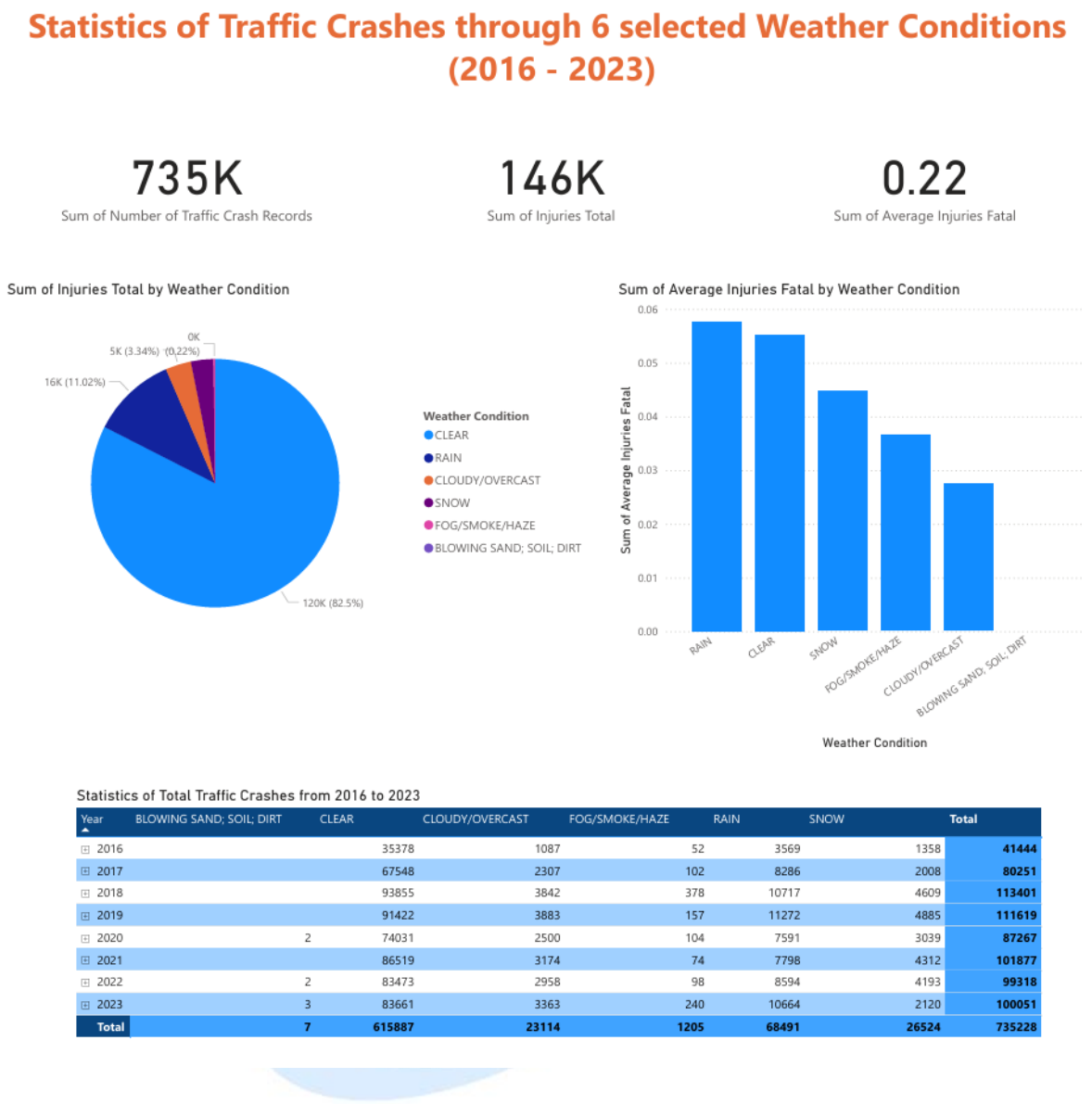
Nhấn **Close and Apply** để hoàn tất.

Bước 6: Chọn các giao diện, biểu đồ để hiển thị kết quả thống kê. "Report 1" ở đây bao gồm một thành phần: Card, Stacked Column Chart, Pie Chart, Matrix. Ở Matrix, ta có thể drill down năm để xem số vụ tai nạn ở từng quý.



Year	BLOWING SAND; SOIL; DIRT	CLEAR	CLOUDY/OVERCAST	FOG/SMOKE/HAZE	RAIN	SNOW	Total
2016		35378	1087	52	3569	1358	41444
1		6137	222	6	522	548	7435
2		7039	246	18	806	95	8204
3		10627	204	4	955	5	11795
4		11575	415	24	1286	710	14010
2017		67548	2307	102	8286	2008	80251
2018		93855	3842	378	10717	4609	113401
2019		91422	3883	157	11272	4885	111619
2020	2	74031	2500	104	7591	3039	87267
2021		86519	3174	74	7798	4312	101877
2022	2	83473	2958	98	8594	4193	99318
2023	3	83661	3363	240	10664	2120	100051
Total	7	615887	23114	1205	68491	26524	735228

*Kết quả "Report 1".



Year	BLOWING SAND; SOIL; DIRT	CLEAR	CLOUDY/OVERCAST	FOG/SMOKE/HAZE	RAIN	SNOW	Total
2016		35378	1087	52	3569	1358	41444
2017		67548	2307	102	8286	2008	80251
2018		93855	3842	378	10717	4609	113401
2019		91422	3883	157	11272	4885	111619
2020	2	74031	2500	104	7591	3039	87267
2021		86519	3174	74	7798	4312	101877
2022	2	83473	2958	98	8594	4193	99318
2023	3	83661	3363	240	10664	2120	100051
Total	**7**	**615887**	**23114**	**1205**	**68491**	**26524**	**735228**

4.1.2.2 Report 2

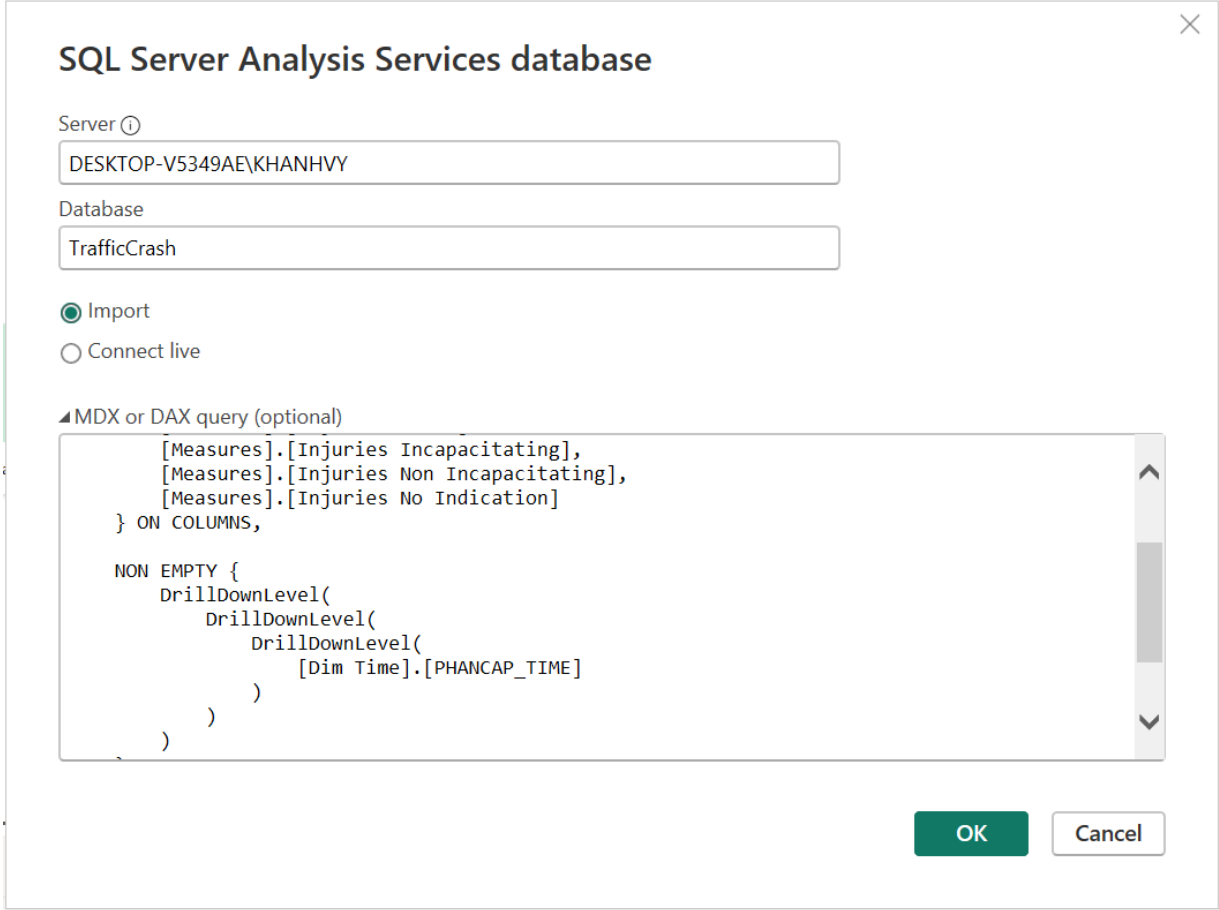
Nội dung: Thống kê thông số thương vong do va chạm giao thông từ năm 2016 đến 2023.

Bước 1: Nhập Server name Database và câu truy vấn MDX. Sau đó nhấn "OK" để xác nhận.

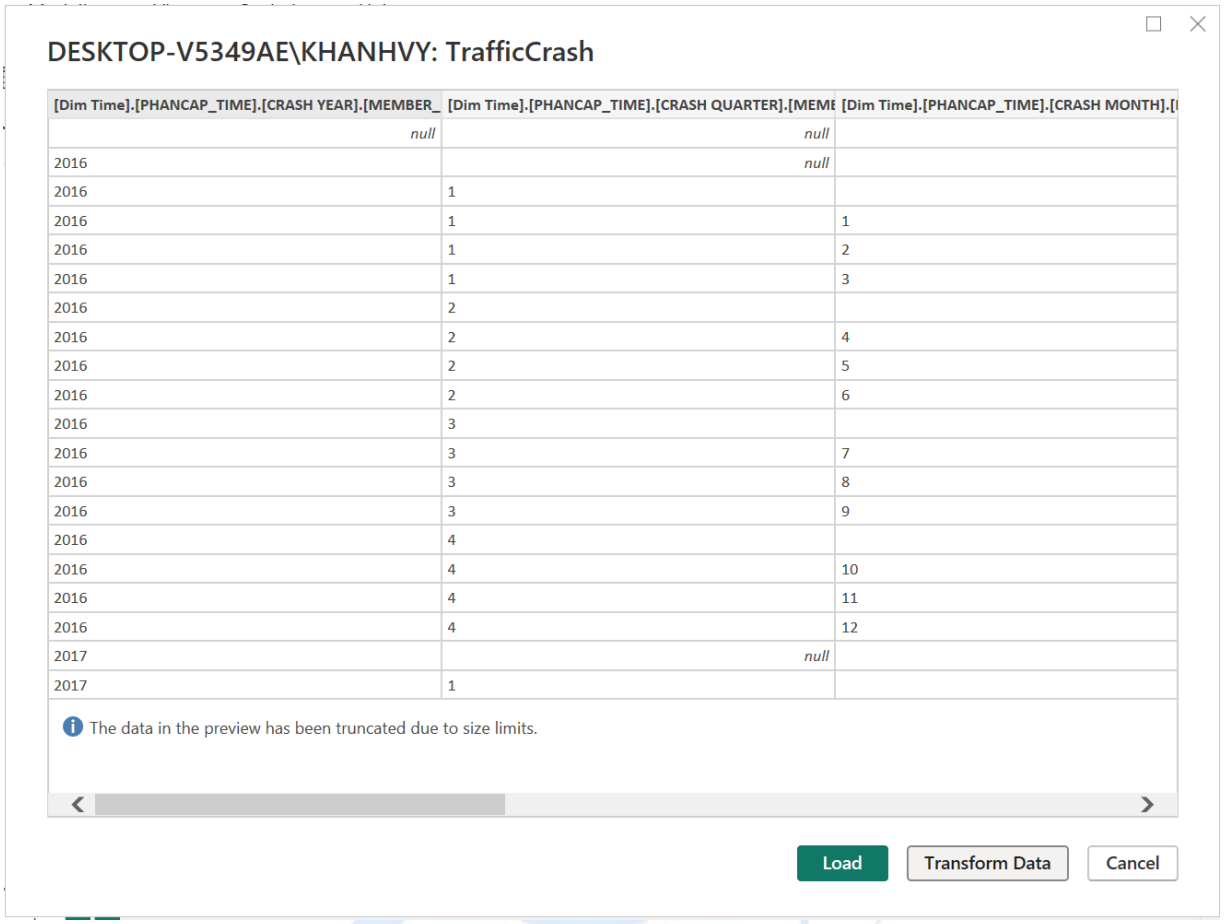
Câu lệnh:

```
SELECT
  NON EMPTY {
    [Measures].[Number of Traffic Crash Records],
```

```
    [Measures].[Injuries Total],
    [Measures].[Injuries Fatal],
    [Measures].[Injuries Incapacitating],
    [Measures].[Injuries Non Incapacitating],
    [Measures].[Injuries No Indication]
  } ON COLUMNS,

  NON EMPTY {
    DrillDownLevel(
      DrillDownLevel(
        DrillDownLevel(
          [Dim Time].[PHANCAP_TIME]
        )
      )
    )
  } ON ROWS
FROM
  [Traffic Crash];
```



Bước 2: Ta tiến hành chọn "Transform data" để biến đổi dữ liệu.



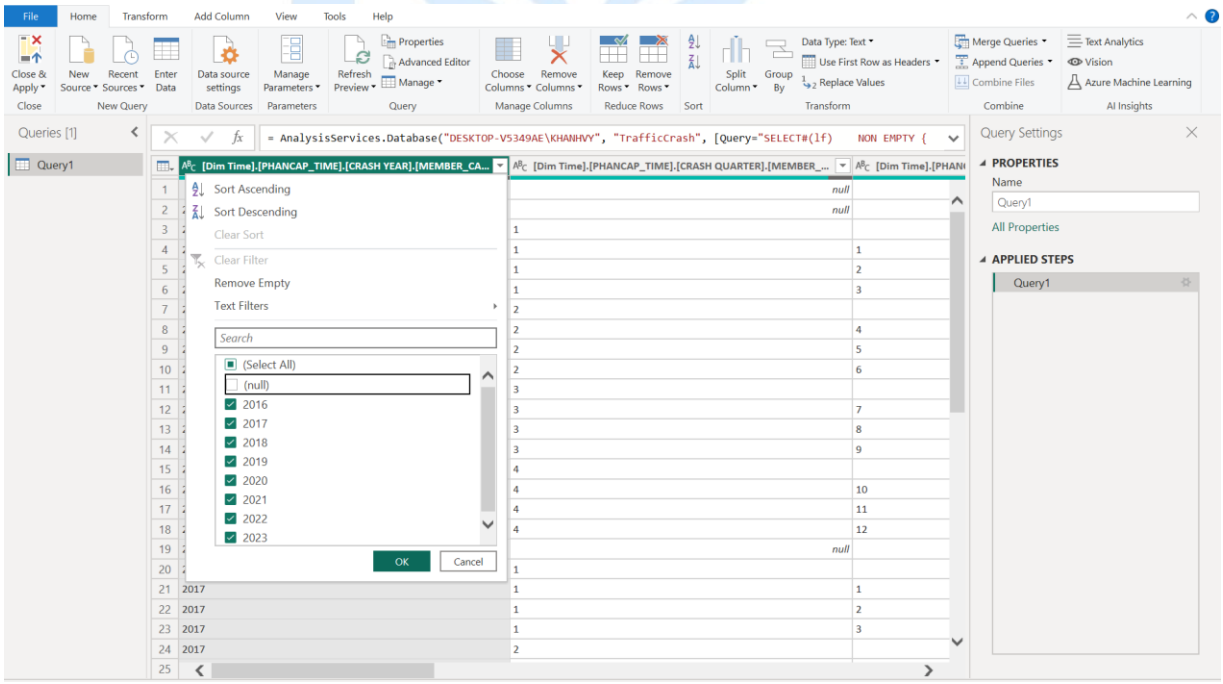
DESKTOP-V5349AE\KHANHVY: TrafficCrash

[Dim Time].[PHANCAP_TIME].[CRASH YEAR].[MEMBER_	[Dim Time].[PHANCAP_TIME].[CRASH QUARTER].[MEMI	[Dim Time].[PHANCAP_TIME].[CRASH MONTH].[
null	null	
2016	null	
2016	1	
2016	1	1
2016	1	2
2016	1	3
2016	2	
2016	2	4
2016	2	5
2016	2	6
2016	3	
2016	3	7
2016	3	8
2016	3	9
2016	4	
2016	4	10
2016	4	11
2016	4	12
2017	null	
2017	1	

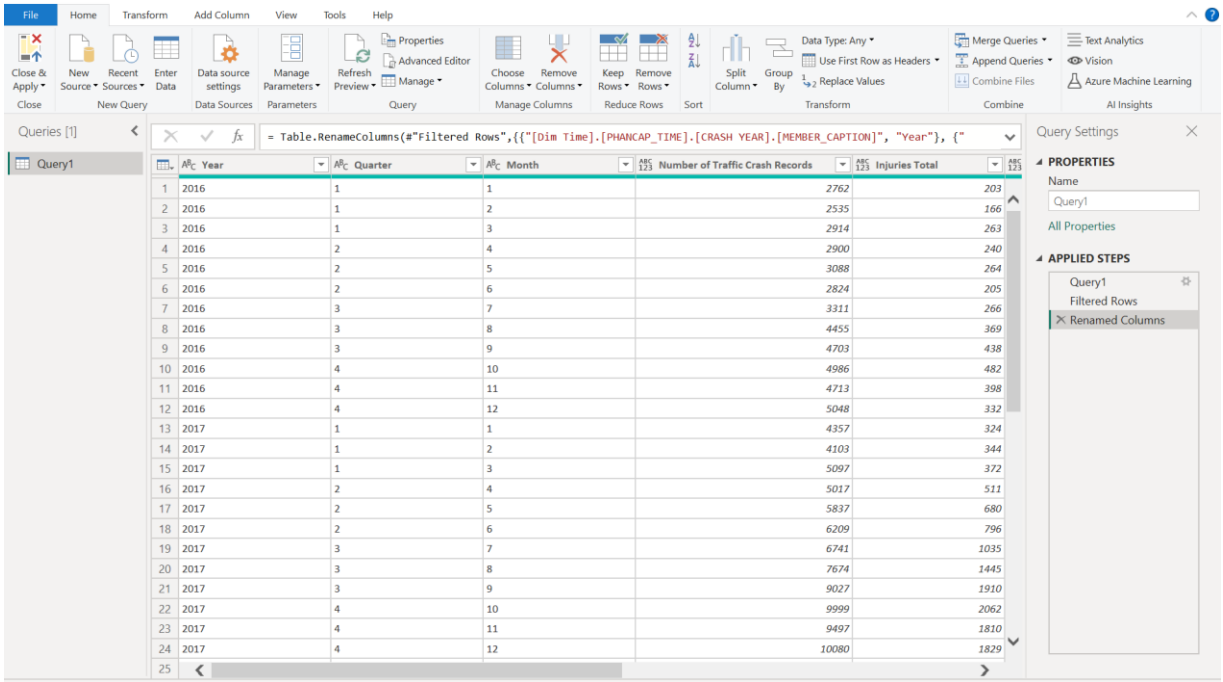
The data in the preview has been truncated due to size limits.

Load | Transform Data | Cancel

Bước 3: Ta có thể lọc bỏ các giá trị ‘null’ ở các cột.

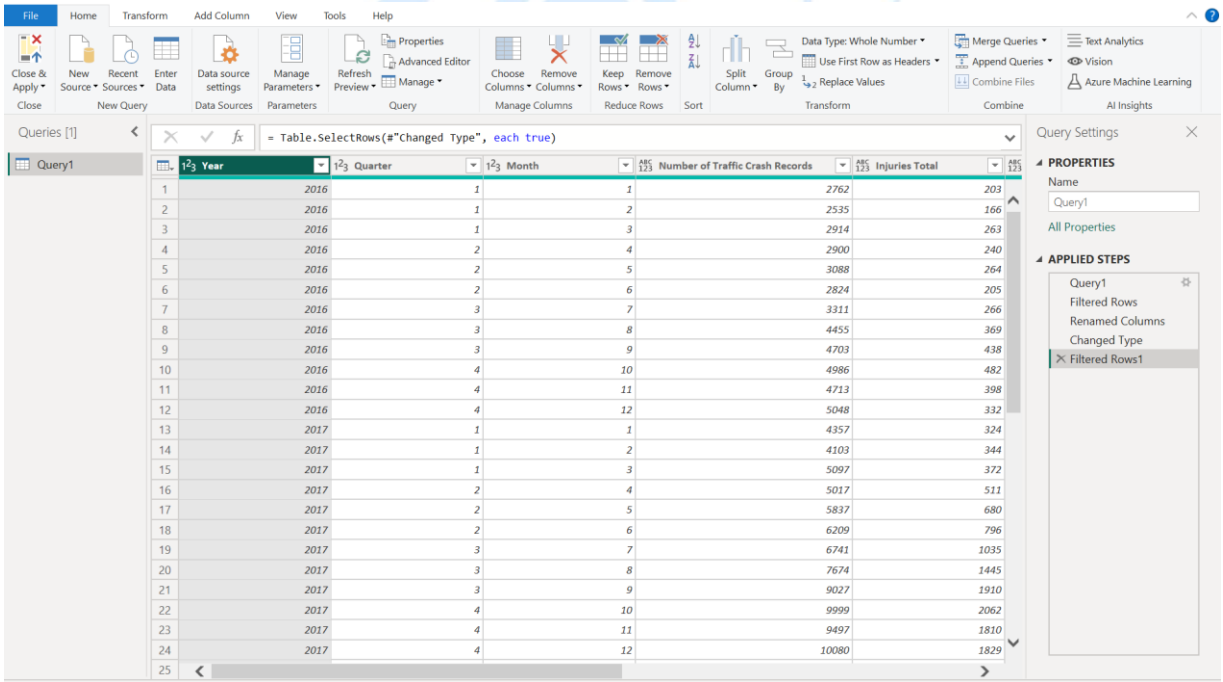


Bước 4: Tiến hành đổi tên cột dữ liệu cho phù hợp.



	Year	Quarter	Month	Number of Traffic Crash Records	Injuries Total
1	2016	1	1	2762	203
2	2016	1	2	2535	166
3	2016	1	3	2914	263
4	2016	2	4	2900	240
5	2016	2	5	3088	264
6	2016	2	6	2824	205
7	2016	3	7	3311	266
8	2016	3	8	4455	369
9	2016	3	9	4703	438
10	2016	4	10	4986	482
11	2016	4	11	4713	398
12	2016	4	12	5048	332
13	2017	1	1	4357	324
14	2017	1	2	4103	344
15	2017	1	3	5097	372
16	2017	2	4	5017	511
17	2017	2	5	5837	680
18	2017	2	6	6209	796
19	2017	3	7	6741	1035
20	2017	3	8	7674	1445
21	2017	3	9	9027	1910
22	2017	4	10	9999	2062
23	2017	4	11	9497	1810
24	2017	4	12	10080	1829

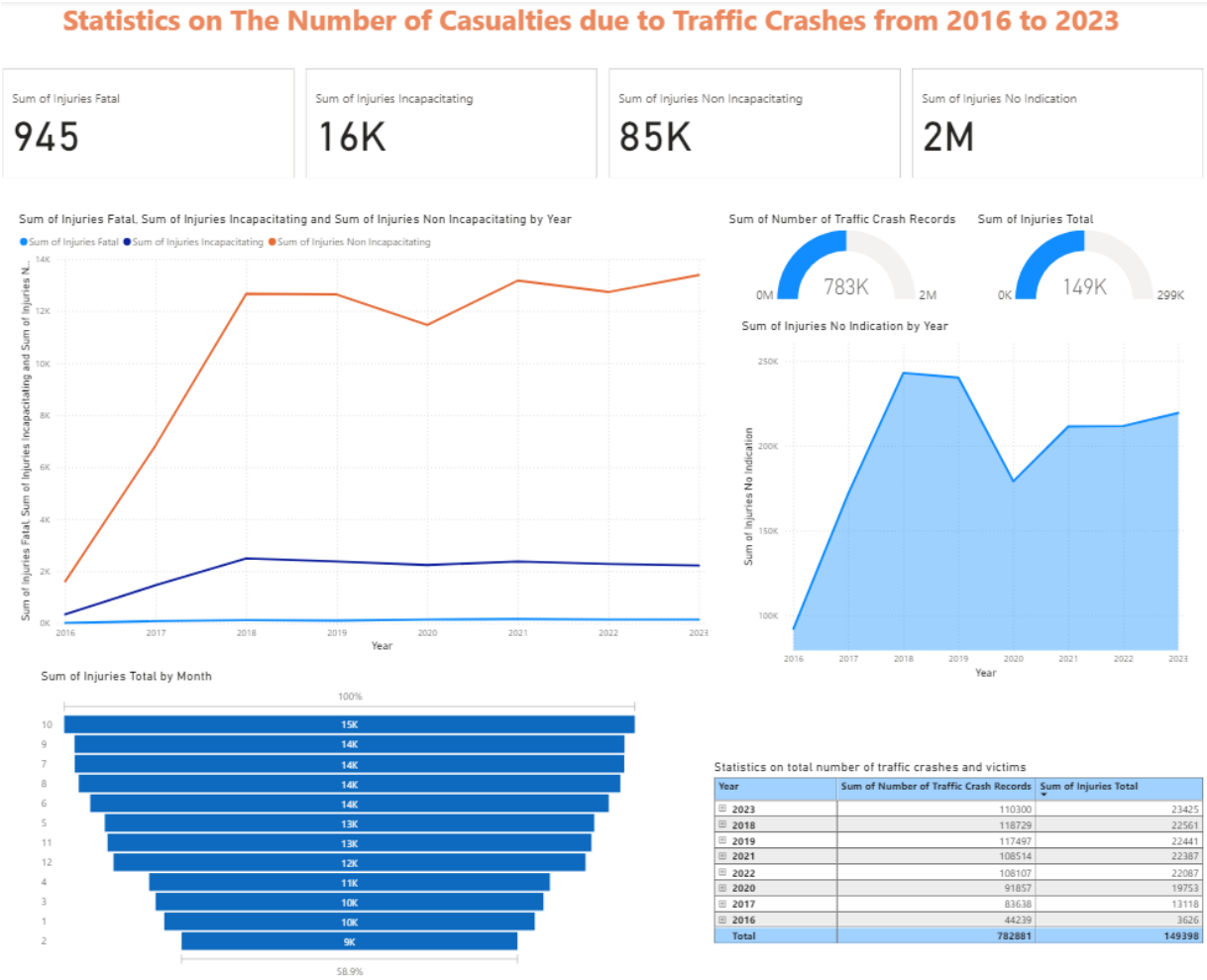
Bước 5: Đổi lại kiểu dữ liệu nếu chưa phù hợp.



	Year	Quarter	Month	Number of Traffic Crash Records	Injuries Total
1	2016	1	1	2762	203
2	2016	1	2	2535	166
3	2016	1	3	2914	263
4	2016	2	4	2900	240
5	2016	2	5	3088	264
6	2016	2	6	2824	205
7	2016	3	7	3311	266
8	2016	3	8	4455	369
9	2016	3	9	4703	438
10	2016	4	10	4986	482
11	2016	4	11	4713	398
12	2016	4	12	5048	332
13	2017	1	1	4357	324
14	2017	1	2	4103	344
15	2017	1	3	5097	372
16	2017	2	4	5017	511
17	2017	2	5	5837	680
18	2017	2	6	6209	796
19	2017	3	7	6741	1035
20	2017	3	8	7674	1445
21	2017	3	9	9027	1910
22	2017	4	10	9999	2062
23	2017	4	11	9497	1810
24	2017	4	12	10080	1829

Sau đó nhấn "Close and Apply".

Bước 6: Chọn các giao diện, biểu đồ để hiển thị kết quả thống kê. "Report 2" ở đây bao gồm một thành phần: Line Chart, Area Chart, Matrix, Card, Gauge. Ở Matrix, ta có thể drill down năm để xem số vụ tai nạn và tổng số nạn nhân ở từng quý.

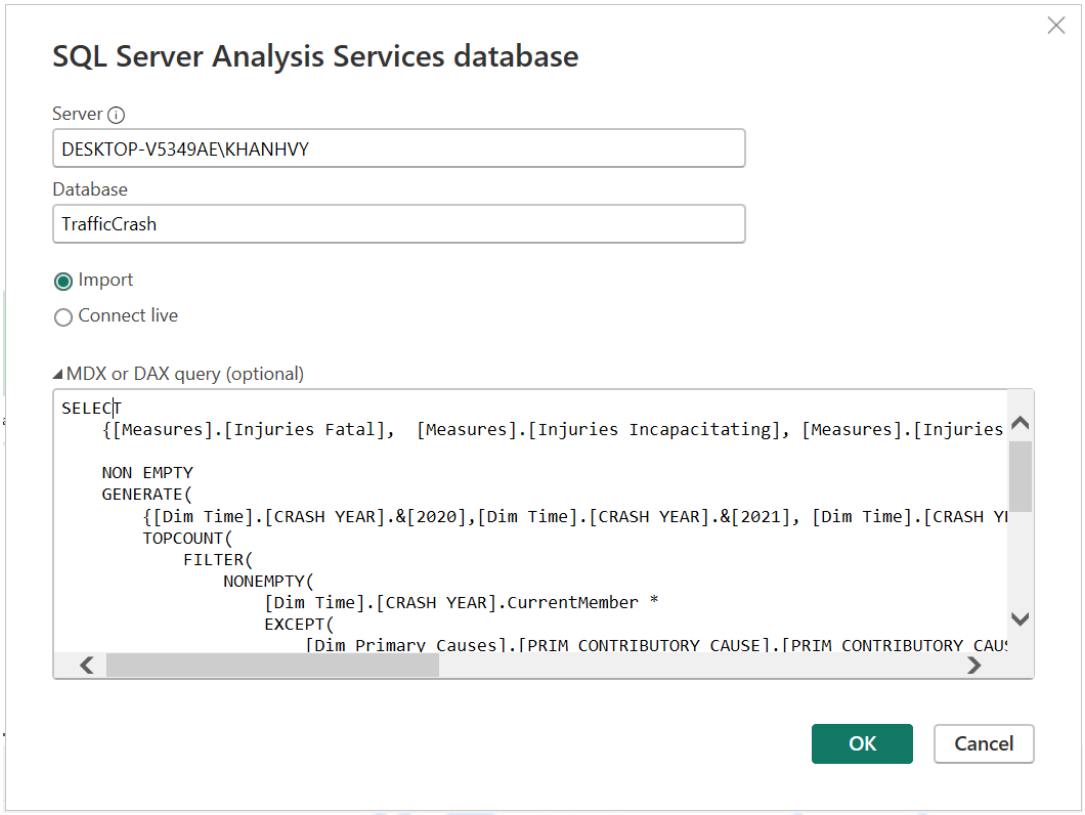


4.1.2.3 Report 3

Nội dung: Trong 10 nguyên nhân gây va chạm giao thông phổ biến nhất ở mỗi năm, thống kê số lượng nạn nhân tử vong, bị thương nặng và bị thương nhẹ ở mỗi nguyên nhân từ năm 2020 đến năm 2023 khi điều kiện thời tiết là "CLEAR".

Ở đây ta sẽ không tính nguyên nhân là "NOT APPLICABLE" và "UNABLE TO DETERMINE" vì đây là hai nguyên nhân không xác định.

Bước 1: Nhập Server name Database và câu truy vấn MDX. Sau đó nhấn "OK" để xác nhận.



Câu lệnh:

SELECT

{[Measures].[Injuries Fatal], [Measures].[Injuries Incapacitating], [Measures].[Injuries Non Incapacitating], [Measures].[Injuries No Indication], [Measures].[Injuries Total],[Measures].[Number of Traffic Crash Records], [Measures].[FatalRatio]} ON COLUMNS,

NON EMPTY

GENERATE(

{[Dim Time].[CRASH YEAR].&[2020],[Dim Time].[CRASH YEAR].&[2021], [Dim Time].[CRASH YEAR].&[2022], [Dim Time].[CRASH YEAR].&[2023]},

TOPCOUNT(

FILTER(

NONEMPTY(

[Dim Time].[CRASH YEAR].CurrentMember *

EXCEPT(

[Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[PRIM CONTRIBUTORY CAUSE].MEMBERS,

{

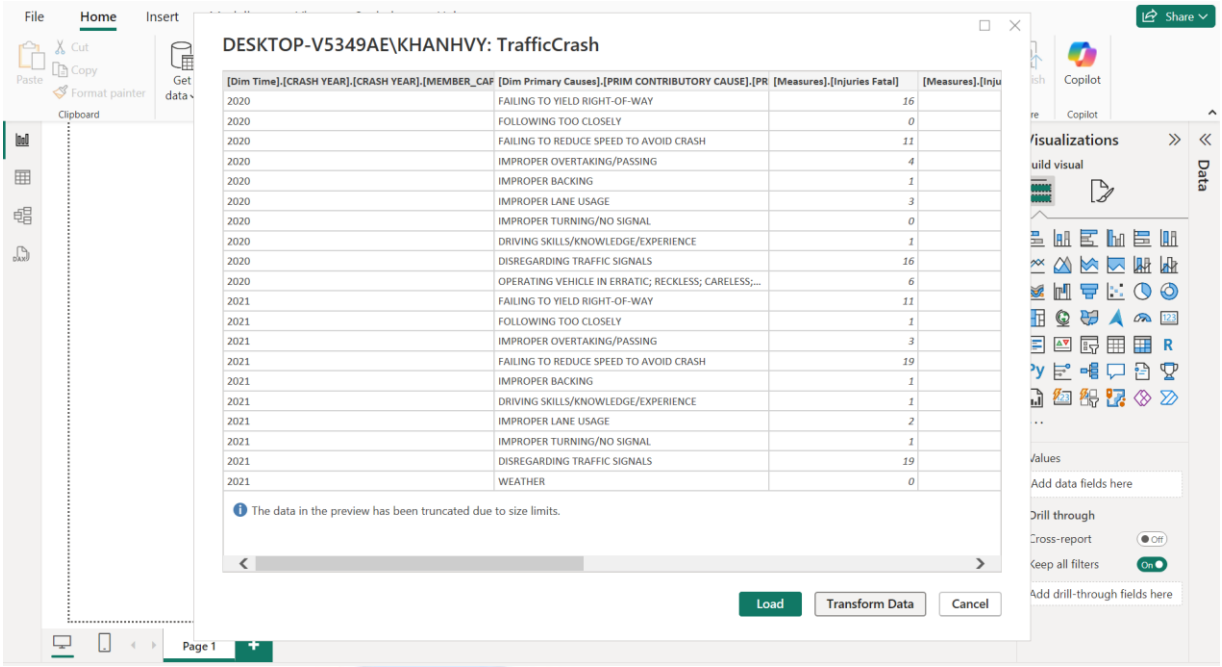
[Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[NOT APPLICABLE],

[Dim Primary Causes].[PRIM CONTRIBUTORY CAUSE].[UNABLE TO DETERMINE]

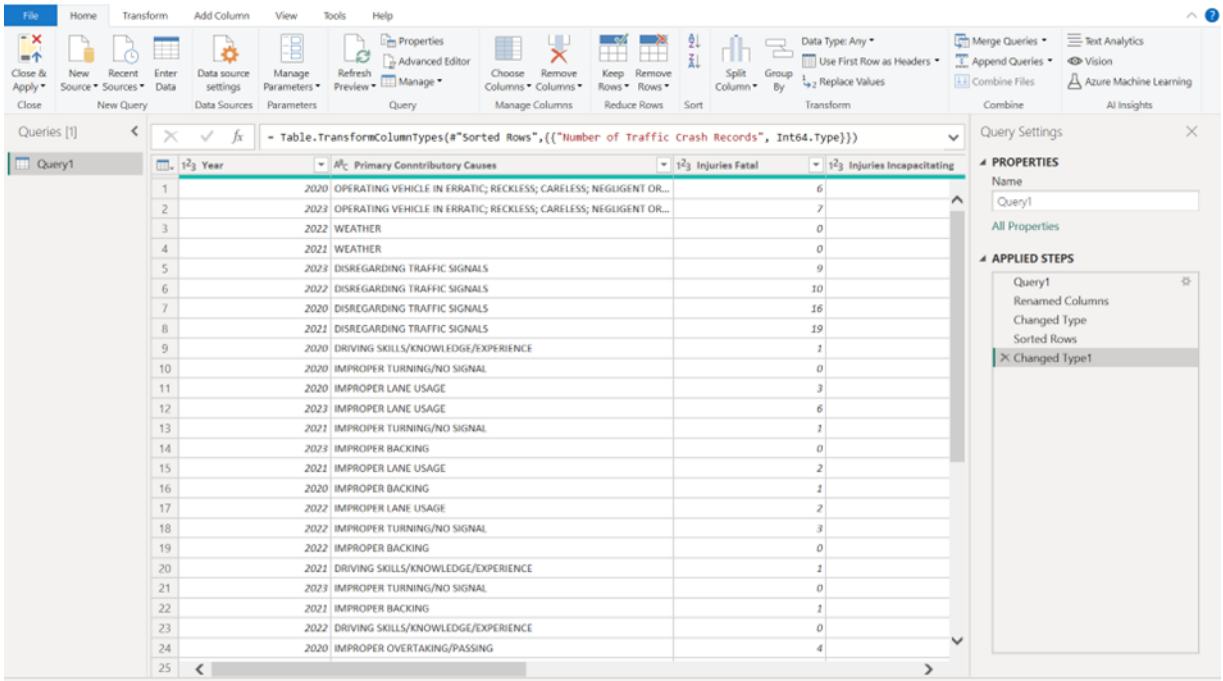
}

```
            )
          ),
          [Dim Weather Conditions].[WEATHER CONDITION].&[CLEAR]
        ),
        10,
        [Measures].[Number of Traffic Crash Records]
      )
  ) ON ROWS
FROM [Traffic Crash];
```

Bước 2: Ta tiến hành chọn "Transform data" để biến đổi dữ liệu.

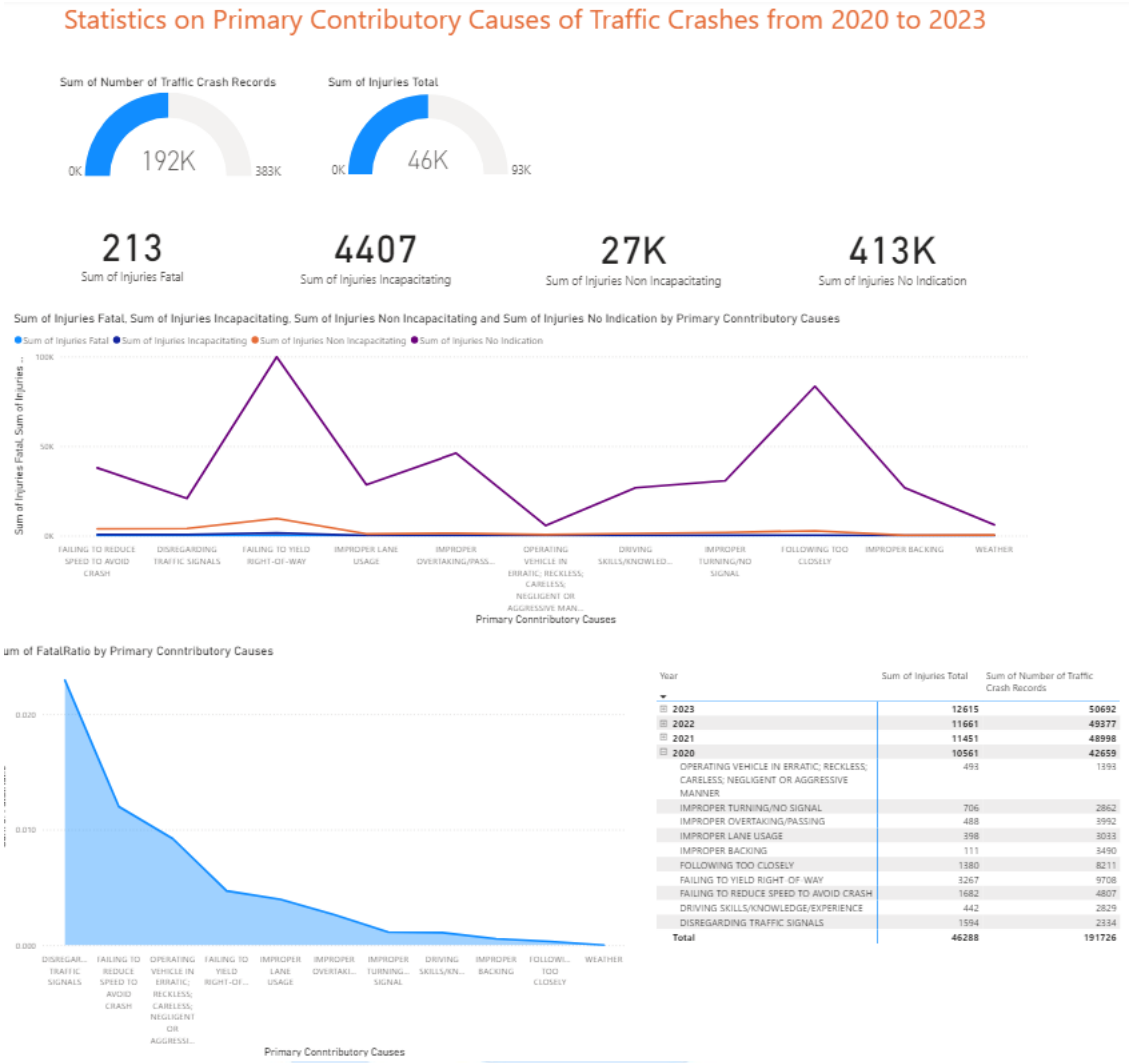


Bước 3: Tiến hành đổi tên cột dữ liệu và kiểu dữ liệu ở các cột cho phù hợp.



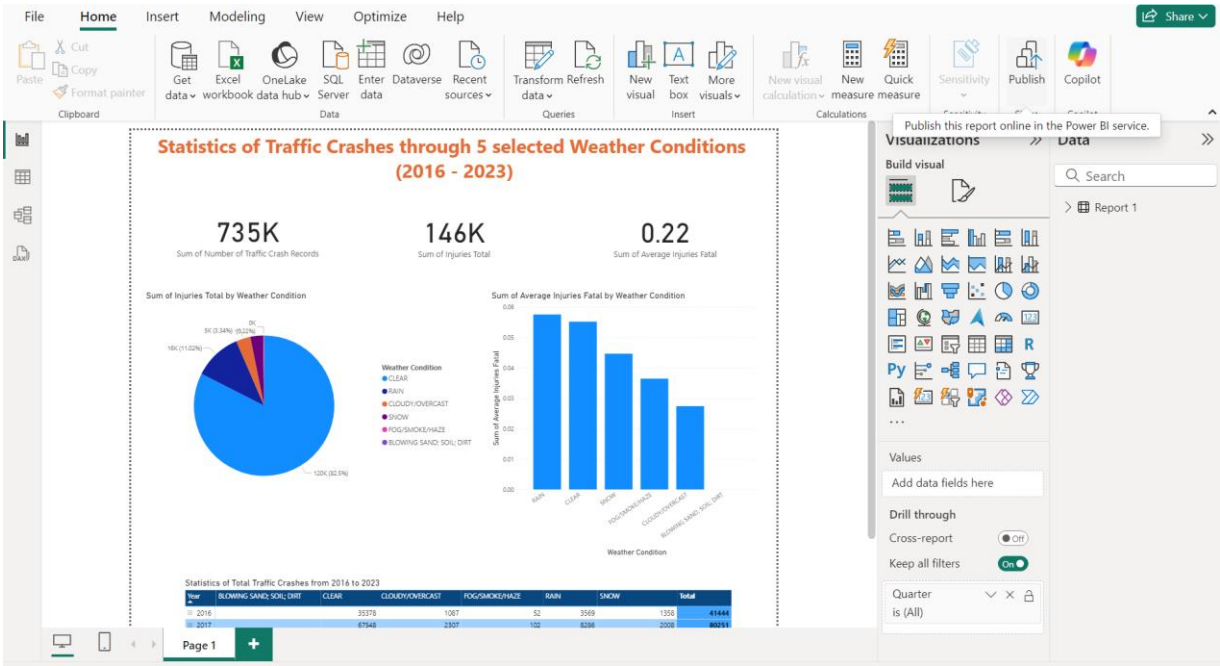
Sau đó nhấn “Close and Apply”.

Bước 4: Chọn các giao diện, biểu đồ để hiển thị kết quả thống kê. “Report 3” ở đây bao gồm một thành phần: Line Chart, Area Chart, Matrix, Funnel, Card, Gauge. Ở Matrix, ta có thể drill down năm để xem số vụ tai nạn và tổng số nạn nhân ở từng quý.

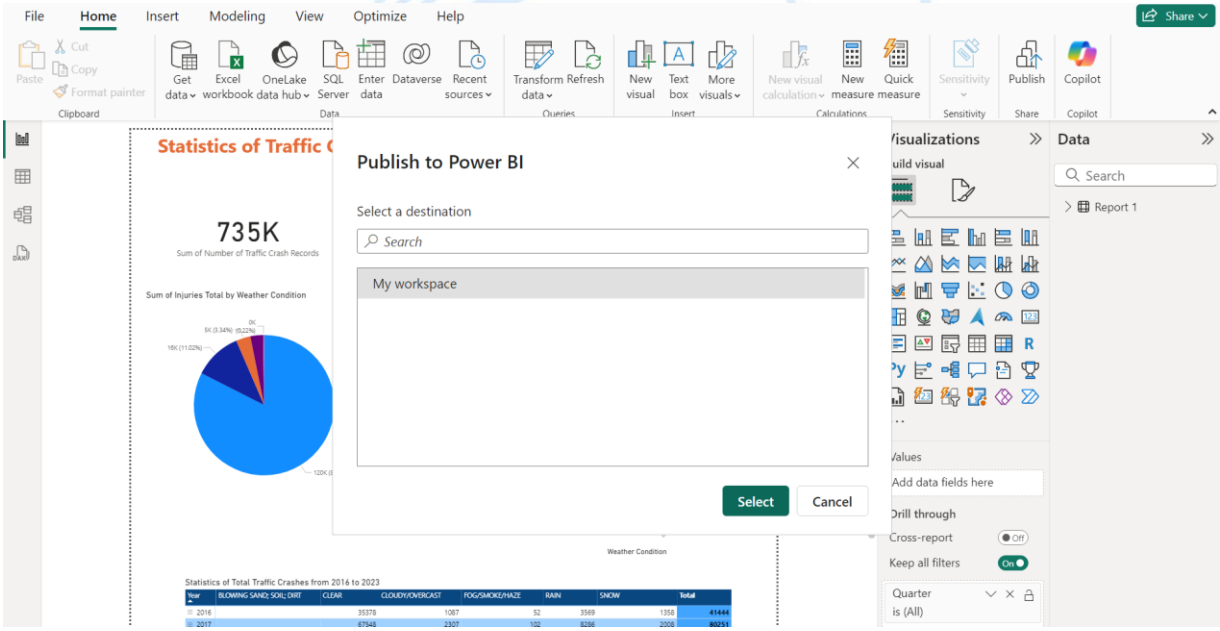


4.1.3 Public các report đã tạo

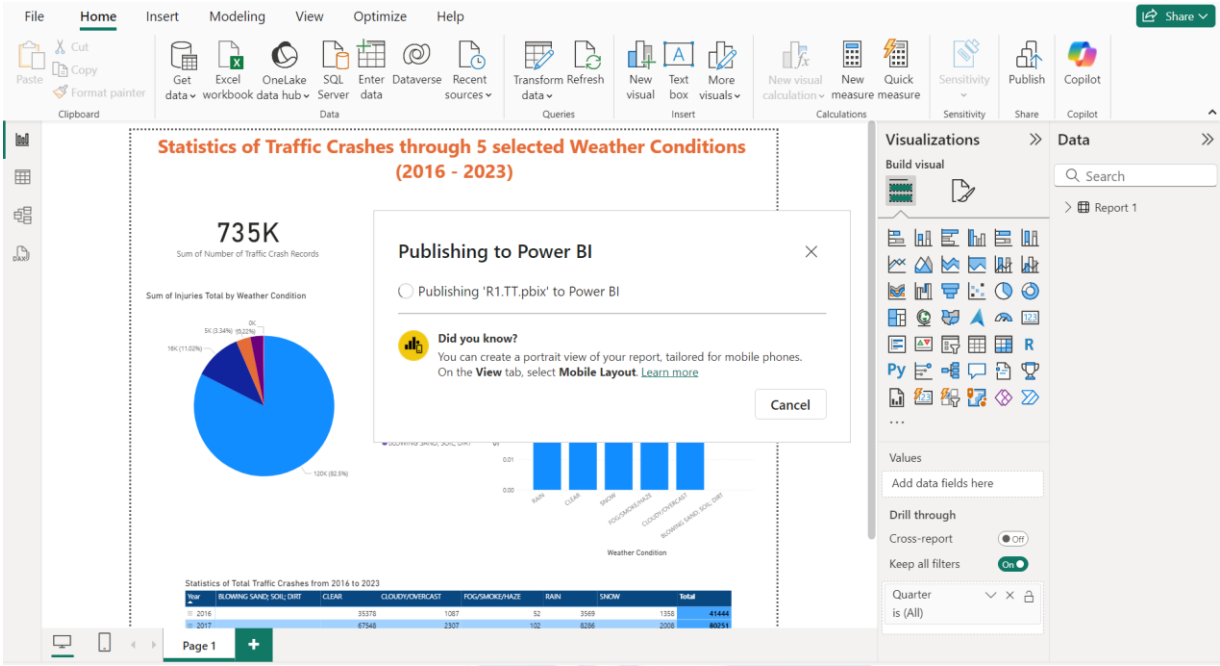
Bước 1: Tại thanh công cụ, chọn **Publish** để xuất bản online các report đã tạo.



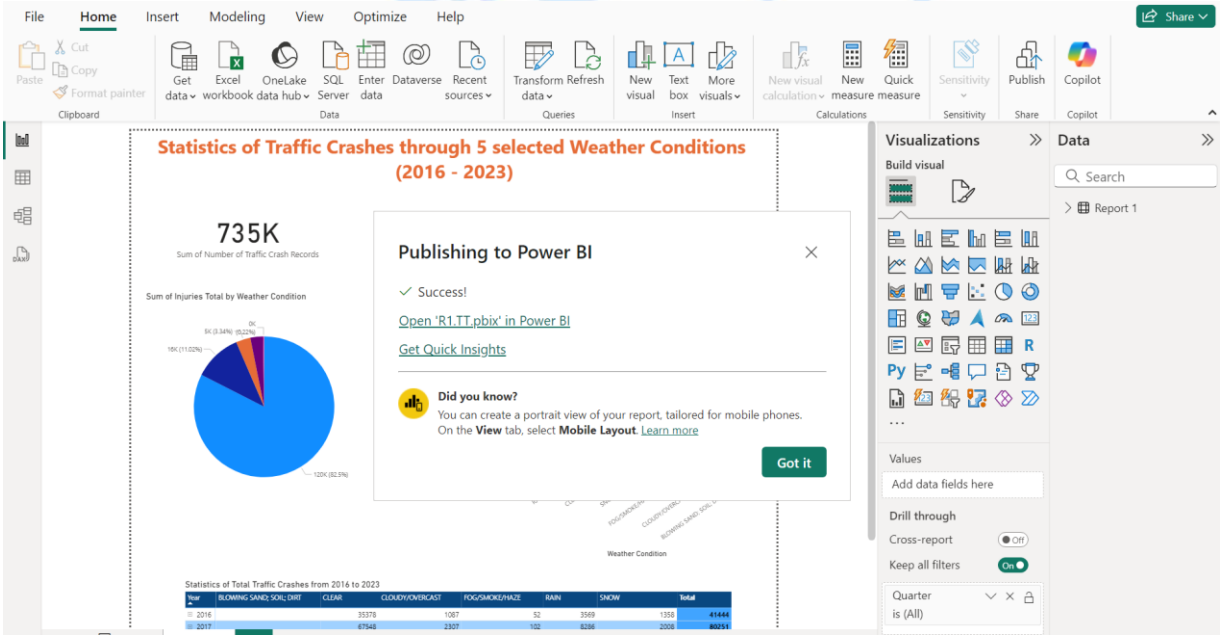
Bước 2: Tiếp theo, ta tiến hành chọn một địa điểm để tiến hành publish. Sau đó nhấn **Select**.



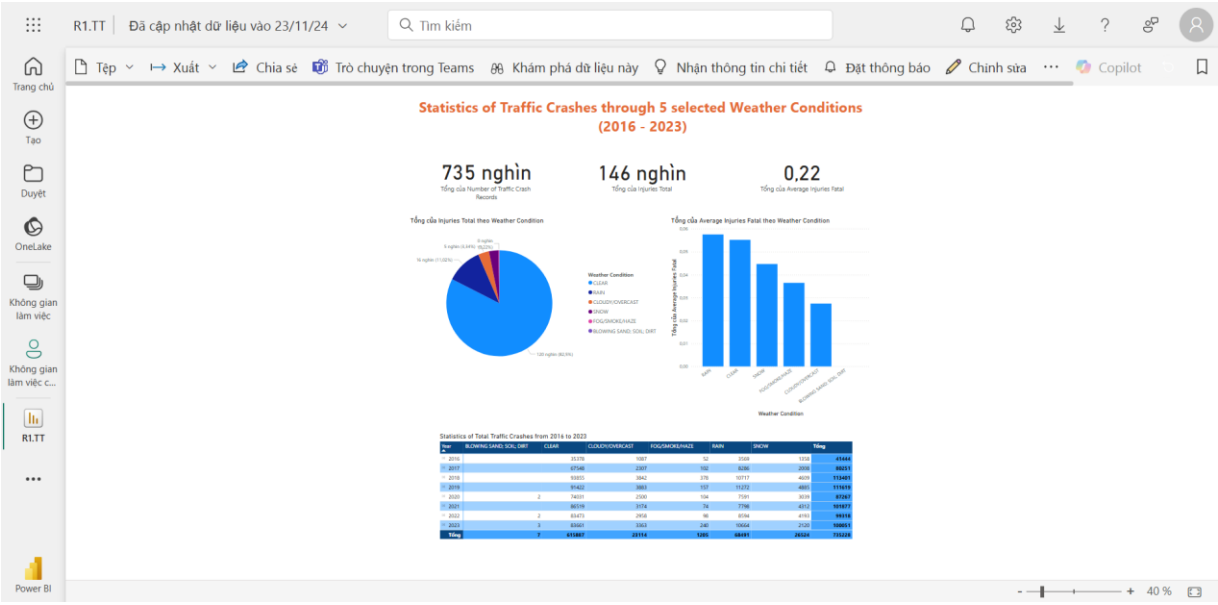
Bước 3: Hệ thống đang thực hiện quá trình publish các report lên server của PowerBI.



Bước 4: Hoàn thành publish report.



Ta có thể cập nhật, thao tác, xuất các báo cáo.



4.2 Tạo project SSRS trên Locker

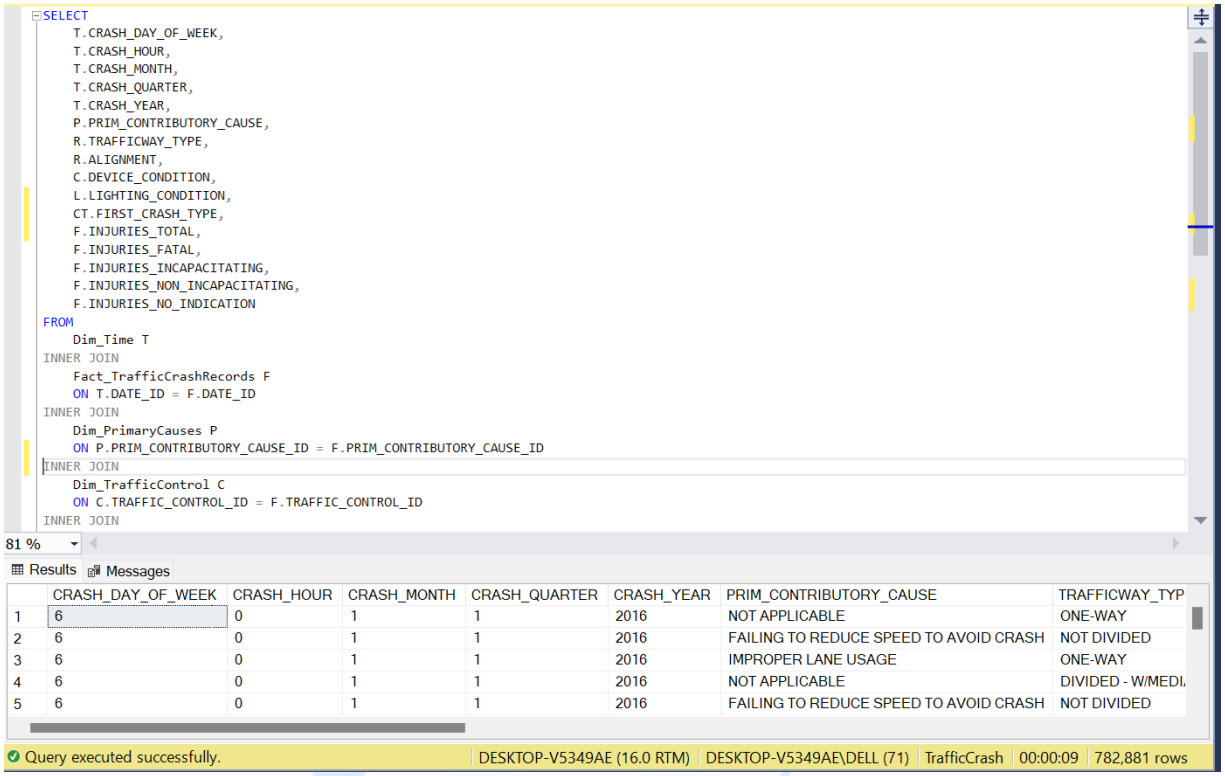
4.2.1 Trích xuất dữ liệu từ kho dữ liệu dưới dạng .csv

Bước 1: Kết nối tới **Database Engine** chứa database đã tạo ở quá trình SSIS.

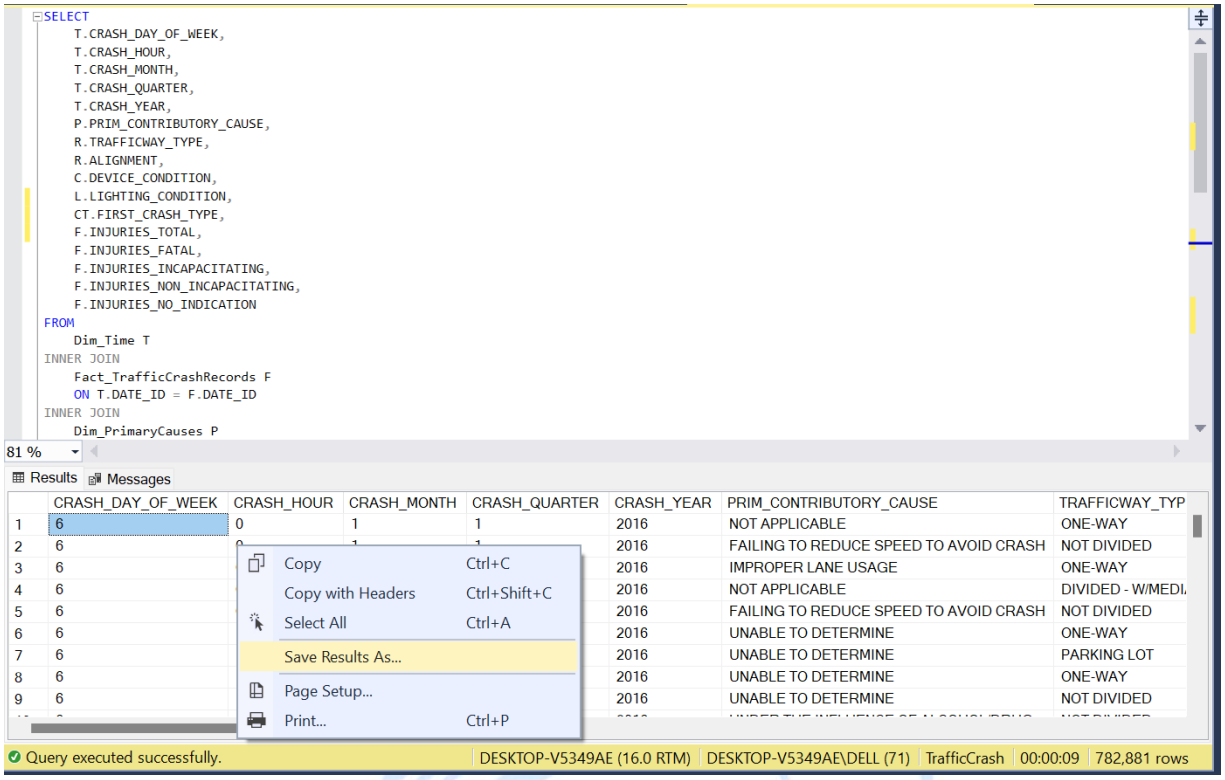
Bước 2: Thực thi câu lệnh như sau:

```
SELECT
    T.CRASH_DAY_OF_WEEK,
    T.CRASH_HOUR,
    T.CRASH_MONTH,
    T.CRASH_QUARTER,
    T.CRASH_YEAR,
    P.PRIM_CONTRIBUTORY_CAUSE,
    R.TRAFFICWAY_TYPE,
    R.ALIGNMENT,
    C.DEVICE_CONDITION,
    L.LIGHTING_CONDITION,
    CT.FIRST_CRASH_TYPE,
    F.INJURIES_TOTAL,
    F.INJURIES_FATAL,
    F.INJURIES_INCAPACITATING,
    F.INJURIES_NON_INCAPACITATING,
    F.INJURIES_NO_INDICATION
FROM
    Dim_Time T
INNER JOIN
    Fact_TrafficCrashRecords F
    ON T.DATE_ID = F.DATE_ID
INNER JOIN
    Dim_PrimaryCauses P
    ON P.PRIM_CONTRIBUTORY_CAUSE_ID = F.PRIM_CONTRIBUTORY_CAUSE_ID
INNER JOIN
    Dim_TrafficControl C
    ON C.TRAFFIC_CONTROL_ID = F.TRAFFIC_CONTROL_ID
INNER JOIN
    Dim_RoadAlignment R
    ON R.ROAD_ALIGNMENT_ID = F.ROAD_ALIGNMENT_ID
INNER JOIN
```

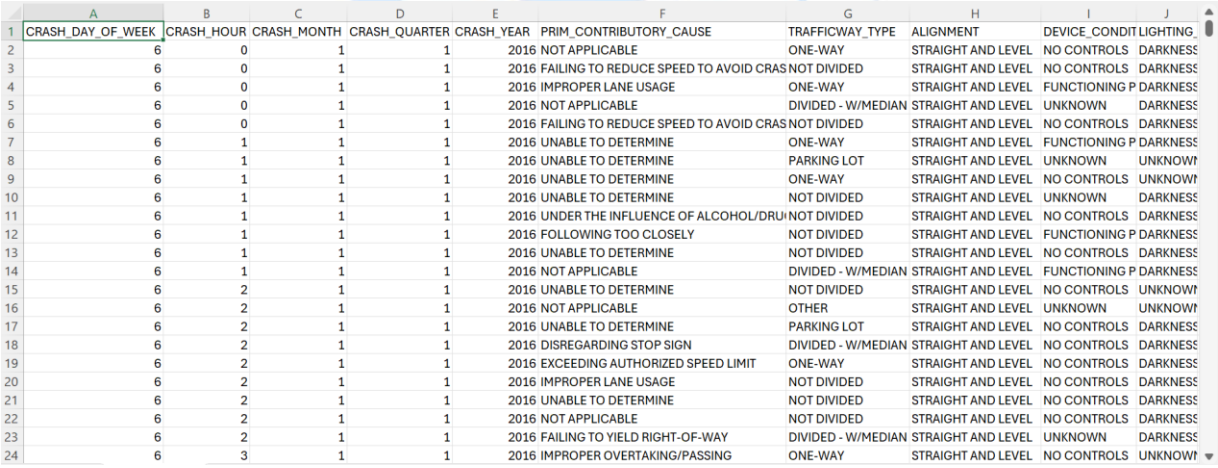
```
    Dim_LightingConditions L
    ON L.LIGHTING_CONDITION_ID = F.LIGHTING_CONDITION_ID
INNER JOIN
       Dim_FirstCrashTypes CT
       ON CT.FIRST_CRASH_TYPE_ID = F.FIRST_CRASH_TYPE_ID;
```



Bước 3: Nhấn chuột phải vào vùng kết quả, chọn **Save Result As**. Sau đó lựa chọn vị trí lưu và nhấn **Save** để lưu file .csv.



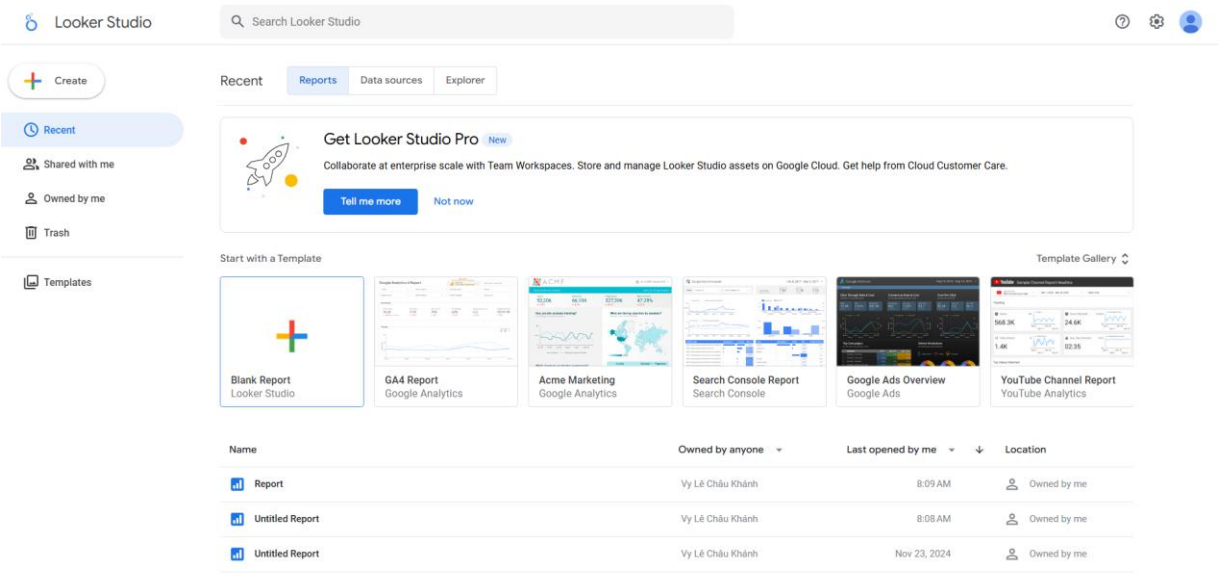
Bước 4: Kiểm tra lại file .csv đã lưu.



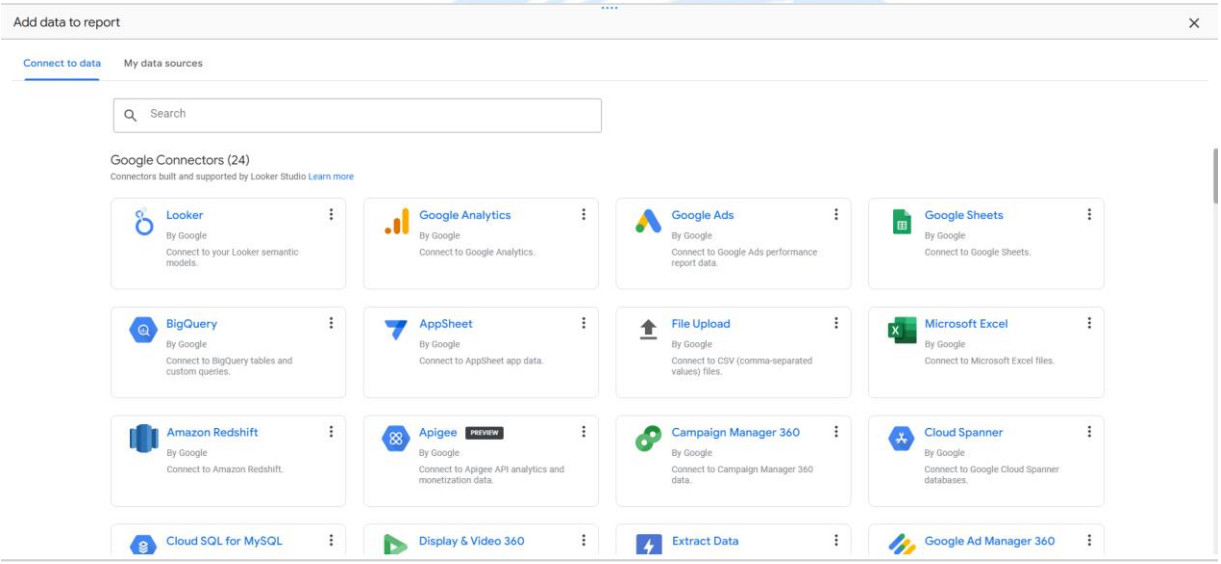
	A	B	C	D	E	F	G	H	I	J
1	CRASH_DAY_OF_WEEK	CRASH_HOUR	CRASH_MONTH	CRASH_QUARTER	CRASH_YEAR	PRIM_CONTRIBUTORY_CAUSE	TRAFFICWAY_TYPE	ALIGNMENT	DEVICE_CONDIT	LIGHTING_
2	6	0	1	1	2016	NOT APPLICABLE	ONE-WAY	STRAIGHT AND LEVEL	NO CONTROLS	DARKNESS
3	6	0	1	1	2016	FAILING TO REDUCE SPEED TO AVOID CRAS	NOT DIVIDED	STRAIGHT AND LEVEL	NO CONTROLS	DARKNESS
4	6	0	1	1	2016	IMPROPER LANE USAGE	ONE-WAY	STRAIGHT AND LEVEL	FUNCTIONING P	DARKNESS
5	6	0	1	1	2016	NOT APPLICABLE	DIVIDED - W/MEDIAN	STRAIGHT AND LEVEL	UNKNOWN	DARKNESS
6	6	0	1	1	2016	FAILING TO REDUCE SPEED TO AVOID CRAS	NOT DIVIDED	STRAIGHT AND LEVEL	NO CONTROLS	DARKNESS
7	6	1	1	1	2016	UNABLE TO DETERMINE	ONE-WAY	STRAIGHT AND LEVEL	FUNCTIONING P	DARKNESS
8	6	1	1	1	2016	UNABLE TO DETERMINE	PARKING LOT	STRAIGHT AND LEVEL	UNKNOWN	UNKNOWN
9	6	1	1	1	2016	UNABLE TO DETERMINE	ONE-WAY	STRAIGHT AND LEVEL	NO CONTROLS	UNKNOWN
10	6	1	1	1	2016	UNABLE TO DETERMINE	NOT DIVIDED	STRAIGHT AND LEVEL	UNKNOWN	DARKNESS
11	6	1	1	1	2016	UNDER THE INFLUENCE OF ALCOHOL/DRU	NOT DIVIDED	STRAIGHT AND LEVEL	NO CONTROLS	DARKNESS
12	6	1	1	1	2016	FOLLOWING TOO CLOSELY	NOT DIVIDED	STRAIGHT AND LEVEL	FUNCTIONING P	DARKNESS
13	6	1	1	1	2016	UNABLE TO DETERMINE	NOT DIVIDED	STRAIGHT AND LEVEL	NO CONTROLS	DARKNESS
14	6	1	1	1	2016	NOT APPLICABLE	DIVIDED - W/MEDIAN	STRAIGHT AND LEVEL	FUNCTIONING P	DARKNESS
15	6	2	1	1	2016	UNABLE TO DETERMINE	NOT DIVIDED	STRAIGHT AND LEVEL	NO CONTROLS	UNKNOWN
16	6	2	1	1	2016	NOT APPLICABLE	OTHER	STRAIGHT AND LEVEL	UNKNOWN	UNKNOWN
17	6	2	1	1	2016	UNABLE TO DETERMINE	PARKING LOT	STRAIGHT AND LEVEL	NO CONTROLS	DARKNESS
18	6	2	1	1	2016	DISREGARDING STOP SIGN	DIVIDED - W/MEDIAN	STRAIGHT AND LEVEL	NO CONTROLS	DARKNESS
19	6	2	1	1	2016	EXCEEDING AUTHORIZED SPEED LIMIT	ONE-WAY	STRAIGHT AND LEVEL	NO CONTROLS	DARKNESS
20	6	2	1	1	2016	IMPROPER LANE USAGE	NOT DIVIDED	STRAIGHT AND LEVEL	NO CONTROLS	DARKNESS
21	6	2	1	1	2016	UNABLE TO DETERMINE	NOT DIVIDED	STRAIGHT AND LEVEL	NO CONTROLS	DARKNESS
22	6	2	1	1	2016	NOT APPLICABLE	NOT DIVIDED	STRAIGHT AND LEVEL	NO CONTROLS	DARKNESS
23	6	2	1	1	2016	FAILING TO YIELD RIGHT-OF-WAY	DIVIDED - W/MEDIAN	STRAIGHT AND LEVEL	UNKNOWN	DARKNESS
24	6	3	1	1	2016	IMPROPER OVERTAKING/PASSING	ONE-WAY	STRAIGHT AND LEVEL	NO CONTROLS	UNKNOWN

4.2.2 Upload và lọc dữ liệu trên Locker Studio

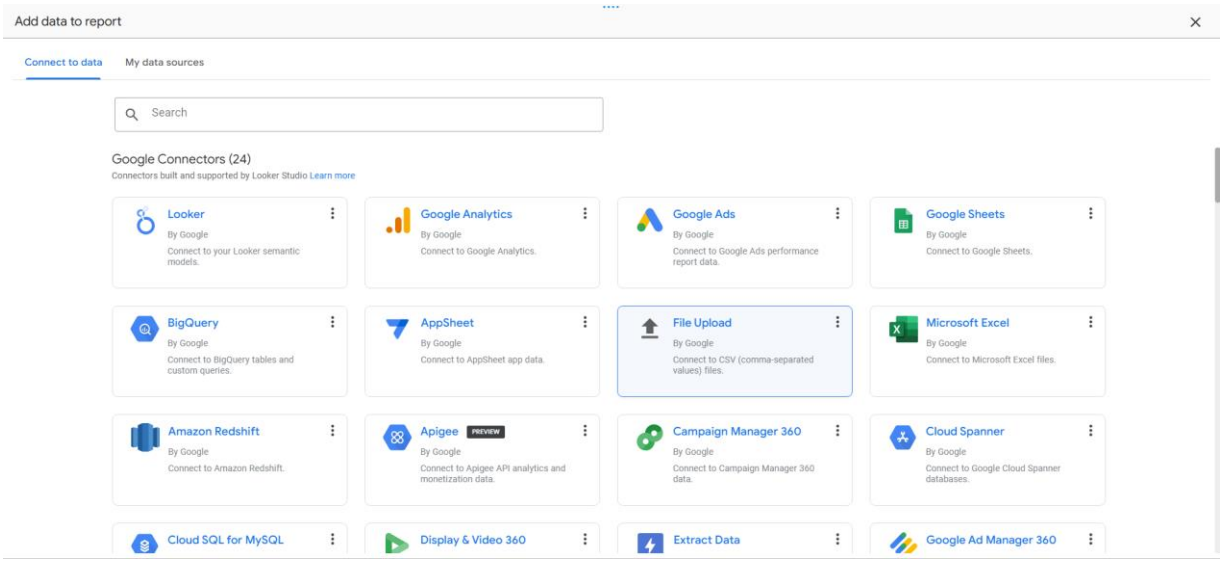
Bước 1: Truy cập vào https://lookerstudio.google.com/. Thực hiện đăng nhập. Sau đó, chọn **Blank Report**.



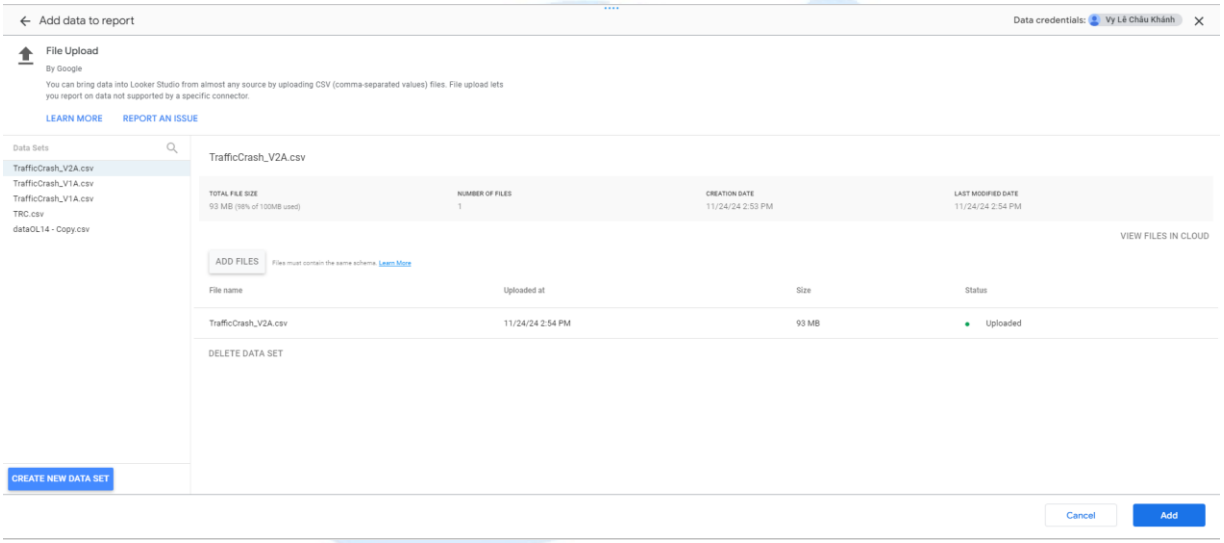
Bước 2: Tiếp theo, một cửa sổ yêu cầu thêm nguồn dữ liệu vào báo cáo sẽ hiện lên.



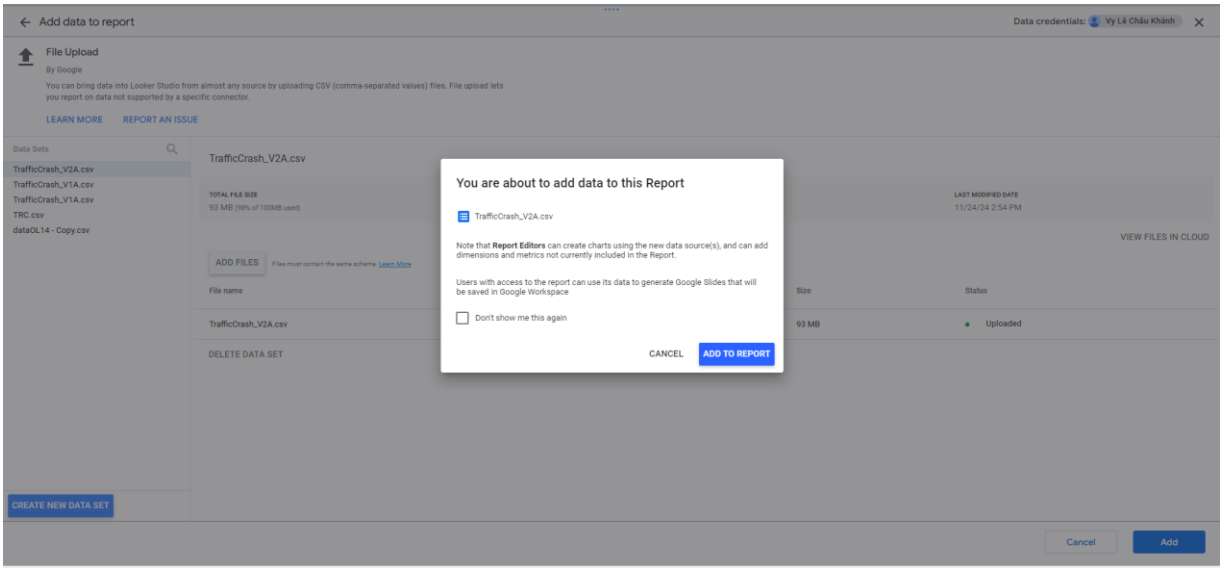
Bước 3: Sau đó, ta chọn **File Upload**.



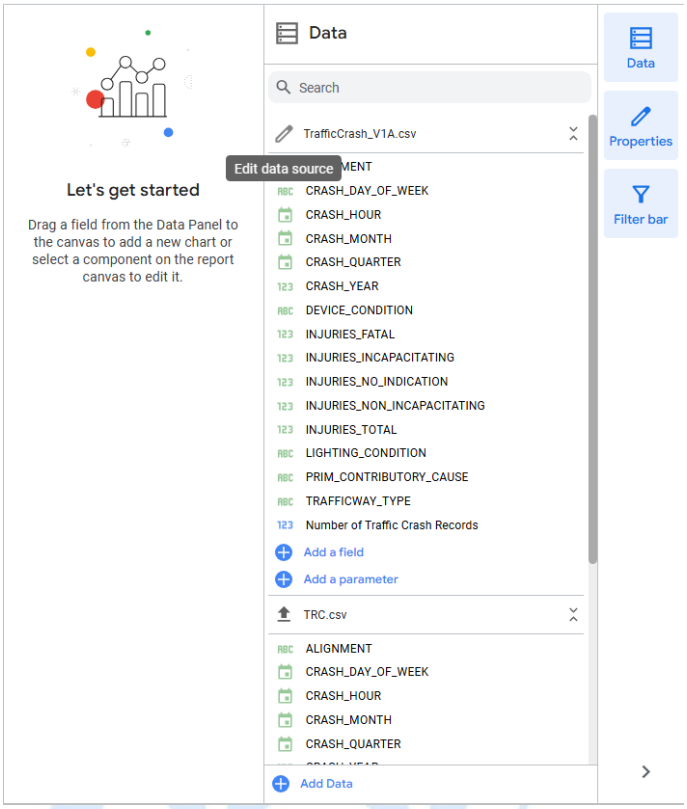
Bước 4: Tải lên file .csv đã lưu. Chọn **Add** để hoàn tất quá trình file.



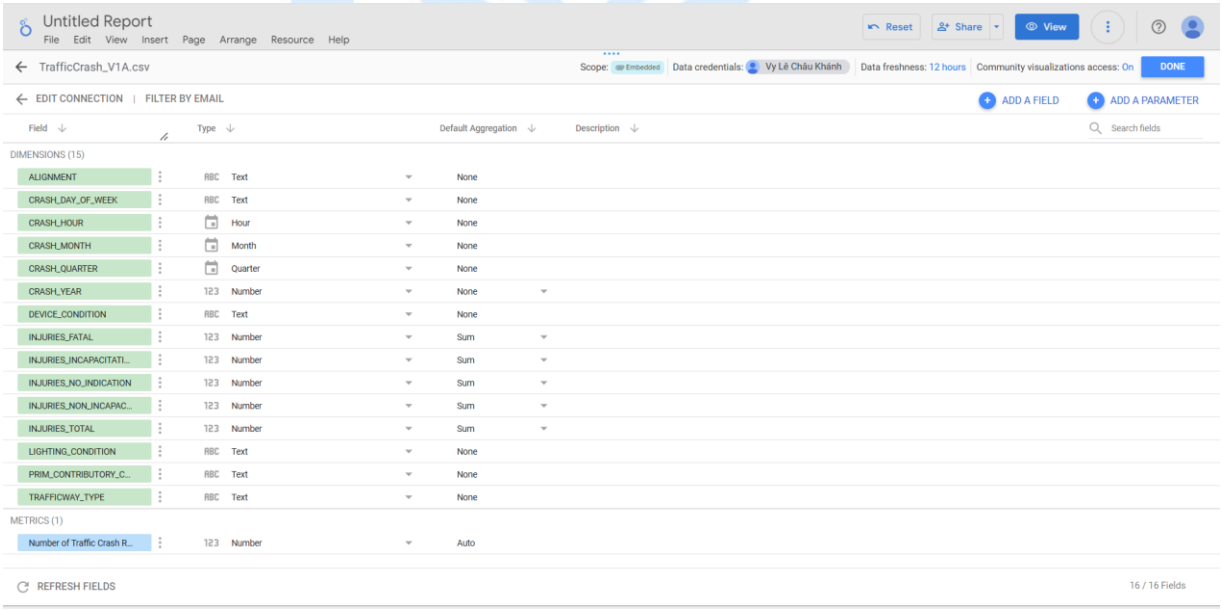
Bước 5: Một cửa sổ nhỏ sẽ hiện lên. Chọn **Add to report**.



Bước 6: Ở cửa sổ làm việc chính của Looker Studio, ta có thể chọn **Edit data source** để chỉnh sửa dữ liệu.



Sau đó nhấn **Done** để hoàn thành.



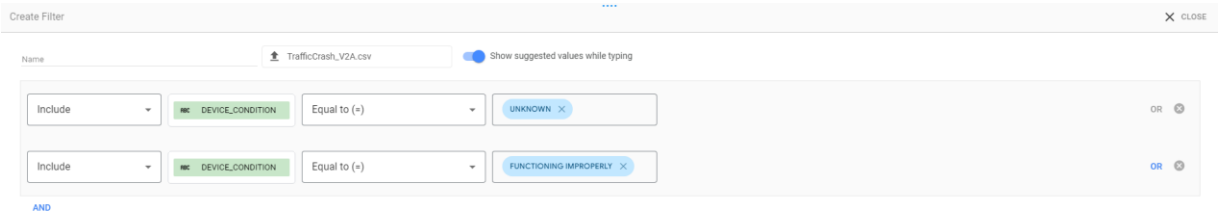
4.2.3 Thực hiện tạo báo cáo trên Locker Studio

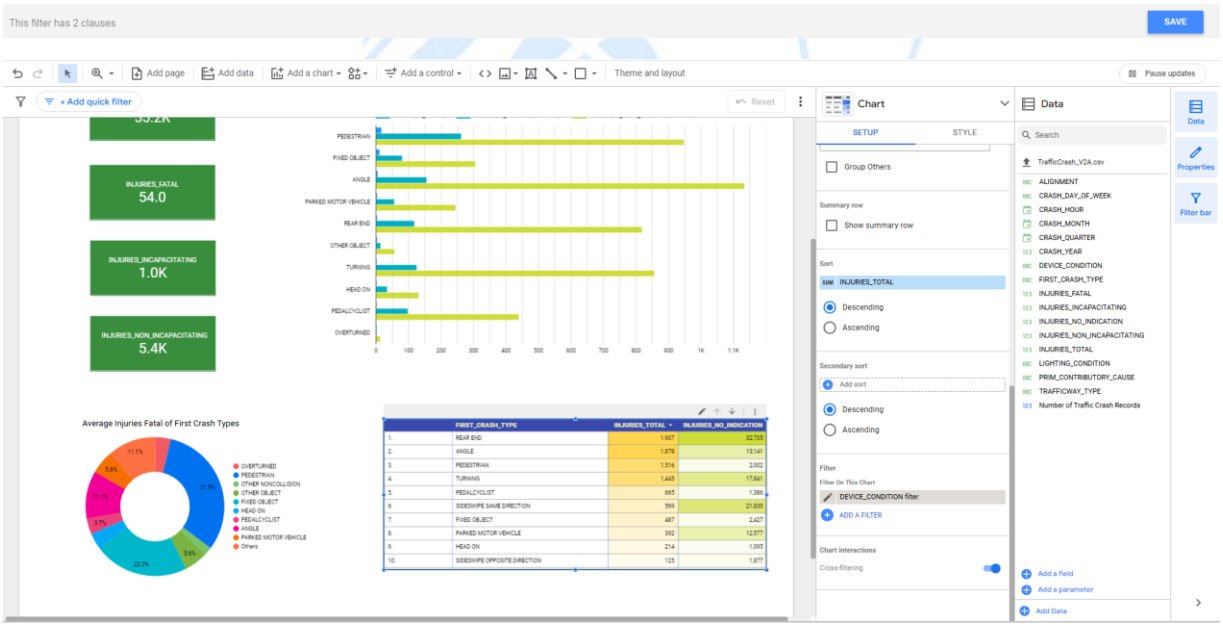
4.2.3.1 Report 1

Nội dung: Thống kê thông tin các loại tác nhân va chạm giao thông cùng các thông số thương vong trong điều kiện thiết bị điều khiển bị lỗi, bị hư ('FUNCTIONING IMPROPERLY') hoặc không rõ hoạt động các thiết bị điều khiển ('UNKNOWN').

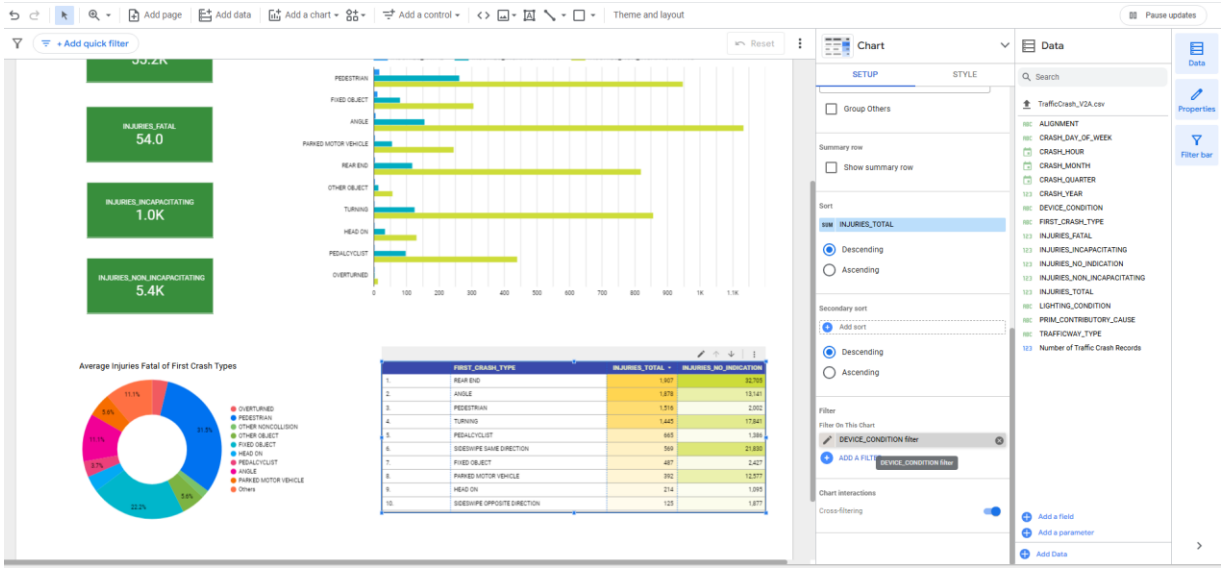
Bước 1: Thiết kế report, kéo các thuộc tính cần để tạo báo biểu, sau đó chọn kiểu biểu đồ muốn trực quan dữ liệu.

Bước 2: Tạo filter để chọn các vụ tai nạn có tác nhân va chạm là 'FUNCTIONING IMPROPERLY' hoặc 'UNKNOWN'.

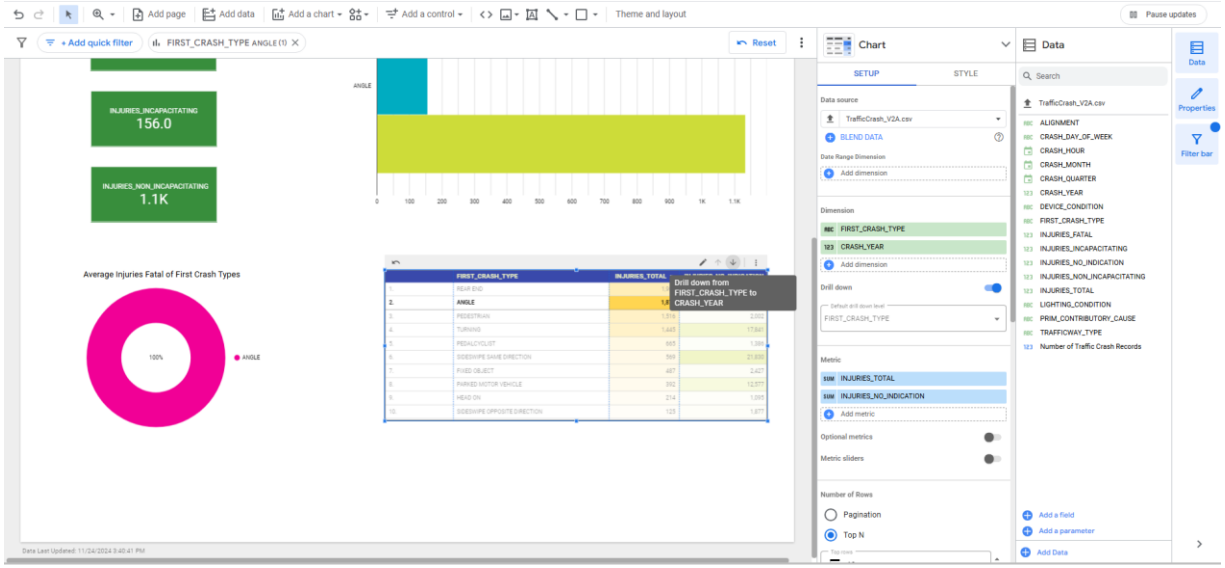




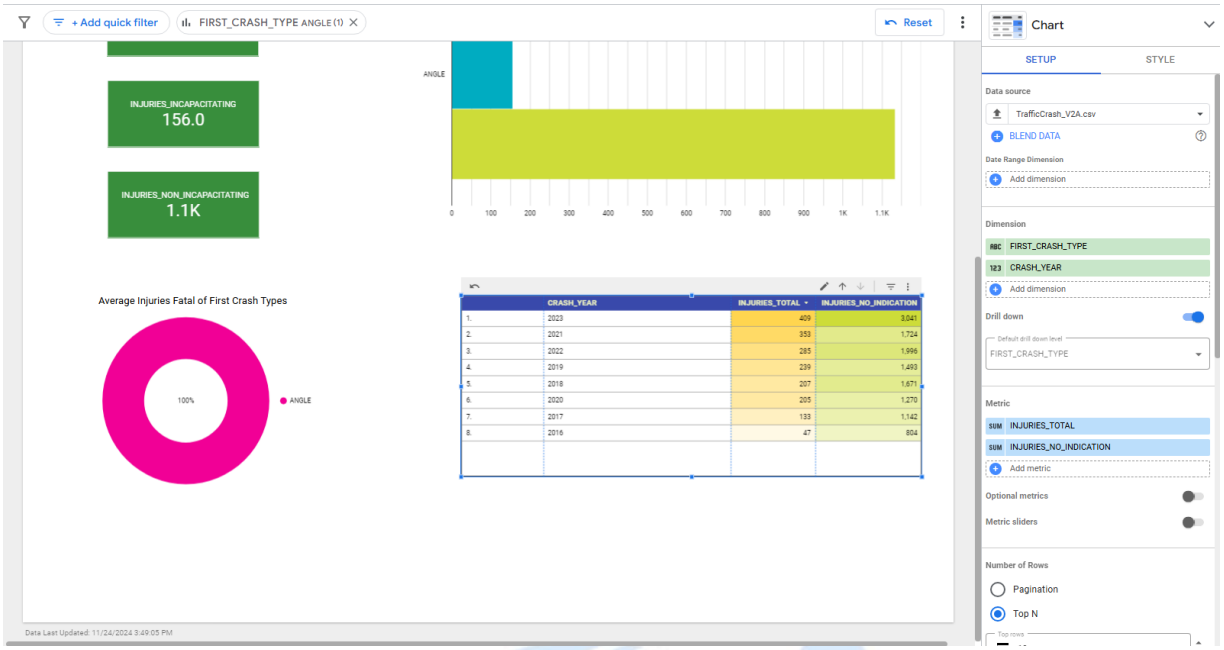
Bước 3: Thêm filter vào các biểu đồ.



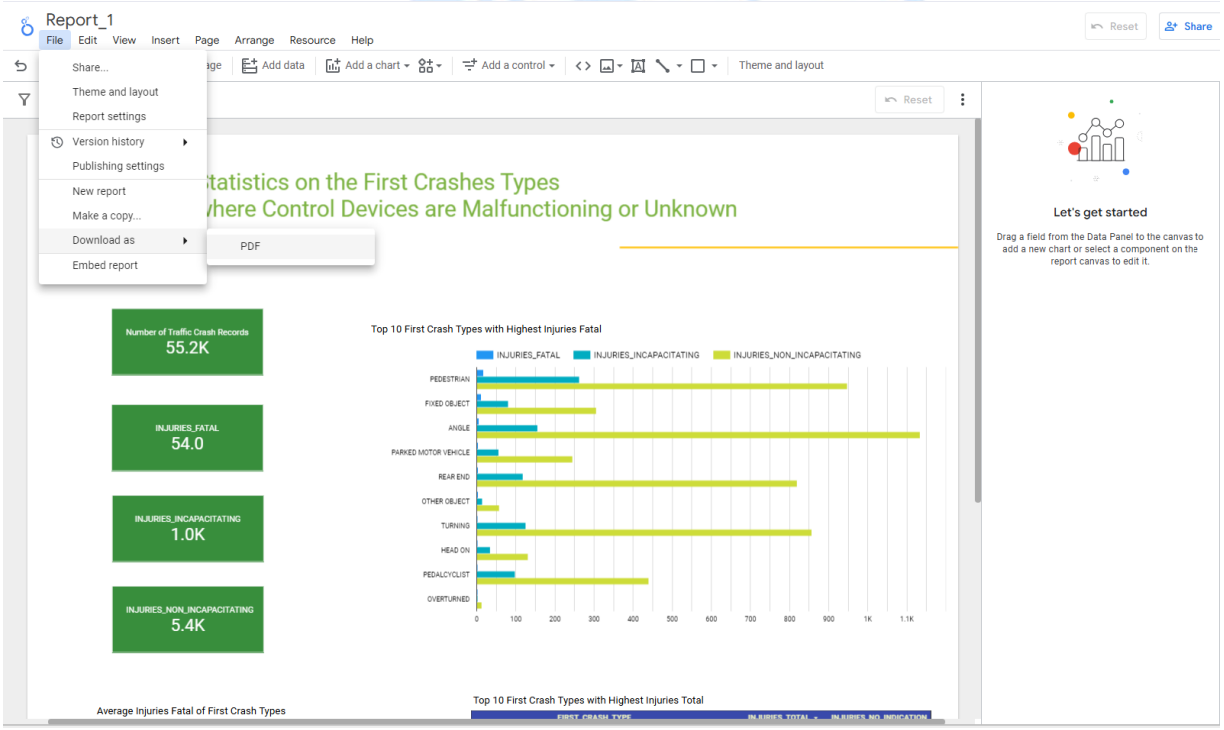
Bước 4: Ở matrix, ta thêm drill down để xem thông số thống kê cho từng loại tác nhân ở từng năm.



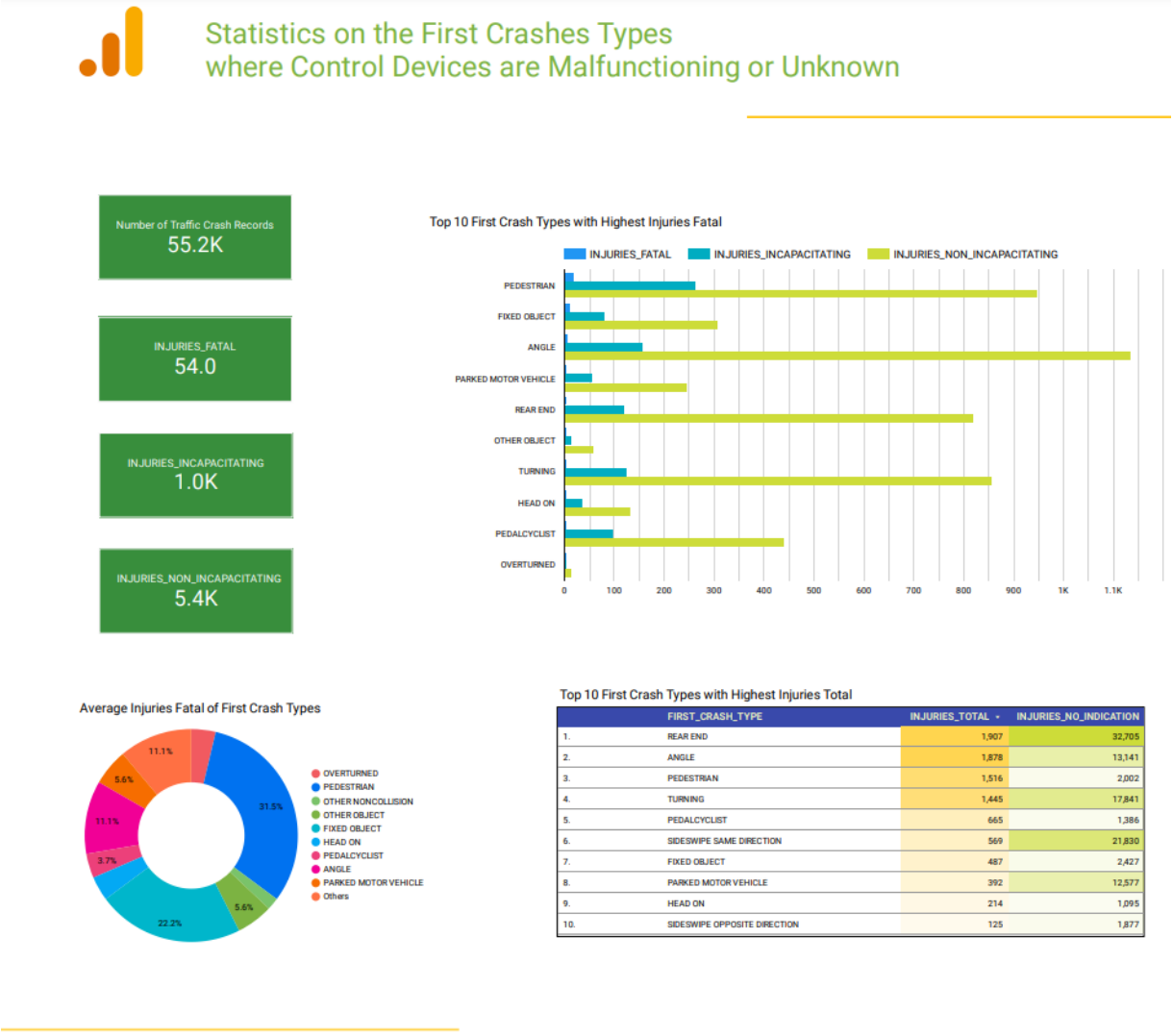
Kết quả drill down:



Bước 5: Để xuất báo cáo dưới dạng PDF, ta chọn File → Download as → PDF.



Bước 6: Xem báo cáo dưới dạng PDF đã tải về.



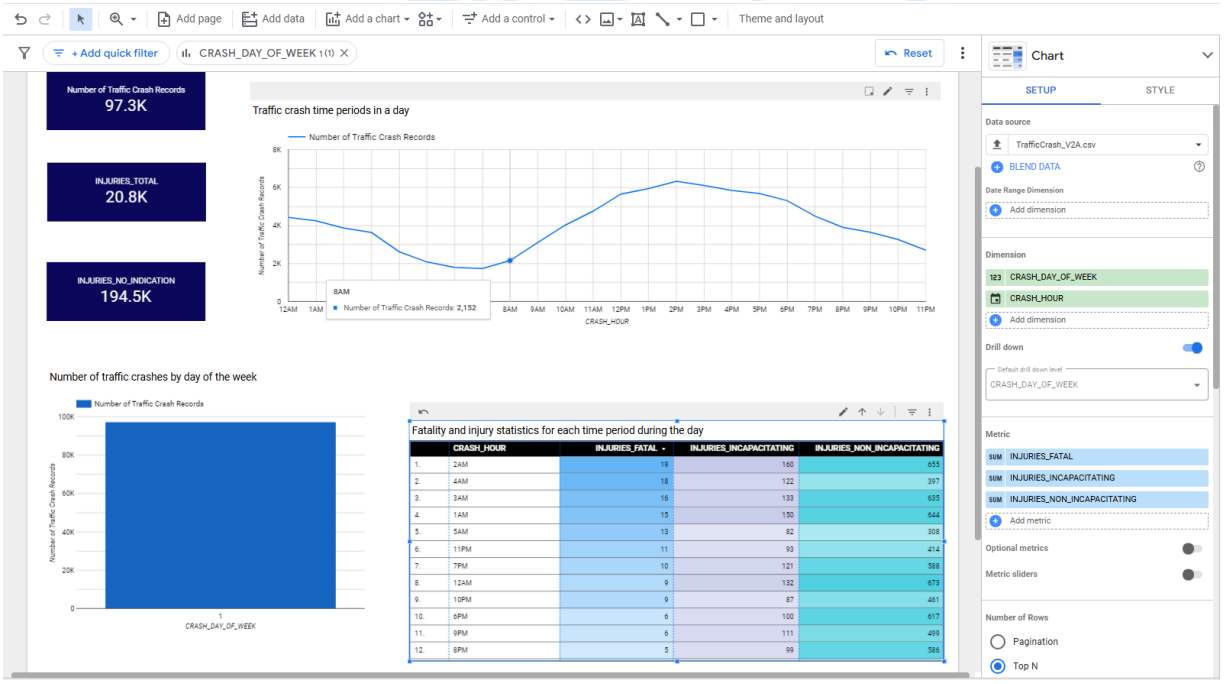
4.2.3.2 Report 2

Nội dung: Thống kê thông tin về các khoảng thời gian xảy ra tai nạn giao thông ở các khung giờ và các ngày trong tuần.

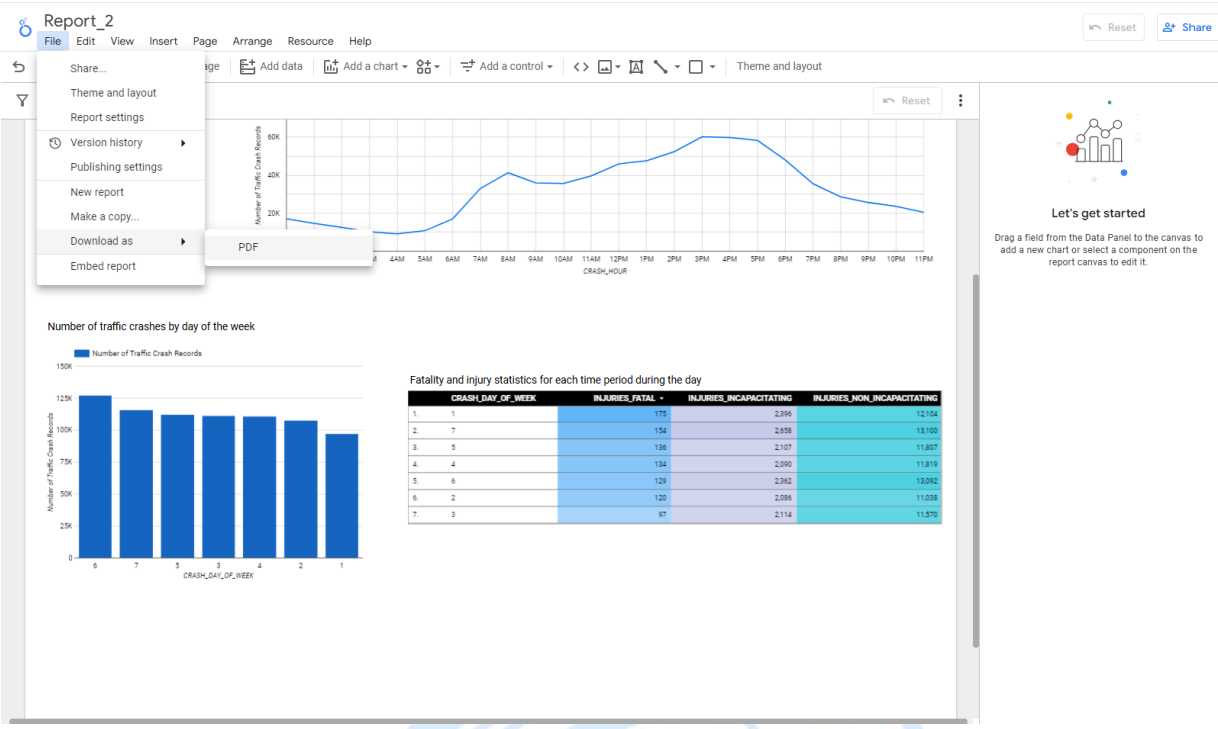
Bước 1: Thiết kế report, kéo các thuộc tính cần để tạo báo biểu, sau đó chọn kiểu biểu đồ muốn trực quan dữ liệu.

Ở matrix, ta có thể drill down để xem thống kê số lượng thương vong ở từng khung giờ của các ngày trong tuần.

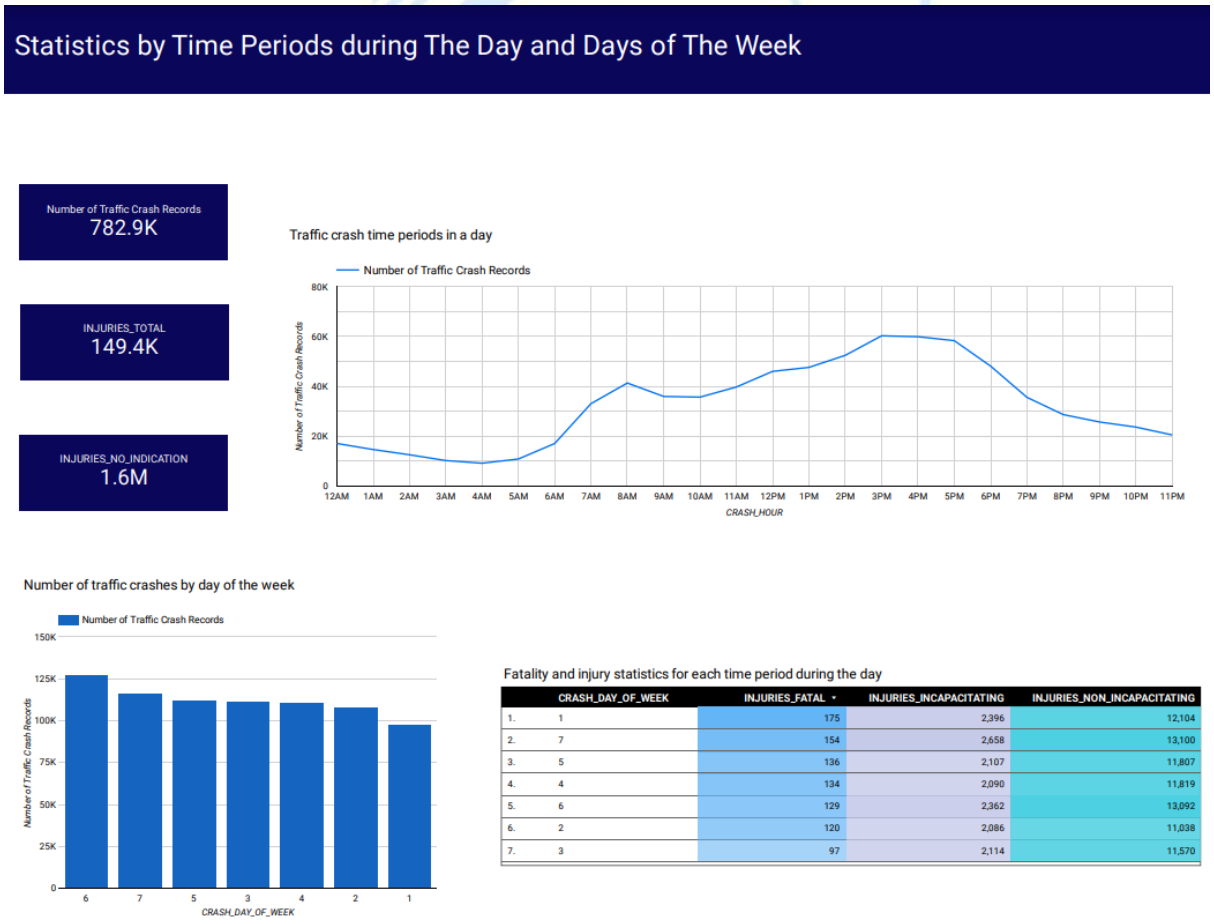
Kết quả khi drill down:



Bước 2: Để xuất báo cáo dưới dạng PDF, ta chọn File → Download as → PDF.



Bước 3: Xem báo cáo dưới dạng PDF đã tải về.

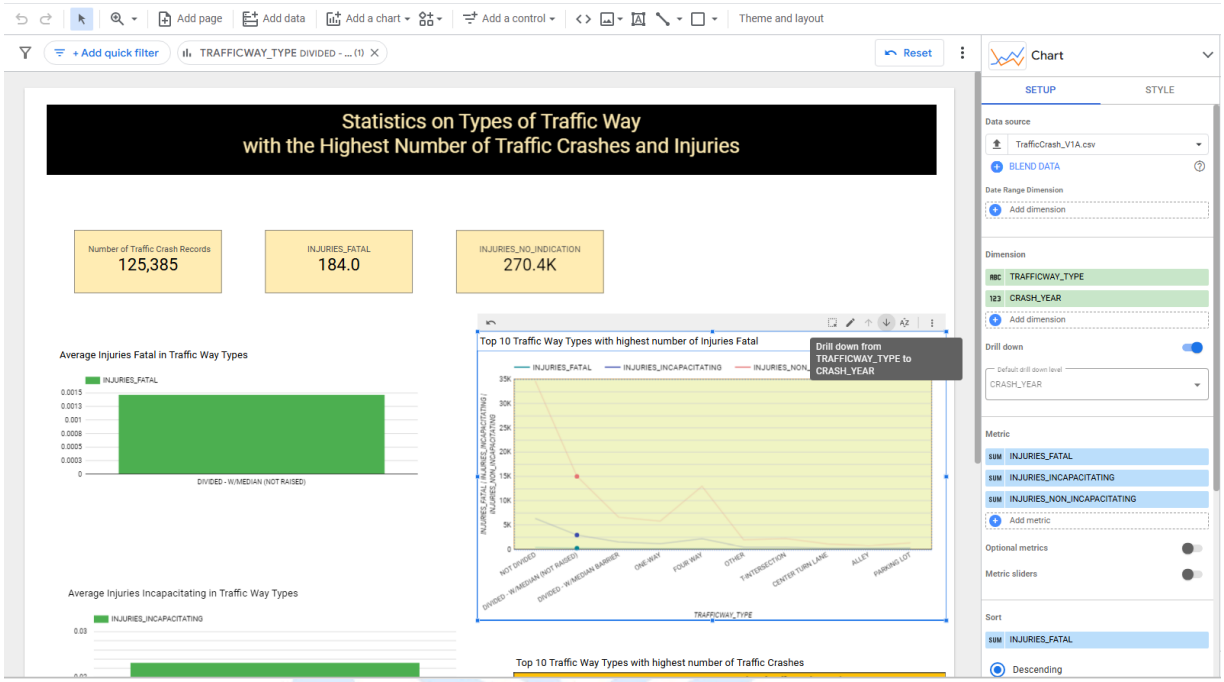


4.2.3.3 Report 3

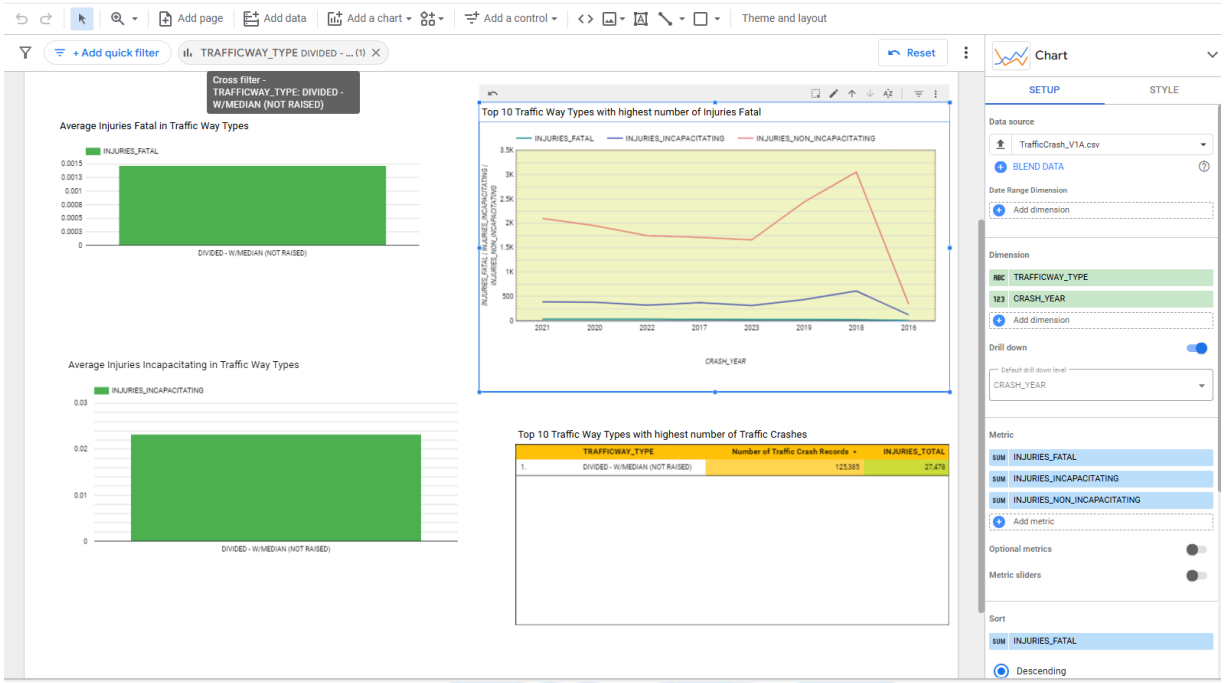
Nội dung: Thống kê top những loại đường giao thông xảy ra nhiều vụ va chạm nhất và số lượng thương vong của các loại.

Bước 1: Thiết kế report, kéo các thuộc tính cần để tạo báo biểu, sau đó chọn kiểu biểu đồ muốn trực quan dữ liệu.

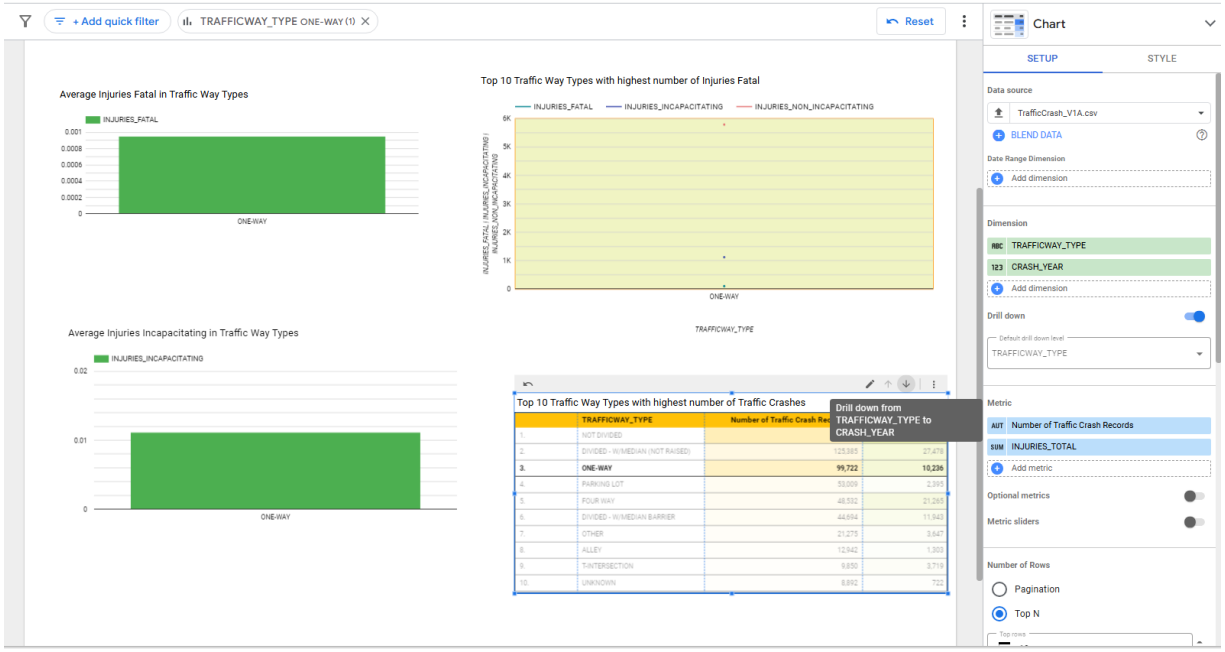
Ở biểu đồ đường, ta có thể drill down để xem thống kê số lượng thương vong ở từng loại đường theo năm.



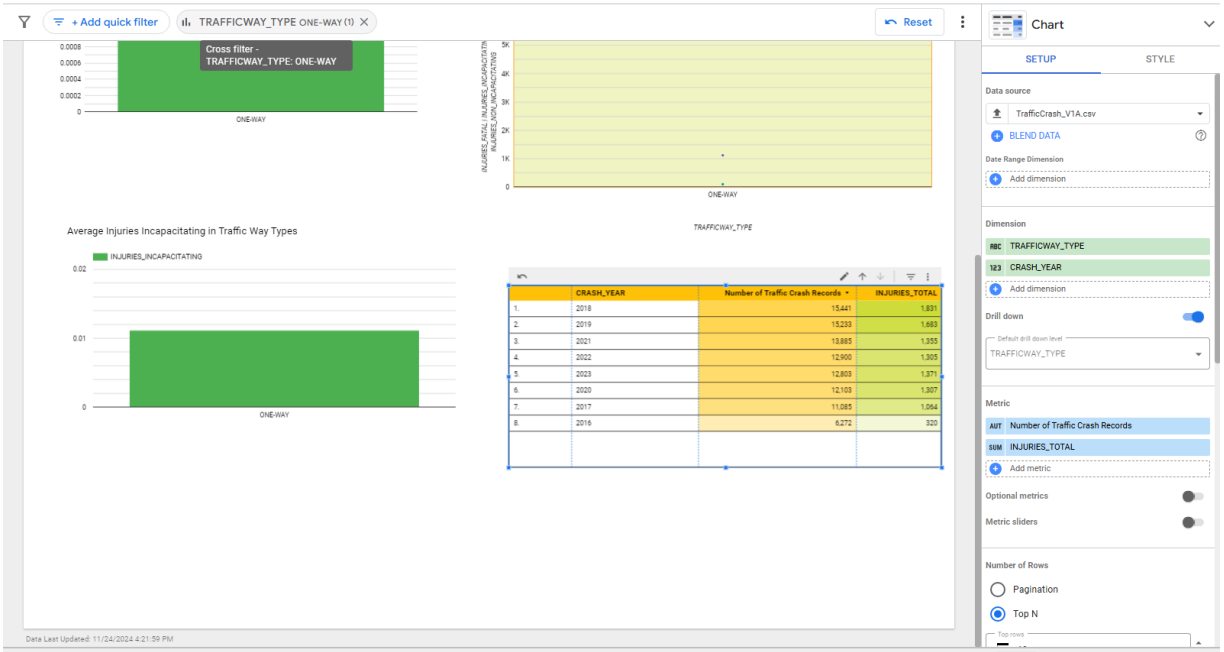
Kết quả sao khi drill down.



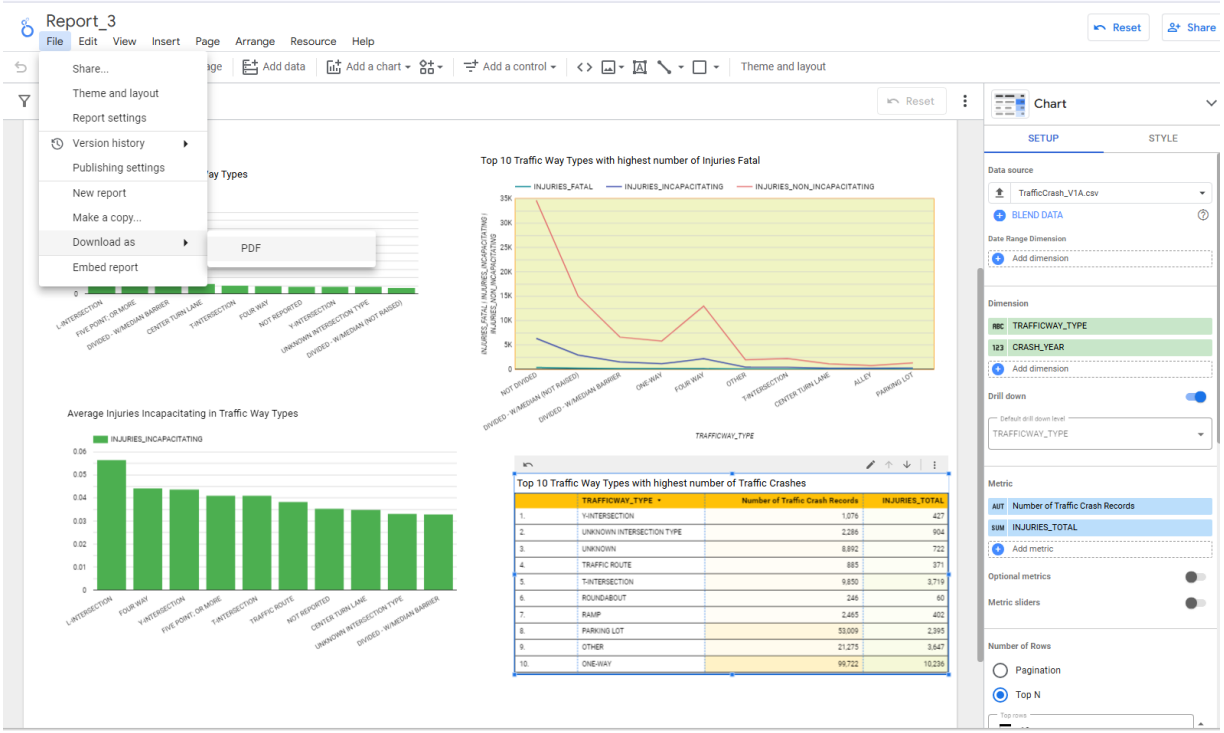
Ở matrix, ta cũng có thể drill down để xem thống kê ở từng loại đường cho từng năm.



Kết quả khi drill down.



Bước 2: Để xuất báo cáo dưới dạng PDF, ta chọn File → Download as → PDF.

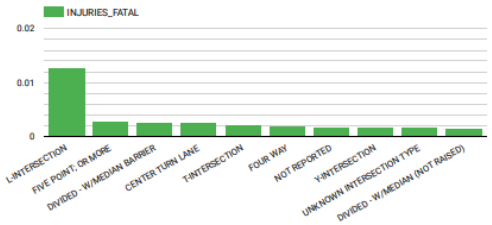


Bước 3: Xem báo cáo dưới dạng PDF đã tải về.

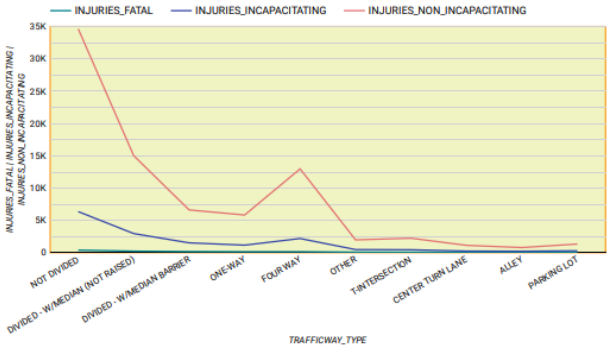
Statistics on Types of Traffic Way with the Highest Number of Traffic Crashes and Injuries

Number of Traffic Crash Records	INJURIES_FATAL	INJURIES_NO_INDICATION
782,881	945.0	1.6M

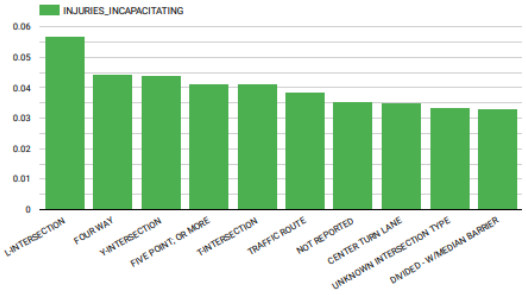
Average Injuries Fatal in Traffic Way Types



Top 10 Traffic Way Types with highest number of Injuries Fatal



Average Injuries Incapacitating in Traffic Way Types



Top 10 Traffic Way Types with highest number of Traffic Crashes

	TRAFFICWAY_TYPE	Number of Traffic Crash Records	INJURIES_TOTAL
1.	NOT DIVIDED	341,614	61,641
2.	DIVIDED - W/MEDIAN (NOT RAISED)	125,385	27,478
3.	ONE-WAY	99,722	10,236
4.	PARKING LOT	53,009	2,395
5.	FOUR WAY	48,532	21,265
6.	DIVIDED - W/MEDIAN BARRIER	44,694	11,943
7.	OTHER	21,275	3,647
8.	ALLEY	12,942	1,303
9.	T-INTERSECTION	9,850	3,719
10.	UNKNOWN	8,892	722

CHƯƠNG 5: QUÁ TRÌNH KHAI THÁC DỮ LIỆU (DATA MINING)

5.1 Mô tả bộ dữ liệu

5.1.1 Tổng quan dữ liệu

- Cơ sở dữ liệu Chicago Traffic Crashes là bộ dữ liệu chứa thông tin về các vụ va chạm giao thông ở Chicago – Mỹ. Bộ dữ liệu bao gồm các thuộc tính như ID, Ngày xảy ra va chạm, Điều kiện thời tiết, Điều kiện ánh sáng, Tác nhân trực tiếp gây ra va chạm, Nguyên nhân, Các mức độ thiệt hại, Số lượng tử vong, Số lượng bị thương nặng và nhiều thuộc tính khác.
- Nguồn dữ liệu: https://www.kaggle.com/datasets/anoopjohny/traffic-crashes-crashes

5.1.2 Mô tả bộ dữ liệu

Mô tả 48 thuộc tính bộ dữ liệu

STT	Tên thuộc tính	Kiểu	Ý nghĩa
1	CRASH_RECORD_ID	int	Mã định danh duy nhất cho mỗi bản ghi tai nạn.
2	CRASH_DATE_EST_I	boolean	Ngày dự kiến xảy ra tai nạn.
3	CRASH_DATE	timestamp	Ngày tai nạn xảy ra.
4	POSTED_SPEED_LIMIT	int	Giới hạn tốc độ được quy định tại địa điểm xảy ra tai nạn.
5	TRAFFIC_CONTROL_DEVICE	varchar	Thiết bị kiểm soát giao thông tại địa điểm xảy ra tai nạn.
6	DEVICE_CONDITION	varchar	Mô tả trạng thái hoặc điều kiện của các thiết bị kiểm soát giao thông tại thời điểm xảy ra tai nạn
7	WEATHER_CONDITION	varchar	Điều kiện thời tiết tại thời điểm xảy ra tai nạn.
8	LIGHTING_CONDITION	varchar	Điều kiện chiếu sáng tại thời điểm xảy ra tai nạn.
9	FIRST_CRASH_TYPE	varchar	Loại sự cố chính.
10	TRAFFICWAY_TYPE	varchar	Loại đường giao thông nơi xảy ra sự cố.
11	LANE_CNT	int	Số làn đường tại địa điểm xảy ra tai nạn.
12	ALIGNMENT	varchar	Cách mà đường hoặc tuyến đường được bố trí hoặc thiết kế theo một chiều cụ thể (STRAIGHT AND LEVEL (Thẳng và bằng phẳng), CURVE ON GRADE (Đường cong có độ dốc), CURVE, LEVEL (cong, bằng phẳng))

13	ROADWAY_SURFACE_COND	varchar	Tình trạng mặt đường.
14	ROAD_DEFECT	varchar	Những khiếm khuyết trên đường.
15	REPORT_TYPE	varchar	Loại báo cáo sự cố.
16	CRASH_TYPE	varchar	Loại tai nạn
17	INTERSECTION_RELATED_I	boolean	Vụ va chạm có liên quan đến ngã tư hay không.
18	NOT_RIGHT_OF_WAY_I	boolean	Chỉ báo vi phạm quyền ưu tiên đi trước.
19	HIT_AND_RUN_I	boolean	Chỉ báo va chạm rồi bỏ chạy.
20	DAMAGE	varchar	Mức độ thiệt hại.
21	DATE_POLICE_NOTIFIED	timestamp	Ngày cảnh sát được thông báo.
22	PRIM_CONTRIBUTORY_CAUSE	varchar	Nguyên nhân trực tiếp và quan trọng nhất dẫn đến vụ tai nạn.
23	SEC_CONTRIBUTORY_CAUSE	varchar	Nguyên nhân thứ hai của vụ tai nạn.
24	STREET_NO	int	Số nhà hoặc số đường nơi tai nạn xảy ra
25	STREET_DIRECTION	varchar(1)	Hướng đi của đường, chẳng hạn như N (Bắc), S (Nam), E (Đông), hoặc W (Tây).
26	STREET_NAME	varchar	Tên con đường nơi tai nạn xảy ra
27	BEAT_OF_OCCURRENCE	int	Cảnh sát phong tỏa nơi xảy ra vụ tai nạn.
28	PHOTOS_TAKEN_I	boolean	Có ảnh được chụp ở hiện trường.
29	STATEMENTS_TAKEN_I	boolean	Cho biết liệu có lời khai nào được ghi nhận tại hiện trường vụ tai nạn hay không.
30	DOORING_I	boolean	Chỉ ra liệu vụ tai nạn có liên quan đến sự cố mở cửa hay không
31	WORK_ZONE_I	boolean	Cho biết liệu có khu vực làm việc tại địa điểm xảy ra tai nạn hay không.
32	WORK_ZONE_TYPE	varchar	Mô tả loại khu vực làm việc nếu có.
33	WORKERS_PRESENT_I	boolean	Cho biết liệu công nhân có mặt tại địa điểm xảy ra tai nạn hay không.
34	NUM_UNITS	int	Số lượng phương tiện hoặc đơn vị tham gia vào vụ tai nạn.
35	MOST_SEVERE_INJURY	varchar	Chấn thương nghiêm trọng nhất được báo cáo trong vụ tai nạn
36	INJURIES_TOTAL	int	Tổng số lượng chấn thương trong vụ tai nạn

37	INJURIES_FATAL	int	Số lượng chấn thương dẫn đến cái chết
38	INJURIES_INCAPACITATING	int	Số lượng chấn thương khiến nạn nhân không thể thực hiện các hoạt động bình thường trong một khoảng thời gian
39	INJURIES_NON_INCAPACITATING	int	Số lượng chấn thương không ảnh hưởng nghiêm trong đến khả năng hoạt động của người bị thương.
40	INJURIES_NO_INDICATION	int	Số lượng chấn thương không có biểu hiện đáng chú ý (không có chấn thương)
41	INJURIES_REPORTED_NOT_EVIDENT	int	Thương tích được báo cáo không rõ ràng
42	INJURIES_UNKNOWN	int	Thương tích không rõ
43	CRASH_HOUR	int	Giờ xảy ra tai nạn
44	CRASH_DAY_OF_WEEK	int	Thứ tự ngày trong tuần của ngày xảy ra tai nạn
45	CRASH_MONTH	int	Tháng xảy ra tai nạn
46	LATITUDE	int	Vĩ độ: Tọa độ địa lý của địa điểm xảy ra tai nạn.
47	LONGITUDE	int	Kinh độ: Tọa độ địa lý của địa điểm máy bay rơi.
48	LOCATION	varchar	Điểm địa lý (vĩ độ và kinh độ) nơi máy bay rơi.

5.2 Phân tích tổng quan bộ dữ liệu

5.2.1 Thống kê mô tả

Đầu tiên, ta cần tính toán đại lượng thống kê mô tả: Count, Mean, Min, Max trên bộ dữ liệu.

Bước 1: Import các thư viện cần thiết.

```
import pandas as pd
import numpy as np
import seaborn as sns
import category_encoders as ce

import matplotlib.pyplot as plt
%matplotlib inline

from scipy.stats import pearsonr
from sklearn.compose import make_column_transformer
from sklearn.pipeline import Pipeline, make_pipeline
import sklearn.linear_model as skl_lm
import sklearn.model_selection as skl_ms
import sklearn.preprocessing as skl_pre
from sklearn.impute import KNNImputer
from sklearn.metrics import mean_squared_error, mean_absolute_error, mean_absolute_percentage_error, r2_score
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Ridge
from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import AdaBoostRegressor
from xgboost import XGBRegressor
from lightgbm import LGBMRegressor
from sklearn.model_selection import cross_validate, KFold
from sklearn.metrics import make_scorer
```

Tiến hành đọc dữ liệu từ file csv.

```
# Đọc file CSV
df = pd.read_csv('Traffic_Crashes_-_Crashes.csv')
```

Bước 2: Để thống kê mô tả trên bộ dữ liệu ta dùng phương thức **describe()**. Cụ thể là cung cấp các chỉ số thống kê như mean (trung bình), max (giá trị lớn nhất), min (giá trị nhỏ nhất), std (độ lệch chuẩn) và số lượng quan sát.

```
df.describe()
```

Kết quả:

df.describe()

	POSTED_SPEED_LIMIT	LANE_CNT	STREET_NO	BEAT_OF_OCCURRENCE	NUM_UNITS	INJURIES_TOTAL	INJURIES_FATAL	INJURIES_INCAPACITATING
count	794956.000000	1.990060e+05	794956.000000	794951.000000	794956.000000	793215.000000	793215.000000	793215.000000
mean	28.402547	1.333032e+01	3688.085435	1242.916397	2.034919	0.189448	0.001195	0.020032
std	6.179272	2.961623e+03	2887.990566	705.268208	0.452475	0.565654	0.037456	0.165673
min	0.000000	0.000000e+00	0.000000	111.000000	1.000000	0.000000	0.000000	0.000000
25%	30.000000	2.000000e+00	1248.000000	714.000000	2.000000	0.000000	0.000000	0.000000
50%	30.000000	2.000000e+00	3201.000000	1211.000000	2.000000	0.000000	0.000000	0.000000
75%	30.000000	4.000000e+00	5600.000000	1822.000000	2.000000	0.000000	0.000000	0.000000
max	99.000000	1.191625e+06	451100.000000	6100.000000	18.000000	21.000000	4.000000	10.000000

```
df.describe()
```

INJURIES_INCAPACITATING	INJURIES_NON_INCAPACITATING	INJURIES_REPORTED_NOT_EVIDENT	INJURIES_NO_INDICATION	INJURIES_UNKNOWN	CRASH_HOUR
793215.000000	793215.000000	793215.000000	793215.000000	793215.0	794956.000000
0.020032	0.107025	0.061195	2.004313	0.0	13.202800
0.165673	0.422095	0.318451	1.158715	0.0	5.568727
0.000000	0.000000	0.000000	0.000000	0.0	0.000000
0.000000	0.000000	0.000000	1.000000	0.0	9.000000
0.000000	0.000000	0.000000	2.000000	0.0	14.000000
0.000000	0.000000	0.000000	2.000000	0.0	17.000000
10.000000	21.000000	15.000000	61.000000	0.0	23.000000

```
df.describe()
```

JURIES_REPORTED_NOT_EVIDENT	INJURIES_NO_INDICATION	INJURIES_UNKNOWN	CRASH_HOUR	CRASH_DAY_OF_WEEK	CRASH_MONTH	LATITUDE	LONGITUDE
793215.000000	793215.000000	793215.0	794956.000000	794956.000000	794956.000000	789567.000000	789567.000000
0.061195	2.004313	0.0	13.202800	4.122344	6.787978	41.854788	-87.673450
0.318451	1.158715	0.0	5.568727	1.980411	3.405398	0.337523	0.686153
0.000000	0.000000	0.0	0.000000	1.000000	1.000000	0.000000	-87.936193
0.000000	1.000000	0.0	9.000000	2.000000	4.000000	41.782429	-87.721709
0.000000	2.000000	0.0	14.000000	4.000000	7.000000	41.874743	-87.674094
0.000000	2.000000	0.0	17.000000	6.000000	10.000000	41.924336	-87.633300
15.000000	61.000000	0.0	23.000000	7.000000	12.000000	42.022780	0.000000

5.2.2 Kiểm tra dữ liệu ban đầu

Tiến hành kiểm tra thông tin của bộ dữ liệu:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 794956 entries, 0 to 794955
Data columns (total 48 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   CRASH_RECORD_ID               794956 non-null  object
 1   CRASH_DATE_EST_I              59606 non-null   object
 2   CRASH_DATE                    794956 non-null  object
 3   POSTED_SPEED_LIMIT            794956 non-null  int64
 4   TRAFFIC_CONTROL_DEVICE        794956 non-null  object
 5   DEVICE_CONDITION              794956 non-null  object
 6   WEATHER_CONDITION             794956 non-null  object
 7   LIGHTING_CONDITION            794956 non-null  object
 8   FIRST_CRASH_TYPE              794956 non-null  object
 9   TRAFFICWAY_TYPE               794956 non-null  object
 10  LANE_CNT                      199006 non-null  float64
 11  ALIGNMENT                     794956 non-null  object
 12  ROADWAY_SURFACE_COND          794956 non-null  object
 13  ROAD_DEFECT                   794956 non-null  object
 14  REPORT_TYPE                   771698 non-null  object
 15  CRASH_TYPE                    794956 non-null  object
 16  INTERSECTION_RELATED_I        182215 non-null  object
 17  NOT_RIGHT_OF_WAY_I            36727 non-null   object
 18  HIT_AND_RUN_I                 248626 non-null  object
 19  DAMAGE                        794956 non-null  object
 20  DATE_POLICE_NOTIFIED          794956 non-null  object
 21  PRIM_CONTRIBUTORY_CAUSE       794956 non-null  object
 22  SEC_CONTRIBUTORY_CAUSE        794956 non-null  object
 23  STREET_NO                     794956 non-null  int64
 24  STREET_DIRECTION              794952 non-null  object
 25  STREET_NAME                   794955 non-null  object
 26  BEAT_OF_OCCURRENCE            794951 non-null  float64
 27  PHOTOS_TAKEN_I                10371 non-null   object
 28  STATEMENTS_TAKEN_I            17557 non-null   object
 29  DOORING_I                     2461 non-null    object
 30  WORK_ZONE_I                   4603 non-null    object
 31  WORK_ZONE_TYPE                3574 non-null    object
 32  WORKERS_PRESENT_I             1178 non-null    object
 33  NUM_UNITS                     794956 non-null  int64
 34  MOST_SEVERE_INJURY            793204 non-null  object
 35  INJURIES_TOTAL                793215 non-null  float64
 36  INJURIES_FATAL                793215 non-null  float64
 37  INJURIES_INCAPACITATING       793215 non-null  float64
 38  INJURIES_NON_INCAPACITATING   793215 non-null  float64
 39  INJURIES_REPORTED_NOT_EVIDENT 793215 non-null  float64
 40  INJURIES_NO_INDICATION        793215 non-null  float64
 41  INJURIES_UNKNOWN              793215 non-null  float64
 42  CRASH_HOUR                    794956 non-null  int64
 43  CRASH_DAY_OF_WEEK             794956 non-null  int64
 44  CRASH_MONTH                   794956 non-null  int64
 45  LATITUDE                      789567 non-null  float64
 46  LONGITUDE                     789567 non-null  float64
 47  LOCATION                      789567 non-null  object
dtypes: float64(11), int64(6), object(31)
memory usage: 291.1+ MB
```

Xem tổng số hàng và cột.

```
print('This data has {} rows and {} columns'.format(df.shape[0], df.shape[1]))
```

```
This data has 794956 rows and 48 columns
```

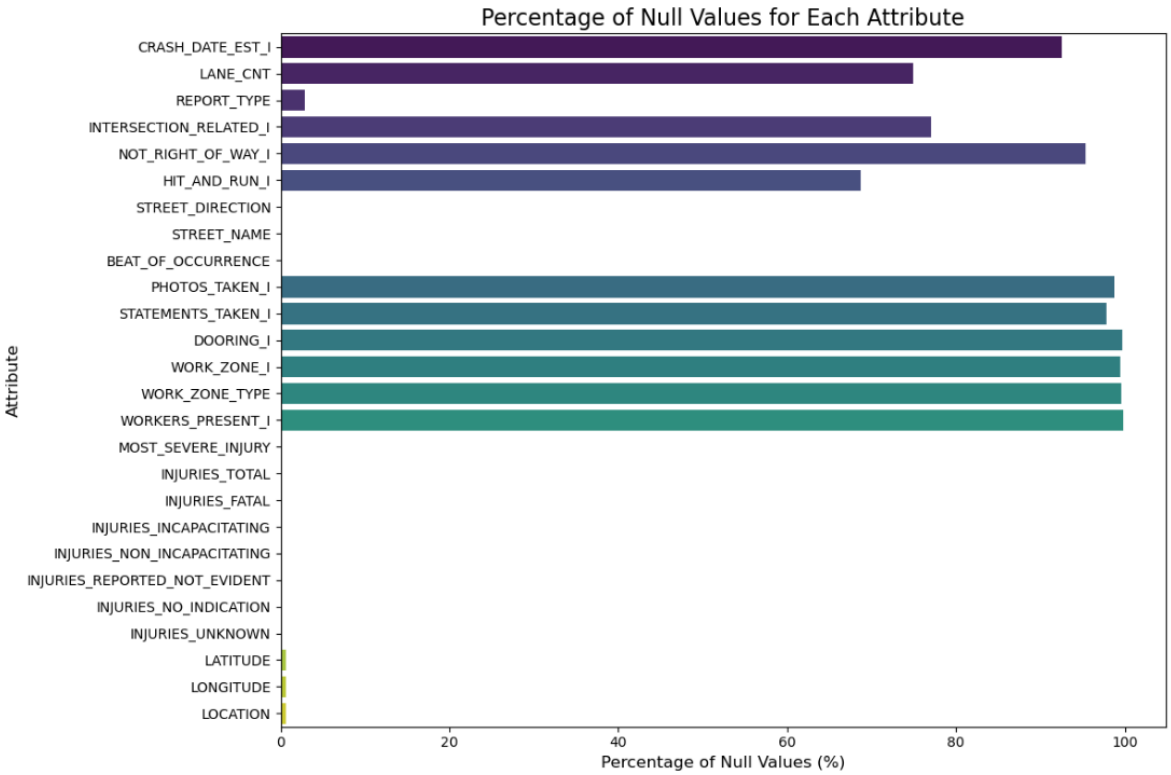
Xem những giá trị bị khuyết.

```python
print("Dataset missing values: \n", df.isna().sum())
```

```
Dataset missing values:
 CRASH_RECORD_ID                      0
CRASH_DATE_EST_I                735350
CRASH_DATE                           0
POSTED_SPEED_LIMIT                   0
TRAFFIC_CONTROL_DEVICE               0
DEVICE_CONDITION                     0
WEATHER_CONDITION                    0
LIGHTING_CONDITION                   0
FIRST_CRASH_TYPE                     0
TRAFFICWAY_TYPE                      0
LANE_CNT                        595950
ALIGNMENT                            0
ROADWAY_SURFACE_COND                 0
ROAD_DEFECT                          0
REPORT_TYPE                      23258
CRASH_TYPE                           0
INTERSECTION_RELATED_I          612741
NOT_RIGHT_OF_WAY_I              758229
HIT_AND_RUN_I                   546330
DAMAGE                               0
DATE_POLICE_NOTIFIED                 0
PRIM_CONTRIBUTORY_CAUSE              0
SEC_CONTRIBUTORY_CAUSE               0
STREET_NO                            0
STREET_DIRECTION                     4
STREET_NAME                          1
BEAT_OF_OCCURRENCE                   5
PHOTOS_TAKEN_I                  784585
STATEMENTS_TAKEN_I              777399
DOORING_I                       792495
WORK_ZONE_I                     790353
WORK_ZONE_TYPE                  791382
WORKERS_PRESENT_I               793778
NUM_UNITS                            0
 MOST_SEVERE_INJURY                 1752
 INJURIES_TOTAL                     1741
 INJURIES_FATAL                     1741
 INJURIES_INCAPACITATING            1741
 INJURIES_NON_INCAPACITATING        1741
 INJURIES_REPORTED_NOT_EVIDENT      1741
 INJURIES_NO_INDICATION             1741
 INJURIES_UNKNOWN                   1741
 CRASH_HOUR                            0
 CRASH_DAY_OF_WEEK                     0
 CRASH_MONTH                           0
 LATITUDE                           5389
 LONGITUDE                          5389
 LOCATION                           5389
 dtype: int64
```

Vẽ đồ thị thống kê (%) null của các thuộc tính có giá trị null.

```python
# Tính toán tỷ lệ phần trăm giá trị null cho từng cột
null_percentage = df.isnull().mean() * 100

# Lọc các cột có giá trị null (tỷ lệ phần trăm > 0)
null_percentage = null_percentage[null_percentage > 0]

# Vẽ đồ thị thống kê tỷ lệ phần trăm cột null
plt.figure(figsize=(12, 8))
sns.barplot(y=null_percentage.index, x=null_percentage.values, palette='viridis')
plt.title('Percentage of Null Values for Each Attribute', fontsize=16)
plt.ylabel('Attribute', fontsize=12)
plt.xlabel('Percentage of Null Values (%)', fontsize=12)
plt.tight_layout()


plt.show()
```



5.3 Khám phá dữ liệu

Tập dữ liệu **TrafficCrash** bao gồm tổng cộng 48 thuộc tính. Nhóm sẽ tiến hành khám phá dữ liệu (EDA) để phân tích chi tiết đơn biến và hai biến (phân loại vụ va chạm – có bị thương hay không) khi cần thiết, từ đó hiểu rõ hơn về đặc điểm và mối quan hệ của các biến trong tập dữ liệu.

Cụ thể, nhóm sẽ thực hiện các nội dung sau:

1. **Khám phá dữ liệu đơn biến**:
    - Phân tích từng thuộc tính riêng biệt trong tập dữ liệu để xác định loại dữ liệu (số học hoặc phân loại), phân bố của các giá trị và kiểm tra sự tồn tại của giá trị thiếu hoặc bất thường.
2. **Khám phá mối quan hệ giữa các thuộc tính và phân loại va chạm**:
    - Phân tích các yếu tố có thể ảnh hưởng trực tiếp đến phân loại thương vong, ví dụ như PRIM_CONTRIBUTORY_CAUSE, TRAFFIC_CONTROL_DEVICE, WEATHER_CONDITION, v.v.
    - Sử dụng các biểu đồ để quan sát mối quan hệ giữa các thuộc tính này và
3. **Xử lý giá trị thiếu và bất thường**:
    - Loại bỏ hoặc thay thế các giá trị thiếu (missing values), cũng như xử lý các giá trị bất thường (outliers) để đảm bảo dữ liệu chất lượng trước khi đưa vào mô hình phân tích.

Sau khi hoàn thành quá trình khám phá dữ liệu, nhóm sẽ tiếp tục bước tiếp theo trong việc xây dựng mô hình và đánh giá kết quả, nhằm đạt được những dự đoán chính xác và đáng tin cậy.

5.3.1 Loại bỏ thuộc tính không cần thiết

Khi phân tích dữ liệu để thực hiện bài toán phân lớp loại va chạm, việc loại bỏ một số thuộc tính nhất định là cần thiết nhằm tránh dư thừa, giảm độ phức tạp và tập trung vào các đặc trưng thực sự hữu ích cho mô hình.Tiến hành loại bỏ các thuộc tính không đầy đủ và các thuộc tính không có tác dụng cho mô hình phân loại.

5.3.1.1 CRASH_RECORD_ID

Lý do loại bỏ: Đây là mã định danh duy nhất của vụ va chạm, không mang ý nghĩa hoặc ảnh hưởng đến kết quả phân lớp.

5.3.1.2 CRASH_DATE_EST_I và CRASH_DATE

Lý do loại bỏ: Thời gian xảy ra vụ va chạm thường không ảnh hưởng trực tiếp đến khả năng xảy ra tai nạn. Thời gian nên được chia lớp thành thời gian trong ngày, ngày trong tuần, bản thân ngày giờ cụ thể không hữu ích.

5.3.1.3 LANE_CNT

Lý do loại bỏ: Số làn đường có thể ảnh hưởng đến xác suất xảy ra tai nạn nhưng không phải lúc nào cũng có sẵn hoặc có ý nghĩa trực tiếp trong bài toán phân lớp. Ở bộ dữ liệu hiện tại, thuộc tính này thiếu dữ liệu cho nhiều bản ghi (null > 80%), nên loại bỏ.

5.3.1.4 REPORT_TYPE

Lý do loại bỏ: Loại báo cáo không cung cấp thông tin trực tiếp về nguyên nhân hay hậu quả của vụ tai nạn, và thường mang tính hành chính hơn là phân tích.

5.3.1.5 INTERSECTION_RELATED_I

Lý do loại bỏ: Biến này chỉ định vụ va chạm có liên quan đến giao lộ hay không. Tuy nhiên, do số dữ trống quá lớn (null>80%) nên biến này không hữu ích cho mô hình phân loại.

5.3.1.6 NOT_RIGHT_OF_WAY_I

Lý do loại bỏ: Thông tin về vi phạm quyền ưu tiên có thể hữu ích, nhưng dữ liệu không đầy đủ (null > 90%) nên được loại bỏ để tránh gây nhiễu.

5.3.1.7 HIT_AND_RUN_I

Lý do loại bỏ: Va chạm có phải là "bỏ chạy" hay không là yếu tố hậu kỳ và không ảnh hưởng đến mục tiêu phân lớp chính.

5.3.1.8 DATE_POLICE_NOTIFIED

Lý do loại bỏ: Ngày cảnh sát được thông báo không liên quan đến đặc điểm xảy ra vụ va chạm hoặc mục tiêu phân lớp.

5.3.1.9 SEC_CONTRIBUTORY_CAUSE

Lý do loại bỏ: Thuộc tính có giá trị "Không xác định", "Không đề cập" rất nhiều. Ngoài ra, nguyên nhân thứ cấp của vụ va chạm có thể là thông tin dư thừa nếu đã có thông tin về nguyên nhân chính.

5.3.1.10 STREET_NO, STREET_DIRECTION, STREET_NAME

Lý do loại bỏ: Các thông tin chi tiết về tên đường, số đường không có ý nghĩa trực tiếp trong bài toán phân loại vụ va chạm và chỉ gây nhiễu cho mô hình, nên được loại bỏ.

5.3.1.11 BEAT_OF_OCCURRENCE

Lý do loại bỏ: Đây là thông tin mã hóa khu vực, thường không ảnh hưởng đến đặc điểm vụ va chạm trừ khi phân tích tập trung vào vị trí địa lý cụ thể.

5.3.1.12 PHOTOS_TAKEN_I, STATEMENTS_TAKEN_I

Lý do loại bỏ: Thông tin về việc chụp ảnh hoặc lấy lời khai là yếu tố hậu kỳ, không ảnh hưởng đến kết quả phân loại va chạm.

5.3.1.13 DOORING_I

Lý do loại bỏ: Va chạm liên quan đến mở cửa xe, biến này có số lượng rỗng khá lớn (null>96%) nên cần loại bỏ.

5.3.1.14 WORK_ZONE_I, WORK_ZONE_TYPE, WORKERS_PRESENT_I

Lý do loại bỏ: Các thông tin về khu vực làm việc chỉ hữu ích trong bài toán phân tích tai nạn tại các công trường. Nếu không nằm trong phạm vi bài toán, nên loại bỏ.

5.3.1.15 INJURIES_TOTAL, INJURIES_FATAL, INJURIES_INCAPACITATING, INJURIES_NON_INCAPACITATING, INJURIES_REPORTED_NOT_EVIDENT, INJURIES_NO_INDICATION, INJURIES_UNKNOWN

Lý do loại bỏ: Các thông tin về thương tích là kết quả của vụ va chạm chứ không phải nguyên nhân để góp phần dự đoán phân loại va chạm. Chúng không nên được sử dụng nếu bài toán phân lớp nhằm xác định khả năng xảy ra của loại tai nạn, tránh gây ra over-fitting.

5.3.1.14 LATITUDE, LONGITUDE, LOCATION

Lý do loại bỏ: Dữ liệu tọa độ và vị trí cụ thể có thể quá chi tiết và không cần thiết cho bài toán phân loại vụ va chạm mà nhóm hướng đến.

5.3.1.15 MOST_SEVERE_INJURY

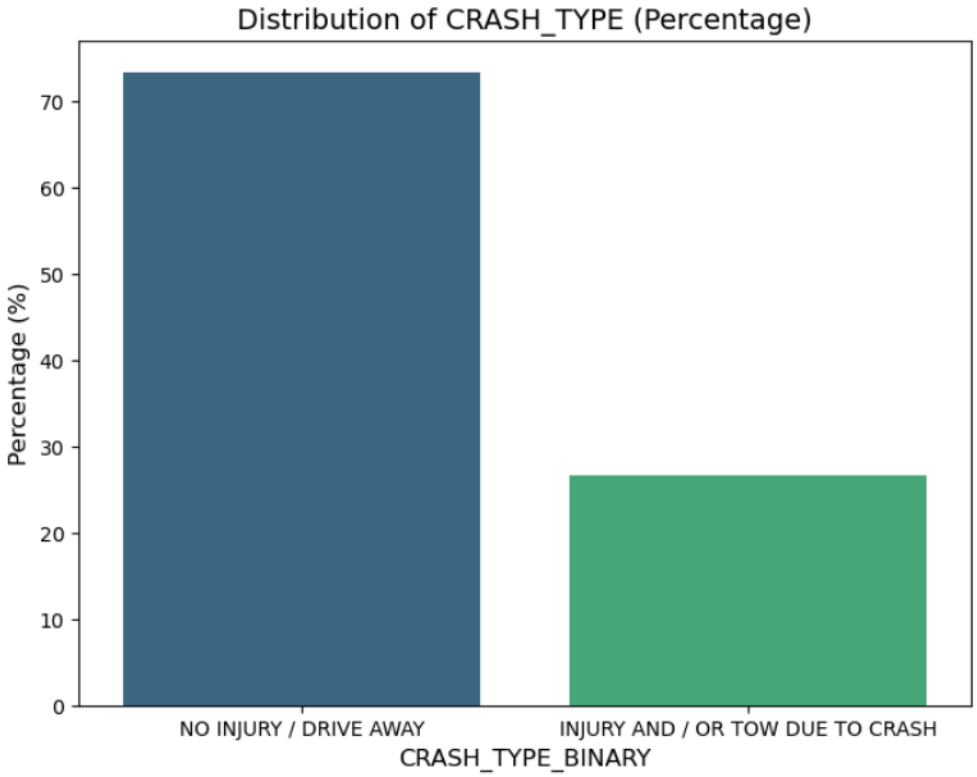
Lý do loại bỏ: Tương tự như thông tin thương tích, đây là yếu tố hậu quả và không phù hợp để làm đặc trưng đầu vào.

5.3.2 Lựa chọn thuộc tính

5.3.2.1 CRASH_TYPE

Đây là biến mục tiêu để phân lớp loại va chạm có thể xảy ra.

Kiểm tra các giá trị trong cột “CRASH_TYPE”.

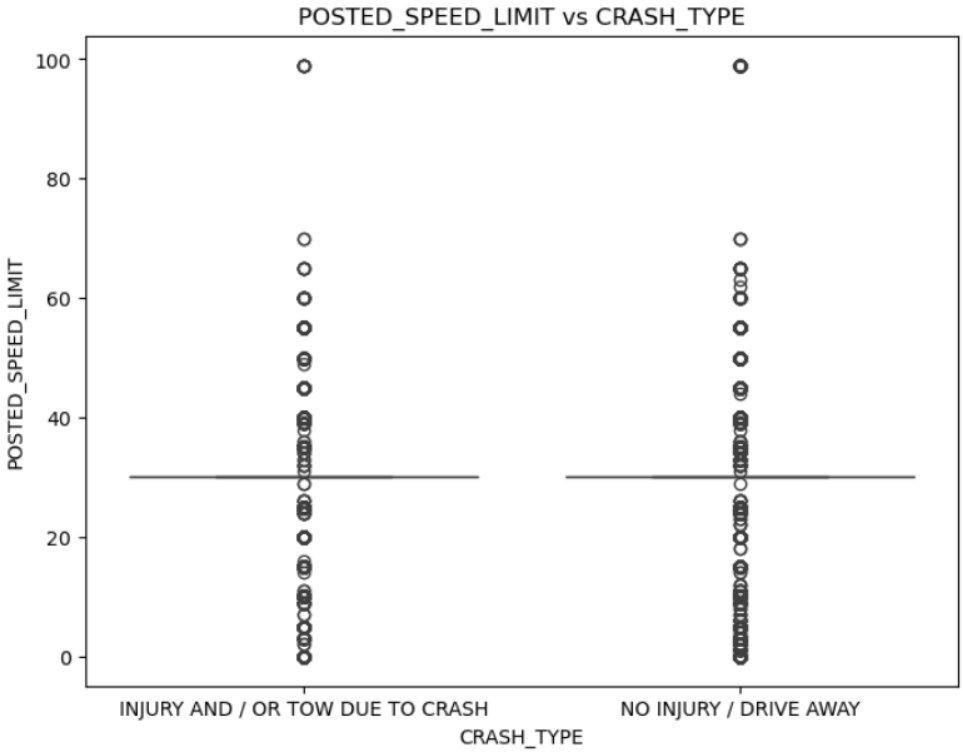


Bài toán mà nhóm hướng đến là bài toán phân lớp nhị phân, biến mục tiêu là “CRASH_TYPE” với hai lớp là: “NO INJURY / DRIVE AWAY” và “INJURY AND / OR TOW DUE TO CRASH”.

Dễ dàng nhận thấy rằng có sự mất cân bằng giữa hai giá trị. Do đó, ta cần lựa chọn các kỹ thuật, thuật toán phù hợp cho bài toán phân lớp loại va chạm.

5.3.2.2 POSTED_SPEED_LIMIT

Phân tích sự phân bố giữa các lớp (CRASH_TYPE) và POSTED_SPEED_LIMIT bằng boxplot.



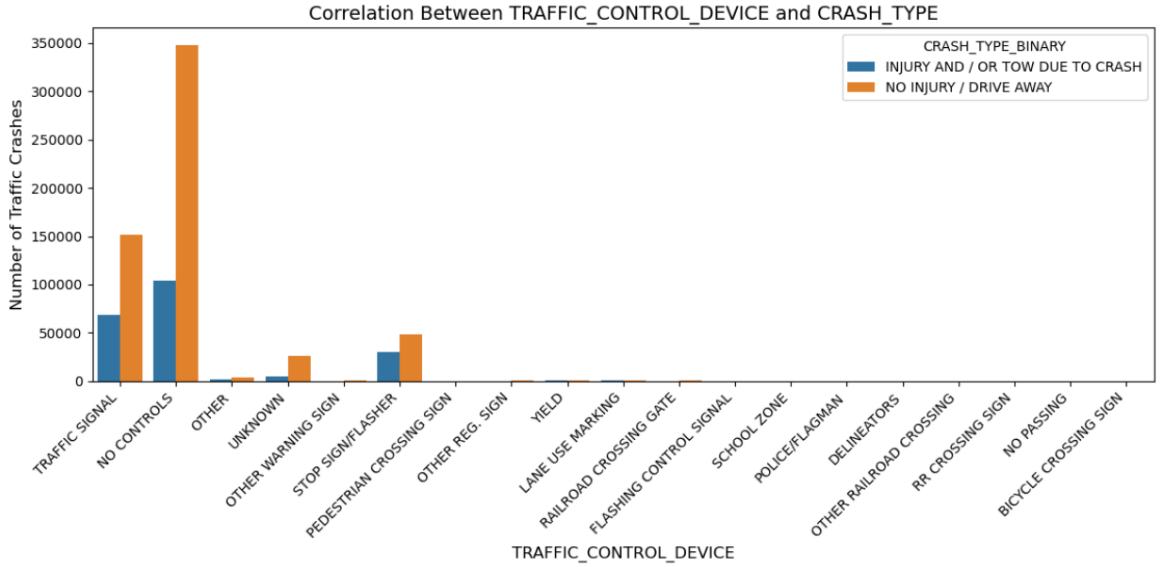
- Cả hai lớp đều có nhiều **ngoại lệ** (dấu chấm ngoài các whisker của boxplot). Điều này có thể chỉ ra rằng có một số vụ tai nạn xảy ra với tốc độ giới hạn rất cao, đặc biệt là trong lớp "INJURY AND / OR TOW DUE TO CRASH", với tốc độ trên 80 hoặc thậm chí 100.
- Các boxplot của hai lớp có sự **chồng lắp** khá nhiều (tức là các giá trị tốc độ giới hạn của cả hai lớp không tách biệt rõ rệt). Điều này có thể là một dấu hiệu cho thấy rằng tốc độ giới hạn không phải là yếu tố duy nhất hoặc không phải yếu tố chính quyết định đến mức độ nghiêm trọng của tai nạn.

Với một bài toán như tai nạn giao thông, các yếu tố như thời tiết, ánh sáng, nguyên nhân có thể tương tác phức tạp với nhau. Khi kết hợp với các thuộc tính khác như TRAFFIC_CONTROL_DEVICE, WEATHER_CONDITION, LIGHTING_CONDITION, mô hình có thể học được các mối quan hệ ẩn và cung cấp dự đoán chính xác hơn.

POSTED_SPEED_LIMIT vẫn là một thuộc tính có giá trị và có thể giúp cải thiện kết quả phân loại khi kết hợp với các yếu tố khác.

5.3.2.3 TRAFFIC_CONTROL_DEVICE

TRAFFIC_CONTROL_DEVICE là một trong những thuộc tính quan trọng trong bài toán phân loại va chạm giao thông. Thuộc tính này mô tả các thiết bị điều khiển giao thông như đèn tín hiệu, biển báo giao thông, v.v. Những thiết bị này có thể ảnh hưởng trực tiếp đến việc phân loại khả năng xảy ra các loại tai nạn.

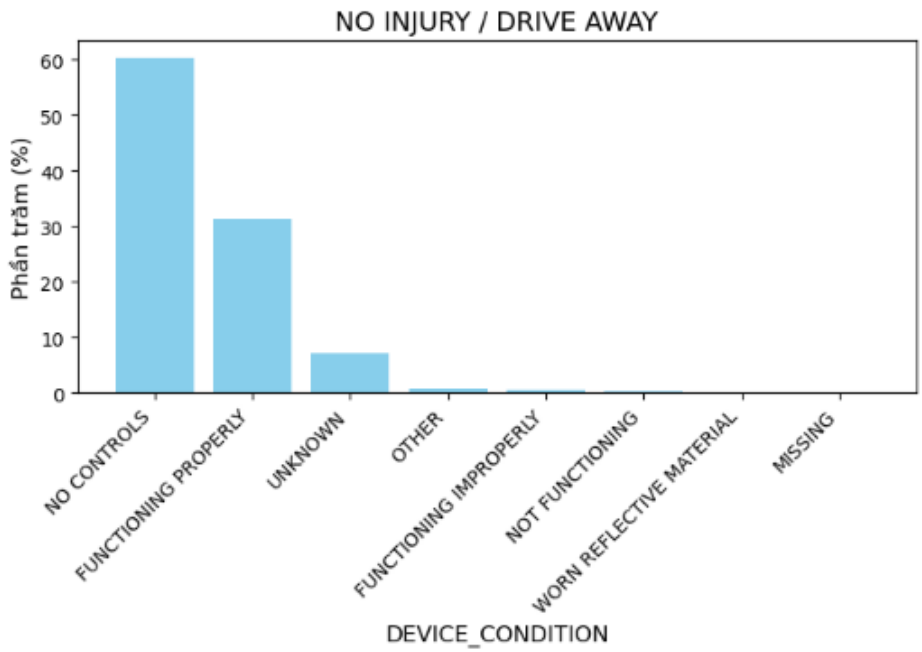


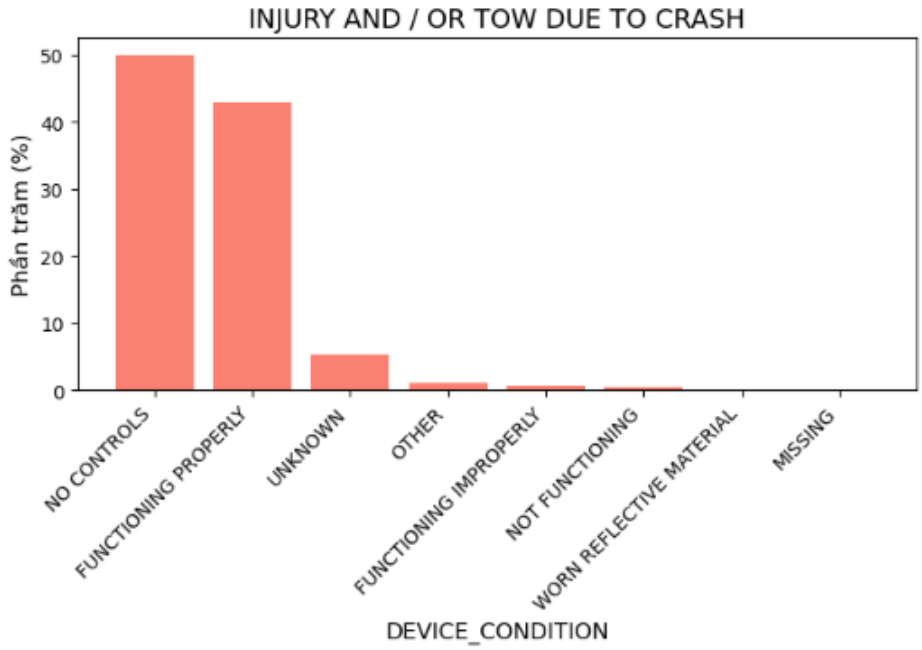
Dữ liệu có sự mất cân bằng giữa giá trị (NO INJURY / DRIVE AWAY) gấp ~2.75 lần giá trị (INJURY AND / OR TOW DUE TO CRASH), nhưng ta có thể thấy rõ có sự chênh lệnh khi dữ liệu (INJURY AND / OR TOW DUE TO CRASH) của 'STOP SIGN/FLASHER' chiếm phần trăm so với giá trị (NO INJURY / DRIVE AWAY) cao hơn các cột khác, điều đó chứng minh rằng đây là loại thiết bị điều khiển giao thông có tỉ lệ các vụ va chạm nghiêm trọng cao hơn các loại thiết bị khác.

Chọn thuộc tính TRAFFIC_CONTROL_DEVICE cho bài toán phân lớp CRASH_TYPE. Thuộc tính này giúp mô hình phân biệt giữa các tai nạn nghiêm trọng và không nghiêm trọng, đồng thời cải thiện khả năng phân loại khi kết hợp với các thuộc tính khác.

5.3.2.4 DEVICE_CONDITION

DEVICE_CONDITION phản ánh tình trạng hoạt động của các thiết bị kiểm soát giao thông (đèn tín hiệu, biển báo, vòng xuyến, v.v.). Đây là một yếu tố quan trọng, vì tình trạng không hoạt động hoặc hỏng hóc của thiết bị có thể làm tăng nguy cơ tai nạn, đặc biệt là ở các giao lộ hoặc khu vực có mật độ giao thông cao.





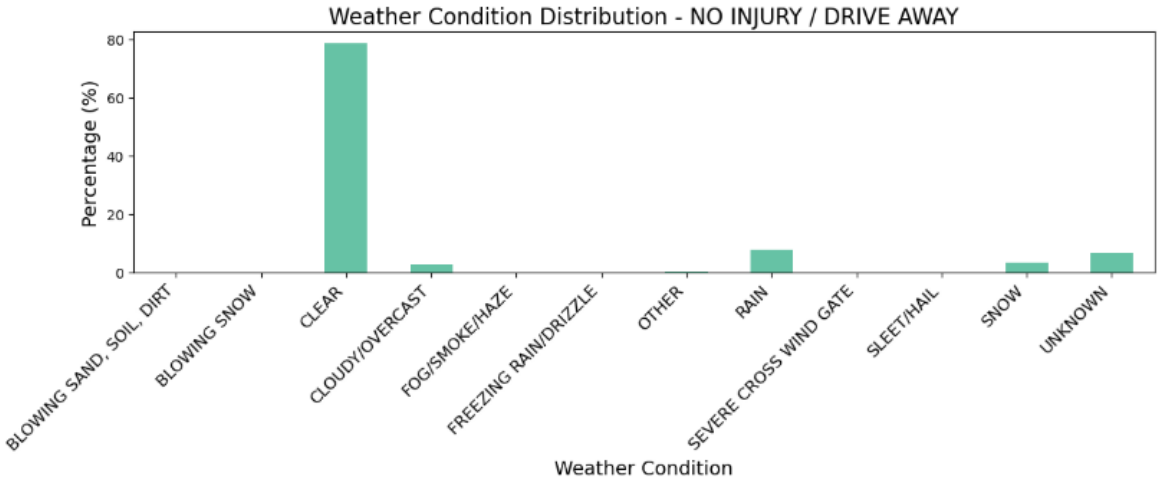
Trong cả hai loại tai nạn, tình trạng thiết bị "NO CONTROLS" (không có điều khiển) chiếm tỷ lệ cao. Điều này có thể phản ánh rằng trong nhiều trường hợp tai nạn, thiết bị điều khiển (như biển báo, đèn tín hiệu, hoặc các thiết bị điều khiển giao thông khác) không có hoặc không hoạt động, dẫn đến sự mất kiểm soát hoặc gây tai nạn.

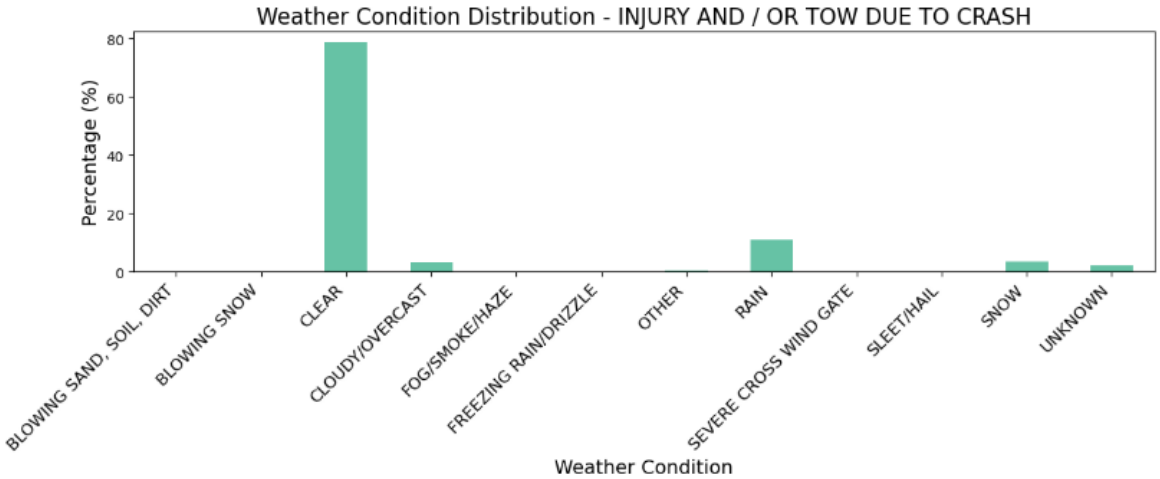
Tuy nhiên, trạng thái "FUNCTIONING PROPERLY" (hoạt động tốt) chiếm phần trăm cao trong loại tai nạn nghiêm trọng cho thấy rằng mặc dù thiết bị có thể hoạt động bình thường, vẫn có những yếu tố khác ngoài thiết bị (như yếu tố con người hoặc điều kiện khác) góp phần vào tai nạn.

Có sự khác biệt trong sự phân bố của DEVICE_CONDITION trong CRASH_TYPE, do đó, nhóm quyết định chọn thuộc tính cho mô hình phân lớp.

5.3.2.5 WEATHER_CONDITION

Điều kiện thời tiết là một yếu tố góp phần vào tình trạng thương vong ở các vụ va chạm giao thông. Thời tiết xấu như mưa hoặc tuyết có thể làm giảm tầm nhìn và độ bám đường, dẫn đến tăng khả năng xảy ra tai nạn nghiêm trọng cao hơn.



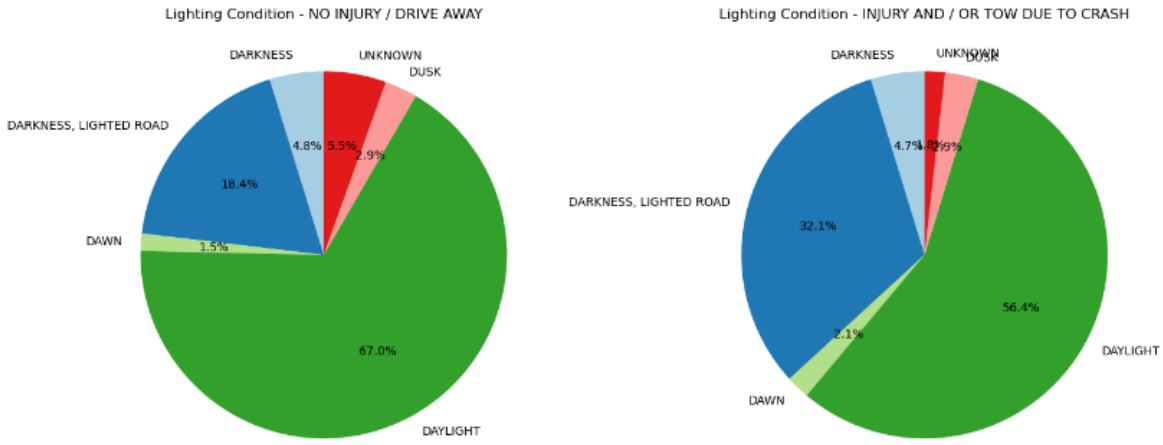


Hầu hết các vụ tai nạn trong loại này có điều kiện thời tiết là "CLEAR" (trời quang đãng), chiếm tỷ lệ gần 80%. Các điều kiện thời tiết khác như "CLOUDY/OVERCAST", "RAIN", và "SNOW" chiếm tỷ lệ nhỏ.

Điều kiện thời tiết có thể ảnh hưởng mạnh đến khả năng xảy ra tai nạn và mức độ nghiêm trọng của tai nạn. Trong các loại tai nạn có chấn thương hoặc cần phải kéo đi, điều kiện thời tiết như mưa hoặc tuyết có thể làm tăng nguy cơ xảy ra tai nạn nghiêm trọng hơn. Chọn thuộc tính WEATHER_CONDITION cho mô hình.

5.3.2.6 LIGHTING_CONDITION

Điều kiện ánh sáng ảnh hưởng trực tiếp đến tầm nhìn và khả năng nhận diện các chướng ngại vật hoặc các phương tiện khác, từ đó tác động đến khả năng gây tai nạn.



Vào ban ngày, tỉ lệ xảy ra các loại va chạm nghiêm trọng có xu hướng thấp hơn ban đêm.

Darkness, Lighted Road (Đêm, đường có ánh sáng): Mặc dù có ánh sáng từ đèn đường, nhưng nó có thể không đủ để cung cấp tầm nhìn rõ ràng trong mọi tình huống. Các sự cố xảy ra vào ban đêm thường là do tầm nhìn hạn chế, thời gian phản ứng chậm và các yếu tố khác tác động khi điều khiển phương tiện vào ban đêm. Chính vì vậy, các vụ tai nạn trong điều kiện này có xu hướng nghiêm trọng hơn.

Đây là một thuộc tính cần thiết cho bài toán phân loại vụ tai nạn.

5.3.2.7 FIRST_CRASH_TYPE

FIRST_CRASH_TYPE là một thuộc tính quan trọng để xác định tác nhân va chạm. Việc phân tích phần trăm của từng loại va chạm giúp ta hiểu rõ hơn về tính chất và đặc điểm của các vụ tai nạn trong các nhóm khác nhau, từ đó có thể đưa ra các biện pháp cải thiện an toàn giao thông.

```
Crash Type Distribution - NO INJURY / DRIVE AWAY:
FIRST_CRASH_TYPE
PARKED MOTOR VEHICLE            26.142663
REAR END                        24.399968
SIDESWIPE SAME DIRECTION        18.311422
TURNING                         13.116429
ANGLE                            8.590874
FIXED OBJECT                     3.006977
SIDESWIPE OPPOSITE DIRECTION     1.484626
REAR TO FRONT                    1.177344
OTHER OBJECT                     0.883951
REAR TO SIDE                     0.666864
PEDALCYCLIST                     0.619537
HEAD ON                          0.563980
PEDESTRIAN                       0.400222
OTHER NONCOLLISION               0.280189
REAR TO REAR                     0.262013
ANIMAL                           0.086080
OVERTURNED                       0.005830
TRAIN                            0.001029
Name: proportion, dtype: float64

Crash Type Distribution - INJURY AND / OR TOW DUE TO CRASH:
FIRST_CRASH_TYPE
TURNING                         17.443184
ANGLE                           17.104623
REAR END                        16.996492
PARKED MOTOR VEHICLE            15.298495
FIXED OBJECT                     9.347008
PEDESTRIAN                       7.589043
SIDESWIPE SAME DIRECTION         6.797180
PEDALCYCLIST                     4.000397
HEAD ON                          1.647472
OTHER OBJECT                     1.283886
SIDESWIPE OPPOSITE DIRECTION     1.215890
OTHER NONCOLLISION               0.416944
REAR TO FRONT                    0.307868
REAR TO SIDE                     0.267260
OVERTURNED                       0.212958
ANIMAL                           0.033998
REAR TO REAR                     0.019832
TRAIN                            0.017471
Name: proportion, dtype: float64
```

Việc tính toán phần trăm cho mỗi loại va chạm trong các nhóm tai nạn giúp chúng ta nhìn nhận được sự phân bố và mối liên hệ giữa điều kiện va chạm và mức độ nghiêm trọng của tai nạn.

Trong nhóm "NO INJURY / DRIVE AWAY", tai nạn phổ biến chủ yếu là va chạm nhẹ, đặc biệt là với các phương tiện đỗ bên đường hoặc va chạm phía sau (rear-end). Những vụ tai nạn này không gây thương tích nghiêm trọng, thường chỉ cần xử lý và không yêu cầu cứu hộ.

Trong nhóm "INJURY AND / OR TOW DUE TO CRASH", các vụ tai nạn thường nghiêm trọng hơn, với tỷ lệ cao ở các loại va chạm như TURNING và ANGLE. Những vụ va chạm này có thể dẫn đến thương tích nghiêm trọng và yêu cầu cứu hộ do mức độ va chạm mạnh hơn.

Từ đó, có thể thấy rằng những va chạm xảy ra khi xe quay đầu hoặc va chạm tại giao lộ (turning, angle) có xu hướng gây ra thương tích hoặc yêu cầu cứu hộ nhiều hơn các loại va chạm khác. Chọn thuộc tính cho mô hình phân loại.

5.3.2.8 TRAFFICWAY_TYPE

TRAFFICWAY_TYPE giúp xác định xem loại đường có phải là yếu tố quan trọng ảnh hưởng đến mức độ nghiêm trọng của tai nạn hay không. Từ đó giúp đưa ra các biện pháp phòng ngừa tai nạn hợp lý, như cải thiện các tuyến đường không chia làn hoặc nâng cấp các ngã tư để giảm nguy cơ xảy ra tai nạn nghiêm trọng.

```
Trafficway Type Distribution - NO INJURY / DRIVE AWAY:
TRAFFICWAY_TYPE
NOT DIVIDED                        44.103248
DIVIDED - W/MEDIAN (NOT RAISED)    15.559427
ONE-WAY                            13.906241
PARKING LOT                         8.656549
DIVIDED - W/MEDIAN BARRIER          4.799743
FOUR WAY                            4.167517
OTHER                               2.862424
ALLEY                               1.798768
UNKNOWN                             1.416380
T-INTERSECTION                      0.884466
CENTER TURN LANE                    0.590730
DRIVEWAY                            0.375529
RAMP                                0.311569
UNKNOWN INTERSECTION TYPE           0.214000
Y-INTERSECTION                      0.100998
FIVE POINT, OR MORE                 0.096540
TRAFFIC ROUTE                       0.077335
NOT REPORTED                        0.041325
ROUNDABOUT                          0.024006
L-INTERSECTION                      0.013204
Name: proportion, dtype: float64

Trafficway Type Distribution - INJURY AND / OR TOW DUE TO CRASH:
TRAFFICWAY_TYPE
NOT DIVIDED                        41.961668
DIVIDED - W/MEDIAN (NOT RAISED)    17.454988
FOUR WAY                           11.489336
ONE-WAY                             9.699734
DIVIDED - W/MEDIAN BARRIER          8.277969
OTHER                               2.398255
T-INTERSECTION                      2.228266
PARKING LOT                         1.760326
ALLEY                               1.249888
CENTER TURN LANE                    1.103509
UNKNOWN INTERSECTION TYPE           0.492495
UNKNOWN                             0.463219
RAMP                                0.321562
FIVE POINT, OR MORE                 0.241289
Y-INTERSECTION                      0.230901
TRAFFIC ROUTE                       0.205403
DRIVEWAY                            0.190293
NOT REPORTED                        0.141657
ROUNDABOUT                          0.050524
L-INTERSECTION                      0.038720
Name: proportion, dtype: float64
```

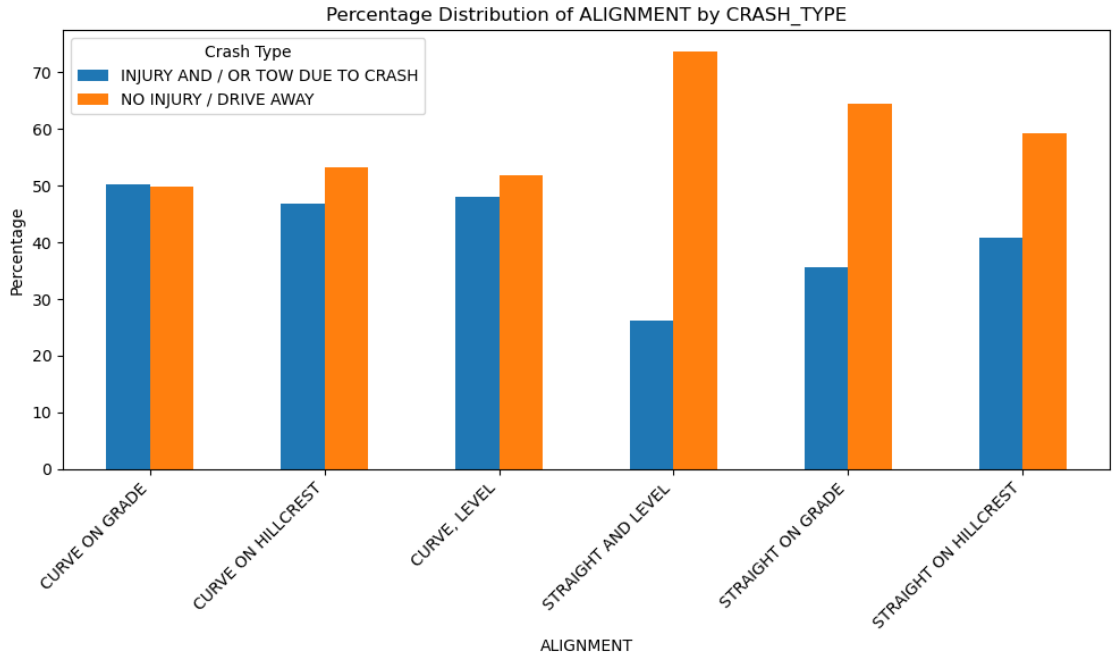
Loại đường không phân cách “NOT DEVICE” chiếm tỉ lệ lớn trong các vụ va chạm.

Mức độ nghiêm trọng: Các vụ tai nạn trong nhóm “INJURY AND / OR TOW DUE TO CRASH” thường xảy ra trên các loại đường có tính chất phức tạp hơn, như FOUR WAY, T-INTERSECTION, và các loại đường có phân cách làn (như DIVIDED - W/MEDIAN BARRIER), nơi có nhiều yếu tố tác động đến khả năng xảy ra tai nạn nghiêm trọng (ví dụ: giao cắt, tốc độ cao, va chạm với các vật cản).

Tình trạng tai nạn: Ngược lại, nhóm NO INJURY / DRIVE AWAY có tỷ lệ cao hơn trên các tuyến đường không chia làn, như NOT DIVIDED, và các khu vực như bãi đậu xe, cho thấy rằng các tai nạn trong nhóm này thường ít nghiêm trọng hơn và có thể chỉ là va quẹt hoặc tai nạn nhẹ.

5.3.2.9 ALIGNMENT

ALIGNMENT là một yếu tố quan trọng trong phân tích va chạm giao thông. Nó đề cập đến cách đường giao thông được xây dựng và thiết kế, chẳng hạn như đường thẳng, cong, hay cua gấp. ALIGNMENT có thể ảnh hưởng đến sự xuất hiện và mức độ nghiêm trọng của các vụ tai nạn giao thông.

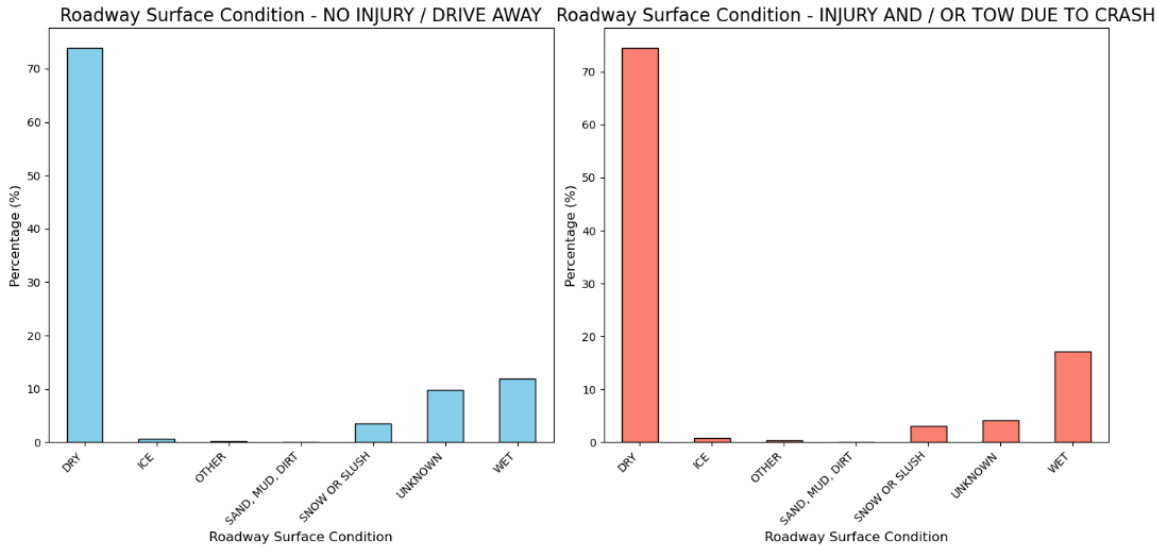


Ở các loại thiết kế đường thẳng thường có tỉ lệ các loại va chạm giao thông nghiêm trọng thấp hơn so với các loại thiết kế kiến trúc đường cong. Trên đường thẳng, các người tham gia giao thông có thể duy trì tốc độ ổn định và dễ dàng kiểm soát phương tiện của mình, không có sự thay đổi đột ngột về hướng đi, giúp dễ dàng điều khiển xe. Ở curve on grade - khi đường cong và độ dốc kết hợp, có thể gây khó khăn cho việc điều khiển phương tiện và làm tăng nguy cơ tai nạn. Vì thế cần đòi hỏi sự tính toán kỹ lưỡng để đảm bảo an toàn cho người tham gia giao thông.

Đây là thuộc tính quan trọng giúp phân loại các loại va chạm giao thông.

5.3.2.10ROADWAY_SURFACE_COND

Tình trạng bề mặt đường có thể đóng vai trò là một yếu tố trong việc phân lớp tai nạn, vì điều này ảnh hưởng trực tiếp đến khả năng lái xe và mức độ nguy hiểm của tai nạn.



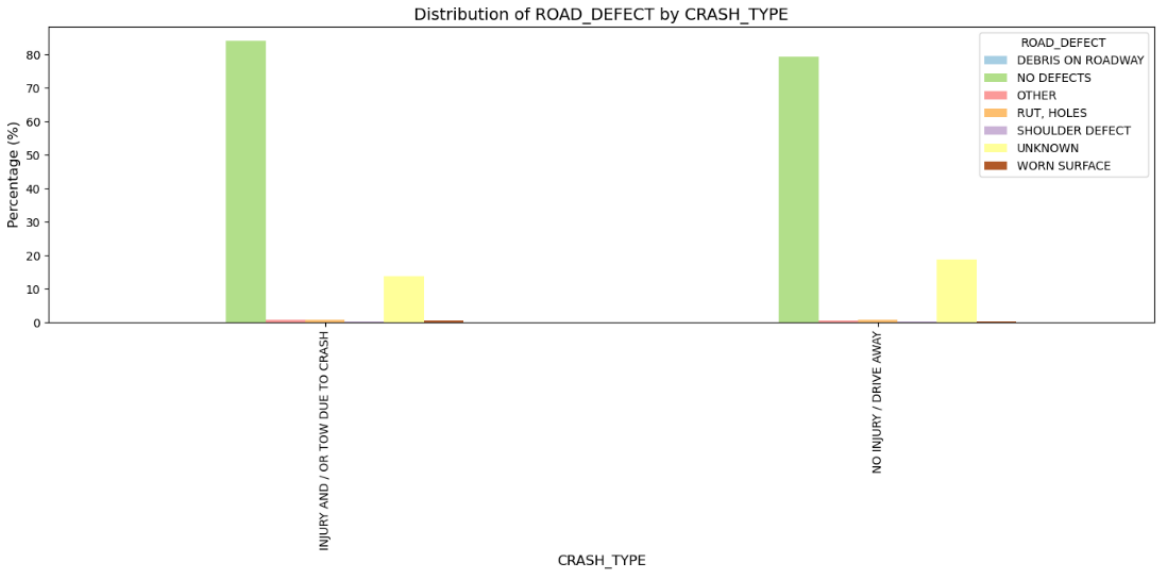
Điều kiện mặt đường khô phổ biến ở hai loại tai nạn do các điều kiện thời tiết thuận lợi, ở Chicago thời tiết mưa-ẩm thường không quá nhiều.

Qua sơ đồ ta thấy điều kiện mặt đường là "UNKNOWN" có một sự khác biệt trong tỉ lệ giữa hai loại tai nạn giao thông. Bên cạnh đó, khi điều kiện mặt đường là ướt, sự giảm độ bám làm tăng khả năng mất kiểm soát và sự nguy hiểm trong việc điều khiển phương tiện. Những yếu tố này có thể dẫn đến tỷ lệ tai nạn nghiêm trọng cao hơn.

Chọn thuộc tính này cho bài toán phân lớp.

5.3.2.11 ROAD_DEFECT

ROAD_DEFECT đại diện cho tình trạng bất thường của đường



Trong nhiều khu vực, mặt đường được duy trì và bảo dưỡng tốt. Do đó, khi xảy ra các loại tai nạn, yếu tố mặt đường thường kết hợp các yếu tố khác điều kiện thời tiết, nguyên nhân va chạm để đánh giá được mức độ tai nạn.

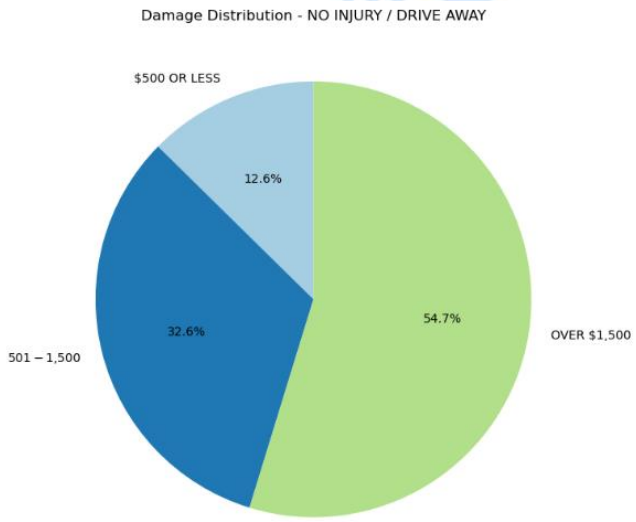
Mặc dù "No Defect" (không có khuyết tật mặt đường) thường được coi là yếu tố giúp giảm thiểu tai nạn, nhưng tỷ lệ tai nạn nghiêm trọng vẫn có thể cao do các yếu tố như hành vi lái xe của con người, điều kiện thời tiết, thiết kế đường không tối ưu. Do đó, việc duy trì tình trạng mặt đường tốt chỉ là một yếu tố trong tổng thể các yếu tố cần được xem xét để giảm thiểu tai nạn giao thông.

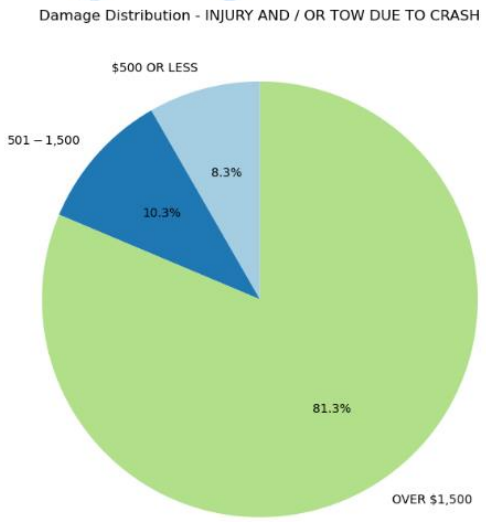
"Worn Surface" có tỷ lệ tai nạn nghiêm trọng cao hơn là do bề mặt mòn làm giảm độ bám đường, tăng độ trơn trượt, và có thể dẫn đến những tai nạn nghiêm trọng hơn, đặc biệt trong điều kiện thời tiết xấu hoặc khi đường không còn đủ độ bền.

Chọn thuộc tính này cho bài toán phân lớp.

5.3.2.12 DAMAGE

Mức độ thiệt hại không chỉ giúp xác định mức độ nghiêm trọng mà còn hỗ trợ trong việc phân tích sự tương quan giữa các loại tai nạn.





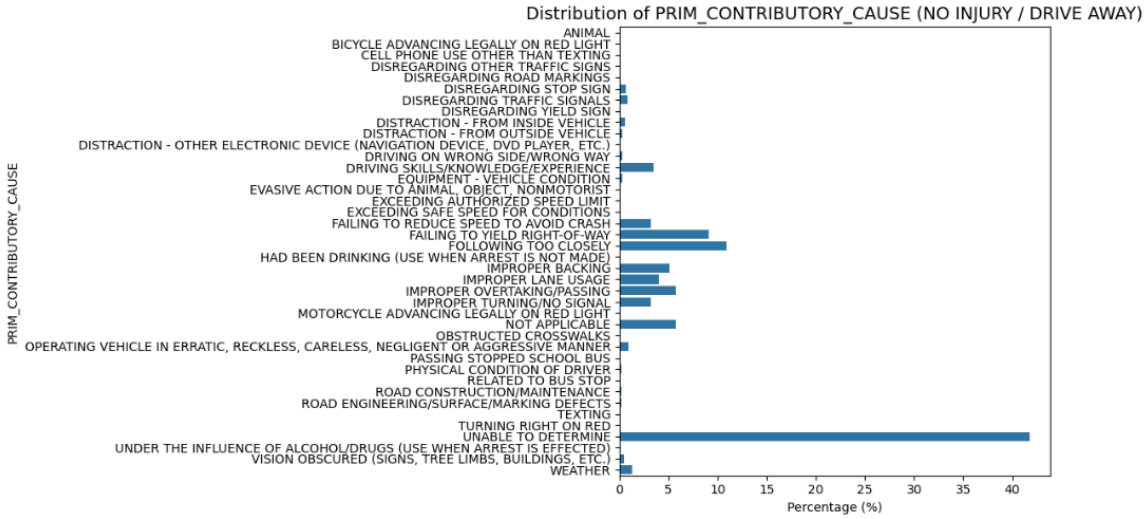
Với nhóm No Injury / Drive Away: Các vụ việc thuộc loại này có thiệt hại lớn hơn $1,500 - chiếm hơn một nửa tổng số vụ.Tuy nhiên, tỷ lệ các vụ thiệt hại thấp ($500 trở xuống) và trung bình ($501–$1,500) cũng đáng kể.

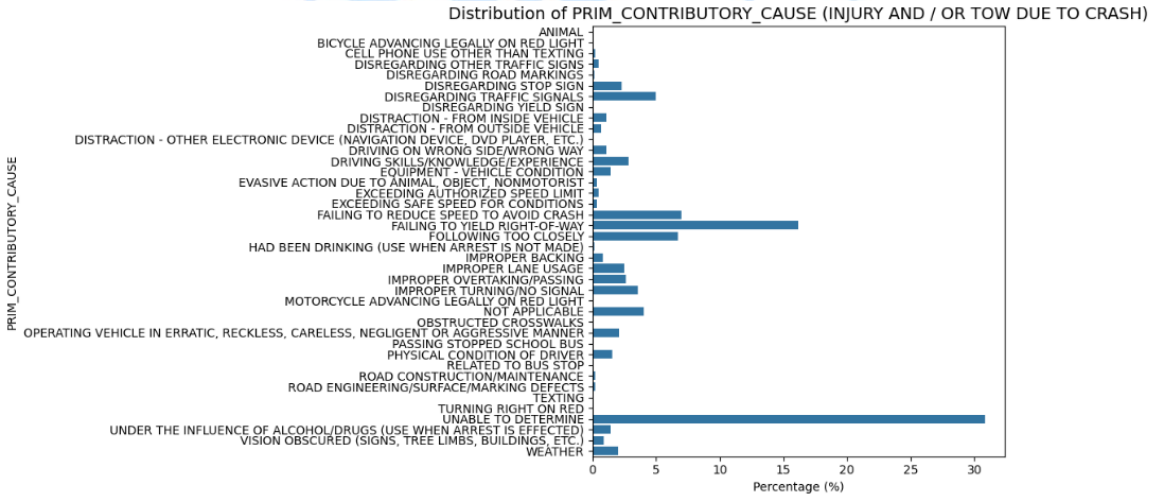
Injury and/or Tow Due to Crash: Phần lớn các vụ thuộc loại này gây thiệt hại nghiêm trọng (trên $1,500) – trên 80%. Tỷ lệ thiệt hại thấp ($500 trở xuống) và trung bình ($501–$1,500) rất thấp, điều này phản ánh rằng các vụ va chạm nghiêm trọng thường có mức độ thiệt hại cao hơn.

DAMAGE có sự khác biệt rõ rệt giữa các loại CRASH_TYPE. Chọn thuộc tính này cho mô hình.

5.3.2.13 PRIM_CONTRIBUTORY_CAUSE

PRIM_CONTRIBUTORY_CAUSE (Nguyên nhân chính đóng góp) là một thuộc tính quan trọng trong phân tích và phân loại các vụ tai nạn, vì nó mô tả nguyên nhân chính dẫn đến tai nạn.





Qua đồ thị, ta có thể dễ dàng nhận ra có mốt tương quan giữa các loại tai nạn và nguyên nhân.

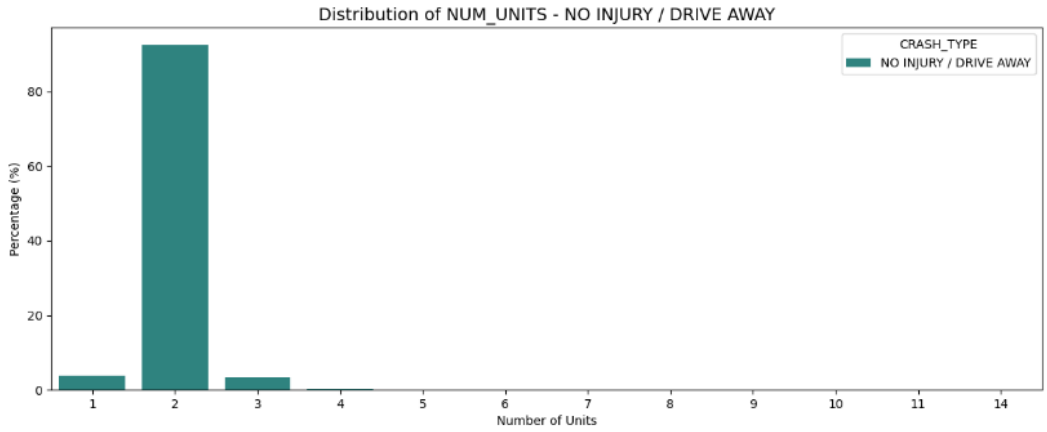
Các nguyên nhân như FALLING TO YIELD RIGHT-OF-WAY, EQUIPMENT - VEHICLE CONDITION, DISREGARDING STOP SIGN, DISREGARDING TRAFFIC SIGNAL có tỉ lệ tai nạn thuộc loại INJURY AND/OR TOW DUE TO CRASH cao hơn so với NO INJURY / DRIVE AWAY vì các nguyên nhân này thường liên quan đến những tình huống giao thông nguy hiểm hơn và dễ dẫn đến các sự cố nghiêm trọng hơn.

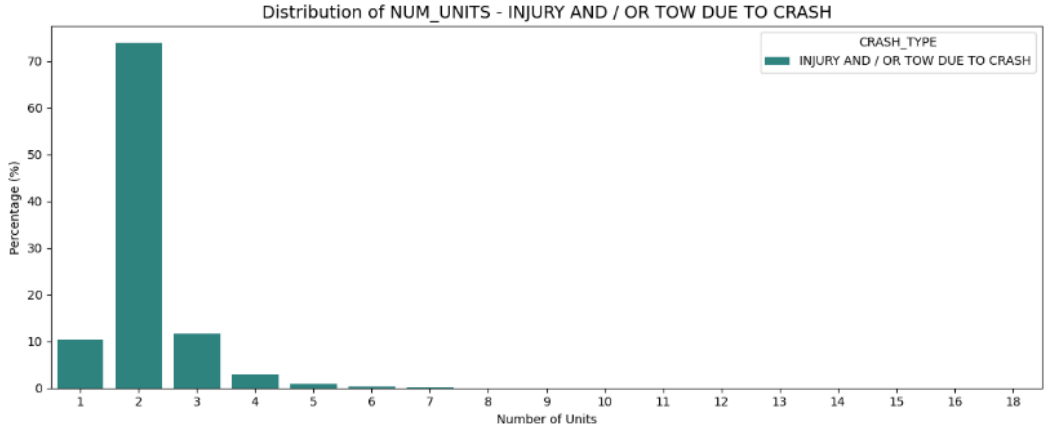
- Falling to Yield Right-of-Way (Không nhường quyền ưu tiên): Nguyên nhân này liên quan đến việc không tuân thủ các quy tắc giao thông về quyền ưu tiên. Khi một người điều khiển phương tiện không nhường quyền ưu tiên cho phương tiện khác, điều này có thể dẫn đến một vụ va chạm mạnh, đặc biệt là khi xe đang đi với tốc độ cao.
- Equipment - Vehicle Condition (Hư hỏng thiết bị - Tình trạng phương tiện): Khi phương tiện gặp vấn đề về thiết bị hoặc tình trạng kỹ thuật, như phanh hỏng, lốp xe mòn, hoặc động cơ gặp trục trặc, nó có thể làm cho người điều khiển không kiểm soát được phương tiện của mình, gây ra tai nạn. Những sự cố này có thể xảy ra mà không có dấu hiệu cảnh báo trước và dẫn đến các vụ va chạm nghiêm trọng.
- Disregarding Stop Sign (Phớt lờ biển báo dừng): Khi lái xe không tuân thủ biển báo dừng, họ có thể chạy vào giao lộ mà không giảm tốc độ hoặc quan sát các phương tiện khác. Điều này có thể dẫn đến các vụ va chạm mạnh, đặc biệt khi có phương tiện khác đang đi qua giao lộ.
- Disregarding Traffic Signal (Phớt lờ tín hiệu giao thông): Phớt lờ tín hiệu đèn giao thông có thể dẫn đến việc một phương tiện vượt qua giao lộ trong khi đèn tín hiệu báo đỏ, dẫn đến va chạm với các phương tiện khác đang di chuyển theo tín hiệu xanh. Việc này có thể tạo ra các tình huống va chạm mạnh, nhất là ở các giao lộ quan trọng hoặc đông đúc, nơi các phương tiện di chuyển với tốc độ cao.

Các nguyên nhân như phớt lờ quyền ưu tiên, sự cố thiết bị phương tiện, và phớt lờ biển báo dừng hoặc tín hiệu giao thông có tỉ lệ INJURY AND/OR TOW DUE TO CRASH cao hơn. Thuộc tính giúp ích trong việc phân loại các vụ tai nạn theo các nguyên nhân khác nhau. Chọn thuộc tính cho mô hình phân lớp.

5.3.2.14 NUM_UNITS

Thể hiện số lượng phương tiện, người tham gia hoặc đơn vị liên quan đến vụ tai nạn.





Trong giao thông, va chạm giữa 2 phương tiện là tình huống phổ biến nhất vì nó dễ dàng xảy ra do các yếu tố như điều kiện giao thông, sự thiếu chú ý, hoặc vi phạm giao thông. Vì vậy, tai nạn loại NO INJURY / DRIVE AWAY chủ yếu xảy ra giữa 2 phương tiện.

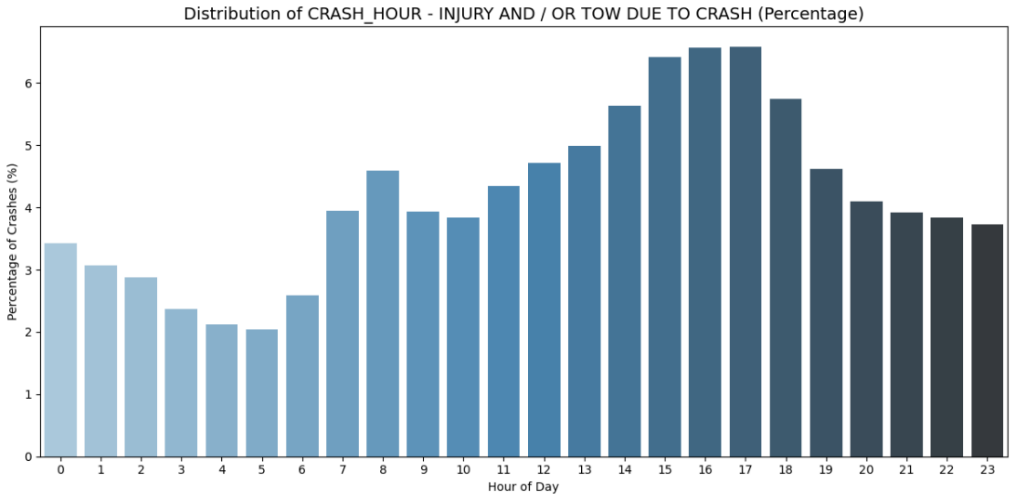
Ngược lại, các vụ tai nạn nghiêm trọng với INJURY AND / OR TOW DUE TO CRASH có thể xảy ra trong điều kiện phức tạp hơn, như tai nạn giữa nhiều phương tiện, có thể xảy ra trong các tình huống tắc nghẽn giao thông hoặc khi một phương tiện bị mất kiểm soát và va vào các phương tiện khác.

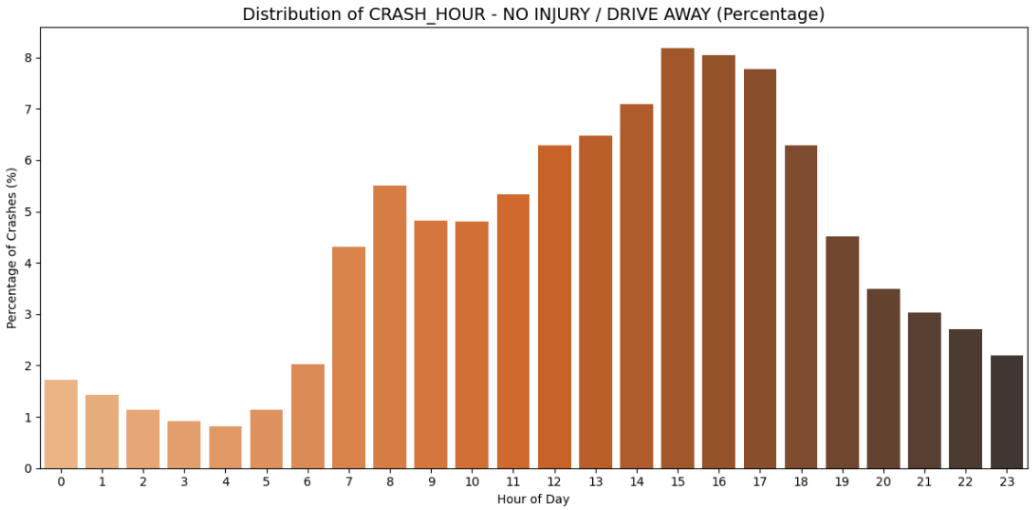
Sự phân bố NUM_UNITS (số lượng phương tiện liên quan trong mỗi vụ tai nạn) giữa hai loại tai nạn INJURY AND / OR TOW DUE TO CRASH và NO INJURY / DRIVE AWAY có sự khác biệt rõ rệt, và điều này có thể phản ánh các yếu tố khác nhau trong quy trình xảy ra và mức độ nghiêm trọng của các vụ tai nạn.

Ta chọn thuộc tính này cho bài toán phân lớp.

5.3.2.15 CRASH_HOUR

Các vụ tai nạn giao thông có thể phân bổ theo các khoảng thời gian trong ngày, và những đặc điểm này thường phản ánh các yếu tố khác nhau như tần suất giao thông, hành vi lái xe, điều kiện môi trường, và các yếu tố khác ảnh hưởng đến sự xuất hiện và mức độ nghiêm trọng của tai nạn.





Từ 12h đến 19h là khoảng thời gian trung bình mà nhiều người di chuyển giữa các công việc, ăn trưa, và quay lại nhà sau một ngày làm việc. Giờ này bao gồm cả giờ ăn trưa (12h - 14h) và giờ tan ca (16h - 19h), là những giờ cao điểm với lưu lượng giao thông dày đặc.Sự gia tăng trong lượng xe cộ (ô tô, xe máy, xe buýt) trên đường làm tăng khả năng xảy ra các vụ tai nạn giao thông, đặc biệt là ở các giao lộ, điểm đen hoặc các khu vực có mật độ giao thông cao.

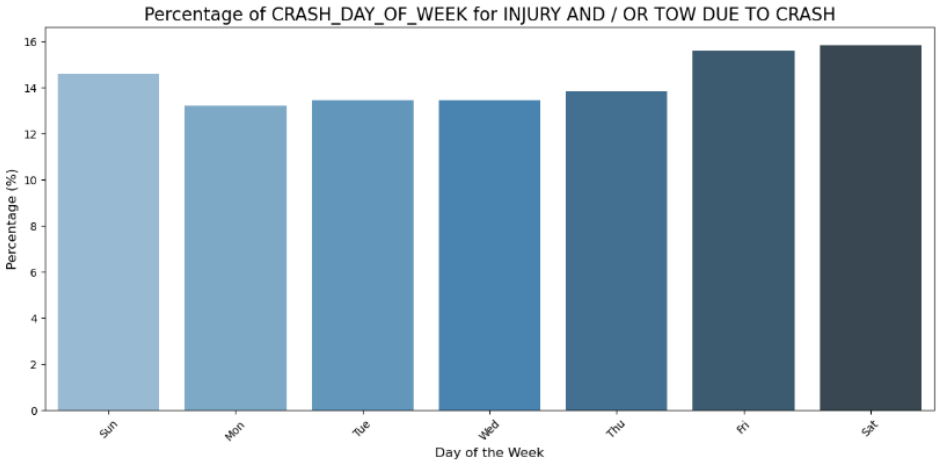
Sau một ngày dài làm việc, nhiều tài xế có thể mệt mỏi, thiếu sự tập trung và không phản ứng nhanh chóng với các tình huống khẩn cấp. Điều này đặc biệt rõ ràng vào ban đêm khi mức độ tỉnh táo và sự cảnh giác của người lái xe giảm đi. Từ 21h đến 6h sáng, nhiều tài xế có thể lái xe trong tình trạng say rượu, thuốc, hoặc các chất kích thích khác, làm giảm khả năng điều khiển phương tiện và tăng nguy cơ tai nạn nghiêm trọng. Vào ban đêm, giao thông ít hơn, và người lái xe có thể cảm thấy an toàn hơn, dẫn đến việc lái xe với tốc độ cao. Tốc độ cao kết hợp với tình trạng thiếu ánh sáng và tầm nhìn hạn chế có thể gây ra những vụ tai nạn nghiêm trọng.

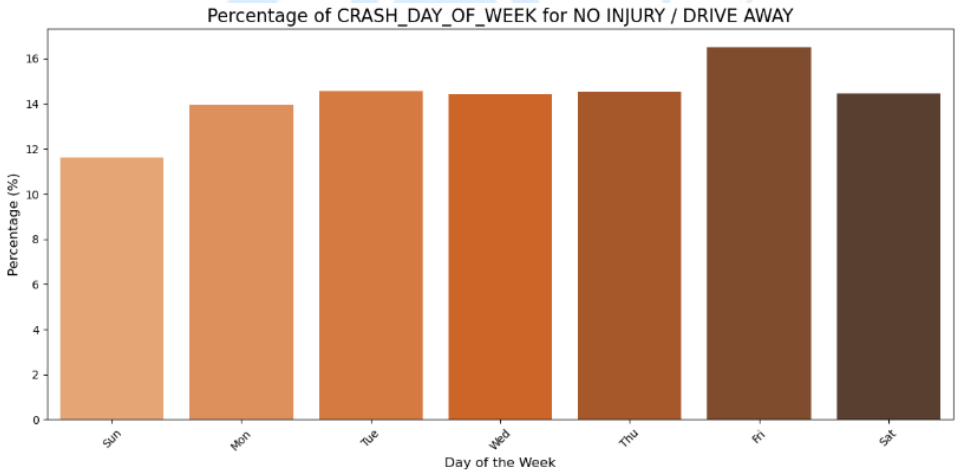
Tỉ lệ tai nạn nghiêm trọng cao hơn vào ban đêm (21h - 6h sáng) là do sự mệt mỏi, say rượu, tốc độ cao, và điều kiện đường xá nguy hiểm. Những tai nạn này có xu hướng nghiêm trọng hơn.

Có sự chênh lệch tỉ lệ ở khung giờ tai nạn cho mỗi loại tai nạn. Chọn thuộc tính này cho mô hình phân lớp.

5.3.2.16 CRASH_DAY_OF_WEEK

Phân tích tần suất tai nạn theo ngày trong tuần giúp nhận diện các xu hướng quan trọng, chẳng hạn như các ngày cuối tuần có thể là thời gian nguy hiểm hơn.





Thứ 7, Chủ nhật có số lượng vụ va chạm nhiều hơn các ngày khác. Trong khi đó, các vụ tai nạn giao thông nghiêm trọng ở Chủ Nhật và Thứ Hai chiếm tỉ lệ cao hơn các vụ tai nạn nhẹ.

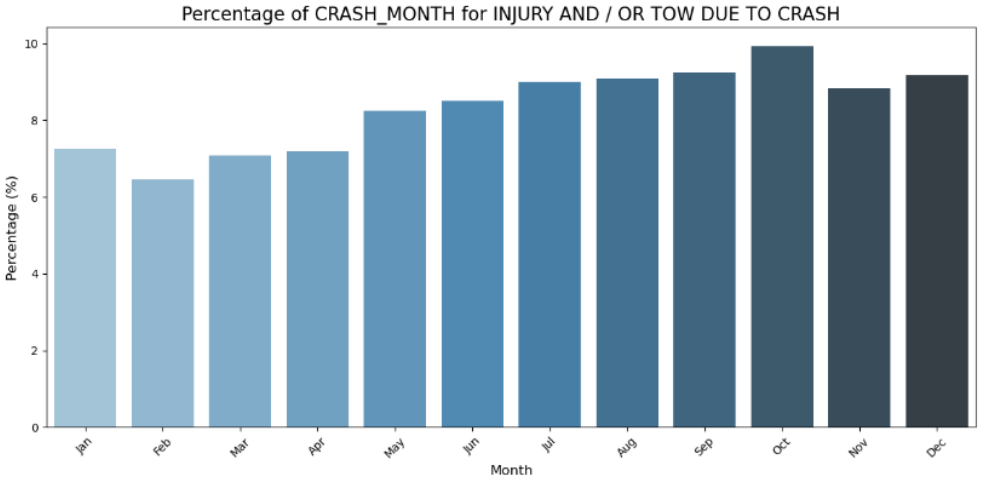
Chủ nhật có thể là ngày cuối tuần với số lượng giao thông lớn hơn, đặc biệt là trong các khu vực đô thị hoặc các khu vực có các điểm du lịch, khu nghỉ dưỡng có thể dẫn đến tình trạng tắc nghẽn giao thông, làm tăng khả năng xảy ra tai nạn. Những vụ tai nạn xảy ra trong điều kiện giao thông dày đặc hoặc không kiểm soát có thể nghiêm trọng hơn.

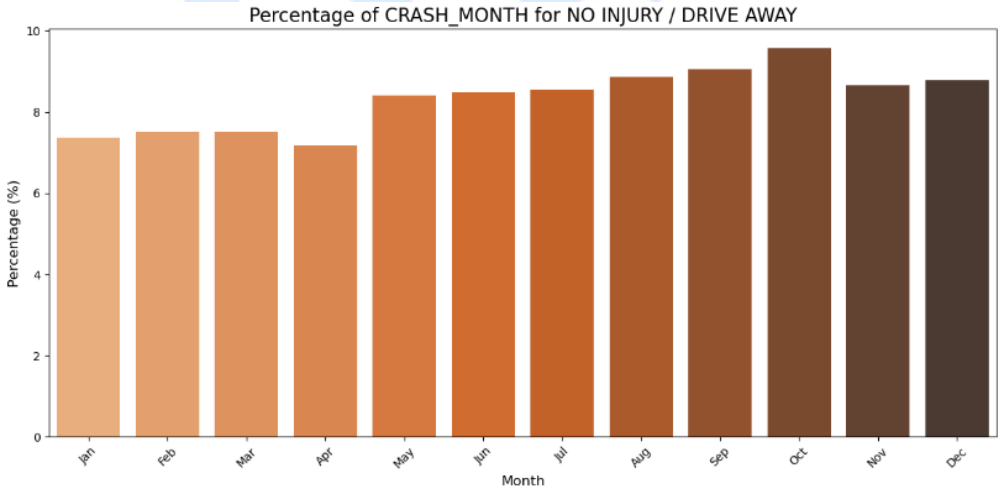
Thứ Hai có thể có một số yếu tố làm tăng nguy cơ tai nạn nghiêm trọng. Sau một kỳ nghỉ cuối tuần, một số người có thể cảm thấy mệt mỏi, hoặc thậm chí là vội vã khi di chuyển đến nơi làm việc, trường học, hoặc các hoạt động khác, hoặc sự thay đổi trong lịch trình làm việc cũng có thể góp phần vào các vụ tai nạn nghiêm trọng. Các lái xe có thể lái xe với tốc độ cao hoặc không tập trung, gây ra tai nạn nghiêm trọng.

Có sự khác biệt tuy không quá nhiều giữa thứ tự các ngày trong tuần và các loại tai nạn. Chọn thuộc tính này cho bài toán phân lớp.

5.3.2.17CRASH_MONTH

Việc phân tích dữ liệu theo tháng giúp xác định những tháng có số lượng tai nạn cao, từ đó có thể thực hiện các biện pháp phòng ngừa phù hợp.





Tháng 10 có số vụ tai nạn cao ở Chicago do sự kết hợp của nhiều yếu tố như thay đổi thời tiết (từ mùa thu sang mùa đông), tăng mật độ giao thông từ các sự kiện và hoạt động ngoài trời, điều kiện ánh sáng yếu vào buổi tối, và hành vi lái xe không an toàn. Những yếu tố này khiến nguy cơ xảy ra tai nạn tăng cao trong tháng 10.

Số lượng vụ tai nạn nghiêm trọng trong tháng 12 ở Chicago có thể tăng do sự kết hợp của các yếu tố sau: điều kiện thời tiết mùa đông khắc nghiệt, mật độ giao thông cao, hành vi lái xe không an toàn (say rượu hoặc mệt mỏi), và sự gia tăng các hoạt động lễ hội. Những yếu tố này tạo ra môi trường giao thông dễ xảy ra tai nạn, đặc biệt là trong những tháng cuối năm.

Vẫn có sự khác biệt ở tỉ lệ loại tai nạn ở một số tháng. Chọn thuộc tính cho bài toán phân lớp.

5.3.3 Tổng kết

Chọn các thuộc tính ('POSTED_SPEED_LIMIT', 'TRAFFIC_CONTROL_DEVICE', 'DEVICE_CONDITION', 'WEATHER_CONDITION', 'LIGHTING_CONDITION', 'FIRST_CRASH_TYPE', 'TRAFFICWAY_TYPE', 'ALIGNMENT', 'ROADWAY_SURFACE_COND', 'ROAD_DEFECT', 'CRASH_TYPE', 'DAMAGE', 'PRIM_CONTRIBUTORY_CAUSE', 'NUM_UNITS', 'CRASH_HOUR', 'CRASH_DAY_OF_WEEK', 'CRASH_MONTH') làm thuộc tính cho bài toán phân lớp nhị phân loại tai nạn, biến mục tiêu là CRASH_TYPE, dữ liệu có sự mất cân bằng.

Nhóm sẽ tiến hành áp dụng các thuật toán và các kỹ thuật xử lý như: Phương pháp tính trọng số lớp (Class Weighting), SMOTE (Synthetic Minority Oversampling Technique).

Phương pháp tính trọng số lớp (Class Weighting)

- **Ưu điểm:**
  - Dễ triển khai: Nhiều thuật toán (như Logistic Regression, Random Forest, SVM, XGBoost) hỗ trợ tham số class_weight mà không cần thay đổi dữ liệu gốc.
  - Hiệu quả với tập dữ liệu nhỏ: Không làm tăng kích thước dữ liệu, nên giảm thiểu nguy cơ overfitting.
  - Phù hợp cho dữ liệu mất cân bằng nhẹ hoặc vừa phải: Điều chỉnh trọng số lớp đảm bảo lớp thiểu số được chú ý mà không làm thay đổi cấu trúc dữ liệu.
  - Không thay đổi phân phối dữ liệu: Dữ liệu gốc không bị nhân bản hay bổ sung, giúp duy trì tính xác thực của dữ liệu.
- **Nhược điểm:**
  - Kém hiệu quả với dữ liệu mất cân bằng nghiêm trọng: Khi tỷ lệ lớp thiểu số quá nhỏ, điều chỉnh trọng số có thể không đủ để cải thiện hiệu suất.
  - Phụ thuộc vào thuật toán: Một số thuật toán không hỗ trợ trực tiếp trọng số lớp (hoặc cần tính toán thủ công).
  - Có thể gây overfitting ở lớp thiểu số: Trọng số quá cao cho lớp thiểu số có thể dẫn đến dự đoán sai, tăng dương tính giả.

SMOTE (Synthetic Minority Oversampling Technique)

- **Ưu điểm:**

- Giải quyết mất cân bằng dữ liệu nghiêm trọng: Tạo mẫu mới cho lớp thiểu số bằng cách nội suy từ các điểm dữ liệu hiện có.
- Giữ nguyên phân phối không gian dữ liệu: SMOTE tạo ra các điểm mới bằng cách kết hợp điểm hiện có và hàng xóm gần nhất, giữ cấu trúc lớp thiểu số.
- Cải thiện hiệu suất mô hình: Tăng cường dữ liệu lớp thiểu số giúp mô hình học được nhiều hơn từ lớp này, cải thiện Recall và AUC.

❖ **Nhược điểm:**

- Có thể tạo dữ liệu không chính xác: SMOTE dựa trên nội suy tuyến tính, dễ tạo ra điểm không thực tế, đặc biệt với dữ liệu không tuyến tính hoặc nhiễu.
- Tăng kích thước dữ liệu: Làm tăng khối lượng dữ liệu, dẫn đến chi phí tính toán cao và nguy cơ overfitting.
- Không hiệu quả nếu dữ liệu gốc quá nhỏ: Nếu lớp thiểu số có rất ít điểm, các điểm mới sẽ không đủ đa dạng, gây sai lệch mô hình.
- Không xử lý được vấn đề chồng lấn lớp (class overlap): Nếu các lớp chồng lấn nhau, SMOTE có thể làm tăng lỗi phân loại do tạo thêm các điểm trong vùng chồng lấn.

5.4 Thuật toán cho mô hình

5.4.1 Cơ sở lý thuyết

5.4.1.1 Precision (Độ chính xác)

Precision đo lường tỷ lệ các dự đoán dương tính (positive predictions) mà thực sự đúng. Nói cách khác, nó cho biết trong số những mẫu mà mô hình dự đoán là dương tính, bao nhiêu mẫu thực sự là dương tính.

Công thức:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **True Positives (TP):** Số lượng mẫu thực sự thuộc lớp dương tính mà mô hình dự đoán đúng.
- **False Positives (FP):** Số lượng mẫu thuộc lớp âm tính nhưng mô hình lại dự đoán là lớp dương tính.

Precision cao có nghĩa là khi mô hình dự đoán dương tính, nó rất ít khi dự đoán sai (False Positives thấp).

5.4.1.2 Recall

Recall đo lường tỷ lệ các mẫu dương tính thực sự mà mô hình có thể phát hiện. Nó cho biết trong số các mẫu thực sự thuộc lớp dương tính, bao nhiêu mẫu mô hình đã dự đoán chính xác.

Công thức:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- **True Positives (TP):** Số lượng mẫu thực sự thuộc lớp dương tính mà mô hình dự đoán đúng.
- **False Negatives (FN):** Số lượng mẫu thực sự thuộc lớp dương tính nhưng mô hình lại dự đoán là lớp âm tính.

Recall cao có nghĩa là mô hình phát hiện được hầu hết các mẫu thực sự thuộc lớp dương tính (ít bỏ sót mẫu).

5.4.1.3 F1-Score

F1-Score kết hợp cả hai chỉ số trên để cung cấp một cái nhìn tổng quát về hiệu suất của mô hình, đặc biệt khi dữ liệu không cân bằng. F1-Score cao chỉ ra rằng mô hình có sự cân bằng tốt giữa độ chính xác và khả năng thu hồi.

Công thức:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1-Score cao cho thấy mô hình không chỉ chính xác mà còn có khả năng thu hồi tốt các mẫu dương tính.

5.4.1.4 Accuracy (Độ chính xác tổng thể)

Accuracy là tỷ lệ của tất cả các dự đoán đúng so với tổng số dự đoán. Nó là chỉ số đơn giản nhất và được sử dụng rộng rãi để đánh giá mô hình, nhưng không phải lúc nào cũng phản ánh chính xác hiệu suất mô hình, đặc biệt là khi dữ liệu mất cân bằng.

Công thức:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}}$$

- **True Positives (TP):** Số lượng mẫu thực sự thuộc lớp dương tính mà mô hình dự đoán đúng.
- **True Negatives (TN):** Số lượng mẫu thực sự thuộc lớp âm tính mà mô hình dự đoán đúng.
- **False Positives (FP):** Số lượng mẫu thuộc lớp âm tính nhưng mô hình dự đoán là lớp dương tính.
- **False Negatives (FN):** Số lượng mẫu thuộc lớp dương tính nhưng mô hình lại dự đoán là lớp âm tính.

Accuracy cao có nghĩa là mô hình dự đoán đúng hầu hết các mẫu. Tuy nhiên, **accuracy** có thể bị sai lệch trong trường hợp dữ liệu mất cân bằng, vì nếu lớp dương tính chiếm tỷ lệ rất nhỏ, mô hình có thể đạt accuracy cao chỉ bằng cách dự đoán tất cả các mẫu là lớp âm tính mà không cần phân biệt.

5.4.1.5 ROC AUC Score

ROC AUC (Area Under the Curve) là diện tích dưới đường cong ROC (Receiver Operating Characteristic). Đường cong ROC thể hiện mối quan hệ giữa True Positive Rate (TPR) và False Positive Rate (FPR).

Công thức tính TPR và FPR:

❖ True Positive Rate (TPR) (hay còn gọi là Sensitivity hoặc Recall):

$$TPR = \frac{TP}{TP + FN}$$

Trong đó:

- **TP**: True Positives (Số lượng trường hợp dương tính được phân loại chính xác).
- **FN**: False Negatives (Số lượng trường hợp dương tính bị phân loại sai).

❖ False Positive Rate (FPR):

$$FPR = \frac{FP}{FP + TN}$$

Trong đó:

- **FP**: False Positives (Số lượng trường hợp âm tính bị phân loại sai).
- **TN**: True Negatives (Số lượng trường hợp âm tính được phân loại chính xác).

ROC AUC đo diện tích dưới đường cong ROC, một giá trị từ 0 đến 1:

- **AUC = 1**: Mô hình phân loại hoàn hảo.
- **AUC = 0.5**: Mô hình phân loại không tốt hơn việc phân loại ngẫu nhiên.
- **AUC < 0.5**: Mô hình phân loại tồi tệ hơn so với phân loại ngẫu nhiên.

5.4.1.6 Log Loss

Log Loss (hay còn gọi là Logarithmic Loss hoặc Cross-Entropy Loss) đo mức độ chính xác của mô hình dựa trên xác suất dự đoán cho các lớp.

Công thức tính Log Loss:

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)\right]$$

Trong đó:

- **N**: Số lượng mẫu.
- **yiy_iyi**: Nhãn thực tế của mẫu thứ iii (0 hoặc 1 cho bài toán phân loại nhị phân).
- **pip_ipi**: Xác suất dự đoán của mô hình cho lớp dương tính (tức là xác suất mẫu iii thuộc lớp dương tính).

Log Loss càng thấp càng tốt, và Log Loss bằng 0 có nghĩa là mô hình dự đoán chính xác 100% các xác suất.

5.4.1.7 Matthews Correlation Coefficient (MCC)

MCC là một chỉ số đánh giá hiệu suất phân loại nhị phân, đặc biệt hữu ích khi dữ liệu bị mất cân bằng.

Công thức tính MCC:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Trong đó:

- **TP**: True Positives.

- **TN**: True Negatives.
- **FP**: False Positives.
- **FN**: False Negatives.

Giải thích:

- **MCC = 1**: Mô hình phân loại hoàn hảo.
- **MCC = 0**: Mô hình phân loại không tốt hơn so với ngẫu nhiên.
- **MCC = -1**: Mô hình phân loại sai hoàn toàn.

5.4.1.8 Balanced Accuracy

Balanced Accuracy được sử dụng để đánh giá mô hình trong trường hợp dữ liệu không cân bằng, bằng cách tính trung bình của **True Positive Rate (TPR)** và **True Negative Rate (TNR)**.

Công thức tính Balanced Accuracy:

$$\text{Balanced Accuracy} = \frac{TPR + TNR}{2}$$

Trong đó:

- **True Positive Rate (TPR)** (hay Recall)

$$\frac{TP}{TP+FN}$$

- **True Negative Rate (TNR)**

$$\frac{TN}{TN+FP}$$

Giải thích:

- **Balanced Accuracy = 1**: Mô hình phân loại hoàn hảo cho cả lớp dương tính và âm tính.
- **Balanced Accuracy = 0.5**: Mô hình phân loại không tốt hơn việc phân loại ngẫu nhiên.

5.4.1.9 Tổng kết

Khi dữ liệu mất cân bằng (imbalanced), nghĩa là một lớp (class) có số lượng mẫu lớn hơn nhiều so với lớp còn lại, các chỉ số đánh giá mô hình có thể phản ánh không chính xác hiệu suất của mô hình, đặc biệt nếu chỉ sử dụng các chỉ số như accuracy. Do đó, việc lựa chọn các chỉ số đánh giá phù hợp là rất quan trọng.

- **ROC AUC**: Đo lường khả năng phân biệt giữa các lớp dương tính và âm tính qua nhiều ngưỡng.
- **Log Loss**: Đo mức độ chính xác của mô hình trong việc dự đoán xác suất cho mỗi lớp.
- **MCC**: Đánh giá mô hình trong các tình huống dữ liệu không cân bằng.
- **Balanced Accuracy**: Đảm bảo đánh giá mô hình phân loại công bằng khi dữ liệu không cân bằng.

Mức độ quan trọng của các chỉ số khi dữ liệu mất cân bằng:

- Quan trọng nhất: Matthews Correlation Coefficient (MCC) và Balanced Accuracy. Cả hai chỉ số này sẽ cung cấp cái nhìn chính xác về hiệu suất mô hình, đặc biệt khi một lớp chiếm ưu thế rõ rệt.
- Quan trọng tiếp theo: ROC AUC Score. AUC giúp đánh giá khả năng phân biệt giữa các lớp mà không bị ảnh hưởng mạnh bởi sự mất cân bằng lớp.
- Ít quan trọng hơn (nhưng vẫn hữu ích): Log Loss, mặc dù nó có thể bị ảnh hưởng trong các trường hợp dữ liệu mất cân bằng, nhưng nếu được sử dụng kết hợp với các chỉ số khác, nó có thể giúp đánh giá mức độ sai lệch của dự đoán xác suất.

5.4.2 Thuật toán Logistic Regression

Lý do lựa chọn: Logistic Regression là một thuật toán phân lớp đơn giản, hiệu quả và dễ triển khai cho các bài toán phân lớp nhị phân. Nó phù hợp với các bài toán mà mối quan hệ giữa các đặc trưng và nhãn lớp có thể được mô hình hóa theo cách tuyến tính, và nó cung cấp các lợi ích lớn trong việc giải thích kết quả dự đoán.

Logistic Regression có độ phức tạp tính toán thấp, vì vậy thời gian huấn luyện và dự đoán tương đối nhanh so với các thuật toán phức tạp khác như Random Forest. Điều này làm cho Logistic Regression trở thành một sự lựa chọn hợp lý khi có dữ liệu lớn hoặc yêu cầu tốc độ xử lý nhanh.

Quy trình:

Bước 1: Biến đổi cột biến mục tiêu thành nhị phân và mã hóa các cột dạng chuỗi thành số nguyên.

```python
# Biến đổi cột 'CRASH_TYPE' thành nhị phân
df['CRASH_TYPE'] = df['CRASH_TYPE'].apply(lambda x: 1 if x == 'INJURY AND / OR TOW DUE TO CRASH' else 0)

# Mã hóa các cột dạng chuỗi thành số nguyên
categorical_cols = [
    'POSTED_SPEED_LIMIT', 'TRAFFIC_CONTROL_DEVICE', 'DEVICE_CONDITION', 'WEATHER_CONDITION',
    'LIGHTING_CONDITION', 'FIRST_CRASH_TYPE', 'TRAFFICWAY_TYPE', 'ALIGNMENT', 'NUM_UNITS',
    'ROADWAY_SURFACE_COND', 'ROAD_DEFECT', 'DAMAGE', 'PRIM_CONTRIBUTORY_CAUSE',
    'CRASH_HOUR', 'CRASH_DAY_OF_WEEK', 'CRASH_MONTH'
]
```

Sau đó, ta tiến hành tách dữ liệu thành đầu vào (X) và nhãn mục tiêu (y) và chia dữ liệu thành tập huấn luyện và kiểm tra.

```python
# Tách dữ liệu thành đầu vào (X) và nhãn mục tiêu (y)
X = df.drop('CRASH_TYPE', axis=1)  # X: Các biến độc lập
y = df['CRASH_TYPE']               # y: Nhãn mục tiêu

# Chia dữ liệu thành tập huấn luyện và kiểm tra
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42, stratify=y
)
```

Bước 2: Thiết lập mô hình Logistic Regression.

Đầu tiên, ta sử dụng class_weight='balanced' để xử lý mất cân bằng dữ liệu.

```python
# Sử dụng class_weight='balanced' để xử lý mất cân bằng dữ liệu
logreg_model = LogisticRegression(class_weight='balanced', random_state=42, solver='liblinear')
```

Sau đó, định nghĩa bộ tham số cho GridSearchCV.

```python
# Định nghĩa bộ tham số cho GridSearchCV
param_grid_logreg = {
    'C': [0.1, 1, 10],            # Tham số điều chỉnh mức độ regularization
    'penalty': ['l1', 'l2'],      # Loại phạt (L1 hoặc L2)
    'solver': ['liblinear']       # Solver phù hợp với các lựa chọn penalty
}
```

Tiếp theo, thiết lập GridSearchCV để tìm kiếm tham số tối ưu.

```python
# Thiết lập GridSearchCV để tìm kiếm tham số tối ưu
grid_search_logreg = GridSearchCV(
    estimator=logreg_model,
    param_grid=param_grid_logreg,
    scoring='roc_auc',            # Sử dụng AUC làm tiêu chí đánh giá
    cv=5,                         # Số lượng tập chéo
    n_jobs=-1,                    # Sử dụng toàn bộ CPU
    verbose=1                     # Hiển thị quá trình tìm kiếm
)
```

Huấn luyện mô hình với tập huấn luyện.

```python
# Huấn luyện mô hình với tập huấn luyện
grid_search_logreg.fit(X_train, y_train)
```

Tiến hành lấy mô hình tốt nhất từ GridSearchCV.

```python
# Lấy mô hình tốt nhất từ GridSearchCV
best_logreg_model = grid_search_logreg.best_estimator_
```

Dự đoán trên tập kiểm tra.

```python
# Dự đoán trên tập kiểm tra
y_pred_logreg = best_logreg_model.predict(X_test)              # Nhãn dự đoán
y_prob_logreg = best_logreg_model.predict_proba(X_test)[:, 1]  # Xác suất dự đoán
```

Bước 3: In ra các chỉ số đánh giá mô hình.

```python
print("\nClassification Report (Logistic Regression):")
print(classification_report(y_test, y_pred_logreg))

# Tính các chỉ số đánh giá
accuracy_logreg = accuracy_score(y_test, y_pred_logreg)
auc_score_logreg = roc_auc_score(y_test, y_prob_logreg)
log_loss_logreg = log_loss(y_test, y_prob_logreg)
mcc_logreg = matthews_corrcoef(y_test, y_pred_logreg)
balanced_accuracy_logreg = balanced_accuracy_score(y_test, y_pred_logreg)

# In kết quả đánh giá
print(f"Accuracy (Logistic Regression): {accuracy_logreg:.2f}")
print(f"AUC (Logistic Regression): {auc_score_logreg:.2f}")
print(f"Log Loss (Logistic Regression): {log_loss_logreg:.2f}")
print(f"Matthews Correlation Coefficient (MCC): {mcc_logreg:.2f}")
print(f"Balanced Accuracy (Logistic Regression): {balanced_accuracy_logreg:.2f}")
```

Kết quả thuật toán Logistic Regression:

```
Fitting 5 folds for each of 6 candidates, totalling 30 fits

Best Parameters from GridSearchCV (Logistic Regression):
{'C': 10, 'penalty': 'l2', 'solver': 'liblinear'}

Classification Report (Logistic Regression):
              precision    recall  f1-score   support

           0       0.84      0.66      0.74    174953
           1       0.41      0.66      0.51     63534

    accuracy                           0.66    238487
   macro avg       0.63      0.66      0.62    238487
weighted avg       0.73      0.66      0.68    238487

Accuracy (Logistic Regression): 0.66
AUC (Logistic Regression): 0.72
Log Loss (Logistic Regression): 0.62
Matthews Correlation Coefficient (MCC): 0.28
Balanced Accuracy (Logistic Regression): 0.66
```

Nhận xét:

Mô hình Logistic Regression hiện tại đã đạt hiệu suất trung bình với AUC = 0.72 và Recall khá tốt cho lớp thiểu số.

Tuy nhiên, Precision thấp và MCC = 0.28 cho thấy cần phải cải thiện thêm nếu muốn có mô hình tốt hơn.

Với dữ liệu mất cân bằng (lớp 0: 174953, lớp 1: 63534), Logistic Regression không phải là thuật toán mạnh nhất. Dù đã tối ưu hóa bằng GridSearchCV.

5.4.3 Thuật toán Random Forest

Lý do lựa chọn: Random Forest có khả năng xử lý các bài toán phức tạp với dữ liệu có nhiều đặc trưng (features) hoặc dữ liệu phi tuyến tính (non-linear data).

Do kết hợp nhiều cây quyết định (decision trees), nó giảm thiểu các vấn đề mà một cây đơn lẻ có thể gặp phải, chẳng hạn như overfitting.

Random Forest kết hợp nhiều cây quyết định và sử dụng nguyên tắc số đông (majority voting) để đưa ra dự đoán cuối cùng. Điều này giúp tăng độ chính xác và giảm sai số so với việc sử dụng một cây quyết định đơn lẻ.

Quy trình:

Bước 1: Biến đổi cột biến mục tiêu thành nhị phân và mã hóa các cột dạng chuỗi thành số nguyên.

```
# Thay đổi giá trị của CRASH_TYPE
df['CRASH_TYPE'] = df['CRASH_TYPE'].apply(lambda x: 1 if x == 'INJURY AND / OR TOW DUE TO CRASH' else 0)

# Mã hóa các cột dạng chuỗi thành số (các cột khác ngoài CRASH_TYPE)
categorical_cols = ['POSTED_SPEED_LIMIT', 'TRAFFIC_CONTROL_DEVICE', 'DEVICE_CONDITION', 'WEATHER_CONDITION',
                    'LIGHTING_CONDITION', 'FIRST_CRASH_TYPE', 'TRAFFICWAY_TYPE', 'ALIGNMENT',
                    'ROADWAY_SURFACE_COND', 'ROAD_DEFECT', 'DAMAGE', 'PRIM_CONTRIBUTORY_CAUSE',
                    'CRASH_HOUR', 'CRASH_DAY_OF_WEEK', 'CRASH_MONTH']

label_encoder = LabelEncoder()
for col in categorical_cols:
    df[col] = label_encoder.fit_transform(df[col])
```

Sau đó, ta tiến hành tách dữ liệu thành đầu vào (X) và nhãn mục tiêu (y) và chia dữ liệu thành tập huấn luyện và kiểm tra.

```
# Tách X và y
X = df.drop('CRASH_TYPE', axis=1)  # X là dữ liệu đầu vào
y = df['CRASH_TYPE']  # y là nhãn mục tiêu

# Chia dữ liệu thành bộ huấn luyện và bộ kiểm tra (70% huấn luyện, 30% kiểm tra)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Bước 2: Thiết lập mô hình Random Forest.

Đầu tiên, ta sử dụng class_weight='balanced' để xử lý mất cân bằng dữ liệu.

```
# Tính toán trọng số lớp để xử lý mất cân bằng lớp trong dữ liệu
classes = np.unique(y_train)  # Lấy danh sách các lớp có trong y_train
class_weights = compute_class_weight('balanced', classes=classes, y=y_train)  # Tính toán trọng số lớp
class_weight_dict = dict(zip(classes, class_weights))  # Chuyển trọng số lớp thành từ điển
```

Sau đó, định nghĩa bộ tham số cho GridSearchCV.

```python
# Định nghĩa các tham số cho GridSearchCV
param_grid_rf = {
    'n_estimators': [50, 100, 200],  # Số lượng cây trong rừng
    'max_depth': [3, 6, 10],          # Độ sâu tối đa của cây
    'min_samples_split': [2, 5, 10],  # Số mẫu tối thiểu để chia một node
    'min_samples_leaf': [1, 2, 4],    # Số mẫu tối thiểu để tạo một lá
    'bootstrap': [True, False]        # Sử dụng bootstrap cho việc huấn luyện các cây
}
```

Tiếp theo, thiết lập GridSearchCV để tìm kiếm tham số tối ưu.

```python
# Tạo đối tượng GridSearchCV cho RandomForest
model = RandomForestClassifier(class_weight=class_weight_dict, random_state=42)
grid_search_rf = GridSearchCV(estimator=model, param_grid=param_grid_rf, cv=3, n_jobs=-1, verbose=2)
```

Huấn luyện mô hình.

```python
# Huấn luyện với GridSearchCV
grid_search_rf.fit(X_train, y_train)
```

Tiến hành lấy mô hình tốt nhất từ GridSearchCV.

```python
# In ra các tham số tốt nhất
print("Best parameters found: ", grid_search_rf.best_params_)
```

Dự đoán trên tập kiểm tra.

```python
# Dự đoán với mô hình tốt nhất
y_pred_best = grid_search_rf.best_estimator_.predict(X_test)
```

Bước 3: In ra các chỉ số đánh giá mô hình.

```python
# Hiển thị các báo cáo đánh giá mô hình tốt nhất
print("\nClassification Report for Best Model:")
print(classification_report(y_test, y_pred_best))

# Confusion Matrix cho mô hình tốt nhất
print("\nConfusion Matrix for Best Model:")
print(confusion_matrix(y_test, y_pred_best))

# ROC AUC Score cho mô hình tốt nhất
y_prob_best = grid_search_rf.best_estimator_.predict_proba(X_test)[:, 1]
roc_auc_best = roc_auc_score(y_test, y_prob_best)
print("\nROC AUC Score for Best Model:")
print(roc_auc_best)

# Log Loss cho mô hình tốt nhất
log_loss_value_best = log_loss(y_test, y_prob_best)
print("\nLog Loss for Best Model:")
print(log_loss_value_best)

# Matthews Correlation Coefficient (MCC) cho mô hình tốt nhất
mcc_value_best = matthews_corrcoef(y_test, y_pred_best)
print("\nMatthews Correlation Coefficient (MCC) for Best Model:")
print(mcc_value_best)

# Balanced Accuracy cho mô hình tốt nhất
balanced_accuracy_value_best = balanced_accuracy_score(y_test, y_pred_best)
print("\nBalanced Accuracy for Best Model:")
print(balanced_accuracy_value_best)
```

Kết quả thuật toán Random Forest:

```
Best parameters found:  {'bootstrap': True, 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 50}

Classification Report for Best Model:
              precision    recall  f1-score   support

           0       0.89      0.78      0.83    116676
           1       0.54      0.72      0.62     42316

    accuracy                           0.76    158992
   macro avg       0.71      0.75      0.72    158992
weighted avg       0.79      0.76      0.77    158992


Confusion Matrix for Best Model:
[[90563 26113]
 [11717 30599]]

ROC AUC Score for Best Model:
0.8243634085158027

Log Loss for Best Model:
0.5183585034297793

Matthews Correlation Coefficient (MCC) for Best Model:
0.46065010731187944

Balanced Accuracy for Best Model:
0.7496496446543577
```

Nhận xét:

Mô hình đạt hiệu suất tốt với ROC AUC = 0.82, Balanced Accuracy = 0.75, và F1-Score ổn định cho cả hai lớp.

Precision cho lớp 1 còn khá thấp (0.54), nhưng Recall cao (0.72), phù hợp nếu bài toán ưu tiên phát hiện lớp thiểu số (ví dụ: gian lận, bất thường).

Mô hình phân biệt tốt giữa hai lớp (ROC AUC cao).

MCC chưa cao phản ánh khả năng dự đoán đúng cả hai lớp còn hạn chế.

5.4.4 Thuật toán XGBoost

Lý do lựa chọn:

- XGBoost cung cấp công cụ mạnh mẽ để xử lý dữ liệu mất cân bằng:
  - Tham số scale_pos_weight: Tham số này cho phép điều chỉnh trọng số của các lớp dương (positive class) so với lớp âm (negative class). Khi lớp dương chiếm tỷ lệ nhỏ trong dữ liệu, việc tăng scale_pos_weight giúp mô hình tập trung hơn vào các mẫu thuộc lớp dương, cải thiện khả năng dự đoán cho lớp thiểu số.
  - Độ nhạy với lỗi của gradient boosting: XGBoost xây dựng cây dựa trên việc giảm thiểu lỗi trong các mẫu khó phân loại. Vì thế, các mẫu thuộc lớp thiểu số thường có lỗi lớn hơn, giúp mô hình tập trung học tốt hơn vào các mẫu này.
- XGBoost có hiệu suất cao và khả năng tối ưu hóa mạnh mẽ:

- Hỗ trợ phân lớp không đối xứng: XGBoost sử dụng hàm loss như *logistic loss* và cho phép tùy chỉnh trọng số cho các lớp trong quá trình huấn luyện. Điều này đặc biệt hữu ích cho bài toán mất cân bằng.
- Hiệu suất tốt trên dữ liệu lớn: Với các bài toán phân lớp mất cân bằng, số lượng dữ liệu thường lớn và tỷ lệ lớp thiểu số nhỏ. XGBoost có khả năng xử lý dữ liệu lớn hiệu quả nhờ tối ưu hóa tính toán và sử dụng bộ nhớ tốt.

- Các tham số linh hoạt để tối ưu hóa mô hình: XGBoost cung cấp nhiều tham số có thể được tinh chỉnh để cải thiện hiệu suất trên dữ liệu mất cân bằng:
  - max_depth: Giới hạn độ sâu của cây, tránh overfitting cho lớp thiểu số.
  - min_child_weight: Đảm bảo các nhánh chỉ được chia nếu đủ dữ liệu, giúp giảm noise từ các mẫu hiếm.
  - subsample: Chọn ngẫu nhiên một tỷ lệ mẫu để huấn luyện, giảm overfitting và cải thiện hiệu suất.
  - colsample_bytree: Chọn ngẫu nhiên một tập hợp các đặc trưng để xây dựng cây, giảm tương quan giữa các cây.
  - gamma: Điều chỉnh độ phức tạp của cây; giá trị lớn hơn giúp tránh overfitting.
- XGBoost hỗ trợ cân bằng bias và variance: Khả năng chống overfitting tốt: XGBoost có nhiều cơ chế kiểm soát overfitting, điều này đặc biệt hữu ích khi dữ liệu mất cân bằng dễ dẫn đến việc học thiên lệch.
  - Tham số eta (learning rate) giúp học chậm hơn và ổn định hơn.
  - Regularization (L1 và L2) kiểm soát độ phức tạp của mô hình.

5.4.4.1 Thuật toán XGBoost với Class Weighting

Quy trình:

Bước 1: Biến đổi cột biến mục tiêu thành nhị phân và mã hóa các cột dạng chuỗi thành số nguyên.

```
# Chuyển đổi cột CRASH_TYPE thành nhị phân
df['CRASH_TYPE'] = df['CRASH_TYPE'].apply(lambda x: 1 if x == 'INJURY AND / OR TOW DUE TO CRASH' else 0)

# Mã hóa các cột dạng chuỗi thành số
categorical_cols = ['POSTED_SPEED_LIMIT', 'TRAFFIC_CONTROL_DEVICE', 'DEVICE_CONDITION', 'WEATHER_CONDITION',
                    'LIGHTING_CONDITION', 'FIRST_CRASH_TYPE', 'TRAFFICWAY_TYPE', 'ALIGNMENT',
                    'ROADWAY_SURFACE_COND', 'ROAD_DEFECT', 'DAMAGE', 'PRIM_CONTRIBUTORY_CAUSE',
                    'CRASH_HOUR', 'CRASH_DAY_OF_WEEK', 'CRASH_MONTH']

label_encoder = LabelEncoder()
for col in categorical_cols:
    df[col] = label_encoder.fit_transform(df[col])
```

Sau đó, ta tiến hành tách dữ liệu thành đầu vào (X) và nhãn mục tiêu (y) và chia dữ liệu thành tập huấn luyện và kiểm tra.

```python
# Tách X và y
X = df.drop('CRASH_TYPE', axis=1)  # X là dữ liệu đầu vào
y = df['CRASH_TYPE']  # y là nhãn mục tiêu

# Chia dữ liệu thành tập huấn luyện và kiểm tra
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)
```

Bước 2: Thiết lập mô hình XGBoost với Class Weighting.

Đầu tiên, ta sử dụng class_weight='balanced' để xử lý mất cân bằng dữ liệu.

```python
# Tính trọng số lớp để xử lý mất cân bằng dữ liệu
classes = np.unique(y_train)
class_weights = compute_class_weight('balanced', classes=classes, y=y_train)
class_weight_dict = dict(zip(classes, class_weights))
```

Sau đó, định nghĩa bộ tham số cho GridSearchCV.

```python
# === Thiết lập GridSearchCV cho XGBoost ===
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [3, 5, 7],
    'learning_rate': [0.01, 0.1, 0.2],
    'subsample': [0.8, 1.0],
    'colsample_bytree': [0.8, 1.0],
    'scale_pos_weight': [1, 5, 10],
    'gamma': [0, 0.1, 0.3]
}
```

Tiếp theo, thiết lập GridSearchCV để tìm kiếm tham số tối ưu.

```python
xgb_model = XGBClassifier(random_state=42, use_label_encoder=False, eval_metric='logloss')

grid_search = GridSearchCV(
    estimator=xgb_model,
    param_grid=param_grid,
    scoring='roc_auc',
    cv=3,
    verbose=1,
    n_jobs=-1
)
```

Huấn luyện mô hình.

```python
# Huấn luyện mô hình với GridSearchCV
grid_search.fit(X_train, y_train)
```

Tiến hành lấy mô hình tốt nhất từ GridSearchCV.

```python
# === Mô hình tốt nhất ===
best_xgb_model = grid_search.best_estimator_
print("\nBest Parameters from GridSearchCV:")
print(grid_search.best_params_)
```

Dự đoán trên tập kiểm tra.

```python
# === Dự đoán và đánh giá ===
y_pred_xgb = best_xgb_model.predict(X_test)
y_prob_xgb = best_xgb_model.predict_proba(X_test)[:, 1]
```

Bước 3: In ra các chỉ số đánh giá mô hình.

```python
# Classification Report
print("\nClassification Report (XGBoost):")
print(classification_report(y_test, y_pred_xgb))

# Accuracy
accuracy = accuracy_score(y_test, y_pred_xgb)
print(f"Accuracy (XGBoost): {accuracy:.2f}")

# AUC Score
auc_score = roc_auc_score(y_test, y_prob_xgb)
print(f"AUC (XGBoost): {auc_score:.2f}")

# Log Loss
log_loss_value = log_loss(y_test, y_prob_xgb)
print(f"Log Loss (XGBoost): {log_loss_value:.2f}")

# Matthews Correlation Coefficient (MCC)
mcc_value = matthews_corrcoef(y_test, y_pred_xgb)
print(f"Matthews Correlation Coefficient (MCC): {mcc_value:.2f}")

# Balanced Accuracy
balanced_accuracy_value = balanced_accuracy_score(y_test, y_pred_xgb)
print(f"Balanced Accuracy (XGBoost): {balanced_accuracy_value:.2f}")
```

Kết quả thuật toán XGBoost với Class Weight:

```
Best Parameters from GridSearchCV:
{'colsample_bytree': 0.8, 'gamma': 0, 'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 300, 'scale_pos_weight': 1, 'subsample': 1.0}

Classification Report (XGBoost):
              precision    recall  f1-score   support

           0       0.84      0.92      0.88    174953
           1       0.71      0.52      0.60     63534

    accuracy                           0.82    238487
   macro avg       0.78      0.72      0.74    238487
weighted avg       0.81      0.82      0.81    238487

Accuracy (XGBoost): 0.82
AUC (XGBoost): 0.84
Log Loss (XGBoost): 0.41
Matthews Correlation Coefficient (MCC): 0.49
Balanced Accuracy (XGBoost): 0.72
```

Nhận xét:

- Độ Chính Xác (Accuracy) cao: Đạt được 0.82, đây là một giá trị khá ổn, nhưng độ chính xác không phải là thước đo tốt nhất trong trường hợp dữ liệu mất cân bằng.
- Khả năng phân biệt lớp tốt (AUC-ROC): AUC = 0.84 cho thấy mô hình phân biệt tốt giữa các lớp. AUC cao phản ánh khả năng nhận diện và phân biệt các lớp tốt hơn, đặc biệt là đối với các trường hợp không cân bằng.
- Hiệu suất của lớp 0: Precision: 0.84, Recall: 0.92 và F1-score: 0.88: Các chỉ số này cho thấy mô hình làm việc khá tốt với lớp đa số (lớp 0), đạt được độ chính xác và độ nhạy cao.
- Log Loss thấp (0.41): Log Loss thấp thể hiện rằng mô hình đưa ra các dự đoán xác suất khá chính xác.
- Recall và F1-score chưa cao. Precision ở mức khá: 0.71
- Mặc dù mô hình đạt Accuracy cao, nhưng sự mất cân bằng dữ liệu làm giảm khả năng phân loại chính xác đối với lớp thiểu số.

5.4.4.2 Thuật toán XGBoost với SMOTE

Quy trình:

Bước 1: Biến đổi cột biến mục tiêu thành nhị phân và mã hóa các cột dạng chuỗi thành số nguyên.

```
# Tạo cột nhãn nhị phân
df['CRASH_TYPE'] = df['CRASH_TYPE'].apply(
    lambda x: 1 if x in ['INJURY AND / OR TOW DUE TO CRASH'] else 0
)

# In số lượng mẫu theo nhãn
print("Số lượng mẫu theo nhãn:")
print(df['CRASH_TYPE'].value_counts())

# Tách dữ liệu và nhãn
X = df.drop(columns=['CRASH_TYPE'])
y = df['CRASH_TYPE']

# Mã hóa các cột phân loại (Label Encoding)
label_encoders = {}
for column in X.select_dtypes(include=['object']).columns:
    le = LabelEncoder()
    X[column] = le.fit_transform(X[column].astype(str))
    label_encoders[column] = le
```

Sau đó, ta tiến hành tách dữ liệu thành đầu vào (X) và nhãn mục tiêu (y) và chia dữ liệu thành tập huấn luyện và kiểm tra.

```
# Chia dữ liệu thành tập huấn luyện và kiểm tra
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Bước 2: Thiết lập mô hình XGBoost với SMOTE.

Áp dụng SMOTE cho tập huấn luyện và tính scale_pos_weight và huấn luyện mô hình.

```
# Áp dụng SMOTE chỉ cho tập huấn luyện (train set)
smote = SMOTE(random_state=42)
X_train_res, y_train_res = smote.fit_resample(X_train, y_train)

# Kiểm tra số lượng mẫu sau khi áp dụng SMOTE
print(f"Số lượng mẫu sau khi áp dụng SMOTE - lớp 0: {sum(y_train_res == 0)}, lớp 1: {sum(y_train_res == 1)}")

# Tính toán scale_pos_weight
pos_weight = (len(y_train_res) - sum(y_train_res)) / sum(y_train_res)
print(f"Scale_pos_weight: {pos_weight:.2f}")

# Huấn luyện mô hình XGBoost
model_xgb = xgb.XGBClassifier(
    random_state=42,
    eval_metric='logloss',
    scale_pos_weight=pos_weight  # Thêm trọng số cho lớp 1
)
```

Sau đó, định nghĩa bộ tham số cho GridSearchCV, thiết lập GridSearchCV để tìm kiếm tham số tối ưu.

```
# Tuning tham số với GridSearchCV
param_grid = {
    'learning_rate': [0.01, 0.1, 0.3],
    'max_depth': [3, 6, 10],
    'n_estimators': [50, 100, 200],
    'subsample': [0.7, 0.8, 1.0],
    'colsample_bytree': [0.7, 0.8, 1.0]
}

grid_search = GridSearchCV(estimator=model_xgb, param_grid=param_grid, cv=5, n_jobs=-1, verbose=2)
grid_search.fit(X_train_res, y_train_res)
```

Tiến hành lấy mô hình tốt nhất từ GridSearchCV và dự đoán trên tập kiểm tra.

```
# Dự đoán và đánh giá
best_model = grid_search.best_estimator_
y_pred_xgb = best_model.predict(X_test)
y_prob_xgb = best_model.predict_proba(X_test)[:, 1]  # Dự đoán xác suất lớp 1
```

Bước 3: In ra các chỉ số đánh giá mô hình.

```
# Báo cáo kết quả
print("Classification Report (XGBoost):")
print(classification_report(y_test, y_pred_xgb))
print(f"Accuracy (XGBoost): {accuracy_score(y_test, y_pred_xgb):.2f}")

# AUC Score
auc_score = roc_auc_score(y_test, y_prob_xgb)
print(f"AUC (XGBoost): {auc_score:.2f}")

# Log Loss
log_loss_value = log_loss(y_test, y_prob_xgb)
print(f"Log Loss: {log_loss_value:.2f}")

# Matthews Correlation Coefficient (MCC)
mcc_value = matthews_corrcoef(y_test, y_pred_xgb)
print(f"Matthews Correlation Coefficient (MCC): {mcc_value:.2f}")

# Balanced Accuracy
balanced_accuracy_value = balanced_accuracy_score(y_test, y_pred_xgb)
print(f"Balanced Accuracy: {balanced_accuracy_value:.2f}")
```

Kết quả thuật toán XGBoost với SMOTE:

```
Số lượng mẫu theo nhãn:
CRASH_TYPE
0    583177
1    211779
Name: count, dtype: int64
Số lượng mẫu sau khi áp dụng SMOTE - lớp 0: 408251, lớp 1: 408251
Scale_pos_weight: 1.00
Fitting 5 folds for each of 243 candidates, totalling 1215 fits
Best Parameters: {'colsample_bytree': 0.7, 'learning_rate': 0.1, 'max_depth': 10, 'n_estimators': 200, 'subsample': 0.8}
Classification Report (XGBoost):
              precision    recall  f1-score   support

           0       0.85      0.91      0.88    174926
           1       0.70      0.54      0.61     63561

    accuracy                           0.81    238487
   macro avg       0.77      0.73      0.74    238487
weighted avg       0.81      0.81      0.81    238487

Accuracy (XGBoost): 0.81
AUC (XGBoost): 0.84
Log Loss: 0.42
Matthews Correlation Coefficient (MCC): 0.50
Balanced Accuracy: 0.73
```

Nhận xét: Sau khi áp dụng SMOTE để xử lý mất cân bằng dữ liệu, mô hình XGBoost cho thấy sự thay đổi về hiệu suất phân loại.

- **Accuracy** là 0.81, có sự giảm nhẹ so với khi sử dụng **class weighting (0.82)**. Tuy nhiên, đây là kết quả khá ổn, đặc biệt khi xét đến việc cân bằng lại dữ liệu và đối phó với sự mất cân bằng giữa các lớp.
- **AUC = 0.84**, không thay đổi so với kết quả trước đó (0.84), cho thấy mô hình vẫn giữ được khả năng phân biệt giữa hai lớp khá tốt sau khi áp dụng SMOTE.
- **Log Loss** tăng nhẹ lên **0.42** từ **0.41**, cho thấy rằng mặc dù mô hình có khả năng phân biệt tốt, nhưng sau khi cân bằng lại lớp dữ liệu, có thể mô hình dự đoán một số mẫu không chính xác hơn. Tuy nhiên, mức độ tăng nhẹ này không quá đáng kể.
- **Precision lớp 0**: **0.85**, không thay đổi nhiều so với **0.84** trước đó.

- **Recall lớp 0**: **0.91**, cũng tương đối ổn định.
- **Precision lớp 1**: **0.70**, giảm nhẹ so với **0.71** trước đó. Mặc dù **precision** của lớp 1 giảm một chút, nhưng có sự cải thiện về **recall**.
- **Recall lớp 1**: **0.54**, cải thiện so với **0.52** trước đó, cho thấy SMOTE giúp mô hình nhận diện được nhiều trường hợp của lớp 1 hơn.
- **F1-score lớp 1**: **0.61**, cũng cải thiện so với **0.60** trước đó.
- **MCC** là **0.50**, tăng nhẹ so với **0.49** trước đó, cho thấy sự cải thiện trong việc nhận diện các trường hợp tai nạn gây thương tích nghiêm trọng (lớp 1).
- **Balanced Accuracy** là **0.73**, cải thiện so với **0.72** trước đó. Điều này phản ánh rằng mô hình đã cải thiện khả năng phân loại trong trường hợp mất cân bằng dữ liệu.

5.4.5 Thuật toán LightGBM

Lý do lựa chọn:

- Hiệu suất cao với tập dữ liệu lớn và mất cân bằng: LightGBM được thiết kế tối ưu cho dữ liệu lớn và mất cân bằng nhờ các cơ chế như:
    - Gradient-based One-Side Sampling (GOSS): Tập trung vào các mẫu có gradient lớn (các mẫu khó dự đoán), giảm số lượng mẫu không quan trọng trong quá trình huấn luyện. Điều này giúp cải thiện hiệu quả và duy trì độ chính xác.
    - Exclusive Feature Bundling (EFB): Tự động gộp các đặc trưng không xung đột, giảm kích thước đầu vào mà không mất thông tin.
- Cơ chế xử lý mất cân bằng qua tham số: LightGBM cung cấp tham số scale_pos_weight, cho phép gán trọng số cao hơn cho lớp thiểu số (positive class) trong hàm mất mát (loss function). Điều này giúp mô hình giảm thiểu hiện tượng bias về phía lớp đa số.
- Tốc độ huấn luyện và dự đoán nhanh: Với các thuật toán tối ưu hóa như GOSS và EFB, LightGBM vượt trội so với các thuật toán khác (như XGBoost) về tốc độ huấn luyện và dự đoán, đặc biệt khi làm việc với dữ liệu có số lượng mẫu và đặc trưng lớn.
- Khả năng điều chỉnh tham số tốt cho dữ liệu mất cân bằng: LightGBM cung cấp các tham số giúp mô hình hoạt động tốt với dữ liệu mất cân bằng:
    - scale_pos_weight: Cân bằng trọng số giữa các lớp.
    - is_unbalance: Tự động cân bằng trọng số các lớp dựa trên tỷ lệ mẫu. Khi is_unbalance=True, LightGBM tự điều chỉnh loss function.
    - max_depth, min_data_in_leaf: Giảm overfitting khi lớp thiểu số có số lượng mẫu ít.
- Độ chính xác cao: LightGBM thường đạt độ chính xác cao nhờ khả năng học tốt các tương tác phi tuyến phức tạp.

5.4.5.1 Thuật toán LightGBM với Class Weighting

Quy trình:

Bước 1: Biến đổi cột biến mục tiêu thành nhị phân và mã hóa các cột dạng chuỗi thành số nguyên.

```python
# Chuyển đổi cột CRASH_TYPE thành nhị phân
df['CRASH_TYPE'] = df['CRASH_TYPE'].apply(lambda x: 1 if x == 'INJURY AND / OR TOW DUE TO CRASH' else 0)

# Mã hóa các cột dạng chuỗi thành số
categorical_cols = ['POSTED_SPEED_LIMIT', 'TRAFFIC_CONTROL_DEVICE', 'DEVICE_CONDITION', 'WEATHER_CONDITION',
                    'LIGHTING_CONDITION', 'FIRST_CRASH_TYPE', 'TRAFFICWAY_TYPE', 'NUM_UNITS','ALIGNMENT',
                    'ROADWAY_SURFACE_COND', 'ROAD_DEFECT', 'DAMAGE', 'PRIM_CONTRIBUTORY_CAUSE',
                    'CRASH_HOUR', 'CRASH_DAY_OF_WEEK', 'CRASH_MONTH']

label_encoder = LabelEncoder()
for col in categorical_cols:
    df[col] = label_encoder.fit_transform(df[col])
```

Sau đó, ta tiến hành tách dữ liệu thành đầu vào (X) và nhãn mục tiêu (y) và chia dữ liệu thành tập huấn luyện và kiểm tra.

```python
# Chia dữ liệu thành tập huấn luyện và kiểm tra
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)
```

Bước 2: Thiết lập mô hình LightGBM Forest.

Đầu tiên, ta sử dụng class_weight='balanced' để xử lý mất cân bằng dữ liệu.

```python
# Tính trọng số lớp để xử lý mất cân bằng dữ liệu
classes = np.unique(y_train)
class_weights = compute_class_weight('balanced', classes=classes, y=y_train)
class_weight_dict = dict(zip(classes, class_weights))
```

Sau đó, định nghĩa bộ tham số cho GridSearchCV.

```python
# === Thiết lập GridSearchCV cho LightGBM ===
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [3, 5, 7],
    'learning_rate': [0.01, 0.1, 0.2],
    'subsample': [0.8, 1.0],
    'colsample_bytree': [0.8, 1.0],
    'class_weight': [None, class_weight_dict]
}
```

Tiếp theo, thiết lập GridSearchCV để tìm kiếm tham số tối ưu.

```python
lgb_model = LGBMClassifier(random_state=42)

grid_search = GridSearchCV(
    estimator=lgb_model,
    param_grid=param_grid,
    scoring='roc_auc',
    cv=3,
    verbose=1,
    n_jobs=-1
)
```

Huấn luyện mô hình.

```python
# Huấn luyện mô hình với GridSearchCV
grid_search.fit(X_train, y_train)
```

Tiến hành lấy mô hình tốt nhất từ GridSearchCV.

```python
# === Mô hình tốt nhất ===
best_lgb_model = grid_search.best_estimator_
print("\nBest Parameters from GridSearchCV:")
print(grid_search.best_params_)
```

Dự đoán trên tập kiểm tra.

```python
# === Dự đoán và đánh giá ===
y_pred_lgb = best_lgb_model.predict(X_test)
y_prob_lgb = best_lgb_model.predict_proba(X_test)[:, 1]
```

Bước 3: In ra các chỉ số đánh giá mô hình.

```python
# Classification Report
print("\nClassification Report (LightGBM):")
print(classification_report(y_test, y_pred_lgb))

# Accuracy
accuracy = accuracy_score(y_test, y_pred_lgb)
print(f"Accuracy (LightGBM): {accuracy:.2f}")

# AUC Score
auc_score = roc_auc_score(y_test, y_prob_lgb)
print(f"AUC (LightGBM): {auc_score:.2f}")

# Log Loss
log_loss_value = log_loss(y_test, y_prob_lgb)
print(f"Log Loss (LightGBM): {log_loss_value:.2f}")

# Matthews Correlation Coefficient (MCC)
mcc_value = matthews_corrcoef(y_test, y_pred_lgb)
print(f"Matthews Correlation Coefficient (MCC): {mcc_value:.2f}")

# Balanced Accuracy
balanced_accuracy_value = balanced_accuracy_score(y_test, y_pred_lgb)
print(f"Balanced Accuracy (LightGBM): {balanced_accuracy_value:.2f}")
```

Kết quả thuật toán LightGBM với Class Weghting:

```
Best Parameters from GridSearchCV:
{'class_weight': None, 'colsample_bytree': 1.0, 'learning_rate': 0.2, 'max_depth': 5, 'n_estimators': 300, 'subsample': 0.8}

Classification Report (LightGBM):
              precision    recall  f1-score   support

           0       0.84      0.92      0.88    174953
           1       0.71      0.52      0.60     63534

    accuracy                           0.82    238487
   macro avg       0.78      0.72      0.74    238487
weighted avg       0.81      0.82      0.81    238487

Accuracy (LightGBM): 0.82
AUC (LightGBM): 0.84
Log Loss (LightGBM): 0.41
Matthews Correlation Coefficient (MCC): 0.49
Balanced Accuracy (LightGBM): 0.72
```

Nhận xét:

- **Accuracy** là **0.82**, tương tự như kết quả của XGBoost trước đó (0.82). Đây là một kết quả khá cao, cho thấy mô hình LightGBM hoạt động tốt và phân loại chính xác hầu hết các mẫu.

- **AUC** đạt **0.84**, cho thấy mô hình có khả năng phân biệt giữa hai lớp rất tốt. AUC gần với **0.84** của XGBoost cho thấy khả năng phân biệt của cả hai mô hình là tương đương.
- **Log Loss** là **0.41**, tương đương với kết quả của XGBoost, chỉ ra rằng mô hình đã học được một cách khá hiệu quả khi dự đoán xác suất cho các lớp.
- **MCC** là **0.49**, phản ánh một sự kết hợp tốt giữa độ chính xác và độ nhạy trong các lớp. Đây là một giá trị khá hợp lý, cho thấy mô hình có thể phân biệt khá tốt giữa các lớp.
- **Balanced Accuracy** đạt **0.72**, tương đương với XGBoost, chỉ ra rằng mô hình vẫn duy trì khả năng phân biệt tốt giữa lớp đa số và thiểu số.

5.4.5.2 Thuật toán LightGBM với SMOTE

Quy trình:

Bước 1: Biến đổi cột biến mục tiêu thành nhị phân và mã hóa các cột dạng chuỗi thành số nguyên.

```
# Chuyển đổi cột CRASH_TYPE thành nhị phân
df['CRASH_TYPE'] = df['CRASH_TYPE'].apply(lambda x: 1 if x == 'INJURY AND / OR TOW DUE TO CRASH' else 0)

# Mã hóa các cột phân loại thành số
categorical_cols = ['POSTED_SPEED_LIMIT', 'TRAFFIC_CONTROL_DEVICE', 'DEVICE_CONDITION', 'WEATHER_CONDITION',
                    'LIGHTING_CONDITION', 'FIRST_CRASH_TYPE', 'TRAFFICWAY_TYPE', 'ALIGNMENT',
                    'ROADWAY_SURFACE_COND', 'ROAD_DEFECT', 'DAMAGE', 'PRIM_CONTRIBUTORY_CAUSE',
                    'CRASH_HOUR', 'CRASH_DAY_OF_WEEK', 'CRASH_MONTH']

label_encoder = LabelEncoder()
for col in categorical_cols:
    df[col] = label_encoder.fit_transform(df[col])
```

Sau đó, ta tiến hành tách dữ liệu thành đầu vào (X) và nhãn mục tiêu (y) và chia dữ liệu thành tập huấn luyện và kiểm tra.

```
# Tách X và y
X = df.drop('CRASH_TYPE', axis=1)
y = df['CRASH_TYPE']

# Chia dữ liệu thành tập huấn luyện và kiểm tra
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)
```

Bước 2: Thiết lập mô hình LightGBM với SMOTE.

Ta sử dụng SMOTE để xử lý mất cân bằng dữ liệu.

```
# === Áp dụng SMOTE để cân bằng dữ liệu ===
smote = SMOTE(random_state=42)
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)
```

Sau đó, định nghĩa bộ tham số cho GridSearchCV.

```
# === Thiết lập mô hình LightGBM ===
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [3, 5, 7],
    'learning_rate': [0.01, 0.1, 0.2],
    'subsample': [0.8, 1.0],
    'colsample_bytree': [0.8, 1.0]
}
```

Tiếp theo, thiết lập GridSearchCV để tìm kiếm tham số tối ưu.

```
lgb_model = LGBMClassifier(random_state=42)

grid_search = GridSearchCV(
    estimator=lgb_model,
    param_grid=param_grid,
    scoring='roc_auc',
    cv=3,
    verbose=1,
    n_jobs=-1
)
```

Huấn luyện mô hình.

```
# Huấn luyện mô hình trên dữ liệu đã được SMOTE
grid_search.fit(X_train_resampled, y_train_resampled)
```

Tiến hành lấy mô hình tốt nhất từ GridSearchCV.

```
# Mô hình tốt nhất
best_lgb_model = grid_search.best_estimator_
print("\nBest Parameters from GridSearchCV:")
print(grid_search.best_params_)
```

Dự đoán trên tập kiểm tra.

```
# === Dự đoán và đánh giá ===
y_pred_lgb = best_lgb_model.predict(X_test)
y_prob_lgb = best_lgb_model.predict_proba(X_test)[:, 1]
```

Bước 3: In ra các chỉ số đánh giá mô hình.

```
# Classification Report
print("\nClassification Report (LightGBM):")
print(classification_report(y_test, y_pred_lgb))

# Accuracy
accuracy = accuracy_score(y_test, y_pred_lgb)
print(f"Accuracy (LightGBM): {accuracy:.2f}")

# AUC Score
auc_score = roc_auc_score(y_test, y_prob_lgb)
print(f"AUC (LightGBM): {auc_score:.2f}")

# Log Loss
log_loss_value = log_loss(y_test, y_prob_lgb)
print(f"Log Loss (LightGBM): {log_loss_value:.2f}")

# Matthews Correlation Coefficient (MCC)
mcc_value = matthews_corrcoef(y_test, y_pred_lgb)
print(f"Matthews Correlation Coefficient (MCC): {mcc_value:.2f}")

# Balanced Accuracy
balanced_accuracy_value = balanced_accuracy_score(y_test, y_pred_lgb)
print(f"Balanced Accuracy (LightGBM): {balanced_accuracy_value:.2f}")
```

Kết quả thuật toán LightGBM với SMOTE:

```
Best Parameters from GridSearchCV:
{'colsample_bytree': 1.0, 'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 300, 'subsample': 0.8}

Classification Report (LightGBM):
              precision    recall  f1-score   support

           0       0.88      0.81      0.84    174953
           1       0.57      0.69      0.63     63534

    accuracy                           0.78    238487
   macro avg       0.73      0.75      0.74    238487
weighted avg       0.80      0.78      0.79    238487

Accuracy (LightGBM): 0.78
AUC (LightGBM): 0.83
Log Loss (LightGBM): 0.48
Matthews Correlation Coefficient (MCC): 0.48
Balanced Accuracy (LightGBM): 0.75
```

Nhận xét: Trong thuật toán LightGBM, SMOTE giúp cải thiện recall nhưng có thể gây nhiễu khi tạo mẫu lớp thiểu số, dẫn đến giảm accuracy và precision của lớp này.

- **Accuracy** đạt **0.78**, thấp hơn so với các mô hình như XGBoost (0.82) và LightGBM với **class weighting** (0.82). Điều này có thể là do SMOTE đã tạo ra các mẫu lớp thiểu số, nhưng điều này cũng có thể tạo ra một chút sự nhiễu loạn trong dữ liệu, làm giảm độ chính xác tổng thể.
- **AUC** là **0.83**, cho thấy rằng mô hình vẫn duy trì khả năng phân biệt tốt giữa hai lớp. Mặc dù accuracy giảm, khả năng phân biệt giữa lớp đa số và thiểu số vẫn rất mạnh mẽ.
- **Log Loss** là **0.48**, cao hơn một chút so với các kết quả trước đó. Điều này có thể phản ánh rằng các mẫu lớp thiểu số được tạo ra bằng SMOTE không hoàn toàn phù hợp với mô hình, dẫn đến việc dự đoán có độ chính xác thấp hơn một chút.
- **MCC** là **0.48**, phản ánh sự cân bằng khá tốt giữa độ chính xác và độ nhạy của mô hình, đặc biệt là trong một bài toán phân lớp mất cân bằng. Đây là một giá trị hợp lý và cho thấy mô hình LightGBM vẫn hoạt động tốt khi kết hợp với SMOTE.
- **Balanced Accuracy** đạt **0.75**, cao hơn nhiều so với mô hình LightGBM với **class weighting** (0.72). Điều này cho thấy SMOTE giúp cải thiện khả năng nhận diện lớp thiểu số, đặc biệt là khi lớp này không còn bị thiếu dữ liệu.

5.4.6 Tổng kết

	Accuracy	AUC	Log Loss	MCC	Balanced Accuracy
Logistic Regression	0.66	0.72	0.62	0.28	0.66
Random Forest	0.76	0.82	0.52	0.46	0.75
XGBoost (Class Weighting)	0.82	0.84	0.41	0.49	0.72
XGBoost (SMOTE)	0.81	0.84	0.42	0.50	0.73
LightGBM (Class Weighting)	0.82	0.84	0.41	0.49	0.72
LightGBM (SMOTE)	0.78	0.83	0.48	0.48	0.75

Bài toán phân lớp vụ tai nạn mà nhóm đã trình bày là bài toán phân loại nhị phân trong bối cảnh dữ liệu thường bị mất cân bằng (số lượng mẫu của một lớp lớn hơn lớp còn lại), và khả năng phân biệt giữa các thuộc tính không cao.

A.Logistic Regression:

- **Độ chính xác (Accuracy):** 0.66 - thấp nhất trong tất cả các mô hình.
- **AUC:** 0.72 - chỉ ra mô hình này có khả năng phân biệt yếu giữa các lớp.
- **Log Loss:** 0.62 - khá cao, cho thấy mô hình không dự đoán tốt xác suất.
- **MCC (Matthews Correlation Coefficient):** 0.28 - thấp, biểu thị mô hình này không xử lý tốt sự mất cân bằng dữ liệu.
- **Balanced Accuracy:** 0.66 - thấp nhất, chứng tỏ khả năng cân bằng giữa các lớp không tốt.

***Nhận xét:** Logistic Regression là mô hình cơ bản, nhưng hiệu quả kém khi dữ liệu mất cân bằng.

B.Random Forest:

- **Độ chính xác:** 0.76 - cải thiện rõ rệt so với Logistic Regression.
- **AUC:** 0.82 - tốt hơn, phản ánh khả năng phân biệt các lớp khá ổn định.
- **Log Loss:** 0.52 - thấp hơn, dự đoán xác suất tốt hơn.
- **MCC:** 0.46 - cải thiện lớn, mô hình xử lý dữ liệu mất cân bằng khá tốt.
- **Balanced Accuracy:** 0.75 - cao hơn Logistic Regression.

***Nhận xét:** Random Forest là mô hình ổn định và hiệu quả hơn trong bối cảnh dữ liệu mất cân bằng.

C.XGBoost (Class Weighting):

- **Độ chính xác:** 0.82 - cao nhất, cùng với LightGBM (Class Weighting).
- **AUC:** 0.84 - xuất sắc, thể hiện khả năng phân biệt tốt.
- **Log Loss:** 0.41 - thấp nhất, chỉ ra mô hình dự đoán xác suất chính xác nhất.
- **MCC:** 0.49 - cải thiện rõ rệt, xử lý tốt mất cân bằng dữ liệu.
- **Balanced Accuracy:** 0.72 - không cao bằng Random Forest nhưng vẫn khá tốt.

***Nhận xét:** XGBoost với Class Weighting là một lựa chọn mạnh mẽ với dữ liệu mất cân bằng.

D.XGBoost (SMOTE):

- **Độ chính xác:** 0.81 - cao, gần như tương đương với Class Weighting.
- **AUC:** 0.84 - tương tự Class Weighting.
- **Log Loss:** 0.42 - chỉ cao hơn nhẹ Class Weighting.
- **MCC:** 0.50 - cao nhất, cho thấy mô hình này xử lý mất cân bằng dữ liệu hiệu quả nhất.

- **Balanced Accuracy:** 0.73 - cao hơn Class Weighting.

*__Nhận xét:__ XGBoost với SMOTE tối ưu hóa kết quả nhờ xử lý tốt các lớp thiểu số, phù hợp với dữ liệu mất cân bằng.

E.LightGBM (Class Weighting):

- **Độ chính xác:** 0.82 - ngang bằng XGBoost (Class Weighting).
- **AUC:** 0.84 - ngang ngửa với XGBoost.
- **Log Loss:** 0.41 - tương đương với XGBoost.
- **MCC:** 0.49 - xử lý dữ liệu mất cân bằng tốt.
- **Balanced Accuracy:** 0.72 - tương tự XGBoost.

*__Nhận xét:__ LightGBM với Class Weighting mang lại kết quả tương đồng với XGBoost, cho thấy cả hai thuật toán đều hiệu quả với dữ liệu mất cân bằng.

F.LightGBM (SMOTE):

- **Độ chính xác:** 0.78 - thấp hơn XGBoost (SMOTE).
- **AUC:** 0.83 - chỉ kém hơn nhẹ.
- **Log Loss:** 0.48 - cao hơn các mô hình khác sử dụng SMOTE.
- **MCC:** 0.48 - thấp hơn XGBoost (SMOTE).
- **Balanced Accuracy:** 0.75 - tương đương Random Forest.

*__Nhận xét:__ LightGBM với SMOTE không vượt trội so với XGBoost (SMOTE) và có phần kém hiệu quả hơn.

Kết luận:

- Mô hình hiệu quả nhất: XGBoost (SMOTE) và XGBoost (Class Weighting) đều là lựa chọn tối ưu, với sự vượt trội về độ chính xác, AUC, MCC và Log Loss.
- Mô hình thay thế: LightGBM (Class Weighting) cũng đáng cân nhắc vì có hiệu năng gần như tương đương XGBoost.
- Kém hiệu quả: Logistic Regression không phù hợp với dữ liệu mất cân bằng.

Khuyến nghị: Sử dụng XGBoost (SMOTE) để đảm bảo xử lý tốt các lớp thiểu số và đạt hiệu quả cao nhất.

Mục tiêu hướng đến:

- Phân loại các vụ tai nạn thành hai nhóm:

Nhóm 1: NO INJURY / DRIVE AWAY (không có thương tích hoặc chỉ cần lái xe đi).

Nhóm 2: INJURY AND / OR TOW DUE TO CRASH (có thương tích hoặc cần kéo xe).

Điều này giúp đánh giá nhanh tình trạng của tai nạn, làm cơ sở để ưu tiên xử lý các vụ việc nghiêm trọng hơn.

- Tăng cường hiệu quả quản lý giao thông: Hỗ trợ các cơ quan quản lý giao thông và lực lượng cứu hộ, điều phối nguồn lực cứu hộ, y tế, và sửa chữa giao thông hiệu quả hơn.
- Đưa ra các biện pháp phòng ngừa như cải thiện cơ sở hạ tầng giao thông, cài đặt biển báo, và nâng cao ý thức lái xe.
- Hỗ trợ nghiên cứu và dự đoán tai nạn: Nghiên cứu các yếu tố dẫn đến tai nạn giao thông (điều kiện thời tiết, hành vi lái xe, vị trí tai nạn).

Tóm lại, bài toán phân lớp CRASH_TYPE nhằm mục đích hỗ trợ ra quyết định nhanh, chính xác trong việc quản lý giao thông, cứu hộ và phân tích dữ liệu để xây dựng xã hội an toàn hơn.

Nhìn chung, nhóm nghĩ rằng nhóm đã làm khá tốt dựa trên cấu trúc dữ liệu và nhóm có hài lòng với những kết quả đạt được.

TÀI LIỆU THAM KHẢO

1. Microsoft. (n.d.). *Analysis Services Multidimensional Modeling.* Retrieved from https://docs.microsoft.com/en-us/power-bi/visuals/power-bi-report-visualizations
2. CodePlex. (n.d.). *AdventureWorks for SQL Server 2014.* Retrieved from https://archive.codeplex.com
3. Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (2nd ed.). Wiley.
4. Microsoft. (n.d.). *Power BI report visualizations.* Retrieved from https://docs.microsoft.com/en-us/power-bi/visuals/power-bi-report-visualizations