

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN



MÔN HỌC: Hệ thống thông tin phục vụ trí tuệ kinh doanh

CHỦ ĐỀ: Đồ án thực hành - Xây dựng và khai thác KDL

NHÓM: TTKD-10

GVHD: Hồ Thị Hoàng Vy, Nguyễn Thị Như Anh, Tiết Gia Hồng

Người thực hiện

18120469 – Nguyễn Hoài Nam

18120510 – Cao Xuân Hồng Phúc

18120518 – Phạm Thị Bích Phượng

18120538 – Võ Nguyễn Hồng Sơn

TP Hồ Chí Minh, ngày 10 tháng 1 năm 2022

NHẬN XÉT CỦA GIẢNG VIÊN

Tp. Hồ Chí Minh, ngày ... tháng ... năm 2022

Giảng viên hướng dẫn



Mục lục

I.	THÔNG TIN NHÓM.....	5
II.	BẢNG PHÂN CÔNG CÔNG VIỆC VÀ ĐÁNH GIÁ	5
III.	NỘI DUNG.....	6
1.	Giới thiệu.....	6
1.1	Mô tả đồ án	6
1.2	Mục tiêu đồ án.....	6
2.	Mô tả ý nghĩa các thuộc tính của các nguồn dữ liệu	7
2.1	Accidents.....	7
2.2	Vehicles	8
2.3	Casualties.....	9
2.4	Postcodes.....	11
2.5	PCD_OA_LSOA_MSOA_LAD_AUG21_UK_LU	12
3.	Xác định dữ liệu cho yêu cầu.....	12
4.	Data Flow Architect: NDS + DDS.....	15
5.	Xác định dimension, fact (measure) phù hợp từng yêu cầu phân tích	15
5.1	Phân tích.....	15
5.2	Kết quả.....	34
6.	Quá trình từ Source vào Stage	38
7.	Quá trình từ Stage vào NDS	40
7.1	Phân tích.....	40
7.2	Thực hành	47
8.	Quá trình từ NDS vào DDS	53
8.1	Dim Table	53
8.2	Fact Table	58
9.	Thiết kế và xây dựng Cube	74
10.	Khai thác dữ liệu	77
10.1	MDX.....	77
10.2	Report (Visualize)	82
10.3	Mining.....	93
a.	Mô tả bài toán:	93



b.	Thực hành xây dựng mô hình:	93
c.	Kết quả:	102
11. Điểm cộng.....	106	
11.1 Calculated measure	106	



I. THÔNG TIN NHÓM

MSSV	HỌ TÊN	EMAIL
18120469	Nguyễn Hoài Nam	18120469@student.hcmus.edu.vn
18120510	Cao Xuân Hồng Phúc	18120510@student.hcmus.edu.vn
18120518	Phạm Thị Bích Phượng	18120518@student.hcmus.edu.vn
18120538 (*)	Võ Nguyễn Hồng Sơn	18120538@student.hcmus.edu.vn

(*): Nhóm trưởng

II. BẢNG PHÂN CÔNG CÔNG VIỆC VÀ ĐÁNH GIÁ

Người thực hiện	Công việc thực hiện	Mức độ hoàn thành	Đánh giá của nhóm
18120469 – Nam	<ul style="list-style-type: none"> - Calculated Measure - MDX - Mô tả ý nghĩa thuộc tính - Xác định dữ liệu - Thiết kế DDS - Thiết kế xây dựng Cube 	100%	10/10
18120510 – Phúc	<ul style="list-style-type: none"> - Report(Visualize) – Region Map - Quá trình từ Stage - NDS - MDX - Mô tả ý nghĩa thuộc tính - Xác định dữ liệu - Thiết kế DDS 	100%	10/10

18120518 – Phượng	- Report(Visuazlie) - Quá trình từ NDS - DDS - MDX - Xác định dữ liệu - Mô tả ý nghĩa thuộc tính - Thiết kế DDS	100%	10/10
18120538 – Sơn	- Mining - Quá trình từ Source - Stage - MDX - Mô tả ý nghĩa thuộc tính - Xác định dữ liệu - Thiết kế DDS	100%	10/10

III. NỘI DUNG

1. Giới thiệu

1.1 Mô tả đồ án

Từ kiến thức lý thuyết đã học về KDL, OLAP, ETL, mining..., đồ án **Xây dựng và phân tích dữ liệu về các vụ tai nạn giao thông ở UK từ năm 2011 đến 2014** giúp sinh viên thực hành xây dựng một KDL cụ thể, biết triển khai ETL để rút trích dữ liệu từ nhiều nguồn, biết cách khai thác KDL với report, OLAP, mining, tạo job định kì thực hiện ETL.

1.2 Mục tiêu đồ án

Đồ án này nhằm mục tiêu đạt được các chuẩn đầu ra sau:

- Thiết kế được lược đồ chuẩn hóa, đa chiều(sao, bông tuyết) dựa vào dữ liệu hệ thống tác vụ và yêu cầu phân tích từ tình huống cho trước.
- Triển khai quy trình ETL để rút trích dữ liệu từ nhiều nguồn, biến đổi, làm sạch dữ liệu, nạp dữ liệu vào kho dữ liệu(KDL) sử dụng SSIS.

3. Xây dựng KDL đa chiều sử dụng SSAS và giải thích được lựa chọn phép toán OLAP phù hợp đối với 1 số yêu cầu phân tích.
4. Sử dụng 1 số công cụ biểu diễn dữ liệu(SSRS, powerBI, exel...) để biểu diễn kết quả phân tích, khai thác được(report, dashboard)
5. Sử dụng SSAS và áp dụng các kĩ thuật mining tích hợp để thực hiện khai thác dữ liệu từ KDL xây dựng được.

2. Mô tả ý nghĩa các thuộc tính của các nguồn dữ liệu

2.1 Accidents

STT	Thuộc tính	Kiểu dữ liệu	Mô tả
1	Accident Index	string	Mã vụ tai nạn
2	Police Force	int	Mã lực lượng cảnh sát
3	Accident Severity	int	Mức độ nghiêm trọng: tử vong, nghiêm trọng , nhẹ
4	Day of Week	int	Các ngày thứ trong tuần: Sunday, Monday,...
5	Local Authority (District)	int	Tên khu vực địa phương xảy ra vụ việc
6	Local Authority (Highway Authority - ONS code)	string	Tên đường chính xảy ra tai nạn
7	1st Road Class	int	Đường cấp 1
8	Road Type	int	Loại đường: EX: One way street (Đường 1 chiều)
9	Junction Detail	int	Chi tiết giao lộ: <ul style="list-style-type: none">• Bùng binh• Ngả tư
10	Junction Control	int	Kiểm soát giao lộ: <ul style="list-style-type: none">• Người có quyền• Tín hiệu giao thông
11	2nd Road Class	int	Đường cấp 2
12	Pedestrian Crossing-Human Control	int	Người kiểm soát phân luồng đưa người qua đường
13	Pedestrian Crossing-Physical Facilities	int	Cơ sở vật chất dành cho người đi bộ
14	Light Conditions	int	Các điều kiện ánh sáng Ví dụ: bóng tối – đèn không sáng
15	Weather Conditions	int	Các điều kiện thời tiết Ví dụ: Gió mạnh - mưa

16	Road Surface Conditions	int	Các điều kiện mặt đường Ví dụ: Khô, ẩm
17	Special Conditions at Site	int	Các điều kiện đặc biệt ở hiện trường Ví dụ: Mặt đường bị lỗi
18	Carriageway Hazards	int	Các mối nguy hiểm trên đường đi: Ví dụ: Chó trên đường, tai nạn trước đó
19	Urban or Rural Area	int	Thành thị hay nông thôn
20	Did Police Officer Attend Scene of Accident	int	Cảnh sát có tham dự hiện trường vụ tai nạn không
21	LSOA_of_Accident_Location	string	Mã địa điểm tai nạn
22	Time	DateTime	Thời gian xảy ra vụ tai nạn

2.2 Vehicles

STT	Thuộc tính	Kiểu dữ liệu	Mô tả
1	Accident_Index	String	Số thứ tự chỉ mục các vụ tai nạn
2	Vehicle_Reference	Int	Thứ tự của phương tiện trong vụ tai nạn
3	Vehicle_Type	Int	Các giá trị được biểu diễn dạng số theo bảng UK Accidents – Codebook. Mỗi giá trị tương ứng với một loại xe khác nhau. Ví dụ 1: Pedal cycle, 18 :Tram,....
4	Vehicle_Manoeuvre	Int	Trước lúc tai nạn, phương tiện đã di chuyển như thế nào. 1: Nearside, 7:Offside.
5	Vehicle_Location-Restricted_Lane	Int	Vị trí của xe có vi phạm các làn đường bị cấm đi hay không.
6	Junction_Location	Int	Vị trí xe trên giao lộ. Mỗi giá trị từ 0 đến 8 trong bảng Codebook thể hiện vị trí của xe. -1 là dữ liệu không có hoặc vượt quá khoảng giá trị trên.
7	Skidding_and_Overturning	Int	Phương tiện có bị mất lái trượt hay lật hay ko. 1: trượt đi, 2: trượt và lật , 4: bị rơi một phần của xe và lật.
8	Hit_Object_in_Carriageway	Int	Xe đã tông vào vật thể gì trên tuyến đường đi. Các giá trị 1-12 thể hiện các vật thể mà xe tông phải. Ngoại trừ -1: dữ liệu null hoặc ngoại khoảng giá trị. 0: không.

9	Vehicle_Leaving_Carriageway	Int	Xe văng khỏi đường đi như thế nào. Các giá trị được ghi trong code book. Ví dụ 1:Nearside(về phía bên trái, gần với giải phân cách), 7: Offside(về phía bên phải, gần với vạch kẻ đường ở giữa). Lưu ý: UK đi xe bên tay trái.
10	Hit_Object_off_Carriageway	Int	Xe đã tông vào vật thể gì bên ngoài tuyến đường. Các giá trị biểu diễn bằng số. Mỗi số tương ứng với một vật thể.
11	1st_Point_of_Impact	Int	Điểm đầu tiên của phương tiện bị ảnh hưởng. 0: did not impact, 1:Front, 2:Back, ...
12	Was_Vehicle_Left_Hand_Drive?	Int	Tài xế đã bị mất lái. 1:No, 2:Yes.
13	Journey_Purpose_of_Driver	Int	Mục đích di chuyển của tài xế. Các giá trị biểu diễn bằng số tương đương với. 1: Journey as part of work, 2: Commuting to/from work,...
14	Sex_of_Driver	Int	Giới tính của tài xế. Giới tính được biểu diễn bằng số tương đương với các giá trị sau: 1: Male, 2:Female, 3: not known, -1: Data missing or out of range.
15	Age_of_Driver	Int	Tuổi hiện tại của tài xế.
16	Age_Band_of_Driver	Int	Thang đo độ tuổi của tài xế. Giá trị từ 1-11 thể hiện thang đo độ của tuổi của tài xế trong bảng codebook. Ví dụ tuổi 56: 9.
17	Engine_Capacity_(CC)	Int	Dung tích của động cơ. Đơn vị đo là CC
18	Propulsion_Code	Int	Động cơ sẽ chạy bằng loại nhiên liệu nào. Ví dụ 1:Petrol, 2: Heavy oil, M:undefined.
19	Age_of_Vehicle	Int	Tuổi hiện tại của phương tiện.
20	Driver_IMD_Decile	Int	Thể hiện mức độ khó khăn thiếu thốn của khu vực mà tài xế sinh sống. Các giá trị được biểu diễn từ -1-10 thể hiện ở các mức độ. 1: nhiều nhất là 10%, 9: ít nhất 10-20%
21	Driver_Home_Area_Type	Int	Loại khu vực mà tài xế sinh sống. 1: Urban Area, 2: Small Town, 3:Rural.

2.3 Casualties

STT	Thuộc tính	Kiểu dữ liệu	Mô tả

1	Accident_Index	String	Số thứ tự chỉ mục các vụ tai nạn
2	Vehicle_Reference	Int	Thứ tự của phương tiện của nạn nhân trong vụ tai nạn
3	Casualty_Reference	Int	Số nạn nhân liên quan đến vụ tai nạn.
4	Casualty_Class	Int	Nạn nhân thuộc nhóm người nào. 1: Driver or rider, 2: passenger, 3: Pedestrian.
5	Sex_of_Casualty	Int	Giới tính của nạn nhân. 1:Male, 2:Female, -1: Data missing or out of range
6	Age_of_Casualty	Int	Tuổi của nạn nhân.
7	Age_Band_of_Casualty	Int	Thang đo độ tuổi của nạn nhân. Giá trị từ 1-11 thể hiện thang đo độ của nhóm tuổi của nạn nhân trong bảng codebook. Ví dụ tuổi 56: 9.
8	Casualty_Severity	Int	Mức độ nghiêm trọng của xảy ra với nạn nhân. Các giá trị biểu diễn kiểu số. 1: Fatal ,2:Serious, 3: Slight
9	Pedestrian_Location	Int	Vị trí của đi bộ của nạn nhân. Ví dụ 0: not a Pedestrian, 1: Crossing on pedestrian crossing facility, 2: Crossing in zig-zag approach lines,....
10	Pedestrian_Movement	Int	Nạn nhân đi bộ di chuyển như thế nào. Ví dụ 0: not a Pedestrian,2: Crossing from nearside - masked by parked or stationary vehicle, 1: Crossing from driver's nearside.....
11	Car_Passenger	Int	Nếu nạn nhân đi car. Thì nạn nhân ngồi ở vị trí nào trên xe. Ví dụ các giá trị 0: not car passenger, 1: Front seat passenger, 2: Rear seat passenger, -1: Data missing or out of range.
12	Bus_or_Coach_Passenger	Int	Có phải nạn nhân khi đi xe bus hoặc xe khách không. Ví dụ 0: not a bus or coach passenger(nạn nhân là người đi bộ), 1: Boarding, 2: Alighting(xuống xe),....
13	Pedestrian_Road_Maintenance_Worker	Int	Có phải nạn nhân là công nhân bảo trì đường cho người đi bộ. Mỗi giá

			trị số được biểu diễn có ý nghĩa riêng. 0: no/ not applicable, 1: Yes, 2: not known, -1:Data missing or out of range.
14	Casualty_Type	Int	Loại phương tiện mà nạn nhân sử dụng. Mỗi giá trị số được biểu diễn có ý nghĩa riêng. 0: Pedestrian, 1: Cyclist, 2: Motorcycle 50cc and under rider or passenger,....
15	Casualty_Home_Area_Type	Int	Loại khu vực mà nạn nhân sinh sống. 1: Urban Area, 2: Small Town, 3:Rural.

2.4 Postcodes

STT	Thuộc tính	Kiểu dữ liệu	Mô tả
1	postcode	string	Mã bưu điện
2	easting	int	Hướng bắc - khoảng cách về phía bắc của vĩ độ. (đơn vị: mét)
3	northing	int	Hướng đông - Khoảng cách về phía đông của kinh độ. (đơn vị: mét)
4	latitude	float	Vĩ độ (đơn vị: độ)
5	longitude	float	Kinh độ (đơn vị: độ)
6	city	string	Thành phố
7	county	string	Hạt - đơn vị hành chính cao nhất ở Anh
8	country_code	string	Mã quốc gia - 3 chữ cái
9	country_name	string	Tên quốc gia
10	iso3166-2	string	Mã quốc gia ISO
11	region_code	string	Mã vùng
12	region_name	string	Tên vùng

2.5 PCD_OA_LSOA_MSOA_LAD_AUG21_UK_LU

STT	Thuộc tính	Kiểu dữ liệu	Mô tả
1	pcd7	string	Mã bưu điện 7 ký tự
2	pcd8	string	Mã bưu điện 8 ký tự
3	pcds	string	Mã bưu điện có số ký tự tùy chỉnh
4	dointr	int	Ngày bắt đầu
5	Doterm	int	Ngày kết thúc
6	usertype	bool	Loại người dùng 0 = small user; 1 = large user
7	oa11cd	string	Vùng đầu ra 2011
8	lsoa11cd	string	Vùng đầu ra lớp dưới 2011
9	msoa11cd	string	Vùng đầu ra Lớp giữa 2011
10	ladcd	string	Mã chính quyền đại phương
11	lsoa11nm	string	Địa chỉ Vùng đầu ra lớp dưới 2011
12	msoa11nm	string	Địa chỉ Vùng đầu ra Lớp giữa 2011
13	ladnm		Địa chỉ chính quyền đại phương
14	ladnmw		

3. Xác định dữ liệu cho yêu cầu

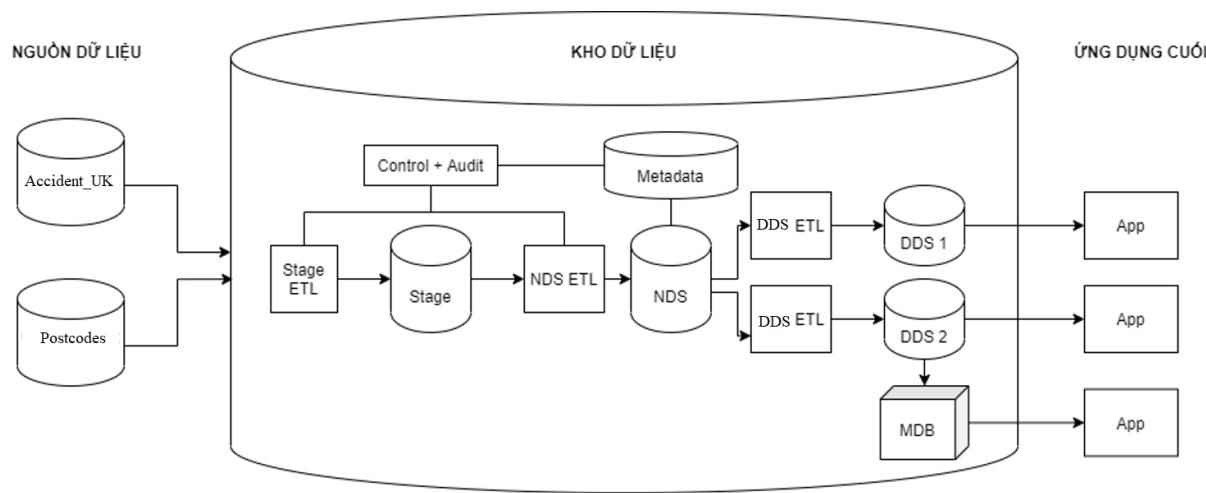
Identifier	Description	Available Data
R01	Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng (Fatal, Serious, Slight) ở các Địa phương (Local_Authority_(District)) trong tất cả các năm.	Dữ liệu bao gồm: <ul style="list-style-type: none"> • Casualty Severity • Local_Authority_(District) • Time

R02	Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa Phương (Local_Authority_(District)) theo các Quý trong từng năm	Đủ dữ liệu bao gồm: <ul style="list-style-type: none">• Casualty Severity• Local_Authority_(District)• Time
R03	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	Đủ dữ liệu bao gồm: <ul style="list-style-type: none">• Casualty Severity• Sex of Casualties• Casualty Type• Age_Band_of_Casualty• Time
R04	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.	Đủ dữ liệu bao gồm: <ul style="list-style-type: none">• Time• Accident Severity• Time
R05	Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng, Vùng (Urban_or_Rural_Area), và Kiểu Đường (Road Type) trong các năm.	Đủ dữ liệu bao gồm: <ul style="list-style-type: none">• Accident Severity• Urban or Rural Area• Road Type• Time

R06	<p>Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng, Loại Nạn Nhân (Casualty Type) và Độ Tuổi trong các năm.</p> <p>Độ tuổi được định nghĩa như sau:</p> <p>Children: 0-15</p> <p>Young adult: 0-17</p> <p>Adult: 18-59 60</p> <p>and over: 60-...</p>	<p>Đủ dữ liệu bao gồm:</p> <ul style="list-style-type: none">• Casualty Severity• Casualty Type• Age_Band_of_Casualty• Time
R07	<p>Tổng hợp số lượng tai nạn theo Mục Đích Hành Trình (Journey Purpose) và Loại Phương Tiện (Vehicle_Type).</p>	<p>Đủ dữ liệu bao gồm:</p> <ul style="list-style-type: none">• Journey_Purpose_of_Driver• Vehicle_Type
R08	<p>Tạo thêm thuộc tính Built-up Road trong table Accidents.</p> <p>Built-up Road có 2 giá trị:</p> <p>Built-up road: Nếu tốc độ giới hạn (Speed Limit) dưới 50 mph</p> <p>Non Built-up road: Nếu tốc độ giới hạn từ 50 mph</p>	<p>ALTER TABLE Accidents ADD Built_up_Road varchar(255);</p> <p>UPDATE Accidents SET Built_up_Road = “Built-up road” WHERE Speed Limit < 50;</p> <p>UPDATE Accidents SET Built_up_Road = “Non Built-up road” WHERE Speed Limit >= 50;</p>
R09	<p>Thống kê số lượng tai nạn theo Mức Độ Nghiêm Trọng, Loại Phương Tiện</p>	<ul style="list-style-type: none">• Accident Severity• Vehicle Type• Built-up Road

	(Vehicle Type), Built-up Road trong các năm.	
--	--	--

4. Data Flow Architect: NDS + DDS



Gồm 3 data store: stage, NDS, DDS

- NDS là cơ sở dữ liệu chuẩn hóa chứa dữ liệu tổng hợp, nhận dữ liệu từ stage qua quá trình ETL. Việc thiết kế càng chi tiết, đạt chuẩn càng cao → phục vụ nhiều yêu cầu phân tích (đạt 3NF hoặc cao hơn)
- MDB csdl đa chiều (multidimensional database).

5. Xác định dimension, fact (measure) phù hợp từng yêu cầu phân tích

5.1 Phân tích

R01. Thông kê số lượng nạn nhân theo Mức Độ Nghiêm Trong (Fatal, Serious, Slight) ở các Địa phương (Local_Authority_(District)) trong tất cả các năm.

a. Phân tích yêu cầu:

Sự kiện: Khi có tai nạn xảy ra.

Bối cảnh:

- Ở đâu: Địa phương xảy ra tai nạn
- Cái gì: Mức độ nghiêm trọng của nạn nhân

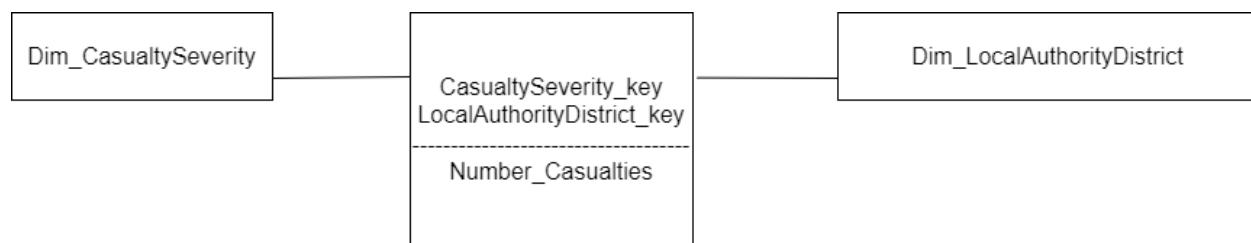
Đo lường (dữ kiện): số lượng nạn nhân

b. Mô hình hóa:

Đo lường → Fact table

Bối cảnh → Dimension Table

Dùng mô hình sao liên kết Dimension Table và Fact Table



c. Fact table:

Các giá trị phải tính toán: NumberCasualties = tổng số nạn nhân theo mức độ nghiêm trọng ở địa phương

Cấp chi tiết dữ liệu (độ mịn):

- Đơn vị nhỏ nhất xảy ra sự kiện: số lượng nạn nhân theo 1 mức độ nghiêm trọng ở 1 địa phương

d. Thiết kế chiều:

Các chiều liên quan sự kiện phân tích:

- Dim_CasualtySeverity
- Dim_LocalAuthorityDistrict

Phân tích lưu giá trị của chiều:

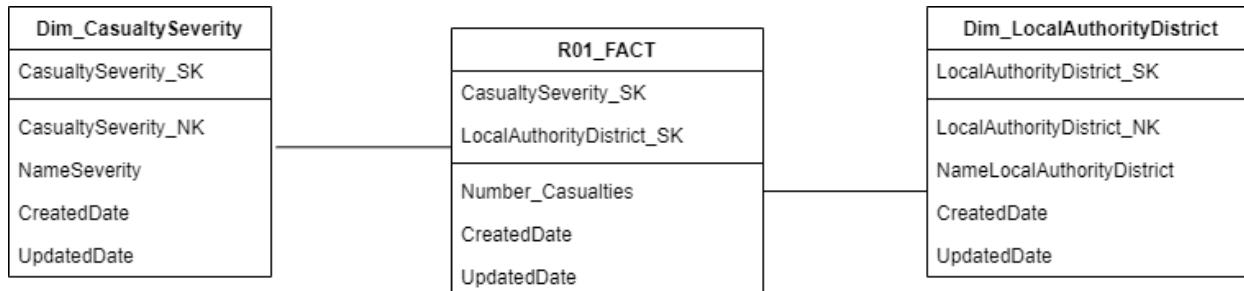
- Chiều thay đổi chậm SDC1

e. Phân cấp dữ liệu:

Dim_CasualtySeverity: Không phân cấp dữ liệu

Dim_LocalAuthorityDistrict: Không phân cấp dữ liệu

f. Kết quả:



R02. Thông kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa Phương (Local_Authority_(District)) theo các Quý trong từng năm.

Phân tích yêu cầu:

Sự kiện: Khi có tai nạn xảy ra.

Bối cảnh:

- Ở đâu: Địa phương xảy ra tai nạn.
- Cái gì: Mức độ nghiêm trọng của nạn nhân
- Khi nào: thời điểm xảy ra vụ tai nạn.

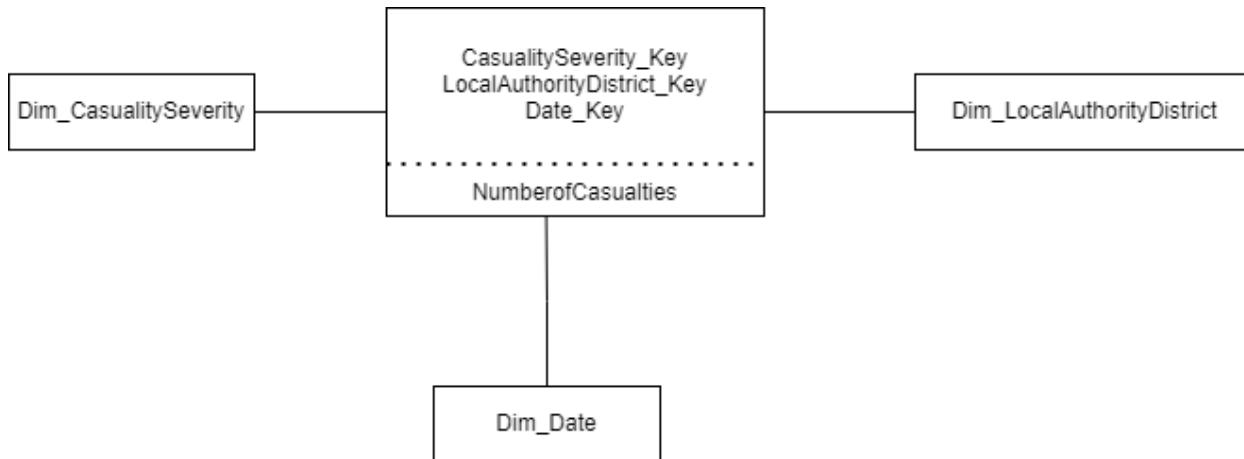
Đo lường (dữ kiện): số lượng nạn nhân.

a. Mô hình hóa:

Đo lường → Fact table

Bối cảnh → Dimension Table

Dùng mô hình sao liên kết Dimension Table và Fact Table



b. Fact table:

Các giá trị phải tính toán: NumberCasualties= tổng số nạn nhân theo mức độ nghiêm trọng ở địa phương trong 1 quý trong từng năm.

Cấp chi tiết dữ liệu (độ mịn):

- Đơn vị nhỏ nhất xảy ra sự kiện: số lượng nạn nhân theo 1 mức độ nghiêm trọng ở 1 địa phương trong 1 quý trong 1 năm.

c. Thiết kế chiều:

Các chiều liên quan sự kiện phân tích:

- Dim_CasualtySeverity
- Dim_LocalAuthorityDistrict.
- Dim_Datetime

Phân tích lưu giá trị của chiều:

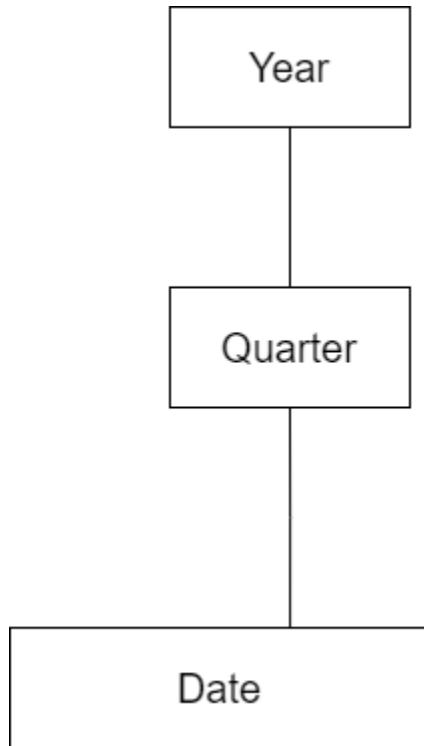
- Chiều thay đổi chậm SDC1.

d. Phân cấp dữ liệu:

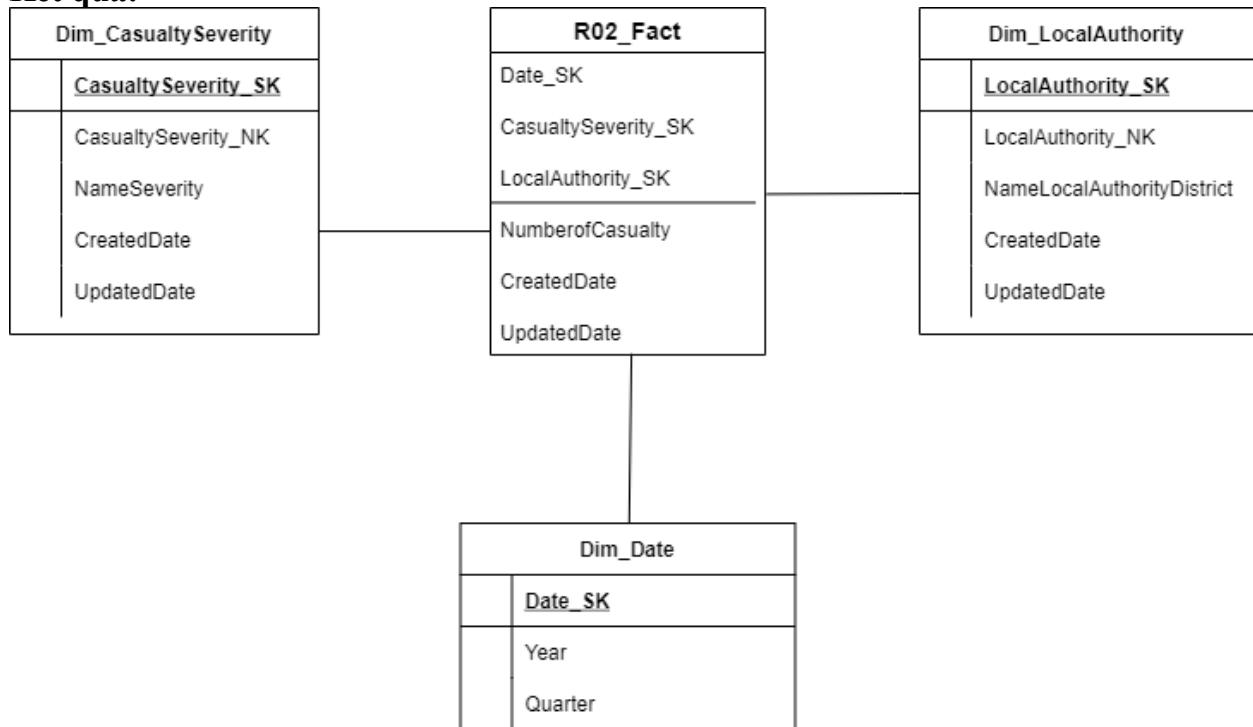
Dim_CasualtySeverity: Không phân cấp dữ liệu

Dim_LocalAuthorityDistrict: Không phân cấp dữ liệu

Dim_Datetime : có phân cấp dữ liệu.



e. Kết quả:



R03. 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

Phân tích yêu cầu:

Sự kiện: Khi có tai nạn xảy ra.

Bối cảnh:

- Cái gì: Giới tính, Nhóm tuổi, Loại nạn nhân, Mức độ nghiêm trọng
- Khi nào: thời điểm năm xảy ra vụ tai nạn.

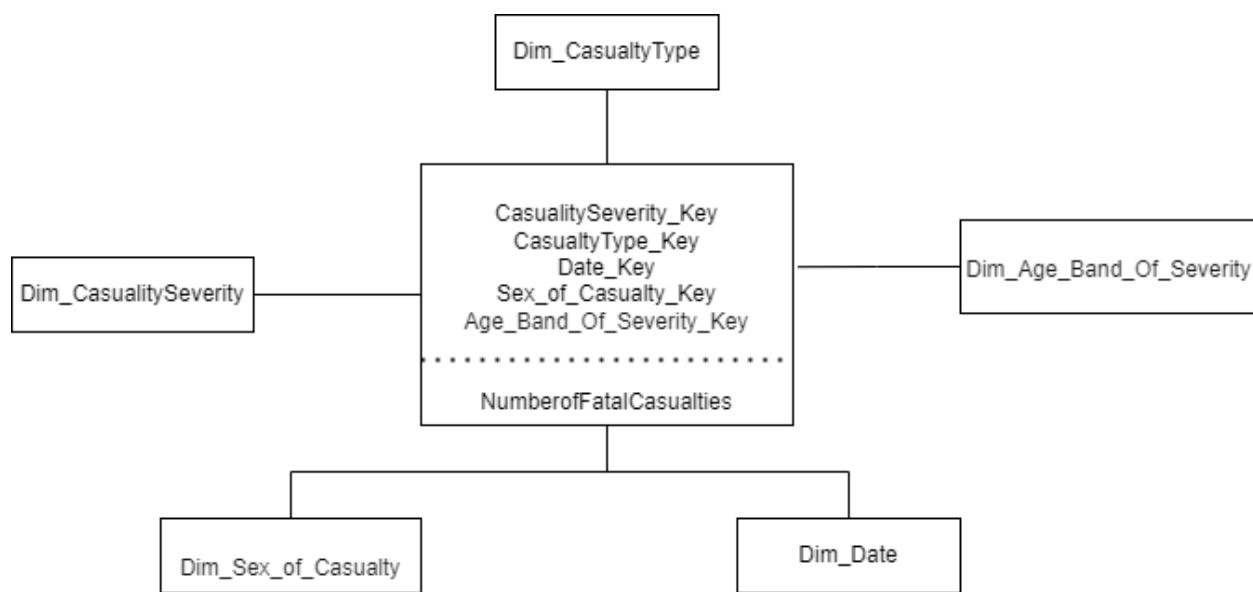
Đo lường (đữ kiện): số lượng người tử vong.

a. **Mô hình hóa:**

Đo lường → Fact table

Bối cảnh → Dimension Table

Dùng mô hình sao liên kết Dimension Table và Fact Table



b. Fact table:

Các giá trị phải tính toán: NumberofFatalCasualties= tổng số nạn nhân tử vong theo Giới Tính, Loại Nạn Nhân (Casualty Type) và Nhóm Tuổi (Age_Band_of_Casualty) theo từng năm

Phân cấp chi tiết dữ liệu (độ mịn):

Đơn vị nhỏ nhất xảy ra sự kiện: số lượng nạn nhân tử vong theo theo từng **Giới Tính**, từng **Loại Nạn Nhân** (Casualty Type) và từng **Nhóm Tuổi** (Age_Band_of_Casualty) theo từng năm.

c. Thiết kế chiều:

Các chiều liên quan sự kiện phân tích:

- Dim_Date.
- Dim_Casualtytype.
- Dim_Age_Band_of_Casualty.
- Dim_Sex_of_Casualty.
- Dim_CasualtySeverity.

Phân tích lưu giá trị của chiều:

- Chiều thay đổi chậm SDC1.

d. Phân cấp dữ liệu:

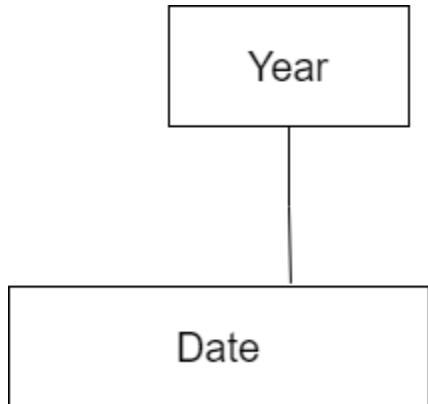
Dim_CasualtySeverity: Không phân cấp dữ liệu

Dim_Age_Band_of_Casualty: Không phân cấp dữ liệu

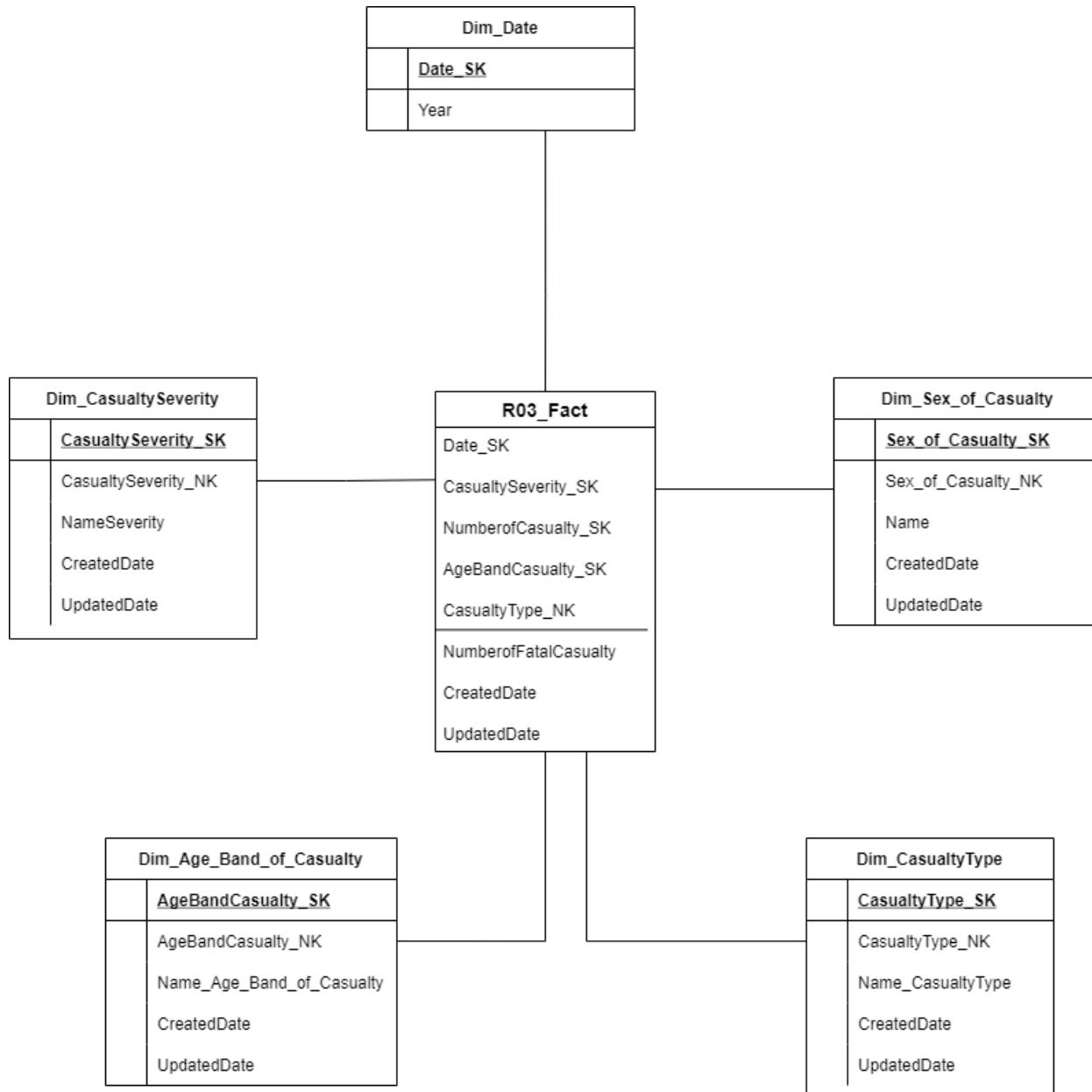
Dim_Casualtytype: Không phân cấp dữ liệu.

Dim_Sex_of_Casualty: không phân cấp dữ liệu.

Dim_Date : có phân cấp dữ liệu.



e. Kết quả:



R04. Thông 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

a. Phân tích yêu cầu:

Sự kiện: Khi có tai nạn xảy ra.

Bối cảnh:

- Cái gì: Mức độ nghiêm trọng của tai nạn
- Khi nào: Thời điểm xảy ra tai nạn

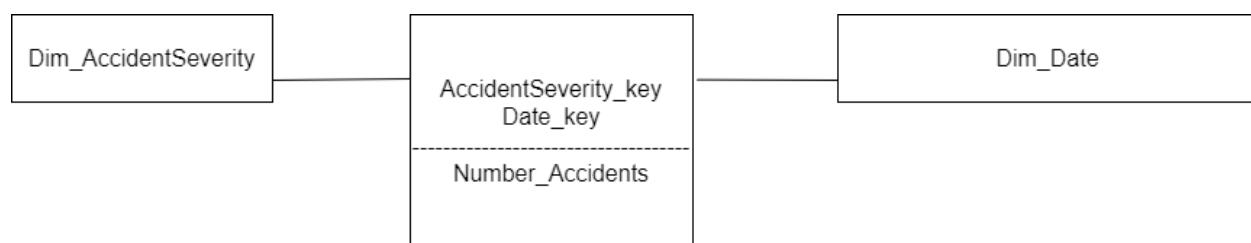
Đo lường (dữ kiện): số lượng tai nạn giao thông

b. Mô hình hóa:

Đo lường → Fact table

Bối cảnh → Dimension Table

Dùng mô hình sao liên kết Dimension Table và Fact Table



c. Fact table:

Các giá trị phải tính toán: NumberAccidents = tổng số tai nạn theo mức độ nghiêm trọng ở từng thời điểm trong ngày.

Cấp chi tiết dữ liệu (độ mịn):

- Đơn vị nhỏ nhất xảy ra sự kiện: số lượng tai nạn theo 1 mức độ nghiêm trọng ở 1 thời điểm trong ngày

d. Thiết kế chiều:

Các chiều liên quan sự kiện phân tích:

- Dim_AccidentSeverity
- Dim_DateTime

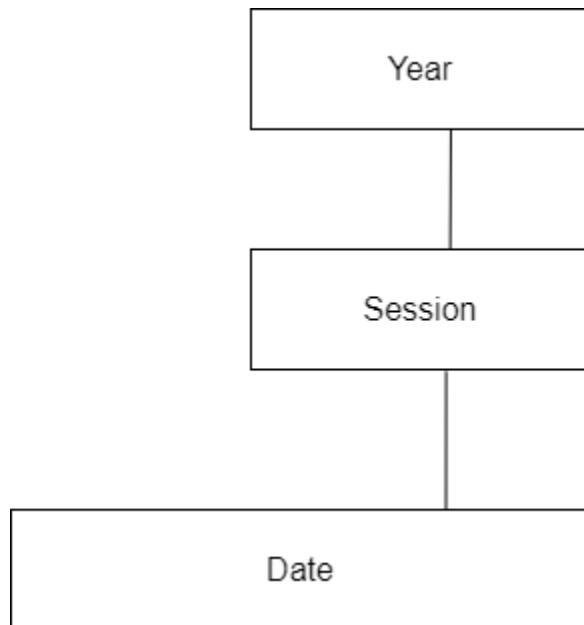
Phân tích lưu giá trị của chiều:

- Chiều thay đổi chậm SDC1

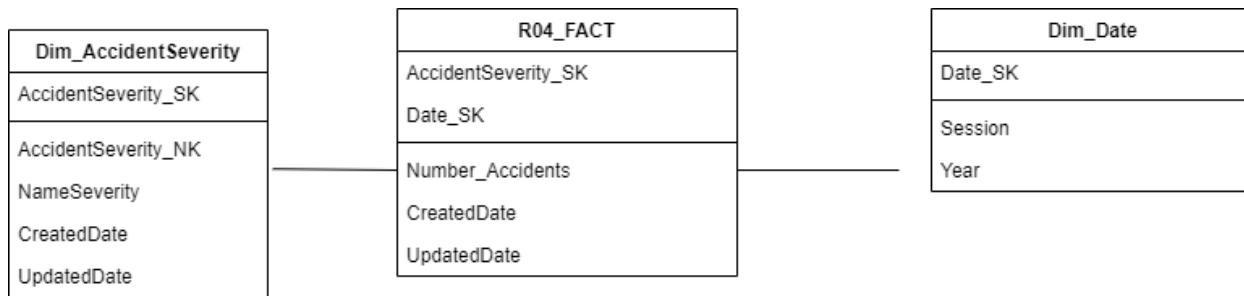
e. Phân cấp dữ liệu:

Dim_AccidentSeverity: Không phân cấp dữ liệu

Dim_DateTime: Có phân cấp dữ liệu



f. Kết quả:



R05. Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng, Vùng (Urban_or_Rural_Area), và Kiểu Đường (Road Type) trong các năm.

a. Phân tích yêu cầu:

Sự kiện: Khi có tai nạn xảy ra.

Bối cảnh:

- Cái gì: Mức độ nghiêm trọng của tai nạn, Kiểu đường xảy ra tai nạn
- Ở đâu: Vùng xảy ra tai nạn

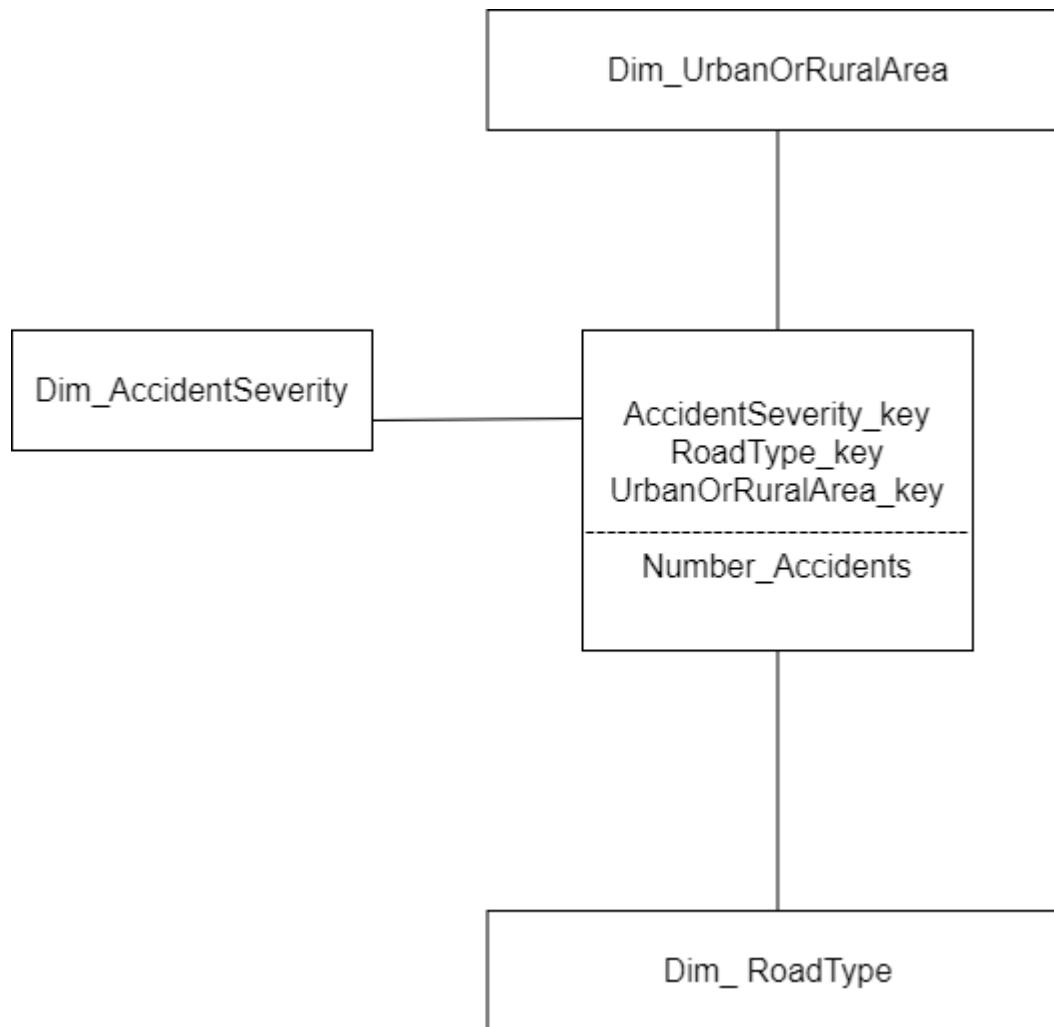
Đo lường (dữ kiện): số lượng tai nạn giao thông

b. Mô hình hóa:

Đo lường → Fact table

Bối cảnh → Dimension Table

Dùng mô hình sao liên kết Dimension Table và Fact Table



c. Fact table:

Các giá trị phải tính toán: NumberAccidents = tổng số tai nạn theo mức độ nghiêm trọng với từng loại kiểu đường ở từng thời điểm trong ngày tại vùng cụ thể.

Cấp chi tiết dữ liệu (độ mịn):

- Đơn vị nhỏ nhất xảy ra sự kiện: số lượng tai nạn theo 1 mức độ nghiêm trọng với 1 loại kiểu đường ở 1 thời điểm trong ngày tại 1 vùng cụ thể.

d. Thiết kế chiều:

Các chiều liên quan sự kiện phân tích:

- Dim_AccidentSeverity
- Dim_UrbanOrRuralArea
- Dim_RoadType

Phân tích lưu giá trị của chiều:

- Chiều thay đổi chậm SDC1

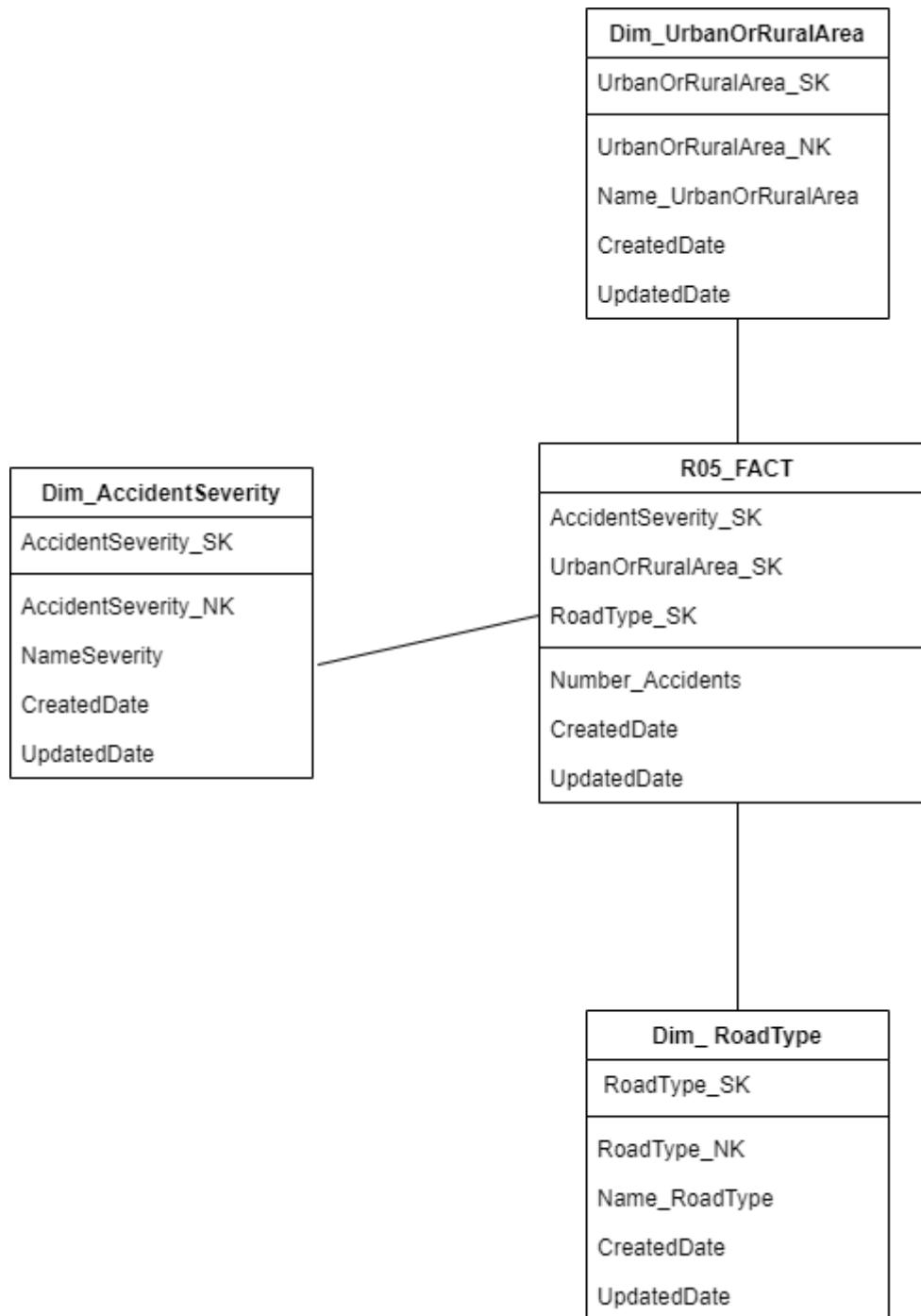
e. Phân cấp dữ liệu:

Dim_AccidentSeverity: Không phân cấp dữ liệu

Dim_UrbanOrRuralArea: Không phân cấp dữ liệu

Dim_RoadType: Không phân cấp dữ liệu

f. Kết quả:



R06. Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng, Loại Nạn Nhân (Casualty Type) và Độ Tuổi trong các năm.

a. Phân tích yêu cầu:

Sự kiện: Khi có tai nạn xảy ra.

Bối cảnh:

- Cái gì: Mức độ nghiêm trọng của nạn nhân, loại nạn nhân, độ tuổi

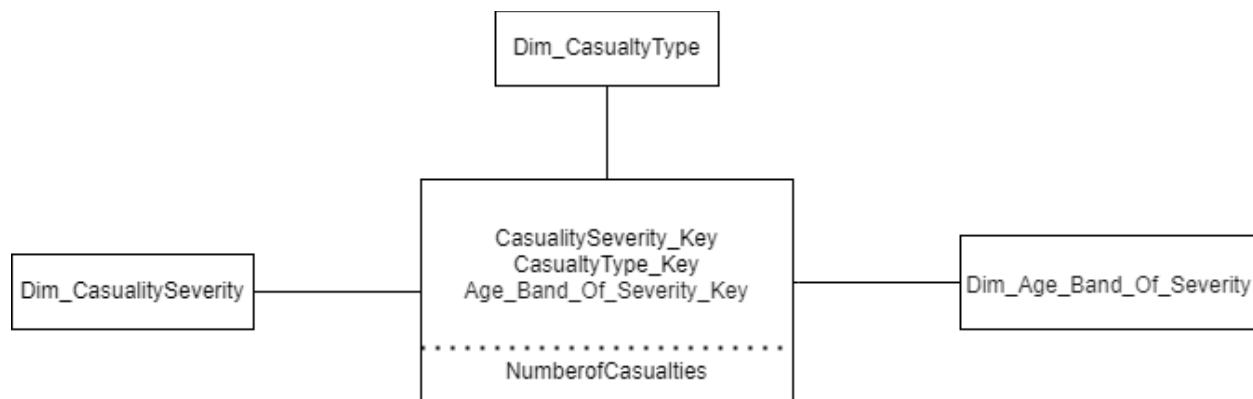
Đo lường (dữ kiện): số lượng nạn nhân

b. Mô hình hóa:

Đo lường → Fact table

Bối cảnh → Dimension Table

Dùng mô hình sao liên kết Dimension Table và Fact Table



c. Fact table:

Các giá trị phải tính toán: NumberCasualties = tổng số nạn nhân

Cấp chi tiết dữ liệu (độ mịn):

- Đơn vị nhỏ nhất xảy ra sự kiện: số lượng nạn nhân theo 1 mức độ nghiêm trọng của 1 độ tuổi thuộc loại nạn nhân nào

d. Thiết kế chiều:

Các chiều liên quan sự kiện phân tích:

- Dim_CasualtySeverity
- Dim_Age_Band_Of_Severity
- Dim_CasualtyType

Phân tích lưu giá trị của chiều:

- Chiều thay đổi chậm SDC1

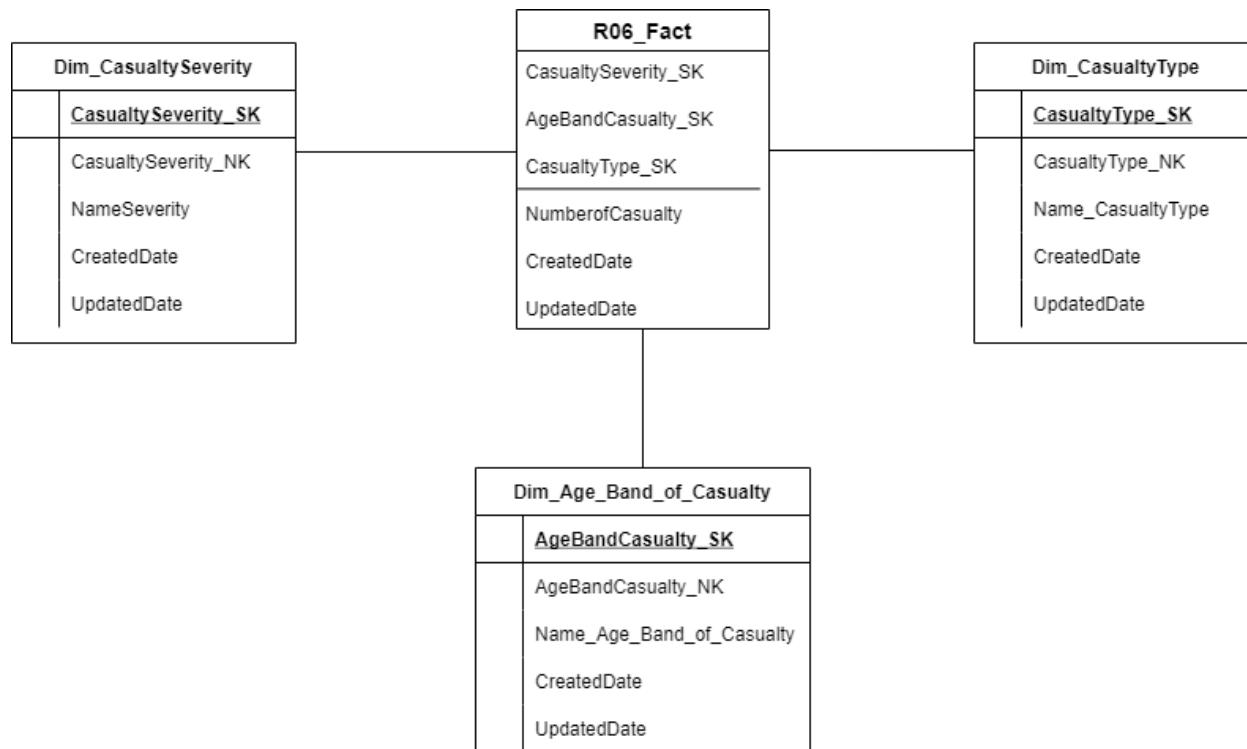
e. Phân cấp dữ liệu:

Dim_CasualtySeverity: Không phân cấp dữ liệu

Dim_Age_Band_Of_Severity: Không phân cấp dữ liệu

Dim_CasualtyType: Không phân cấp dữ liệu

f. Kết quả:



R07. Tổng hợp số lượng phương tiện theo Mục Đích Hành Trình (Journey Purpose) và Loại Phương Tiện (Vehicle_Type).

Phân tích yêu cầu:

Sự kiện: Khi có tai nạn xảy ra.

Bối cảnh:

- Cái gì: Loại Phương Tiện, Mục Đích Hành Trình

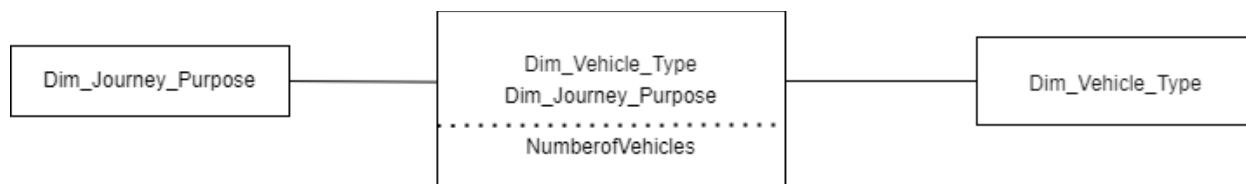
Đo lường (dữ kiện): số lượng phương tiện

a. **Mô hình hóa:**

Đo lường → Fact table

Bối cảnh → Dimension Table

Dùng mô hình sao liên kết Dimension Table và Fact Table



b. **Fact table:**

Các giá trị phải tính toán: NumberVehicles = số phương tiện

Cấp chi tiết dữ liệu (độ mịn):

- Đơn vị nhỏ nhất xảy ra sự kiện: số lượng phương tiện theo 1 mục đích hành trình và loại phương tiện

c. **Thiết kế chiều:**

Các chiều liên quan sự kiện phân tích:

- Dim_Journey_Purpose
- Dim_Vehicle_Type

Phân tích lưu giá trị của chiều:

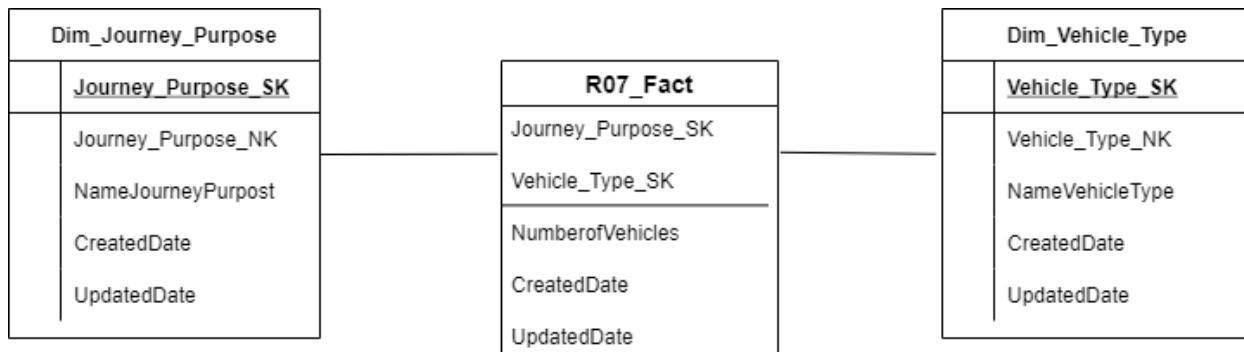
- Chiều thay đổi chậm SDC1

d. **Phân cấp dữ liệu:**

Dim_Journey_Purpose: Không phân cấp dữ liệu

Dim_Dim_Vehicle_Type: Không phân cấp dữ liệu

e. Kết quả:



R09. Thống kê số lượng tai nạn theo Mức Độ Nghiêm Trọng, Loại Phương Tiện(Vehicle Type), Built-up Road trong các năm.

- Accident Severity
- Vehicle Type
- Built-up Road

Note: R8.Tạo thêm thuộc tính Built-up Road trong table Accidents. 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

a. Phân tích yêu cầu:

Sự kiện: Khi có tai nạn xảy ra.

Bối cảnh:

- Cái gì: Mức độ nghiêm trọng của tai nạn, loại phương tiện, Built-up Road

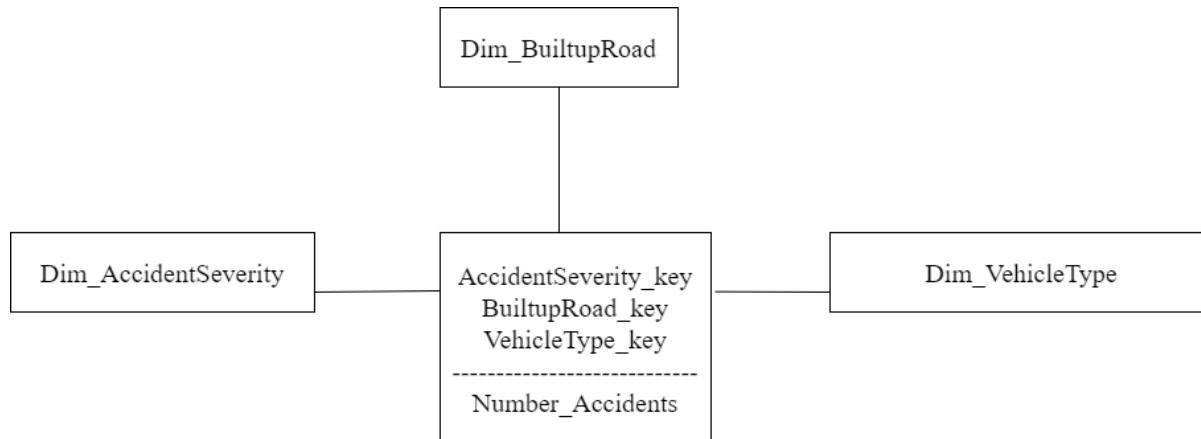
Đo lường (dữ kiện): số lượng tai nạn giao thông

b. Mô hình hóa:

Đo lường → Fact table

Bối cảnh → Dimension Table

Dùng mô hình sao liên kết Dimension Table và Fact Table



c. Fact table:

Các giá trị phải tính toán: NumberAccidents = tổng số tai nạn có cùng mức độ nghiêm trọng (Accident Severity), cùng một loại phương tiện(Vehicle Type) và cùng loại Built-up Road.

Cấp chi tiết dữ liệu (độ mịn):

- Đơn vị nhỏ nhất xảy ra sự kiện: số lượng tai nạn của một loại phương tiện, ở một mức độ nghiêm trọng và một loại Built-up Road.

d. Thiết kế chiều:

Các chiều liên quan sự kiện phân tích:

- Dim_AccidentSeverity
- Dim_BuiltupRoad
- Dim_VehicleType

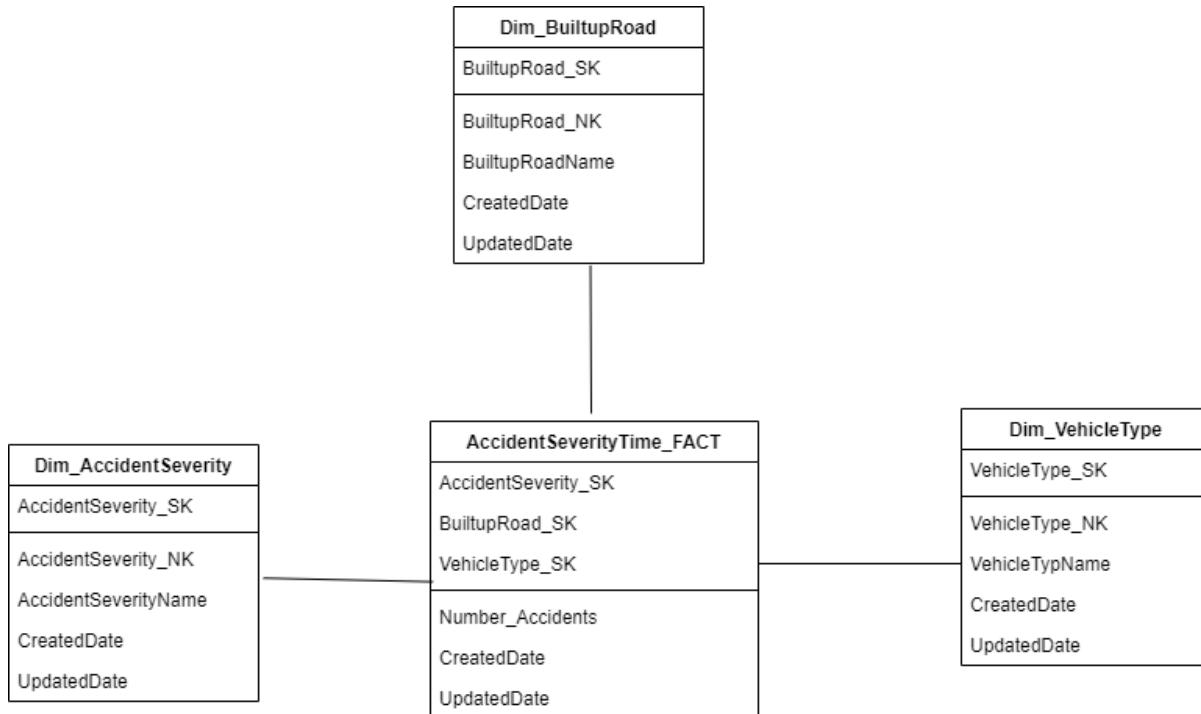
Phân tích lưu giá trị của chiều:

- Chiều thay đổi chậm SDC1

e. Phân cấp dữ liệu:

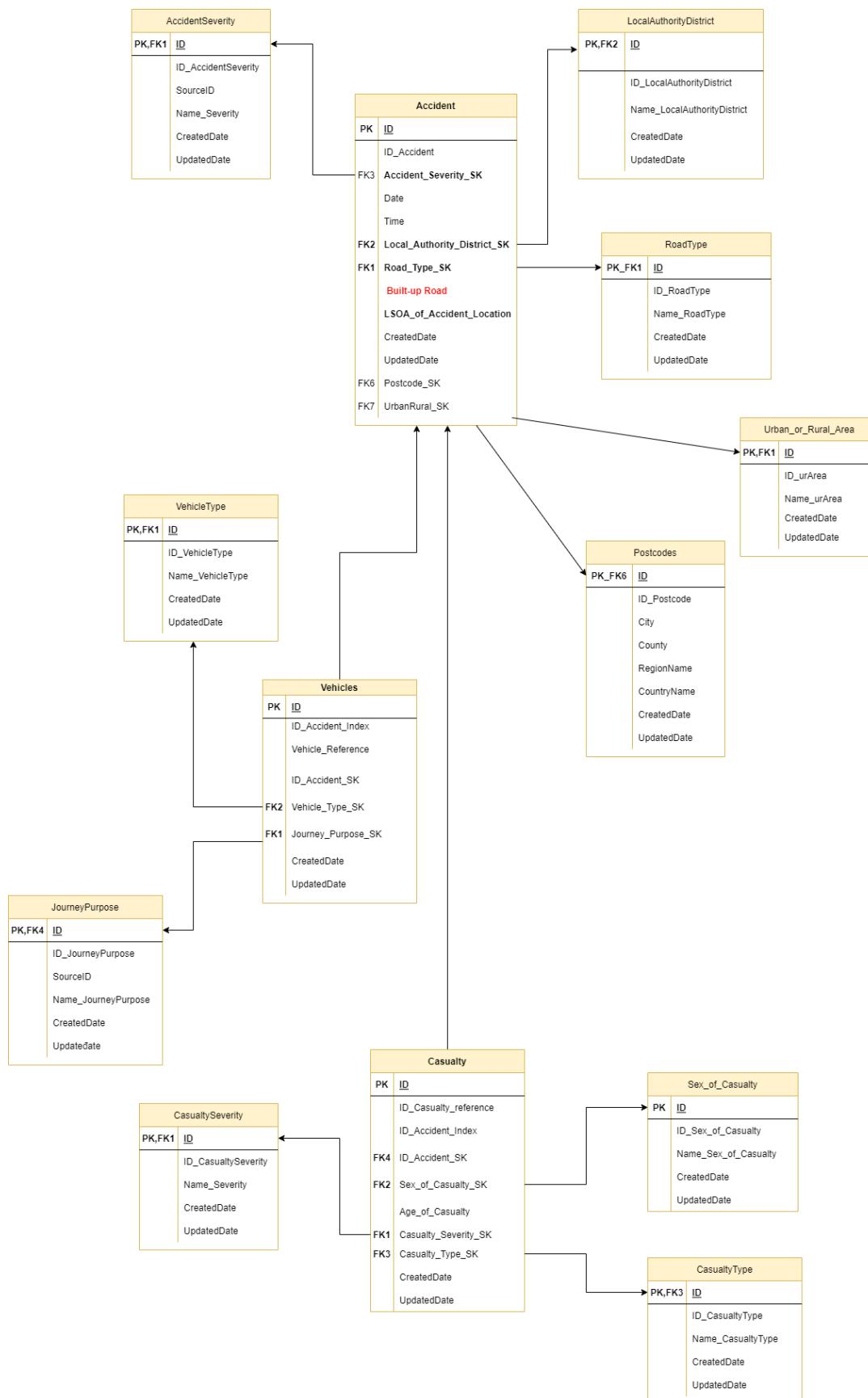
- Dim_AccidentSeverity: Không phân cấp dữ liệu
- Dim_BuiltupRoad: Không phân cấp dữ liệu
- Dim_VehicleType: Không phân cấp dữ liệu

f. Kết quả

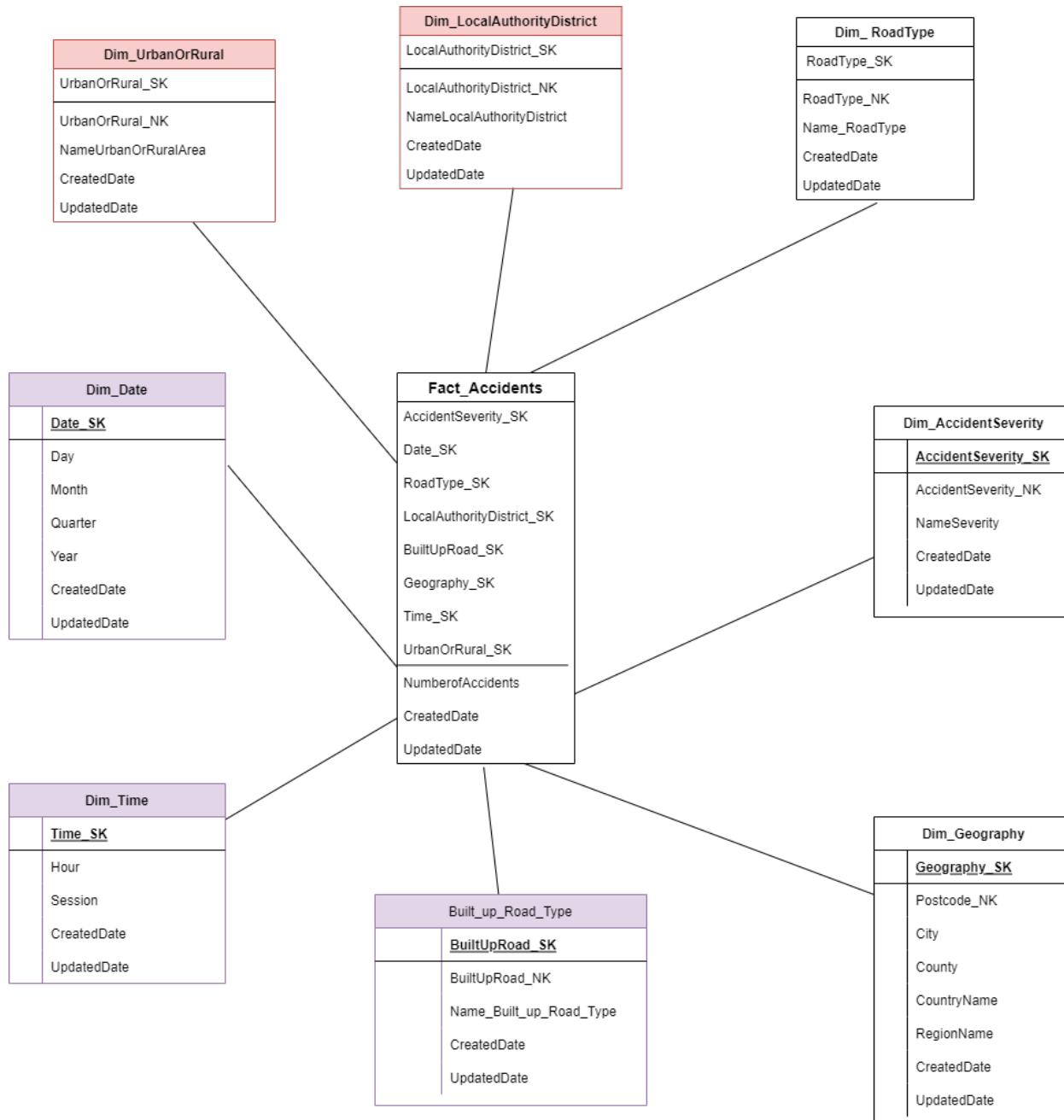


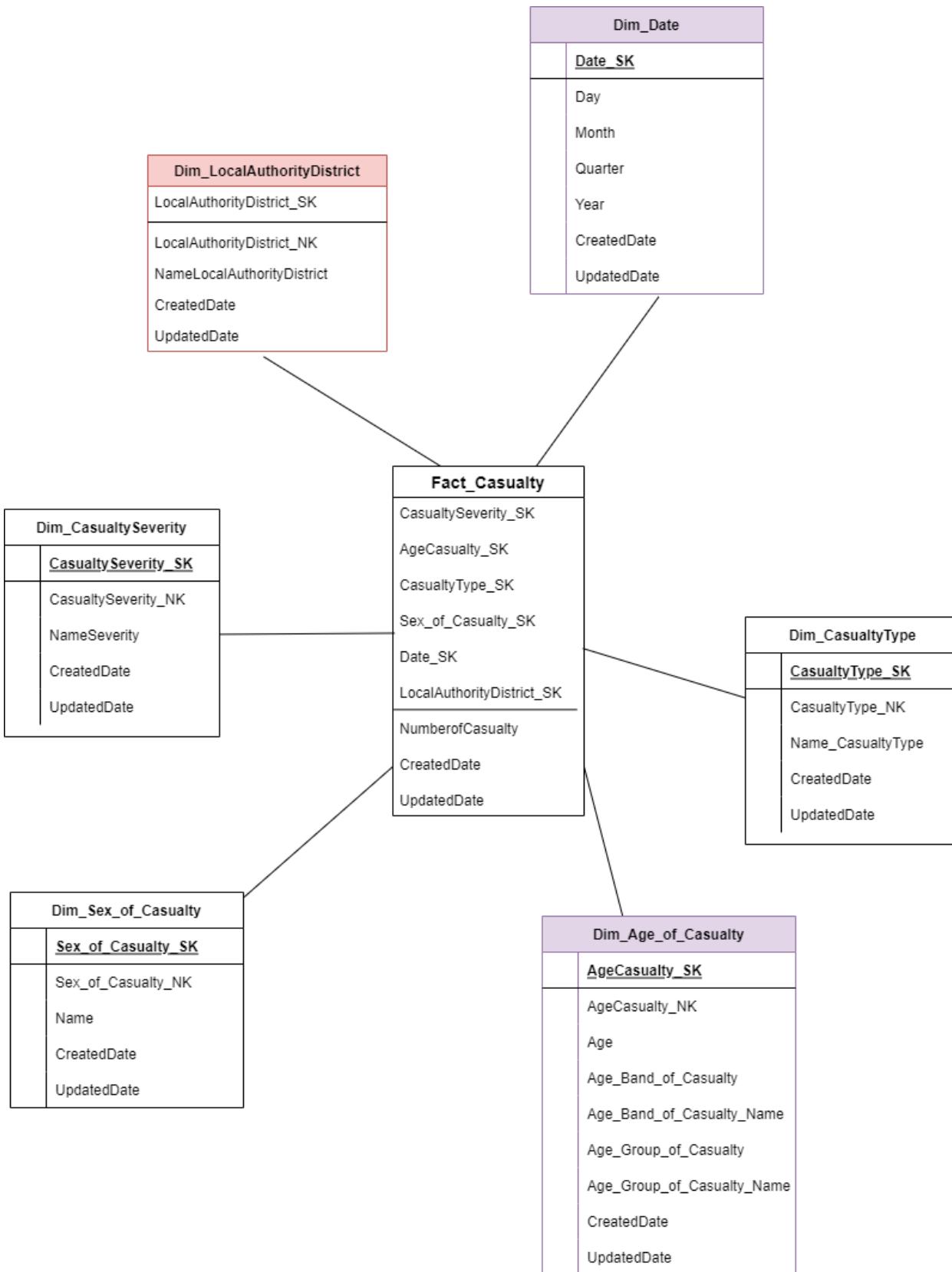
5.2 Kết quả

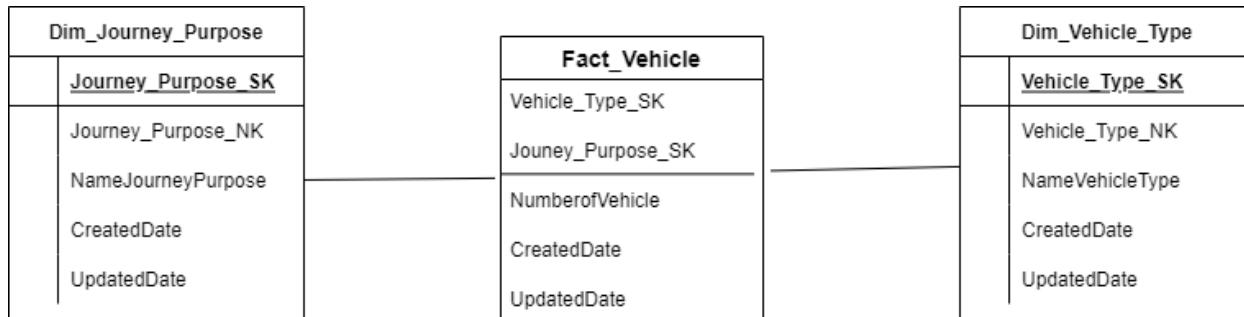
a. Lược đồ NDS



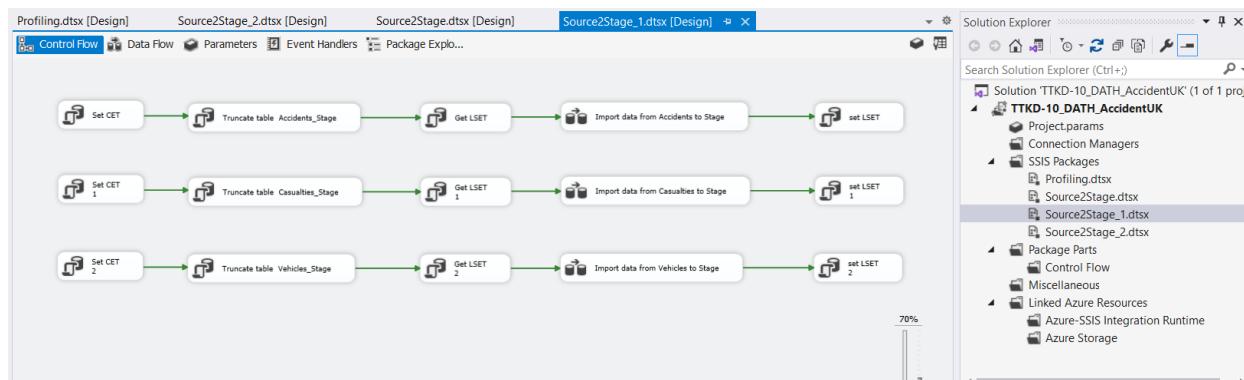
b. Lược đồ DDS



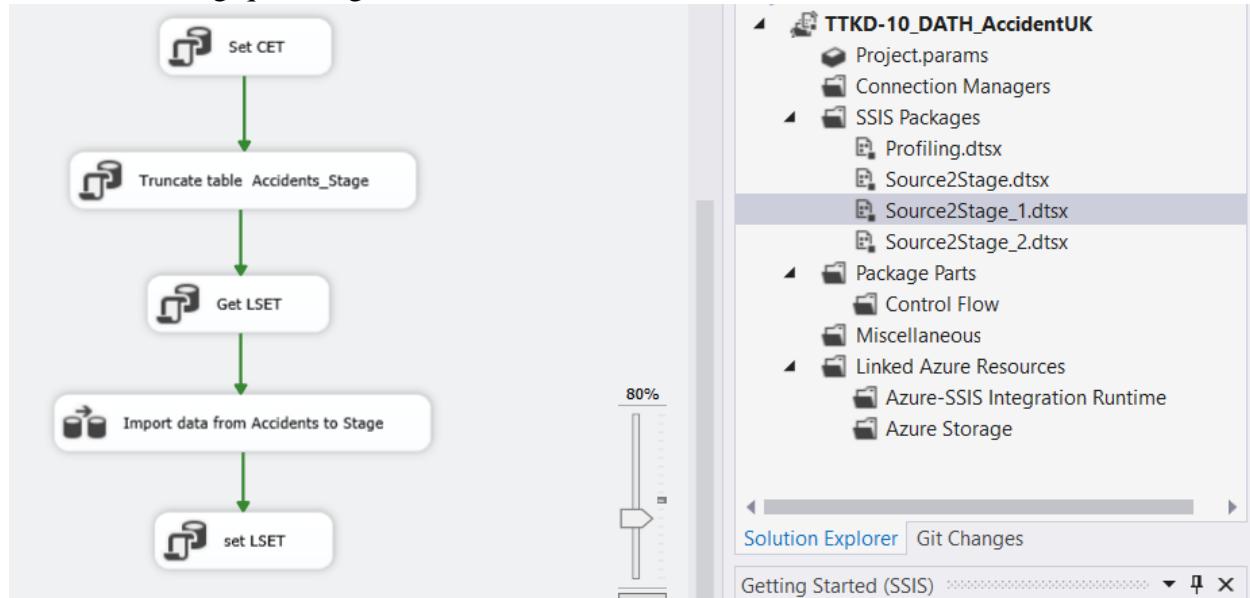




6. Quá trình từ Source vào Stage



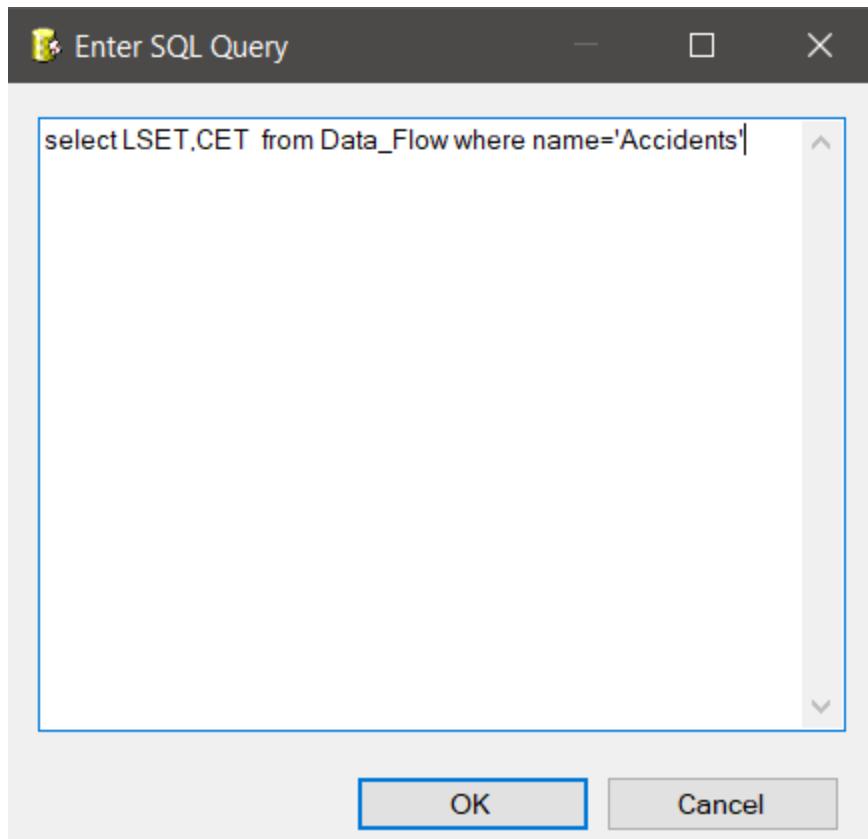
- Minh họa thông qua bảng Accidents



Bước 1: Set CET: lấy thời gian ETL hiện tại

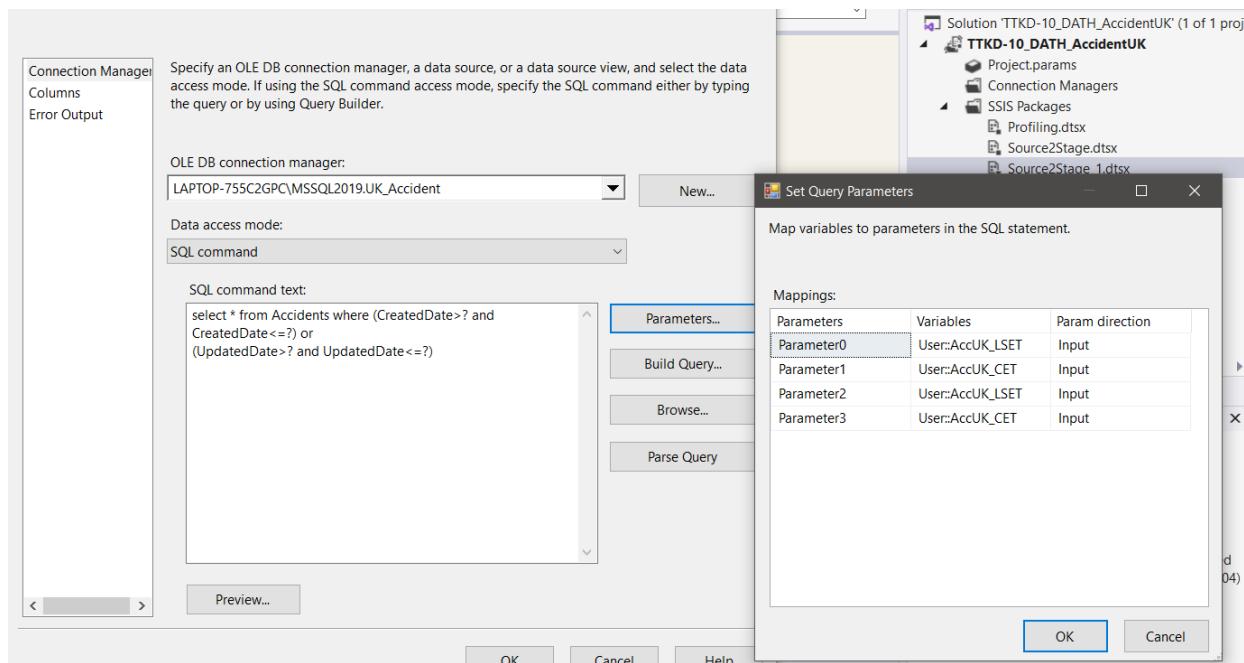
Bước 2: Truncate table Accidents_Stage: Xóa dữ liệu của bảng cần ETL trong stage, minh họa ở đây là bảng Accidents

Bước 3: Get LSET: lấy thời điểm rút trích từ Source vào Stage gần nhất được lưu trong bảng Data_Flow từ database MetaData



Bước 4: Import data from Accidents to Stage: Ta rút trích những dữ liệu mà nằm trong khoảng LSET đến CET để nạp vào stage

Mục đích: chỉ lấy những dòng thay đổi, mới được thêm vào từ lần ETL trước đó hoặc những dòng mới được update từ lần ETL trước đó => tránh trùng lặp dữ liệu.



Bước 5: Cuối cùng gán lại LSET = CET

7. Quá trình từ Stage vào NDS

7.1 Phân tích

a. Accident

Column	Data Type	Key Type	Null Ratio	Description
ID	Int	SK (PK)		Khóa của bảng Accidents (SK)
ID_Accidents	Varchar(13)	NK		Mã tai nạn trong source system (NK)
Accident_Severity_SK	Int	FK		Mức độ nghiêm trọng của vụ tai nạn
Date	Date			Ngày xảy ra tai nạn
Time	Varchar(5)		0,0043% (25 null records)	Thời gian xảy ra tai nạn

Local_Authority_District_SK	Int	FK		Tên khu vực địa phương xảy ra tai nạn
Road_Type_SK	Int	FK		Loại đường xảy ra tai nạn
Name_Built_up_Road	Varchar(50)			Giới hạn tốc độ
LSOA_of_Accident_Location	Varchar(9)		6,6476% (38691 null records)	Mã địa điểm tai nạn
Journey_Purpose_SK	Int	FK		Mục đích chuyến đi
Vehicle_ID_SK	Int	FK		Loại phương tiện
PostCode_SK	Int	FK		Postcodes
CreatedDate	Datetime			Thời gian record được tạo trong NDS
UpdatedDate	Datetime			Thời gian record được cập nhật trong NDS

b. AccidentSeverity

Column	Data Type	Key Type	Null Ratio	Description
ID	Int	SK (PK)		Khóa của bảng AccidentSeverity (SK)
ID_AccidentSeverity	Int	NK		Mã mức độ nghiêm trọng tai nạn trong source system (NK)
NameSeverity	Varchar(7)			Tên mức độ nghiêm trọng
CreatedDate	Datetime			Thời gian record được tạo trong NDS
UpdatedDate	Datetime			Thời gian record được cập nhật trong NDS

c. Casualty

Column	Data Type	Key Type	Null Ratio	Description
ID	Int	SK (PK)		Khóa của bảng Casualty (SK)
ID_Casualty_reference	Int	NK		Mã nạn nhân liên quan đến vụ tai nạn.
ID_Accident_Index	Varchar(50)	NK		Mã tai nạn trong source system
ID_Accident_SK	Int	FK		Khóa ngoại của bảng Accidents
Sex_of_Casualty_SK	Int	FK		Giới tính nạn nhân
Age_of_Casualty	Int			Tuổi nạn nhân
Casualty_Severity_SK	Int	FK		Mức độ nghiêm trọng của nạn nhân
Casualty_Type_SK	Int	FK		Loại nạn nhân
CreatedDate	Datetime			Thời gian record được tạo trong NDS
UpdatedDate	Datetime			Thời gian record được cập nhật trong NDS

d. CasualtySeverity

Column	Data Type	Key Type	Null Ratio	Description
ID	Int	SK (PK)		Khóa của bảng CasualtySeverity (SK)
ID_CasualtySeverity	Int	NK		Mã mức độ nghiêm trọng nạn nhân trong source system (NK)

Name_Severity	Varchar(7)			Tên mức độ nghiêm trọng
CreatedDate	Datetime			Thời gian record được tạo trong NDS
UpdatedDate	Datetime			Thời gian record được cập nhật trong NDS

e. CasualtyType

Column	Data Type	Key Type	Null Ratio	Description
ID	Int	SK (PK)		Khóa của bảng CasualtyType (SK)
ID_CasualtyType	Int	NK		Mã loại nạn nhân trong source system (NK)
Name_CasualtyType	Varchar(57)			Tên loại nạn nhân
CreatedDate	Datetime			Thời gian record được tạo trong NDS
UpdatedDate	Datetime			Thời gian record được cập nhật trong NDS

f. JourneyPurpose

Column	Data Type	Key Type	Null Ratio	Description
ID	Int	SK (PK)		Khóa của bảng JourneyPurpose(SK)
ID_JourneyPurpose	Int	NK		Mã mục đích hành trình trong source system (NK)
Name_JourneyPurpose	Varchar(28)			Tên mục đích hành trình
CreatedDate	Datetime			Thời gian record được tạo trong NDS

UpdatedDate	Datetime			Thời gian record được cập nhật trong NDS
--------------------	----------	--	--	--

g. LocalAuthorityDistrict

Column	Data Type	Key Type	Null Ratio	Description
ID	Int	SK (PK)		Khóa của bảng LocalAuthorityDistrict (SK)
ID_LocalAuthorityDistrict	Int	NK		Mã địa phương (NK)
NameLocalAuthorityDistrict	Varchar(28)			Tên địa phương
CreatedDate	Datetime			Thời gian record được tạo trong NDS
UpdatedDate	Datetime			Thời gian record được cập nhật trong NDS

h. Postcodes

Column	Data Type	Key Type	Null Ratio	Description
ID	Int	SK (PK)		Khóa của bảng Postcodes (SK)
ID_Postcode	Varchar(50)	NK		Mã postcode (NK)
City	Varchar(40)			Tên thành phố
County	Varchar(28)			Tên County
RegionName	Varchar(24)			Tên Vùng
CountryName	Varchar(16)			Tên Nước
CreatedDate	Datetime			Thời gian record được tạo trong NDS

UpdatedDate	Datetime			Thời gian record được cập nhật trong NDS
--------------------	----------	--	--	--

i. RoadType

Column	Data Type	Key Type	Null Ratio	Description
ID	Int	SK (PK)		Khóa của bảng RoadType (SK)
ID_RoadType	Int	NK		Mã loại đường (NK)
Name_RoadType	Varchar(28)			Tên loại đường
CreatedDate	Datetime			Thời gian record được tạo trong NDS
UpdatedDate	Datetime			Thời gian record được cập nhật trong NDS

j. Sex_of_Casualty

Column	Data Type	Key Type	Null Ratio	Description
ID	Int	SK (PK)		Khóa của bảng Sex_of_Casualty (SK)
ID_Sex_of_Casualty	Int	NK		Mã giới tính nạn nhân (NK)
Name_Sex_of_Casualty	Varchar(28)			Tên giới tính nạn nhân
CreatedDate	Datetime			Thời gian record được tạo trong NDS
UpdatedDate	Datetime			Thời gian record được cập nhật trong NDS

k. Urban_or_Rural_Area

Column	Data Type	Key Type	Null Ratio	Description
--------	-----------	----------	------------	-------------

ID	Int	SK (PK)		Khóa của Urban_or_Rural_Area (SK)
ID_UrbanorRuralArea	Int	NK		Mã Area (NK)
Name_urArea	Varchar(11)			Tên Area
CreatedDate	Datetime			Thời gian record được tạo trong NDS
UpdatedDate	Datetime			Thời gian record được cập nhật trong NDS

I. Vehicle

Column	Data Type	Key Type	Null Ratio	Description
ID	Int	SK (PK)		Khóa của bảng Vehicle (SK)
Vehicle_Reference	Int	NK		Mã phương tiện liên quan đến vụ tai nạn
ID_Accident_Index	Varchar(50)	NK		Mã tai nạn trong source system
ID_Accident_SK	Int	FK		Khóa ngoại của bảng Accidents
Journey_Purpose_of_Driver_SK	Int	FK		Mã mục đích hành trình
Vehicle_Type_SK	Int	FK		Mã loại phương tiện
CreatedDate	Datetime			Thời gian record được tạo trong NDS
UpdatedDate	Datetime			Thời gian record được cập nhật trong NDS

m. VehicleType

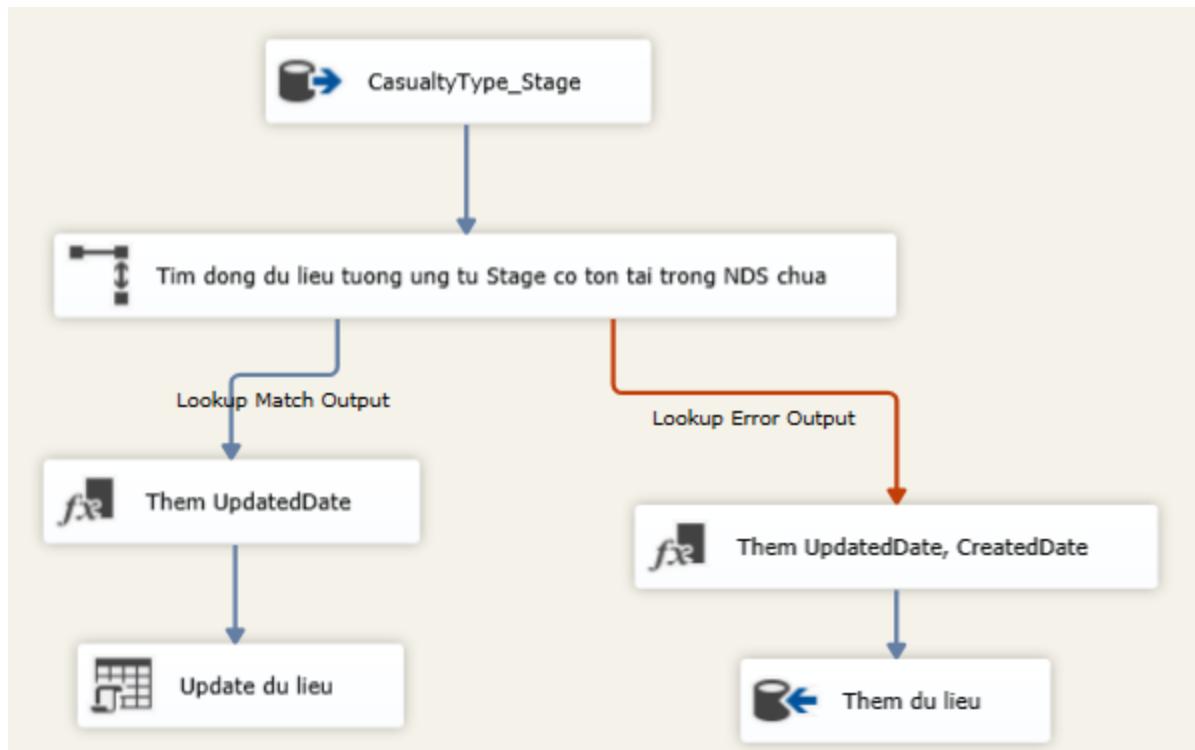
Column	Data Type	Key Type	Null Ratio	Description
ID	Int	SK (PK)		Khóa của VehicleType (SK)
ID_VehicleType	Int	NK		Mã loại phương tiện(NK)
Name_VehicleType	Varchar(37)			Tên loại phương tiện
CreatedDate	Datetime			Thời gian record được tạo trong NDS
UpdatedDate	Datetime			Thời gian record được cập nhật trong NDS

7.2 Thực hành

Các bước thực hiện từ Stage → NDS

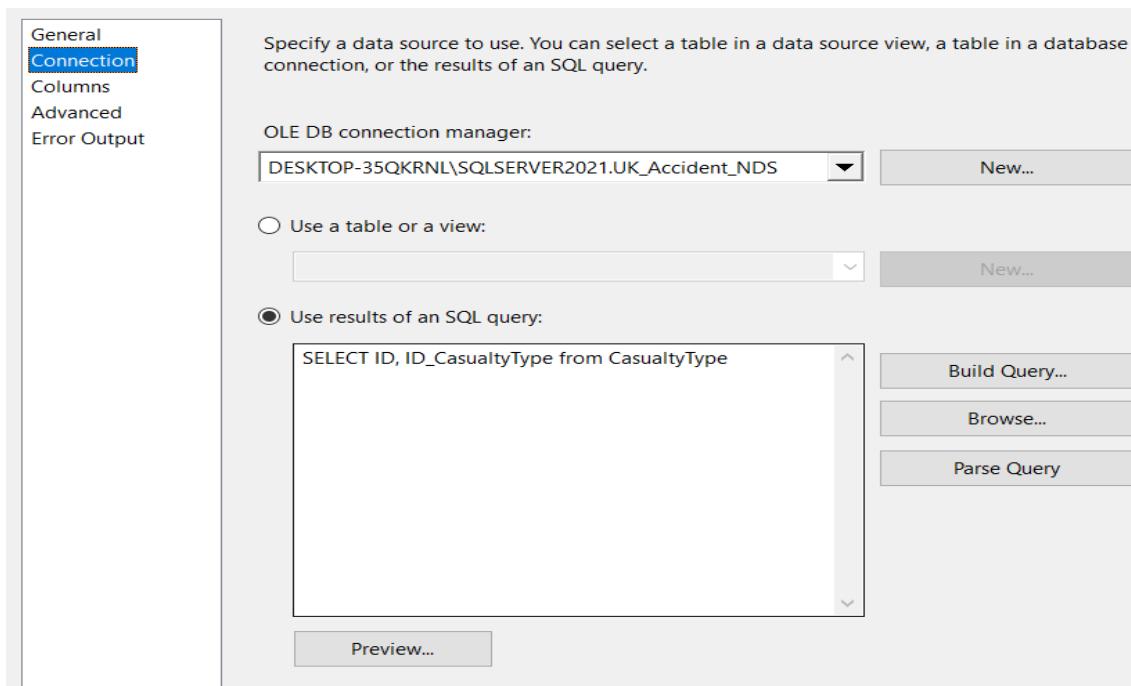
- **Lưu ý:** Chúng ta nên đỗ các bảng không có khóa ngoại trước.
- **Minh họa thông qua 2 ví dụ:**
 - Ví dụ về bảng không có khóa ngoại: **CasualtyType**
 - Ví dụ về bảng có khóa ngoại: **Accident**

Ví dụ 1: CasualtyType

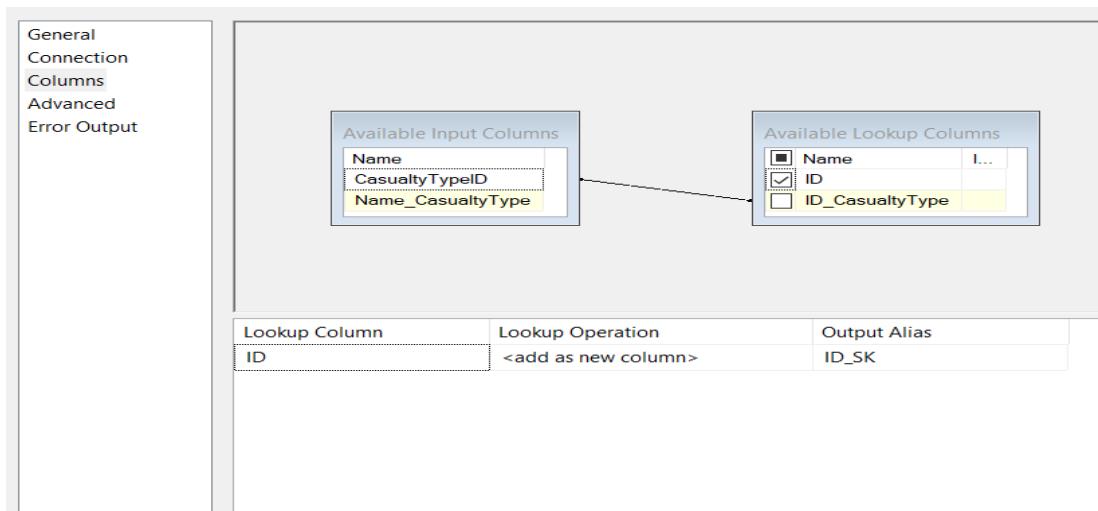


Bước 1: Load dữ liệu từ CasualtyType_stage lên để xử lý

Bước 2: Kiểm tra xem các dòng dữ liệu đã tồn tại trong bảng CasualtyType của NDS hay chưa



Lấy ID, ID_CasualtyType từ CasualtyType trong NDS.

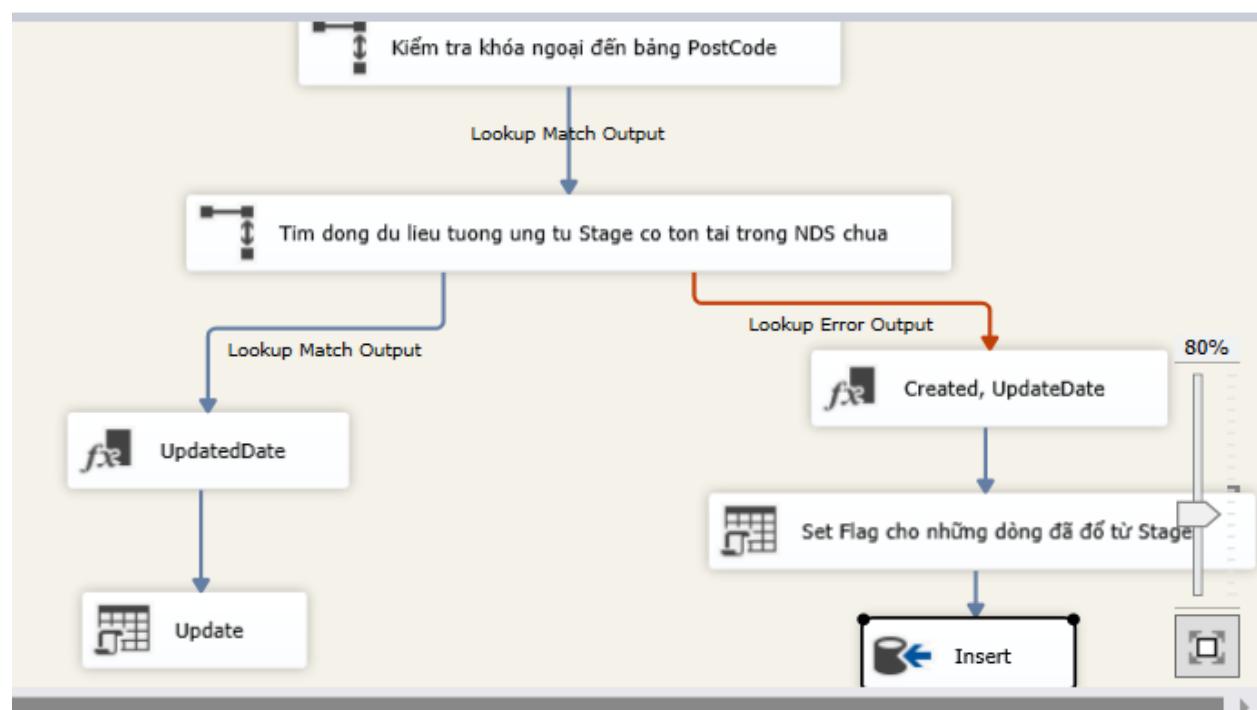
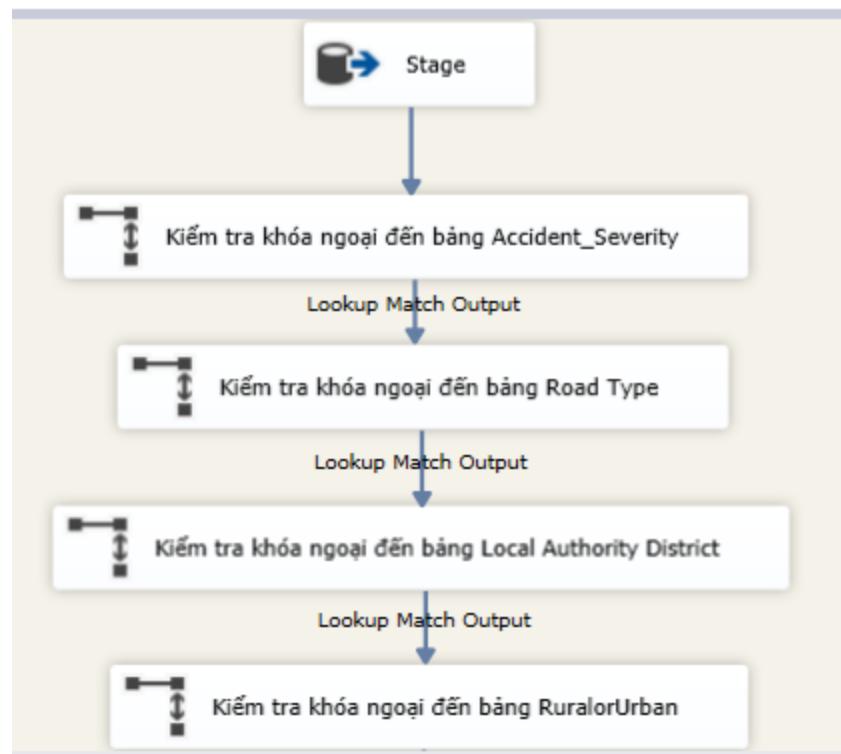


Sau khi lấy được ID và ID_CasualtyType, tiếp theo chúng ta sẽ mapping CasualtyTypeID (NK ở Stage) với ID_CasualtyType (NK ở NDS). Nếu mà mà lookup tìm thấy sẽ trả về ID_SK (SK ở NDS)

Bước 4: Sau khi Lookup xong, có 2 trường hợp xảy ra:

- Nếu đã tồn tại (tìm thấy được giá trị ID_SK) thì tiến hành
 - Thêm cột UpdatedDate là ngày ETL hiện tại, sử dụng hàm getdate()
 - Update dòng dữ liệu trong NDS
- Nếu chưa tồn tại thì tiến hành
 - Thêm cột UpdatedDate và CreatedDate là ngày ETL hiện tại, sử dụng hàm getdate()
 - Insert dòng dữ liệu mới trong NDS

Ví dụ 2: Accident



Bước 1: Load dữ liệu cần thiết từ Accident_stage và các bảng chứa dữ liệu cần thiết như: Accident_Severity, RoadType, Authority District, RuralorUrban và PostCode lên để xử lý.

```
TH 1.sql - DESK...35QKRNL\phuc (56)*  ↵ X
--SELECT
tmp.*

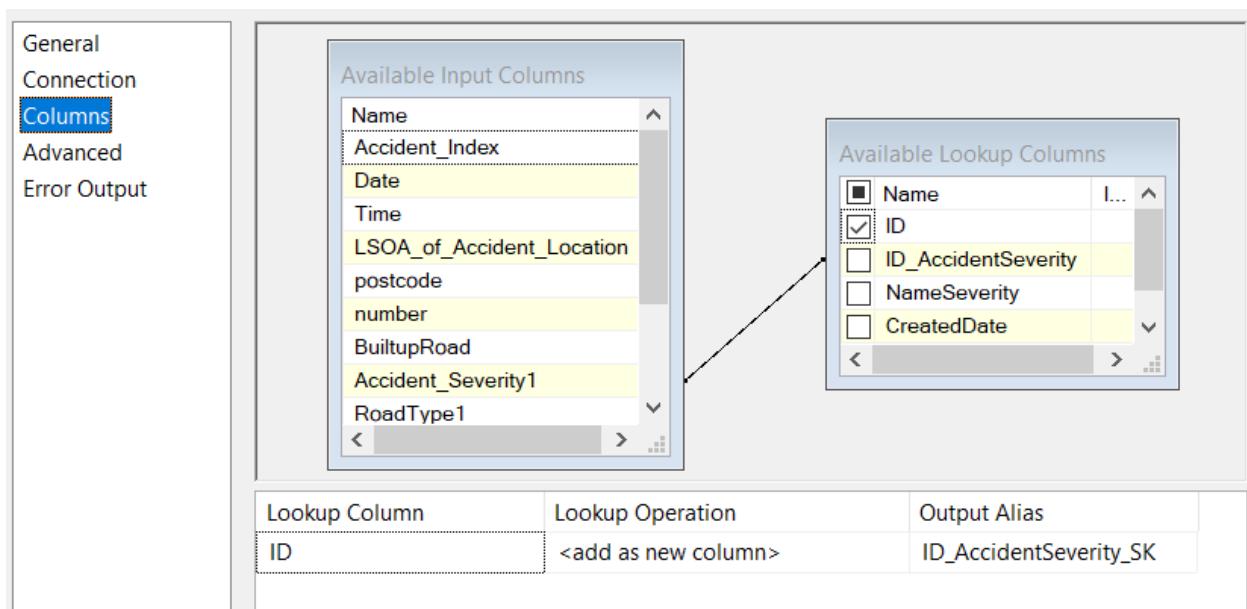
FROM
(
SELECT
    Accidents.Accident_Index,
    CONVERT(DATE,TRIM(Accidents.Date)),103) as Date,
    Accidents.Time,
    Cast(Accidents.Accident_Severity as int) as Accident_Severity,
    Cast(Accidents.Road_Type as int) as RoadType,
    Cast(Accidents.Local_Authority_District as int) as Local_Authority_District ,
    Cast(Accidents.Urban_or_Rural_Area as int) as Urban_or_Rural_Area,
    Case when Cast(Accidents.Speed_limit as int) >50 then 'Built up Road' else 'Non Built up Road' end as BuiltupRoad ,--Built up road
    Accidents.LSOA_of_Accident_Location,
    Cast(Accidents.Number_of_Casualties as int) as Number_of_Casualties,
    Postcodes.Stage.postcode ,
    ROW_NUMBER() OVER ( Partition by
        Accidents.Accident_Index
    ORDER BY geography::Point(Accidents.Latitude, Accidents.Longitude, 4326).STDistance(geography::Point(Postcodes.Stage.Latitude, Postcodes.Stage.Longitude, 4326))) number
    FROM Accidents_stage as Accidents

    -- -- -- -- -
    FROM Accidents_stage as Accidents
    JOIN
        PCD_LSOA_Stage
        ON Accidents.LSOA_of_Accident_Location = PCD_LSOA_Stage.lsoa11cd
    JOIN
        Postcodes_Stage
        ON PCD_LSOA_Stage.pcds like Postcodes_Stage.postcode + '%' where Accidents.Flag is null
) tmp
where tmp.number = 1
```

Bước 2: Bảng Accident có các khóa ngoại là Accident_Severity, RoadType, Authority District, RuralorUrban và PostCode

dbo.Accident
Columns
ID (PK, int, not null)
ID_Accident_Index (varchar(50), not null)
Accident_Severity_SK (FK, int, not null)
Date (date, null)
Time (varchar(50), null)
Local_Authority_District_SK (FK, int, null)
Road_Type_SK (FK, int, null)
Name_Built_up_Road (varchar(50), null)
LSOA_of_Accident_Location (varchar(50), null)
CreatedDate (datetime, null)
UpdatedDate (datetime, null)
PostCode_SK (FK, int, null)
UrbanRural_SK (FK, int, null)
Number_of_Casualties (int, null)

Ta sẽ Lookup xem Accident_Severity có tồn tại trong NDS hay chưa



Lookup Column	Lookup Operation	Output Alias
ID	<add as new column>	ID_AccidentSeverity_SK

Sau khi lấy được ID và ID_AccidentSeverity, tiếp theo chúng ta sẽ mapping AccidentSeverity (ở Stage) với ID_AccidentSeverity (NK ở NDS). Nếu mà mà lookup tìm thấy sẽ trả về ID_AccidentSeverity_SK (SK ở NDS). Tương tự đối với Lookup xem **RoadType**, **Authority District**, **RuralorUrban** và **PostCode** có tồn tại trong NDS hay chưa.

Bước 3: Kiểm tra xem các dòng dữ liệu đã tồn tại trong bảng Accident của NDS hay chưa

Input Column	Lookup Column
Accident_Index	ID_Accident_Index

Sau khi lấy được ID_Accident_Index, tiếp theo chúng ta sẽ mapping Accident_Index (NK ở Stage) với ID_Accident_Index (NK ở NDS).

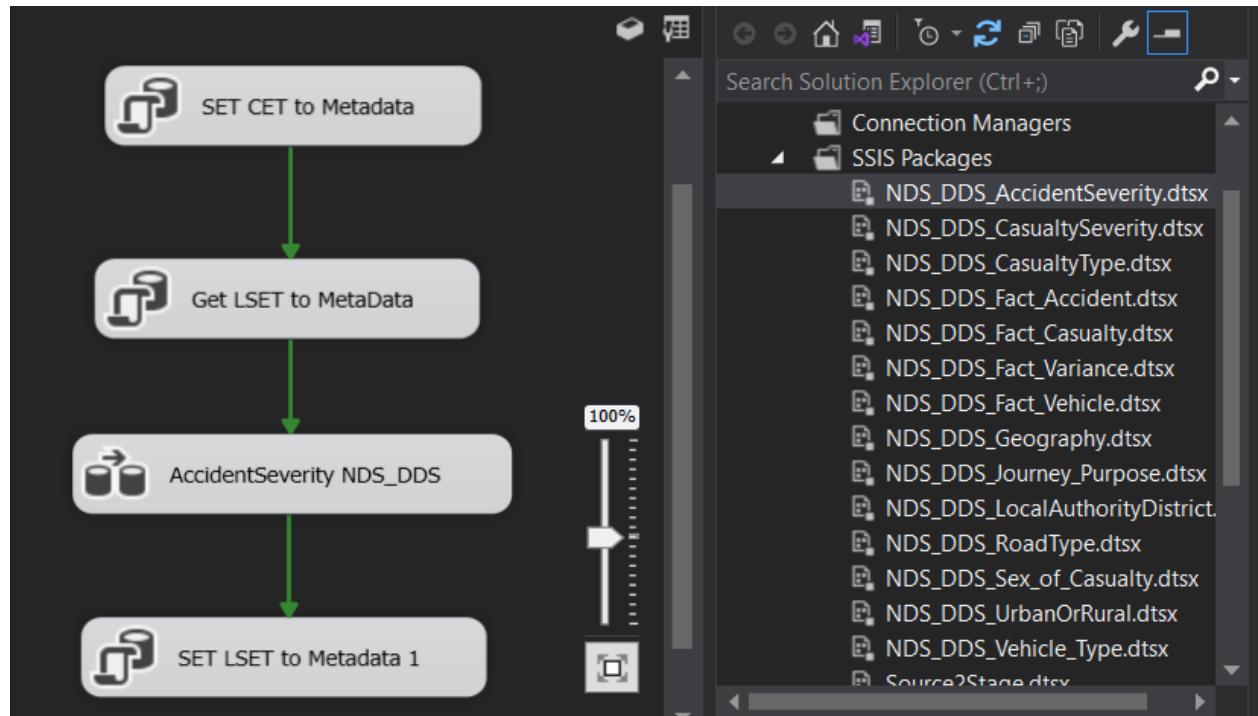
Bước 4: Sau khi Lookup xong, có 2 trường hợp xảy ra:

- Nếu đã tồn tại (tìm thấy được giá trị ID_Accident_Index) thì tiến hành
 - Thêm cột UpdatedDate là ngày ETL hiện tại, sử dụng hàm getdate()
 - Update dòng dữ liệu trong NDS
- Nếu chưa tồn tại thì tiến hành
 - Thêm cột CreateDate và UpdatedDate là ngày ETL hiện tại, sử dụng hàm getdate()
 - Insert dòng dữ liệu mới trong NDS.

8. Quá trình từ NDS vào DDS

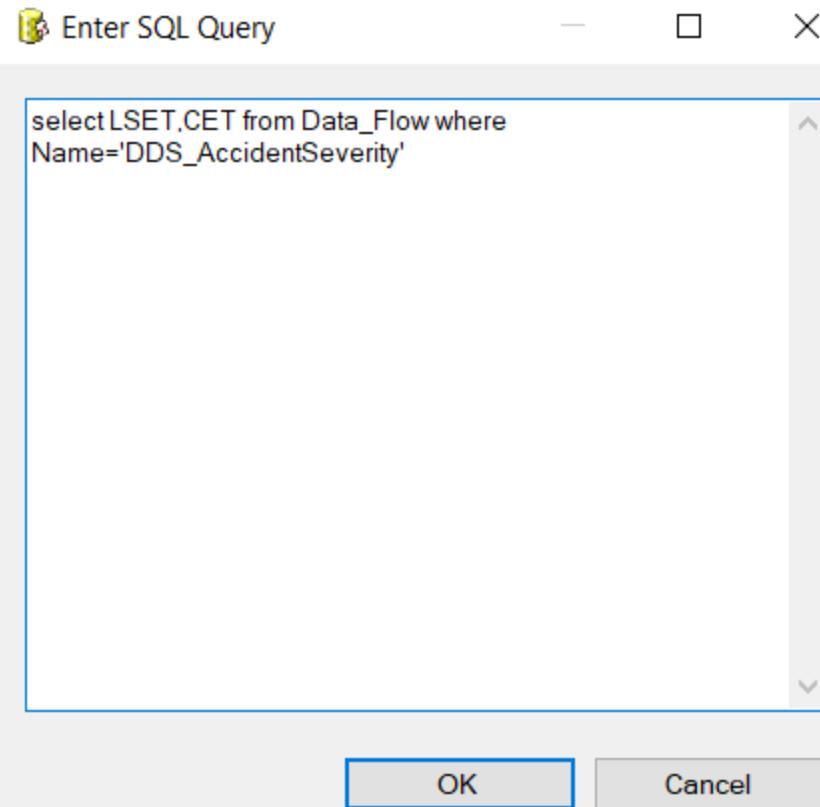
8.1 Dim Table

- Ví dụ về trường hợp *Changing Attribute*(**Type 1 change**): bảng **AccidentSeverity**
AccidentSeverity(NDS) → AccidentSeverity (DDS) – Type 1

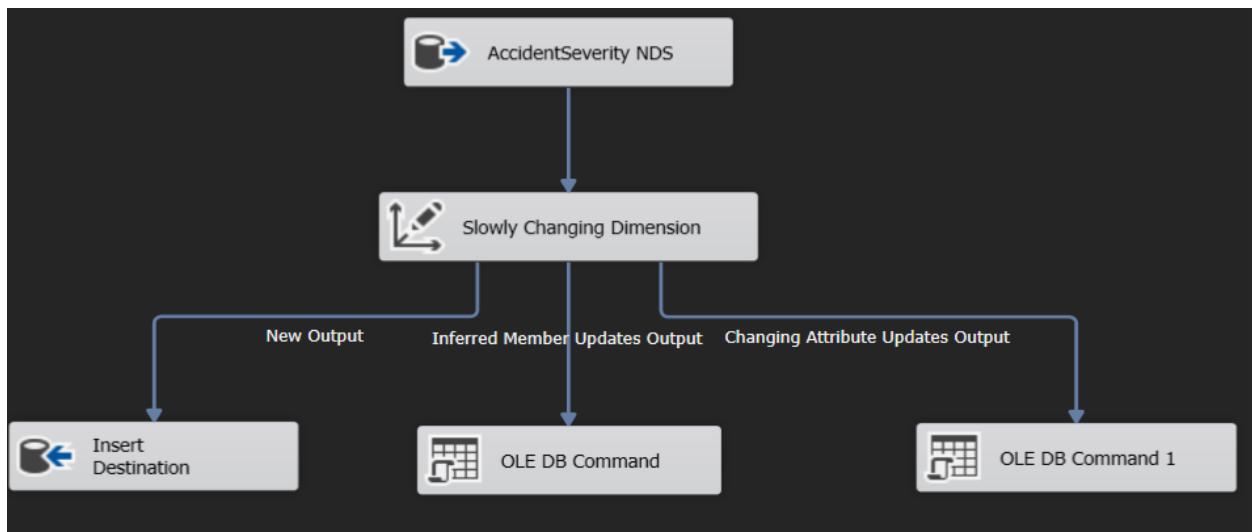


Bước 1: Set CET: lấy thời gian ETL hiện tại

Bước 2: Get LSET: lấy thời điểm rút trích từ NDS vào DDS gần nhất được lưu trong bảng Data_Flow từ database MetaData

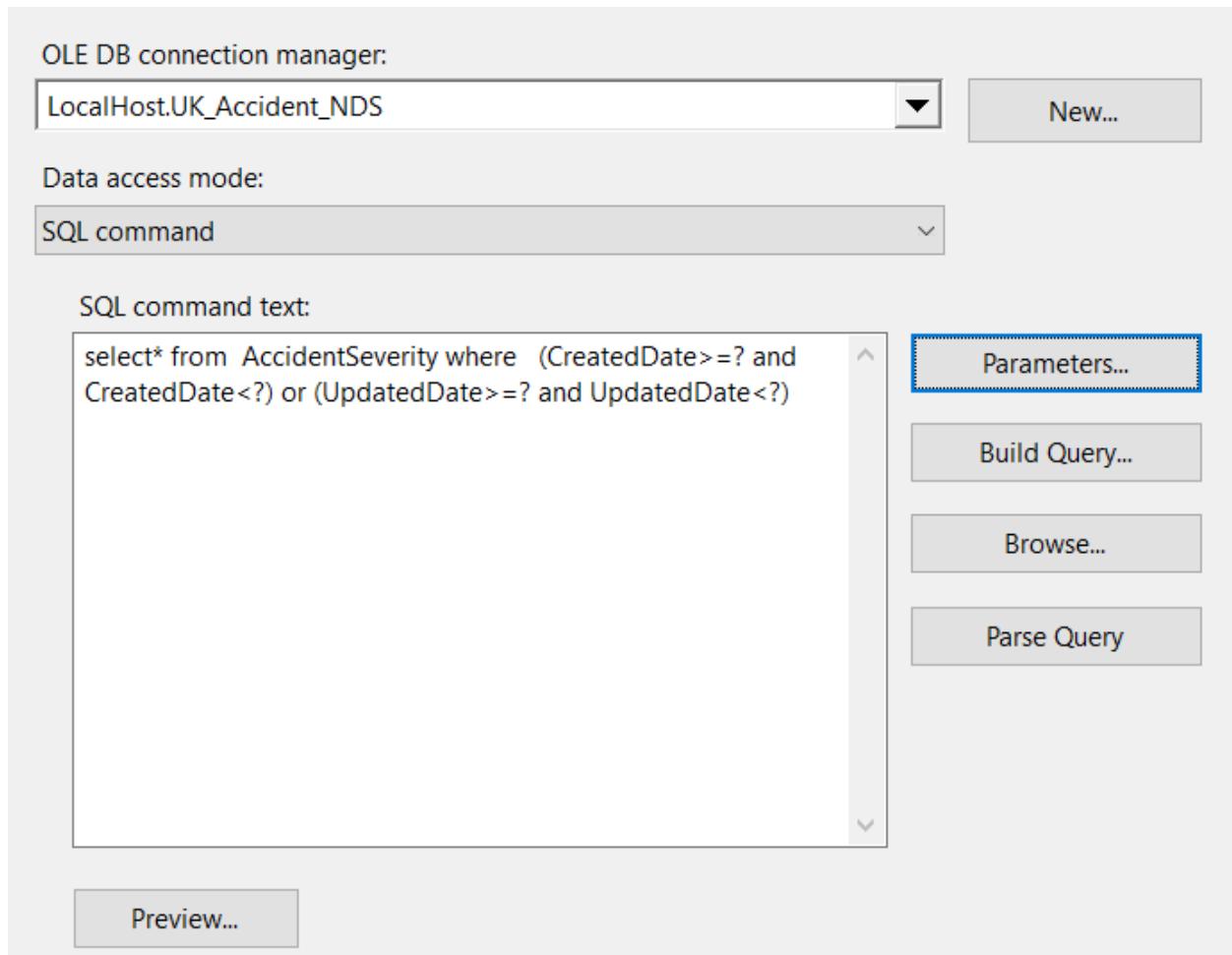


Bước 3: Populate AccidentSeverity, thiết kế Data flow task như hình bên dưới



Bước 3.1: Load dữ liệu từ NDS lên để xử lý, ta rút trích những dữ liệu mà nằm trong khoảng LSET đến CET để nạp vào DDS

Mục đích: chỉ lấy những dòng thay đổi, mới được thêm vào từ lần ETL trước đó hoặc những dòng mới được update từ lần ETL trước đó => tránh trùng lắp dữ liệu.



Bước 3.2: Sử dụng component Slowly Changing Dimension Wizard
Xác định **ID_AccidentSeverity** là **Business key**

Slowly Changing Dimension Wizard

Select a Dimension Table and Keys

Select a dimension table to load and map columns in the transformation input to columns in the

Connection manager:

Table or view:

Input Columns	Dimension Columns	Key Type
ID_AccidentSeverity	AccidentSeverity_NK	Business key
NameSeverity	NameSeverity	Not a key column

Bước 3.3: Thuộc tính NameSeverity sẽ sử dụng Changing attribute (Type 1).

Slowly Changing Dimension Wizard

Slowly Changing Dimension Columns

Manage the changes to column data in your slowly changing dimensions by setting the change type for

Fixed Attribute
Select this type when the value in a column should not change. Changes are treated as errors.

Changing Attribute
Select this type when changed values should overwrite existing values. This is a Type 1 change.

Historical Attribute
Select this type when changes in column values are saved in new records. Previous values are saved in

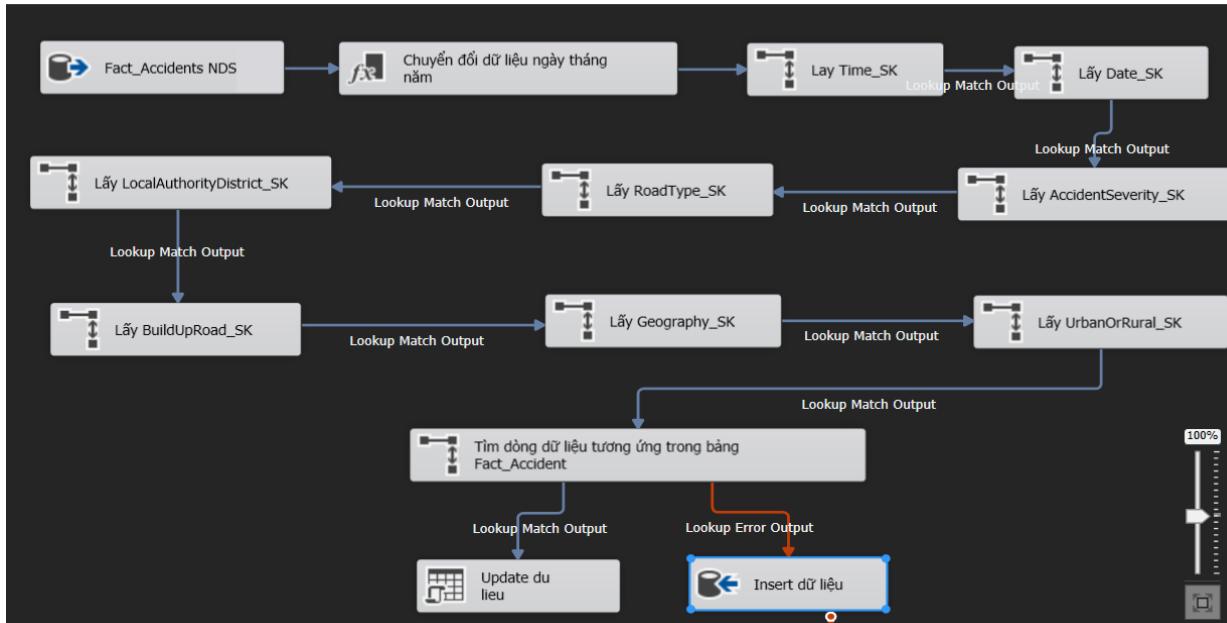
Select a change type for slowly changing dimension columns:

Dimension Columns	Change Type
NameSeverity	Changing attribute

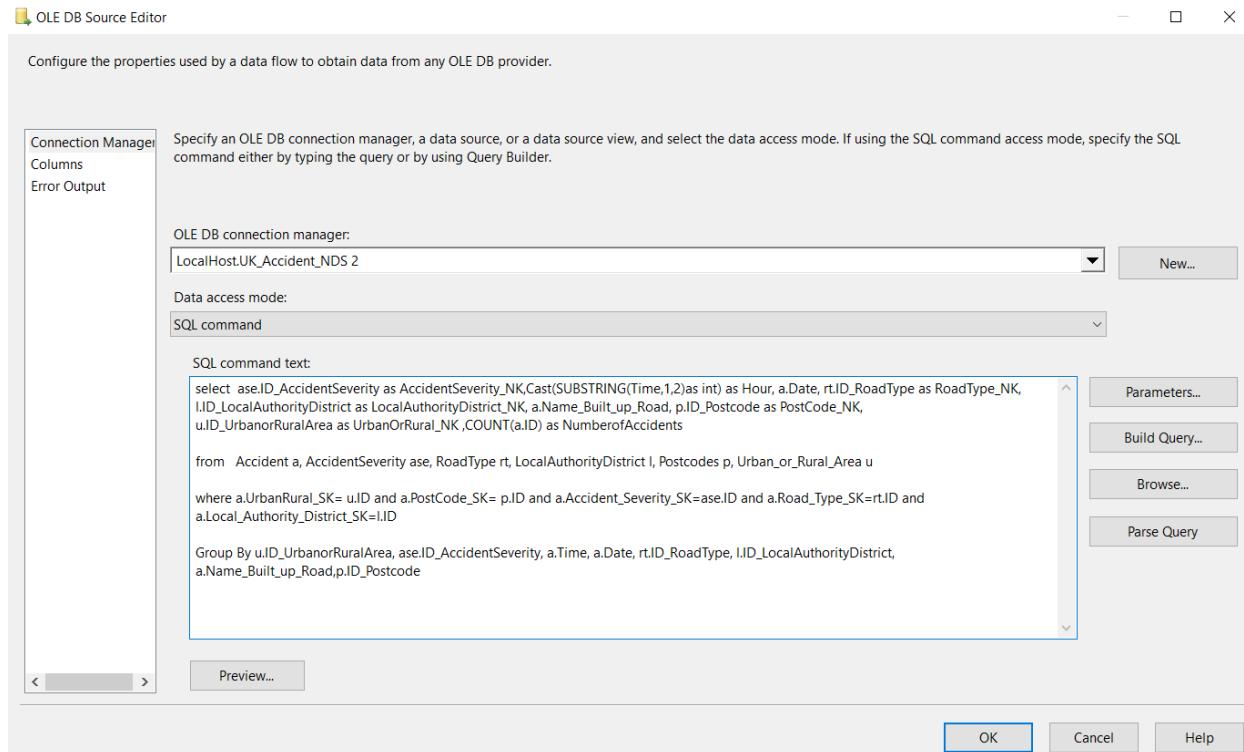
Bước 4: Cuối cùng gán lại LSET = CET

8.2 Fact Table

a. Fact Accident



Bước 1: Load dữ liệu từ các bảng trong NDS để lấy các dữ liệu phù hợp với cấu trúc bảng Fact_Accident trong DDS.



Bước 2: Thêm các thuộc tính Day, Month, Year được lấy ra từ Date (Ngày xảy ra tai nạn) trong dữ liệu đã được load ở bước đầu tiên.

Derived Column Transformation Editor

Specify the expressions used to create new column values, and indicate whether the values update existing columns or populate new columns.

[+]  Variables and Parameters
[+]  Columns

[+]  Mathematical Functions
[+]  String Functions
[+]  Date/Time Functions
[+]  NULL Functions
[+]  Type Casts
[+]  Operators

Description:

Derived Column Name	Derived Column	Expression	Data Type	L...
Day	<add as new column>	DAY(Date)	four-byte signed integer	
Month	<add as new column>	MONTH(Date)	four-byte signed integer	
Year	<add as new column>	YEAR(Date)	four-byte signed integer	



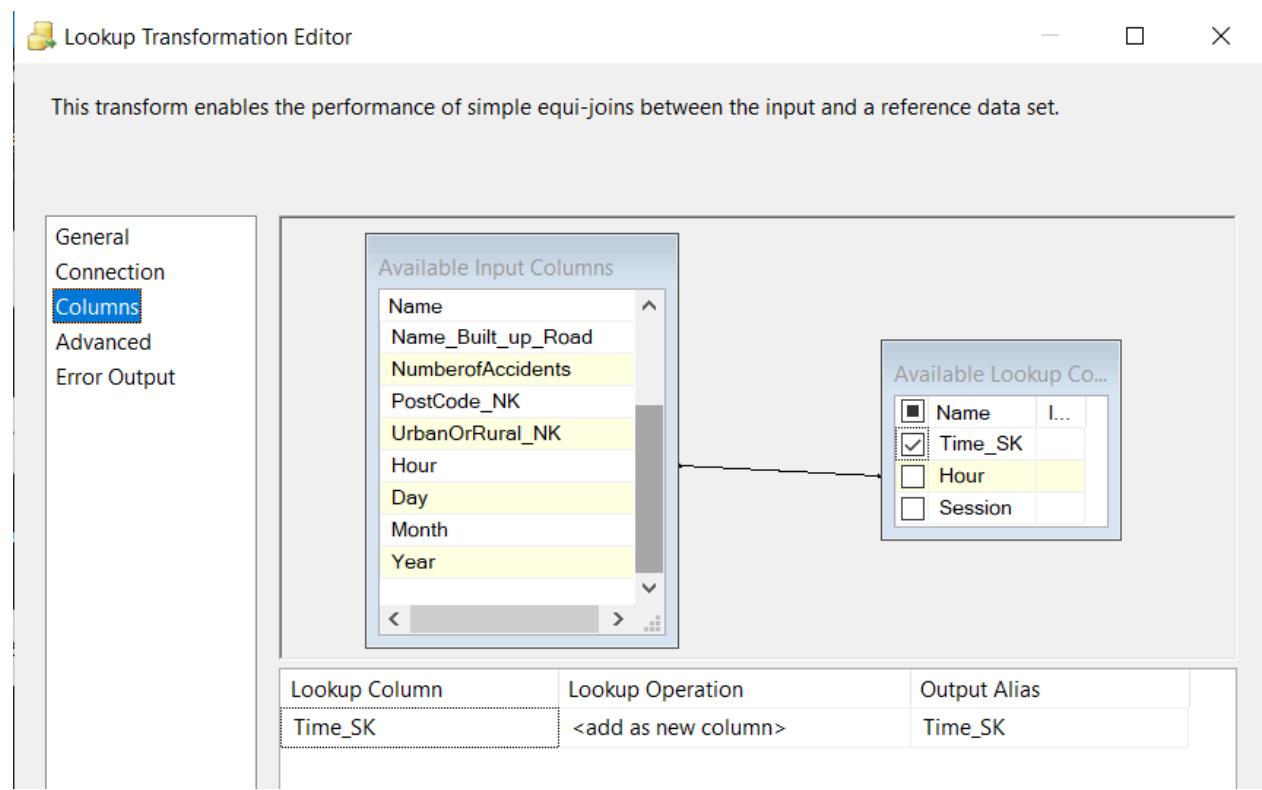
Configure Error Output...

OK

Cancel

Help

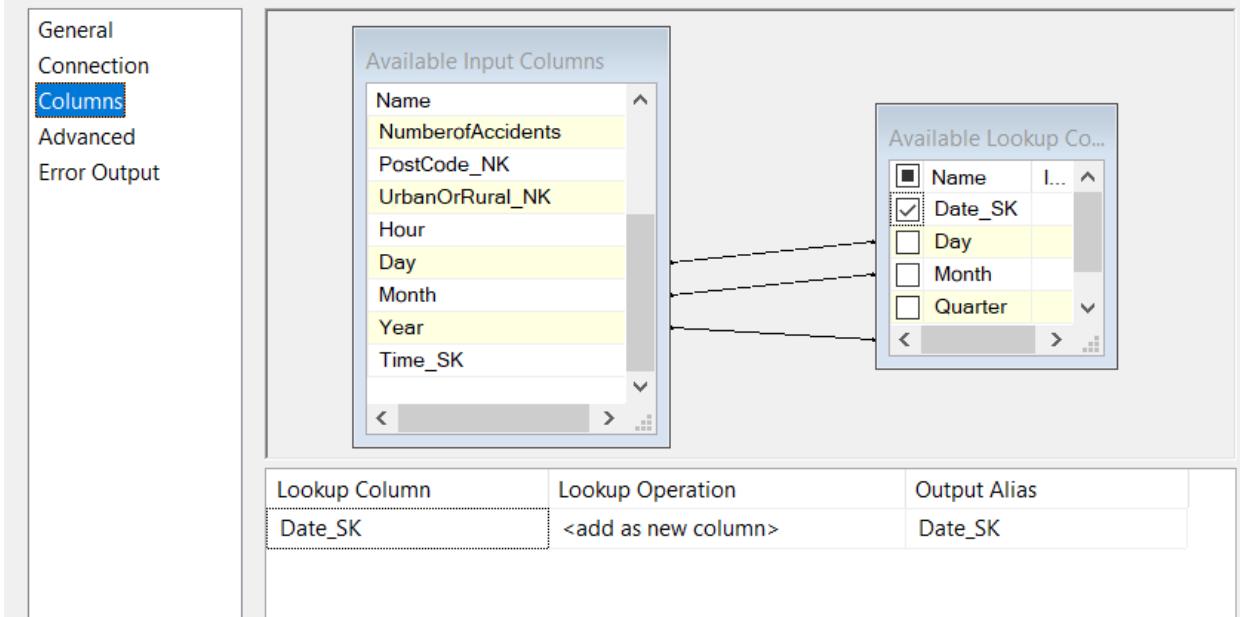
Bước 3: Lookup Time: từ Hour lấy được ở bước 1, ta sẽ mapping để tìm ra
được khóa Time_SK tương ứng trong bảng Dim Time



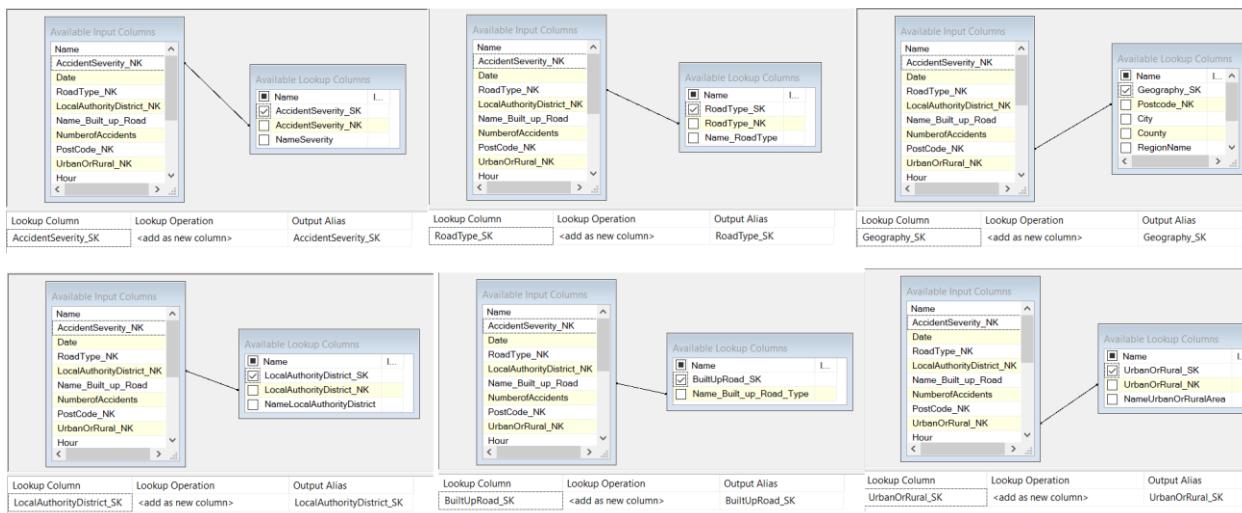
Bước 4: Lookup Date: từ Day, Month, Year ở bước 2, ta sẽ mapping để tìm ra được khóa Date_SK tương ứng trong bảng Dim Date

Lookup Transformation Editor

This transform enables the performance of simple equi-joins between the input and a reference data set.



Bước 5: Lookup các khóa ngoại: đổi với AccidentSeverity_NK, RoadType_NK, Geography_NK, LocalAuthorityDistrict_NK, BuiltUpRoad_NK, UrbanOrRural_NK từ các bảng dimension tương ứng trong DDS và trả về các khóa AccidentSeverity_SK, RoadType_SK, Geography_SK, LocalAuthorityDistrict_SK, BuiltUpRoad_SK, UrbanOrRural_SK



Bước 6: Tìm kiếm dữ liệu đã tồn tại trong DDS chưa

- Nếu dữ liệu chưa tồn tại thì Insert dữ liệu vào bảng Fact_Accident, mapping các giá trị output mà đã lookup ở các bước trên tương ứng với các thuộc tính của bảng Fact_Accident

OLE DB Destination Editor

Configure the properties used to insert data into a relational database using an OLE DB provider.

Connection Manager
Mappings
Error Output

Available Input Columns

Name
AccidentSeverity_NK
Date
RoadType_NK
LocalAuthorityDistrict_NK
Name_Built_up_Road
NumberofAccidents
PostCode_NK
UrbanOrRural_NK
Hour

Available Destination Colu...

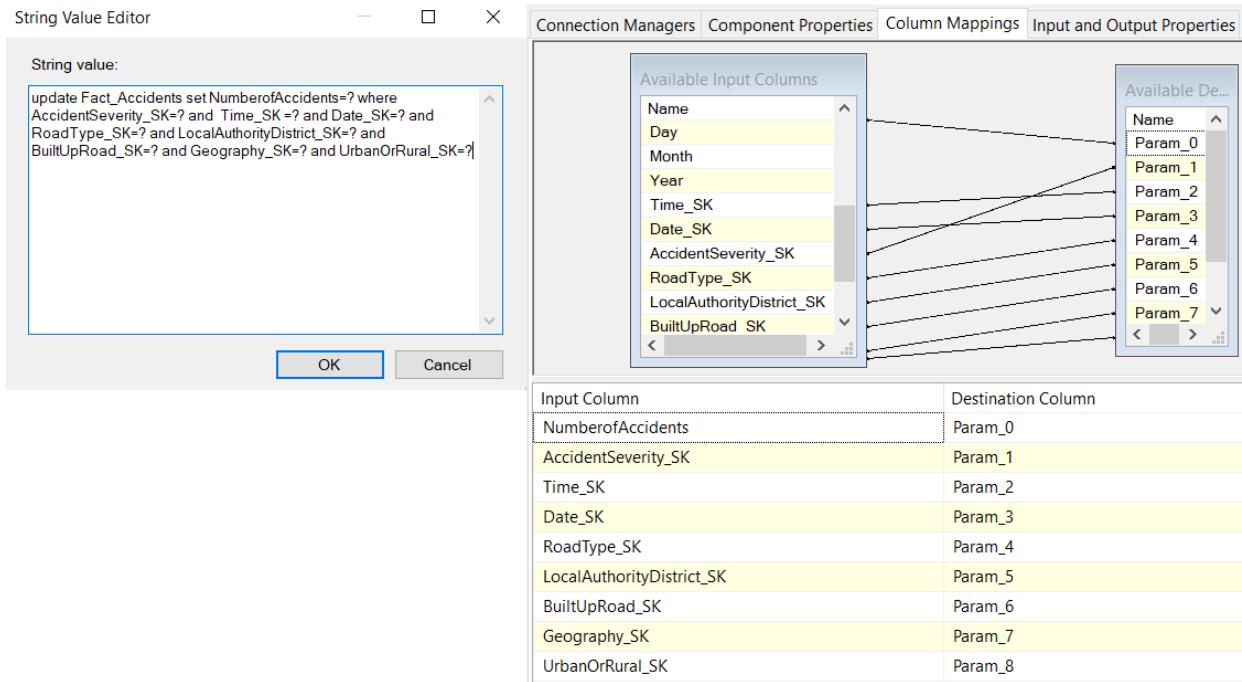
Name
AccidentSeverity_SK
Time_SK
RoadType_SK
LocalAuthorityDistrict_SK
BuiltUpRoad_SK
Geography_SK
NumberofAccidents
Date_SK

Input Column Destination Column

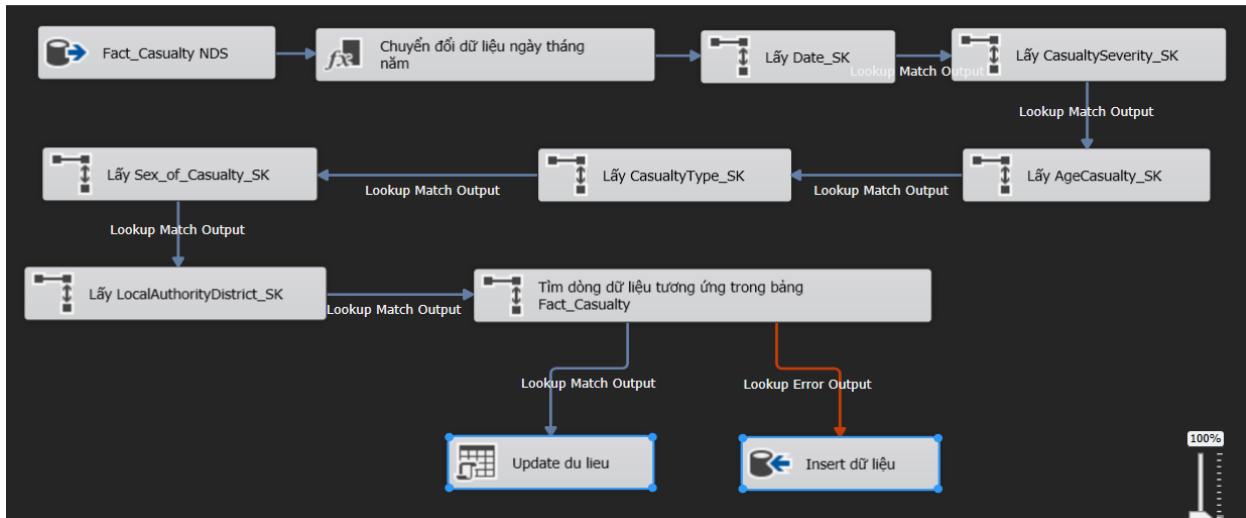
AccidentSeverity_SK	AccidentSeverity_SK
Time_SK	Time_SK
RoadType_SK	RoadType_SK
LocalAuthorityDistrict_SK	LocalAuthorityDistrict_SK
BuiltUpRoad_SK	BuiltUpRoad_SK
Geography_SK	Geography_SK
NumberofAccidents	NumberofAccidents
Date_SK	Date_SK
UrbanOrRural_SK	UrbanOrRural_SK

OK Cancel Help

- Nếu dữ liệu đã tồn tại thì update



b. Fact Casualty



Bước 1: Load dữ liệu từ các bảng trong NDS để lấy các dữ liệu phù hợp với cấu trúc bảng Fact_Casualty trong DDS

OLE DB connection manager:

localhost.UK_Accident_NDS

Data access mode:

SQL command

SQL command text:

```
select cs.ID_CasualtySeverity as CasualtySeverity_NK, c.Age_of_Casualty, ct.ID_CasualtyType as CasualtyType_NK, sc.ID_Sex_of_Casualty as Sex_of_Casualty_NK, a.Date, l.ID_LocalAuthorityDistrict as LocalAuthorityDistrict_NK, COUNT(c.ID) as NumberofCasualty
from Casualty c, CasualtySeverity cs, CasualtyType ct, Sex_of_Casualty sc, Accident a,
LocalAuthorityDistrict l
where c.Casualty_Severity_SK= cs.ID and c.Casualty_Type_SK= ct.ID and
c.Sex_of_Casualty_SK=sc.ID and c.ID_Accident_SK =a.ID and a.Local_Authority_District_SK =l.ID
Group by cs.ID_CasualtySeverity, c.Age_of_Casualty, ct.ID_CasualtyType, sc.ID_Sex_of_Casualty,
a.Date, l.ID_LocalAuthorityDistrict
```

Bước 2: Thêm các thuộc tính Day, Month, Year được lấy ra từ Date (Ngày xảy ra tai nạn) trong dữ liệu đã được load ở bước đầu tiên.

Derived Column Transformation Editor

Specify the expressions used to create new column values, and indicate whether the values update existing columns or populate new columns.

[+]  Variables and Parameters
[+]  Columns

[+]  Mathematical Functions
[+]  String Functions
[+]  Date/Time Functions
[+]  NULL Functions
[+]  Type Casts
[+]  Operators

Description:

Derived Column Name	Derived Column	Expression	Data Type	L
Day	<add as new column>	DAY(Date)	four-byte signed inte...	
Month	<add as new column>	MONTH(Date)	four-byte signed inte...	
Year	<add as new column>	YEAR(Date)	four-byte signed inte...	



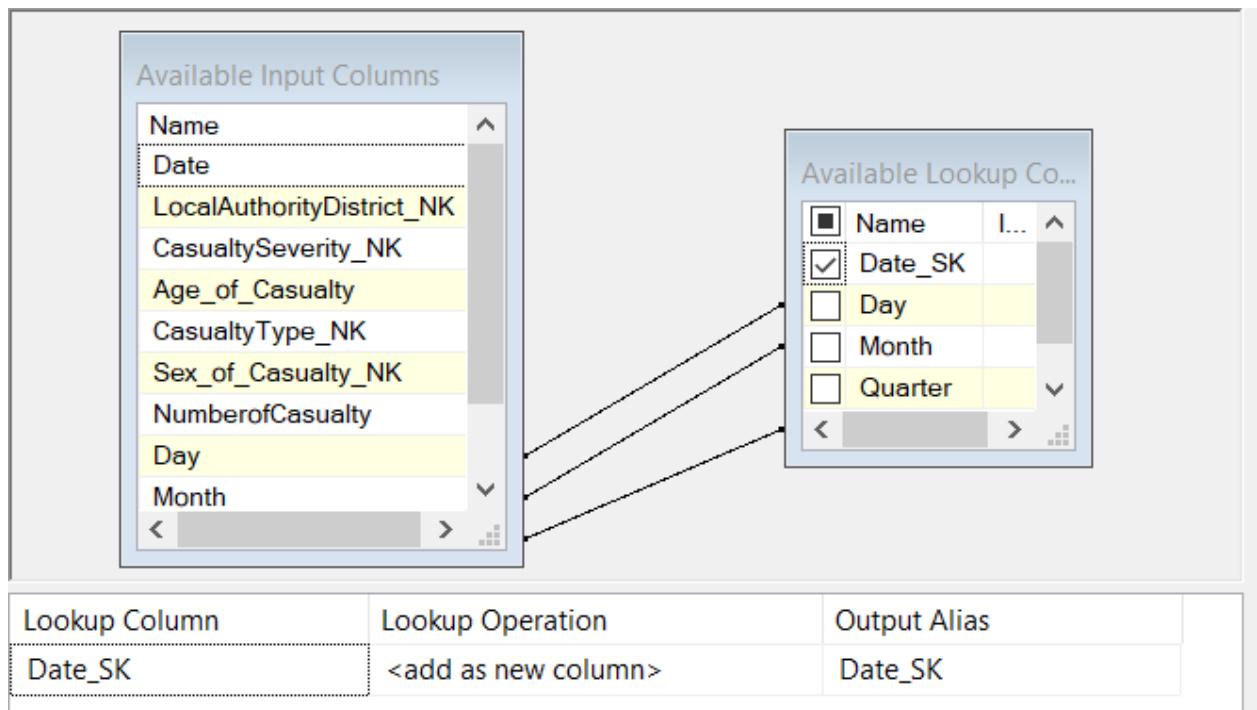
Configure Error Output...

OK

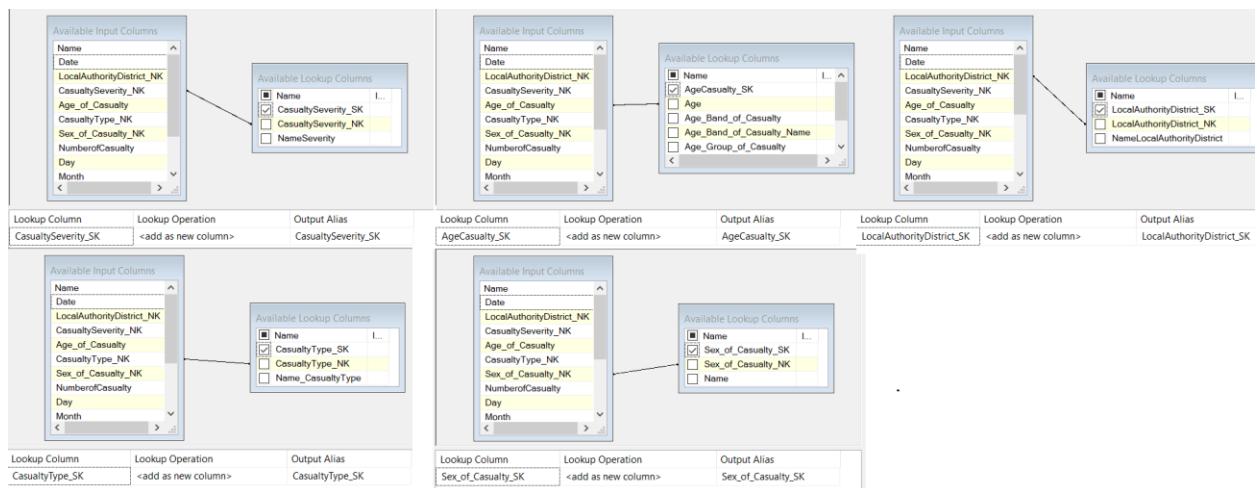
Cancel

Help

Bước 3: Lookup Date: từ Day, Month, Year ở bước 2, ta sẽ mapping để tìm ra được khóa Date_SK tương ứng trong bảng Dim Date

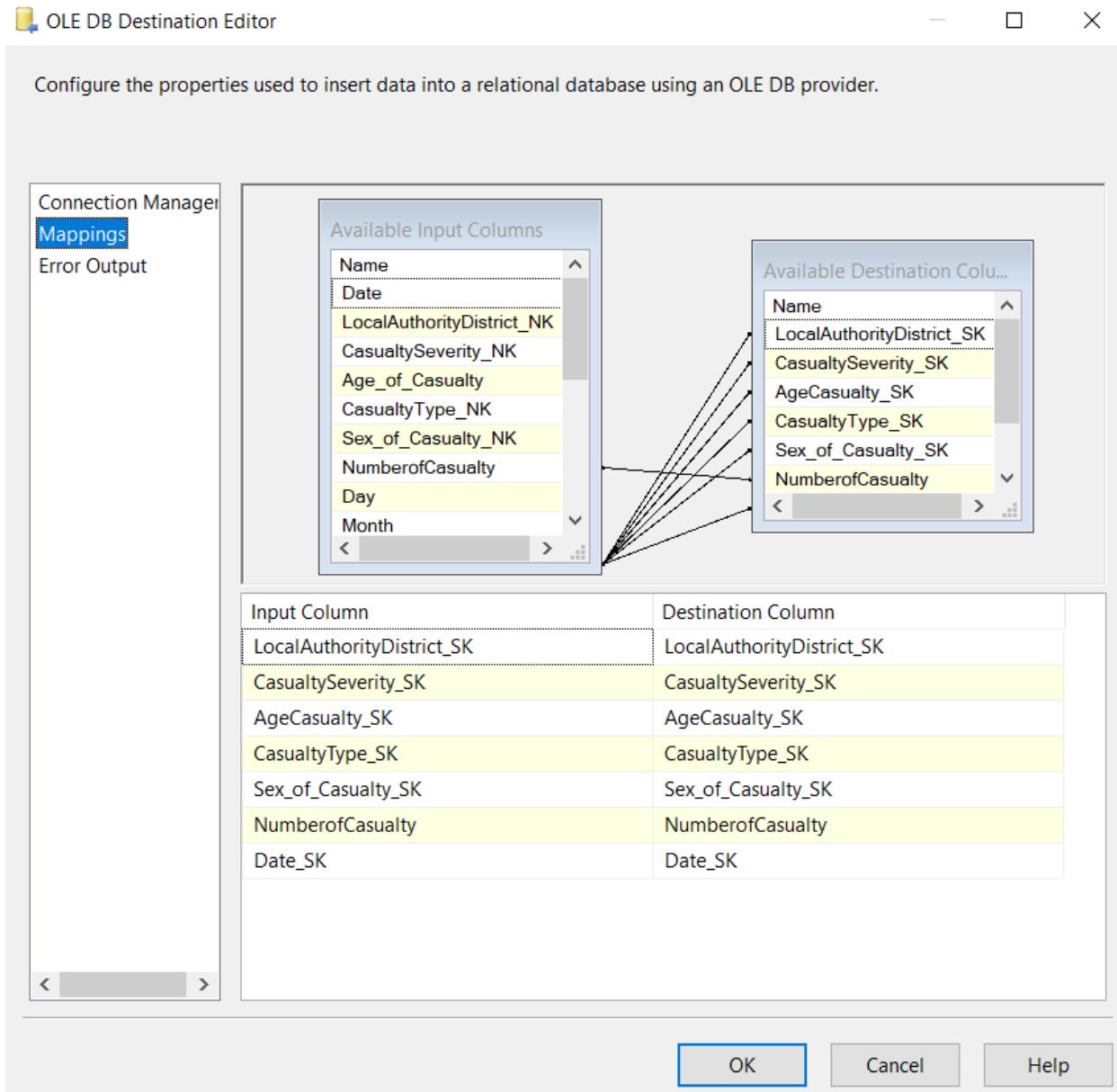


Bước 4: Lookup các khóa ngoại: đổi với CasualtySeverity_NK, AgeCasualty_NK, LocalAuthorityDistrict_NK, CasualtyType_NK, Sex_of_Casualty_NK từ các bảng dimension tương ứng trong DDS và trả về các khóa CasualtySeverity_SK, AgeCasualty_SK, LocalAuthorityDistrict_SK, CasualtyType_SK, Sex_of_Casualty_SK

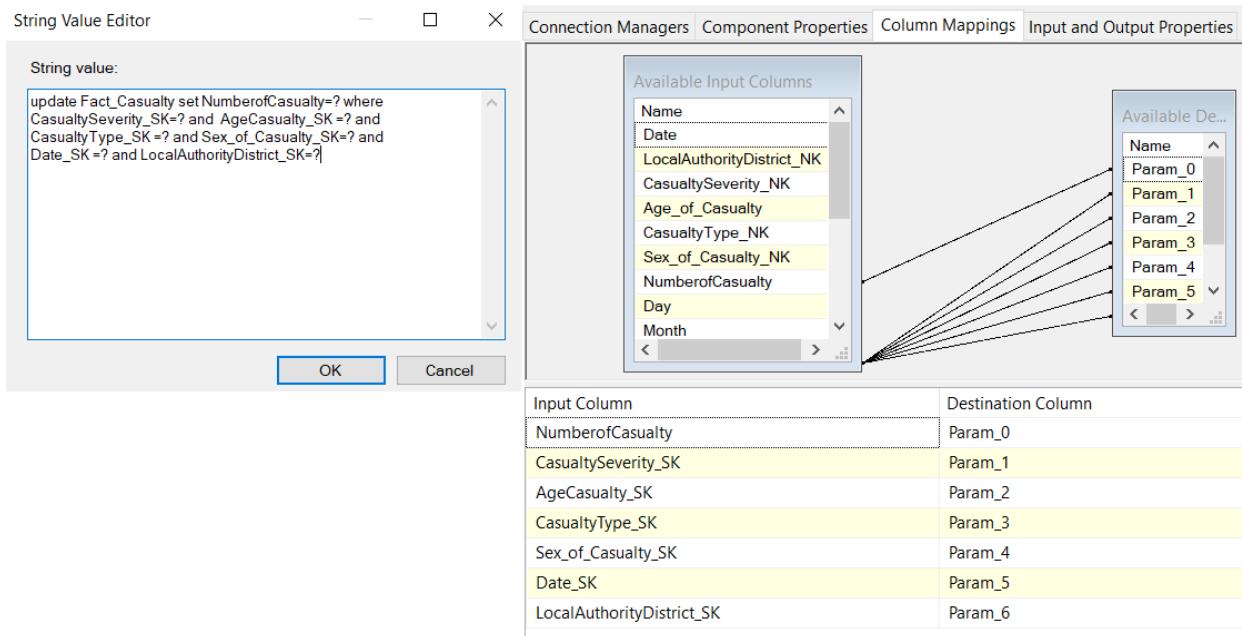


Bước 5: Tìm kiếm dữ liệu đã tồn tại trong DDS chưa

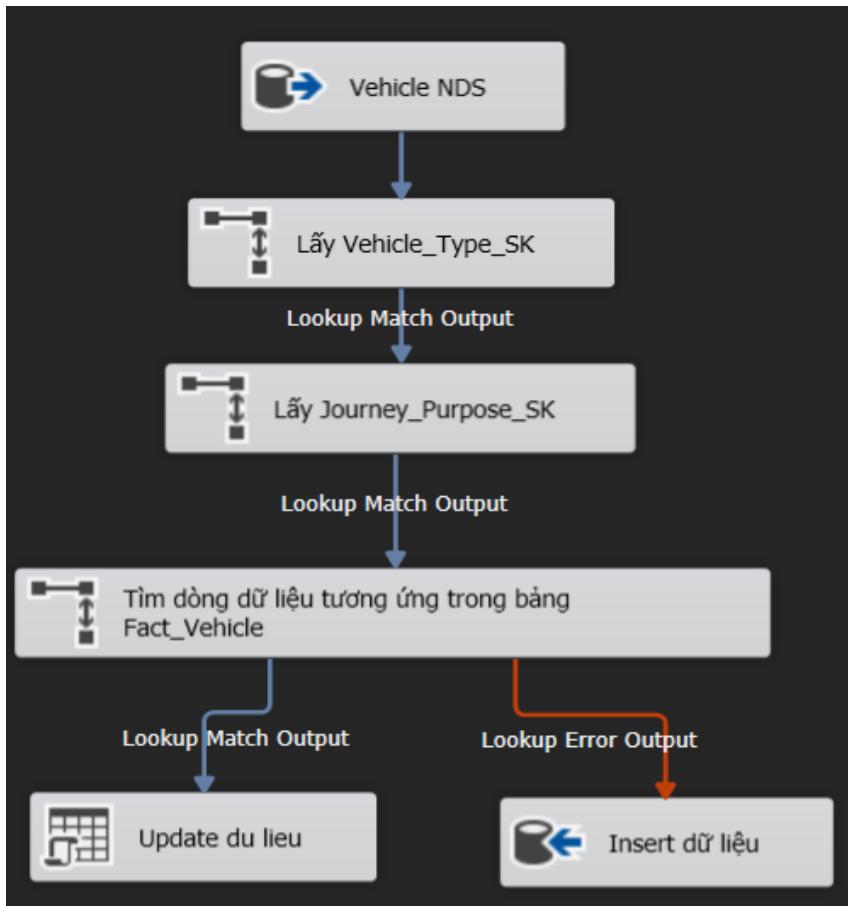
- Nếu dữ liệu chưa tồn tại thì Insert dữ liệu vào bảng Fact_Casualty, mapping các giá trị output mà đã lookup ở các bước trên tương ứng với các thuộc tính của bảng Fact_Casualty



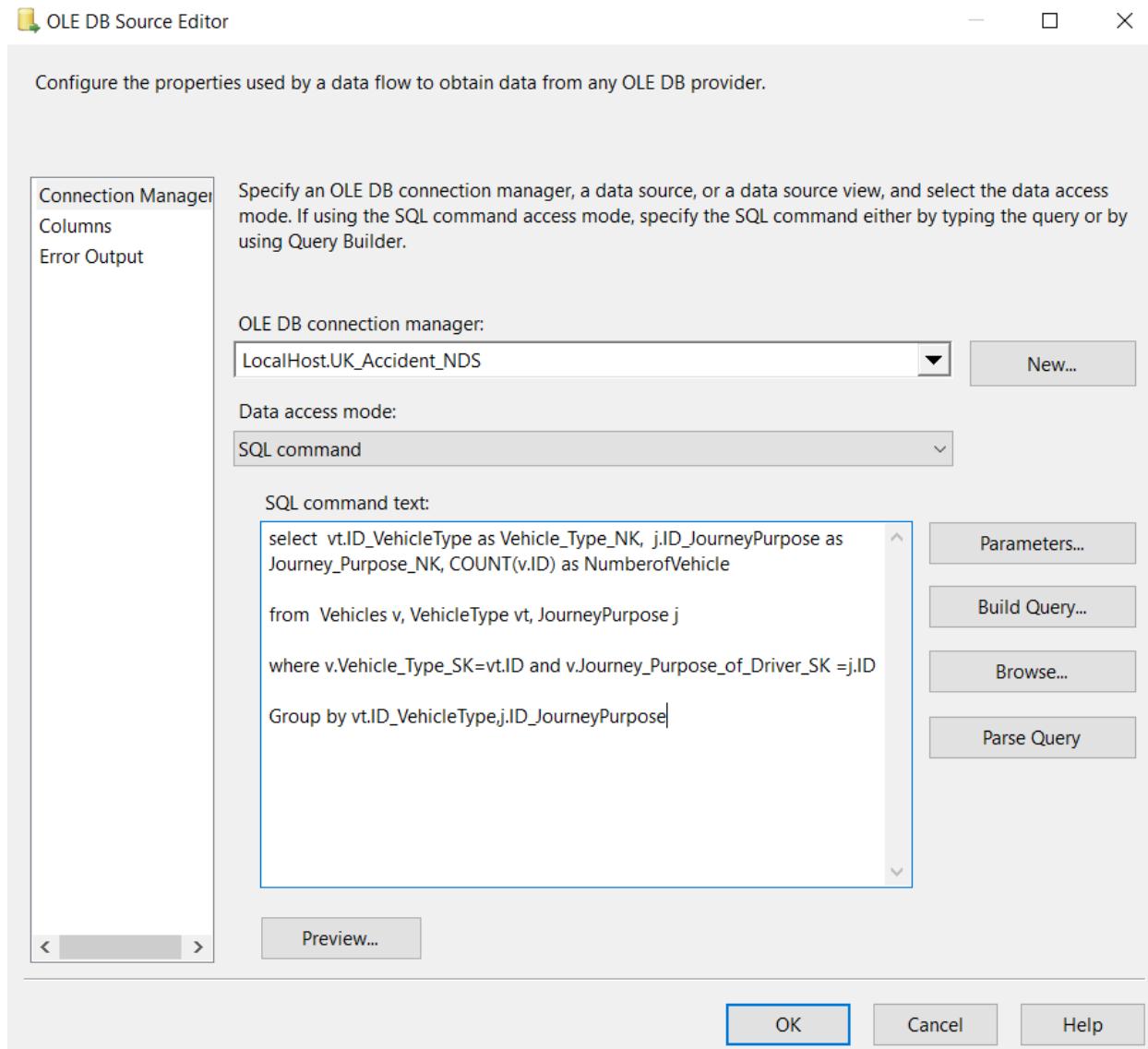
- Nếu dữ liệu đã tồn tại thì update



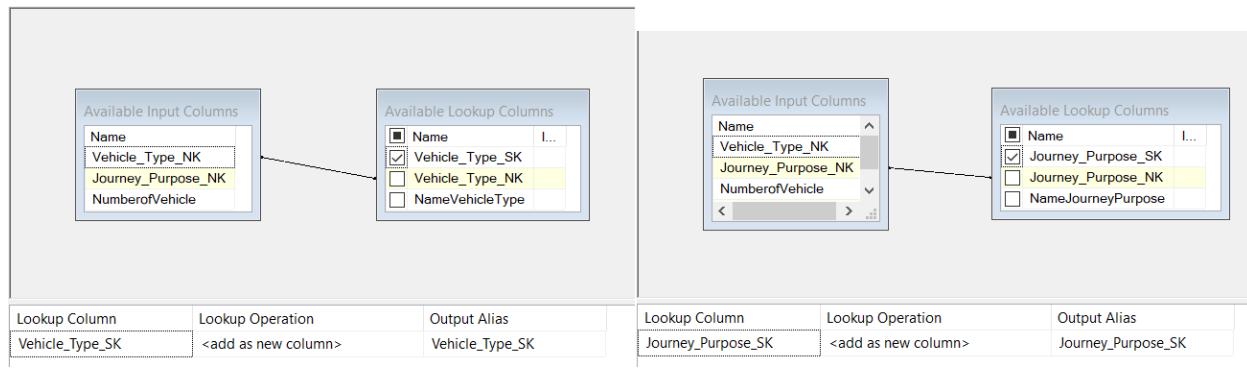
c. Fact Vehicle



Bước 1: Load dữ liệu từ các bảng trong NDS để lấy các dữ liệu phù hợp với cấu trúc bảng Fact_Vehicle trong DDS.



Bước 2: Lookup các khóa ngoại: đổi với Vehicle_Type_NK, Journey_Purpose_NK từ các bảng dimension tương ứng trong DDS và trả về các khóa Vehicle_Type_SK, Journey_Purpose_SK



Bước 3: Tìm kiếm dữ liệu đã tồn tại trong DDS chưa

- Nếu dữ liệu chưa tồn tại thì Insert dữ liệu vào bảng Fact_Vehicle, mapping các giá trị output mà đã lookup ở các bước trên tương ứng với các thuộc tính của bảng Fact_Vehicle

OLE DB Destination Editor

Configure the properties used to insert data into a relational database using an OLE DB provider.

Connection Manager
Mappings
Error Output

Available Input Columns

Name
Vehicle_Type_NK
Journey_Purpose_NK
NumberofVehicle
Vehicle_Type_SK
Journey_Purpose_SK
ErrorCode

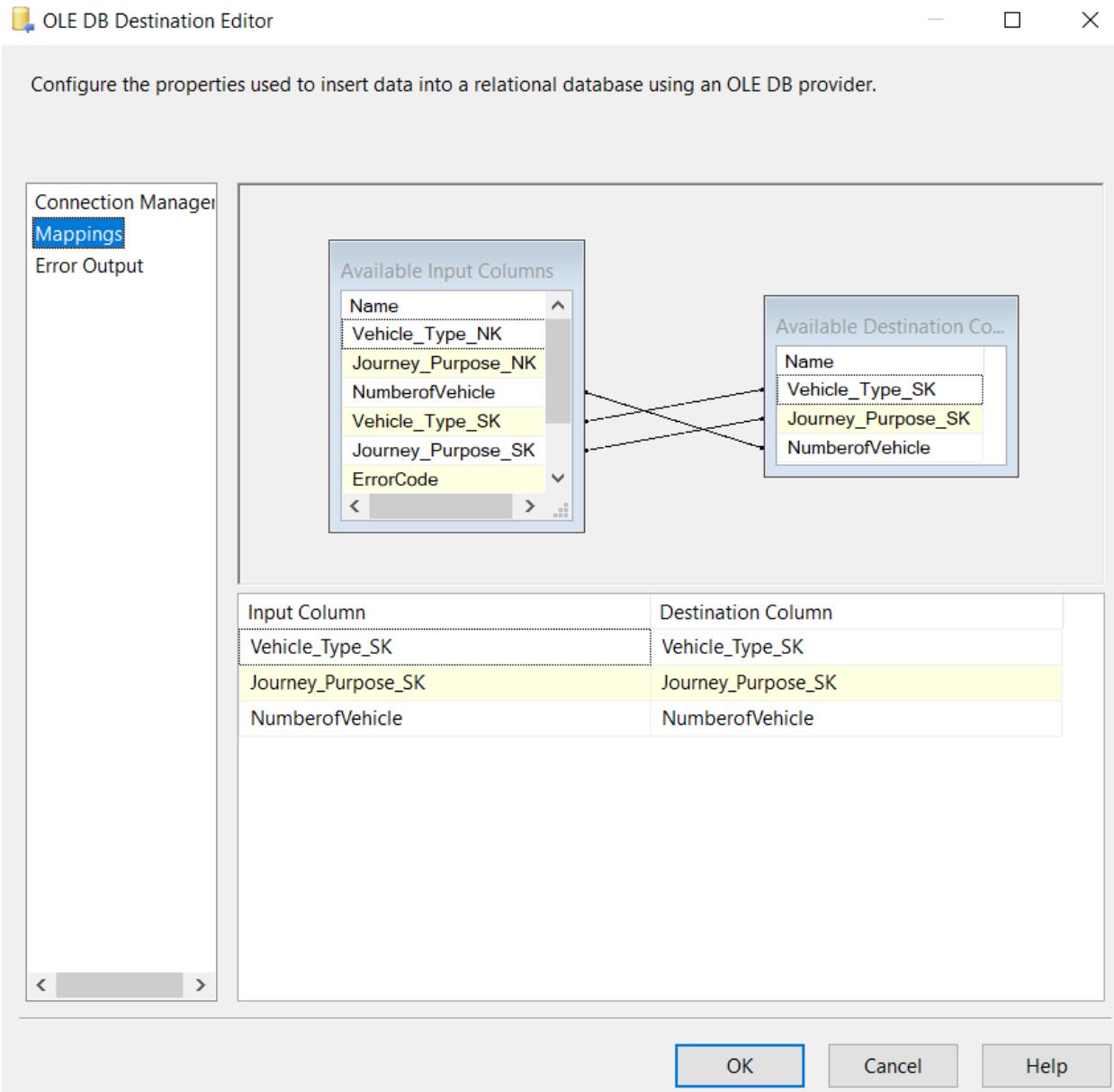
Available Destination Co...

Name
Vehicle_Type_SK
Journey_Purpose_SK
NumberofVehicle

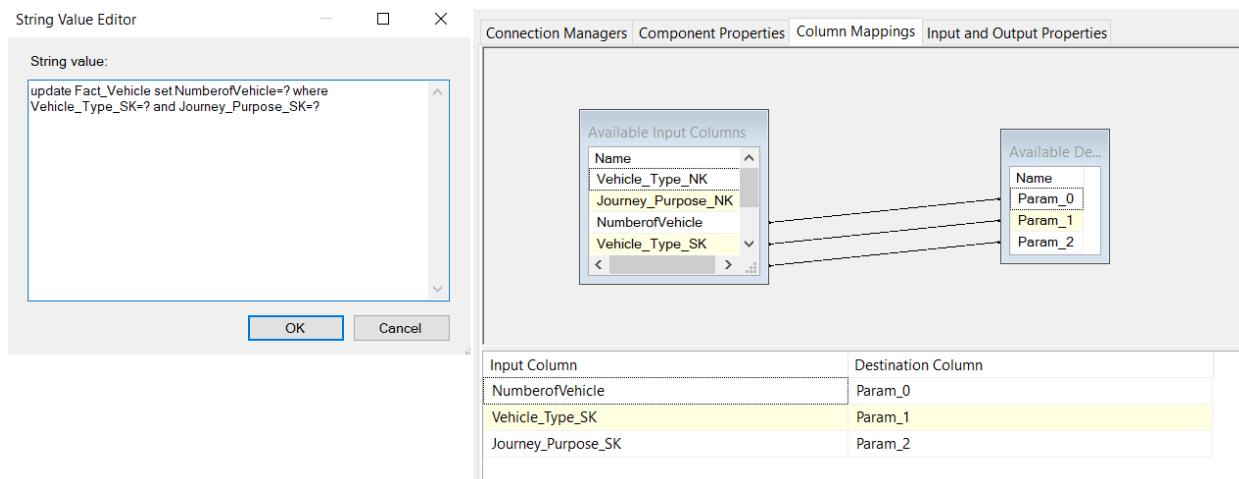
Input Column Destination Column

Vehicle_Type_SK	Vehicle_Type_SK
Journey_Purpose_SK	Journey_Purpose_SK
NumberofVehicle	NumberofVehicle

OK Cancel Help

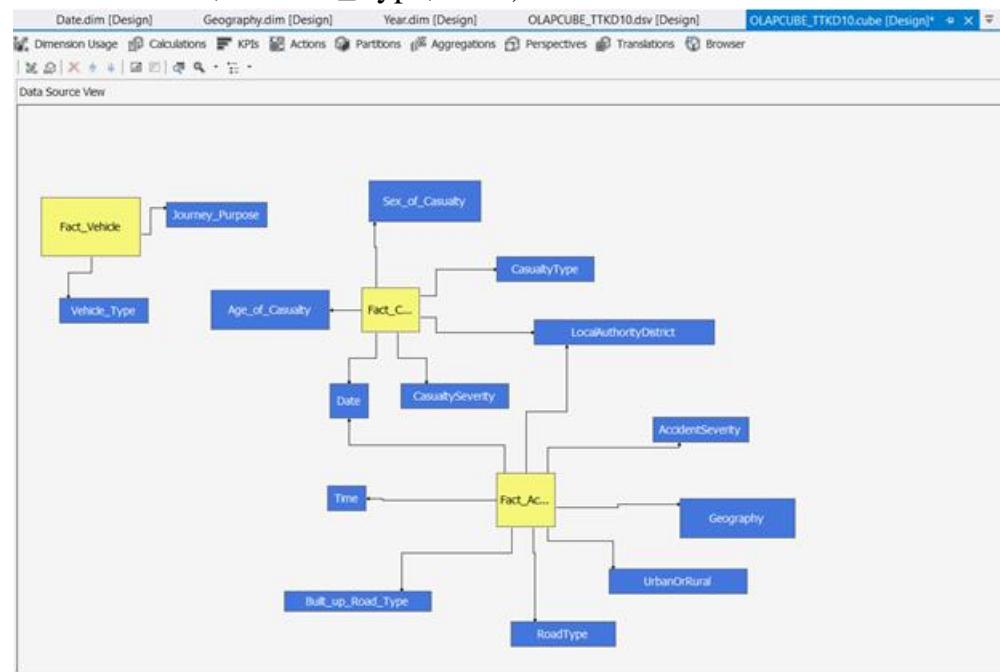


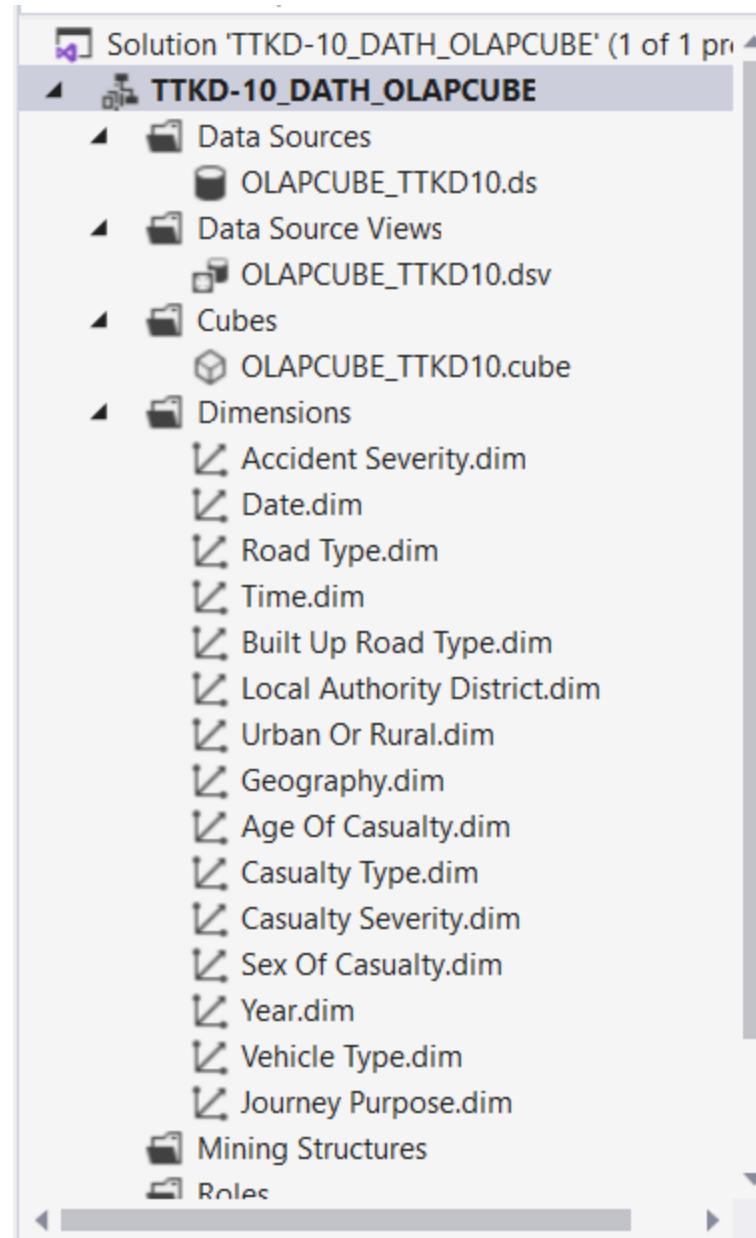
- Nếu dữ liệu đã tồn tại thì update



9. Thiết kế và xây dựng Cube

Thiết kế Cube gồm 3 fact và 15 dimensions (AccidentSeverity, Age_of_Casualty, Built_up_Road_Type, CasualtySeverity, CasualtyType, Date, Geography, Journey_Purpose, LocalAuthorityDistrict, RoadType, Sex_of_Casualty, Time, UrbanOrRural, Vehicle_Type, Year).





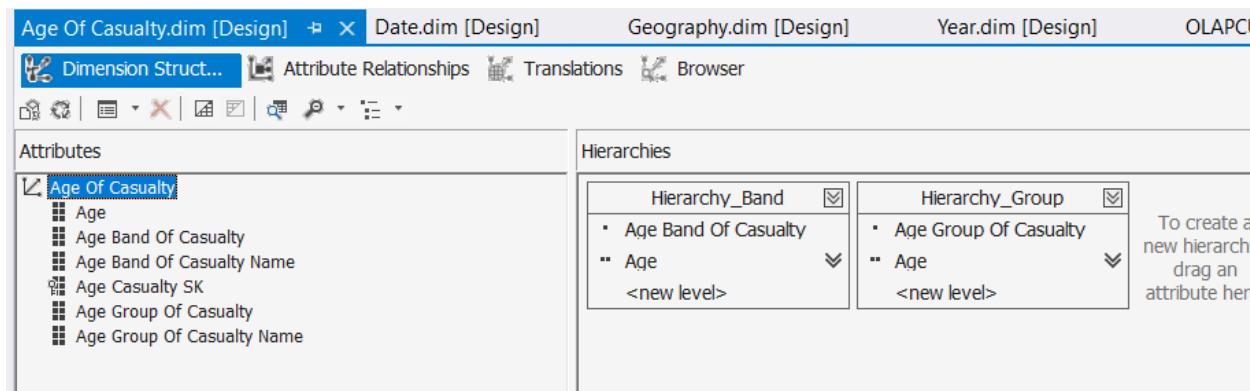
Với 4 chiều được phân cấp (**Date, Geography, Time, Age Of Casualty**) => Hierarchies

The screenshot shows the SSAS Dimension Designer interface with the Date.dim [Design] tab selected. The ribbon menu includes Dimension Struct..., Attribute Relationships, Translations, and Browser. Below the ribbon are standard toolbar icons. The left pane, titled 'Attributes', lists the Date dimension components: Date SK, Day, Month, Quarter, and Year. The right pane, titled 'Hierarchies', displays a hierarchy structure with levels: Year, Quarter, Month, Day, and Date SK, along with a placeholder for creating new levels. A note on the right says: "To create a new hierarchy, drag an attribute here."

Hierarchy Calendar của Dim Date

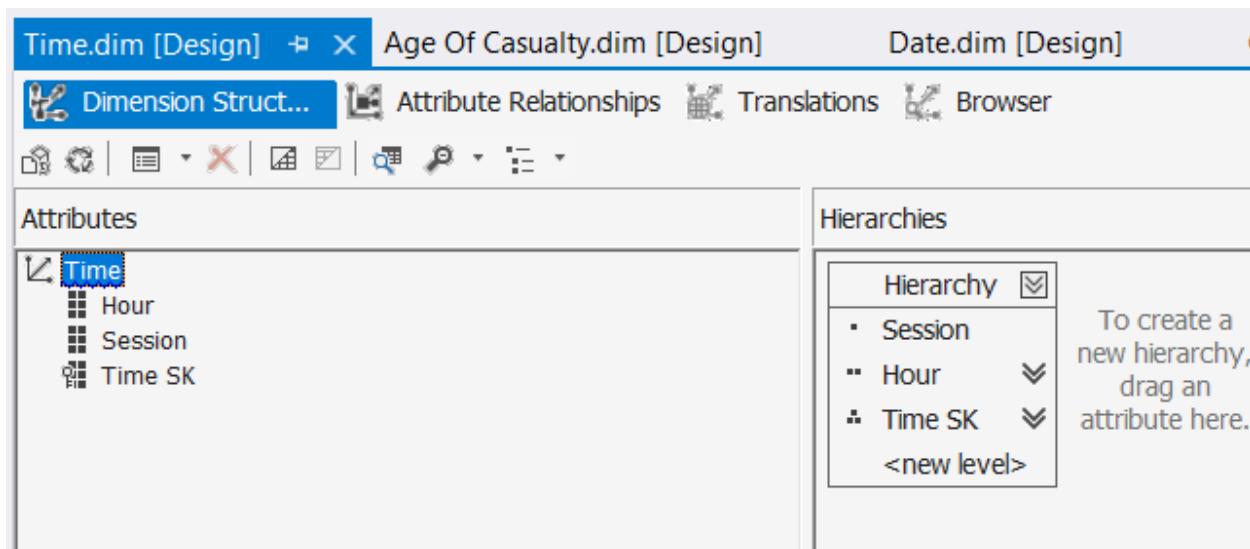
The screenshot shows the SSAS Dimension Designer interface with the Geography.dim [Design] tab selected. The ribbon menu includes Dimension Struct..., Attribute Relationships, Translations, and Browser. Below the ribbon are standard toolbar icons. The left pane, titled 'Attributes', lists the Geography dimension components: City, Country Name, County, Geography SK, and Region Name. The right pane, titled 'Hierarchies', displays a hierarchy structure with levels: Country Name, Region Name, County, and City, along with a placeholder for creating new levels. A note on the right says: "To create a new hierarchy, drag an attribute here."

Hierarchy Calendar của Dim Geography



The screenshot shows the SSAS Dimension Designer interface. The top navigation bar includes tabs for 'Age Of Casualty.dim [Design]', 'Date.dim [Design]', 'Geography.dim [Design]', 'Year.dim [Design]', and 'OLAPCU'. Below the tabs are buttons for 'Dimension Struct...', 'Attribute Relationships', 'Translations', and 'Browser'. A toolbar with various icons is located at the top. The main area is divided into two sections: 'Attributes' and 'Hierarchies'. The 'Attributes' section lists items under 'Age Of Casualty', including 'Age', 'Age Band Of Casualty', 'Age Band Of Casualty Name', 'Age Casualty SK', 'Age Group Of Casualty', and 'Age Group Of Casualty Name'. The 'Hierarchies' section shows two hierarchies: 'Hierarchy_Band' and 'Hierarchy_Group'. 'Hierarchy_Band' contains levels 'Age Band Of Casualty' and 'Age'. 'Hierarchy_Group' contains levels 'Age Group Of Casualty' and 'Age'. A tooltip on the right side of the 'Hierarchies' section says: 'To create a new hierarchy, drag an attribute here.'

Hierarchy Calendar của Dim Age of Casualty



The screenshot shows the SSAS Dimension Designer interface. The top navigation bar includes tabs for 'Time.dim [Design]', 'Age Of Casualty.dim [Design]', 'Date.dim [Design]', and 'OLAPCU'. Below the tabs are buttons for 'Dimension Struct...', 'Attribute Relationships', 'Translations', and 'Browser'. A toolbar with various icons is located at the top. The main area is divided into two sections: 'Attributes' and 'Hierarchies'. The 'Attributes' section lists items under 'Time', including 'Hour', 'Session', and 'Time SK'. The 'Hierarchies' section shows a single hierarchy 'Hierarchy' with levels 'Session', 'Hour', and 'Time SK'. A tooltip on the right side of the 'Hierarchies' section says: 'To create a new hierarchy, drag an attribute here.'

Hierarchy Calendar của Dim Time

10.Khai thác dữ liệu

10.1 MDX

R1. Thông 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

```
--R1: Thông 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 [Casualty Severity].[Name Severity].[Name Severity] on columns,
non empty [Local Authority District].[Name Local Authority District].[Name Local Authority District] on rows
from [OLAPCUBE_TTKD10]
where [Measures].[Numberof Casualty]
```

97 % ▶

	Fatal	Serious	Slight
Adur	5	102	470
Allerdale	29	145	1087
Amber Valley	5	144	1056
Arun	18	209	980
Ashfield	12	175	1165
Ashford	19	164	1363
Aylesbury Vale	21	206	1422
Babergh	13	107	1035
Barking and Dagenham	15	166	2128
Barnet	27	453	4648
Barnsley	23	267	2113
Barrow-in-Furness	5	49	507
Basildon	7	199	1497
Basingstoke and Deane	17	244	1045
Bassetlaw	23	207	1155
Bath and North East Somerset	16	101	1265
Bedford	14	182	1509
Bexley	11	148	1953

R2. Thông kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa Phương (Local_Authority_(District)) theo các Quý trong từng năm.

```
--R2: Thông 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 [Local Authority District].[Name Local Authority District].[Name Local Authority District]*
*[Casualty Severity].[Name Severity].[Name Severity] on rows,
non empty {[Date].[Hierarchy].[Quarter]*[Date].[Year].[Year]} on columns
from [OLAPCUBE_TTKD10]
where [Measures].[Numberof Casualty]
```

97 % ▶

		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
		2011	2011	2011	2011	2012	2012	2012	2012	2013	2013	2013	2013	2014	2014	2014	2014
Adur	Fatal	(null)	(null)	(null)	(null)	1	3	(null)	(null)	1	(null)	5	4	2	6	2	2
Adur	Serious	9	6	10	4	9	8	9	7	10	9	5	4	2	6	2	2
Adur	Slight	48	24	36	42	33	37	37	35	24	32	38	42	10	17	6	9
Allerdale	Fatal	4	1	1	1	4	2	(null)	2	1	2	1	3	2	1	1	1
Allerdale	Serious	7	24	9	14	9	9	1	4	5	13	11	5	6	11	7	10
Allerdale	Slight	67	59	61	76	95	71	3	60	67	48	96	81	80	73	79	71
Amber Valley	Fatal	1	(null)	(null)	1	(null)	(null)	(null)	2	(null)	(null)	1	(null)	(null)	(null)	(null)	(null)
Amber Valley	Serious	9	9	11	10	13	14	10	13	7	8	10	8	4	1	9	8
Amber Valley	Slight	98	88	98	95	75	84	76	83	79	52	65	78	27	12	23	25
Arun	Fatal	1	1	3	1	1	(null)	1	2	(null)	1	2	1	1	(null)	(null)	3
Arun	Serious	14	13	17	18	8	13	22	19	7	19	17	21	4	4	6	7
Arun	Slight	55	72	57	94	63	59	79	60	63	84	67	111	15	29	40	32
Ashfield	Fatal	2	(null)	2	(null)	2	(null)	1	(null)	(null)	2	2	(null)	(null)	1	(null)	(null)
Ashfield	Serious	16	18	12	10	14	19	14	16	3	15	13	13	3	4	4	1
Ashfield	Slight	98	82	95	79	83	82	81	85	108	67	73	83	34	42	44	29
Ashford	Fatal	2	2	(null)	2	2	1	2	4	2	(null)	1	(null)	(null)	1	(null)	(null)
Ashford	Serious	8	9	12	4	7	13	25	17	8	10	27	12	6	3	3	(null)
Ashford	Slight	86	105	105	103	86	92	115	109	87	111	144	115	23	27	29	26

R3. Thông kê số lượng người tử vong theo Giới Tính, Loại Nạn Nhân (Casualty Type) và Nhóm Tuổi (Age_Band_of_Casualty) theo các năm

```
--R3: 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
select non empty
[Sex Of Casualty].[Name].[Name] *
[Casualty Type].[Name Casualty Type].[Name Casualty Type]*
[Age Of Casualty].[Hierarchy_Band].[Age Band Of Casualty] on rows,
non empty [Date].[Hierarchy].[Year] on columns
from [OLAPCUBE_TTKD10]
where ([Measures].[Numberof Casualty],[Casualty Severity].[Casualty Severity SK].&[1])
```

88 %

			2011	2012	2013	2014
Female	Car occupant		1	3	8	3
Female	Car occupant		10	23	23	25
Female	Car occupant		11	55	39	56
Female	Car occupant		2	(null)	2	(null)
Female	Car occupant		3	3	2	1
Female	Car occupant		4	35	21	34
Female	Car occupant		5	31	21	24
Female	Car occupant		6	33	21	29
Female	Car occupant		7	23	23	16
Female	Car occupant		8	22	14	23
Female	Car occupant		9	16	20	18
Female	Cyclist		10	4	1	2
Female	Cyclist		11	(null)	(null)	1
Female	Cyclist		3	2	(null)	(null)
Female	Cyclist		4	(null)	(null)	2
Female	Cyclist		5	3	1	3
Female	Cyclist		6	3	1	5
Female	Cyclist		7	4	1	3
Female	Cyclist		8	3	1	1
Female	Cyclist		9	1	1	(null)

R4. 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.

```
--R4: 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 [Time].[Hierarchy].[Session] on rows,
non empty [Accident Severity].[Name Severity].[Name Severity] on columns
from [OLAPCUBE_TTKD10]
where [Measures].[Numberof Accidents]
```

97 %

	Fatal	Serious	Slight
Afternoon	1350	20127	129295
Evening	1689	20634	121653
Morning	1324	16573	114066
Night	587	3811	15342

R5. Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng, Vùng (Urban_or_Rural_Area), và Kiểu Đường (Road Type) trong các năm.

```
--R5: 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
[Urban Or Rural].[Name Urban Or Rural Area].[Name Urban Or Rural Area]
*
[Road Type].[Name Road Type].[Name Road Type] on rows,
non empty [Accident Severity].[Name Severity].[Name Severity] on columns
from [OLAPCUBE_TTKD10]
where [Measures].[Numberof Accidents]
```

97 %

		Fatal	Serious	Slight
Rural	Dual carriageway	653	3761	25566
Rural	One way street	10	101	515
Rural	Roundabout	38	1180	10268
Rural	Single carriageway	2422	18929	82491
Rural	Slip road	27	284	2558
Rural	Unknown	6	52	442
Urban	Dual carriageway	279	3577	28052
Urban	One way street	41	947	6645
Urban	Roundabout	41	1793	17583
Urban	Single carriageway	1422	30281	203703
Urban	Slip road	8	147	1601
Urban	Unknown	3	93	932

R6. 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-...

```
--R6: 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-...
select non empty
[Casualty Type].[Name Casualty Type].[Name Casualty Type]*
[Age Of Casualty].[Hierarchy_Group].[Age Group Of Casualty] * [Age Of Casualty].[Name Casualty Name].[Age Group Of Casualty Name] on rows
non empty [Casualty Severity].[Name Severity].[Name Severity] on columns
from [OLAPCUBE_TTKD10]
where [Measures].[Numberof Casualty]
```

97 %

			Fatal	Serious	Slight
Agricultural vehicle occupant	-1	Data missing or out of range	(null)	(null)	3
Agricultural vehicle occupant	2	Young adult	(null)	5	20
Agricultural vehicle occupant	3	Adult	7	39	183
Agricultural vehicle occupant	4	60 and over	5	9	41
Bus or coach occupant (17 or more pass seats)	1	Children	(null)	22	828
Bus or coach occupant (17 or more pass seats)	-1	Data missing or out of range	(null)	30	1217
Bus or coach occupant (17 or more pass seats)	2	Young adult	(null)	36	1835
Bus or coach occupant (17 or more pass seats)	3	Adult	6	332	6827
Bus or coach occupant (17 or more pass seats)	4	60 and over	18	495	4686
Car occupant	1	Children	20	246	5636
Car occupant	-1	Data missing or out of range	(null)	258	4741
Car occupant	2	Young adult	110	1418	23779
Car occupant	3	Adult	1545	16280	250461
Car occupant	4	60 and over	695	4738	38857
Cyclist	1	Children	1	34	225
Cyclist	-1	Data missing or out of range	(null)	162	1292
Cyclist	2	Young adult	30	1242	7826
Cyclist	3	Adult	225	7441	39569
Cyclist	4	60 and over	83	877	2562

R7. Tổng hợp số lượng tai nạn theo Mục Đích Hành Trình (Journey Purpose) và Loại Phương Tiện (Vehicle_Type)

--R7: Tổng hợp số lượng tai nạn theo Mục Đích Hành Trình (Journey Purpose) và Loại Phương Tiện (Vehicle_Type)

```
select non empty
[Journey Purpose].[Name Journey Purpose].[Name Journey Purpose] on rows,
non empty [Vehicle Type].[Name Vehicle Type].[Name Vehicle Type] on columns
from [OLAPCUBE_TTKD10]
where [Measures].[Numberof Vehicle]
```

107 %

	Agricultural vehicle	Bus or coach (17 or more pass seats)	Car	Data missing or out of range	Electric motorcycle	Goods 7.5 tonnes mgw and over
Commuting to/from work	25	85	53987	(null)	(null)	133
Data missing or out of range	(null)	(null)	3	(null)	(null)	(null)
Journey as part of work	1244	16709	55702	1	(null)	12903
Not known	332	2329	450591	77	10	1992
Other	18	45	14942	1	(null)	27
Pupil riding to/from school	1	12	916	(null)	(null)	3
Taking pupil to/from school	(null)	182	7675	(null)	(null)	5

R9. Thống kê số lượng tai nạn theo Mức Độ Nghiêm Trọng, Loại Phương Tiện (Vehicle Type), Built-up Road trong các năm.

--R9: Thống kê số lượng tai nạn theo Mức Độ Nghiêm Trọng, Loại Phương Tiện (Vehicle Type), Built-up Road trong các năm.

```
select non empty [Accident Severity].[Name Severity].[Name Severity] on rows,
non empty [Built Up Road Type].[Name Built Up Road Type].[Name Built Up Road Type] on columns
from [OLAPCUBE_TTKD10]
where [Measures].[Numberof Accidents]
```

118 %

	Built up Road	Non Built up Road
Fatal	2355	2595
Serious	15100	46045
Slight	72453	307903

R11. Định nghĩa fact Variance để tính mức độ tăng giảm của TNGT theo đơn vị phần trăm qua các năm

--R11: Định nghĩa fact Variance để tính mức độ tăng giảm của TNGT theo đơn vị phần trăm qua các năm (Calculated Measures)

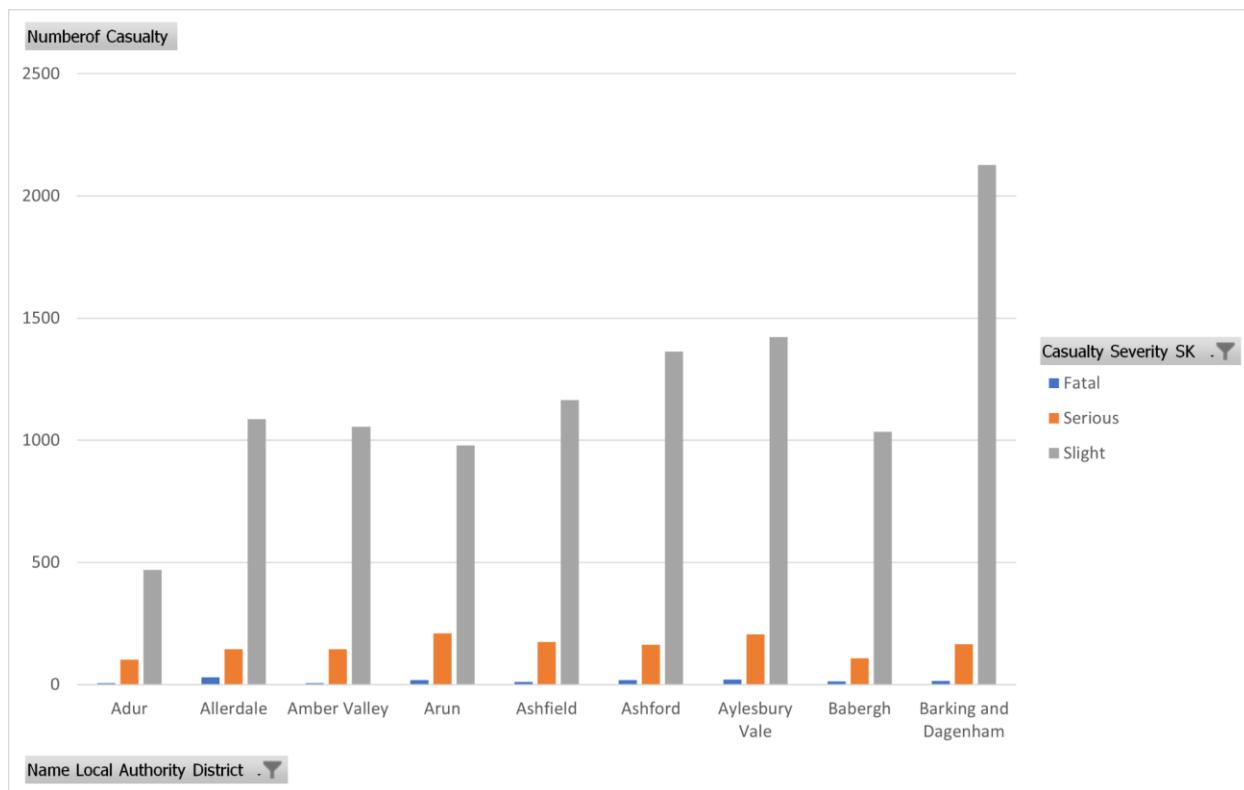
```
with member [Measures].[last Numberof Accidents] as
(ParallelPeriod([Date].[Hierarchy].[Year]
, 0
,[Date].[Hierarchy].CurrentMember
,[Measures].[Numberof Accidents])
/
(ParallelPeriod([Date].[Hierarchy].[Year]
, 1
,[Date].[Hierarchy].CurrentMember
,[Measures].[Numberof Accidents]))
select
non empty [Measures].[last Numberof Accidents] on rows,
non empty[Date].[Hierarchy].[Year] on columns
from [OLAPCUBE_TTKD10]
```

107 %

	2011	2012	2013	2014
last Numberof Accidents	inf	0.892353411135751	1.07833522993507	0.497664385368655

10.2 Report (Visualize)

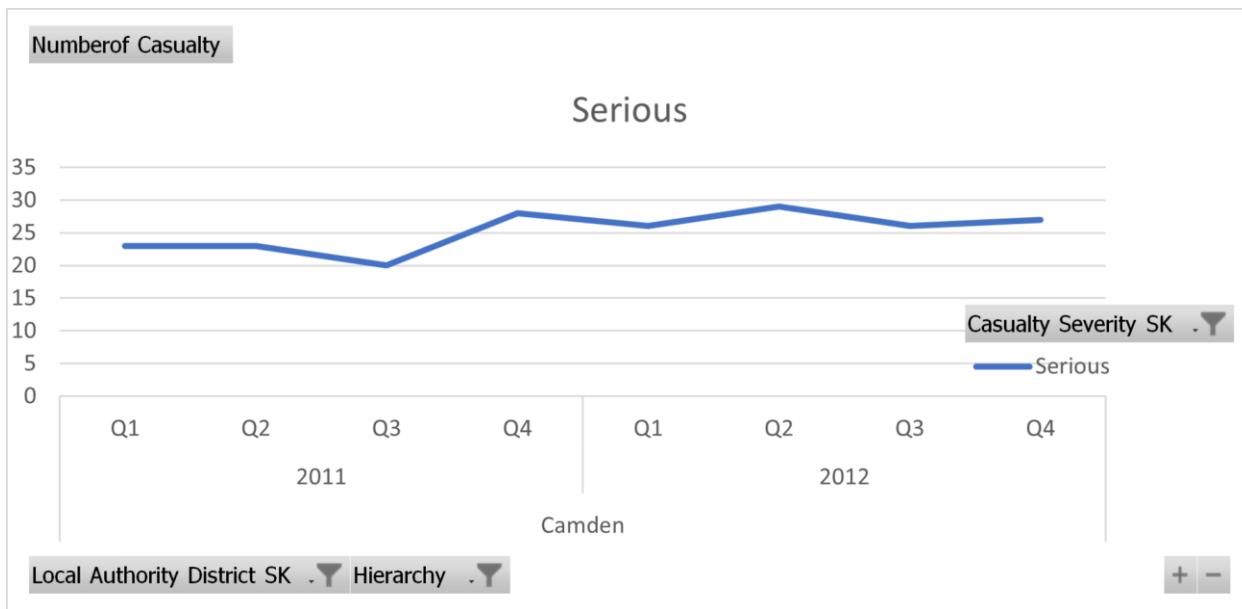
R1. Thông kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng (Fatal, Serious, Slight) ở các Địa phương (Local_Authority_(District)) trong tất cả các năm



Nhận xét: Dựa vào đồ thị trên, ta thấy ở Barking and Dagenham có số lượng nạn nhân ở mức độ nhẹ nhiều nhất so với các địa phương khác

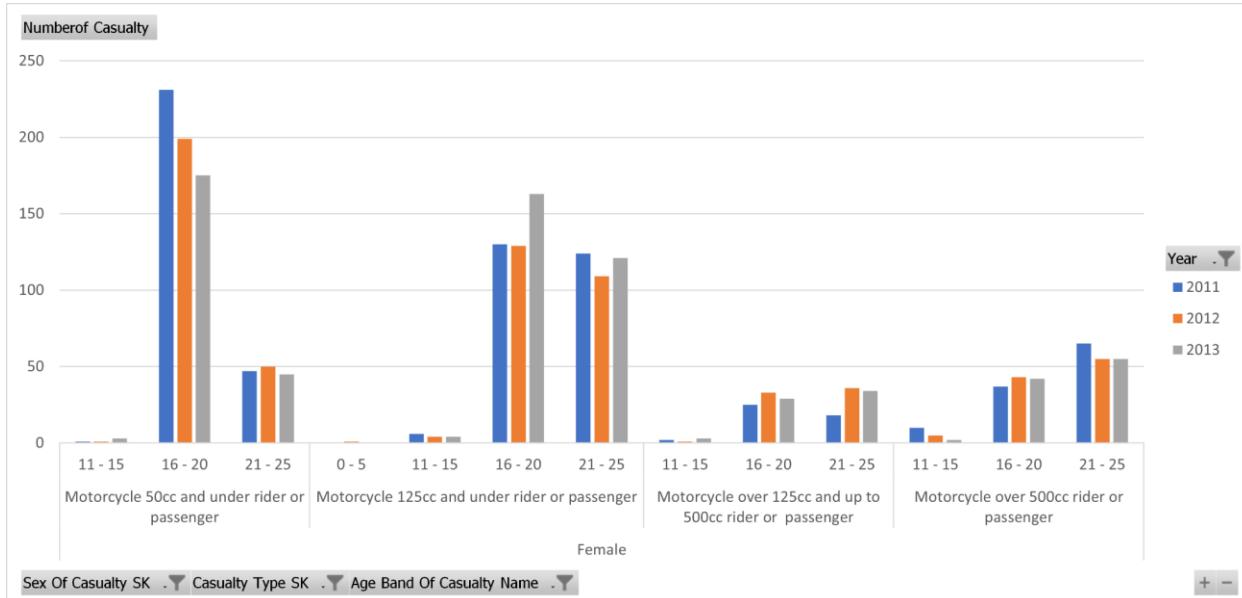
So với các địa phương khác, Adur có số lượng nạn nhân tai nạn giao thông ở mọi mức độ thấp nhất.

R2. Thông kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa Phương (Local_Authority_(District)) theo các Quý trong từng năm.



Nhận xét. Filter số lượng nạn nhân qua các quý trong từng năm của địa phương Camden
 Ta thấy rằng, số lượng nạn nhân tai nạn giao thông ở Camden không thay đổi nhiều qua
 các quý trong năm, và có xu hướng tăng vào quý 4 mỗi năm.

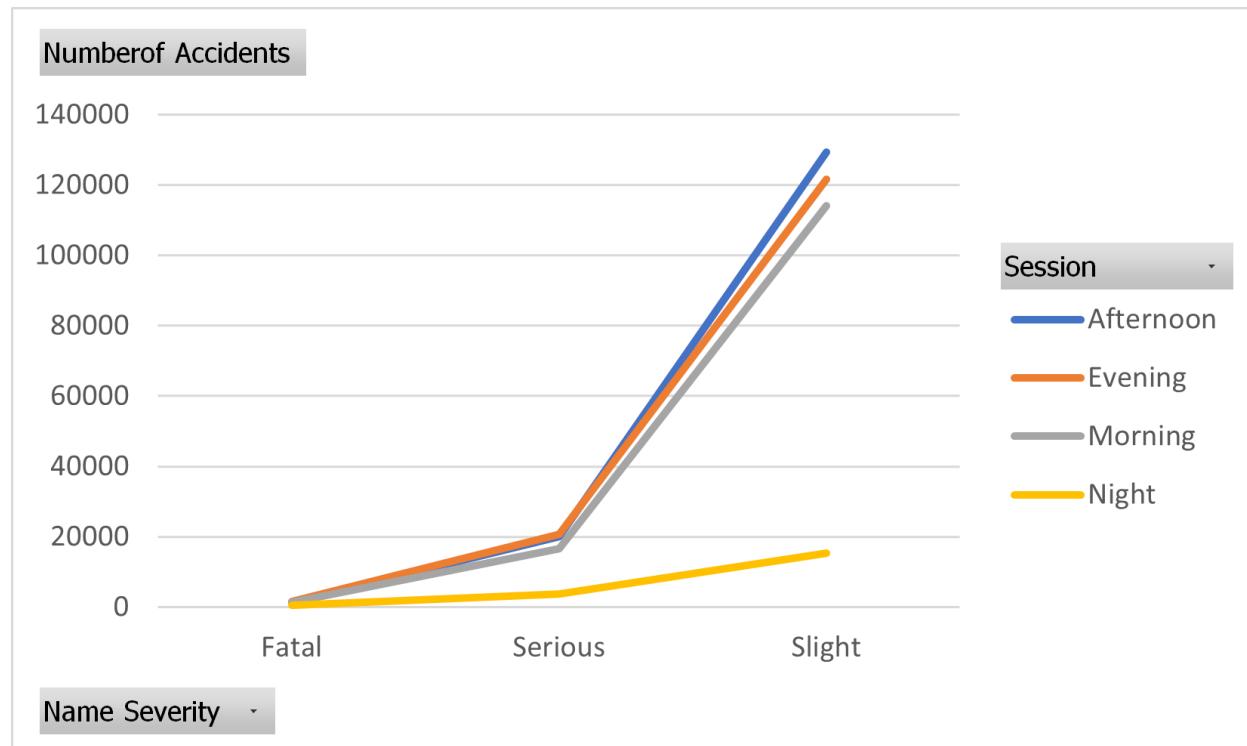
R3. Thống kê số lượng người tử vong theo Giới Tính, Loại Nạn Nhân (Casualty Type) và Nhóm Tuổi (Age_Band_of_Casualty) theo các năm



Nhận xét: Năm 2011, số lượng nạn nhân tử vong rất nhiều, nhiều hơn so với các năm khác, đa số đều ở nhóm tuổi 16-20.

Số lượng người tử vong do tai nạn giao thông, chạy xe từ 50cc đến 125cc rất nhiều.

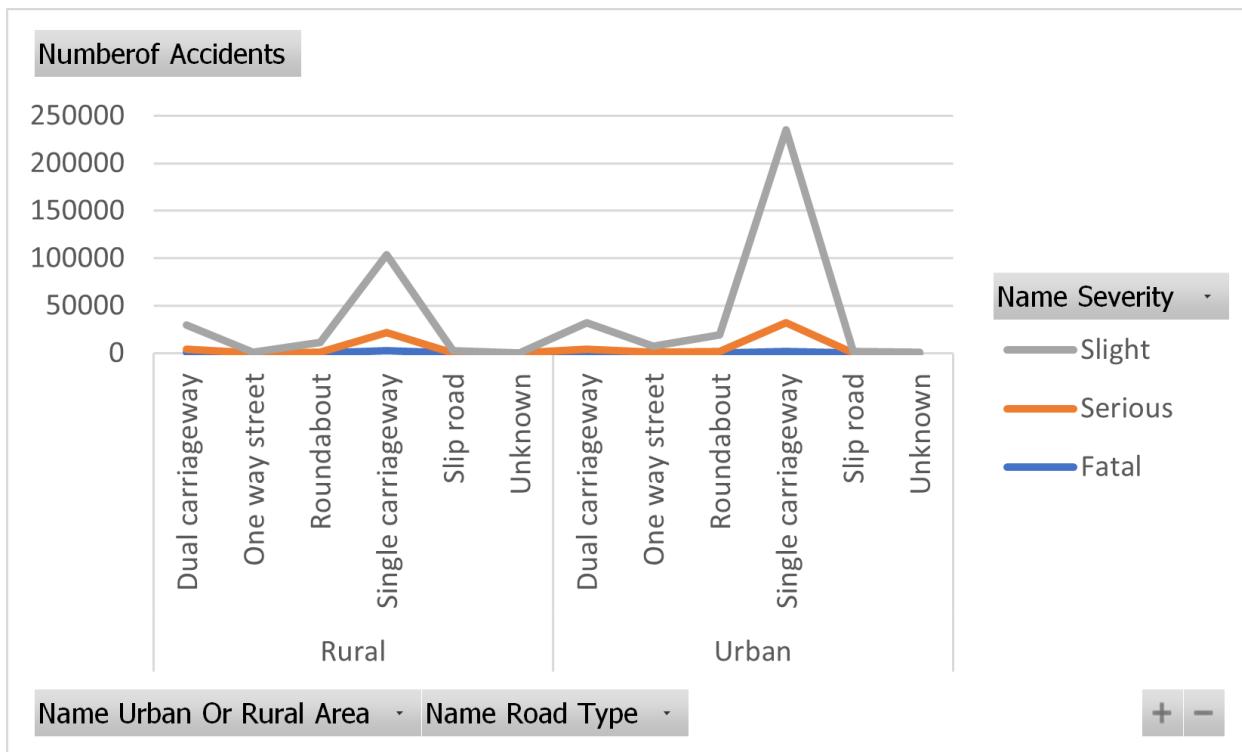
R4. Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng và Thời Điểm Trong Ngày (Morning: 5am-12pm, Afternoon: 12pm-5pm, Evening: 5pm-9pm, Night: 9pm-5am) trong các năm.



Nhận xét: Đa số các tai nạn giao thông đều xảy ra vào buổi sáng, chiều và chiều tối. Nhất là buổi sáng, số lượng tai nạn giao thông ở mức độ nhẹ là cao nhất.

Vào buổi đêm thì ít xảy ra tai nạn giao thông hơn.

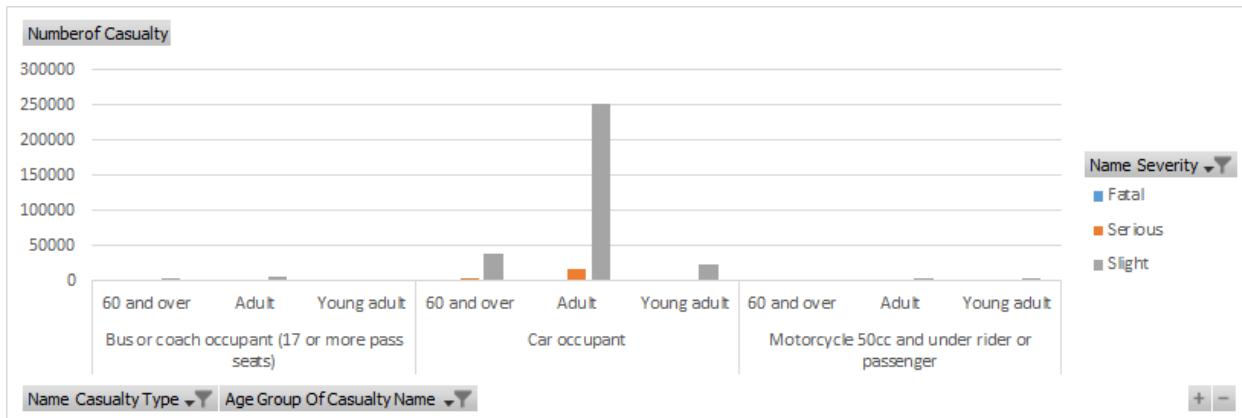
R5. Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng, Vùng (Urban_or_Rural_Area), và Kiểu Đường (Road Type) trong các năm.



Nhân xét: Ở khu vực đô thị, số lượng tai nạn giao thông ở loại đường Single carriageway nhiều nhất so với các loại đường khác, và đa phần đều mức độ nhẹ và nghiêm trọng

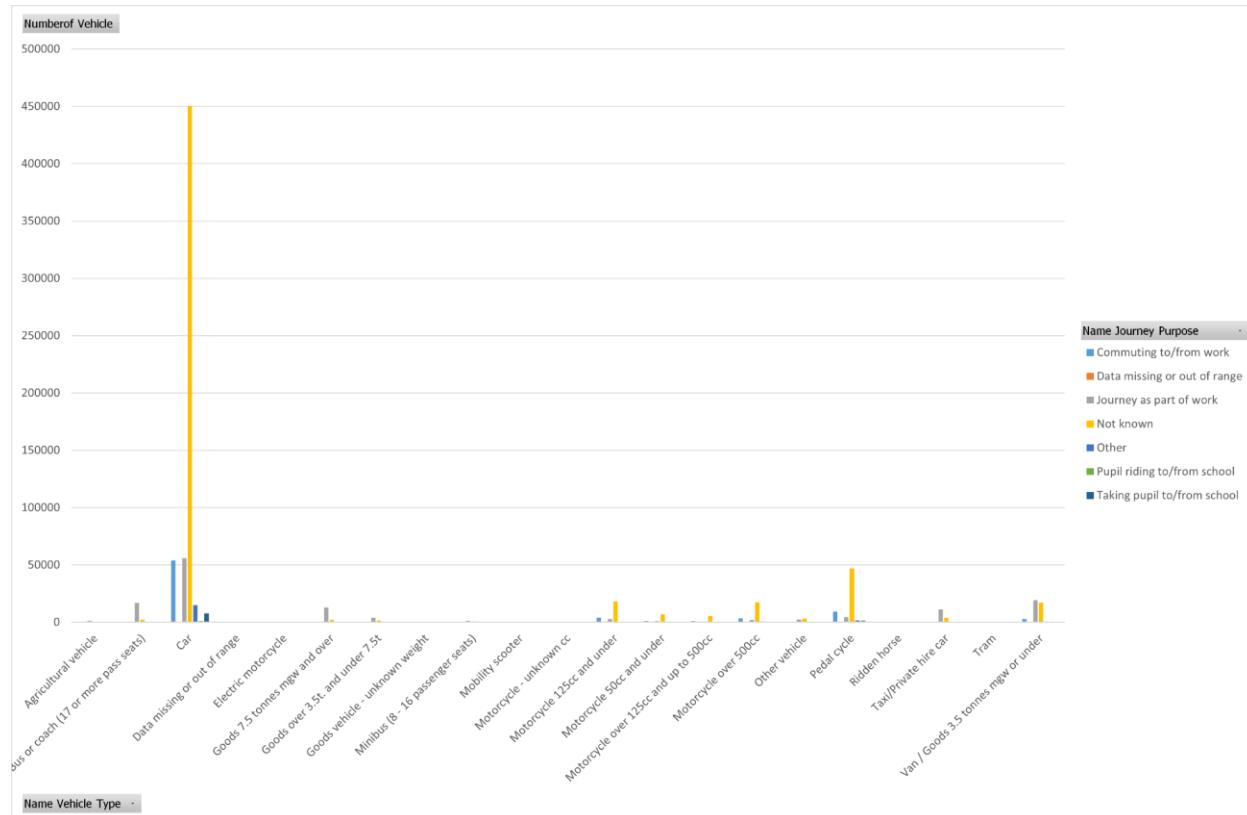
Ở khu vực nông thôn, số lượng tai nạn giao thông thấp hơn so với đô thị, và đa số tai nạn ở loại đường Single carriageway

R6. 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-...



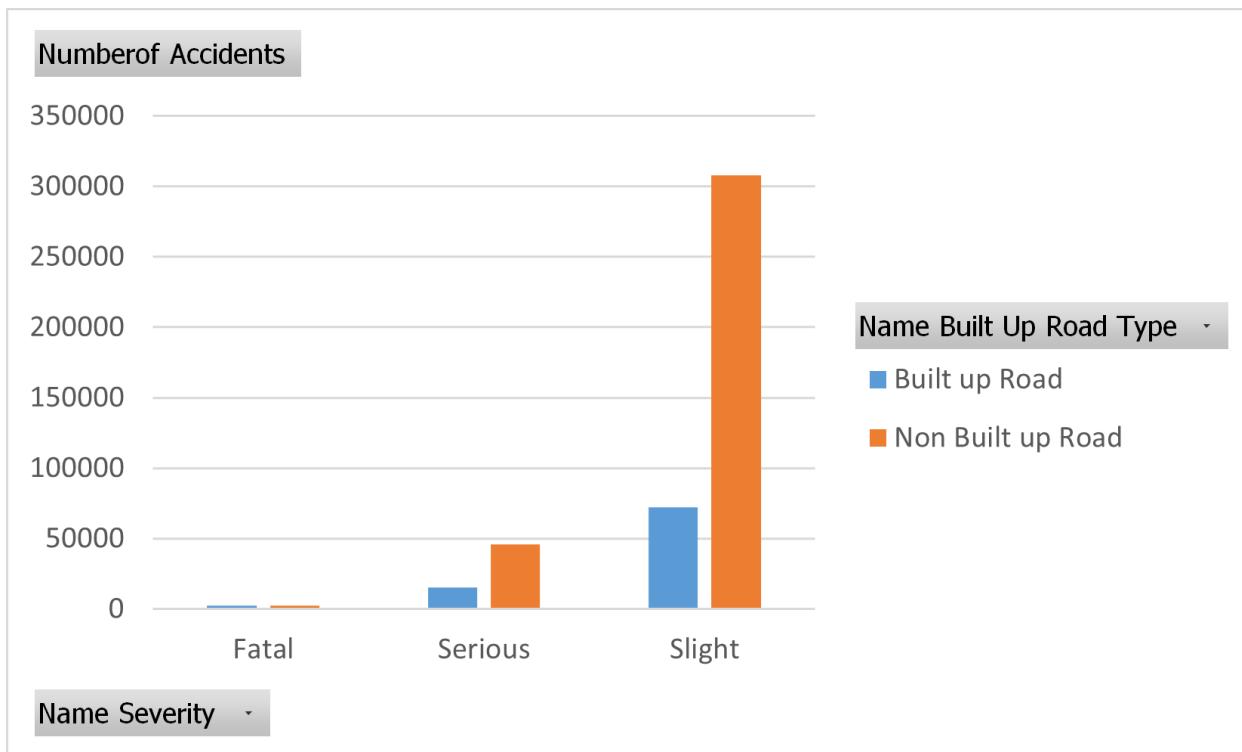
Nhận xét: Số lượng nạn nhân ở độ tuổi Adult (18-59) nhiều nhất ở mỗi loại nạn nhân, nhất là loại **Car occupant**.

R7. Tổng hợp số lượng tai nạn theo Mục Đích Hành Trình (Journey Purpose) và Loại Phương Tiện (Vehicle_Type)



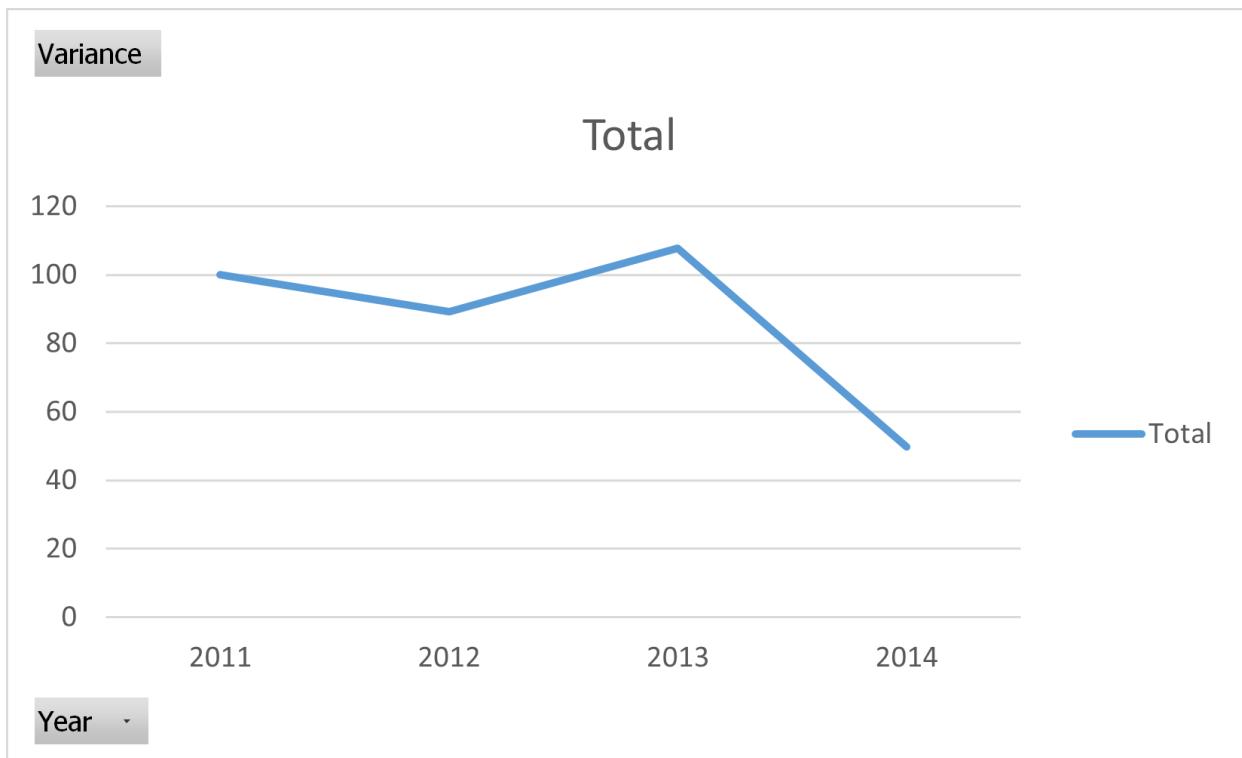
Nhận xét: Đa phần các vụ tai nạn giao thông đều không biết được mục đích hành trình. Số lượng xe hơi xảy ra tai nạn nhiều nhất so với các loại phương tiện khác

R9. Thống kê số lượng tai nạn theo Mức Độ Nghiêm Trọng, Loại Phương Tiện (Vehicle Type), Built-up Road trong các năm.



Nhận xét. Các vụ tai nạn xảy ra trên các con đường có tốc độ giới hạn từ 50 mph (Non Built-up road), nhất là số lượng tai nạn giao thông ở mức độ nhẹ rất nhiều

R11. Định nghĩa fact Variance để tính mức độ tăng giảm của TNGT theo đơn vị phần trăm qua các năm



Nhận xét: Ta thấy tỉ lệ đều ở số dương, chứng tỏ tai nạn giao thông tăng qua các năm, chứ không có giảm.

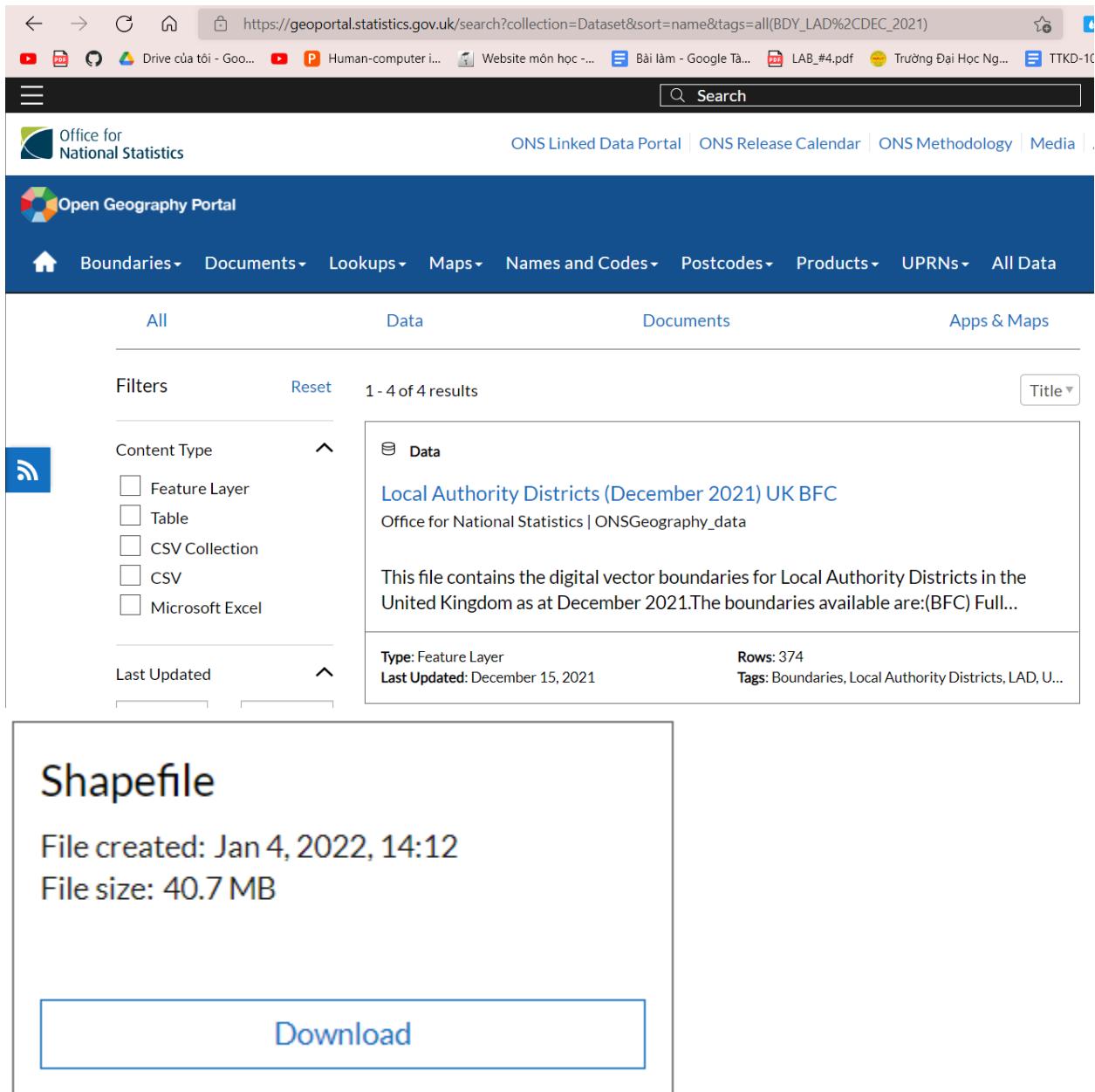
Đặc biệt là năm 2013, số lượng tai nạn giao thông tăng hơn gấp 2 lần (hơn 100%) so với năm 2012.

Năm 2014, số lượng tai nạn giao thông vẫn tăng so với năm 2013, nhưng tỉ lệ tăng khoảng 50% -> tỉ lệ tăng giảm.

R12. Dùng regional map để biểu diễn trực quan (bằng màu sắc) số lượng TNGT ở các vùng trong năm.

- ❖ Các bước thực hiện vẽ region Map:

Bước 1: Tải local authority districts UK Map từ trang Geoportal. Lựa chọn tải về dưới dạng ShapeFile.



The screenshot shows the 'Open Geography Portal' interface. In the search bar, the URL is https://geoportal.statistics.gov.uk/search?collection=Dataset&sort=name&ttags=all(BDY_LAD%2CDEC_2021). The results page displays a single item: 'Local Authority Districts (December 2021) UK BFC'. The item details include: Type: Feature Layer, Last Updated: December 15, 2021, Rows: 374, and Tags: Boundaries, Local Authority Districts, LAD, U... A large blue button labeled 'Download' is prominently displayed at the bottom of the result card.

Bước 2: Truy cập [mapshaper](#) (MapShaper.org). Tải các file đã download về từ GeoPortal.

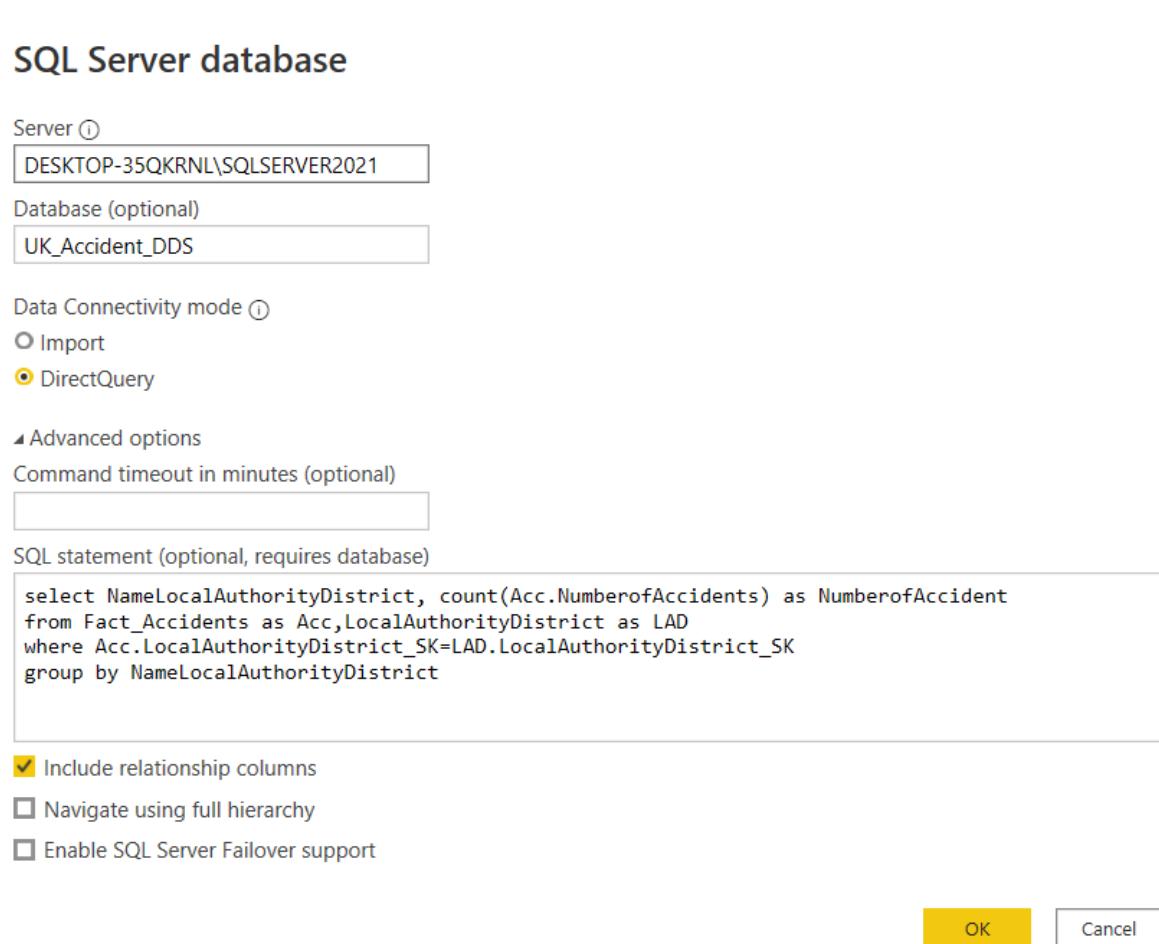


Bước 3: Export Map UK. Xuất dưới dạng TopoJson

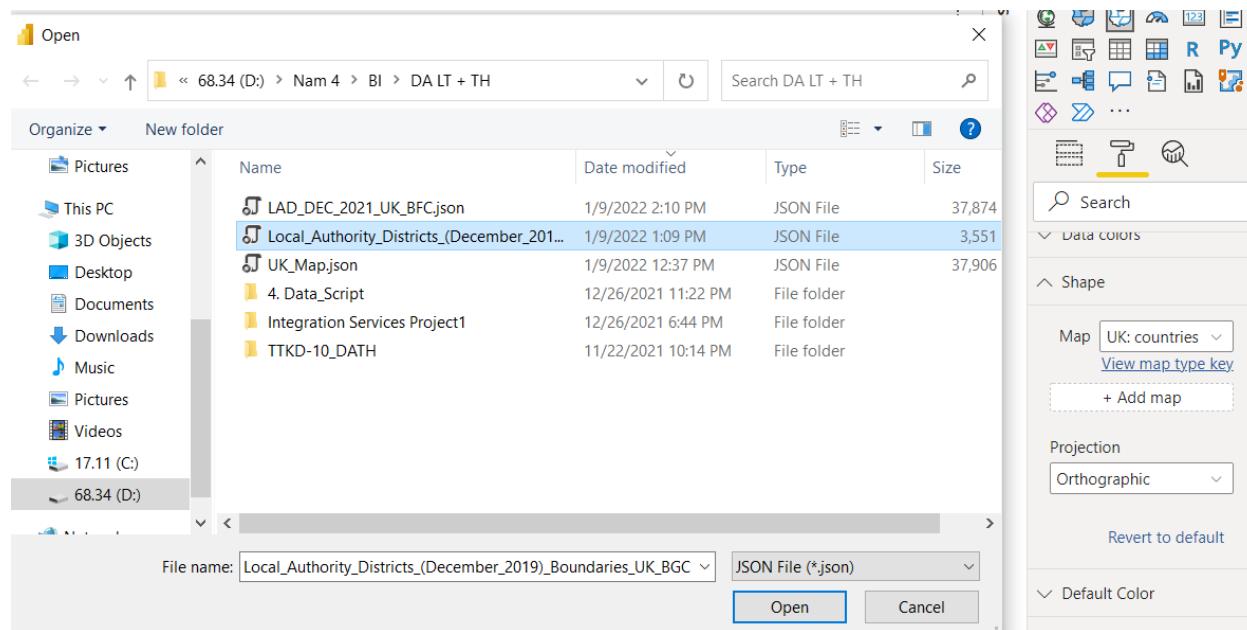
Bước 4: Để vẽ Map trong powerBI. Ta cần enable tính năng Shape Map.

Link enable tính năng: [Use Shape maps in Power BI Desktop \(Preview\) - Power BI | Microsoft Docs.](https://powerbi.microsoft.com/en-us/blog/use-shape-maps-in-power-bi-desktop-preview/)

Bước 5: Load data từ Sql Server bằng câu truy vấn:



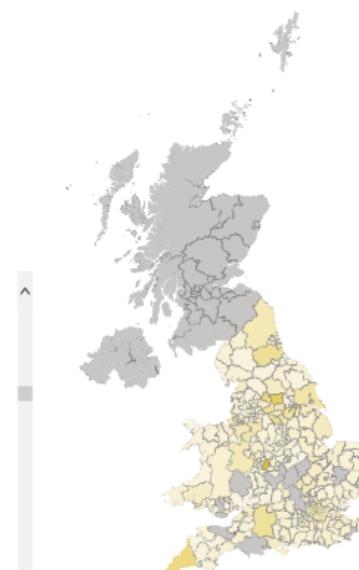
Bước 6: Chọn **NameLocalAuthorityDistrict** và sau đó Load Map vào powerBI. Chọn NumberofAccident để thể hiện phân bố mật độ trên map.



Kết quả:

NumberofAccidents by NameLocalAuthorityDistrict

NameLocalAuthorityDistrict	NumberofAccidents
Carlisle	992
Carmarthenshire	1391
Castle Point	552
Central Bedfordshire	2033
Ceredigion	565
Charnwood	1028
Chelmsford	1190
Cheltenham	531
Cherwell	1361
Cheshire East	2556
Cheshire West and Chester	2100
Total	446313



Nhận xét:

- + Ta thấy map thể hiện số TNGT ở các vùng ở UK.
- + Ở các vùng có số TNGT cao thì màu vàng thể hiện càng đậm. Màu càng nhạt thể hiện số TNGT càng ít.
- + Các vùng thể hiện màu xám thì có thể do PostCode dữ liệu thiếu các vùng cần thiết hoặc không có vụ tai nạn ở các vùng này.

10.3 Mining

a. Mô tả bài toán:

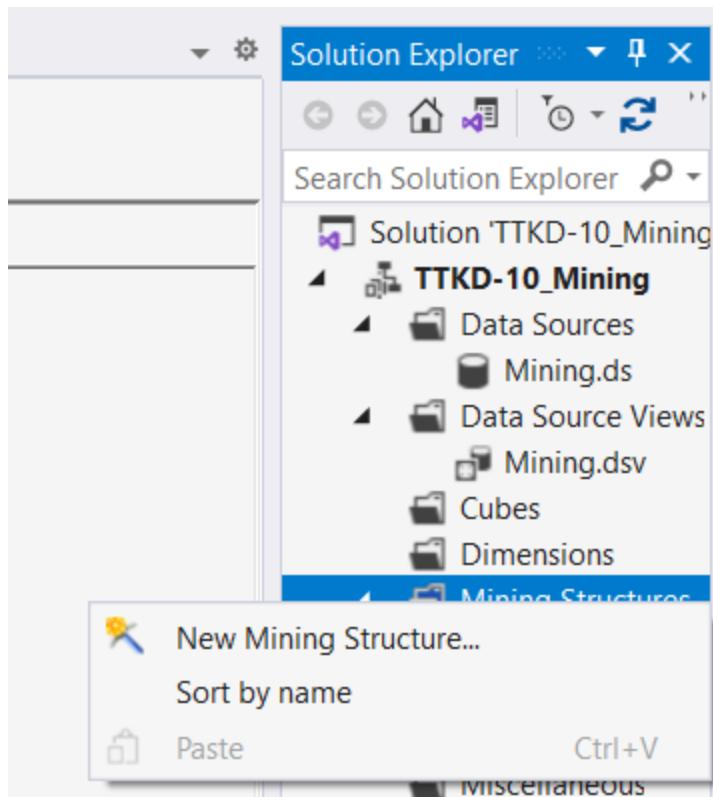
Với cơ sở dữ liệu Accident_UK, xây dựng mô hình cây quyết định hỗ trợ việc dự đoán mức độ nghiêm trọng của một tai nạn dựa trên các yếu tố về loại đường, thời tiết, ánh sáng, bề mặt đường, ở vùng nông thôn hay thành thị và điều kiện đặc biệt.

- Cơ sở dữ liệu sử dụng: Accident_UK
- Bảng dữ liệu: Predict_AccSeverity
- Thuộc tính đầu vào (Input):
 - Name_SpecialCondition
 - Name_LightConditon
 - Name_RoadSurface
 - Name_Weather
 - UrbanRuralName
 - RoadTypeName
- Dữ liệu dự đoán (Predict): AccidentSeverityName

b. Thực hành xây dựng mô hình:

Tạo Mining Structure

- **Bước 1:** Tạo SSAS trong Visual Studio 19
- **Bước 2 :** Trong Solution Explorer, phải chuột vào Mining Structures. Chọn New Mining Structure



- **Bước 3:** Welcome to Data Mining Wizard xuất hiện, nhấn Next.

↗ Data Mining Wizard



Welcome to the Data Mining Wizard

Use this wizard to create a new mining structure and a new mining model. A mining structure is a data structure that represents discovered knowledge based on analysis of OLAP or relational data. A mining model can be used to make predictions, if supported by the data mining technique used to create the mining model.

Click Next to build a mining structure and a mining model, or Cancel to exit the wizard.

Don't show this page again

< Back

Next >

Finish >>

Cancel

- **Bước 4:** Chọn From Existing relation database or data warhouse. Chọn Next.

↗ Data Mining Wizard

— □ ×

Select the Definition Method

Select the method to be used while creating the mining structure definition.

Which method do you use to define the mining structure?

From existing relational database or data warehouse

From existing cube

Description:

This method defines a mining structure based on tables and columns from an existing relational database.

< Back Next > Finish >>| Cancel

- **Bước 5:** Tại hộp thoại Create the Data Mining Structure. Chọn Create mining structure with a mining model. Chọn Microsoft Decision Trees. Nhấn next.

↗ Data Mining Wizard

Create the Data Mining Structure

Specify if mining model should be created and select the most applicable technique.

Create mining structure with a mining model

Which data mining technique do you want to use?

Microsoft Decision Trees

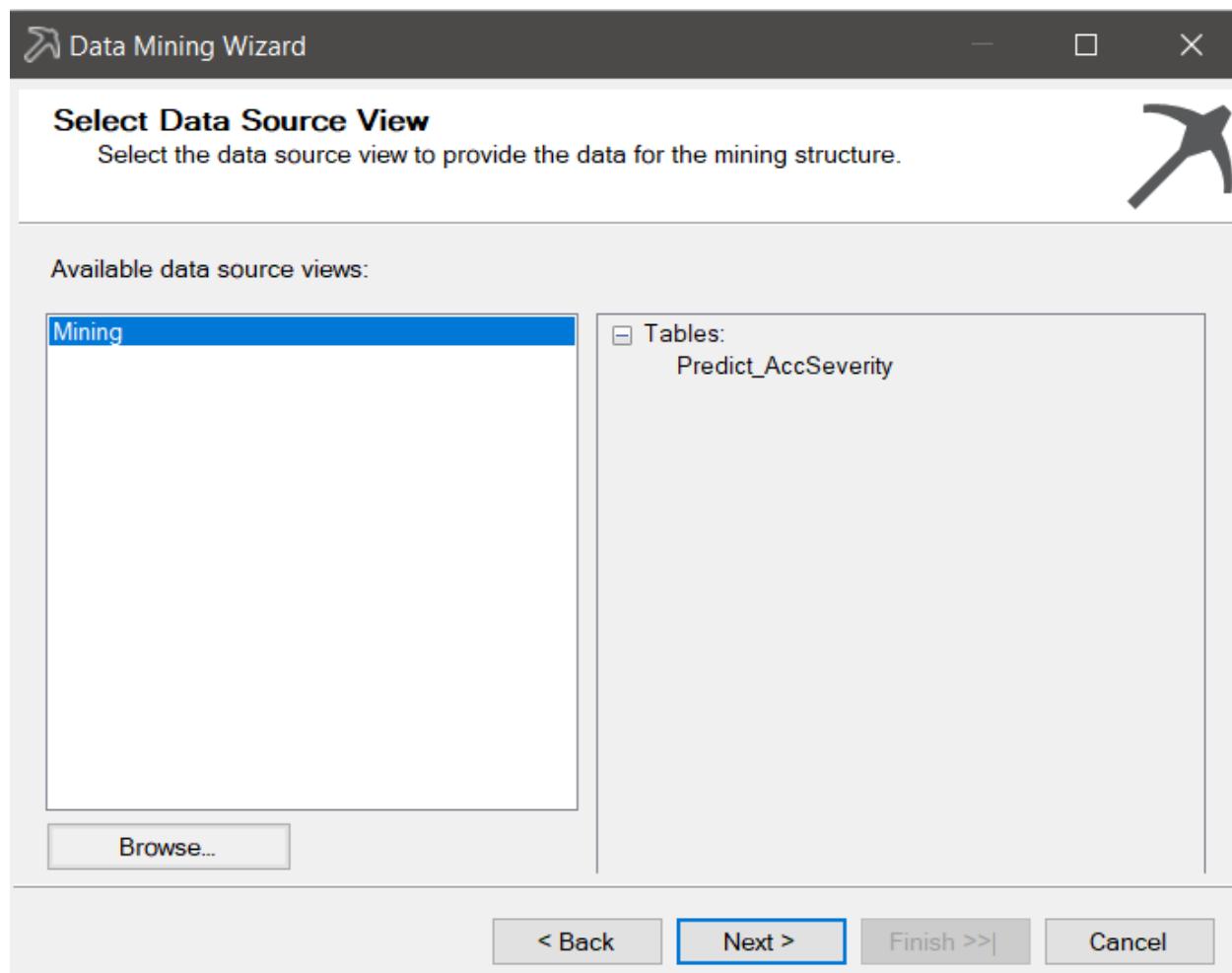
Create mining structure with no models

Description:

The Microsoft Decision Trees algorithm is a classification algorithm that works well for predictive modeling. The algorithm supports the prediction of both discrete and continuous attributes.

< Back Next > Finish >> Cancel

Bước 6: Tại hộp thoại Select Data Source View chọn Mining. Nhấn Next.



Bước 7: Tại hộp thoại Specify Table Types. Tại mục case, chon Predict_AccSeverity

Data Mining Wizard

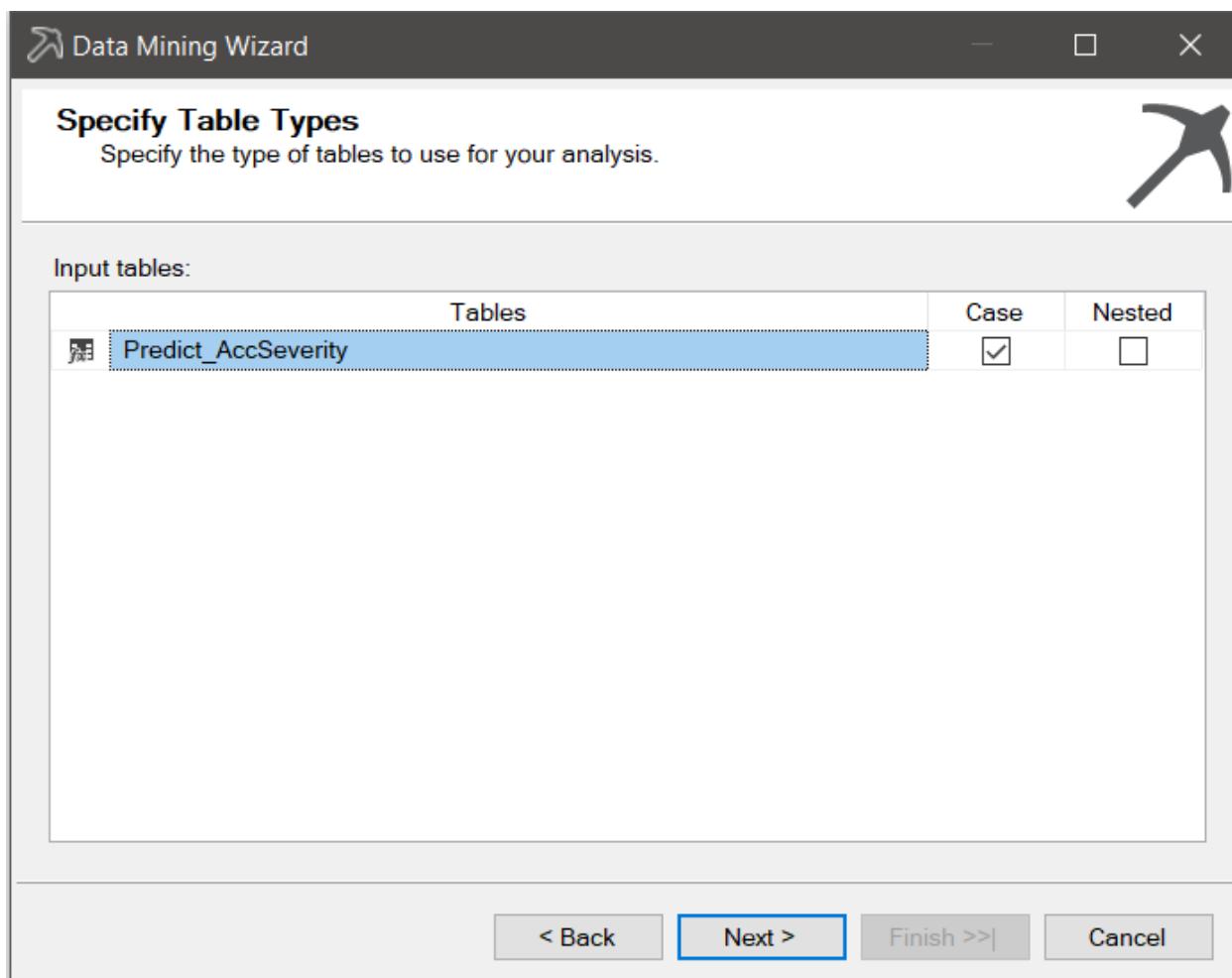
Specify Table Types

Specify the type of tables to use for your analysis.

Input tables:

Tables	Case	Nested
Predict_AccSeverity	<input checked="" type="checkbox"/>	<input type="checkbox"/>

< Back Next > Finish >> Cancel



Bước 8: Tại hộp thoại Specify the Training Data, chọn

- Key : Accident_Index
- 5 Input: Name_SpecialCondition, Name_LightConditon, Name_RoadSurface, Name_Weather, UrbanRuralName, RoadTypeName
- 1 Prediction: Name_AccidentSeverity

Data Mining Wizard

Specify the Training Data

Specify the columns used in your analysis.

Mining model structure:

	Tables/Columns	Key	<input type="checkbox"/> Input	<input type="checkbox"/> Predi...
<input checked="" type="checkbox"/>	Predict_AccSeverity	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Accident_Index	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	AccidentSeverityName	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	Name_LightConditon	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Name_RoadSurface	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Name_SpecialCondition	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Name_Weather	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	RoadTypeName	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	UrbanRuralName	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Recommend inputs for currently selected predictable:

Bước 9: Tại màn hình Specify Columns' Content and Data Type chọn kiểu Content Type và Data type như hình. Nhấn Next

Data Mining Wizard

Specify Columns' Content and Data Type

Specify mining structure columns' content and data type.

Mining model structure:

Columns	Content Type	Data Type
Accident Index	Key	Text
Accident Severity Name	Discrete	Text
Name Light Condition	Discrete	Text
Name Road Surface	Discrete	Text
Name Special Condition	Discrete	Text
Name Weather	Discrete	Text
Road Type Name	Discrete	Text
Urban Rural Name	Discrete	Text

Detect continuous or discrete for numeric columns:

Bước 10: Tạo testing set

Tại màn hình Create Testing Set:

Chọn phần trăm dữ liệu cho testing là 30%

Số trường hợp lớn nhất cho mẫu testing: 100 000 trường hợp

Data Mining Wizard

Create Testing Set

Specify the number of cases to be reserved for model testing.

Percentage of data for testing: %

Maximum number of cases in testing data set:

Description:

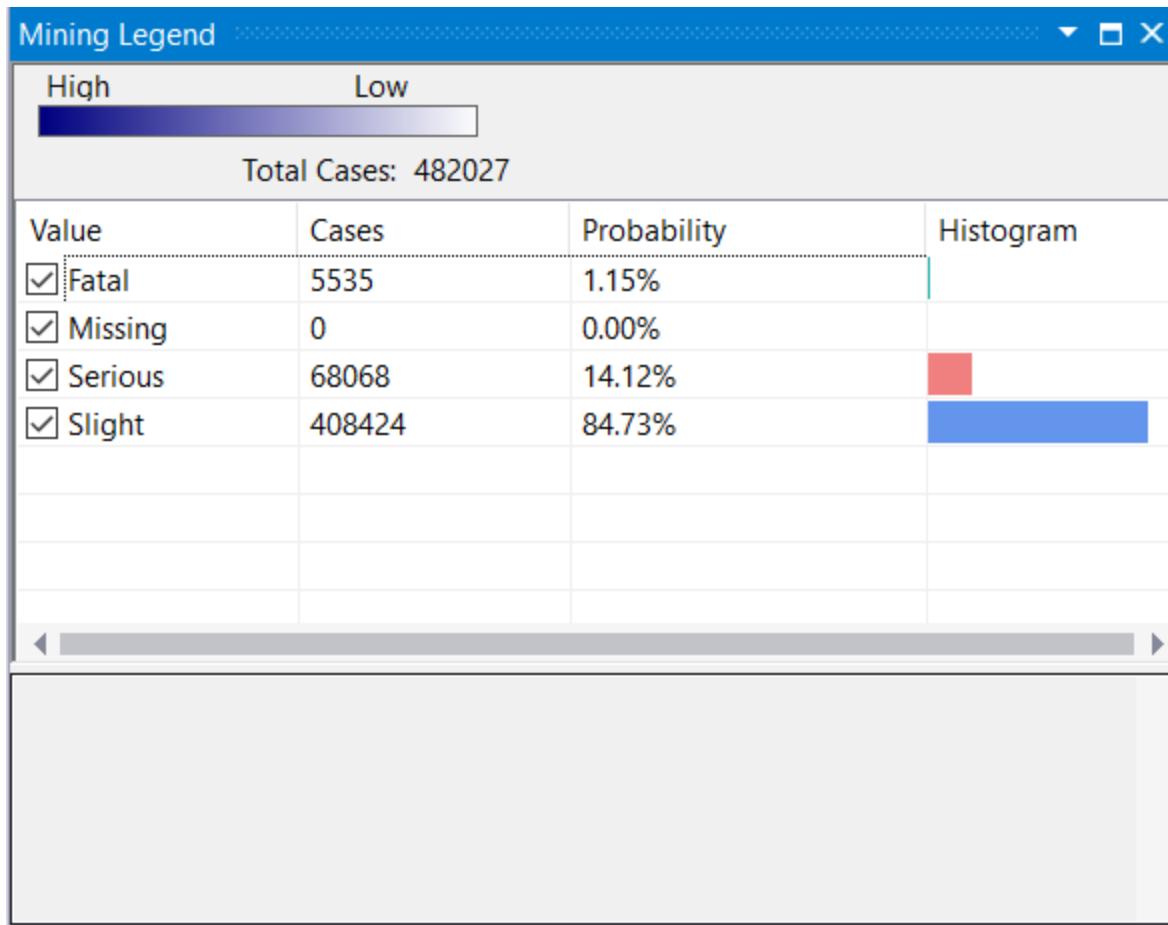
Input data will be randomly split into two sets, a training set and a testing set, based on the percentage of data for testing and maximum number of cases in testing data set you provide. The training set is used to create the mining model. The testing set is used to check model accuracy.

[Percentage of data for testing] specifies percentages of cases reserved for testing set.
[Maximum number of cases in testing data set] limits total number of cases in the testing set.
If both values are specified, both limits are enforced.

< Back Next >| Finish >> Cancel

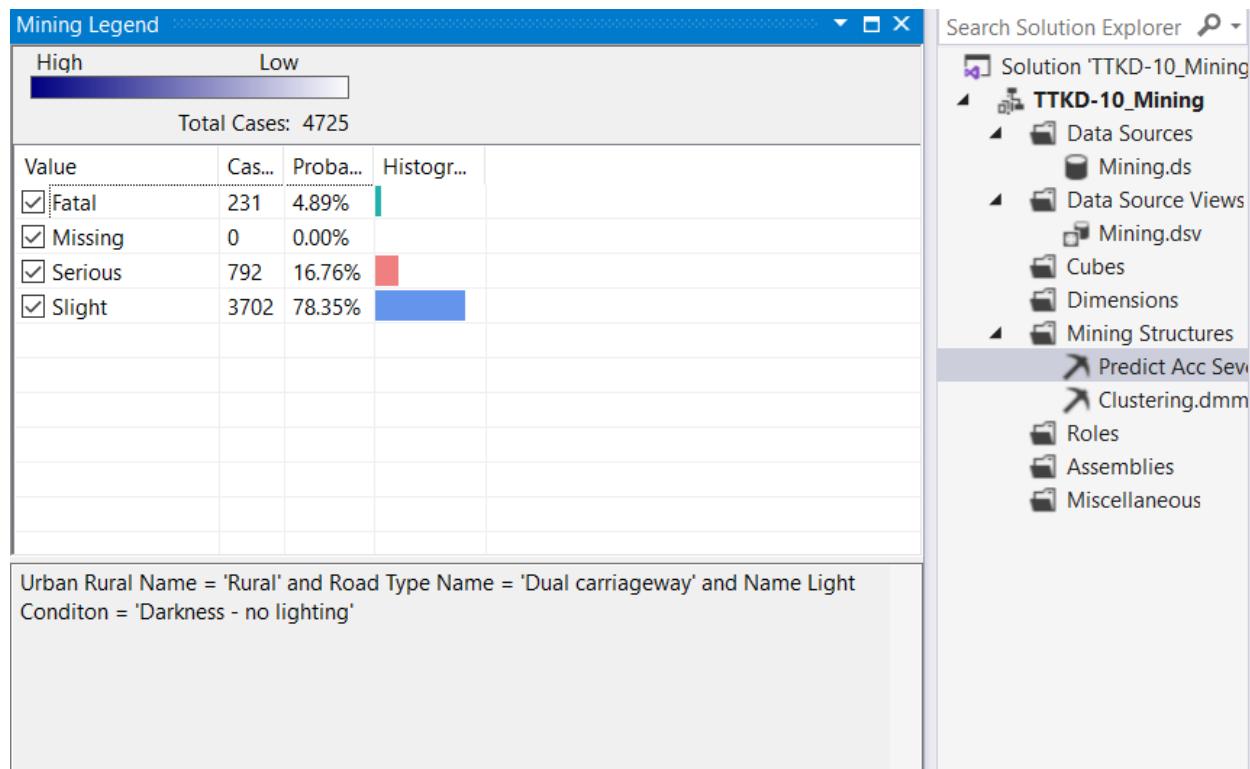
c. Kết quả:

- ❖ Tại tab Mining Model Viewer: Kết quả của mô hình sau khi chạy thành công.



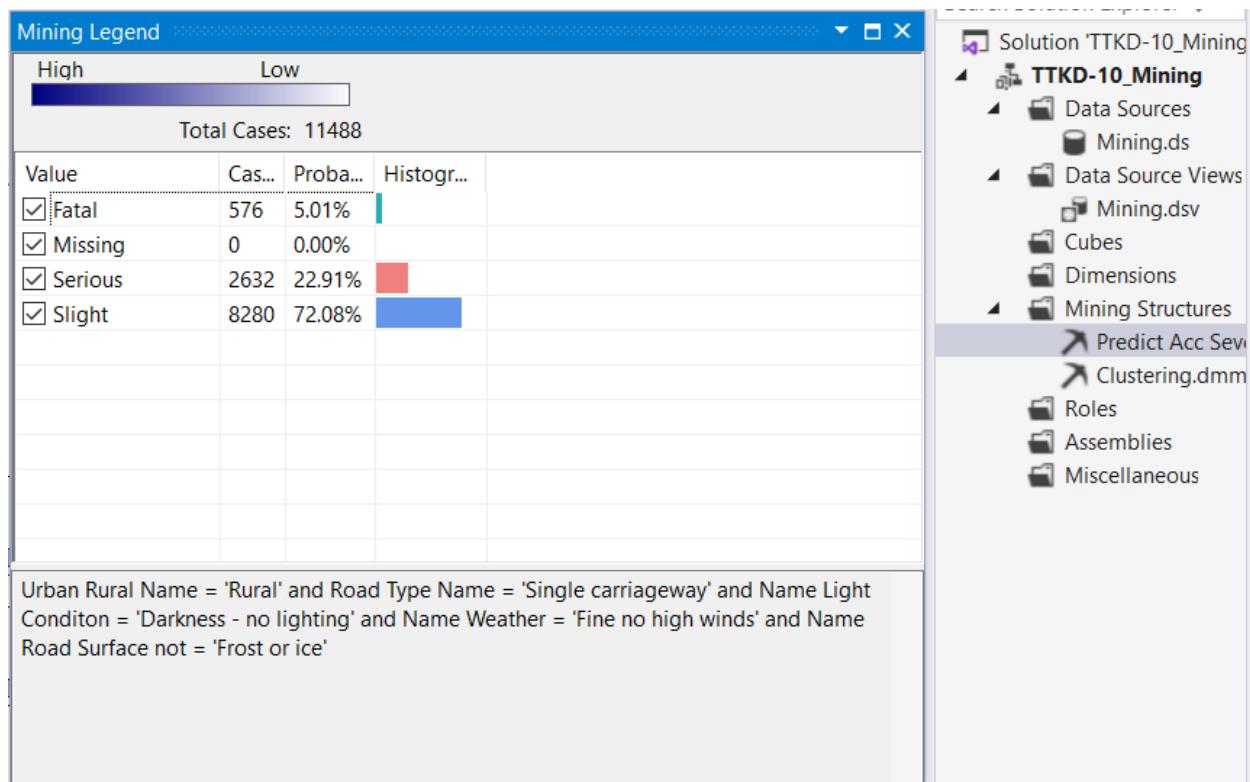
Hình trên thể hiện kết quả tổng quan của bài toán dự đoán: đối với bài toán và tập data nhóm lựa chọn có tổng cộng 482027 trường hợp, trong đó:

- 5535 trường hợp được dự đoán là Fatal chiếm 1,15% trên tổng số
- 68068 trường hợp được dự đoán là Serious chiếm 14,12% trên tổng số
- 408424 trường hợp được dự đoán là Slight chiếm 84,73% trên tổng số
- Không có trường hợp nào missing data.



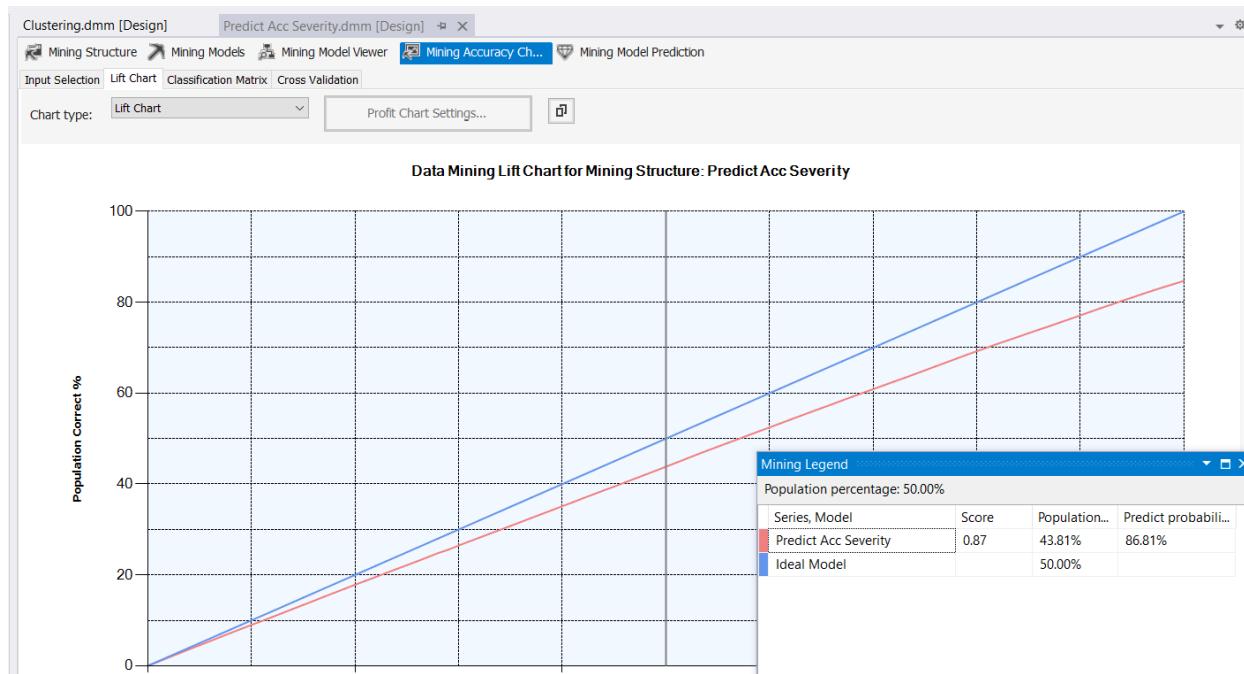
Trong điều kiện về ánh sáng là Darkness – no lighting và loại đường là Dual carriageway (xa lộ hai chiều) và xảy ra ở Rural(nông thôn) thì tỷ lệ xảy ra tai nạn với

- Mức độ Fatal(Tử vong) lên tới 4,89% (231 trường hợp)
- Mức độ Serious(Nghiêm trọng) lên tới 16,76% (792 trường hợp)
- Mức độ Slight(Nhẹ) lên tới 78.35% (3702 trường hợp)



Trong điều kiện về ánh sáng là Darkness – no lighting và loại đường là Single carriageway (xa lộ 1 chiều), thời tiết là Fine no high winds, mặt đường không phải là Frost or ice và xảy ra ở Rural(nông thôn) thì tỷ lệ xảy ra tai nạn với

- Mức độ Fatal(Tử vong) lên tới 5,01% (576 trường hợp)
- Mức độ Serious(Nghiêm trọng) lên tới 22,91% (2632 trường hợp)
- Mức độ Slight(Nhẹ) lên tới 72,08% (8280 trường hợp)



Line chart cho ta thấy tỷ lệ chính xác của thuật toán Microsoft Decision Tree: 0.87 score.

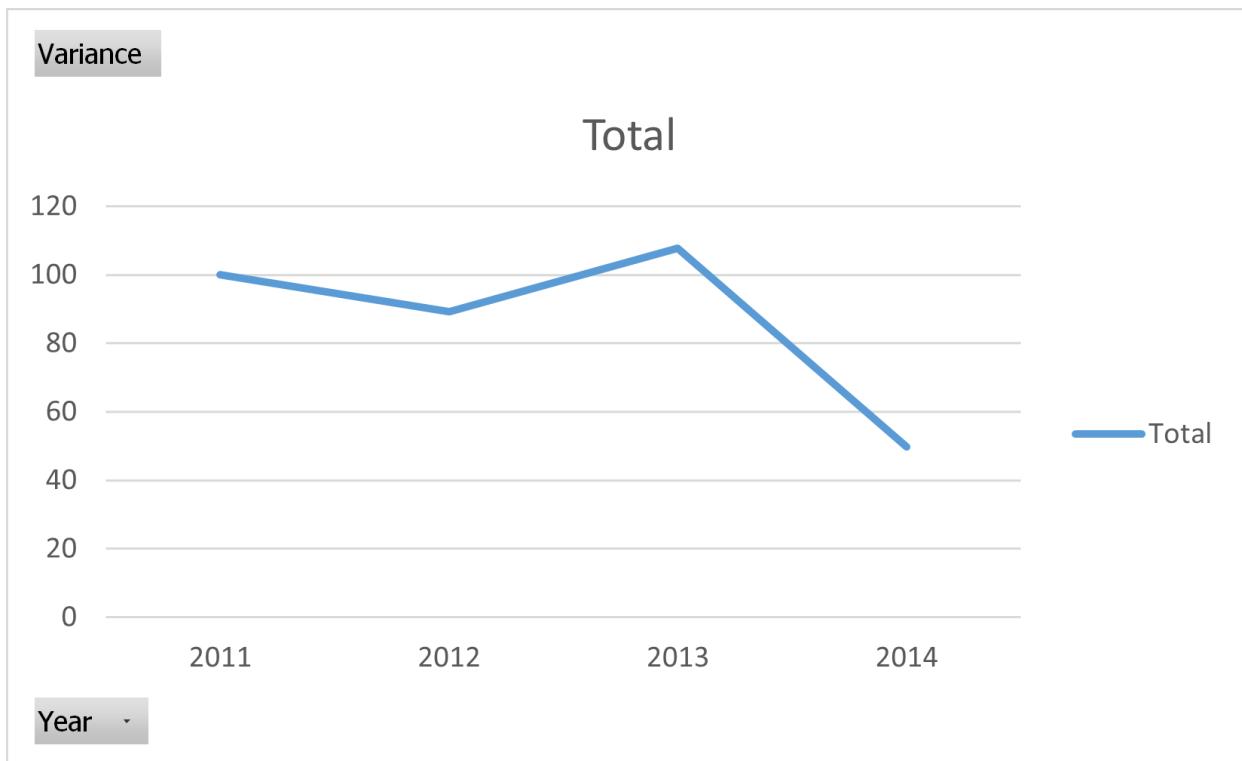
11.Điểm cộng

11.1 Calculated measure

R11. Định nghĩa fact Variance để tính mức độ tăng giảm của TNGT theo đơn vị phần trăm qua các năm

```
--R11: Định nghĩa fact Variance để tính mức độ tăng giảm của TNGT theo đơn vị phần trăm qua các năm (Calculated Measures)
with member [Measures].[last Numberof Accidents] as
(ParallelPeriod([Date].[Hierarchy].[Year]
, 0
,[Date].[Hierarchy].CurrentMember
,[Measures].[Numberof Accidents])
/
(ParallelPeriod([Date].[Hierarchy].[Year]
, 1
,[Date].[Hierarchy].CurrentMember
,[Measures].[Numberof Accidents])
select
non empty [Measures].[last Numberof Accidents] on rows,
non empty[Date].[Hierarchy].[Year] on columns
from [OLAPCUBE_TTKD10]
```

last Numberof Accidents	2011	2012	2013	2014
inf	0.892353411135751	1.07833522993507	0.497664385368655	



107