**BỘ MÔN HỆ THỐNG THÔNG TIN – KHOA CÔNG NGHỆ THÔNG TIN**

**ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TP HCM**

Sinh viên thực hiện:

18127274 – Nguyễn Lê Đức Hoàng

18127241 – Nguyễn Đăng Triều

18127086 – Lê Thị Thùy Dương

ĐỒ ÁN/BÀI TẬP MÔN HỌC

HỌC KỲ I – NĂM HỌC 2021-2022

**HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ TUỆ KINH DOANH**

BẢNG THÔNG TIN CHI TIẾT CÁ NHÂN

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **MSSV** | **Họ Tên** | **Mức độ hoàn thành** | **Đóng góp** |
| 1 | 18127274 | Nguyễn Lê Đức Hoàng | 100% | 25% |
| 2 | 18127241 | Nguyễn Đăng Triều | 100% | 35% |
| 3 | 18127086 | Lê Thị Thùy Dương | 100% | 40% |

Mục lục

[**Chương 1. Tổng quan về đề tài** 4](#_Toc92468884)

[1. Giới thiệu bộ dữ liệu 4](#_Toc92468885)

[2. Mô tả dữ liệu 4](#_Toc92468886)

[3. Dữ liệu liên quan yêu cầu phân tích 17](#_Toc92468887)

[4. Phân tích profiling data 20](#_Toc92468888)

[**Chương 2. Quá trình SSIS** 22](#_Toc92468889)

[1. Source vào Stage 22](#_Toc92468890)

[2. Stage vào NDS 25](#_Toc92468891)

[3. NDS vào DDS 28](#_Toc92468892)

[**Chương 3: OLAP CUBE và MDX** 56](#_Toc92468893)

[1. Tạo ứng dụng SSAS 56](#_Toc92468894)

[2. Tạo Data Source View 60](#_Toc92468895)

[3. Tạo Cube 64](#_Toc92468896)

[4. Sữa bảng chiều 70](#_Toc92468897)

[5. Deploy cube lên server 71](#_Toc92468898)

[6. Tạo mới Measure 72](#_Toc92468899)

[7. Xóa dòng dữ liệu Unknow 73](#_Toc92468900)

[**Chương 4 Thực hiện ngôn ngữ MDX** 74](#_Toc92468901)

[1. Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng (Fatal, Serious, Slight) ở các Địa phương (Local\_Authority\_(District)) trong tất cả các năm. 74](#_Toc92468902)

[2. Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa Phương (Local\_Authority\_(District)) theo các Quý trong từng năm. 75](#_Toc92468903)

[3. Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng và Thời Điểm Trong Ngày (Morning: 5am-12pm, Afternoon: 12pm-5pm, Evening: 5pm-9pm, Night: 9pm-5am) trong các năm. 76](#_Toc92468904)

[4. Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng, Vùng (Urban\_or\_Rural\_Area), và Kiểu Đường (Road Type) trong các năm. 77](#_Toc92468905)

[5. Tổng hợp số lượng tai nạn theo Mục Đích Hành Trình (Journey Purpose) và Loại Phương Tiện (Vehicle\_Type). 78](#_Toc92468906)

[6. Dùng regional map để biểu diễn trực quan (bằng màu sắc) số lượng TNGT ở các vùng trong năm. 79](#_Toc92468907)

[**Chương 5: Trực quan hóa dữ liệu** 81](#_Toc92468908)

[1. Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng (Fatal, Serious, Slight) ở các Địa phương (Local\_Authority\_(District)) trong tất cả các năm. 81](#_Toc92468909)

[2. Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa Phương (Local\_Authority\_(District)) theo các Quý trong từng năm. 81](#_Toc92468910)

[3. Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng và Thời Điểm Trong Ngày (Morning: 5am-12pm, Afternoon: 12pm-5pm, Evening: 5pm-9pm, Night: 9pm-5am) trong các năm. 82](#_Toc92468911)

[4. Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng, Vùng (Urban\_or\_Rural\_Area), và Kiểu Đường (Road Type) trong các năm. 83](#_Toc92468912)

[5. Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng, Loại Nạn Nhân (Casualty Type) và Độ Tuổi trong các năm, Độ Tuổi được định nghĩa như sau: 84](#_Toc92468913)

[6. Tổng hợp số lượng tai nạn theo Mục Đích Hành Trình (Journey Purpose) và Loại Phương Tiện (Vehicle\_Type). 85](#_Toc92468914)

[7. Thống kê số lượng tai nạn theo Mức Độ Nghiêm Trọng, Loại Phương Tiện (Vehicle Type), Built-up Road trong các năm. 86](#_Toc92468915)

[8. Một vài biểu đồ cho tai nạn giao thông 86](#_Toc92468916)

[9. Dùng regional map để biểu diễn trực quan (bằng màu sắc) số lượng TNGT ở các vùng trong năm. 87](#_Toc92468917)

[**Chương 6: Data Mining** 88](#_Toc92468918)

# **Chương 1. Tổng quan về đề tài**

## Giới thiệu bộ dữ liệu

Dữ liệu về các vụ tai nạn giao thông ở UK trong khoảng 3-4 năm.

* Dữ liệu UK Car Accidents 2005 - 2015 (SV chỉ lấy dữ liệu 3-4 năm, hoặc lấy dữ

liệu từ 2011-2014 được cung cấp sẵn):

https://www.kaggle.com/silicon99/dft-accidentdata/discussion/28970?fbclid=IwAR1BvAiy8mEMy01XXAKtxLkX7Kx3kwPt3c3EYhwoxlWq5psikSAB2mVlF8A

* Dữ liệu LSOA-Postcode mapping:

https://geoportal.statistics.gov.uk/datasets/postcode-to-output-area-to-lower-layer-super-output-area-to-middle-layer-super-output-area-to-local-authority-district-august-2021-lookup-in-the-uk/about

* Dữ liệu UK-Postcodes:

<https://github.com/academe/UK-Postcodes/blob/master/postcodes.csv>

1. **Mô tả dữ liệu**

**2.1. Bảng Casualty**

|  |  |  |
| --- | --- | --- |
| **STT** | **Column** | **Description** |
| 1 | Sex of Casualty | Giới tính xảy ra thương vong   * Giá trị 1 đại diện cho Nam * Giá trị 2 đại diện cho Nữ * Giá trị -1 dữ liệu bị thiếu hoặc ngoài phạm vi |
| 2 | Age Band of Casualty | Các nhóm tuổi thương vong   * Giá trị 1 : nhóm tuổi từ 0 đến 5 tuổi * Giá trị 2 : nhóm tuổi từ 6 đến 10 tuổi * Giá trị 3 : nhóm tuổi từ 11 đến 15 tuổi * Giá trị 4 : nhóm tuổi từ 16 đến 20 tuổi * Giá trị 5 : nhóm tuổi từ 21 đến 25 tuổi * Giá trị 6 : nhóm tuổi từ 26 đến 35 tuổi * Giá trị 7 : nhóm tuổi từ 36 đến 45 tuổi * Giá trị 8 : nhóm tuổi từ 46 đến 55 tuổi * Giá trị 9 : nhóm tuổi từ 56 đến 65 tuổi * Giá trị 10 : nhóm tuổi từ 65 đến 75 tuổi * Giá trị 11 : nhóm tuổi từ 75 tuổi trở lên * Giá trị -1 : Mất dữ liệu hoặc nằm ngoài phạm vi |
| 3 | CasualtyType | Các loại nạn nhân thương vong   * Giá trị 0 Người đi bộ * Giá trị 1 người đi xe đạp * Giá trị 2 Xe mô tô từ 50cc trở xuống cho người lái hoặc chở khách * Giá trị 3 Người lái hoặc chở khách từ 125cc trở xuống * Giá trị 4 Xe mô tô trên 125 phân khối đến người lái hoặc chở khách đến 500 phân khối * Giá trị 5 Xe mô tô trên 500cc chở hoặc người lái * Giá trị 8 Taxi / Thuê xe riêng * Giá trị 9 Người ngồi trên ô tô * Giá trị 10 xe buýt nhỏ (8-16 chỗ ngồi) hành khách * Giá trị 11 Người ngồi trên xe buýt hoặc xe khách (17 chỗ ngồi trở lên) * Giá trị 16 Người cưỡi ngựa * Giá trị 17 Người ngồi trên xe công nông * Giá trị 18 Xe điện * Giá trị 19 Xe Van / Xe chở hàng (3,5 tấn mgw trở xuống) * Giá trị 20 Xe chở hàng (người trên 3,5 tấn và dưới 7,5 tấn) * Giá trị 21 Xe chở hàng (7,5 tấn mgw trở lên) * Giá trị 22 Người lái xe tay ga cơ động * Giá trị 23 Người lái xe mô tô điện hoặc chở khách * Giá trị 90 Người ngồi trên xe khác * Giá trị 97 Xe mô tô - người lái hoặc hành khách phân khối không xác định * Giá trị 98 Xe chở hàng (không rõ trọng lượng) |

**2.2. Bảng Accident**

|  |  |  |
| --- | --- | --- |
| **STT** | **Column** | **Description** |
| 1 | Accident Severity | Đây là thuộc tính chỉ ra mức độ nghiêm trọng của 1 vụ tai nạn, gồm có 3 mức độ là(fatal, serious, slight). |
| 2 | Local Authority (District) | Thuộc tính chỉ ra tên của địa phương. |
| 3 | Time | Thời gian của vụ tai nạn.(giờ-phút) |
| 4 | Date | Ngày xảy ra tai nạn(DD/MM/YYYY). |
| 5 | Urban or Rural Area | Thuộc tính lưu thông tin khu vực xảy ra tai nạn.   * 1: Urban - khu đô thị * 2: Rural - nông thôn * 3: Unallocated - chưa phân bố |
| 6 | Road Type | Thuộc tính lưu thông tin loại đường nơi xảy ra tai nạn.   * 1: Roundabout - bùng binh * 2: One way street - đường một chiều * 3: Dual carriageway - đường hai chiều * 6: Single carriageway - đường đơn * 7: Slip road - đường trượt * 9: Unknown - không xác định * 12: One way street/Slip road - đường một chiều/ đường trượt * -1: Data missing or out of range - dữ liệu bị thiếu hoặc nằm ngoài phạm vi |
| 7 | Police Force | Chỉ ra lực lượng cảnh sát làm ở hiện trường tai nạn |
| 8 | Number of Vehicles | Chỉ ra số phương tiện trong vụ tai nạn |
| 9 | Number of Casualties | Chỉ ra số nạn nhân trong vụ tai nạn |
| 10 | Local\_Authority\_(Highway) | Chỉ ra thông tin của xa lộ. |
| 11 | Speed limit | Chỉ ra tốc độ giới hạn trên đường. |
| 12 | Light Conditions | Chỉ ra điều kiện ánh sáng trong lúc xảy ra tai nạn. |
| 13 | Weather Conditions | Chỉ ra điều kiện thời tiết trong lúc xảy ra tai nạn. |
| 14 | Road Surface Conditions | Chỉ ra điều kiện mặt đường trong lúc xảy ra tai nạn |
| 15 | LSOA | Chỉ ra vị trí tai nạn xảy ra |

**2.3. Bảng Vehicle**

|  |  |  |
| --- | --- | --- |
| **STT** | **Column** | **Description** |
| 1 | Journey Purpose of Driver | Chỉ ra mục đích của hành trình của người điều khiển phương tiện. và được chia thành các loại sau đây.   * 1: Journey as part of work * 2: Commuting to/from work - Đi làm đến / đi từ nơi làm việc * 3: Taking pupil to/from school - Đưa học sinh đến / từ trường * 4: Pupil riding to/from school - Cưỡi ngựa đến / từ trường * 5: Other - khác * 6: Not known - chưa biết * 15: Other/Not known (2005-10) - khác/ chưa biết * -1: Data missing or out of range - dữ liệu bị thiếu hoặc ngoài phạm vi |
| 2 | Vehicle Type | Chỉ ra loại phương tiện sử dụng, và được chia thành các loại sau đây.   * 1: Pedal cycle - xe đạp * 2: Motorcycle 50cc and under - Xe mô tô 50cc trở xuống * 3: Motorcycle 125cc and under - Xe mô tô 125cc trở xuống * 4: Motorcycle over 125cc and up to 500cc - Xe mô tô trên 125cc đến 500cc * 5: Motorcycle over 500cc - Xe máy trên 500cc * 8: Taxi/Private hire car - Taxi / thuê xe riêng * 9: Car - Xe hơi * 10: Minibus (8 - 16 passenger seats) - Xe buýt nhỏ (8-16 chỗ ngồi) * 11: Bus or coach (17 or more pass seats) - Xe buýt hoặc xe khách (17 chỗ ngồi trở lên) * 16: Ridden horse - Ngựa hoang * 17: Agricultural vehicle - Xe công nông * 18: Tram - Xe điện * 19: Van / Goods 3.5 tonnes mgw or under - Van / Hàng hóa 3,5 tấn mgw trở xuống * 20: Goods over 3.5t. and under 7.5t - Hàng trên 3,5t. và dưới 7,5t * 21: Goods 7.5 tonnes mgw and over - Hàng hóa 7,5 tấn mgw trở lên * 22: Mobility scooter - Xe tay ga di động * 23: Electric motorcycle - Xe máy điện * 90: Other vehicle - Xe khác * 97: Motorcycle - unknown cc - Xe máy - phân khối không xác định * 98: Goods vehicle - unknown weight - Xe chở hàng - không rõ trọng lượng |
| 3 | Junction Location | Chỉ ra vị trí giao lộ của phương tiện tai nạn, và được chia thành các loại sau đây   * 0: Not at or within 20 metres of junction - Không ở hoặc trong vòng 20 mét của đường giao nhau * 1:Approaching junction or waiting/parked at junction approach - Tiếp cận đường giao nhau hoặc chờ / đỗ ở đường giao nhau * 2:Cleared junction or waiting/parked at junction exit - Đường giao nhau đã thông thoáng hoặc chờ / đỗ ở lối ra đường giao nhau * 3: Leaving roundabout - Rời khỏi bùng binh * 4: Entering roundabout - Đi vào bùng binh * 5: Leaving main road - Ra khỏi đường chính * 6: Entering main road - Vào đường chính * 7: Entering from slip road - Đi vào từ đường trượt * 8: Mid Junction - on roundabout or on main road -Giao lộ Giữa - trên bùng binh hoặc trên đường chính * -1: Data missing or out of range - dữ liệu bị mất hoặc nằm ngoài phạm vi |
| 4 | Skidding and Overturning | Chỉ ra phương tiện bị trượt hoặc lật, và được chia thành các loại sau đây.   * 0: None * 1: Skidded * 2: Skidded and overturned * 3: Jackknifed * 4: Jackknifed and overturned * 5: Overturned * -1: Data missing or out of range |
| 5 | Was Vehicle Left Hand Drive | Chỉ ra người điều khiển phương tiện có lái xe bằng tay trai hay không.   * 1: No * 2: Yes * -1: Data missing or out of range - dữ liệu bị thiếu hoặc nằm ngoài phạm vi |
| 6 | Sex of Driver | Chỉ ra giới tính của người điều khiển phương tiện.   * 1 : Male - Nam * 2 : Female - Nữ * 3:Not known - Chưa biết * -1: Data missing or out of range - dữ liệu bị thiếu hoặc nằm ngoài phạm vi |
| 7 | Age of Driver | Chỉ ra tuổi của người điều khiển phương tiện |
| 8 | Age Band of Driver | Thuộc tính chỉ ra nhóm tuổi của nạn nhân, được lưu trữ theo quy luật code từ 1-11 (code – độ tuổi).   * 1: 0 - 5 * 2: 6 - 10 * 3: 11 – 15 * 4: 16 – 20 * 5: 21 – 25 * 6: 26 – 35 * 7: 36 – 45 * 8: 46 – 55 * 9: 56 – 65 * 10: 66 – 75 * 11: Over 75 * -1: Data missing or out range |
| 9 | Age of Vehicle (manufacture) | Chỉ ra số tuổi của phương tiện. |

**2.4. Bảng Postcodes (UK Postcodes)**

|  |  |  |
| --- | --- | --- |
| **STT** | **Column** | **Description** |
| 1 | city | Thị trấn(town) hoặc thành phố (city) |
| 2 | county | Quận(the county) hoặc khu vực đô thị (city conurbation) hoặc khu vực thành phố (municiple region). |
| 3 | country\_code | Mã quốc gia gồm 3 chữ cái |
| 4 | country\_name | Tên của quốc gia |
| 5 | region\_code | Mã vùng ( không chính thức ) |
| 6 | region\_name | Tên vùng |
| 7 | postcode | Mã bưu điện |
| 8 | easting | Giá trị hướng đông - tối đa 7 chữ số |
| 9 | northing | Giá trị hướng bắc - tối đa 7 chữ số |
| 10 | latitude | Góc vĩ độ tính bằng độ |
| 11 | longitude | Góc kinh độ tính bằng độ |
| 12 | iso3166-2 | Mã quốc gia ISO |

**2.5. Bảng PCD\_OA\_LSOA\_MSOA\_LAD\_AUG21\_UK\_LU**

|  |  |  |
| --- | --- | --- |
|  | **Column** | **Description** |
| 1 | pcd7 | Mã postcode 7 ký tự |
| 2 | pcd8 | Mã postcode 8 ký tự |
| 3 | pcds | Khoảng cách giữa quận và đơn vị ngành |
| 4 | dointr | Ngày postcode có hiệu lực |
| 5 | doterm | Ngày postcode hết hiệu lực |
| 6 | usertype | Loại người dùng |
| 7 | oa11cd | Mã Output Area dân số trung bình khu vực vào khoảng 310 người |
| 8 | lsoa11cd | Mã Lower layer super output area dân số vào khoảng 1500 |
| 9 | msoa11cd | Mã Middle layer super output area dân số vào khoảng 7500 |
| 10 | ladcd | Mã vùng đất |
| 11 | lsoa11nm | Tên của Lower layer super output area |
| 12 | msoa11nm | Tên của Middle layer super output area |
| 13 | ladnm | Tên vùng đất |

## 3. Dữ liệu liên quan yêu cầu phân tích

**3.1 Report the number of calsulities by Severity (Fatal, Serious, Slight) in the LocalAuthority Districts over years**

Các kiểu dữ liệu liên quan:

* Number of Casualities (Casualty table)
* Casualty Severity (Casualty table)
* Local Authority Districts (Accident table)
* Year (Accident table, column Date)

**3.2 Report the number of calsulities by Severity (Fatal, Serious, Slight) in the Local Authority Districts by Quarters in years.**

Các kiểu dữ liệu liên quan:

* Number of Casualities (Casualty table)
* Casualty Severity (Casualty table)
* Local Authority Districts (Accident table)
* Quarters, Years (Accident table, column Date) - Dữ liệu chưa có, tự tạo thêm

**3.3 Report the number of faltal calsulities by Gender, Casualty Type and Age Band over years.**

Các kiểu dữ liệu liên quan:

* Number of Casualities (Casualty table)
* Casualty Severity (Casualty table, Column Casualty Severity, value Fatal) - Dữ liệu bị thiếu sót
* Sex of Casualty(Casualty table)
* Casualty Type (Casualty table)
* Age\_Band\_of\_Casualty (Casualty table)
* Years (Accident table, column Date) - Dữ liệu chưa có, tự tạo thêm

**3.4 Report the number of accidents by Severity and Time of Day (Morning: 5am- 12pm, Afternoon: 12pm-5pm, Evening: 5pm-9pm, Night: 9pm-5am) over years.**

Các kiểu dữ liệu liên quan:

* Number of Accidents(Accident table)
* Accident Severity (Accident tablel)
* Time of Day (Accident table) - Dữ liệu chưa có, tự tạo thêm từ cột Time
* Years (Accident table, column Date) - Dữ liệu chưa có, tự tạo thêm

**3.5 Report the number of accidents by Severity, Urban or Rural Area and Road Type over years.**

Các kiểu dữ liệu liên quan:

* Number of Accidents(Accident table)
* Accident Severity (Accident tablel)
* Urban\_or\_Rural\_Area (Accident table)
* Road\_Type (Accident table)
* Years (Accident table, column Date) - Dữ liệu chưa có, tự tạo thêm

**3.6 Report the number of calsulities by Severity, Casualty Type and Age Group over years, Age Group is defined as below:**

**Children: 0-15**

**Young adult: 0-17**

**Adult: 18-59**

**60 and over: 60-...**

Các kiểu dữ liệu liên quan:

* Number of Casualities (Casualty table)
* Casualty Severity (Casualty tablel)
* Casualty Type (Casualty table)
* Age Group (Casualty table) - Dữ liệu chưa có, tự tạo thêm

**3.7 Report the number of accidents by Journey Purpose and Vehicle Type.**

Các kiểu dữ liệu liên quan:

* Number of Accidents(Accident table)
* Journey\_Purpose\_of\_Driver (Vehicle tablel)
* Vehicle Type (Vehicle table)

**3.8 Create a new attribute Built-up Road in Accidients table. Built-up Road may have 2 values:**

**Built-up road: if Speed Limit below 50 mph**

**Non Built-up road: if Speed Limit above 50 mph**

Các kiểu dữ liệu liên quan:

* Built-up road (Accident table) - Dữ liệu chưa có, tự tạo thêm

**3.9 Report the number of accidents by Severity, Vehicle Type, Built-up Road over years.**

Các kiểu dữ liệu liên quan:

* Number of Accidents(Accident table)
* Accident Severity (Accident tablel)
* Vehicle Type (Vehicle table)
* Built-up road (Accident table)

**3.10 Students design other reports about UK car accident.**

Các kiểu dữ liệu liên quan:

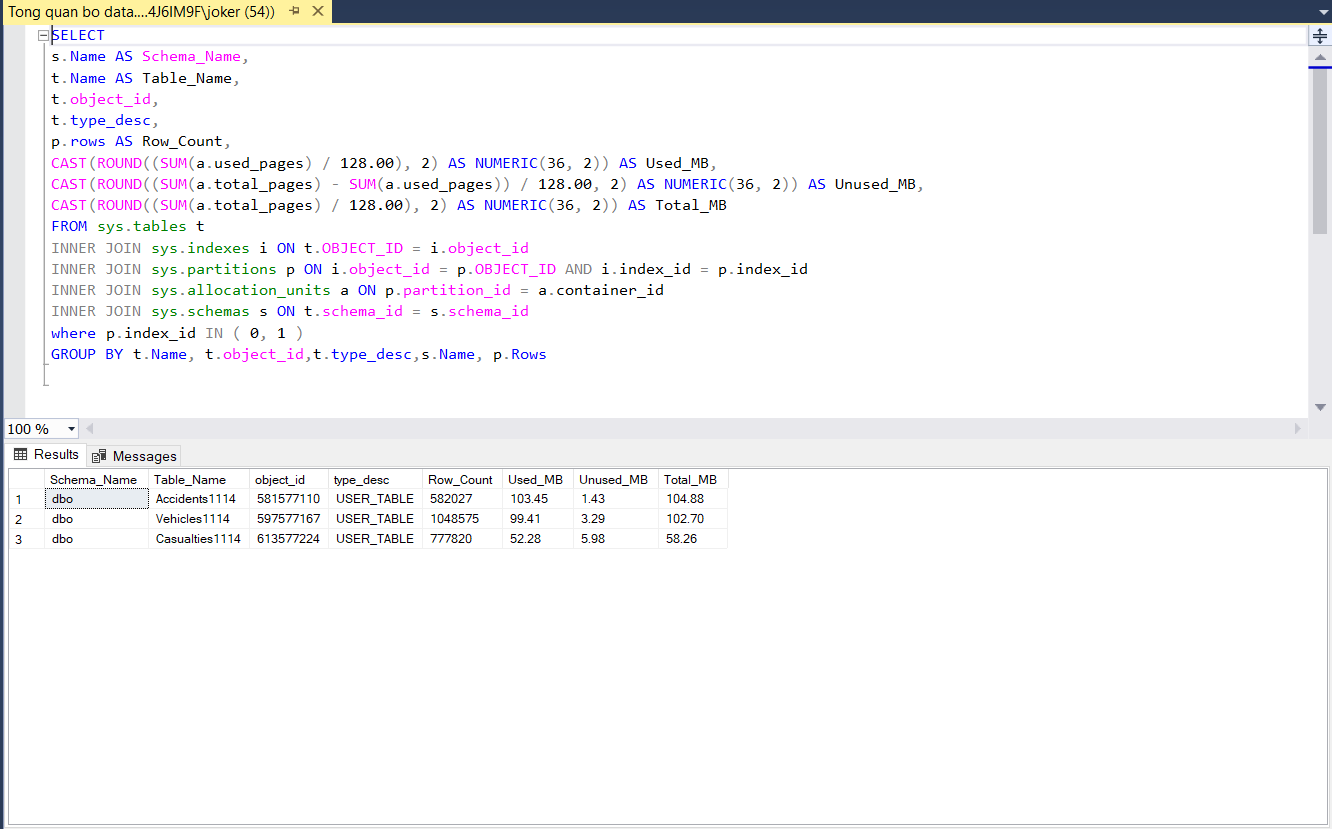
* Number of Accidents
* Accident Severity
* Vehicle Type
* Year, Quater, Month, Session
* Urban or Rural Area
* Sex Of Driver
* Road Type

**3.11 Use regional map to visually represent (by color) the number of car accidents in regions during a year.**

Các kiểu dữ liệu liên quan:

* Number of Accidents(Accident table)
* Region
* Country

1. **Phân tích profiling data**

****

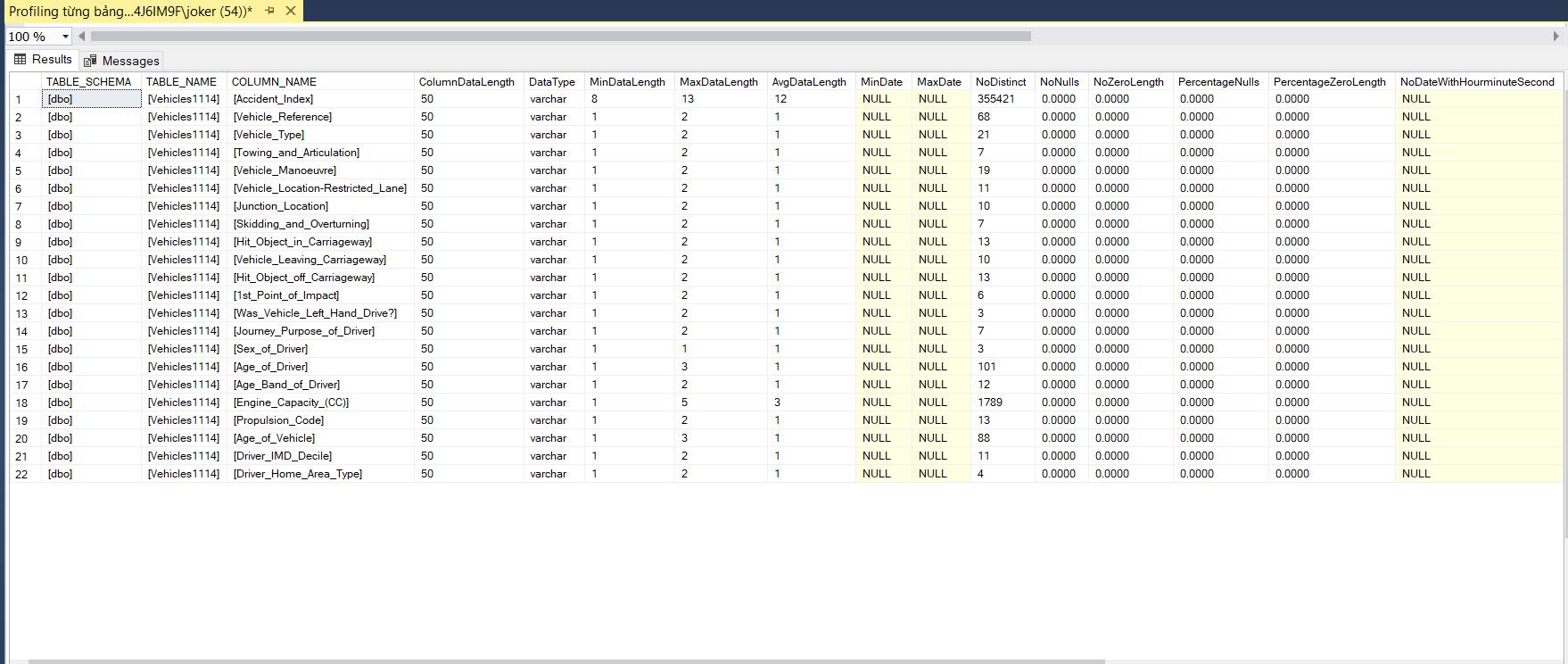
File Accidents1114 có 582027 dòng

File Vehicles1114 có 1048575 dòng

File Casualties1114 có 777820 dòng



Data Accidents1114 gồm 32 cột tất cả cột đều không bị missing value, ngoài ra còn có thông tin về min, max từng cột, độ dài cột, các giá trị duy nhất, …



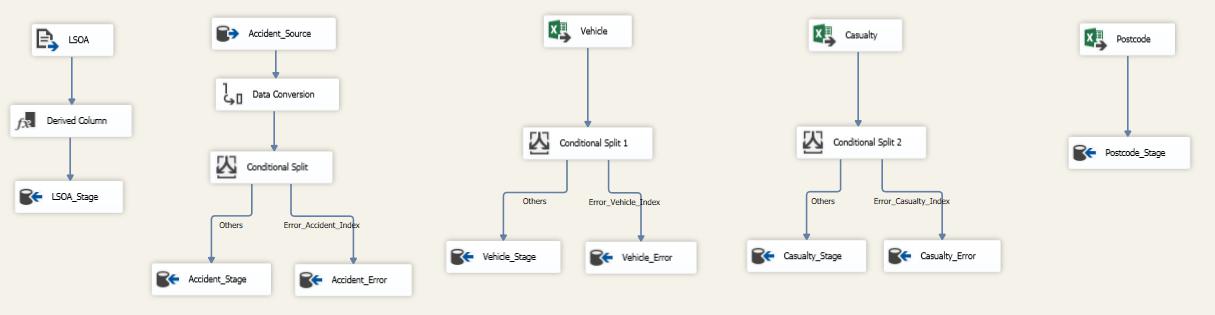
Data Vehicles1114 gồm 22 cột tất cả cột đều không bị missing value, ngoài ra còn có thông tin về min, max từng cột, độ dài cột, các giá trị duy nhất, …



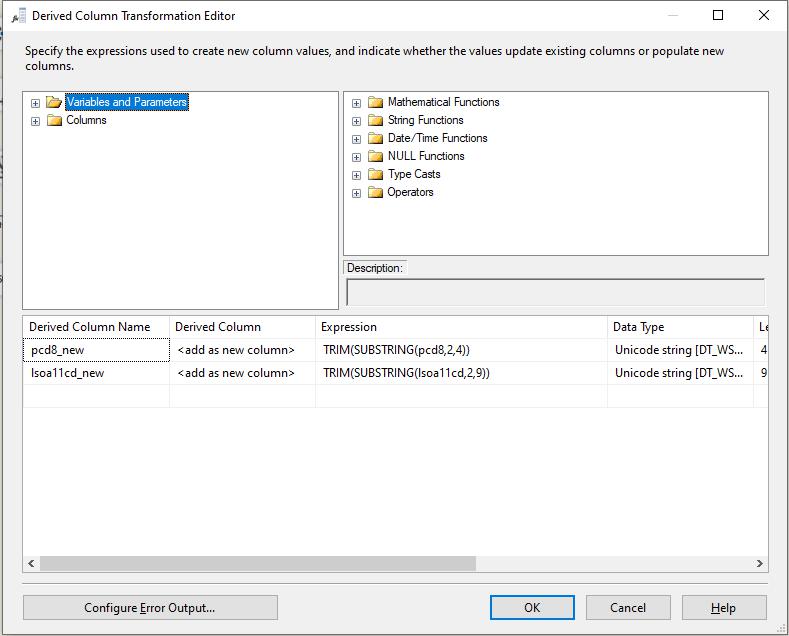
Data Vehicles1114 gồm 15 cột tất cả cột đều không bị missing value, ngoài ra còn có thông tin về min, max từng cột, độ dài cột, các giá trị duy nhất, …

# **Chương 2. Quá trình SSIS**

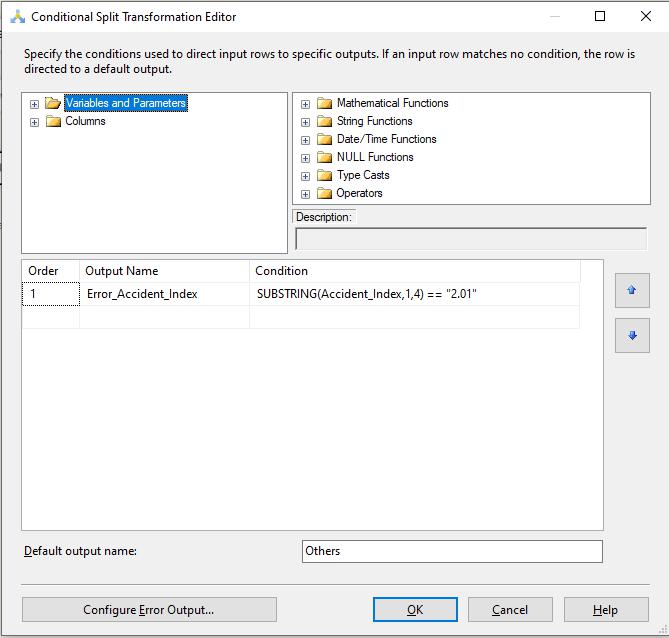
## Source vào Stage

****

* Ở bảng LSOA khi đổ từ Source vào Stage chúng em dùng derive column để xử lý 2 cột lsoa11cd và pcd8 vì khi đổ LSOA bằng flat file data xuất hiện ký tự “ ” ở 2 đầu dữ liệu. Vì vậy đã chúng em dùng Derived Column để làm sạch dữ liệu
* Cột pcd8: Dùng substring từ vị trí thứ 2 để tránh dấu “ và lấy đi 4 ký tự đầu của chuỗi để so khớp với data từ bảng Postcode và dùng trim để xóa khoảng cách do các chuỗi postcode có độ dài không cố định từ 2-4 ký tự.
* Cột lsoa11cd: Tương tự xử lý 9 ký tự của LSOA.

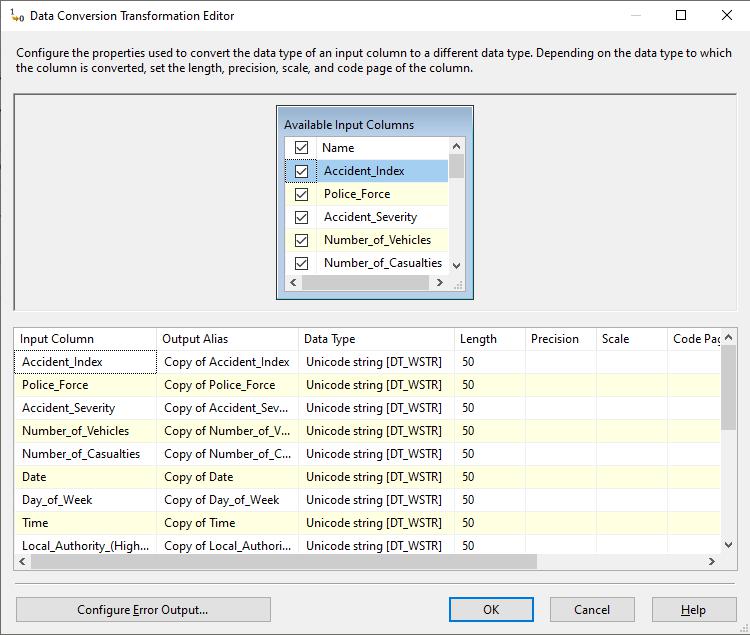


* Ở 3 bảng tiếp theo gồm Accident, Casualty và Vehicle thì chúng em dùng conditional split để loại bỏ các cột có thuộc tính Accident\_Index bị lỗi định dạng từ source và đưa vào database riêng có tên Error. Dấu hiệu nhận biết là 4 ký tự đầu có dạng “2.01”

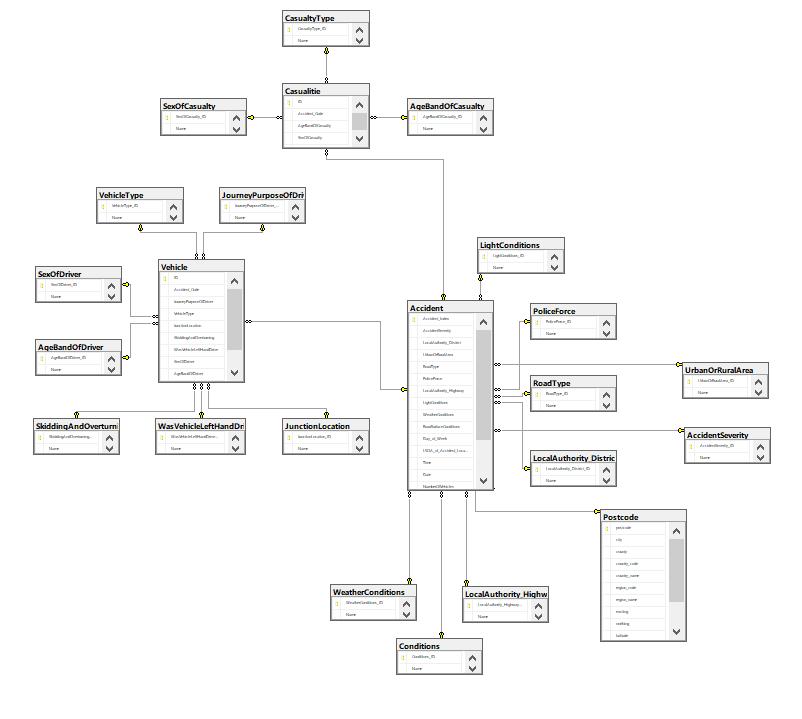


Quá trình conditional split

* Đặc biệt riêng bảng Accident chúng em không dùng file excel để làm source đổ dữ liệu do data bị lỗi cột Date. Vì vậy team đã chọn cách import file excel vào thẳng sql server và dùng data từ đó để đổ vào Stage.



1. **Stage vào NDS**



Thiết kế diagram dành cho NDS

****

Control flow của NDS

* 2 Data flow đầu tiên là để đổ data từ 2 file postcode và UK Accident dùng để tham chiếu khóa ngoại khi đổ các bảng chính phía sau.
* Accident\_NDS:
* Kiểm tra hợp lệ các khóa ngoại của bảng Accident
* Gộp bảng Accident và LSOA thông qua khóa ngoại LSOA để lấy ra cột pcd8 trong bảng LSOA
* Tiếp tục dùng cột pcd8 để kiểm tra khóa ngoại postcode trong bảng Postcode. Sau đó chia dữ liệu về 2 bảng gồm những dòng có pcd8 hợp lệ và không hợp lệ
* Với những dòng pcd8 không tồn tại thì sẽ chuyển về null và gộp lại với những dòng hợp lệ
* Sau khi thỏa hết điều kiện thì sẽ được đổ vào NDS
* Casualty\_NDS và Vehicle\_NDS:
* Chỉ đơn giản kiểm tra hợp lệ của các khóa ngoại trước khi đổ vào NDS

1. **NDS vào DDS**

**3.1. Lượt đồ sao DDS**

* **Nhu cầu phân tích**

1 Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng (Fatal, Serious, Slight) ở các Địa phương (Local\_Authority\_(District)) trong tất cả các năm. Bảng

2 Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa Phương (Local\_Authority\_(District)) theo các Quý trong từng năm.

3 Thống kê số lượng người tử vong theo Giới Tính, Loại Nạn Nhân (Casualty Type) và Nhóm Tuổi (Age\_Band\_of\_Casualty) theo các năm.

4 Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng và Thời Điểm Trong Ngày (Morning: 5am-12pm, Afternoon: 12pm-5pm, Evening: 5pm-9pm, Night: 9pm-5am) trong các năm.

5 Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng, Vùng (Urban\_or\_Rural\_Area), và Kiểu Đường (Road Type) trong các năm.

6 Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng, Loại Nạn Nhân (Casualty Type) và Độ Tuổi trong các năm, Độ Tuổi được định nghĩa như sau:

Children: 0-15

Young adult: 0-17

Adult: 18-59

60 and over: 60-...

7 Tổng hợp số lượng tai nạn theo Mục Đích Hành Trình (Journey Purpose) và Loại Phương Tiện (Vehicle\_Type).

8 Tạo thêm thuộc tính Built-up Road trong table Accidients. Built-up Road có 2

giá trị: Built-up road: Nếu tốc độ giới hạn (Speed Limit) dưới 50 mph Non Built-up road: Nếu tốc độ giới hạn từ 50 mph

9 Thống kê số lượng tai nạn theo Mức Độ Nghiêm Trọng, Loại Phương Tiện (Vehicle Type), Built-up Road trong các năm.

10 Thống kê số lượng tai nạn theo Mức Độ Nghiêm Trọng, Loại Phương Tiện (Vehicle Type), Built-up Road trong các năm.

11 Định nghĩa fact Variance để tính mức độ tăng giảm của TNGT theo đơn vị phần trăm qua các năm.

12 Xây dựng đồ thị/ biểu đồ cho các bảng thống kê ở trên.

13 Dùng regional map để biểu diễn trực quan (bằng màu sắc) số lượng TNGT ở các vùng trong năm.

* **Mô hình hóa thiết kế chiều**

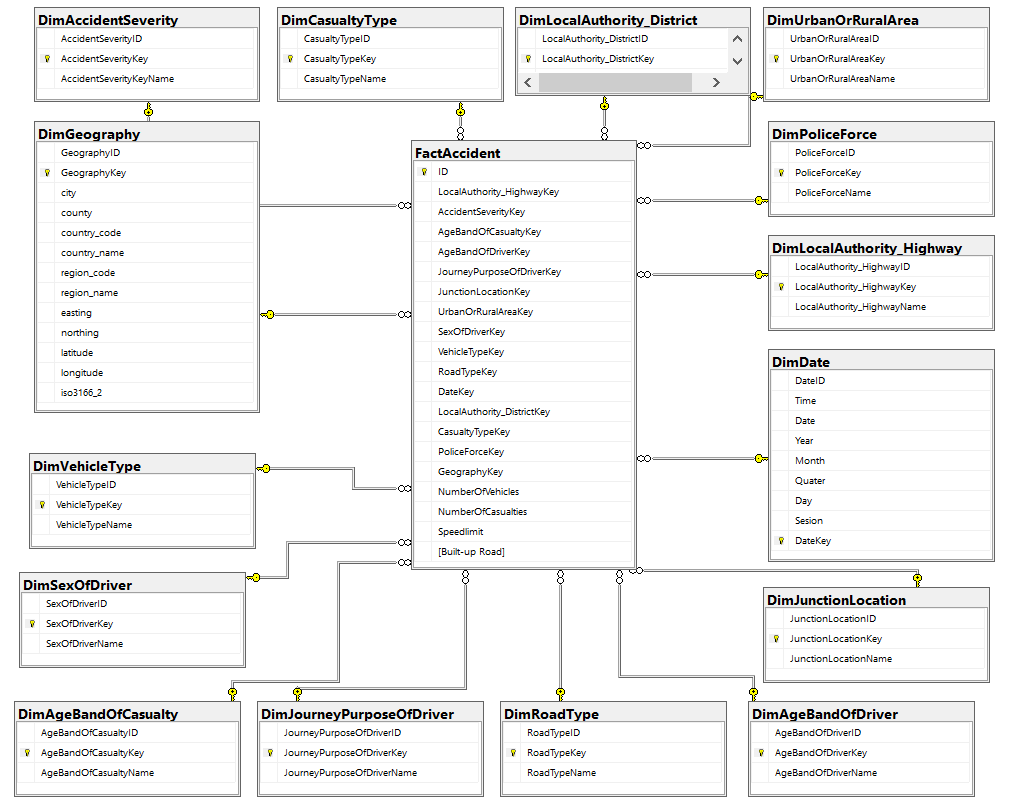
Các chiều liên quan sự kiện phân tích

* DimAccidentServity
* DimCasualtyType
* DimAgeBandOfCasualty
* DimAgeBandOfDriver
* DimDate
* DimGeography
* DimJourneyPurposeOfDriver
* DimJunctionLocation
* DimLocalAuthority\_District
* DimLocalAuthority\_Highway
* DimPoliceForce
* DimRoadType
* DimSexOfDriver
* DimUrbanOrRuralArea
* DimVehicleType
* FactAccident
* **Mô hình hóa thiết kế Fact**

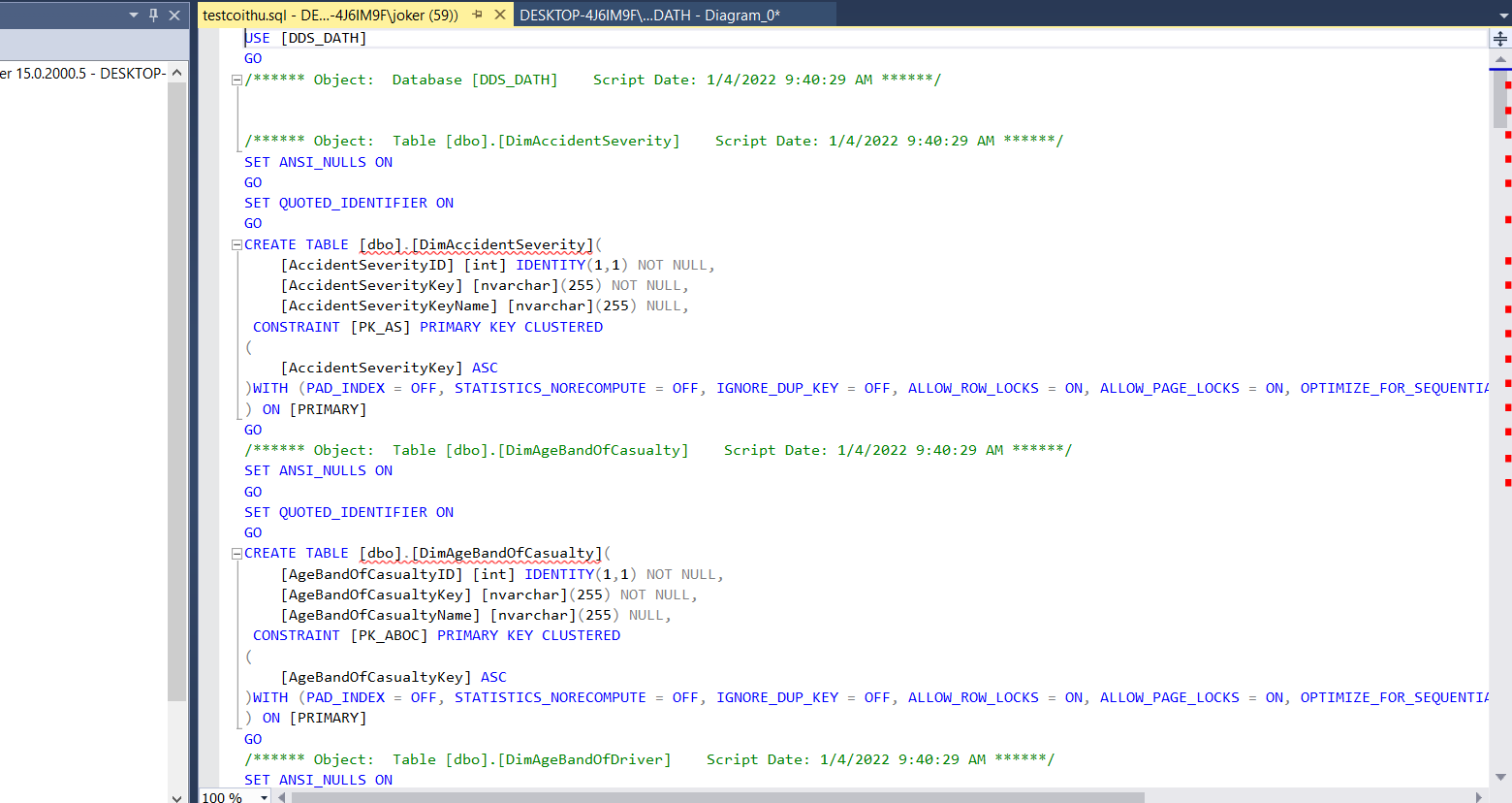
Các giá trị có sẵn: NumberOfVehicles, NumberOfCasualties, Speedlimit

Các giá trị tính toán thêm: [Built-up Road]

* **Lượt đồ sao**

****

**3.2. Tạo DataBase và đổ bảng vào**

****

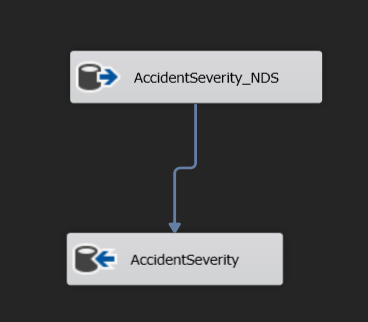
**3.3. Tạo bảng Dim**

* **DimAccidentSeverity**

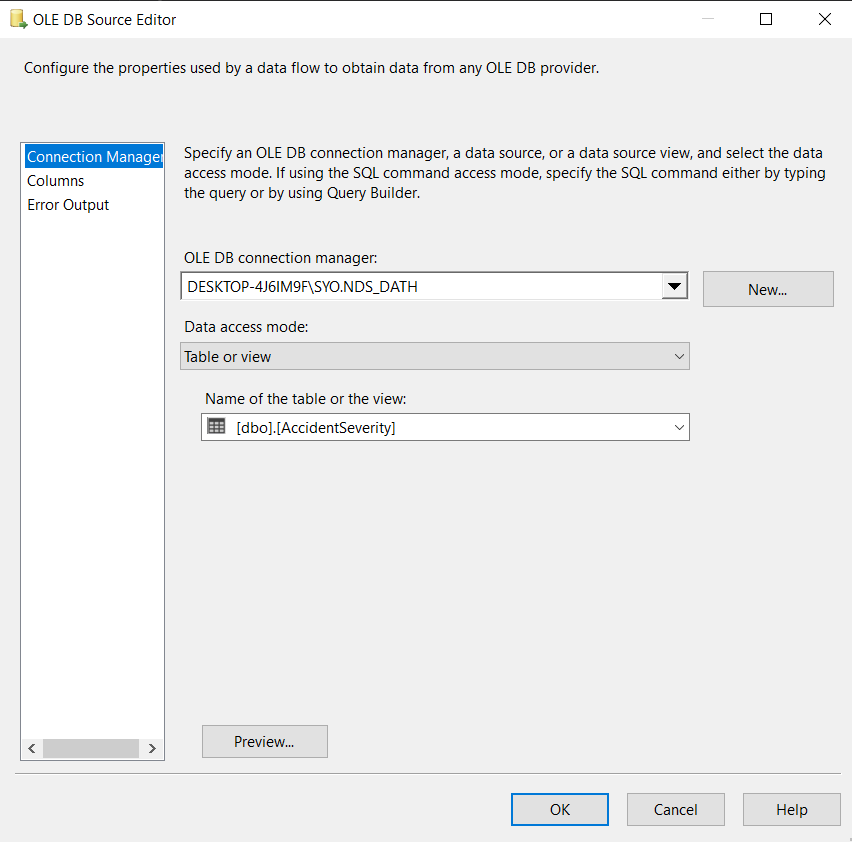
Kéo thả Data Flow Task. Sau đó đổi tên Data Flow thành AccidentSeverity.



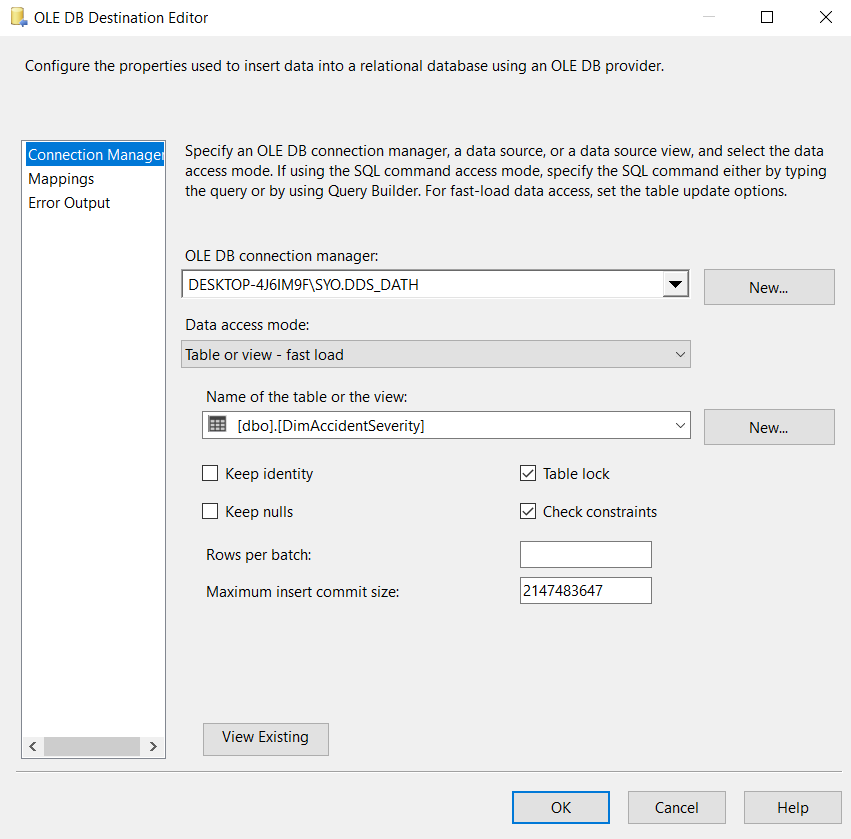
Chọn Flat File Source, OLE DB Destination. Đổi tên Flat File Source thành AccidentSeverity\_NDS. ­ Đổi tên OLE DB Destination thành AccidentSeverity.

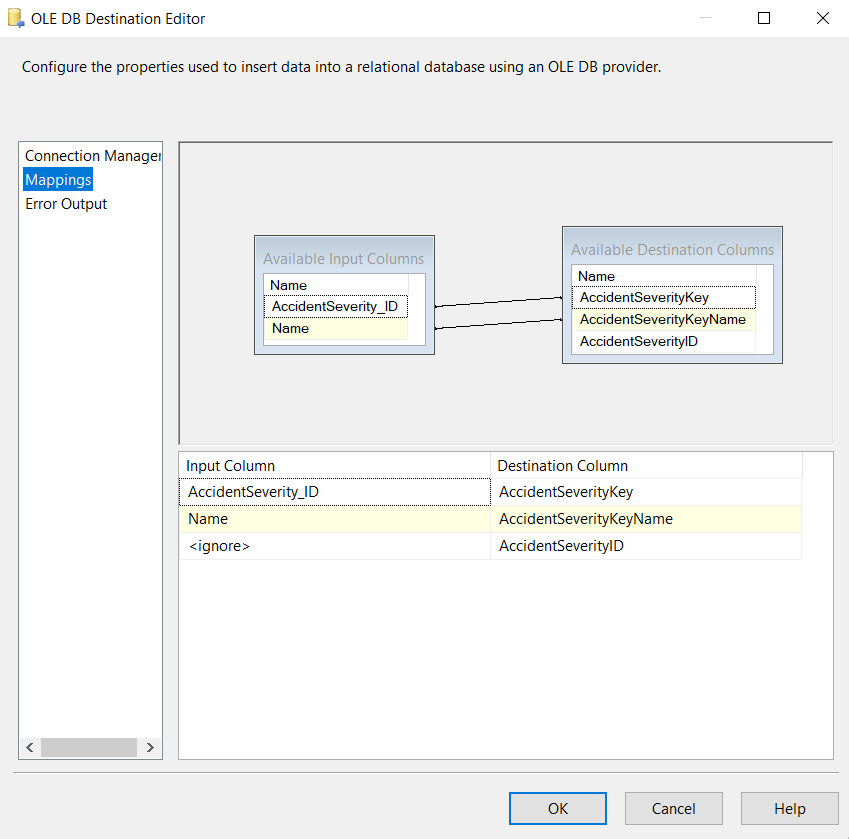


Nhấp chuột vào AccidentSeverity\_NDS chọn DB Server và Table sau đó nhấn OK.



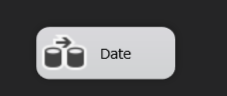
Nhấp chuột vào AccidentSeverity chọn DB Server và Table



Sang mục Mapping tiến hành mapping nguồn dữ liệu NDS với bảng Dim sau đó chọn OK.  


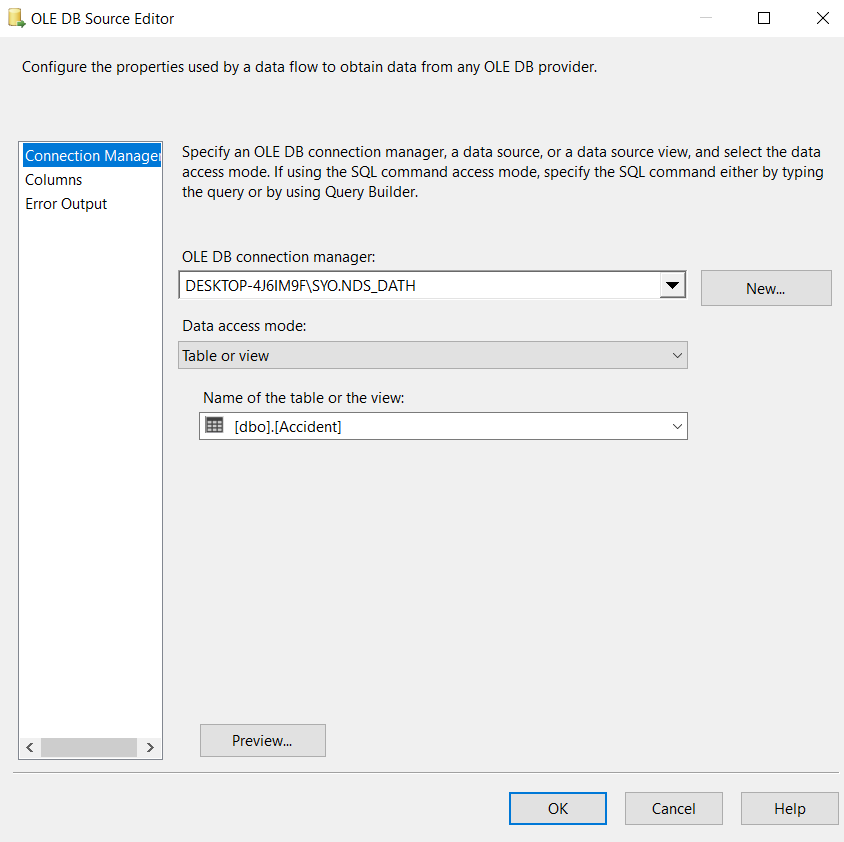
* Tương tự cho các bảng còn lại: DimAgeBandOfCasualty, DimAgeBandOfDriver, DimCasualtyType, DimGeography, DimJourneyPurposeOfDriver, DimJunctionLocation, DimLocalAuthority\_District, DimLocalAuthority\_Highway, DimPoliceForce, DimRoadType, DimSexOfDriver, DimUrbanOrRuralArea, DimVehicleType.
* **DimDate**

Kéo thả Data Flow Task. Sau đó đổi tên Data Flow thành Date.

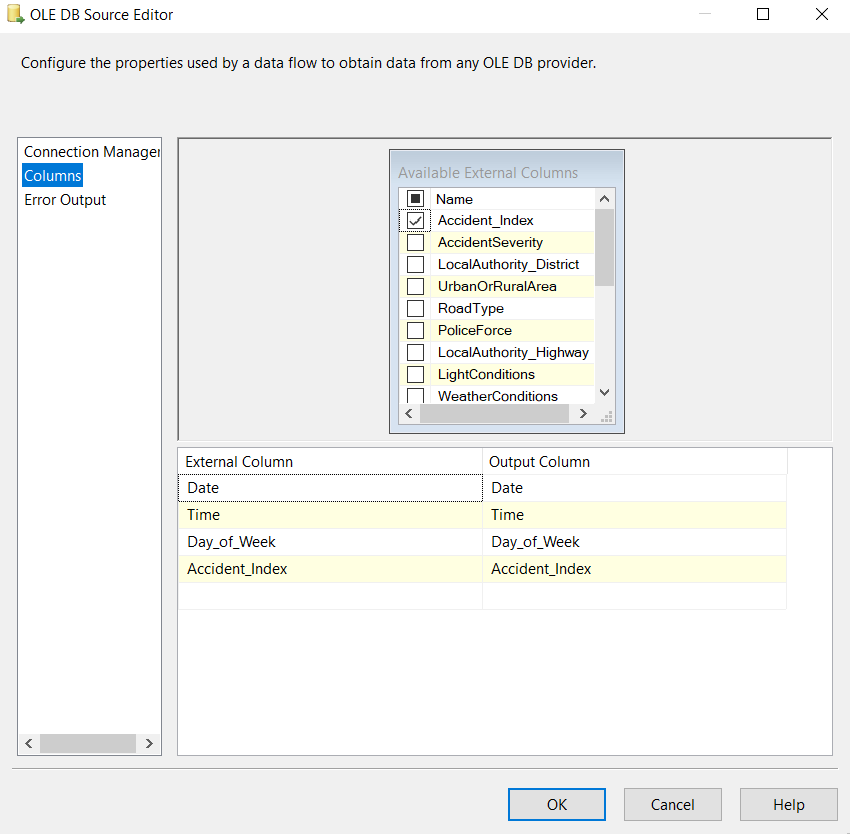


Chọn Flat File Source,Data Conversion,Derived Column, Sort, OLE DB Destination. Đổi tên Flat File Source thành Accident\_NDS. ­ Đổi tên OLE DB Destination thành Date.

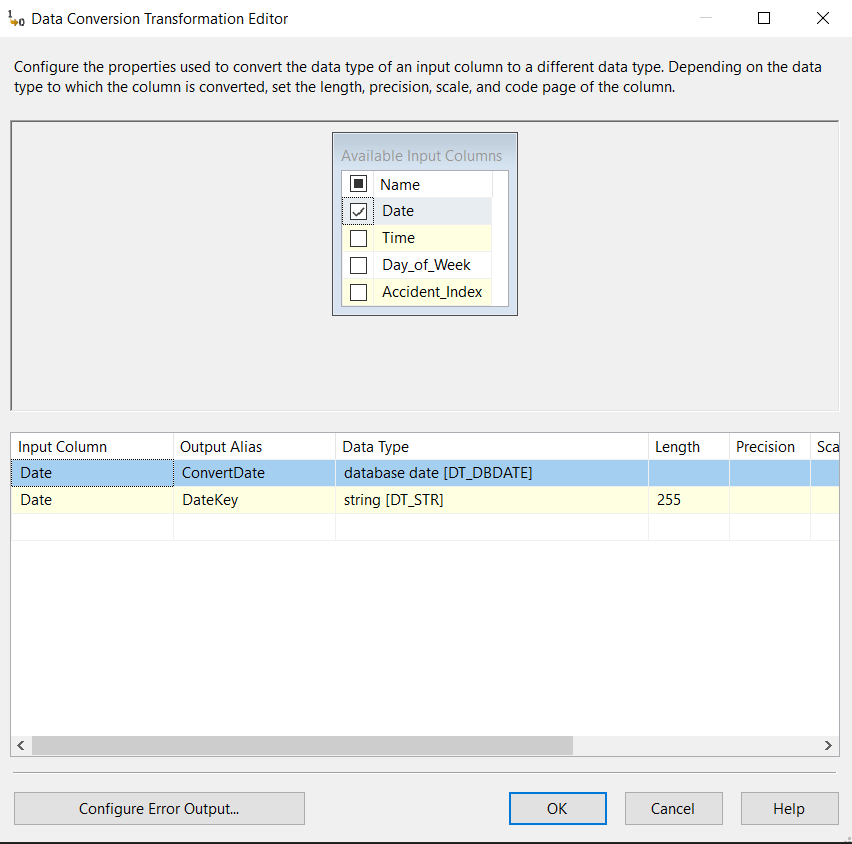
Nhấp chuột vào Accident\_NDS chọn DB Server và Table



Tại tab Column chọn các cột cần sử dụng sau đó nhấn OK.

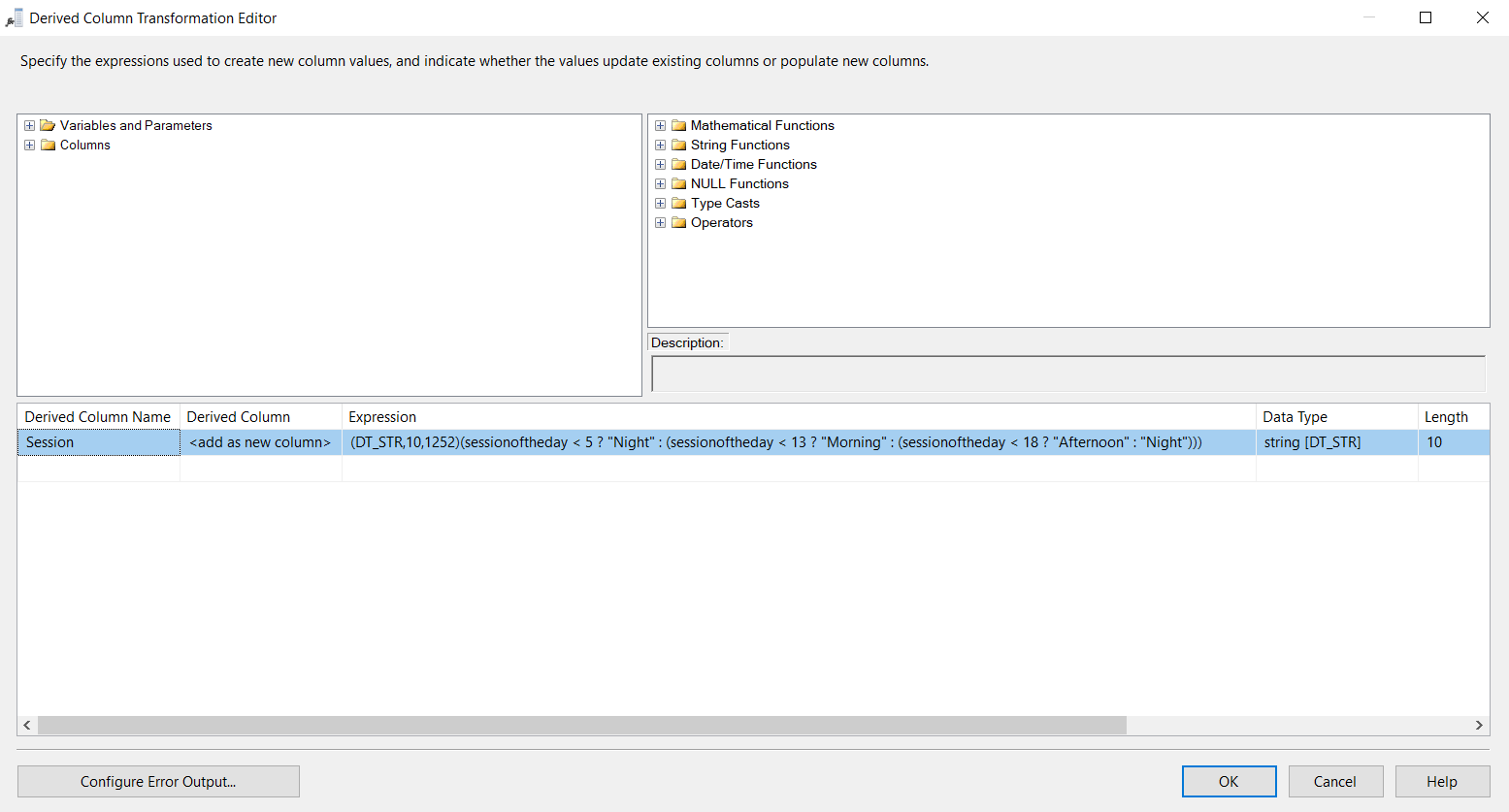


Mở Data Conversion để tiến hành biến đổi lại dữ liệu, chuyển đổi Date từ dạng nvarchar sang kiểu date và lưu là ConvertDate, chuyển Date từ dạng nvarchar sang dạng varchar và lưu là DateKey.

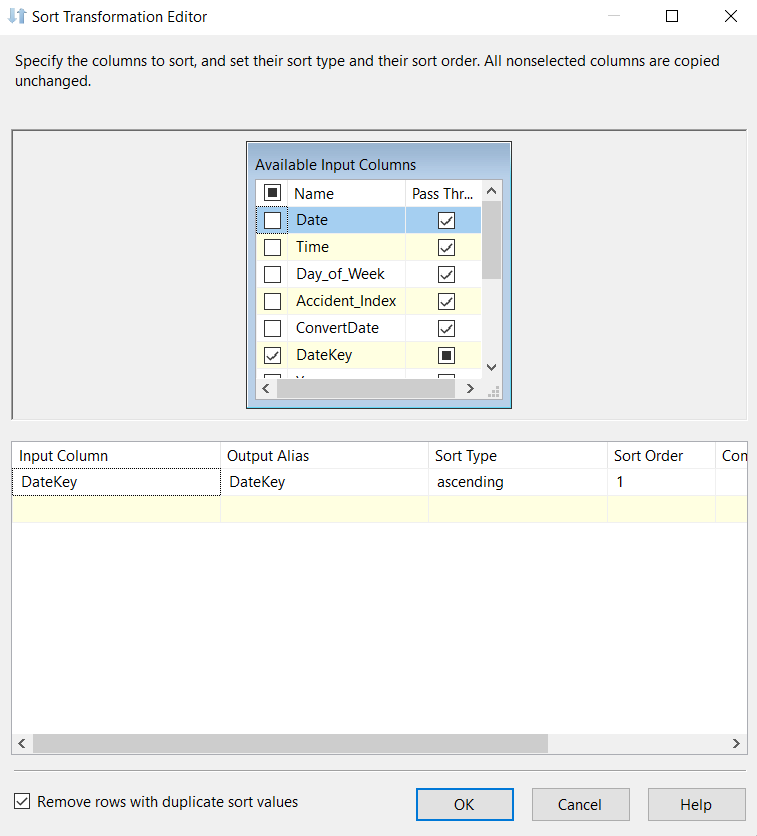


Mở Derived Column để tiến hành tính toán một số giá trị, lấy ra Year, Quater, Month, Day từ ConvertDate đã chuyển ở bước Data Conversion bằng các hàm Date/Time Functions. Lấy ra Session dùng để lưu các buổi trong một ngày.

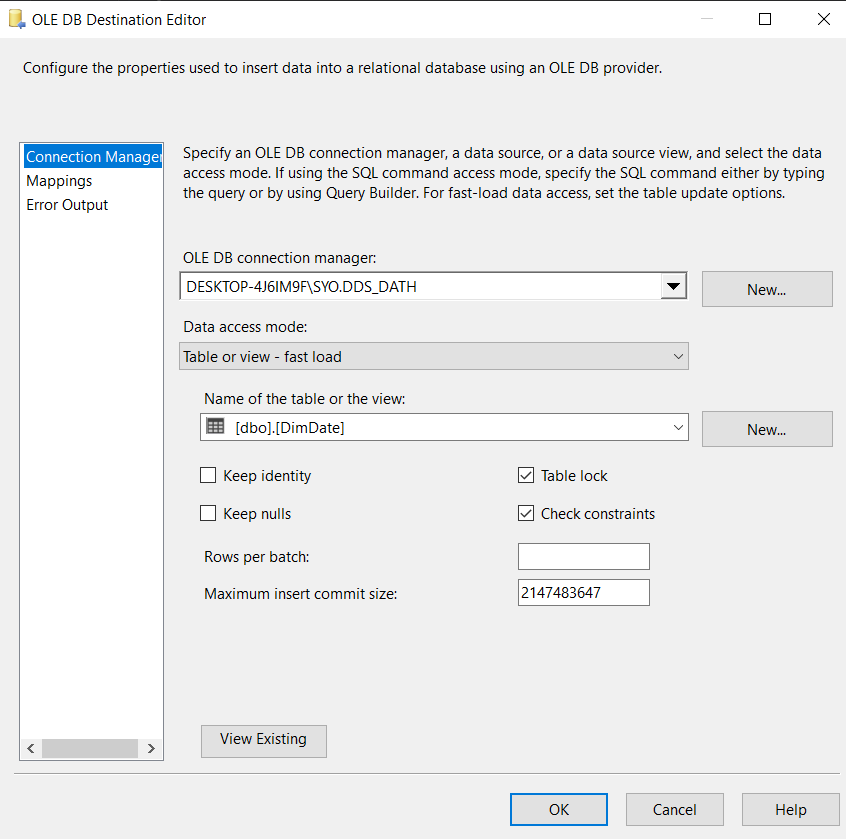




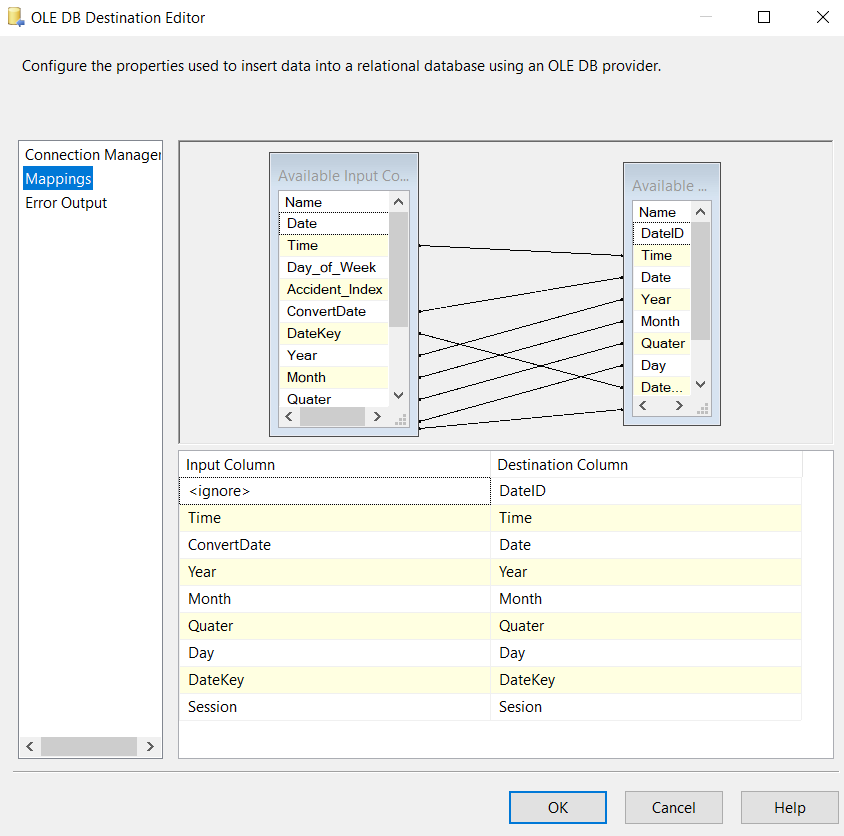
Sau khi biến đổi tính toán thì tiến hành Sort để nhằm xóa các giá trị bị trùng lặp.



Nhấp chuột vào OLE DB Destination để đưa dữ liệu vào DimDate, chọn DB và Table

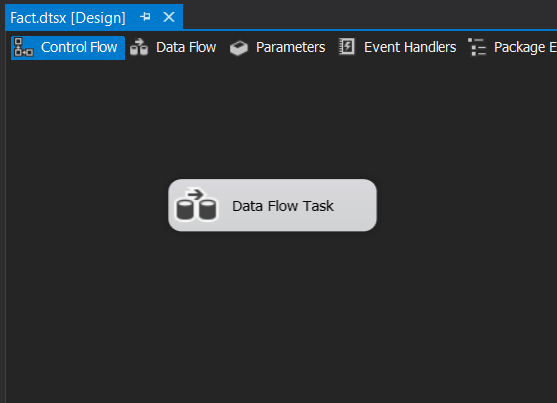


Sang mục Mapping tiến hành mapping nguồn dữ liệu NDS với bảng Dim sau đó chọn OK.

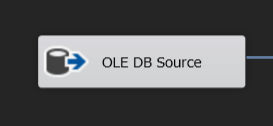


* **FactAccident**

Kéo thả Data Flow Task để tiến hành xây Data Flow cho bảng Fact.



Kéo task OLE DB Source để tiến hành nạp source từ NDS vào



Chọn OLE DB Source sau đó tiến hành kết nối DB và viết truy vấn để lấy các dữ liệu cần thiết từ các bảng của NDS

…. Code

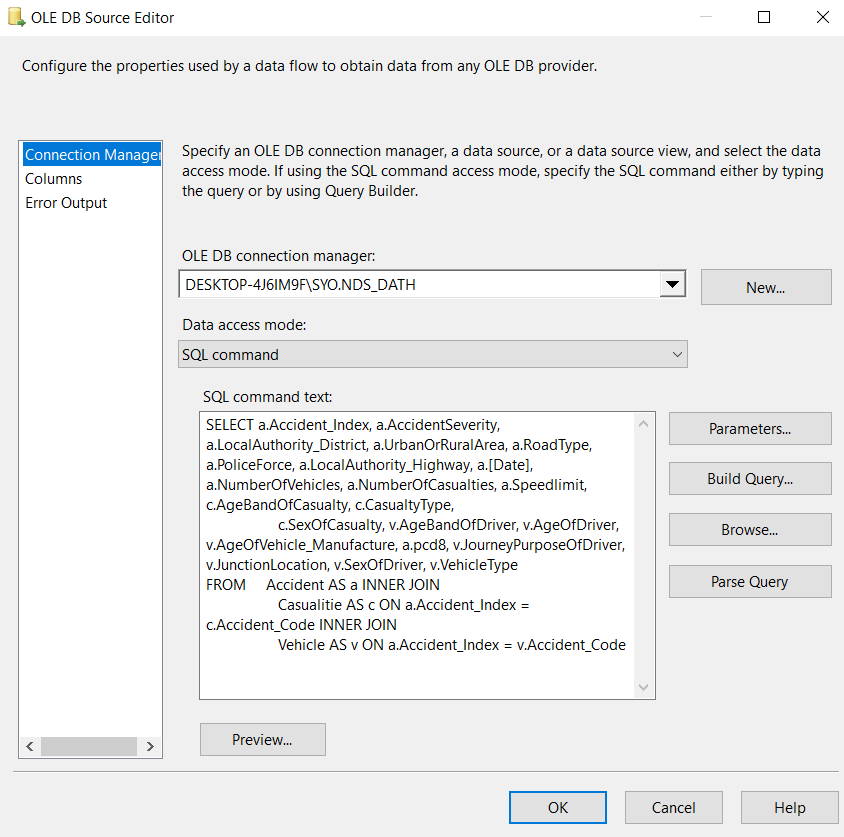
SELECT a.Accident\_Index, a.AccidentSeverity, a.LocalAuthority\_District, a.UrbanOrRuralArea, a.RoadType, a.PoliceForce, a.LocalAuthority\_Highway, a.[Date], a.NumberOfVehicles, a.NumberOfCasualties, a.Speedlimit, c.AgeBandOfCasualty, c.CasualtyType,

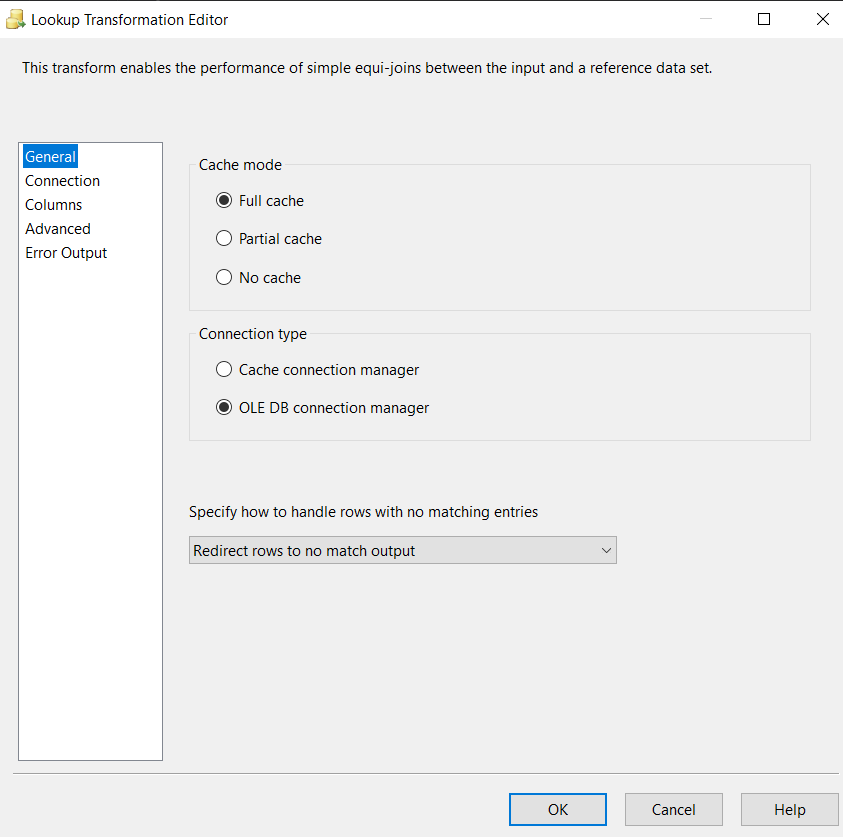
c.SexOfCasualty, v.AgeBandOfDriver, v.AgeOfDriver, v.AgeOfVehicle\_Manufacture, a.pcd8, v.JourneyPurposeOfDriver, v.JunctionLocation, v.SexOfDriver, v.VehicleType

FROM Accident AS a INNER JOIN

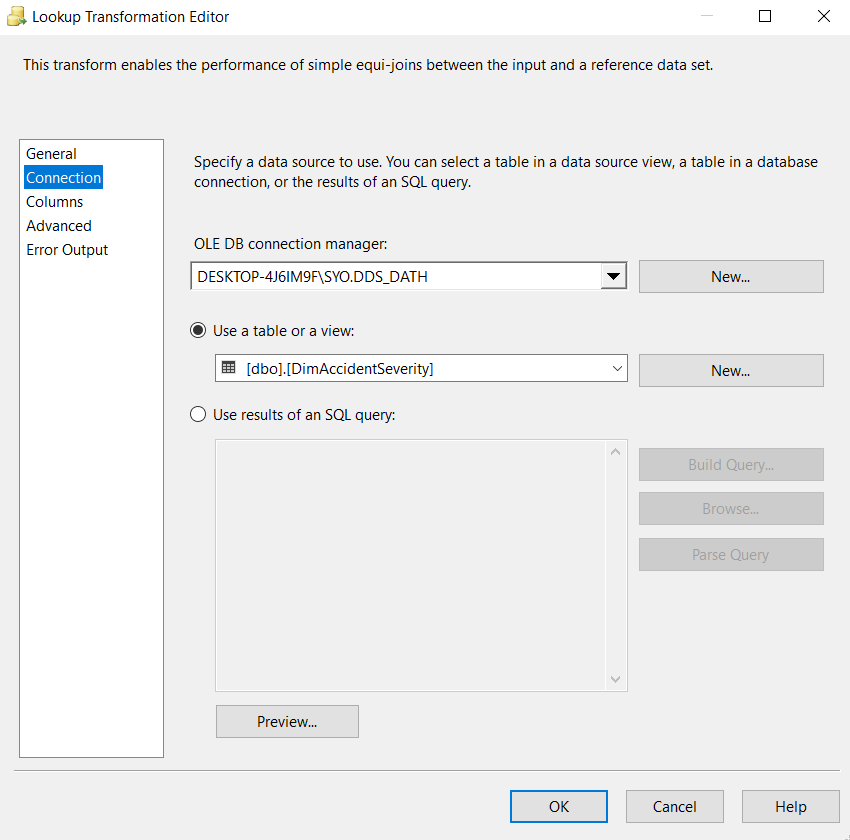
Casualitie AS c ON a.Accident\_Index = c.Accident\_Code INNER JOIN

Vehicle AS v ON a.Accident\_Index = v.Accident\_Code

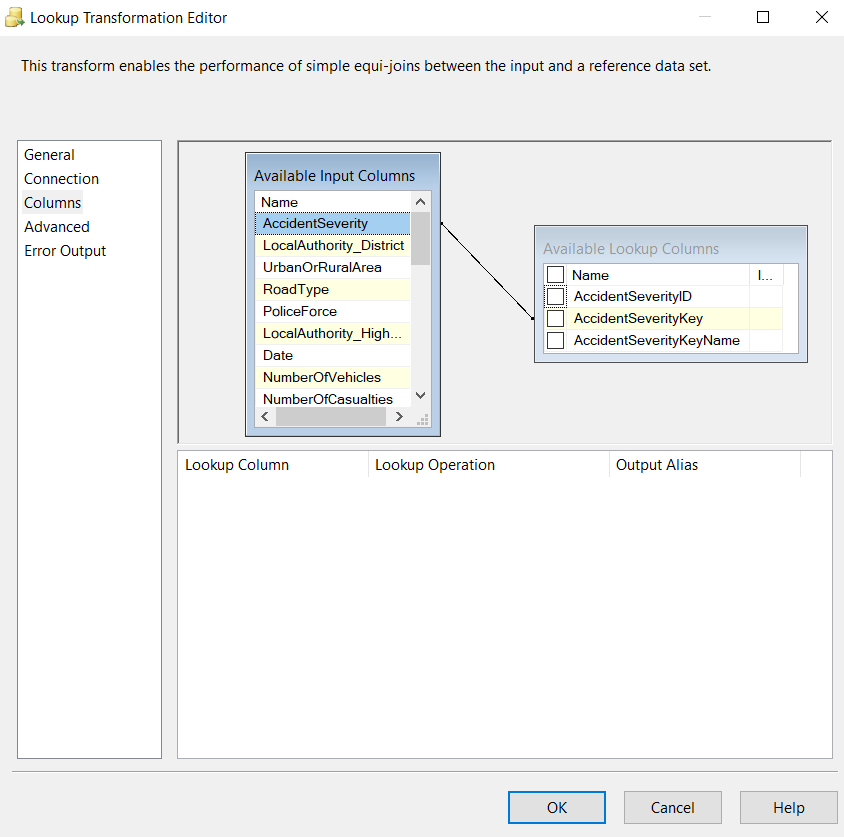


Sau khi đưa dữ liệu từ NDS vào tiến hành Lookup cho bảng DimAccidentServity

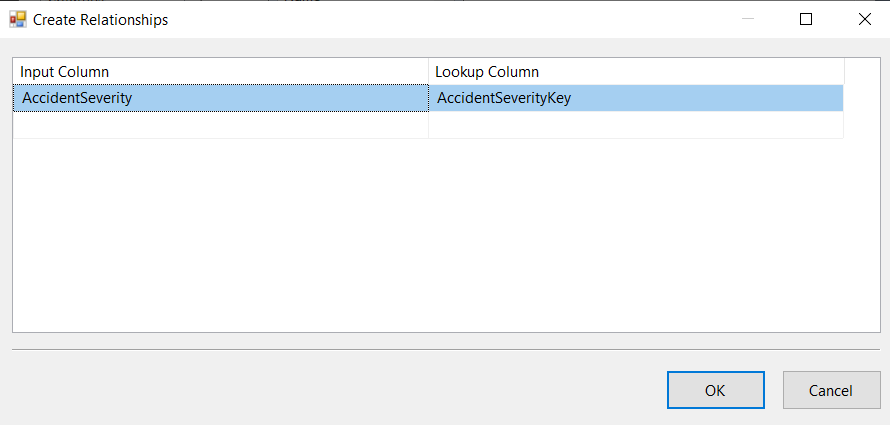
Connect đến bảng



Tại mục Column chuột phải vào khoảng trống các Table chọn Edit Mapping



Tiến hành mapping



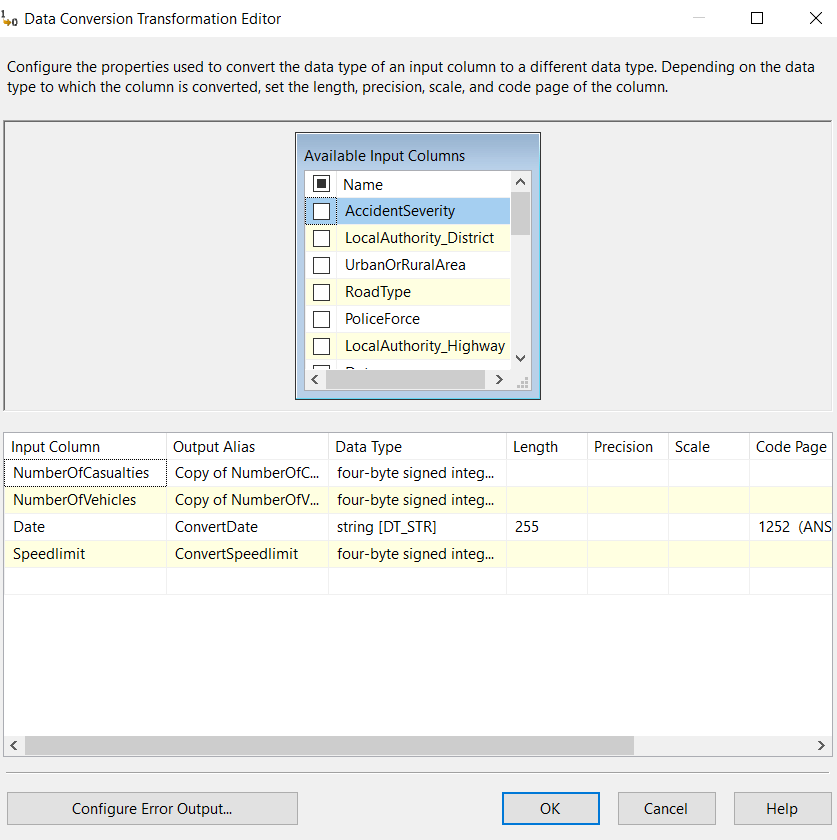
Tương tự cho các bảng còn lại

Sau khi Lookup các bảng tiến hành Conversion lại dữ liệu.

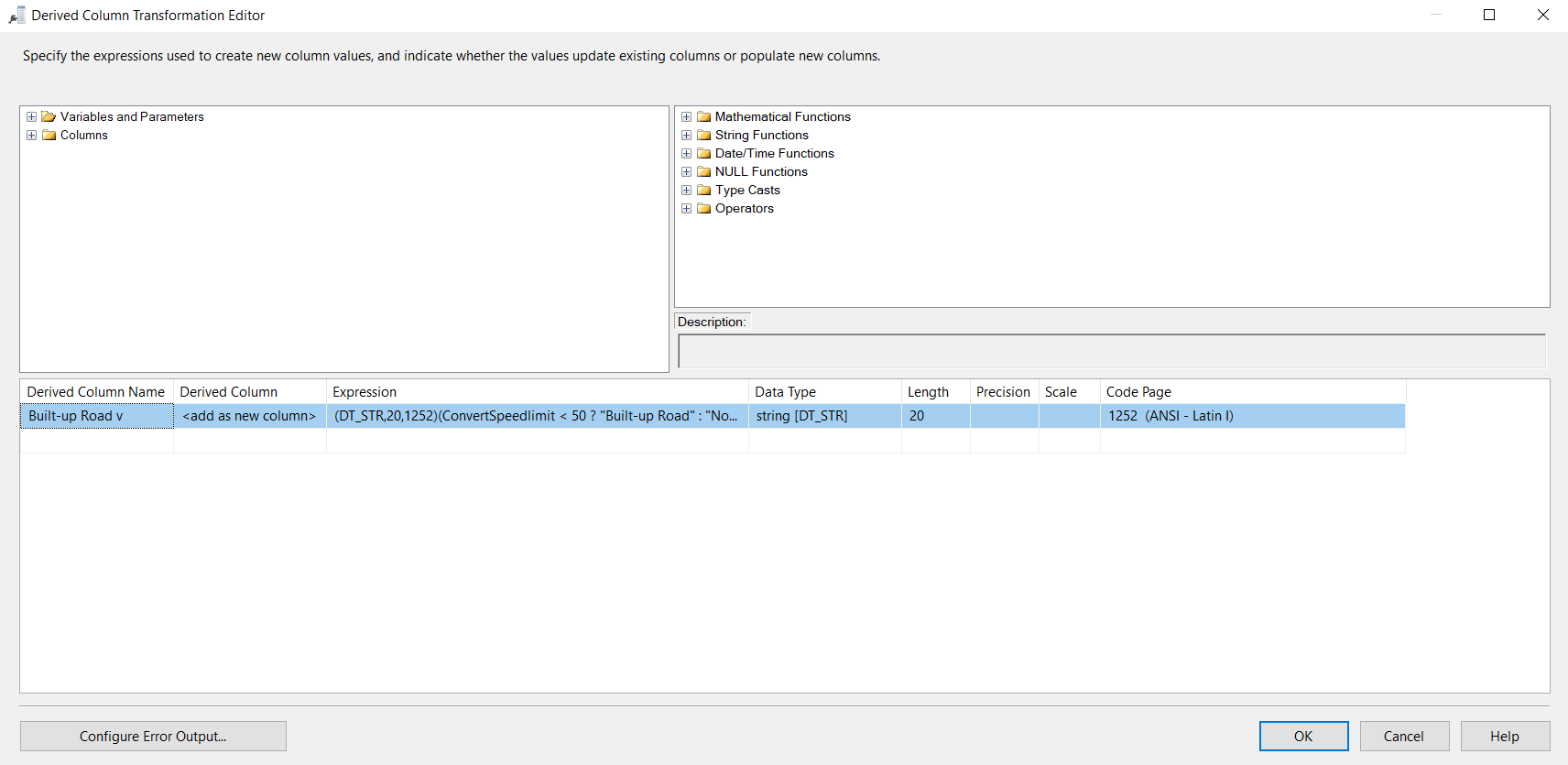
Biến đổi Number Of Vehicle và Number Of Casualty từ dạng nvarchar sang dạng int.

Biến đổi Date từ dạng nvarchar sang dạng varchar.

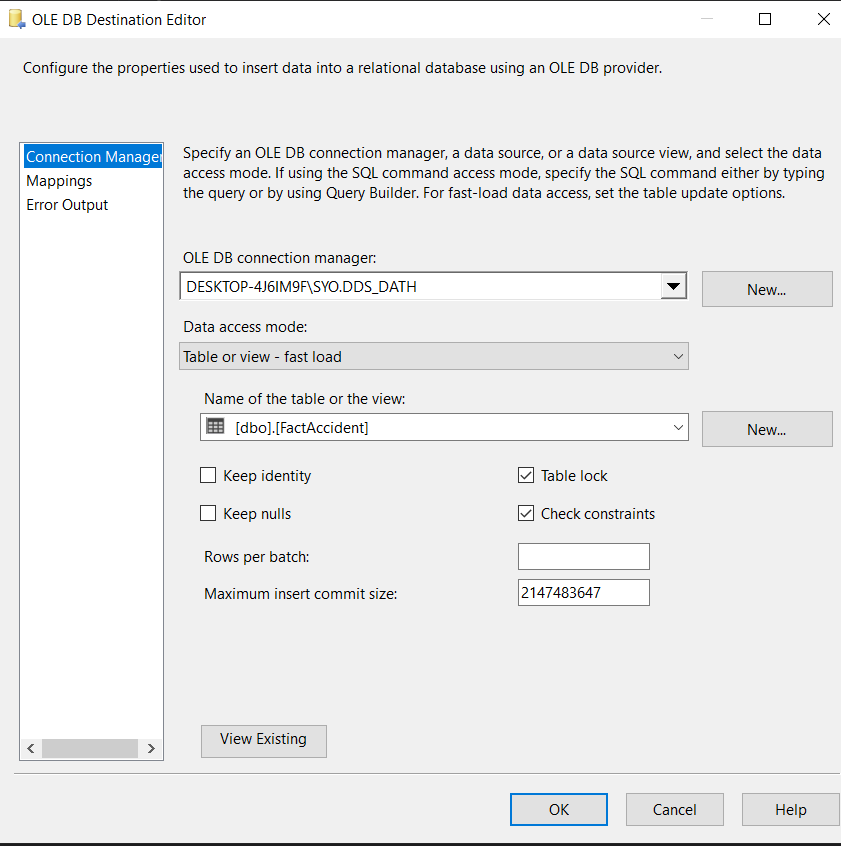
Biến đổi Speedlimit từ dạng nvarchar sang dạng int.



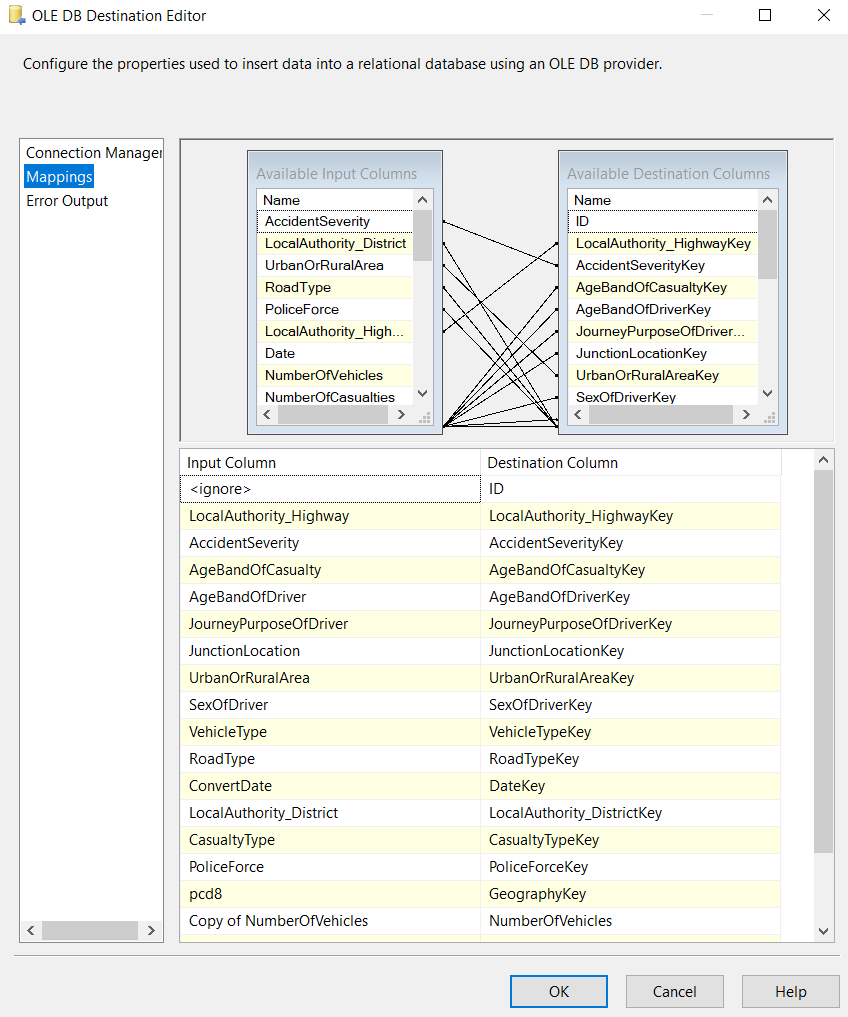
Sử dụng Derived Column để chuyển thuộc tính Build Up Road có giá trị Speedlimit < 50 và > = 50



Cuối cũng đổ dữ liệu vào Bảng Fact. Tạo một OLE DB Destination. Chọn DB và Table



Mapping lại dữ liệu



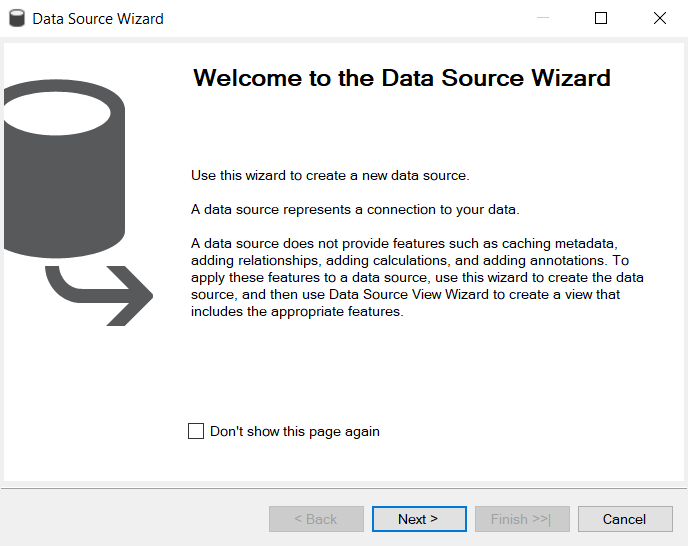
# **Chương 3: OLAP CUBE và MDX**

## Tạo ứng dụng SSAS

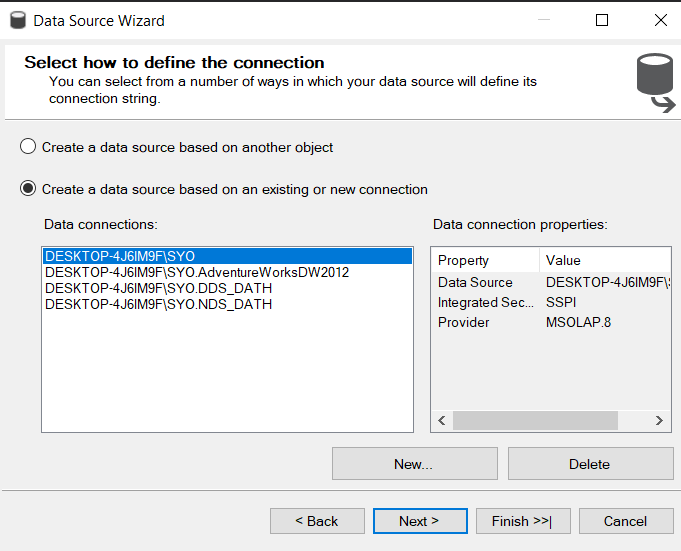
****

1. **Tạo Data Source**

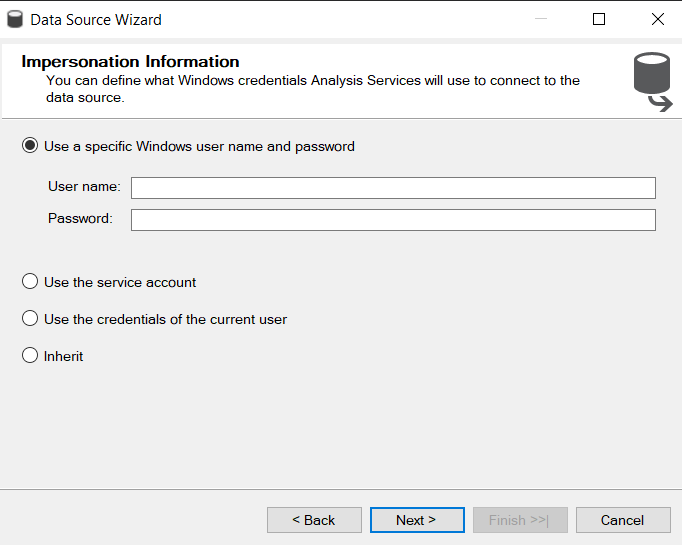
Chuột phải vào DataSource và chọn New Data Source.Chọn Next để tiếp tục.



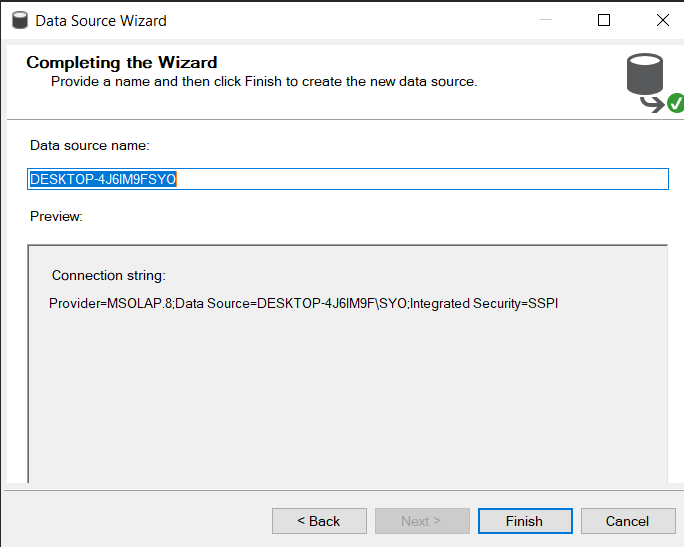
Connect đến Data có sẵn hoặc tạo connect mới. Connect đến DDS đã tạo trước đó.



Nhập User Name và pass của Window

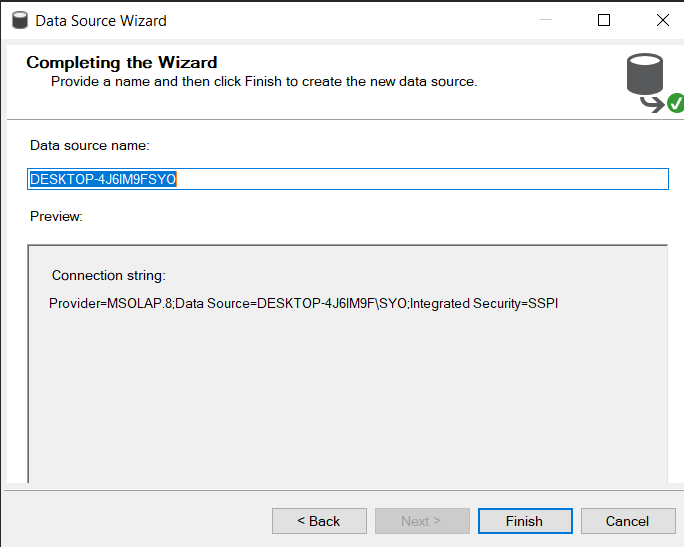


Sau đó chọn Finish.

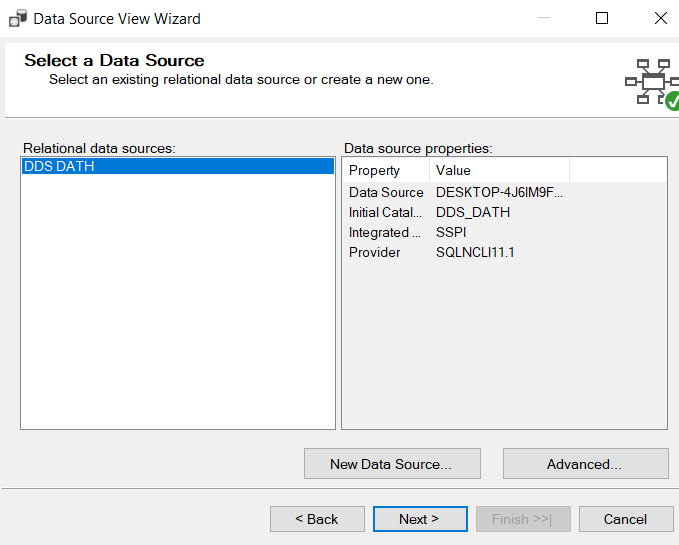


## Tạo Data Source View

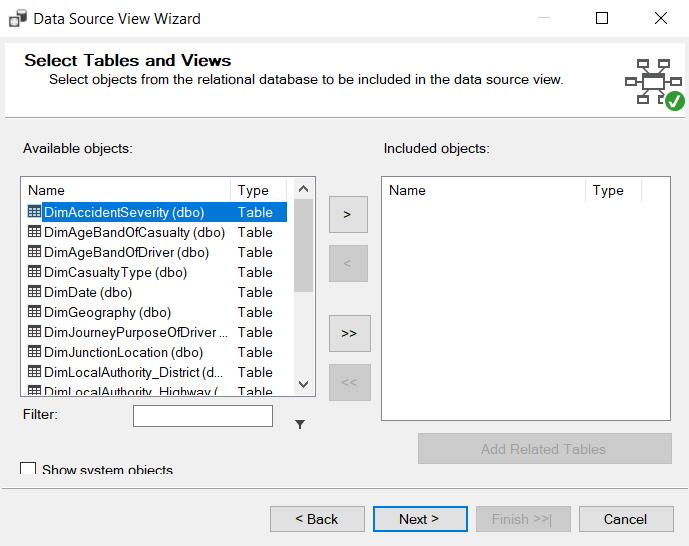
Chọn chuột phải vào Data Source Views và New Data Source View để tiến hành tạo data mới. Sau đó chọn Next.



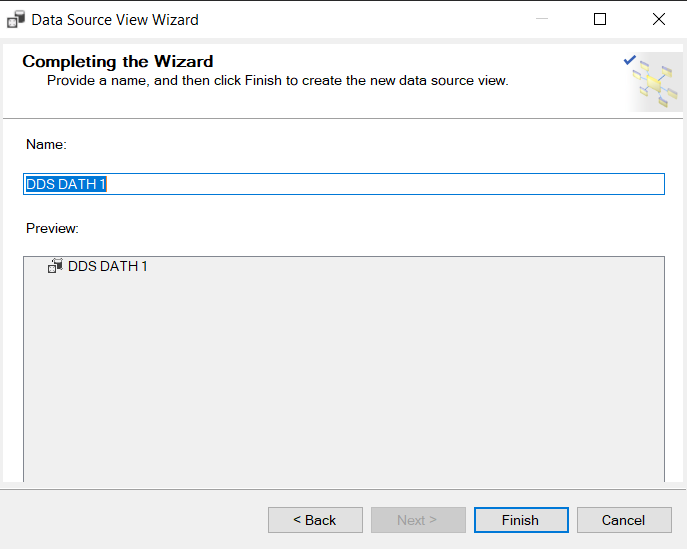
Chọn data source đã tạo từ bước Data Source trước đó hoặc new một data mới



Chọn ra các bảng cần sử dụng và next ( Lấy hết các bảng từ DDS)

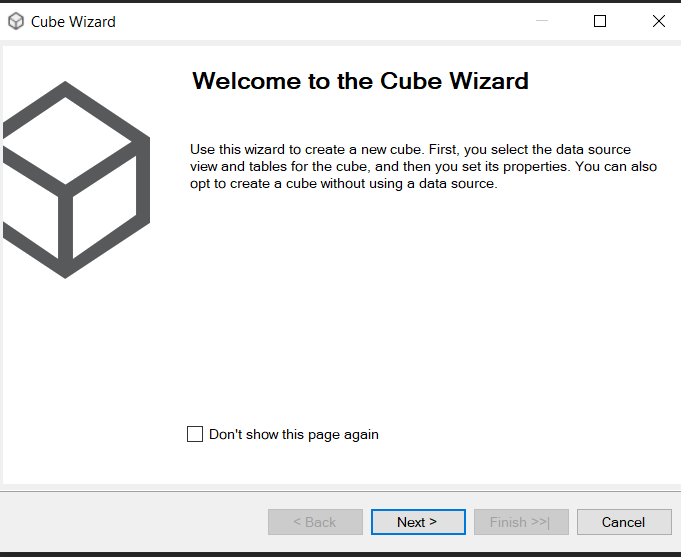


Sau đó chọn Finish

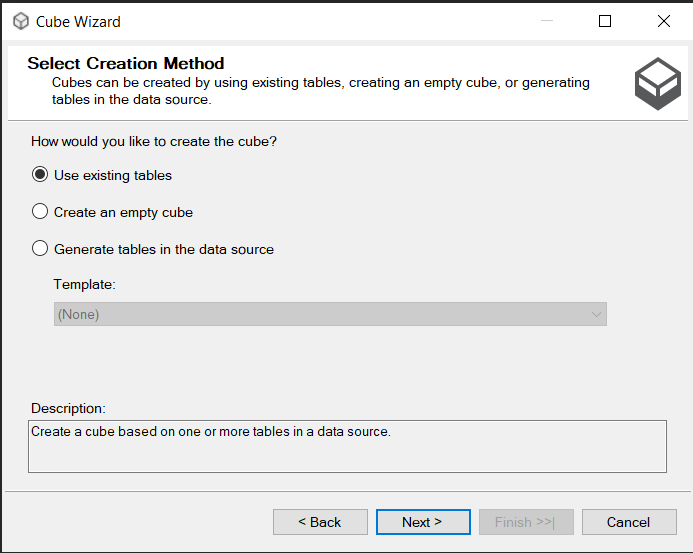


1. **Tạo Cube**

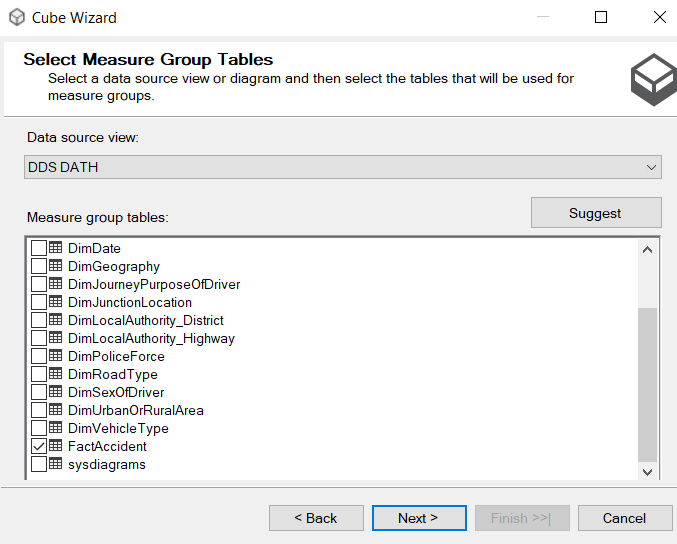
Chuột phải vào Cubes và chọn New Cube sau đó Next



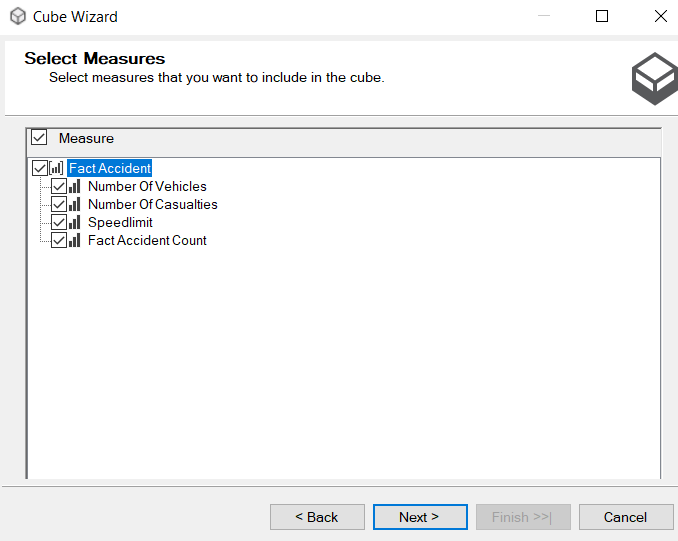
Sử dụng table đã tồn tại và chon Next



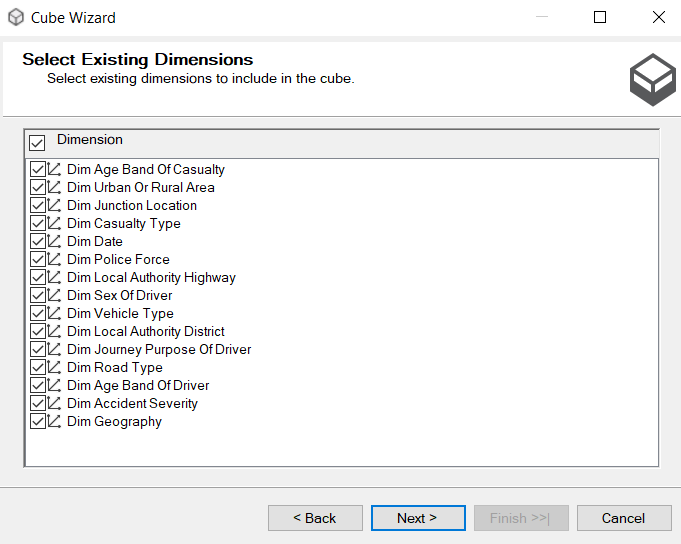
Chọn nút Suggest và chọn bảng Fact



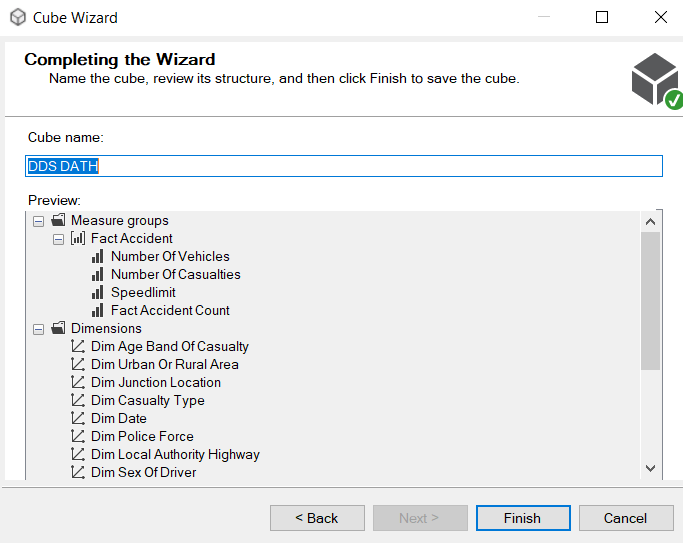
Chon các Measure cho Fact



Chon các bảng Dim



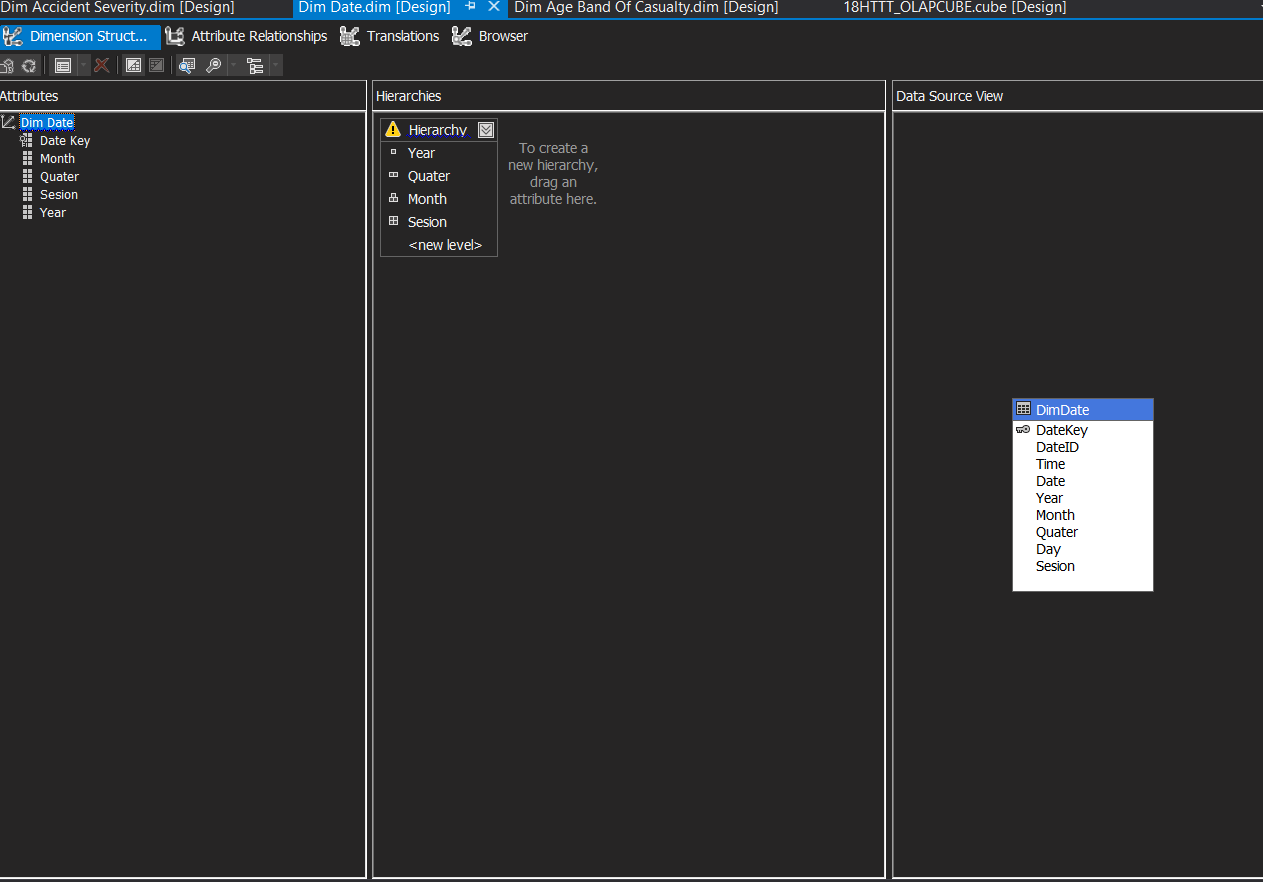
Sau đó chọn Finish



1. **Sữa bảng chiều**

Với bảng Dim\_Date

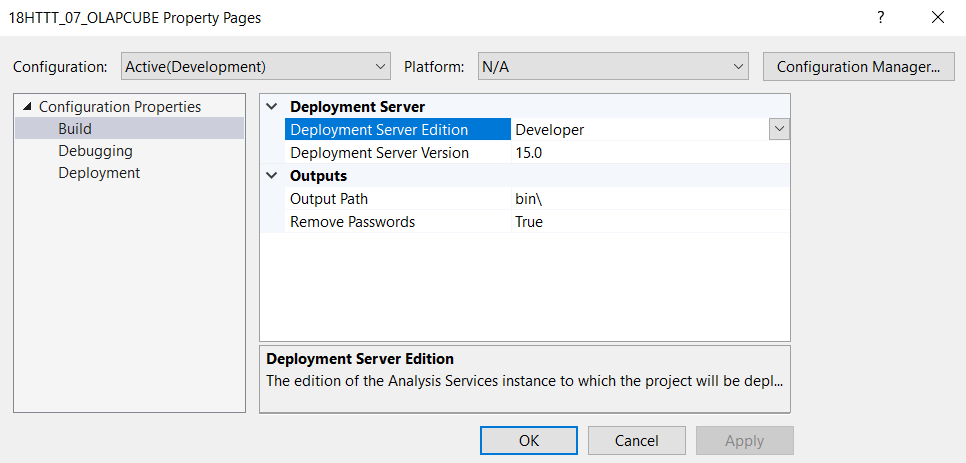
Trong Folder Dimensions. Nhấp đúp chuột vào Dimdate - Trong Data Source View, di chuyển các trường trong trong Dimdate vào Dim date trong Attributes.



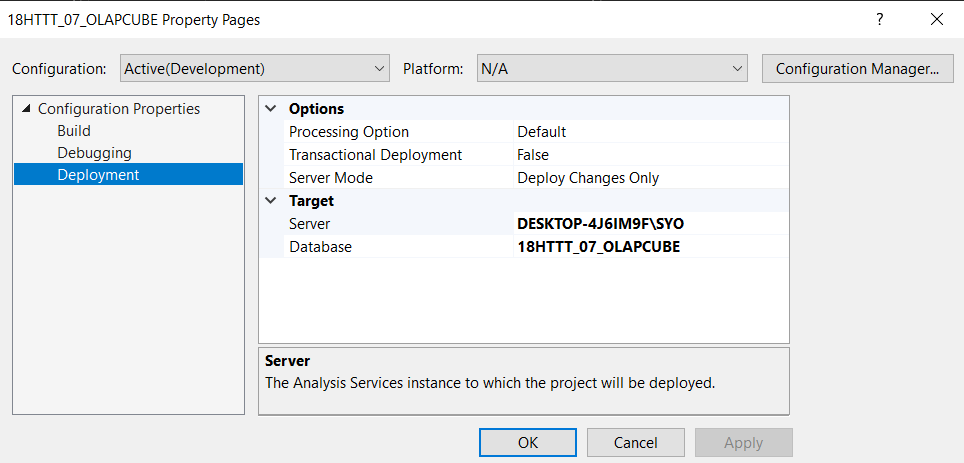
Tương tự cho các bảng còn lại

## Deploy cube lên server

Trong pane Solution Explorẻ , chuột phải vào tên project và chọn Properties



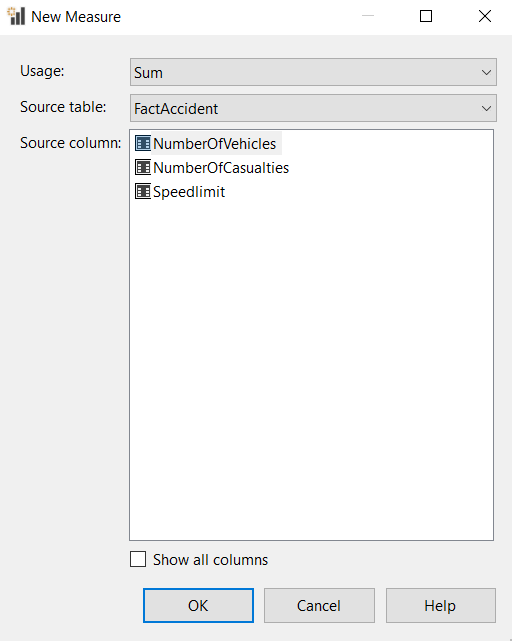
Tại mục Deployment đổi server thành server name của SQL Server sau đó nhấn Start để tiến hành deploy.



## Tạo mới Measure

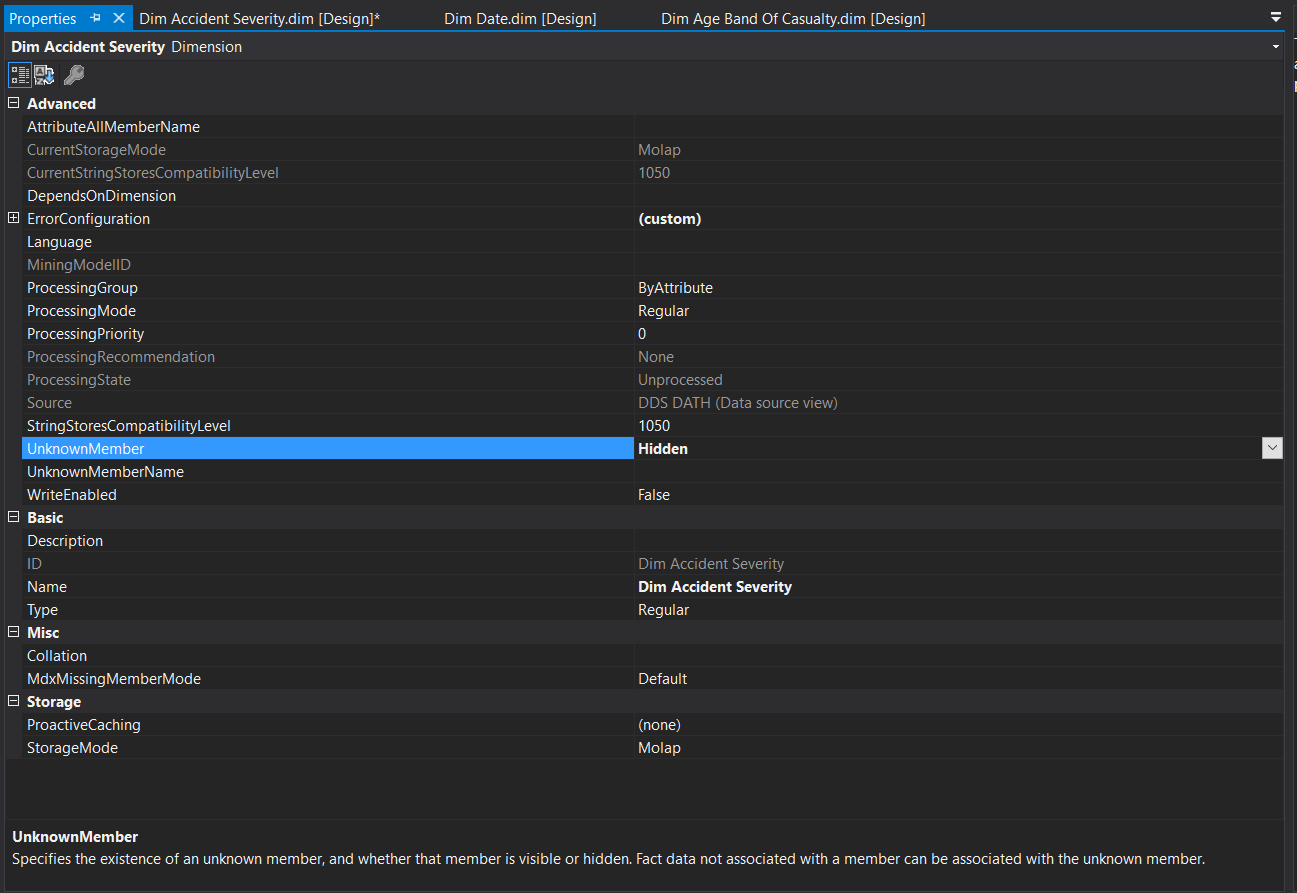
Chuyển sang Cube Designer bằng cách đúp chuột vào 18HTTT\_OLAPCUBE.cube.

Trong tab Cube Structure ở ô Measure, nhấp phải vào Fact và chon New Measure



1. **Xóa dòng dữ liệu Unknow**

Chọn bảng DimAccidentServity, chuột phải vào tên bảng ở mục Attribute và chon Properties.Chỉnh sữa ở dùng UnknowMemberName, đổi Visible thành Hidden.



Tương tự cho các bảng còn lại.

# **Chương 4 Thực hiện ngôn ngữ MDX**

## Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng (Fatal, Serious, Slight) ở các Địa phương (Local\_Authority\_(District)) trong tất cả các năm.

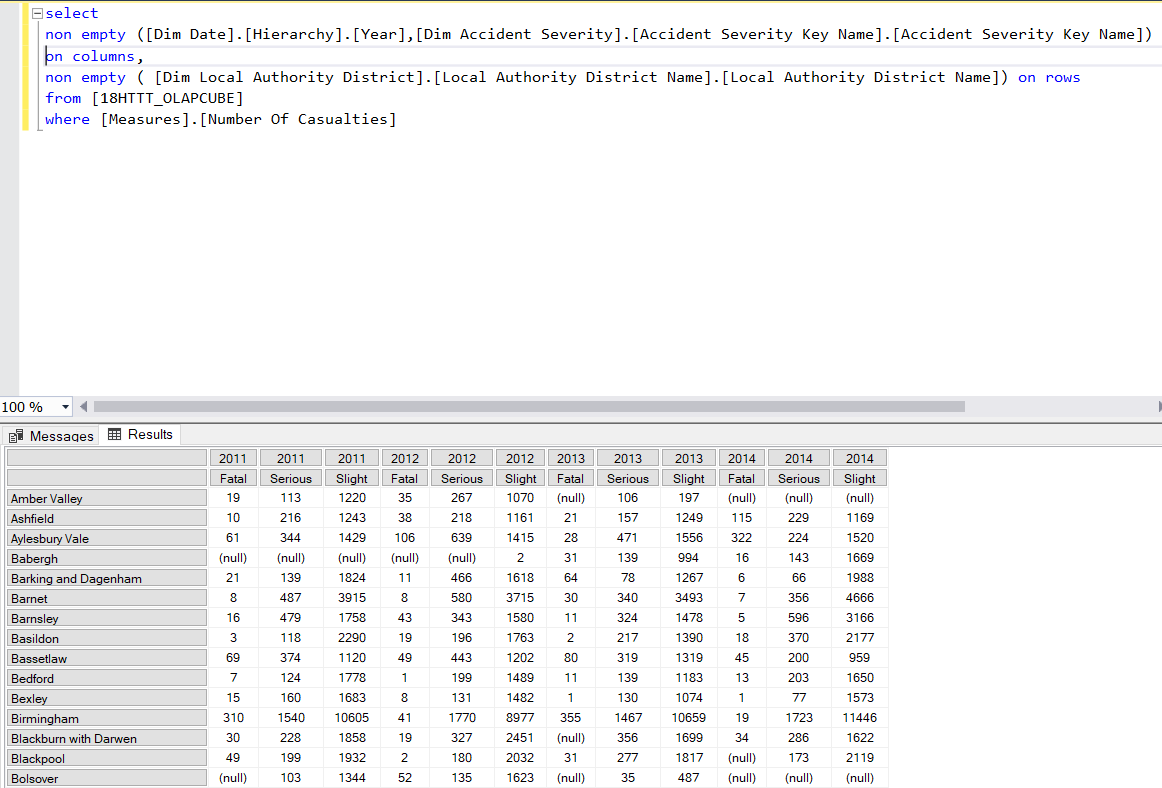
select

non empty ([Dim Date].[Hierarchy].[Year],[Dim Accident Severity].[Accident Severity Key Name].[Accident Severity Key Name]) on columns,

non empty ( [Dim Local Authority District].[Local Authority District Name].[Local Authority District Name]) on rows

from [18HTTT\_OLAPCUBE]

where [Measures].[Number Of Casualties]



* + - 1. **Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa Phương (Local\_Authority\_(District)) theo các Quý trong từng năm.**

select

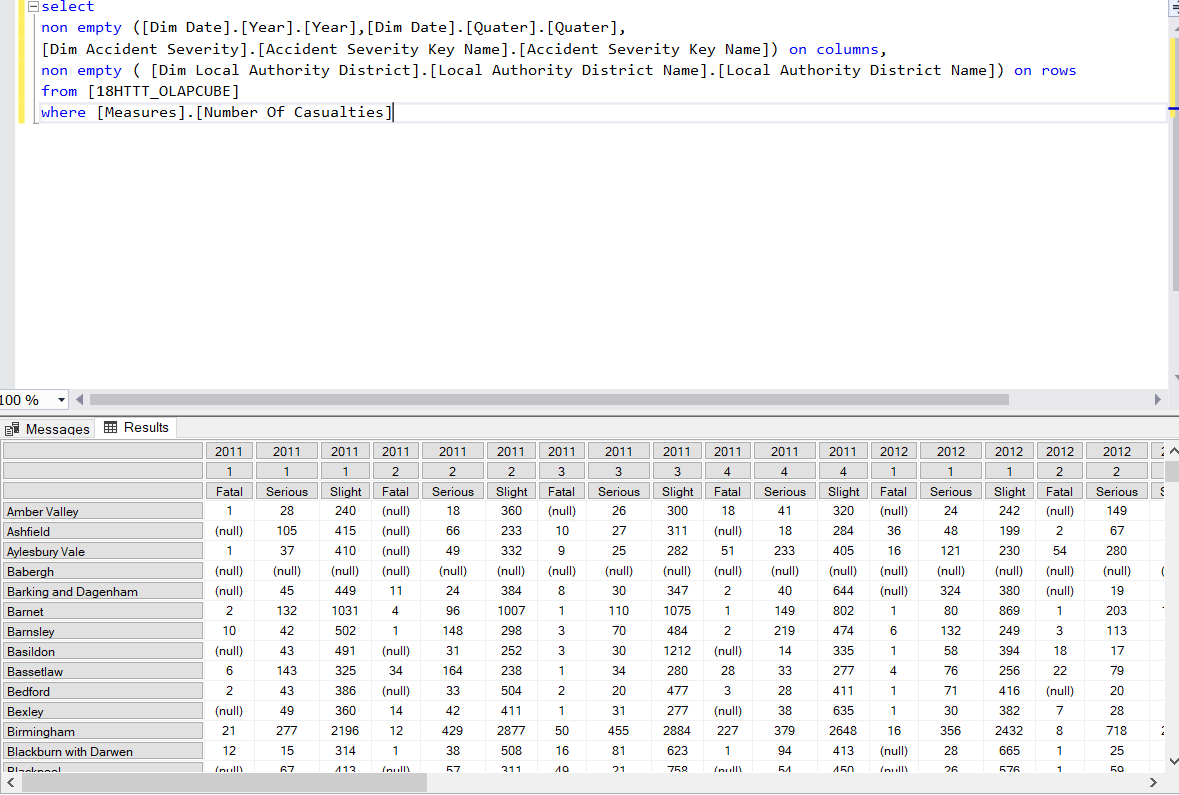
non empty ([Dim Date].[Year].[Year],[Dim Date].[Quater].[Quater],

[Dim Accident Severity].[Accident Severity Key Name].[Accident Severity Key Name]) on columns,

non empty ( [Dim Local Authority District].[Local Authority District Name].[Local Authority District Name]) on rows

from [18HTTT\_OLAPCUBE]

where [Measures].[Number Of Casualties]



* + - 1. **Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng và Thời Điểm Trong Ngày (Morning: 5am-12pm, Afternoon: 12pm-5pm, Evening: 5pm-9pm, Night: 9pm-5am) trong các năm.**

select

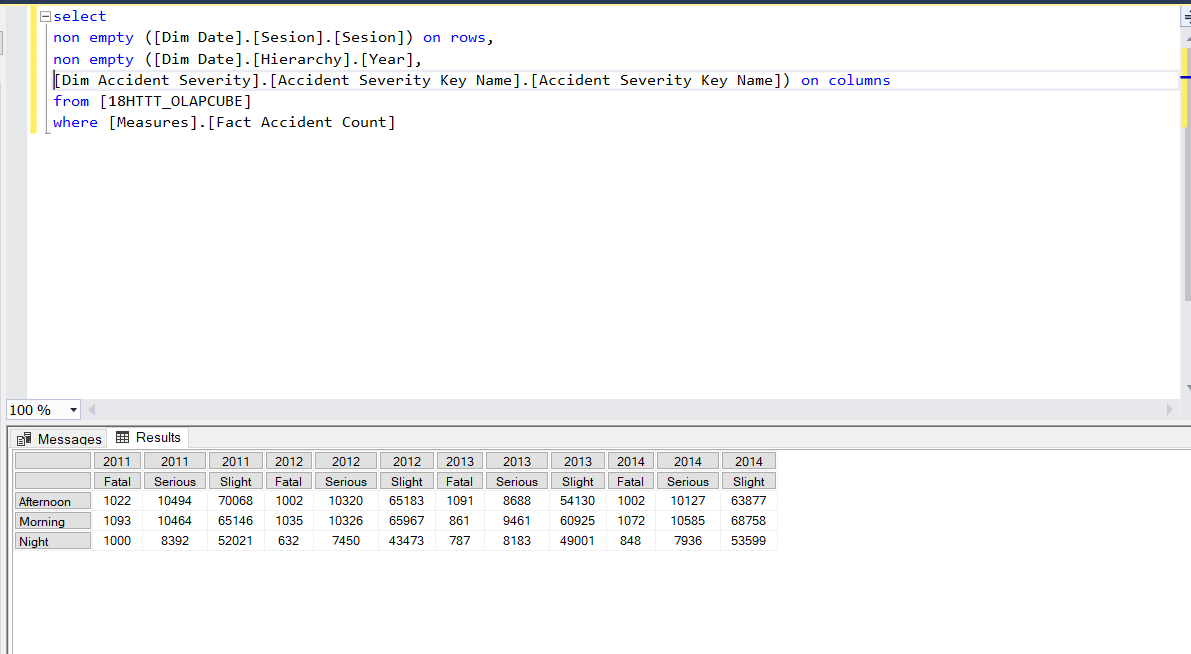
non empty ([Dim Date].[Sesion].[Sesion]) on rows,

non empty ([Dim Date].[Hierarchy].[Year],

[Dim Accident Severity].[Accident Severity Key Name].[Accident Severity Key Name]) on columns

from [18HTTT\_OLAPCUBE]

where [Measures].[Fact Accident Count]



* + - 1. **Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng, Vùng (Urban\_or\_Rural\_Area), và Kiểu Đường (Road Type) trong các năm.**

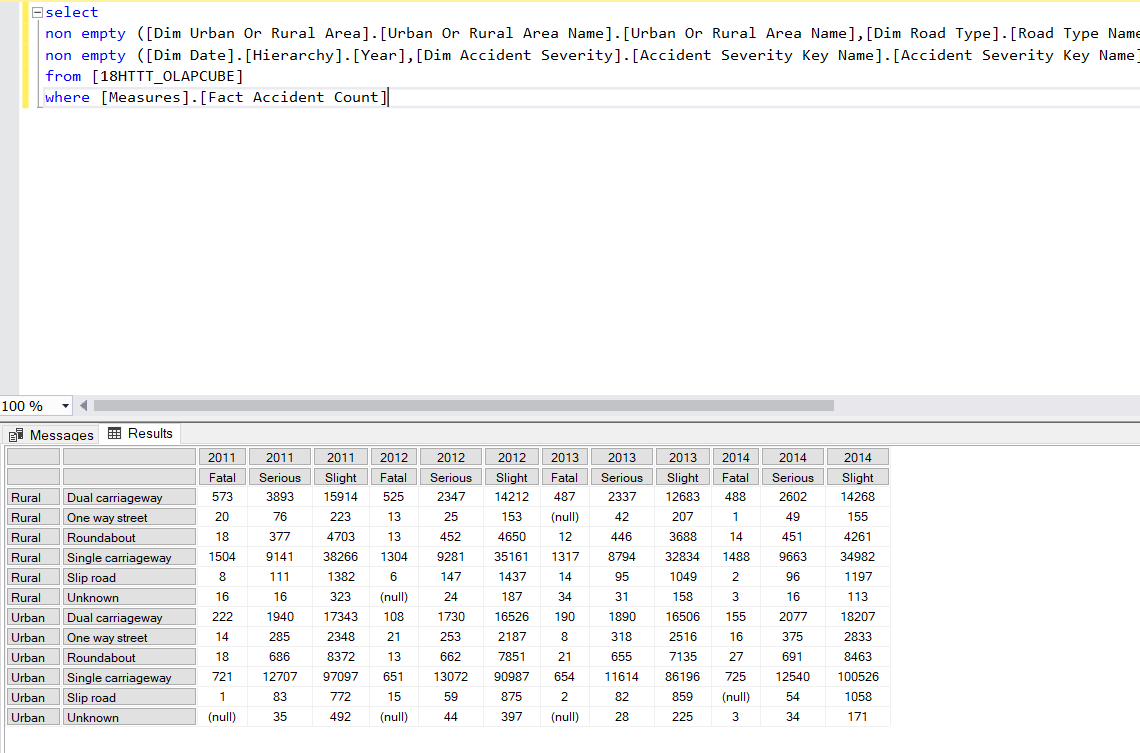
select

non empty ([Dim Urban Or Rural Area].[Urban Or Rural Area Name].[Urban Or Rural Area Name],[Dim Road Type].[Road Type Name].[Road Type Name]) on rows,

non empty ([Dim Date].[Hierarchy].[Year],[Dim Accident Severity].[Accident Severity Key Name].[Accident Severity Key Name] ) on columns

from [18HTTT\_OLAPCUBE]

where [Measures].[Fact Accident Count]



* + - 1. **Tổng hợp số lượng tai nạn theo Mục Đích Hành Trình (Journey Purpose) và Loại Phương Tiện (Vehicle\_Type).**

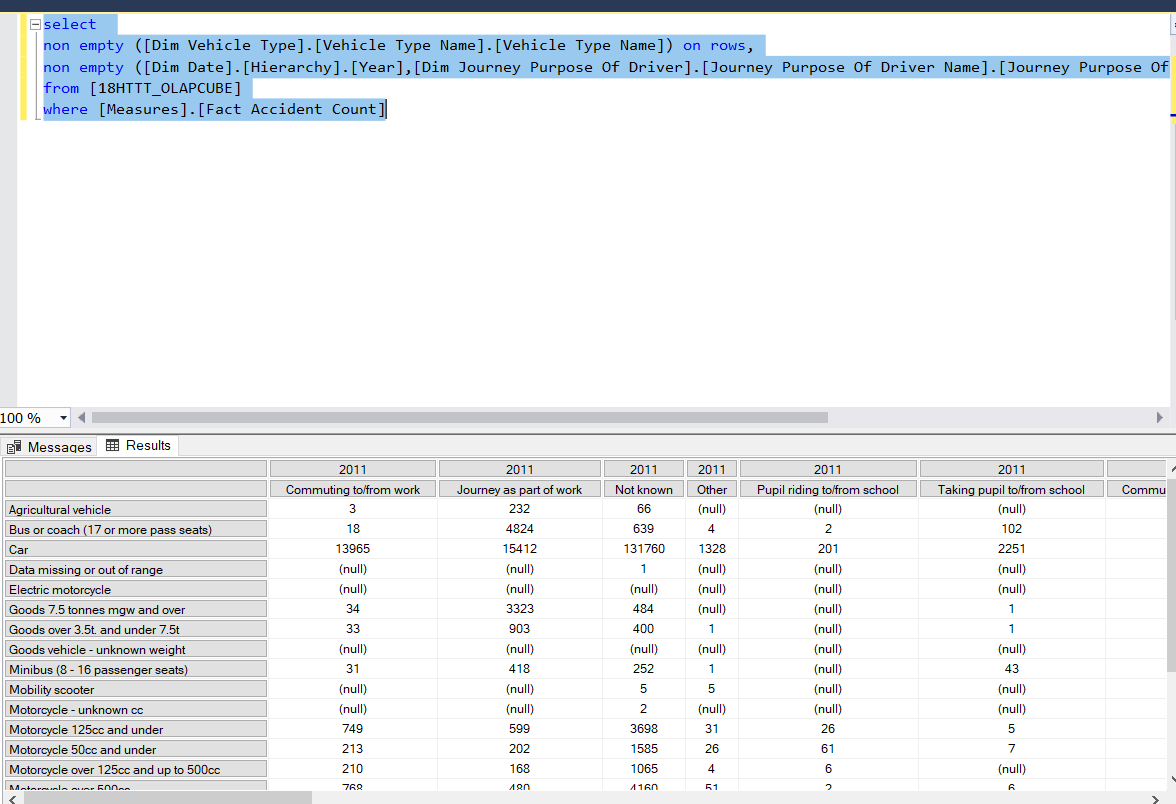
select

non empty ([Dim Vehicle Type].[Vehicle Type Name].[Vehicle Type Name]) on rows,

non empty ([Dim Date].[Hierarchy].[Year],[Dim Journey Purpose Of Driver].[Journey Purpose Of Driver Name].[Journey Purpose Of Driver Name] ) on columns

from [18HTTT\_OLAPCUBE]

where [Measures].[Fact Accident Count]



* + - 1. **Dùng regional map để biểu diễn trực quan (bằng màu sắc) số lượng TNGT ở các vùng trong năm.**

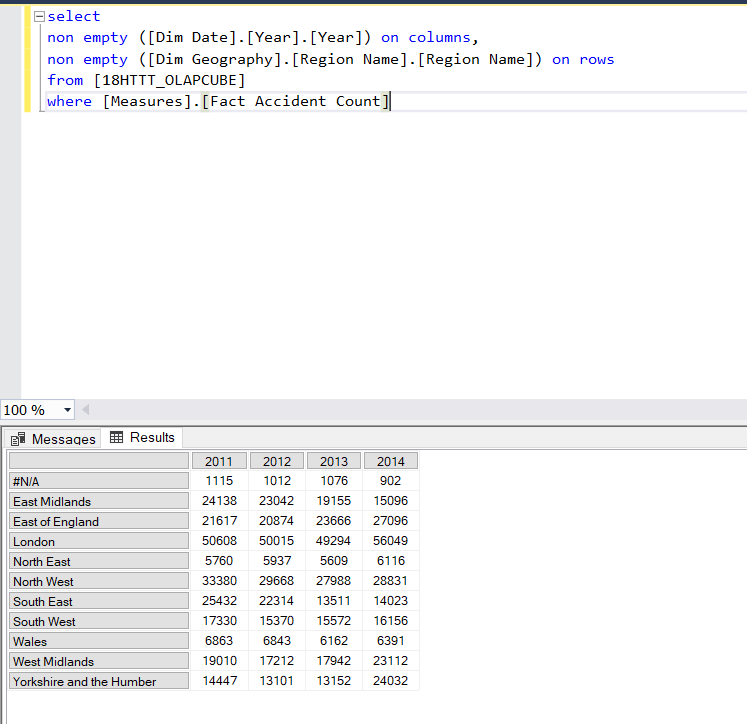
select

non empty ([Dim Date].[Year].[Year]) on columns,

non empty ([Dim Geography].[Region Name].[Region Name]) on rows

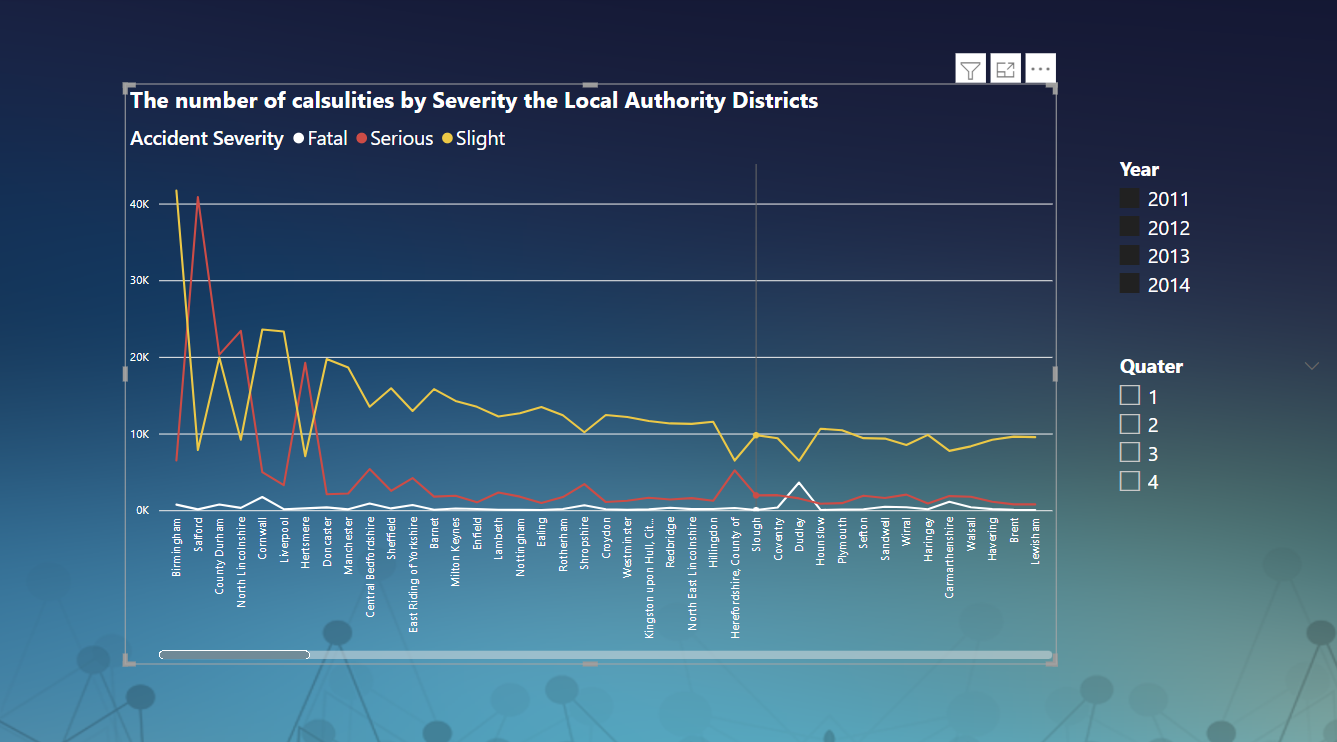
from [18HTTT\_OLAPCUBE]

where [Measures].[Fact Accident Count]



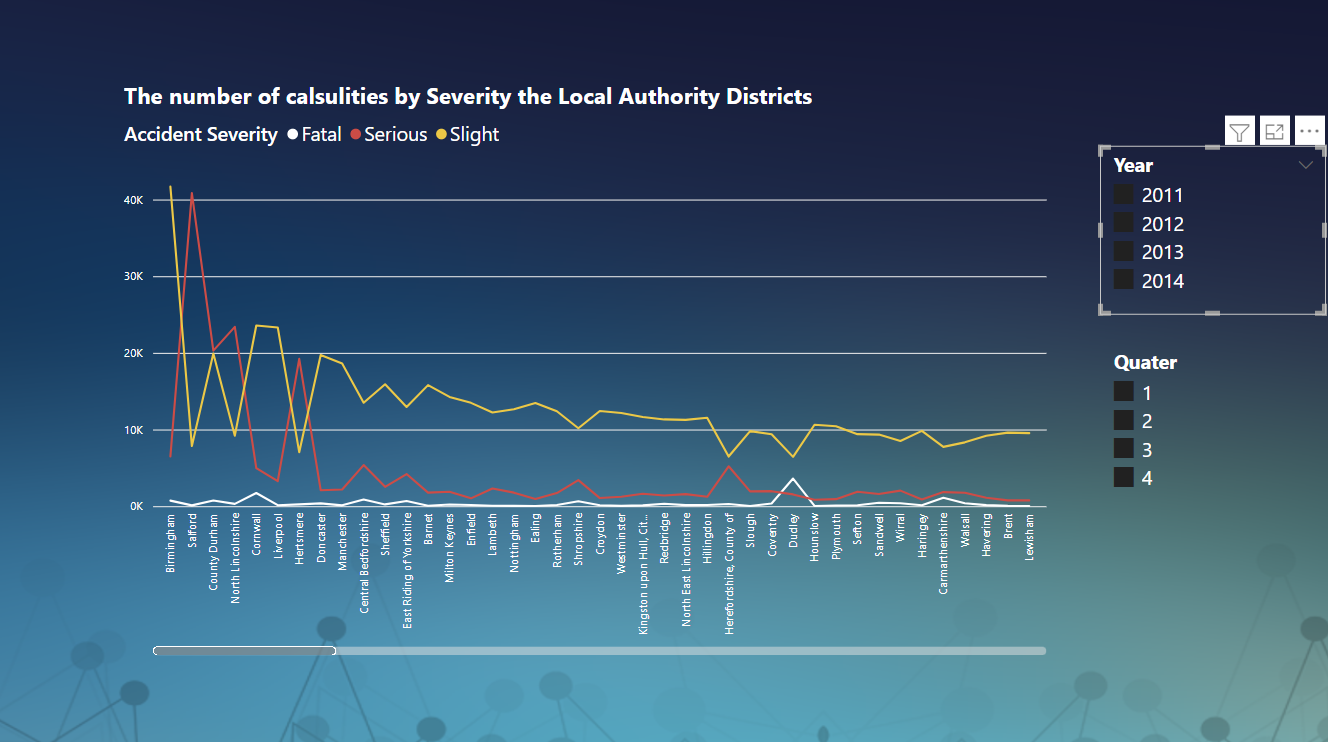
# **Chương 5: Trực quan hóa dữ liệu**

## Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng (Fatal, Serious, Slight) ở các Địa phương (Local\_Authority\_(District)) trong tất cả các năm.



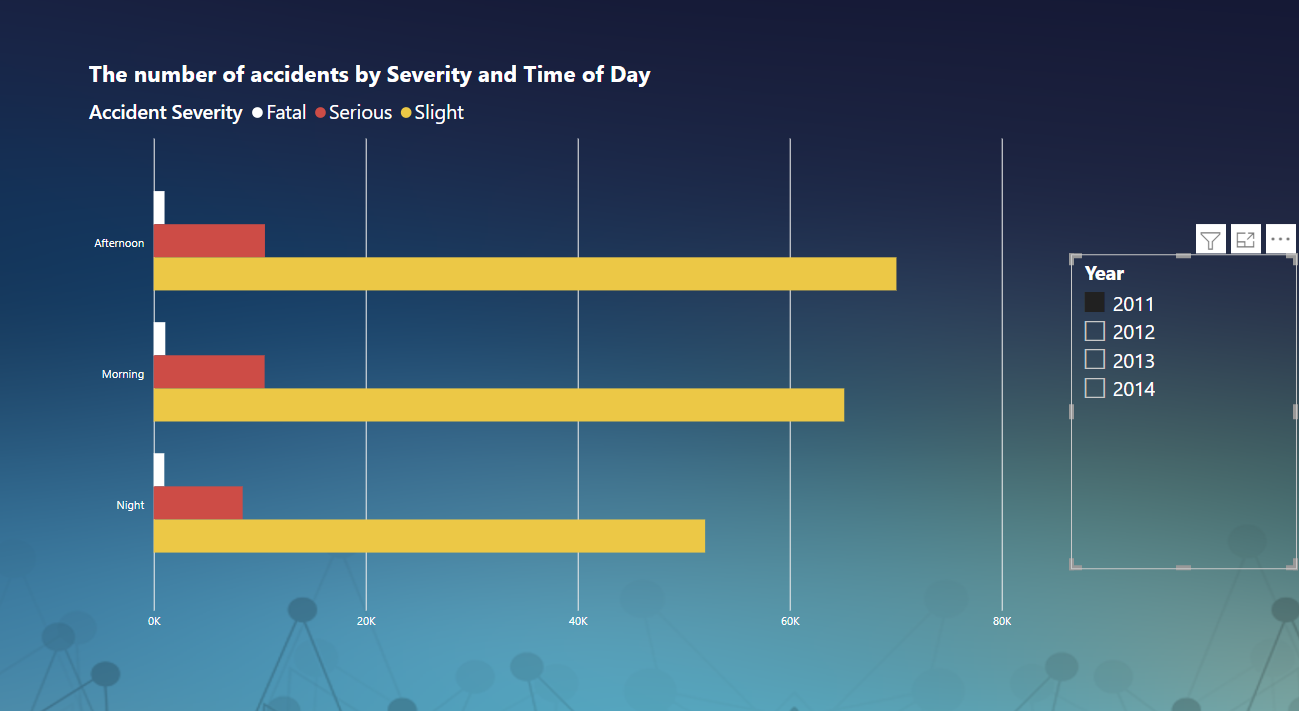
Từ năm 2012 đến 2014 Mức độ gây ra tử vong ở East Dorset là cao nhất, mức độ nhẹ ở Birmingham là cao nhất, mức độ nghiêm trọng ở Central Bedfordshire

**Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa Phương (Local\_Authority\_(District)) theo các Quý trong từng năm.**



Từ năm 2012 đến 2014 qua các quý, Mức độ gây ra tử vong ở Uttlestford là cao nhất, mức độ nhẹ ở Birmingham là cao nhất, mức độ nghiêm trọng ở Salfard

**Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng và Thời Điểm Trong Ngày (Morning: 5am-12pm, Afternoon: 12pm-5pm, Evening: 5pm-9pm, Night: 9pm-5am) trong các năm.**



Năm 2012 Số lượng tai nạn gây ra tử vong, nhẹ hoặc nghiêm trọng đều vào khoảng buổi sáng

**Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng, Vùng (Urban\_or\_Rural\_Area), và Kiểu Đường (Road Type) trong các năm.**



Năm 2012 đa phần tai nạn loại nhẹ và nghiêm trọng diển ra ở thành phố còn mức độ gây tử vong lại diễn ra ở vùng nông thôn nhiều hơn

**Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng, Loại Nạn Nhân (Casualty Type) và Độ Tuổi trong các năm, Độ Tuổi được định nghĩa như sau:**

Children: 0-15

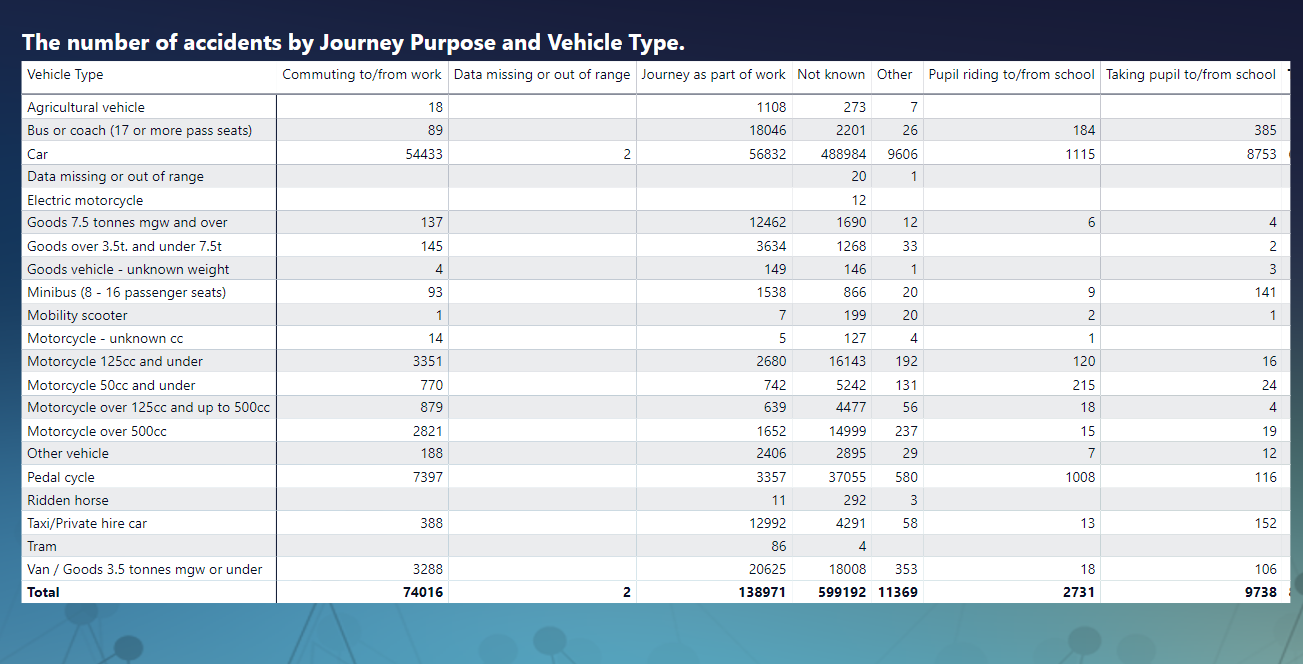
Young adult: 0-17

Adult: 18-59

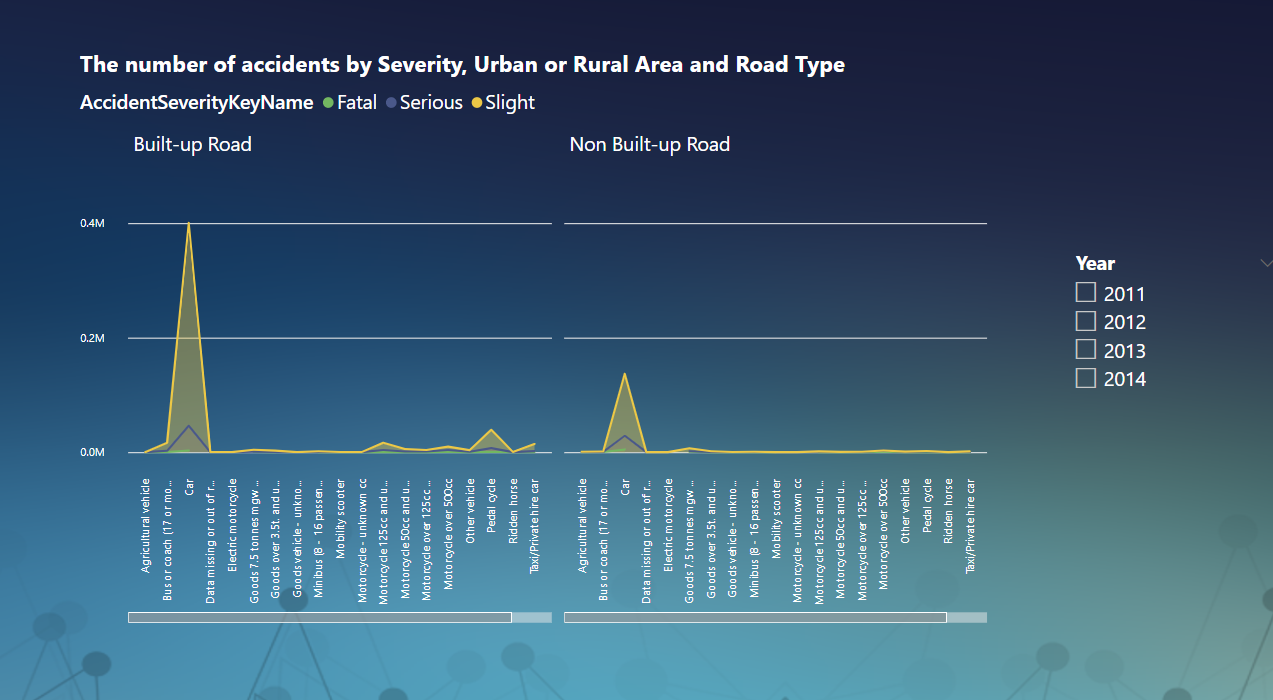
1. nd over: 60-...



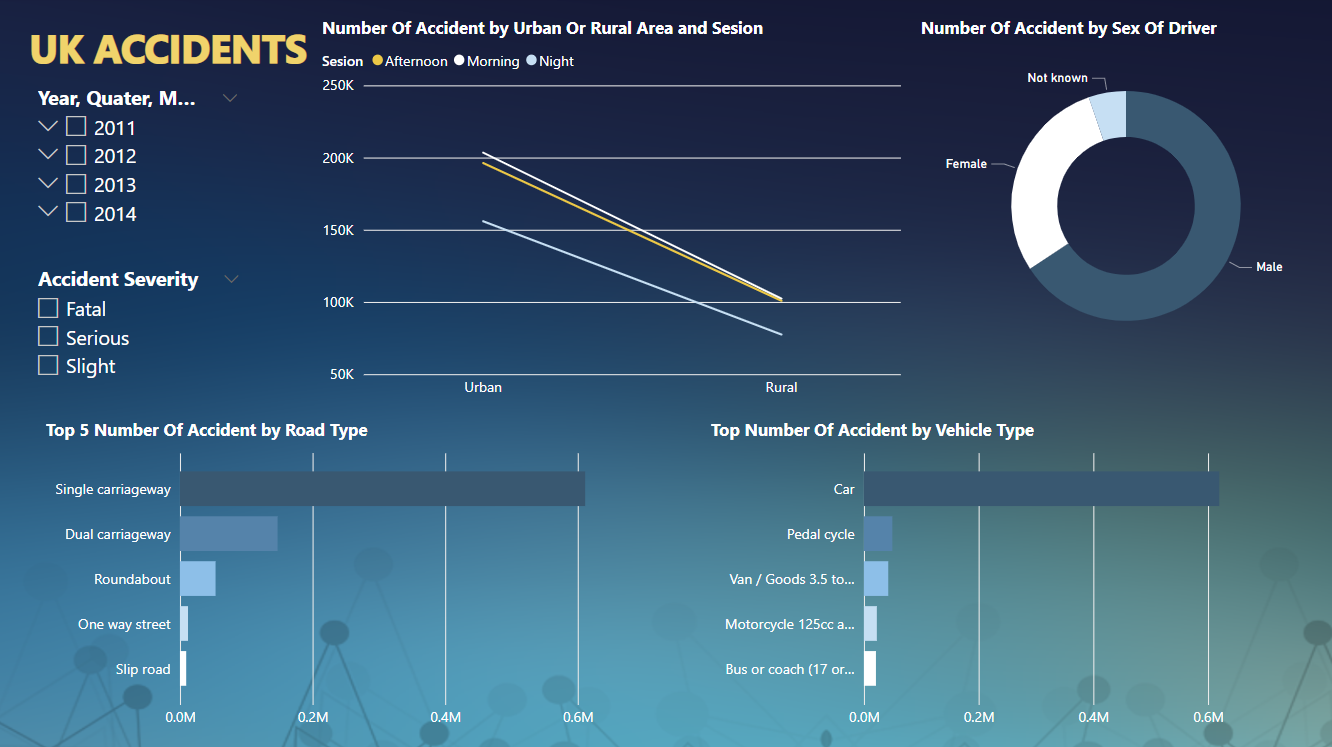
**Tổng hợp số lượng tai nạn theo Mục Đích Hành Trình (Journey Purpose) và Loại Phương Tiện (Vehicle\_Type).**



* + - 1. **Thống kê số lượng tai nạn theo Mức Độ Nghiêm Trọng, Loại Phương Tiện (Vehicle Type), Built-up Road trong các năm.**

****

1. **Một vài biểu đồ cho tai nạn giao thông**



## Dùng regional map để biểu diễn trực quan (bằng màu sắc) số lượng TNGT ở các vùng trong năm.

****

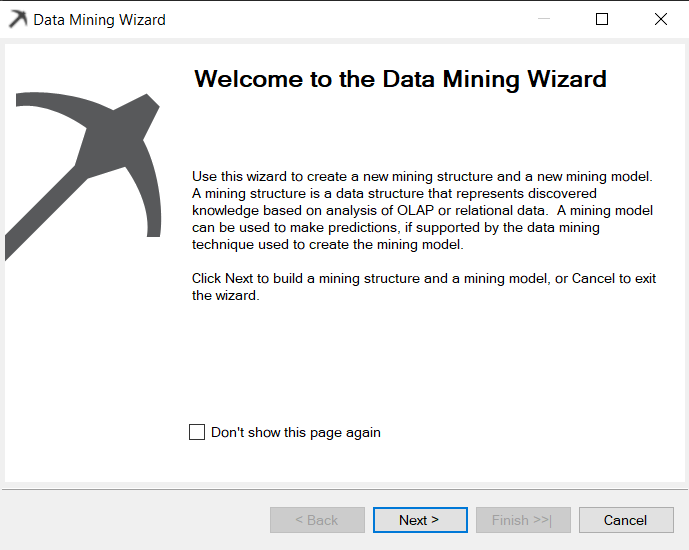
# **Chương 6: Data Mining**

Phân tích cho yêu cầu tỉ lệ xảy ra tai nạn ở vùng thành thị và nông thôn đồng thời dự đoán tỉ lệ tai nạn một người xảy ra tai nạn giao thông bằng Decision Free.

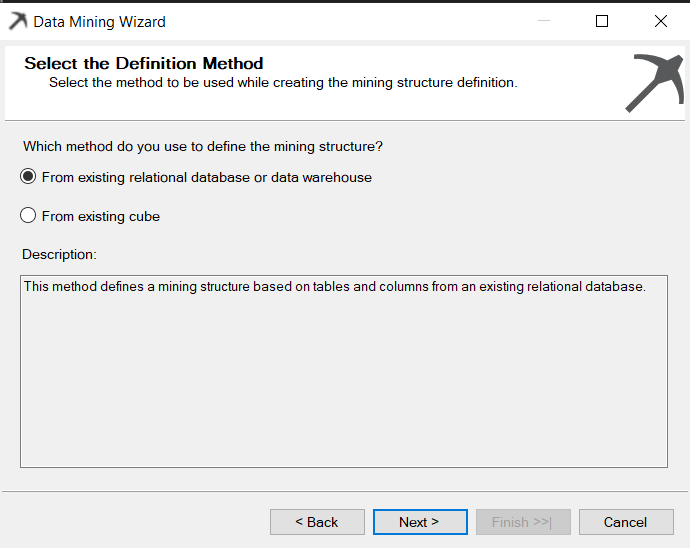
1. **Thực hiện quá trình Data mining**

* Tạo Data Source: Tương tự bước tạo data source ở mục Tạo Data Source của Olap Cube ( Lấy NDS làm dữ liệu nguồn )
* Tạo Data Source View: Tương tự bước tạo data source ở mục Tạo Data Source của Olap Cube
* Tạo Mining Structure

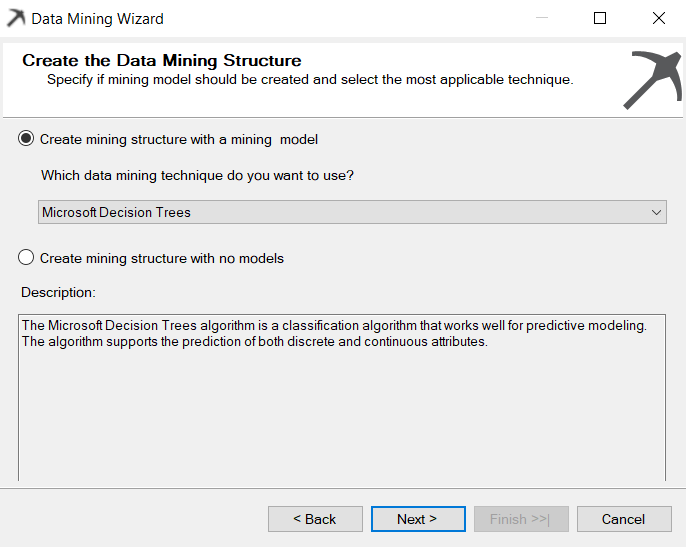
Chuột phải vào mining structures chọn New Data Structures sau đó Next



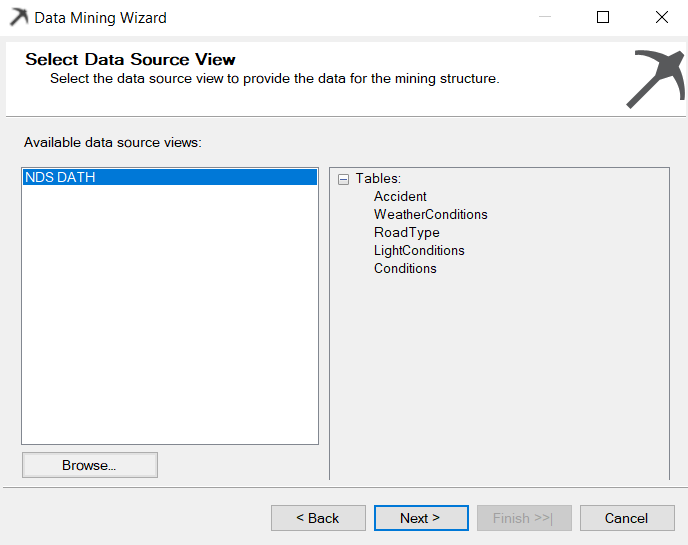
Chọn data có sẵn



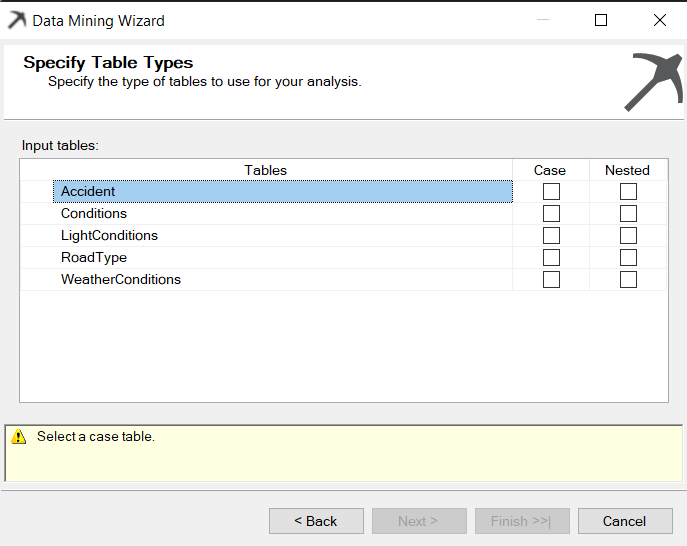
Chọn loại Data mining structure



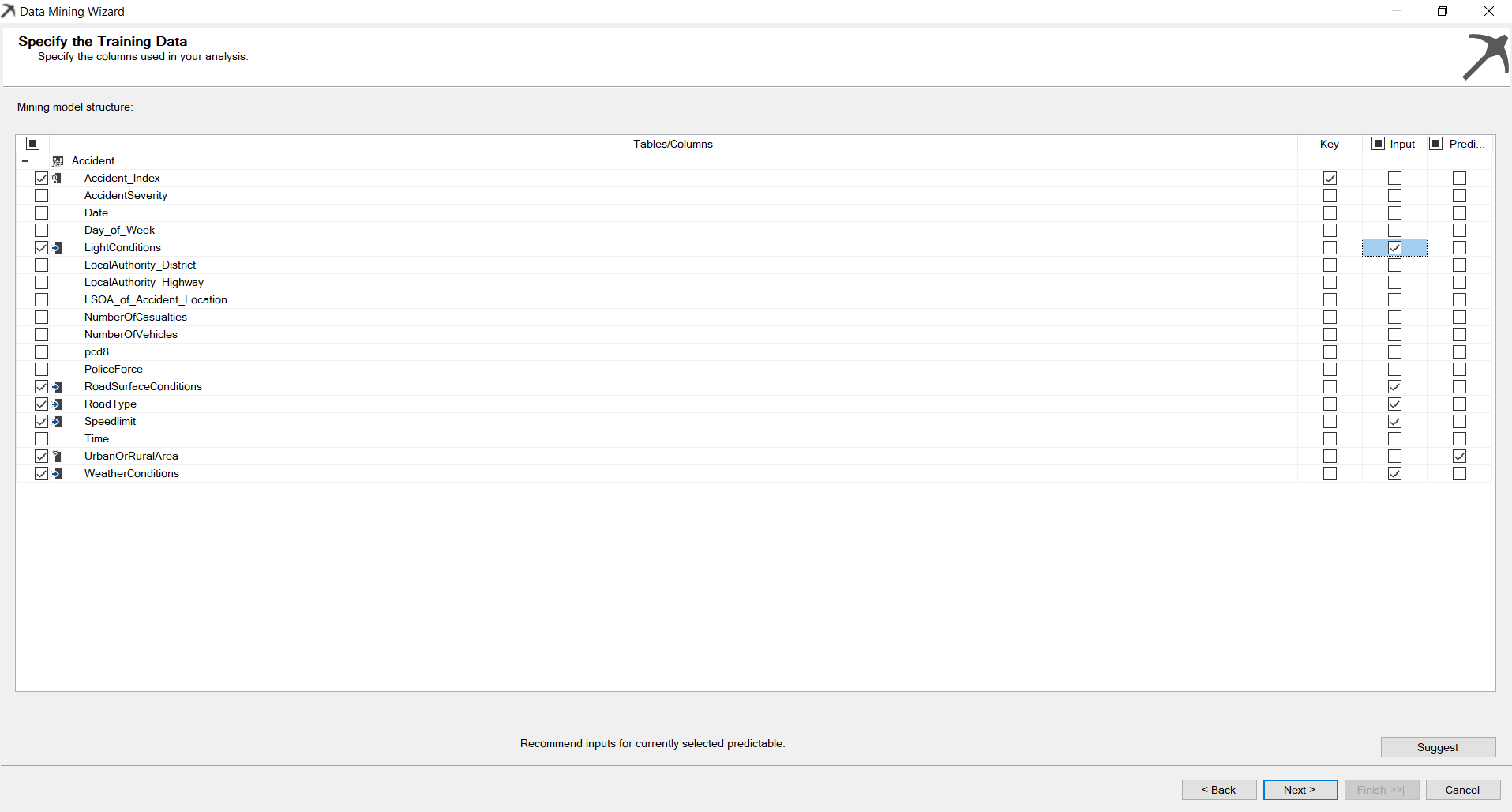
Connect đến CSDL tạo ở Data Source View



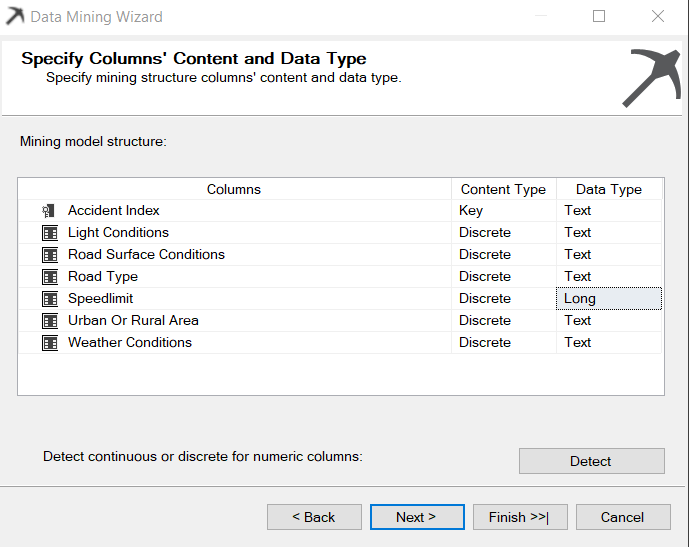
Tiếp theo chọn khóa chính



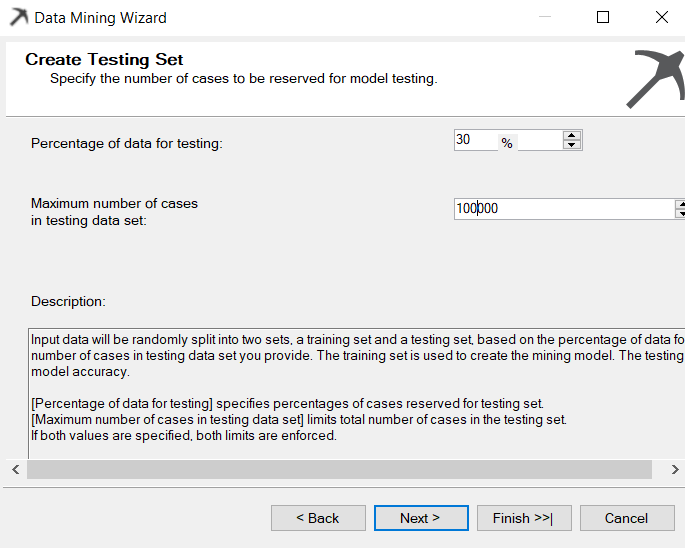
Chọn các Input đầu vào và biến muốn dự đoán



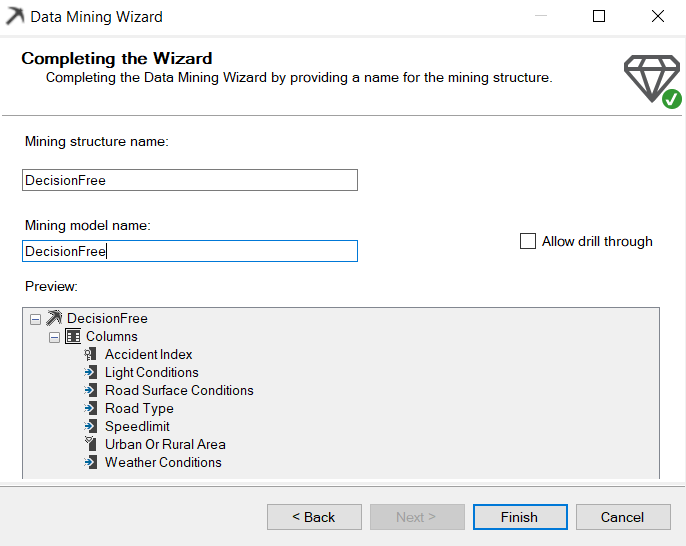
Thay đổi kiểu dữ liệu



Chia tập train test và số trường hợp lớn nhất cho mẫu testing là 100000



Đổi tên cho mining structure

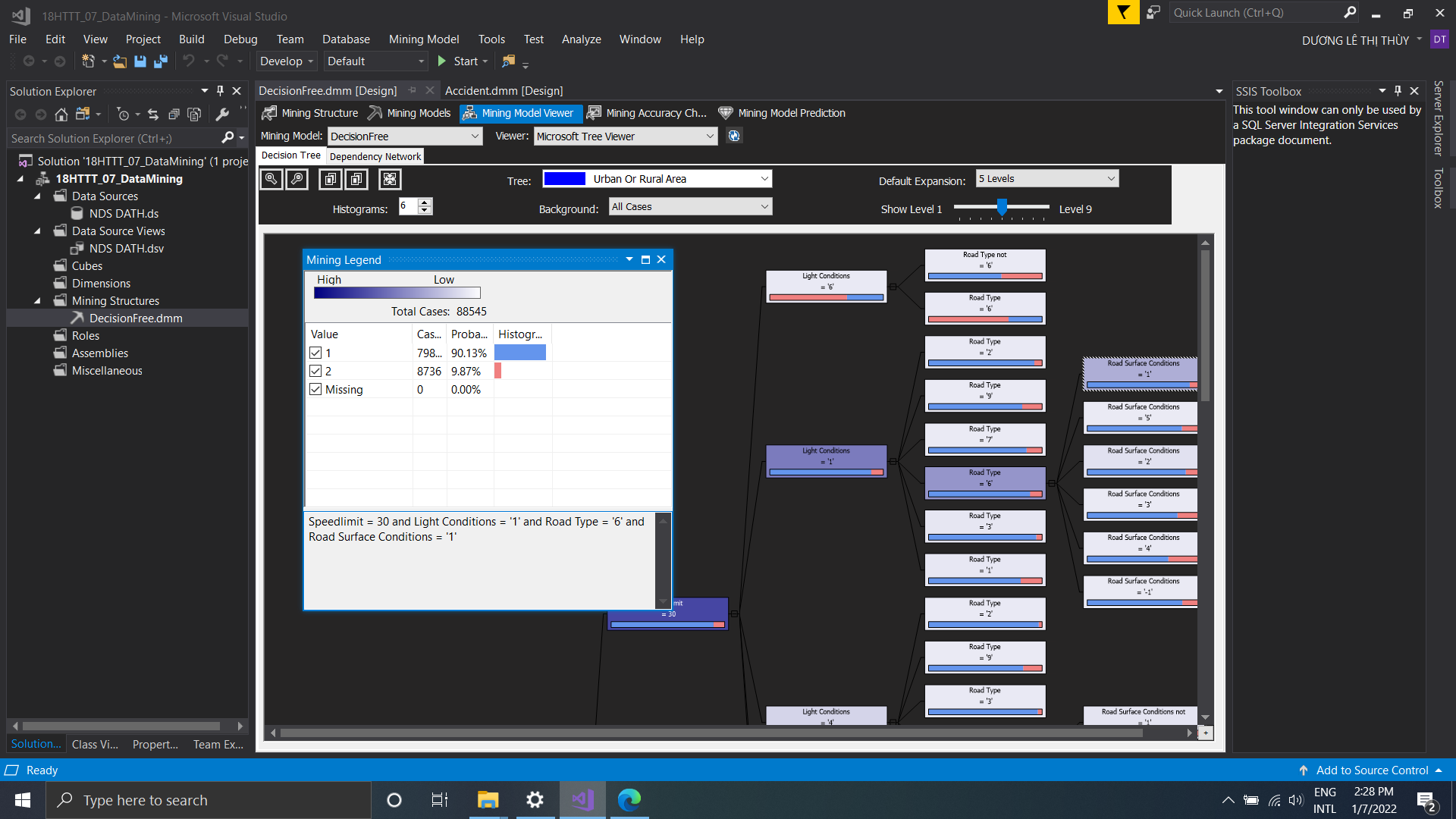


Sau khi Finish tiến hành Deploy để chạy Model

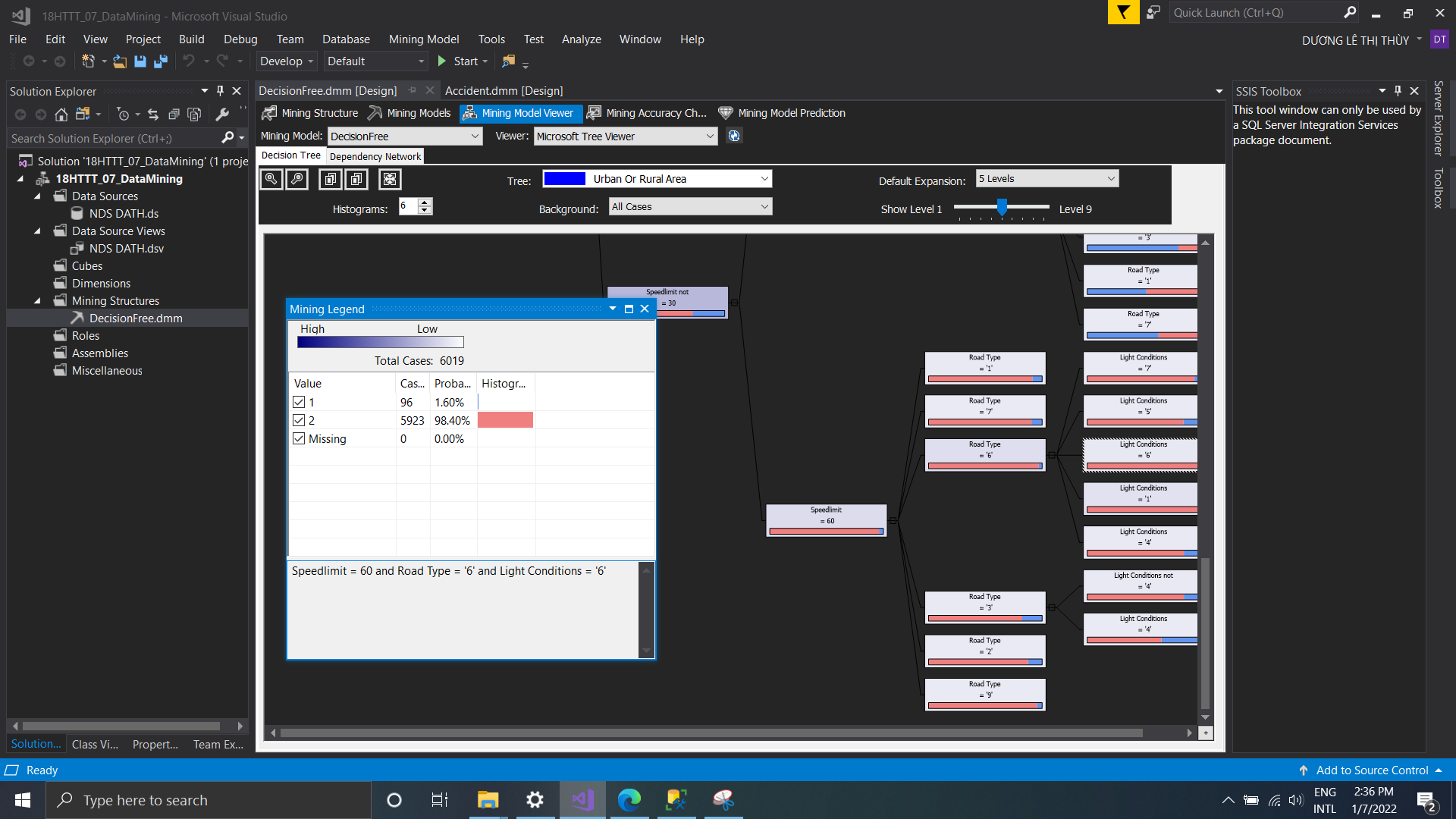
1. **Desicion Free**

Xem thông tin của 5 lớp

Trong điều kiện tốc độ giới hạn là 30, điều kiện ánh sáng là 1, Road Type là 6, Road Surface là 1 thì tỉ lệ xảy ra tai nạn giao thông của thành thị là 90.13%



Trong điều kiện tốc độ giới hạn là 60, Road Type là 6, Light Condition là 6 tỉ lệ xảy ra tai nạn ở nông thôn là 98.40%



Dependency Network của thuật toán Decision Free cho thấy những thuộc tính ảnh hưởng đến việc xảy ra tai nạn giao thông.

Những thuộc tính có ảnh hưởng giảm dần: Speedlimit -> RoadType -> Light Conditions -> Road Surface Condition -> Weather Condition 