

BỘ GIÁO DỤC VÀ ĐÀO TẠO  
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH  
KHOA CÔNG NGHỆ THÔNG TIN



ĐỒ ÁN CUỐI KỲ MÔN HỌC  
**KHO DỮ LIỆU**

ĐỀ TÀI:

**XÂY DỰNG KHO DỮ LIỆU QUẢN LÝ CÁC CA TAI NẠN  
GIAO THÔNG CỦA NUỚC MỸ TỪ 2016 - 2023**

**GVHD: Nguyễn Văn Thành**

**Lớp HP: DAWH430784**

**Nhóm thực hiện: Nhóm 2**

**Học kỳ: II**

**Năm học: 2024 – 2025**

*Thành phố Hồ Chí Minh, tháng 5 năm 2025*

## DANH SÁCH THÀNH VIÊN

<i>Dương Minh Hiếu</i>	22133018
<i>Trương Trọng Đại Long</i>	22133033
<i>Nguyễn Ngọc Minh Nhật</i>	22133039
<i>Nguyễn Quang Hoàng Phát</i>	22133040
<i>Nguyễn Đức Cao Thắng</i>	22133053

## PHÂN CÔNG NHIỆM VỤ

Nhóm: 2

Lớp: DAWH430784

**Đề tài: Thiết kế kho dữ liệu với Dữ liệu Các Ca Tai Nạn Giao Thông US từ 2016 - 2023**

Nhiệm vụ	Minh Hiếu	Cao Thăng	Đại Long	Minh Nhật	Hoàng Phát
Tìm kiếm dữ liệu	X	X	X	X	X
Hiểu tập dữ liệu	X	X			
Xác định Business Process	X	X		X	
Xác định bảng Dim	X	X		X	
Xác định bảng Fact		X		X	
Đẩy dữ liệu từ CSV – SQL Server	X	X	X	X	X
Tạo nguồn kết nối dữ liệu	X	X	X	X	X
Staging và load và các dim, fact				X	
Nhập dữ liệu vào SSAS và tạo Data Cube			X		
Phân tích SSAS			X		
Đặt các câu hỏi	X	X	X	X	X
Trả lời các câu hỏi bằng SSAS			X		
Trả lời các câu hỏi bằng Pivot Table	X				
Report PowerBI					X
Viết báo cáo và trình bày	X	X	X	X	X

# MỤC LỤC

<b>I. Giới thiệu</b>	1
1. Mục tiêu	1
2. Lý do chọn đề tài	1
3. Chuẩn bị	2
<b>II. Các bước thực hiện xây dựng hệ thống kho dữ liệu</b>	2
1. Dữ liệu	2
1.1. Tìm kiếm dữ liệu	2
1.2. Mô tả tập dữ liệu	3
1.3 Tiêu xử lý dữ liệu	8
2. Xây dựng cấu trúc hệ thống	11
2.1. Business requirement	11
2.2. Thiết kế mức cao (high level design)	12
2.2.1 Xác định mức độ chi tiết (declare the grain)	12
2.2.2 Xác định các dimension (identify the dimensions)	12
2.2.3. Xác định bảng Fact	12
2.3. Thiết kế mức chi tiết	13
3. Xây dựng kho dữ liệu bằng SSIS	15
3.1. Load dữ liệu từ excel vào database Source	15
3.2. Cấu hình connection tới Source, Stage và DataWarehouse	16
3.3. Tạo các table stage để chuẩn bị cho việc staging dữ liệu:	17
3.4. Load dữ liệu từ Source vào Stage cho các bảng Dim	17
3.5. Thực hiện tạo stage cho bảng Fact	19
3.6. Đổ dữ liệu từ stage vào các bảng Dim và Fact	22
3.7. Nạp dữ liệu từ Stage vào DataWarehouse cho các bảng Dim	22
3.8. Nạp dữ liệu FactAccident	23
4. Tạo Cube và thực hiện truy vấn bằng SSAS	25
4.1. Tạo Data Source từ kho dữ liệu database DWAccident	25
4.2.Tạo Data Source View	25
4.3. Tạo cube, thêm measure và các dim cần thiết,tạo các phân cấp	25
4.4. Truy vấn các câu hỏi mà nhóm đưa ra bằng công cụ SSAS,	26
4.5. Truy vấn các câu hỏi mà nhóm đưa ra bằng công cụ Pivot Tables	33
5. Thiết kế Dashboard Power BI	40
5.1. Khởi đầu hiểu bức tranh toàn cảnh về tai nạn giao thông theo vị trí	40
5.2. Tai nạn giao thông khi có các điều kiện tác động	44
5.3. Thời gian kê câu chuyện gì? Dashboard "Thống kê tai nạn theo thời gian"	49
5.4. Kết luận: Hành động từ dữ liệu	52
<b>III. Kết luận</b>	53
<b>TÀI LIỆU THAM KHẢO</b>	54

# I. Giới thiệu

## 1. Mục tiêu

Báo cáo này được thực hiện nhằm xây dựng một kho dữ liệu hiệu quả để quản lý và phân tích dữ liệu các ca tai nạn giao thông tại Hoa Kỳ trong giai đoạn từ năm 2016 đến năm 2023. Trong bối cảnh an toàn giao thông đường bộ là một ưu tiên quan trọng của xã hội hiện đại, kho dữ liệu được thiết kế không chỉ để thu thập và lưu trữ thông tin mà còn để hỗ trợ các cơ quan quản lý nhà nước, các nhà nghiên cứu và các tổ chức liên quan trong việc đưa ra các quyết định chính sách phù hợp. Hệ thống này cho phép phân tích xu hướng tai nạn, xác định các khu vực có nguy cơ cao và đề xuất các biện pháp phòng ngừa hiệu quả nhằm giảm thiểu thiệt hại về người và tài sản. Hơn nữa, kho dữ liệu được xây dựng với mục tiêu trở thành một hệ thống phân tán, đảm bảo khả năng xử lý nhanh chóng và chính xác các truy vấn phức tạp, đáp ứng nhu cầu phân tích dữ liệu lớn trong thời đại công nghệ Big Data. Tính bảo mật, độ chính xác và khả năng mở rộng của hệ thống cũng được đặt lên hàng đầu để đáp ứng các yêu cầu phân tích dài hạn, đồng thời hỗ trợ tích hợp với các công nghệ mới trong tương lai, từ đó phục vụ tốt nhất cho các mục tiêu chiến lược của chính phủ Mỹ và các tổ chức liên quan đến an toàn giao thông.

## 2. Lý do chọn đề tài

Việc xây dựng kho dữ liệu quản lý các ca tai nạn giao thông tại Hoa Kỳ trong giai đoạn 2016-2023 xuất phát từ nhu cầu cấp thiết trong việc nâng cao hiệu quả quản lý và phân tích dữ liệu liên quan đến an toàn giao thông. Tai nạn giao thông không chỉ gây ra thiệt hại nghiêm trọng về con người và tài sản mà còn đặt ra thách thức lớn đối với các cơ quan quản lý trong việc hoạch định chính sách và triển khai các biện pháp can thiệp kịp thời. Dữ liệu tai nạn giao thông từ năm 2016 đến 2023 chứa đựng lượng thông tin phong phú, bao gồm các yếu tố như thời gian, địa điểm, điều kiện thời tiết, đặc điểm phương tiện và hành vi của tài xế, tạo cơ hội để phân tích sâu rộng nhằm phát hiện các mô hình và nguyên nhân gốc rễ của tai nạn. Việc chọn đề tài này không chỉ đáp ứng nhu cầu thực tiễn trong việc tổ chức và khai thác dữ liệu mà còn góp phần vào sự phát triển của các phương pháp phân tích dữ liệu lớn, tận dụng công nghệ kho dữ liệu hiện đại để cung cấp thông tin giá trị cho các nhà hoạch định chính sách, nhà nghiên cứu và cộng đồng. Ngoài ra, đề tài mang ý nghĩa xã hội sâu sắc khi góp phần cải thiện an toàn giao thông, giảm thiểu tai nạn và nâng cao chất lượng cuộc sống cho người dân Mỹ.

### **3. Chuẩn bị**

Để thực hiện báo cáo và xây dựng kho dữ liệu quản lý các ca tai nạn giao thông tại Hoa Kỳ từ năm 2016 đến 2023, trước tiên, nhóm thực hiện đã tiến hành thu thập dữ liệu từ các nguồn đáng tin cậy, bao gồm các cơ quan quản lý giao thông, cơ sở dữ liệu công cộng của chính phủ Mỹ và các tổ chức nghiên cứu an toàn giao thông. Các tập dữ liệu này được kiểm tra và làm sạch để đảm bảo tính nhất quán, chính xác và đầy đủ trước khi đưa vào kho dữ liệu. Tiếp theo, việc nghiên cứu các công nghệ và công cụ phù hợp để xây dựng kho dữ liệu đã được thực hiện, bao gồm việc lựa chọn hệ quản trị cơ sở dữ liệu SQL Server 2022 để lưu trữ và quản lý dữ liệu, SQL Server Integration Services (SSIS) để thực hiện các quy trình ETL (Extract, Transform, Load), và SQL Server Analysis Services (SSAS) để xây dựng các khối phân tích đa chiều (OLAP cubes) nhằm hỗ trợ phân tích dữ liệu phức tạp. Visual Studio 2022 được sử dụng làm môi trường phát triển chính để thiết kế và triển khai các gói SSIS, SSAS, đồng thời đảm bảo tích hợp mượt mà giữa các thành phần của hệ thống. Để trực quan hóa và báo cáo dữ liệu, Power BI đã được chọn để tạo các báo cáo tương tác và bảng điều khiển (dashboards), giúp các bên liên quan dễ dàng khai thác thông tin.

## **II. Các bước thực hiện xây dựng hệ thống kho dữ liệu**

### **1. Dữ liệu**

#### **1.1. Tìm kiếm dữ liệu**

Quá trình xây dựng kho dữ liệu quản lý các ca tai nạn giao thông tại Hoa Kỳ từ năm 2016 đến năm 2023 bắt đầu bằng việc tìm kiếm và lựa chọn nguồn dữ liệu phù hợp, đảm bảo tính đầy đủ, chính xác và phù hợp với mục tiêu phân tích. Sau khi xem xét nhiều nguồn dữ liệu tiềm năng, nhóm thực hiện đã quyết định sử dụng tập dữ liệu US-Accidents có sẵn trên nền tảng Kaggle, một nguồn dữ liệu công khai được công nhận rộng rãi trong cộng đồng nghiên cứu. Tập dữ liệu này bao gồm thông tin về khoảng 7.7 triệu vụ tai nạn giao thông trên 49 bang của Hoa Kỳ, được thu thập từ tháng 2 năm 2016 đến tháng 3 năm 2023 thông qua nhiều API cung cấp dữ liệu sự kiện giao thông theo thời gian thực. Các API này tổng hợp thông tin từ các cơ quan quản lý giao thông Hoa Kỳ, cơ quan thực thi pháp luật, camera giao thông và cảm biến trên mạng lưới đường bộ, đảm bảo dữ liệu phản ánh đầy đủ các khía cạnh của tai nạn giao thông, bao gồm thời gian, địa điểm, điều kiện môi trường, đặc điểm phương tiện và các yếu tố liên quan. Để phù hợp với phạm vi nghiên cứu từ năm 2016 đến năm 2023, nhóm đã lọc dữ liệu để chỉ sử dụng các bản ghi trong khoảng thời gian này, loại bỏ các bản ghi ngoài khung thời gian nhằm đảm bảo tính tập trung và nhất quán.

Tập dữ liệu US-Accidents được lựa chọn không chỉ vì khối lượng dữ liệu lớn mà còn vì tính đa dạng và giá trị ứng dụng cao trong các phân tích liên quan đến an toàn giao thông. Dữ liệu này hỗ trợ nhiều mục đích nghiên cứu, từ dự đoán tai nạn theo thời gian thực, phân tích các điểm nóng tai nạn, đánh giá thương vong, đến khám phá các quy luật nhân quả và tác động của các yếu tố môi trường như lượng mưa hoặc điều kiện thời tiết. Ngoài ra, tập dữ liệu cung cấp một phiên bản mẫu với 500,000 bản ghi, giúp nhóm thực hiện dễ dàng thử nghiệm và đánh giá ban đầu trước khi xử lý toàn bộ dữ liệu. Tuy nhiên, nhóm cũng nhận thấy một số hạn chế của tập dữ liệu, chẳng hạn như việc thiếu dữ liệu cho một số ngày do vấn đề kết nối mạng trong quá trình thu thập, đòi hỏi phải thực hiện các bước làm sạch và xử lý dữ liệu bổ sung. Tập dữ liệu được phân phối dưới giấy phép Creative Commons Attribution-Noncommercial-ShareAlike (CC BY-NC-SA 4.0), phù hợp với mục đích nghiên cứu học thuật của dự án, và nhóm cam kết tuân thủ các yêu cầu trích dẫn, bao gồm việc tham chiếu các bài báo khoa học liên quan của tác giả Sobhan Moosavi và cộng sự. Quá trình tìm kiếm dữ liệu đã đặt nền tảng vững chắc cho các bước tiếp theo trong việc xây dựng kho dữ liệu, đảm bảo rằng hệ thống có thể khai thác hiệu quả thông tin từ một nguồn dữ liệu đáng tin cậy và toàn diện.

## 1.2. Mô tả tập dữ liệu

Tập dữ liệu "US Accidents" chứa thông tin chi tiết về các vụ tai nạn giao thông tại Mỹ. Dữ liệu bao gồm nhiều thuộc tính liên quan đến thời gian, vị trí, điều kiện thời tiết, và các yếu tố hạ tầng giao thông tại thời điểm xảy ra tai nạn,... Tập dữ liệu này có thể được sử dụng để phân tích các yếu tố ảnh hưởng đến tai nạn giao thông, từ đó hỗ trợ việc đưa ra các biện pháp phòng ngừa và cải thiện an toàn giao thông.

STT	Tên thuộc tính	Ý nghĩa
1	ID	ID duy nhất của tai nạn giao thông
2	SEVERITY	Mức độ nghiêm trọng của tai nạn
3	START_TIME	Thời gian bắt đầu của tai nạn
4	END_TIME	Thời gian kết thúc của tai nạn
5	START_LAT	Vĩ độ của vị trí bắt đầu tai nạn

STT	Tên thuộc tính	Ý nghĩa
6	START_LNG	Kinh độ của vị trí bắt đầu tai nạn
7	END_LAT	Vĩ độ của vị trí kết thúc tai nạn
8	END_LNG	Kinh độ của vị trí kết thúc tai nạn
9	DISTANCE (mi)	Khoảng cách của tai nạn tính bằng dặm
10	DESCRIPTION	Mô tả về vụ tai nạn
11	NUMBER	Số nhà gần vị trí tai nạn
12	STREET	Tên đường nơi xảy ra tai nạn
13	SIDE	Vị trí của tai nạn trên đường (trái
14	CITY	Thành phố nơi xảy ra tai nạn
15	COUNTY	Quận nơi xảy ra tai nạn
16	STATE	Tiểu bang nơi xảy ra tai nạn
17	ZIPCODE	Mã bưu chính của vị trí tai nạn
18	COUNTRY	Tên quốc gia (trong trường hợp này là Mỹ)
19	TIMEZONE	Múi giờ địa phương của vị trí tai nạn
20	AIRPORT_CODE	Mã sân bay gần vị trí tai nạn
21	WEATHER_TIMESTAMP	Thời gian thu thập thông tin thời tiết tại thời điểm tai nạn

STT	Tên thuộc tính	Ý nghĩa
22	TEMPERATURE (°F)	Nhiệt độ tính bằng độ Fahrenheit
23	WIND_CHILL (°F)	Nhiệt độ gió tính bằng độ Fahrenheit
24	HUMIDITY (%)	Độ ẩm tính bằng phần trăm
25	PRESSURE (in)	Áp suất khí quyển tính bằng inches
26	VISIBILITY (mi)	Tầm nhìn tính bằng dặm
27	WIND_DIRECTION	Hướng gió
28	WIND_SPEED (mph)	Tốc độ gió tính bằng dặm/giờ
29	PRECIPITATION (in)	Lượng mưa tính bằng inches
30	WEATHER_CONDITION	Điều kiện thời tiết (như mưa
31	AMENITY	Có cửa hàng tiện ích (như cây xăng
32	BUMP	Có chướng ngại vật trên đường không
33	CROSSING	Có đường giao nhau không
34	GIVE_WAY	Có biển báo nhường đường không
35	JUNCTION	Có ngã tư không
36	NO_EXIT	Có đường cụt không
37	RAILWAY	Có đường sắt cắt ngang không

STT	Tên thuộc tính	Ý nghĩa
38	ROUNABOUT	Có vòng xuyến không
39	STATION	Có trạm xe buýt hoặc trạm dừng chân không
40	STOP	Có biển STOP trên đường không
41	TRAFFIC_CALMING	Có biện pháp hạn chế tốc độ trên đường không
42	TRAFFIC_SIGNAL	Có đèn giao thông không
43	TURNING_LOOP	Có lối rẽ hay không
44	SUNRISE_SUNSET	Thời điểm tai nạn là ban ngày hay ban đêm (mặt trời mọc hay lặn)
45	CIVIL_TWILIGHT	Thời điểm tai nạn có ánh sáng dân sự không (hoàng hôn hoặc bình minh)
46	NAUTICAL_TWILIGHT	Thời điểm tai nạn có ánh sáng hàng hải không
47	ASTRONOMICAL_TWILIGHT	Thời điểm tai nạn có ánh sáng thiên văn không

Tập dữ liệu bổ sung của "US Accidents" cung cấp thông tin chi tiết hơn về các đặc điểm của phương tiện, hành vi lái xe, và các yếu tố liên quan đến vụ tai nạn. Dữ liệu bao gồm các thuộc tính như độ tuổi tài xế, loại phương tiện, sức chứa động cơ, và các yếu tố kỹ thuật khác. Tập dữ liệu này hỗ trợ phân tích sâu hơn về vai trò của tài xế và phương tiện trong các vụ tai nạn, từ đó giúp cải thiện an toàn giao thông.

STT	Tên thuộc tính	Ý nghĩa	Ghi chú
1	Accident_Index	Mã chỉ số duy nhất của vụ tai nạn	dùng để nhận diện từng sự kiện.

2	Age_Band_of_Driver	Dải tuổi của tài xế (ví dụ: dưới 25	25-34
3	Age_of_Vehicle	Độ tuổi của phương tiện (thời gian từ khi phương tiện được sản xuất đến thời điểm xảy ra tai nạn)	
4	Driver_Home_Area_Type	Loại khu vực nơi tài xế sinh sống (ví dụ: thành thị	nông thôn)
5	Driver_IMD_Decile	Chỉ số IMD (Index of Multiple Deprivation) của tài xế	được chia thành 10 phân vị (decile) để đánh giá mức độ bất bình đẳng xã hội
6	Engine_Capacity_(CC)	Dung tích động cơ của phương tiện	tính bằng cc (centimet khối)
7	Hit_Object_in_Carriageway	Đối tượng bị va chạm trên làn đường (ví dụ: xe khác	chướng ngại vật)
8	Journey_Purpose_of_Driver	Mục đích chuyến đi của tài xế (ví dụ: công việc	cá nhân)
9	Junction_Location	Vị trí của tai nạn tại ngã tư hoặc giao lộ	
10	make	Hãng sản xuất phương tiện (ví dụ: Toyota	Ford)
11	model	Mô hình hoặc loại phương tiện (ví dụ: Camry	F-150)
12	Propulsion_Code	Loại động cơ hoặc hệ thống truyền động của phương tiện (ví dụ: xăng	diesel)

13	Sex_of_Driver	Giới tính của tài xế	
14	Skidding_and_Overturning	Tình trạng trượt bánh hoặc lật xe	
15	Towing_and_Articulation	Tình trạng kéo xe hoặc khớp nối (ví dụ: xe kéo rơ-mooc)	
16	Vehicle_Leaving_Carriageway	Phương tiện rời khỏi làn đường	
17	Vehicle_Location_Restricted_Lane	Vị trí phương tiện trong làn đường hạn chế	
18	Vehicle_Manoeuvre	Hành động hoặc thao tác của phương tiện (ví dụ: rẽ)	dừng)
19	Vehicle_Reference	Mã tham chiếu của phương tiện trong vụ tai nạn	
20	Vehicle_Type	Loại phương tiện (ví dụ: ô tô	xe máy
21	Was_Vehicle_Left_Hand_Drive	Phương tiện có vô-lăng bên trái không	
22	X1st_Point_of_Impact	Điểm va chạm đầu tiên trên phương tiện	
23	Year	Năm sản xuất của phương tiện	

### 1.3 Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một bước quan trọng trong quá trình xây dựng kho dữ liệu quản lý các ca tai nạn giao thông tại Hoa Kỳ từ năm 2016 đến năm 2023. Mục tiêu của bước này là đảm bảo rằng dữ liệu được thu thập từ các nguồn như tập dữ liệu US-Accidents và Vehicle được làm sạch, chuẩn hóa và sẵn sàng cho việc tích hợp vào kho dữ liệu.

Quá trình tiền xử lý bao gồm việc xử lý các giá trị thiếu, chuẩn hóa định dạng, loại bỏ các giá trị không hợp lệ, và đảm bảo tính nhất quán của dữ liệu để hỗ trợ các phân tích phức tạp sau này. Dựa trên mã nguồn được cung cấp, phần tiền xử lý dữ liệu

được thực hiện trên hai tập dữ liệu chính: Vehicle.csv và US\_Accidents\_Dec21\_update.csv, với các bước cụ thể như sau:

- Xử lý dữ liệu trong tập Vehicle.csv

Tập dữ liệu Vehicle.csv chứa thông tin chi tiết về các phương tiện và tài xế liên quan đến các vụ tai nạn giao thông. Dữ liệu này bao gồm các thuộc tính như dải tuổi của tài xế (Age\_Band\_of\_Driver), độ tuổi của phương tiện (Age\_of\_Vehicle), khu vực sinh sống của tài xế (Driver\_Home\_Area\_Type), giới tính của tài xế (Sex\_of\_Driver), và vị trí giao lộ (Junction\_Location). Tuy nhiên, tập dữ liệu này tồn tại nhiều giá trị thiếu (NaN) và các giá trị không hợp lệ như "Data missing or out of range", "Not known", hoặc các giá trị không phù hợp với định dạng mong muốn (ví dụ: "15-Nov", "10-Jun"). Để giải quyết các vấn đề này, quá trình tiền xử lý đã được thực hiện như sau:

- Xử lý giá trị thiếu (NaN):
  - + Đối với các cột như Age\_Band\_of\_Driver, Driver\_Home\_Area\_Type, Sex\_of\_Driver, và Junction\_Location, các giá trị thiếu được điền bằng cách chọn ngẫu nhiên các giá trị hợp lệ từ danh sách các giá trị duy nhất (unique\_values) trong cùng cột. Phương pháp này sử dụng hàm np.random.choice của thư viện NumPy để đảm bảo tính ngẫu nhiên và giữ được phân bố tự nhiên của dữ liệu.
  - + Ví dụ, với cột Age\_Band\_of\_Driver, các giá trị thiếu được thay thế bằng các giá trị hợp lệ (như "26-35", "36-45", v.v.) được chọn ngẫu nhiên từ danh sách các giá trị đã được làm sạch.
- Loại bỏ và thay thế giá trị không hợp lệ:
  - + Một số giá trị trong các cột như Age\_Band\_of\_Driver, Driver\_Home\_Area\_Type, Sex\_of\_Driver, và Junction\_Location chứa các giá trị không hợp lệ như "Data missing or out of range", "15-Nov", "10-Jun", "0 - 5", "Not known", hoặc "None". Những giá trị này được xác định trong danh sách lst\_remove và bị loại bỏ khỏi danh sách các giá trị hợp lệ (lst\_Age\_Band\_Of\_Driver, lst\_Driver\_Home\_Area\_Type, v.v.).
  - + Sau đó, các hàng chứa giá trị không hợp lệ được thay thế bằng cách chọn ngẫu nhiên một giá trị từ danh sách các giá trị hợp lệ, đảm bảo rằng dữ liệu vẫn giữ được tính đại diện và phù hợp với mục đích phân tích.
- Xử lý cột Age\_of\_Vehicle:
  - + Đối với cột Age\_of\_Vehicle, các giá trị thiếu (NaN) được thay thế bằng cách chọn ngẫu nhiên một giá trị từ danh sách các giá trị hợp lệ (lst\_Age\_of\_Vehicle). Điều này đảm bảo rằng tất cả các bản ghi đều có giá trị

hợp lệ cho độ tuổi của phương tiện, giúp duy trì tính toàn vẹn của dữ liệu khi phân tích các yếu tố liên quan đến phương tiện.

- Lưu trữ dữ liệu đã xử lý:

Sau khi hoàn tất các bước tiền xử lý, tập dữ liệu đã được làm sạch được lưu lại vào tệp Vehicle.csv mới bằng lệnh `df.to_csv('./Vehicle.csv', index=False)`. Việc này đảm bảo rằng dữ liệu đã được chuẩn hóa và sẵn sàng để tích hợp vào kho dữ liệu.

- Xử lý dữ liệu trong tập US\_Accidents\_Dec21\_update.csv

Tập dữ liệu US\_Accidents\_Dec21\_update.csv chứa thông tin chi tiết về các vụ tai nạn giao thông tại Hoa Kỳ, bao gồm các thuộc tính như thời gian bắt đầu (Start\_Time), thời gian kết thúc (End\_Time), nhiệt độ (Temperature(F)), độ ẩm (Humidity(%)), áp suất khí quyển (Pressure(in)), tốc độ gió (Wind\_Speed(mph)), lượng mưa (Precipitation(in)), và tầm nhìn (Visibility(mi)). Tương tự như tập Vehicle.csv, tập dữ liệu này cũng gặp vấn đề về giá trị thiếu và định dạng không hợp lệ, đặc biệt là trong các cột liên quan đến thời gian và điều kiện thời tiết. Các bước tiền xử lý được thực hiện như sau:

- Xử lý giá trị không hợp lệ trong cột thời gian (Start\_Time và End\_Time):
  - + Một số bản ghi trong các cột Start\_Time và End\_Time chứa giá trị không hợp lệ như "1/0/00", không phù hợp với định dạng ngày tháng chuẩn (MM/DD/YY). Để xử lý, các giá trị này được thay thế bằng cách tạo ngẫu nhiên các ngày tháng hợp lệ trong khoảng thời gian từ năm 2017 đến 2023 (dựa trên danh sách năm [17, 19, 21, 23]).
  - + Cụ thể, cho mỗi bản ghi có giá trị "1/0/00":
    - Năm (y) được chọn ngẫu nhiên từ danh sách [17, 19, 21, 23].
    - Tháng (m) được chọn ngẫu nhiên từ 1 đến 12.
    - Ngày (d) được chọn ngẫu nhiên dựa trên số ngày hợp lệ của tháng (ví dụ: tháng 2 có tối đa 28 ngày, các tháng 4, 6, 9, 11 có tối đa 30 ngày, và các tháng còn lại có tối đa 31 ngày).
    - Giá trị ngày tháng mới được tạo dưới dạng chuỗi "MM/DD/YY" và thay thế vào các cột Start\_Time và End\_Time tương ứng.
  - + Phương pháp này đảm bảo rằng tất cả các bản ghi có thời gian hợp lệ, phù hợp với phạm vi nghiên cứu từ năm 2016 đến năm 2023.
- Xử lý giá trị thiếu trong các cột thời tiết:
  - + Các cột liên quan đến điều kiện thời tiết như Temperature(F), Wind\_Chill(F), Humidity(%), Pressure(in), Wind\_Speed(mph), Precipitation(in), và Visibility(mi) chứa các giá trị thiếu (NaN). Để xử lý, các giá trị thiếu được thay

thế bằng cách chọn ngẫu nhiên từ danh sách các giá trị hợp lệ trong cùng cột (ví dụ: lst\_Tem cho Temperature(F), lst\_Wind cho Wind\_Chill(F), v.v.).

- + Phương pháp chọn ngẫu nhiên này giúp duy trì phân bố tự nhiên của dữ liệu thời tiết mà không làm sai lệch các phân tích sau này. Ví dụ, nếu một bản ghi thiếu giá trị Temperature(F), giá trị được thay thế bằng một nhiệt độ ngẫu nhiên từ danh sách các giá trị đã có trong cột.
- Lưu trữ dữ liệu đã xử lý:
- + Sau khi hoàn tất các bước tiền xử lý, tập dữ liệu đã được làm sạch được lưu lại vào tệp Accident.csv bằng lệnh df.to\_csv('./Accident.csv', index=False). Điều này đảm bảo rằng dữ liệu đã được chuẩn hóa và sẵn sàng để tích hợp vào kho dữ liệu.

Quá trình tiền xử lý dữ liệu được thực hiện dựa trên các phương pháp đơn giản nhưng hiệu quả, tập trung vào việc đảm bảo tính nhất quán và đầy đủ của dữ liệu. Việc sử dụng các giá trị ngẫu nhiên để điền giá trị thiếu giúp duy trì phân bố tự nhiên của dữ liệu, tránh làm sai lệch các phân tích thống kê sau này. Tuy nhiên, phương pháp này cũng có một số hạn chế, chẳng hạn như có thể không phản ánh chính xác các đặc điểm cụ thể của từng bản ghi (ví dụ: nhiệt độ thực tế tại một địa điểm cụ thể). Để cải thiện, trong tương lai, nhóm có thể xem xét sử dụng các phương pháp tiên tiến hơn như nội suy dựa trên các giá trị lân cận hoặc mô hình học máy để dự đoán các giá trị thiếu.

Quá trình tiền xử lý đã tạo ra hai tập dữ liệu sạch (Vehicle.csv và Accident.csv) với các giá trị thiếu được điền đầy đủ và các giá trị không hợp lệ được loại bỏ hoặc thay thế. Dữ liệu này sau đó có thể được sử dụng trong các quy trình ETL (Extract, Transform, Load) để tích hợp vào kho dữ liệu, hỗ trợ các phân tích phức tạp như xác định xu hướng tai nạn, đánh giá các điểm nóng tai nạn, và đề xuất các biện pháp cải thiện an toàn giao thông. Việc chuẩn hóa dữ liệu cũng đảm bảo rằng kho dữ liệu có thể tích hợp với các công cụ như SQL Server Integration Services (SSIS), SQL Server Analysis Services (SSAS), và Power BI để tạo ra các báo cáo và bảng điều khiển trực quan, đáp ứng nhu cầu của các nhà hoạch định chính sách và nghiên cứu.

## 2. Xây dựng cấu trúc hệ thống

### 2.1. Business requirement

Dự án muốn xây dựng hệ thống phân tích dữ liệu tai nạn giao thông tại Mỹ theo mô hình Kimball, giúp cơ quan chức năng nhận diện yếu tố rủi ro, thời gian, địa điểm tai nạn để đề xuất giải pháp nâng cao an toàn và giảm thiểu thương vong.

## 2.2. Thiết kế mức cao (high level design)

### 2.2.1 Xác định mức độ chi tiết (declare the grain)

- **Business Process Name:** "Accident Analysis"
- **Fact Table:** "FactAccident" là bảng chứa dữ liệu thực tế về tai nạn.
- **Fact Grain Type:** Xác định loại dữ liệu trong bảng Fact, ở đây là dạng giao dịch ("Transaction").
- **Granularity:** Mỗi hàng trong bảng FactAccident đại diện cho một vụ tai nạn duy nhất.

### 2.2.2 Xác định các dimension (identify the dimensions)

- **DimDate:** Thời gian xảy ra tai nạn
- **DimLocation:** Vị trí tai nạn
- **DimDriver:** Thông tin tài xế
- **DimVehicle:** Đặc điểm phương tiện
- **DimRoadFeature:** Đặc điểm đường
- **DimSpeedLimit:** Giới hạn tốc độ tại nơi xảy ra tai nạn.
- **DimTwilight:** Điều kiện ánh sáng
- **DimWeather:** Thời tiết khi tai nạn xảy ra

### 2.2.3. Xác định bảng Fact

**NumberOfCasualties:** Số người bị thương hoặc tử vong trong tai nạn  
*Loại dữ liệu:* Có thể cộng dồn (Additive)

**NumberOfVehicles:** Số lượng phương tiện tham gia vào tai nạn  
*Loại dữ liệu:* Có thể cộng dồn (Additive)

**Distance:** Khoảng cách bị ảnh hưởng bởi tai nạn  
*Loại dữ liệu:* Bán cộng dồn (Semi-additive)

**Duration\_Minutes:** Thời gian diễn ra tai nạn tính bằng phút  
*Loại dữ liệu:* Bán cộng dồn (Semi-additive)

**AccidentSeverity:** Mức độ nghiêm trọng của tai nạn  
*Loại dữ liệu:* Không cộng dồn (Non-additive)

**Visibility:** Tầm nhìn tại thời điểm xảy ra tai nạn  
*Loại dữ liệu:* Không cộng dồn (Non-additive)

**RoadConditionSeverity:** Mức độ nghiêm trọng của điều kiện đường vào thời điểm xảy ra tai nạn

*Loại dữ liệu:* Không cộng dồn (Non-additive)

**WeatherImpactLevel:** Mức độ ảnh hưởng của thời tiết đến tai nạn

Loại dữ liệu: Không cộng dồn (Non-additive)

Ta sẽ điền vào **Detailed Bus Matrix** worksheet như hình bên dưới.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	#REF!														
2	Business Process Name	Fact Table	Fact Grain Type	Granularity	Facts	DimDate	DimLocation	DimDriver	DimVehidle	DimRoadFeature	DimSpeedLimit	DimNight	DimWeather		
3	Accident Analysis	FactAccident	Transaction	One row per accident	NumberOfCasualties,NumberOfVehicles,Distance,Duration_Minutes,AccidentSeverity,Visibility,RoadConditionSeverity,WeatherImpactLevel	x	x	x	x	x	x	x	x	x	x
4															

### 2.3. Thiết kế mức chi tiết

#### DimDate

- Hoàn thành table definition của worksheet

A	B	C	D	E
1 Table Name	DimDate			
2 Table Type	Dimension			
3 Display Name	Date			
4 Database Schema				
5 Table Description	Date dimension chưa có một hàng cho mỗi ngày.			
6 Comment	The Date dimension is derived; it is not populated from any transaction system.			
7 Biz Filter Logic				
8 Size	365/year			
9 Generate Script?	Y			
10				

- Hoàn thành basic column information

11	Column Name	Display Name	Description	Unknown Member	Example Values	SCD Type
12	DateKey	DateKey	Khóa chính thay thế cho ngày và giờ	-1	200411231301	
13	FullDate	FullDate	Ngày đầy đủ theo định dạng SQL		23/11/2023	
14	Year	Year	Năm	0	2000	1
15	Quarter	Quarter	Quý trong năm (1 đến 4)	0	1, 2, 3, 4	1
16	Month	Month	Tháng trong năm (1 đến 12)	0	1, 2, 3, 4, ...	1
17	Day	Day	Ngày trong tháng (1 đến 31)	0	12	1
18	Hour	Hour	Giờ trong ngày (0 đến 23)	0	13	1
19	Minute	Minute	Phút trong giờ (0 đến 59)	0	4	1
20	Weekday	Weekday	Ngày trong tuần (1 đến 7)	0	1, 2, ...	1
21						
22						

- Hoàn thành target table information và source data information

11	Column Name	Display Name	Datatype	Size	Precision	Key?	FK To	NULL?	Default Value	Source System	Source Schema	Source Table	Source Field Name	Source Datatype
12	DateKey	DateKey	int			PK	N	N	0	Derived				
13	FullDate	FullDate	date				N	0		Derived	stgUSAccidentsDate			DATE
14	Year	Year	int				N	0		Derived	stgUSAccidentsDate			int
15	Quarter	Quarter	int				N	0		Derived	stgUSAccidentsDate			int
16	Month	Month	int				N	0		Derived	stgUSAccidentsDate			int
17	Day	Day	int				N	0		Derived	stgUSAccidentsDate			int
18	Hour	Hour	int				N	0		Derived	stgUSAccidentsDate			int
19	Minute	Minute	int				N	0		Derived	stgUSAccidentsDate			int
20	Weekday	Weekday	int				N	0		Derived	stgUSAccidentsDate			int
21														
22														

#### FactAccident

- Hoàn thành table definition của worksheet

1	<b>Table Name</b>	FactAccident
2	Table Type	Fact
3	Display Name	Accident
4	Database Schema	
5	Table Description	
6	Comment	
7	Biz Filter Logic	
8	Size	
9	Generate Script?	Y

- **Hoàn thành basic column information**

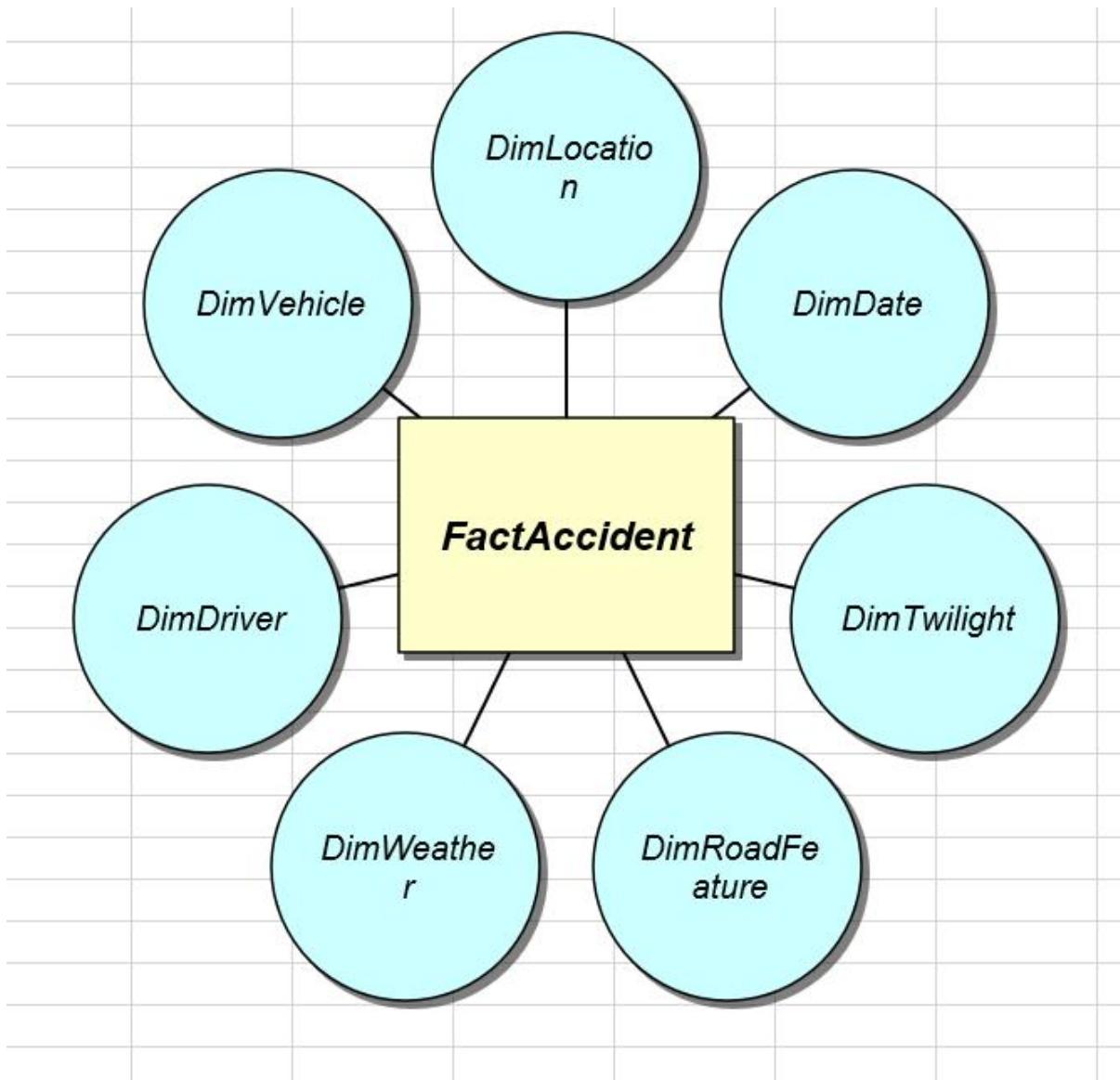
Column Name	Display Name	Description	Example Values	Display Folder	ETL Rules
ID	ID		A-101, ...	key	
LocationKey	LocationKey	Key to location	1, 2, 3	key	Key lookup
DateKey	DateKey	Key to date	1, 2, 3	key	Key lookup
TwilightKey	TwilightKey	Key to twilight	1, 2, 3	key	Key lookup
RoadFeatureKey	RoadFeatureKey	Key to road feature	1, 2, 3	key	Key lookup
WeatherKey	WeatherKey	Key to weather	1, 2, 3	key	Key lookup
DriverKey	DriverKey	Key to Driver	1, 2, 3	key	Key lookup
SpeedLimitKey	SpeedLimitKey	Key to speed limit	1, 2, 3	key	Key lookup
VehicleKey	VehicleKey	Key to vehicle	1, 2, 3	key	Key lookup
NumberOfCasualties	NumberOfCasualties	số người bị thương			
NumberOfVehicles	NumberOfVehicles	số phương tiện liên quan			
AccidentSeverity	AccidentSeverity	mức độ nghiêm trọng (1-3 hoặc nhẹ/vừa/nặng)	1,2,3		
Distance	Distance	Khoảng cách	1,0		
Visibility	Visibility	Điều kiện ánh sáng từ DimTwilight			
Duration Minutes	Duration Minutes	Thời lượng của vụ tai nạn			
RoadConditionSeverity	RoadConditionSeverity	Mức độ ảnh hưởng của điều kiện mặt đường			
WeatherImpactLevel	WeatherImpactLevel	Mức độ ảnh hưởng của thời tiết từ DimWeatherCondition			

- **Hoàn thành target table information và source data information**

Column Name	Display Name	Target					Source			Source Field Name	Source Datatype		
		Datatype	Size	Precision	Key?	FK To	NULL?	Default Value	Source System	Source Schema	Source Table		
ID	ID	varchar	10		PK		N		USAAccidents	dbo	StageAccident	D	varchar
LocationKey	LocationKey	int			FK	DimLocation.LocationKey	N		USAAccidents	dbo	DimLocation	LocationKey	int
DateKey	DateKey	int			FK	DimDate.DateKey	N		USAAccidents	dbo	DimDate	DateKey	int
TwilightKey	TwilightKey	int			FK	DimTwilight.TwilightKey	N		USAAccidents	dbo	DimTwilight	TwilightKey	int
RoadFeatureKey	RoadFeatureKey	int			FK	DimRoadFeature.RoadFeatureKey	N		USAAccidents	dbo	DimRoadFeature	RoadFeatureKey	int
WeatherKey	WeatherKey	int			FK	DimWeather.WeatherKey	N		USAAccidents	dbo	DimWeather	WeatherKey	int
DriverKey	DriverKey	int			FK	DimDriver.DriverKey	N		USAAccidents	dbo	DimDriver	DriverKey	int
SpeedLimitKey	SpeedLimitKey	int			FK	DimSpeed.SpeedLimKey	N		USAAccidents	dbo	DimSpeedLimit	SpeedLimKey	int
VehicleKey	VehicleKey	int			FK	DimVehicle.VehicleKey	N		USAAccidents	dbo	DimVehicle	VehicleKey	int
NumberOfCasualties	NumberOfCasualties	tinyint							stgUSAAccident			NumberOfCasualties	tinyint
NumberOfVehicles	NumberOfVehicles	tinyint							stgUSAAccident			NumberOfVehicles	tinyint
AccidentSeverity	AccidentSeverity								stgUSAAccident			AccidentSeverity	
Distance	Distance	float							stgUSAAccident			Distance	float
Visibility	Visibility	float							stgUSAAccident			visibility_condition	float
Duration_Minutes	Duration_Minutes	int							stgUSAAccident			Duration_Minutes	int
RoadConditionSeverity	RoadConditionSeverity	int							stgUSAAccident			road_condition_severity	int
WeatherImpactLevel	WeatherImpactLevel	int							stgUSAAccident			weather_impact_level	int

Các dimension còn lại: DimLocation, DimDriver, DimRoadFeature, SpeedLimit, DimTwilight, DimVehicle, DimWeather được thể hiện trong file USAccidentsDW-Detailed-Dimensional-Modeling.xlsx

Biểu đồ Star Schema



### 3. Xây dựng kho dữ liệu bằng SSIS

#### 3.1. Load dữ liệu từ excel vào database Source

Dữ liệu sau khi được load vào SQL Server:

Object Explorer

```

SQLQuery1.sql - (L-3Q8RQB5\user (53)) 43 ×
43     ,[Turning_Loop]
44     ,[Sunrise_Sunset]
45     ,[Civil_Twilight]
46     ,[Nautical_Twilight]
47     ,[Astronomical_Twilight]
48     ,[Road_Surface_Conditions]
49     ,[Road_Type]
50     ,[Special_Conditions_at_Site]
51     ,[Speed_limit]
52     ,[Number_of_Casualties]
53   FROM [Accident_Source].[dbo].[Accident]
54

```

Results Messages

ID	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance_mi	Description	Number	Street
A-1	2	2016-02-08 00:37:00.0000000	2016-02-08 00:37:00.0000000	40.108906989235	-83.09285792602898	40.112060546875	-83.011857980957	3.23000001907249	Between Semmill Rd/Sect 20 and OH-315/Odentangy Riv	NULL	Out
A-10	2	2016-02-11 08:42:00.0000000	2016-02-11 14:42:00.0000000	41.1262397761113	-81.6529922485352	41.1192092895508	-81.6365814208984	0.829989925204468	At I-71/Exit 26 - Accident	NULL	Out
A-100	2	2016-03-23 07:04:00.0000000	2016-03-23 13:04:00.0000000	37.179104858380	-121.55943145752	37.17181386484375	-121.685351967676	0.34400001168251	At Flynn Rd - Accident Left lane blocked	NULL	I-77
A-1000	2	2016-05-26 10:23:00.0000000	2016-05-26 16:23:00.0000000	33.0725262897949	-117.07288611816	33.0558798217773	-117.089236755371	0.53200005245209	At Vaca Rancho Pky - Accident	NULL	I-58
A-10000	2	2016-05-20 16:18:00.0000000	2016-05-20 22:18:00.0000000	40.109539031982	-75.294853210492	40.110610961914	-75.286979675293	0.42898992024838	At Germantown Pike/Ext 333 - Accident	NULL	Penn
A-100000	2	2016-08-16 11:06:00.0000000	2021-08-16 13:06:00.0000000	28.4477424621582	-81.473915100077	28.45037002563	-81.474479675293	0.16200000476837	Slow traffic on FL-482 from Sand Lake Rd (FL-482) ext 1	NULL	I-4 E
A-1000001	2	2021-08-16 07:01:00.0000000	2021-08-16 19:18:00.0000000	39.3246063232422	-74.95037085742	39.3247438523438	-74.950544116211	0.01498999647239	MD DOT - TCC South Crash on NJ 73 northbound Remb.	1153	State
A-1000002	2	2021-08-21 00:52:00.0000000	2021-08-21 14:08:00.0000000	40.9033708537422	-111.141352942383	40.9028528500488	-111.1413543701211	0.07800001693254	Incident on MAIN ST SB near MAIN ST Drive with caution	301	S.Mr
A-1000003	2	2021-04-25 15:43:00.0000000	2021-04-25 18:01:00.0000000	43.5004310670791	-116.13646697998	43.4962805701661	-116.131912231445	0.365989966214	Incident on I-84 EB near BLACKS CREEK REST AREA E	NULL	I-84
A-1000004	2	2021-10-13 23:13:00.0000000	2021-10-14 00:28:00.0000000	32.90478515625	-96.7252197265625	32.8973274230957	-96.713973990234	0.830989970436096	Incident on LBJ FWY EB near FOREST LN Drive with co	NULL	I-83
A-1000005	2	2021-05-21 06:34:00.0000000	2021-05-21 18:21:00.0000000	33.8465134765625	-118.289184570313	33.8456802368184	-118.290367126465	0.081000002384186	Stationary traffic from I-10 S to W Benton St due to accid.	845	Del I
A-1000006	2	2021-06-06 18:30:00.0000000	2021-06-06 22:44:00.0000000	32.4378038530031	-106.514262226563	32.4382095336914	-106.47457286377	2.328989961653	Incident on US-70 near MM 169 Right lane blocked Exp	NULL	US-70
A-1000007	2	2021-10-16 07:45:00.0000000	2021-10-16 19:04:00.0000000	36.07423047019018	-88.981544494289	36.071578794922	-88.983024597168	0.201000002050679	Incident on COLEY DAVIS RD near POPLAR RIDGE DR	7701	Popl
A-1000008	4	2021-03-20 19:50:00.0000000	2021-03-20 21:53:00.0000000	39.10671997070	-76.9459304080957	39.109455108642	-76.9472732543945	0.11000001430511	MD 198 EAST/WEST AT LIONS DEN RD	3034	Wini
A-1000009	2	2021-04-16 18:02:00.0000000	2021-04-16 24:48:00.0000000	34.9407958984375	-82.262052590332	34.9408950805664	-82.2618026753398	0.01798999251598	Incident on WADE HAMPTON BLVD near DILL AVE Driv	NULL	Dill /
A-100001	2	2016-05-20 16:33:00.0000000	2016-05-20 22:33:00.0000000	42.4001197814941	-71.094238281216	42.3981592070998	-71.0920028866523	0.17700001027884	At I-93 Hwy Ln (North) - Accident	NULL	I-93

Query executed successfully.

Ready

SQLQuery2.sql - (L-3Q8RQB5\user (78)) 43 × SQLQuery1.sql - (L-3Q8RQB5\user (53))

```

1 SELECT TOP (1000) [ID]
2     ,[Age_Band_of_Driver]
3     ,[Age_Band_of_Vehicle]
4     ,[Driver_Home_Area_Type]
5     ,[Driver_IMD_Decile]
6     ,[Engine_Capacity_CC]
7     ,[Hit_Object_in_Carriageway]
8     ,[Hit_Object_off_Carriageway]
9     ,[Journey_Purpose_of_Driver]
10    ,[Junction_Location]
11    ,[make]
12    ,[model]
13    ,[Propulsion_Code]
14    ,[Sex_of_Driver]
15    ,[Skidding_and_Overtaking]
16    ,[Towing_and_Articulation]
17    ,[Vehicle_leaving_Carriageway]

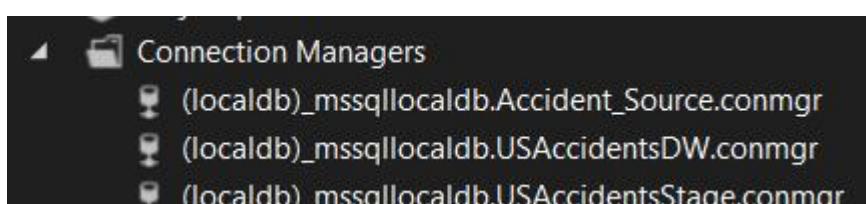
```

Results Messages

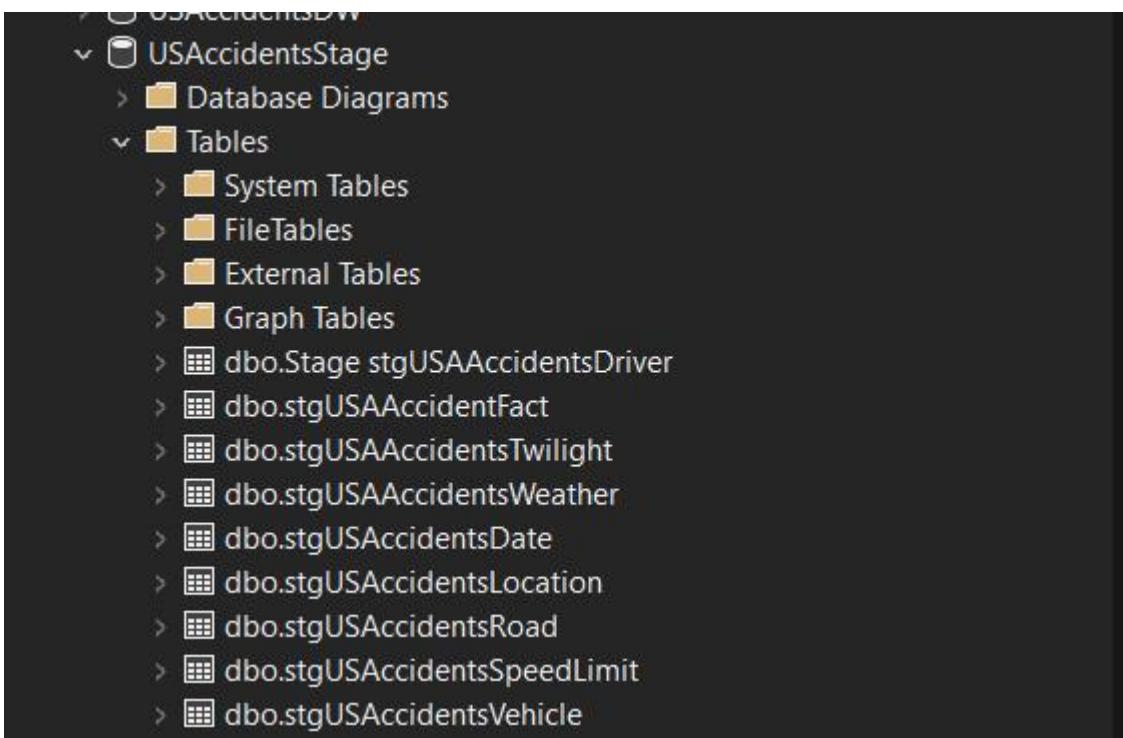
ID	Age_Band_of_Driver	Age_Band_of_Vehicle	Driver_Home_Area_Type	Driver_IMD_Decile	Engine_Capacity_CC	Hit_Object_in_Carriageway	Hit_Object_off_Carriageway	Journey_Purpose_of_Driver	Junction_Location	make
A-1	26 - 35	3	Urban area	4	1588	None	None	Data missing or out of range	Approaching junction or waiting/parked at junction	ROVER
A-2	26 - 35	37	Urban area	3	NULL	None	None	Data missing or out of range	Cleared junction or waiting/parked at junction exit	BMW
A-3	26 - 35	4	Rural	NULL	998	None	None	Data missing or out of range	Entering roundabout	NISSAN
A-4	66 - 75	95	Urban area	NULL	NULL	None	None	Data missing or out of range	Approaching junction or waiting/parked at junction	LONDON TAXI
A-5	26 - 35	1	Urban area	4	124	None	None	Data missing or out of range	Entering main road	PIAGGIO
A-6	36 - 45	10	Urban area	NULL	1781	None	None	Data missing or out of range	Not at or within 20 metres of junction	VOLKSWAGEN
A-7	26 - 35	28	Urban area	4	NULL	None	None	Data missing or out of range	Cleared junction or waiting/parked at junction exit	PIAGGIO
A-8	36 - 45	56	Urban area	8	NULL	None	None	Data missing or out of range	Leaving main road	BMW
A-9	46 - 55	3	Rural	NULL	2685	None	None	Data missing or out of range	Mid Junction - on roundabout or on main road	MERCEDES
A-10	26 - 35	4	Urban area	6	2300	None	None	Data missing or out of range	Entering main road	VOLKSWAGEN
A-11	21 - 25	3	Urban area	8	2402	None	None	Data missing or out of range	Leaving roundabout	FORD
A-12	36 - 45	2	Small town	NULL	8268	None	None	Data missing or out of range	Entering from slip road	DENIS
A-13	36 - 45	11	Urban area	NULL	998	None	None	Data missing or out of range	Leaving roundabout	VAUXHALL
A-14	21 - 25	6	Small town	NULL	NULL	None	None	Data missing or out of range	Entering from slip road	DENIS
A-15	66 - 75	1	Urban area	10	6570	None	None	Data missing or out of range	Leaving main road	MERCEDES
A-16	46 - 55	53	Urban area	NULL	NULL	None	None	Data missing or out of range	Not at or within 20 metres of junction	MAN
A-17	28 - 35	6	Urban area	6	1984	None	None	Data missing or out of range	Leaving main road	VOLKSWAGEN
A-18	16 - 20	85	Urban area	7	NULL	None	None	Data missing or out of range	Not at or within 20 metres of junction	PIAGGIO
A-19	66 - 75	6	Urban area	5	2684	None	None	Data missing or out of range	Mid Junction - on roundabout or on main road	LONDON TAXI
A-20	36 - 45	2	Rural	NULL	2979	None	None	Data missing or out of range	Entering from slip road	BMW

Query executed successfully.

### 3.2. Cấu hình connection tới Source, Stage và DataWarehouse



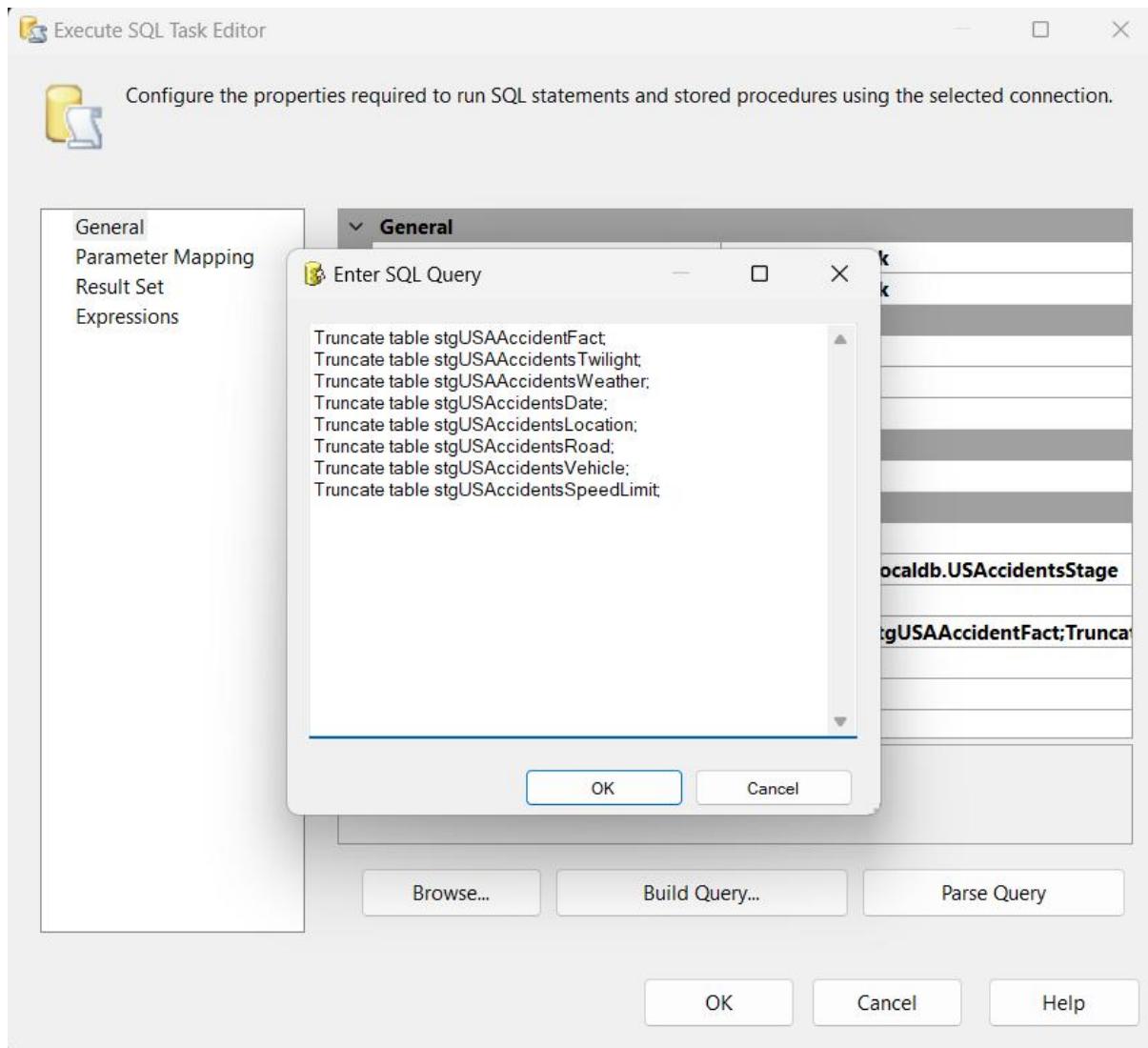
### 3.3. Tạo các table stage để chuẩn bị cho việc staging dữ liệu:



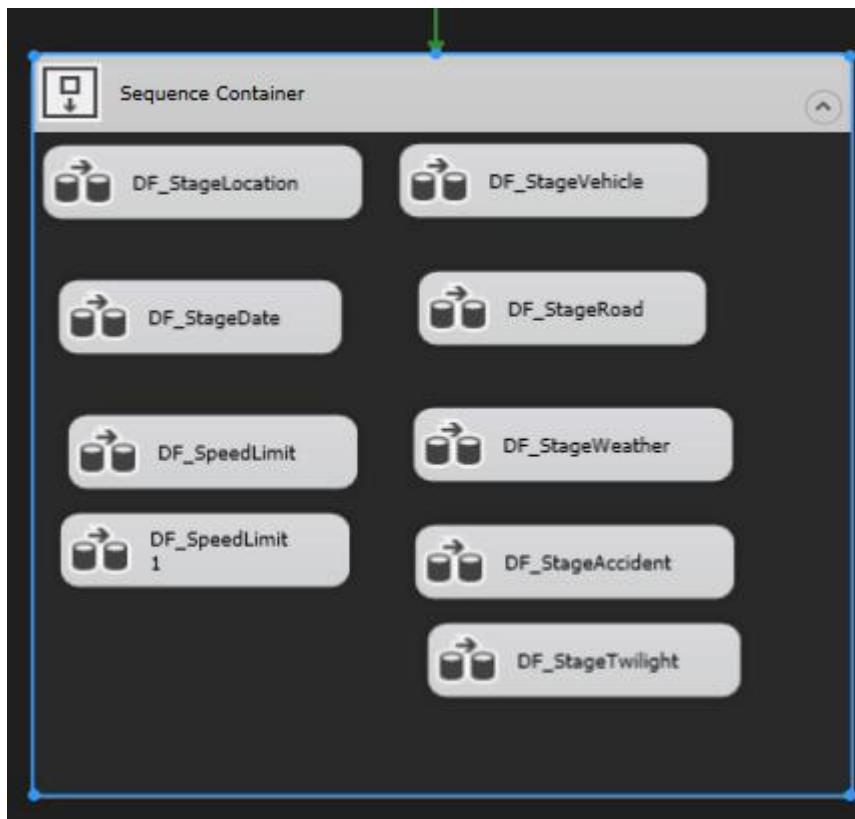
Lưu ý: Các bảng stage này sẽ được tạo trong bước chọn destination của dataflow “Source => Stage”

### 3.4. Load dữ liệu từ Source vào Stage cho các bảng Dim

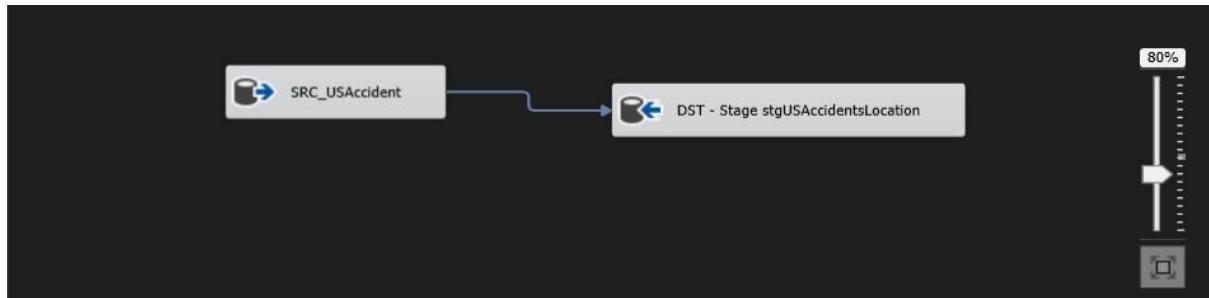
- Bước 1: Truncate tables trong Stage. Tạo một “Execute SQL Task Editor” để thực hiện truncate tất cả bảng trong stage.



- Bước 2: Tạo một “Sequence Container” chứa tất cả các “DataFlow” dùng để nạp dữ liệu từ Source vào.



Bước 3: Tạo DataFlow thêm dữ liệu từ Source vào Stage cho tất cả các bảng Dim.



Bước 4: Thực hiện tương tự cho các bảng Dim khác

### 3.5. Thực hiện tạo stage cho bảng Fact

Việc tạo bảng stage cho Fact được lấy dữ liệu từ Source và có truy vấn SQL như sau:

```

SELECT
    a.ID as LocationID,
    CAST(YEAR(start_time) AS VARCHAR(4)) +
    RIGHT('0' + CAST(MONTH(start_time) AS VARCHAR(2)), 2) +
    RIGHT('0' + CAST(DAY(start_time) AS VARCHAR(2)), 2) +
    RIGHT('0' + CAST(DATEPART(HOUR, start_time) AS VARCHAR(2)), 2) +
    RIGHT('0' + CAST(DATEPART(MINUTE, start_time) AS VARCHAR(2)), 2) +
    AS DateID,

```

```

a.ID as TwilightID,
a.ID as RoadFeatureID,
a.ID as WeathertID,
a.ID as DriverID,
a.ID as SpeedLimitID,
v.ID as VehicleID,
a.Number_of_Casualties as NumberOfCasualties,
v.Vehicle_Reference as NumberOfVehicles,
a.Severity as AccidentSeverity,
a.Distance_mi as Distance,
    a.Visibility_mi as Visibility,
DATEDIFF(MINUTE, Start_Time, End_Time) AS Duration_Minutes,

```

CASE

WHEN Road\_Surface\_Conditions = 'Dry' THEN 1 -- Điều kiện đường khô ráo, ít nguy hiểm

WHEN Road\_Surface\_Conditions = 'Wet or damp' THEN 2 -- Đường ướt hoặc ẩm, có nguy cơ trơn trượt

WHEN Road\_Surface\_Conditions = 'Frost or ice' THEN 3 -- Có băng hoặc sương giá, nguy hiểm hơn

WHEN Road\_Surface\_Conditions = 'Snow' THEN 4 -- Tuyết phủ đường, hạn chế tầm nhìn & kiểm soát xe

WHEN Road\_Surface\_Conditions = 'Flood over 3cm. deep' THEN 5 -- Đường ngập sâu, rất nguy hiểm

ELSE 0 -- Data missing or out of range

END AS RoadConditionSeverity,

CASE

-- Ít ảnh hưởng (1)

WHEN Weather\_Condition LIKE '%Clear%' OR

Weather\_Condition LIKE '%Fair%' THEN 1

-- Ảnh hưởng nhẹ (2)

WHEN Weather\_Condition LIKE '%Cloudy%' OR

Weather\_Condition LIKE '%Partly Cloudy%' OR

Weather\_Condition LIKE '%Scattered Clouds%' OR

Weather\_Condition LIKE '%Mostly Cloudy%' OR

Weather\_Condition LIKE '%Haze%' THEN 2

-- Ảnh hưởng trung bình (3)

WHEN Weather\_Condition LIKE '%Mist%' OR

Weather\_Condition LIKE '%Drizzle%' OR

Weather\_Condition LIKE '%Rain Shower%' OR

Weather\_Condition LIKE '%Light Rain%' OR

Weather\_Condition LIKE '%Fog%' THEN 3

-- Ánh hưởng lớn (4)

WHEN Weather\_Condition LIKE '%Heavy Rain%' OR

Weather\_Condition LIKE '%Snow%' OR

Weather\_Condition LIKE '%Wintery Mix%' OR

Weather\_Condition LIKE '%Ice Pellets%' OR

Weather\_Condition LIKE '%Sand%' OR

Weather\_Condition LIKE '%Dust%' OR

Weather\_Condition LIKE '%Blowing Snow%' OR

Weather\_Condition LIKE '%Freezing Rain%' THEN 4

-- Ánh hưởng nghiêm trọng (5)

WHEN Weather\_Condition LIKE '%Thunder%' OR

Weather\_Condition LIKE '%T-Storm%' OR

Weather\_Condition LIKE '%Squalls%' OR

Weather\_Condition LIKE '%Heavy Thunderstorms%' OR

Weather\_Condition LIKE '%Tornado%' OR

Weather\_Condition LIKE '%Volcanic Ash%' THEN 5

-- Không xác định (0)

ELSE 0

END AS WeatherImpactSeverity

FROM Accident a

JOIN Vehicle v

ON a.ID = v.ID

Kết quả của bảng Stage FactAccident:

Result	Messages	LocationID	DateTime	TwilightID	RoadFeatureID	WeatherID	DriverID	SpeedLimitID	VehicleID	NumberOfCasualties	NumberOfVehicles	AccidentSeverity	Distance	Visibility	Duration_Minutes	RoadConditionSeverity	WeatherImpactSeverity
1	A-1	201602080037	201602080037	A-1	A-1	A-1	A-1	A-1	A-1	1	2	3	3.23000001907349	10	360	2	3
2	A-10	201602081516	201602081516	A-10	A-10	A-10	A-10	A-10	A-10	5	1	2	0.82599987504465	0.5	360	1	4
3	A-100	201602110842	201602110842	A-100	A-100	A-100	A-100	A-100	A-100	1	2	2	0.98299980449677	9	360	2	0
4	A-1000	201603207024	201603207024	A-1000	A-1000	A-1000	A-1000	A-1000	A-1000	1	2	2	0.34400001168251	10	360	1	1
5	A-10000	201605261023	201605261023	A-10000	A-10000	A-10000	A-10000	A-10000	A-10000	1	1	2	0.532000005245209	10	360	2	0
6	A-100000	201605261618	201605261618	A-100000	A-100000	A-100000	A-100000	A-100000	A-100000	1	2	2	0.42899990224838	10	360	2	1
7	A-1000000	202108161106	202108161106	A-1000000	A-1000000	A-1000000	A-1000000	A-1000000	A-1000000	1	3	2	0.16200000476837	10	122	1	2
8	A-1000001	202108061701	202108061701	A-1000001	A-1000001	A-1000001	A-1000001	A-1000001	A-1000001	1	1	2	0.01499999647239	10	137	2	1
9	A-1000002	202106210052	202106210052	A-1000002	A-1000002	A-1000002	A-1000002	A-1000002	A-1000002	1	1	2	0.0780000016093254	10	796	1	0
10	A-1000003	202104251543	202104251543	A-1000003	A-1000003	A-1000003	A-1000003	A-1000003	A-1000003	1	1	2	0.3659999966214	10	138	1	1
11	A-1000004	202110132313	202110132313	A-1000004	A-1000004	A-1000004	A-1000004	A-1000004	A-1000004	1	2	2	0.830999970436096	7	75	1	0
12	A-1000005	202105210634	202105210634	A-1000005	A-1000005	A-1000005	A-1000005	A-1000005	A-1000005	2	3	2	0.081000002384186	10	707	1	1
13	A-1000006	202106061830	202106061830	A-1000006	A-1000006	A-1000006	A-1000006	A-1000006	A-1000006	1	2	2	2.328999961853	10	254	1	1
14	A-1000007	202110180629	202110180629	A-1000007	A-1000007	A-1000007	A-1000007	A-1000007	A-1000007	1	2	2	0.201000002600679	10	76	1	1
15	A-1000008	202103201950	202103201950	A-1000008	A-1000008	A-1000008	A-1000008	A-1000008	A-1000008	1	3	4	0.111000001430511	10	123	1	1
16	A-1000009	202104161802	202104161802	A-1000009	A-1000009	A-1000009	A-1000009	A-1000009	A-1000009	1	2	2	0.0179999992251393	10	226	1	1
17	A-100001	201605201633	201605201633	A-100001	A-100001	A-100001	A-100001	A-100001	A-100001	1	1	2	0.177000001072884	10	360	1	2
18	A-1000010	202105050554	202105050554	A-1000010	A-1000010	A-1000010	A-1000010	A-1000010	A-1000010	3	2	2	2.09699998365173	9	244	1	1
19	A-1000011	202112231341	202112231341	A-1000011	A-1000011	A-1000011	A-1000011	A-1000011	A-1000011	1	3	2	1.644000005340576	10	302	2	1
20	A-1000012	202103081414	202103081414	A-1000012	A-1000012	A-1000012	A-1000012	A-1000012	A-1000012	2	1	2	0.0469999983906748	10	117	2	1
21	A-1000013	20210308695	20210308695	A-1000013	A-1000013	A-1000013	A-1000013	A-1000013	A-1000013	1	1	2	0.0469999983906748	10	75	1	2
22	A-1000014	202112231939	202112231939	A-1000014	A-1000014	A-1000014	A-1000014	A-1000014	A-1000014	2	1	2	0.535000026226044	2	126	1	4
23	A-1000015	202105041543	202105041543	A-1000015	A-1000015	A-1000015	A-1000015	A-1000015	A-1000015	1	2	2	0.39199989748001	10	99	1	1
24	A-1000016	202109231258	202109231258	A-1000016	A-1000016	A-1000016	A-1000016	A-1000016	A-1000016	2	2	2	3.00699998648242	4	163	2	5
25	A-1000017	202105190511	202105190511	A-1000017	A-1000017	A-1000017	A-1000017	A-1000017	A-1000017	1	1	2	0.0900000035762787	10	1080	1	1
26	A-1000018	202109291446	202109291446	A-1000018	A-1000018	A-1000018	A-1000018	A-1000018	A-1000018	1	3	2	1.5549994754791	10	96	2	1
27	A-1000019	202105061951	202105061951	A-1000019	A-1000019	A-1000019	A-1000019	A-1000019	A-1000019	1	1	2	0.788999974727631	10	74	1	1
28	A-100002	201605201644	201605201644	A-100002	A-100002	A-100002	A-100002	A-100002	A-100002	1	1	4	0.95300000095906	10	360	1	1
29	A-1000022	202111300853	202111300853	A-1000020	A-1000020	A-1000020	A-1000020	A-1000020	A-1000020	1	1	2	0.3089998546467	10	125	2	1
30	A-1000021	202108160823	202108160823	A-1000021	A-1000021	A-1000021	A-1000021	A-1000021	A-1000021	3	1	2	0.430000007152557	8	127	2	1
31	A-1000022	202112132151	202112132151	A-1000022	A-1000022	A-1000022	A-1000022	A-1000022	A-1000022	1	2	2	0.66299987602234	10	76	1	1
32	A-1000023	202112031736	202112031736	A-1000023	A-1000023	A-1000023	A-1000023	A-1000023	A-1000023	1	1	2	0.028000000642673	10	78	1	1
33	A-1000024	202111131615	202111131615	A-1000024	A-1000024	A-1000024	A-1000024	A-1000024	A-1000024	1	2	2	1.070999979727284	10	78	2	1
34	A-1000025	202107020744	202107020744	A-1000025	A-1000025	A-1000025	A-1000025	A-1000025	A-1000025	1	3	2	4.15999987471211	10	223	2	2
35	A-1000026	202111132234	202111132234	A-1000026	A-1000026	A-1000026	A-1000026	A-1000026	A-1000026	1	1	2	3.6300001444092	10	117	2	1

### 3.6. Đỗ dữ liệu từ stage vào các bảng Dim và Fact

Sau khi Run trong SSIS và thành công ta sẽ thực hiện kiểm tra bằng cách vào SQL Server và in 5 dòng đầu tiên của tất cả các bảng.

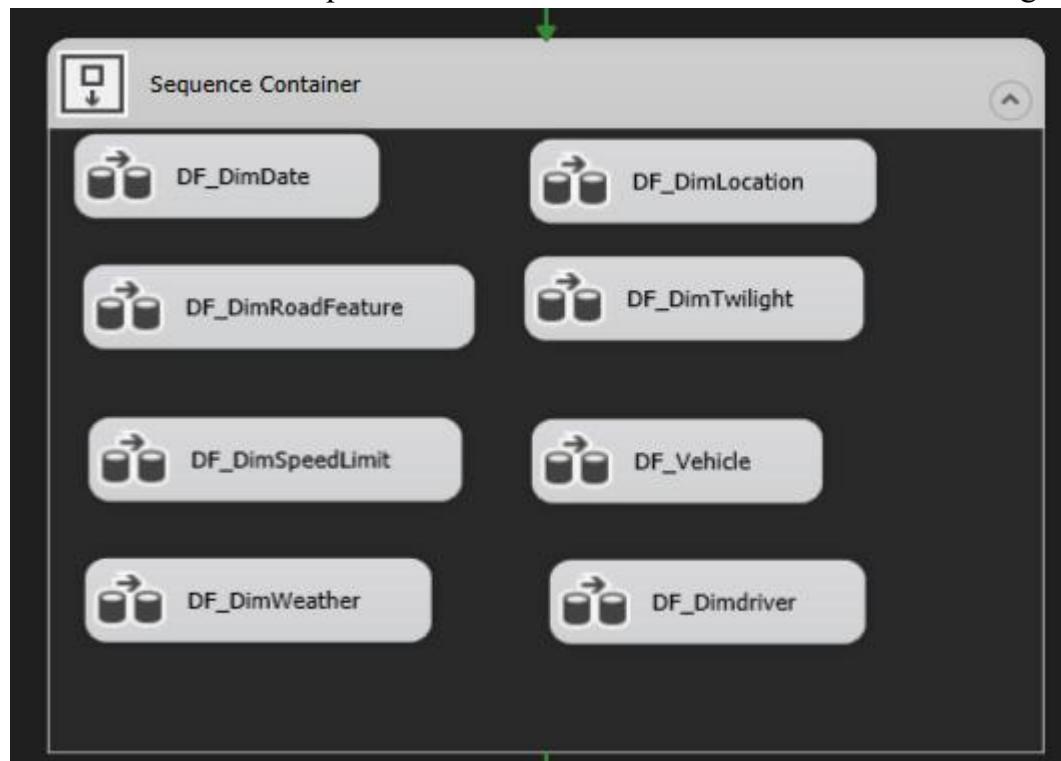
The screenshot shows a results grid with five tables:

- VehicleKey**: Columns include VehicleKey, ID, EngineCapacityCC, PropulsionCode, Make, Model, JunctionLocation, VehicleManoeuvre, VehicleLocationRestrictedLane, X1stPointofImpact, WasVehicleLeftHandDrive, Year. Data rows show vehicles like A-417972 (Petrol, HYUNDAI ACCENT COUPE I), A-417973 (Heavy oil, MERCEDES A180 CDI ELEGANCE SE), etc.
- WeatherKey**: Columns include WeatherKey, ID, Temperature(F), Wind\_Chill(F), Humidity(%), Pressure(in), Visibility(m), WindDirection, WindSpeed(mph), Precipitation(in), WeatherCondition. Data rows show weather conditions like Clear, Light Rain, Partly Cloudy, Fair, Mostly Cloudy.
- RoadFeatureKey**: Columns include ID, LocationKey, DateKey, TwilightKey, RoadFeatureKey, WeatherKey, DriverKey, SpeedLimitKey, VehicleKey, NumberOfCasualties, NumberOfVehicles, AccidentSeverity, Distance, Visibility, Duration\_Minutes, RoadConditionSeverity. Data rows show road features and driver statistics.
- DateKey**: Columns include DateKey, FullDate, Year, Quarter, Month, Day, Hour, Minute, Weekday. Data rows show dates from 2016-01-14 to 2016-02-08.
- LocationKey**: Columns include ID, Number, Street, Side, City, County, State, Zipcode, Country. Data rows show locations like Raton, Colfax, NM, US.

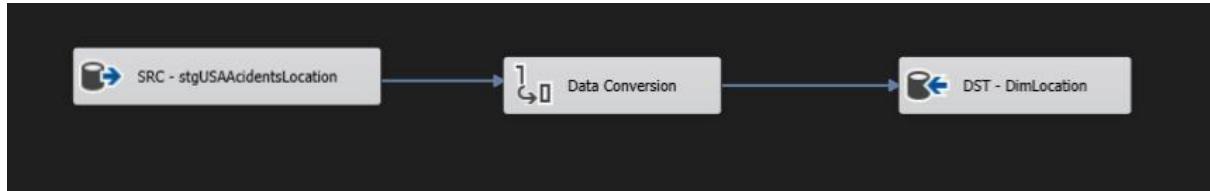
Từ kết quả cho thấy, quy trình nạp dữ liệu từ Source sang Stage được diễn ra thành công.

### 3.7. Nạp dữ liệu từ Stage vào DataWarehouse cho các bảng Dim

Bước 1: Tạo sequence container chứa tất cả data flow của các bảng Dim



Bước 2: Cấu hình cho mỗi dataflow lấy dữ liệu từ Stage và nạp vào DW. Dùng Data Conversion để chuyển đổi kiểu dữ liệu phù hợp.



Bước 4: Thực hiện tương tự cho các DF Dim khác

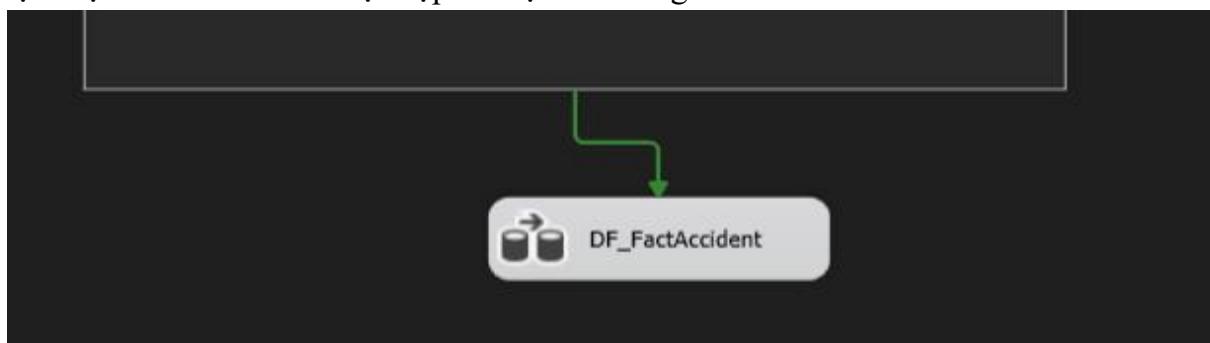
### 3.8. Nạp dữ liệu FactAccident

Bước 1: Kiểm tra dữ liệu các bảng Dim đã được nạp vào thành công chưa:

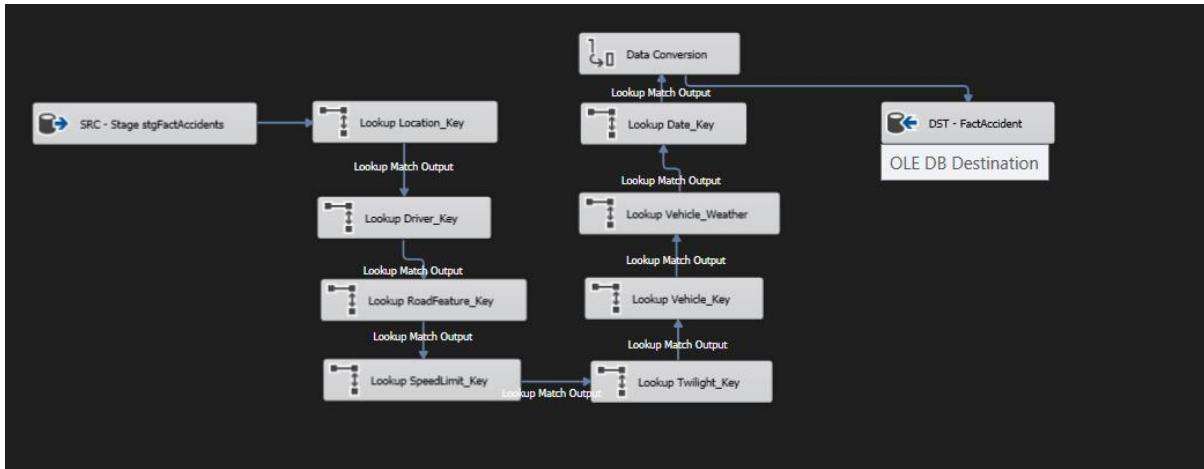
SQLQuery1.sql - (l...-3Q8RQB5\user (59)) [StageFact.sql - (l...-3Q8RQB5\user (60)*]														
Results		Messages												
	VehicleKey	ID	EngineCapacityCC	PropulsionCode	Make	Model	JunctionLocation			VehicleManoeuvre	Vehicle.LocationRestrictedLane	X1stPointofImpact	WasVehicleLeftHandDrive	Year
1	1	A-17972	1341	Petrol	HYUNDAI	ACCENT COUPE I	Mid Junction - on roundabout or on main road	Going ahead other	0	Front	No	2007		
2	2	A-17973	1891	Heavy oil	MERCEDES	A180 CDI ELEGANCE SE	Approaching junction or waiting/parked at junction	Going ahead other	0	Front	No	2007		
3	3	A-17974	1783	Petrol	VOLVO	V40 I	Mid Junction - on roundabout or on main road	Waiting to turn right	0	Back	No	2007		
4	4	A-17975	1783	Petrol	VOLVO	V40 S	Entering main road	Going ahead other	0	Front	No	2007		
5	5	A-17976	1870	Heavy oil	VOLVO	V40 D SE	Mid Junction - on roundabout or on main road	Slowing or stopping	0	Back	No	2007		
	WeatherKey	ID	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	WindDirection	WindSpeed(mph)	Precipitation(in)	WeatherCondition			
1	A-23141	1	75.9000001528789	63	30	30.0400009155273	10	South	19.6000003814697	0	Clear			
2	2	A-231410	54	54	93	28.700004577637	4	SW	6	0.00999999977648288	Light Rain			
3	3	A-231411	57	57	94	29.9400005340576	8	CALM	0	0	Partly Cloudy			
4	4	A-231412	79	79	28	29.25	10	VAR	5	0	Fair			
5	5	A-231413	34	30	87	29.599998474121	10	NNW	5	0	Mostly Cloudy			
	ID	LocationKey	DateKey	TwilightKey	RoadFeatureKey	WeatherKey	DriverKey	DriverLimitKey	VehicleKey	NumberOfCasualties	NumberOfVehicles	AccidentSeverity	Distance	Visibility
1	A-128674	985488	201603311529	829494	882092	864970	995541	902924	103395	2	2	0	360	1
2	2	A-128675	985489	201603311540	829495	882093	864971	995542	902925	103396	1	1	0.6219...	10
3	3	A-128676	985490	201603311536	829496	882094	864972	995543	902926	103397	2	2	0.4510...	10
4	4	A-128677	985491	201603311542	829497	882095	864973	995544	902927	103398	1	1	0	10
5	5	A-128678	985492	201603311549	829498	882096	864974	995545	902928	103399	1	1	0.6639...	10
	name	principal_id	diagram_id	version	definition									
	DateKey	FullDate	Year	Quarter	Month	Day	Hour	Minute	Weekday					
1	201601142018	2016-01-14	2016	1	1	14	20	18	5					
2	201602080037	2016-02-08	2016	1	2	8	0	37	2					
3	201602080558	2016-02-08	2016	1	2	8	5	56	2					
4	201602080615	2016-02-08	2016	1	2	8	6	15	2					
5	201602080651	2016-02-08	2016	1	2	8	6	51	2					
	LocationKey	ID	Number	Street	Side	City	County	State	Zipcode	Country				
1	A-148060	NULL	I-25 N	R	Relon	Coffey	NM	87740	US					
2	2	A-148061	NULL	I-45 N	R	Houston	Harris	TX	77017	US				
3	3	A-148062	NULL	I-61	R	Houston	Harris	TX	77026	US				
4	4	A-148063	NULL	I-45 N	R	Houston	Harris	TX	77037	US				
5	5	A-148064	6242	Nort	R	Fort W.	Tarrant	TX	76137	US				
	DriverKey	ID	AgeBandOffDriver	SexOffDriver	DriverIMDDecile	DriverHomeAreaType	JourneyPurposeofDriver							
1	1	A-142015	36 - 45	Male	NULL	Urban area	Other/Not known (2005-10)							
2	2	A-142016	56 - 65	Male	NULL	Urban area	Journey as part of work							
3	3	A-142017	Over 75	Male	NULL	Urban area	Other/Not known (2005-10)							
4	4	A-142018	26 - 35	Male	NULL	Urban area	Journey as part of work							

Query executed successfully.

Bước 2: Sau khi xác nhận dữ liệu các bảng Dim trong DW đã được hoàn thiện, tạo một Data Flow cho việc nạp dữ liệu của bảng Fact.



Bước 3: Cấu hình Data flow để nạp dữ liệu. Nguồn là Stage và đích là DW. Trong quá trình đó chúng ta phải dùng đến lookup để lấy các giá trị key cho mỗi Dimension.



### Bước 5: Sau khi chạy thành công, kiểm tra lại trong SQL Server

SQLQuery1.sql - (L-3Q8RQB5user (59)) \* X StageFact.sql - (L-3Q8RQB5user (60)) \* X

```

1 | select * from FactAccident

```

Results Messages

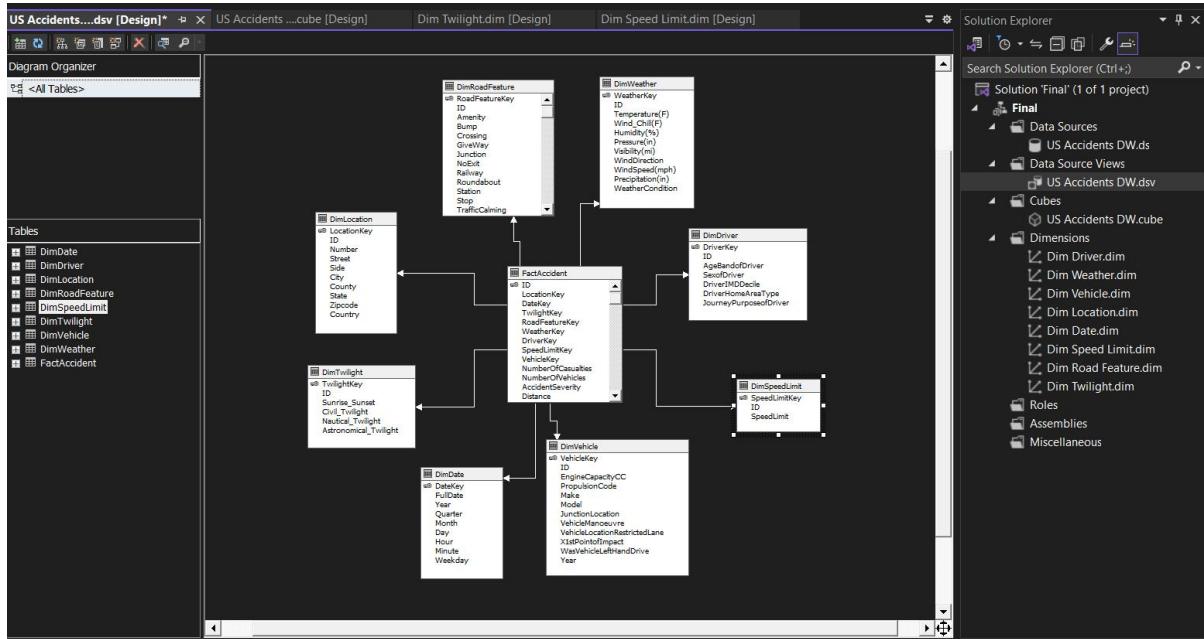
ID	LocationKey	DateKey	TwilightKey	RoadFeatureKey	WeatherKey	DriverKey	SpeedLimitKey	VehicleKey	NumberofCasualties	NumberOfVehicles	AccidentSeverity	Distance	Visibility	Duration_Minutes	RoadConditionSeverity	
1	A-128674	985488	201608311529	829494	882092	864970	995541	90224	103395	1	2	0	10	360	1	
2	A-128675	985489	201608311540	829495	882093	864971	995542	90225	103396	1	2	0	10	360	1	
3	A-128676	985490	201608311536	829496	882094	864972	995543	90226	103397	2	2	0	10	360	1	
4	A-128677	985491	201608311542	829497	882095	864973	995544	90227	103398	1	2	0	10	360	1	
5	A-128678	985492	201608311549	829498	882096	864974	995545	90228	103399	1	1	2	0	10	360	1
6	A-128679	985493	201608311550	829499	882097	864975	995546	90229	103400	3	1	2	0	10	360	1
7	A-128680	985494	201612091800	829500	882098	864976	458039	90230	844373	2	1	3	0	10	360	2
8	A-128680	985495	201608311559	829501	882099	864977	995547	90231	103401	1	1	2	0	10	360	1
9	A-128681	985496	201608311602	829502	882100	864978	995548	90232	103402	1	2	2	0	10	360	1
10	A-128682	985497	201608311559	829503	882101	864979	995549	90233	103403	2	1	2	0	10	360	1
11	A-128683	985498	201608311603	829504	882102	864980	995550	90234	103404	1	2	2	0	10	360	1
12	A-128684	985499	201608311601	829505	882103	864981	995551	90235	103405	1	1	2	0	10	360	1
13	A-128685	985500	201608311624	829506	882104	864982	995552	90236	103406	1	1	2	0	10	360	1
14	A-128686	985501	201608311620	829507	882105	864983	995553	90237	103407	1	1	2	0	10	360	1
15	A-128687	985502	201608311624	829508	882106	864984	995554	90238	103408	2	1	2	0	10	360	1
16	A-128688	985503	201608311635	829509	882107	864985	995555	90239	103409	1	5	2	0	10	360	1
17	A-128689	985504	201608311635	829510	882108	864986	995556	90240	103410	1	6	2	0	10	360	1
18	A-128690	985505	201612091757	829511	882109	864987	458040	90241	844374	1	1	2	0	10	360	1
19	A-128690	985506	201608311633	829512	882110	864988	995557	90242	103411	1	8	2	0	10	360	1

Executing query... (localdb)\mssqllocaldb (15...) DESKTOP-3Q8RQB5\user (60) USAccidentsDW 00:00:03 0 rows

## 4. Tạo Cube và thực hiện truy vấn bằng SSAS

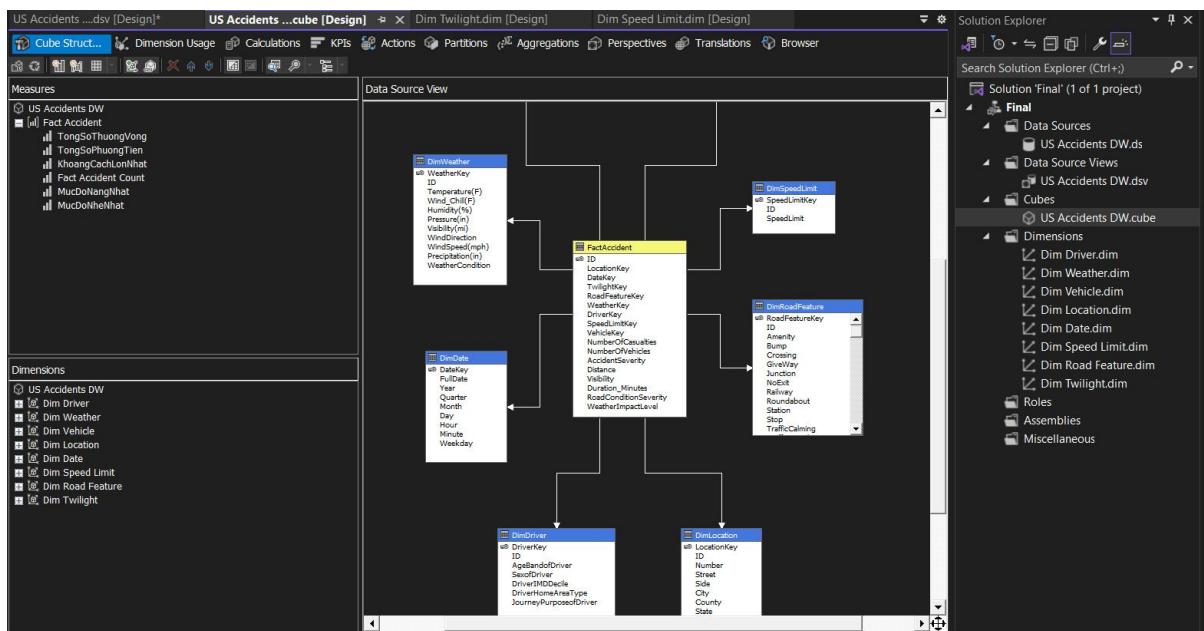
### 4.1. Tạo Data Source từ kho dữ liệu database DWAccident

### 4.2. Tạo Data Source View



### 4.3. Tạo cube, thêm measure và các dim cần thiết, tạo các phân cấp

#### Cube



#### Phân cấp

The screenshot displays two windows of the Microsoft Analysis Services (SSAS) Dimension Designer:

- Dim Date.dim [Design] Window:**
  - Attributes:** Shows attributes for Dim Date, including Date Key, Day, Hour, Minute, Month, Quarter, Weekday, and Year.
  - Hierarchies:** Shows three hierarchies: DateHierarchy, Y-Q, and TimeHierarchy. DateHierarchy has levels Year, Month, and Day. Y-Q has levels Year, Quarter, and <new level>. TimeHierarchy has levels Hour and Minute.
  - Data Source View:** Shows the DimDate table with attributes FullDate, Year, Quarter, Month, Day, Hour, Minute, and Weekday.
- Solution Explorer:** Shows the project 'Final' containing Data Sources (US Accidents DW.ds), Data Source Views (US Accidents DW.dsv), Cubes (US Accidents DW.cube), Dimensions (Dim Driver.dim, Dim Weather.dim, Dim Vehicle.dim, Dim Location.dim, Dim Date.dim, Dim Speed Limit.dim, Dim Road Feature.dim, Dim Twilight.dim), Roles, Assemblies, and Miscellaneous.
- Dim Location.dim [Design] Window:**
  - Attributes:** Shows attributes for Dim Location, including City, County, and Location Key.
  - Hierarchies:** Shows one hierarchy named Hierarchy with levels State, County, and City. A placeholder message says "To create a new hierarchy, drag an attribute here."
  - Data Source View:** Shows the DimLocation table with attributes LocationKey, ID, Number, Street, Side, City, County, State, Zipcode, and Country.

#### 4.4. Truy vấn các câu hỏi mà nhóm đưa ra bằng công cụ SSAS,

Danh sách câu hỏi:

- Thống kê số vụ tai nạn giao thông và thương vong theo năm từ 2016 đến 2023 là bao nhiêu?

Screenshot of the SSAS Model Designer interface showing the 'US Accidents ...cube [Design]' tab selected. The left pane displays the 'Metadata' tree with nodes like 'Fact Accident', 'KPIs', and 'Dim Date'. The right pane shows a table with columns 'Year', 'Fact Accident Count', and 'TongSoThuongVong'.

Year	Fact Accident Count	TongSoThuongVong
2016	64320	86930
2017	113732	150621
2018	1	1
2019	13338	18642
2020	13	21
2021	244753	336860
2023	13213	18383

2. Thống kê số vụ tai nạn giao thông , thương vong, tổng số phương tiện liên quan theo từng quý qua các năm

Screenshot of the SSAS Model Designer interface showing the 'US Accidents ...cube [Design]' tab selected. The left pane displays the 'Metadata' tree with nodes like 'US Accidents DW', 'Measures', and 'Dim Date'. The right pane shows a table with columns 'Year', 'Quarter', 'Fact Accident Count', 'TongSoThuongVong', and 'TongSoPhuongTien'.

Year	Quarter	Fact Accident Count	TongSoThuongVong	TongSoPhuongTien
2016	1	1053	1250	1490
2016	2	10948	14058	16656
2016	3	19457	26150	29348
2016	4	32862	45472	50442
2017	1	48826	65303	76385
2017	2	45848	60476	73224
2017	3	15806	20342	24839
2017	4	3252	4500	5040
2018	3	1	1	1
2019	1	3323	4660	5107
2019	2	3270	4573	5141
2019	3	3350	4703	5184
2019	4	3395	4706	5285
2020	3	1	5	1
2020	4	12	16	17
2021	1	15231	20917	23752
2021	2	51111	70195	79967
2021	3	67483	93061	105596
2021	4	110928	152687	173670
2023	1	3315	4569	5183
2023	2	3276	4447	5089
2023	3	3280	4616	5137
2023	4	3342	4751	5193

3. Thông kê số vụ tai nạn giao thông và thương vong theo ngày trong tuần (1-Chủ nhật ...; 7 - Thứ 7)

Weekday	Fact Accident Count	TongSoThuongVong
1	36587	49956
2	67833	92084
3	72940	99007
4	73588	100149
5	75105	102300
6	80549	109130
7	42768	58832

4. Số vụ tai nạn giao thông theo giờ trong ngày (0-23h) thay đổi như thế nào?

Hour	Fact Accident Count
0	58805
1	4728
10	16616
11	17593
12	21734
13	22672
14	27515
15	32402
16	33228
17	33917
18	24522
19	15358
2	4595
20	10415
21	8533
22	7769
23	6765
3	5079
4	6118
5	11133
6	17313
7	22457
8	22219
9	17884

## 5. Top 3 giờ xảy ra nhiều vụ tai nạn nhất trong ngày

The screenshot shows the SSAS Cube Design interface. In the top navigation bar, the active tab is "US Accidents ...cube [Design]". The main area displays a table titled "Fact Accident Count" with two columns: "Hour" and "Fact Accident Count". The data shows three rows: Hour 0 with 58805 accidents, Hour 16 with 33228 accidents, and Hour 17 with 33917 accidents. A filter expression "Top3\_Tai\_Nan\_Nhieu\_Nhat" is applied to the "Hour" dimension. The left sidebar shows the cube's structure, including dimensions like "Dim Date" and "Fact Accident", and measures like "Fact Accident Count".

Hour	Fact Accident Count
0	58805
16	33228
17	33917

## 6. Top 3 giờ xảy ra ít vụ tai nạn nhất trong ngày

The screenshot shows the SSAS Cube Design interface. In the top navigation bar, the active tab is "US Accidents ...cube [Design]". The main area displays a table titled "Fact Accident Count" with two columns: "Hour" and "Fact Accident Count". The data shows three rows: Hour 1 with 4728 accidents, Hour 2 with 4595 accidents, and Hour 3 with 5079 accidents. A filter expression "Top3\_Tai\_Nan\_It\_Nhat" is applied to the "Hour" dimension. The left sidebar shows the cube's structure, including dimensions like "Dim Date" and "Fact Accident", and measures like "Fact Accident Count".

Hour	Fact Accident Count
1	4728
2	4595
3	5079

## 7. Thống kê số vụ tai nạn ở nông thôn, thành thị.

8. Thống kê tổng quãng đường bị ảnh hưởng qua từng năm của 3 tiểu bang (LA, NY, WA)

State	Year	TongQuangDuongAnhHuong
CA	2016	13791.5930054006
CA	2017	14476.0050031974
CA	2019	2757.24599723658
CA	2020	0.280000001192093
CA	2021	49170.256009966
CA	2023	2523.6310012266
NY	2016	1373.17200154834
NY	2017	5073.30699580302
NY	2019	491.760000913404
NY	2021	8172.24300195905
NY	2023	465.752000045963
WA	2016	1460.86099597125
WA	2017	1406.77400184702
WA	2019	155.592998762615
WA	2021	2108.78300128225
WA	2023	141.474999728613

9. Thống kê số người tử vong, số vụ tai nạn theo loại đường và tốc độ từ 30-70mi/h

Screenshot of the SSAS Management Studio interface showing the US Accidents cube [Design] tab. The left pane displays the model structure with nodes like Dim Twilight.dim [Design], Dim Date.dim [Design], Dim Location.dim [Design], and US Accidents ...cube [Design]. The right pane shows a table with data filtered by Speed Limit between 30 and 70.

Dimension	Hierarchy	Operator	Filter Expression	Param...
Dim Speed Limit	Speed Limit	Range (Inclusive)	30 : 70	
<Select dimension>				
Road Type	Speed Limit	Fact Accident Count	TongSoThuongVong	
Dual carriageway	30	19288	26130	
Dual carriageway	40	10207	14734	
Dual carriageway	50	5086	7420	
Dual carriageway	60	3511	5259	
Dual carriageway	70	30112	47860	
One way street	30	9280	11135	
One way street	40	172	244	
One way street	50	73	99	
One way street	60	212	276	
One way street	70	115	169	
Roundabout	30	17723	22121	
Roundabout	40	4221	5305	
Roundabout	50	1112	1424	
Roundabout	60	3831	4917	
Roundabout	70	1984	2576	
Single carriageway	30	236811	302045	
Single carriageway	40	21077	30952	
Single carriageway	50	6078	9593	
Single carriageway	60	67538	105010	
Single carriageway	70	2	2	
Slip road	30	1400	1791	
Slip road	40	300	402	
Slip road	50	348	492	
Slip road	60	632	910	

## 10. Thông kê số vụ tai nạn, thương vong theo điều kiện thời tiết

Screenshot of the SSAS Management Studio interface showing the US Accidents cube [Design] tab. The left pane displays the model structure with nodes like Dim Twilight.dim [Design], Dim Date.dim [Design], Dim Location.dim [Design], and US Accidents ...cube [Design]. The right pane shows a table with data filtered by Weather Condition.

Dimension	Hierarchy	Operator	Filter Expression	Param...
<Select dimension>				
Weather Condition	Fact Accident Count	TongSoThuongVong		
Blowing Dust	3	6		
Blowing Dust / Windy	13	20		
Blowing Snow	17	21		
Blowing Snow / Windy	4	7		
Clear	61011	81066		
Cloudy	37155	51065		
Cloudy / Windy	686	930		
Drizzle	246	357		
Drizzle / Windy	3	3		
Drizzle and Fog	12	15		
Dust Whirls	1	1		
Fair	139521	192452		
Fair / Windy	1659	2240		
Fog	4720	6527		
Fog / Windy	34	43		
Freezing Rain / Windy	1	3		
Funnel Cloud	1	7		
Haze	5659	7749		
Haze / Windy	73	90		
Heavy Drizzle	25	35		
Heavy Ice Pellets	2	2		
Heavy Rain	1919	2605		
Heavy Rain / Windy	99	140		
Heavy Rain.Shower / Windy	1	1		

## 11. Thông kê số vụ tai nạn và tổng số thương vong theo DateHierarchi (Y-M-D)

Screenshot of the SSAS Dimension Designer interface showing the 'Dim Speed Limit.dim [Design]' tab selected. The left pane displays the dimension structure with nodes like 'KPIs', 'Dim Date', and 'Dim Driver'. The right pane shows a data grid with columns: Year, Month, Day, Fact Accident Count, and TongSoThuongVong. The data is as follows:

Year	Month	Day	Fact Accident Count	TongSoThuongVong
2016	9	27	400	525
2016	9	28	385	528
2016	9	29	328	456
2016	9	3	49	70
2016	9	30	332	429
2016	9	4	38	59
2016	9	5	138	174
2016	9	6	491	654
2016	9	7	402	532
2016	9	8	480	679
2016	9	9	470	645
2017	1	1	61	78
2017	1	10	924	1194
2017	1	11	848	1157
2017	1	12	905	1288
2017	1	13	772	1045
2017	1	14	146	210
2017	1	15	136	190
2017	1	16	603	765
2017	1	17	798	1052
2017	1	18	448	598

## 12. Thống kê số vụ tai nạn, tổng số thương vong, tổng số phương tiện liên quan theo LocationHierarchy (State-County-City)

Screenshot of the SSAS Dimension Designer interface showing the 'US Accidents...ube [Design]\*' tab selected. The left pane displays the dimension structure with nodes like 'Weekday', 'Year', 'Dim Driver', 'Dim Location', and 'Dim Vehicle'. The right pane shows a data grid with columns: State, County, City, Fact Accident Count, TongSoThuongVong, and TongSoPhuongTien. The data is as follows:

State	County	City	Fact Accident Count	TongSoThuongVong	TongSoPhuongTien
AL	Autauga	Autaugaville	1	1	1
AL	Autauga	Deatsville	8	12	14
AL	Autauga	Marbury	3	3	3
AL	Autauga	Prattville	23	30	36
AL	Baldwin	Bay Minette	8	11	14
AL	Baldwin	Daphne	77	116	122
AL	Baldwin	Elberta	1	1	2
AL	Baldwin	Fairhope	2	2	4
AL	Baldwin	Loxley	10	19	19
AL	Baldwin	Perdido	2	2	2
AL	Baldwin	Robertsdale	18	27	26
AL	Baldwin	Spanish Fort	2	3	3
AL	Barbour	Midway	1	2	2
AL	Bibb	Brent	1	1	1
AL	Bibb	Centreville	3	3	5
AL	Blount	Altoona	4	7	4
AL	Blount	Blountsville	9	11	14
AL	Blount	Cleveland	3	4	9
AL	Blount	Hayden	10	11	19
AL	Blount	Locust Fork	2	4	2
AL	Blount	Oneonta	10	16	15

## 4.5. Truy vấn các câu hỏi mà nhóm đưa ra bằng công cụ Pivot Tables

Danh sách câu hỏi:

- Thống kê số vụ tai nạn giao thông và thương vong theo năm từ 2016 đến 2023 là bao nhiêu?

	A	B	C	D	E	F	G	H	I	J
1			1. Thống kê số vụ tai nạn giao thông và thương vong theo năm từ 2016 đến 2023 là bao nhiêu?							
2										
3	Row Labels	Fact Accident Count	TongSoThuongVong							
4	2016	64320	86930							
5	2017	113732	150621							
6	2018	1	1							
7	2019	13338	18642							
8	2020	13	21							
9	2021	244753	336860							
10	2023	13213	18383							
11	Grand Total	449370	611458							
12										

- Thống kê số vụ tai nạn giao thông , thương vong, tổng số phương tiện liên quan theo từng quý qua các năm

	A	B	C	D	E	F	G	H	I
1			2. Thống kê số vụ tai nạn giao thông , thương vong, tổng số phương tiện liên quan theo từng quý qua các năm						
2									
3	Row Labels	Fact Accident Count	TongSoThuongVong	TongSoPhuongTien					
4	2016								
5	1	1053	1250	1490					
6	2	10948	14058	16656					
7	3	19457	26150	29348					
8	4	32862	45472	50442					
9	2017								
10	1	48826	65303	76385					
11	2	45848	60476	73224					
12	3	15806	20342	24839					
13	4	3252	4500	5040					
14	2018	1	1	1					
15	2019								
16	1	3323	4660	5107					
17	2	3270	4573	5141					
18	3	3350	4703	5184					
19	4	3395	4706	5285					
20	2020								
21	3	1	5	1					
22	4	12	16	17					
23	2021	244753	336860	382985					
24	2023	13213	18383	20602					
25	Grand Total	449370	611458	701747					
26									

- Thống kê số vụ tai nạn giao thông và thương vong theo ngày trong tuần (1-Chủ nhật ...; 7 - Thứ 7)

1		3. Thống kê số vụ tai nạn giao thông và thương vong theo ngày trong tuần (1-Chủ nhật ...; 7 - Thứ 7)
2		
3	Row Labels	Fact Accident Count TongSoThuongVong
4	1	36587 49956
5	2	67833 92084
6	3	72940 99007
7	4	73588 100149
8	5	75105 102300
9	6	80549 109130
10	7	42768 58832
11	Grand Total	449370 611458
12		
13		

4. Số vụ tai nạn giao thông theo giờ trong ngày (0-23h) thay đổi như thế nào?

1		4. Số vụ tai nạn giao thông theo giờ trong ngày (0-23h) thay đổi như thế nào?
2		
3	Row Labels	Fact Accident Count
4	0	58805
5	1	4728
6	10	16616
7	11	17593
8	12	21734
9	13	22672
10	14	27515
11	15	32402
12	16	33228
13	17	33917
14	18	24522
15	19	15358
16	2	4595
17	20	10415
18	21	8533
19	22	7769
20	23	6765
21	3	5079
22	4	6118
23	5	11133
24	6	17313
25	7	22457
26	8	22219
27	9	17884
28	Grand Total	449370
29		

5. Top 3 giờ xảy ra nhiều vụ tai nạn nhất trong ngày

A	B	C	D	E	F
1		5. Top 3 giờ xảy ra nhiều vụ tai nạn nhất trong ngày			
2					
3	Row Labels	Fact Accident Count			
4	0	58805			
5	16	33228			
6	17	33917			
7					
8					

6. Top 3 giờ xảy ra ít vụ tai nạn nhất trong ngày

	A	B	C	D	E
1			6.Top 3 giờ xảy ra ít vụ tai nạn nhất trong ngày		
2					
3	<b>Row Labels</b>	<b>Fact Accident Count</b>			
4	1	4728			
5	2	4595			
6	3	5079			
7					

7. Thống kê số vụ tai nạn ở nông thôn, thành thị.

1		7. Thống kê số vụ tai nạn ở nông thôn, thành thị.	
2			
3	<b>Row Labels</b>	<b>Fact Accident Count</b>	
4	Rural	71340	
5	Small town	61054	
6	Urban area	316976	
7	<b>Grand Total</b>	<b>449370</b>	
8			

8. Thống kê tổng quãng đường bị ảnh hưởng qua từng năm của 3 tiểu bang (LA, NY, WA)

1		
2		8. Thống kê tổng quãng đường bị ảnh hưởng qua từng năm của 3 tiểu bang (LA, NY, WA)
3		
4	Row Labels	TongQuangDuongAnhHuong
5	2016	
6	LA	1128.447998
7	NY	1373.172002
8	WA	1460.860996
9	2017	
10	LA	1039.732996
11	NY	5073.306996
12	WA	1406.774002
13	2019	
14	LA	157.5599999
15	NY	491.7600009
16	WA	155.5929988
17	2021	
18	LA	3176.380003
19	NY	8172.243002
20	WA	2108.783001
21	2023	
22	LA	129.3650004
23	NY	465.752
24	WA	141.4749997
25	Grand Total	26481.206
26		

9. Thống kê số người tử vong, số vụ tai nạn theo loại đường và tốc độ từ 30-70mi/h

2		9. Thống kê số người tử vong, số vụ tai nạn theo loại đường và tốc độ từ 30-70mi/h		
3				
4	Row Labels	Fact Accident Count	TongSoThuongVong	
5	▪ Dual carriageway	68204	101403	
6	▪ One way street			
7	30	9280	11135	
8	40	172	244	
9	50	73	99	
10	60	212	276	
11	70	115	169	
12	▪ Roundabout			
13	30	17723	22121	
14	40	4221	5305	
15	50	1112	1424	
16	60	3831	4917	
17	70	1984	2576	
18	▪ Single carriageway			
19	30	236811	302045	
20	40	21077	30952	
21	50	6078	9593	
22	60	67538	105010	
23	70	2	2	
24	▪ Slip road			
25	30	1400	1791	
26	40	300	402	
27	50	348	492	

10. Thống kê số vụ tai nạn, thương vong theo điều kiện thời tiết

Row Labels	Fact Accident Count	TongSoThuongVong
	10240	13903
Blowing Dust	3	6
Blowing Dust / Windy	13	20
Blowing Snow	17	21
Blowing Snow / Windy	4	7
Clear	61011	81066
Cloudy	37155	51065
Cloudy / Windy	686	930
Drizzle	246	357
Drizzle / Windy	3	3
Drizzle and Fog	12	15
Dust Whirls	1	1
Fair	139521	192452
Fair / Windy	1659	2240
Fog	4720	6527
Fog / Windy	34	43
Freezing Rain / Windy	1	3
Funnel Cloud	1	7
Haze	5659	7749
Haze / Windy	73	90
Heavy Drizzle	25	35
Heavy Ice Pellets	2	2
Heavy Rain	1919	2605
Heavy Rain / Windy	99	140

11. Thống kê số vụ tai nạn và tổng số thương vong theo DateHierarchi (Y-M-D)

11. Thống kê số vụ tai nạn và thương vong theo [DateHierarchy](Y-M-D)

Row Labels	Fact Accident Count	TongSoThuongVong
2016		
+ 1	7	13
+ 10	7828	10795
+ 11	7908	10749
+ 12	17126	23928
+ 2	402	447
+ 3	644	790
+ 4	2154	2602
+ 5	3431	4299
+ 6	5363	7157
+ 7	5624	7978
+ 8	7062	9234
+ 9	6771	8938
2017	113732	150621
2018		
+ 9		
28	1	1
2019	13338	18642
2020	13	21
2021		
+ 1		
1	36	46
10	26	31
11	40	60
12	42	52

12. Thống kê số vụ tai nạn, tổng số thương vong, tổng số phương tiện liên quan theo LocationHierarchy (State-County-City)

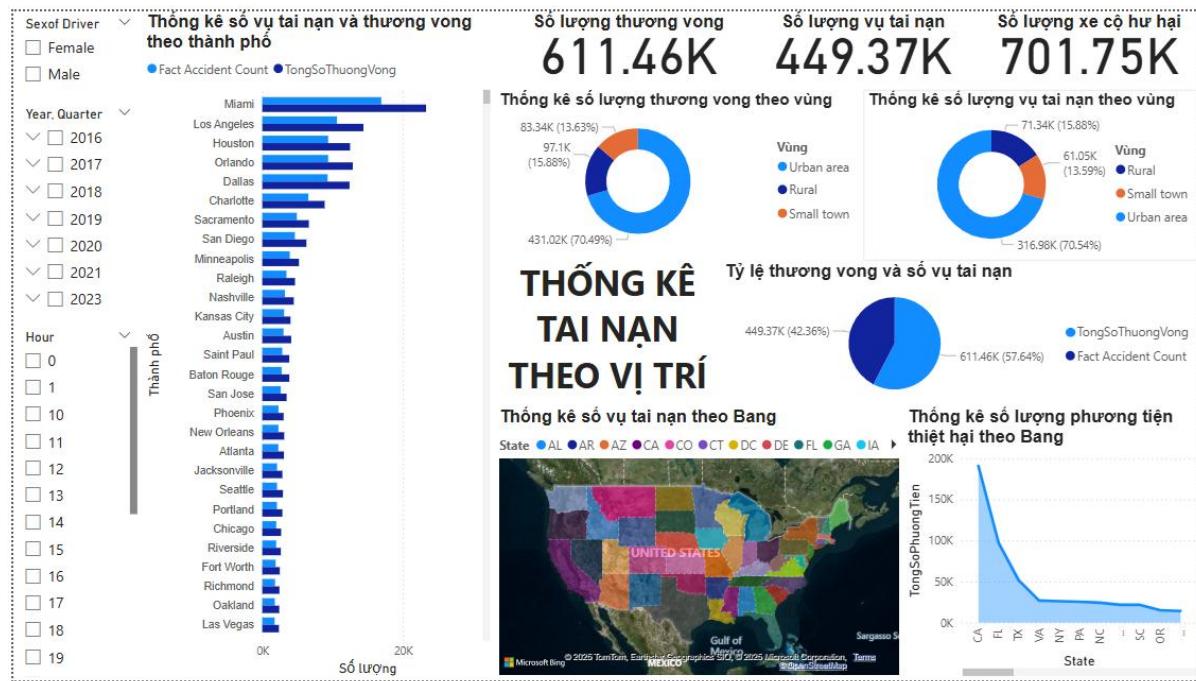
1		12. Thống kê số vụ tai nạn và thương vong theo [LocationHierarchy](State-County - City)	
2		Row Labels	Fact Accident Count TongSoThuongVong
3		⊕ AL	2521 3506
4		⊕ AR	1461 2022
5		⊕ AZ	
6		⊕ Apache	92 124
7		⊕ Cochise	49 59
8		⊕ Coconino	255 333
9		⊕ Gila	80 109
10		⊕ Graham	12 13
11		⊕ Greenlee	5 5
12		⊕ La Paz	84 105
13		⊕ Maricopa	4620 6173
14		⊕ Mohave	192 278
15		⊕ Navajo	104 129
16		⊕ Pima	1719 2433
17		⊕ Pinal	234 322
18		⊕ Santa Cruz	10 12
19		⊕ Yavapai	353 475
20		⊕ Yuma	24 32
21		⊕ CA	122951 167758
22		⊕ CO	3427 4589
23		⊕ CT	5176 7035
24		⊕ DC	1261 1728
25		⊕ DE	776 1066
26		⊕ FL	62330 85400
27		⊕ GA	6024 8200

1		12. Thống kê số vụ tai nạn và thương vong theo [LocationHierarchy](State-County - City)	
2		Row Labels	Fact Accident Count TongSoThuongVong
3		⊕ AL	2521 3506
4		⊕ AR	1461 2022
5		⊕ AZ	
6		⊕ Apache	92 124
7		⊕ Cochise	
8		Benson	3 3
9		Bisbee	2 2
10		Bowie	2 2
11		Cochise	3 6
12		Douglas	2 2
13		Dragoon	4 4
14		Huachuca City	6 6
15		San Simon	17 19
16		Sierra Vista	1 1
17		Tombstone	2 4
18		Willcox	7 10
19		⊕ Coconino	255 333
20		⊕ Gila	80 109
21		⊕ Graham	12 13
22		⊕ Greenlee	5 5
23		⊕ La Paz	84 105
24		⊕ Maricopa	4620 6173
25		⊕ Mohave	192 278
26		⊕ Navajo	104 129
27		⊕ Pima	1719 2433

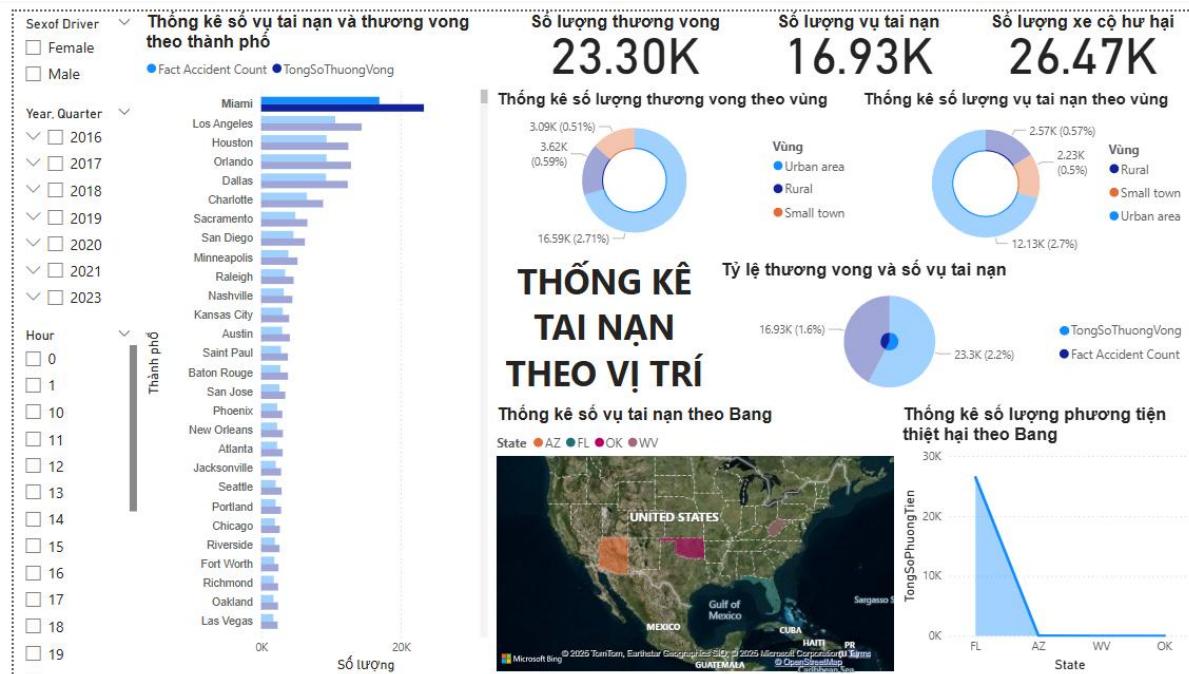
## 5. Thiết kế Dashboard Power BI

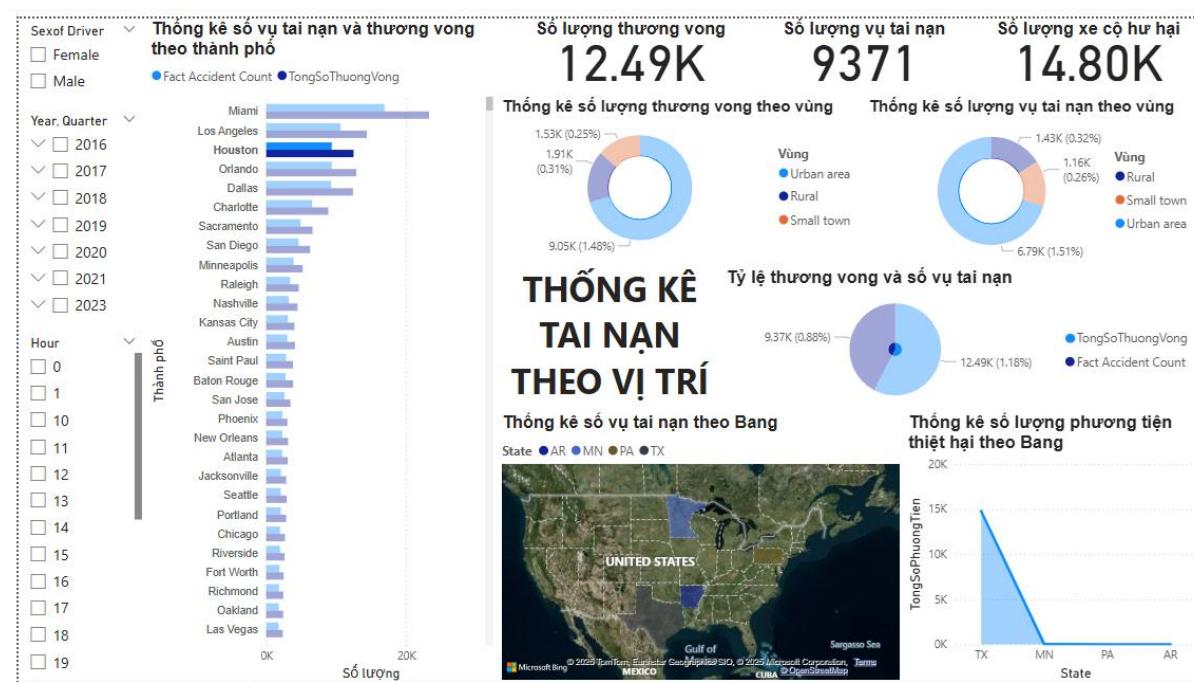
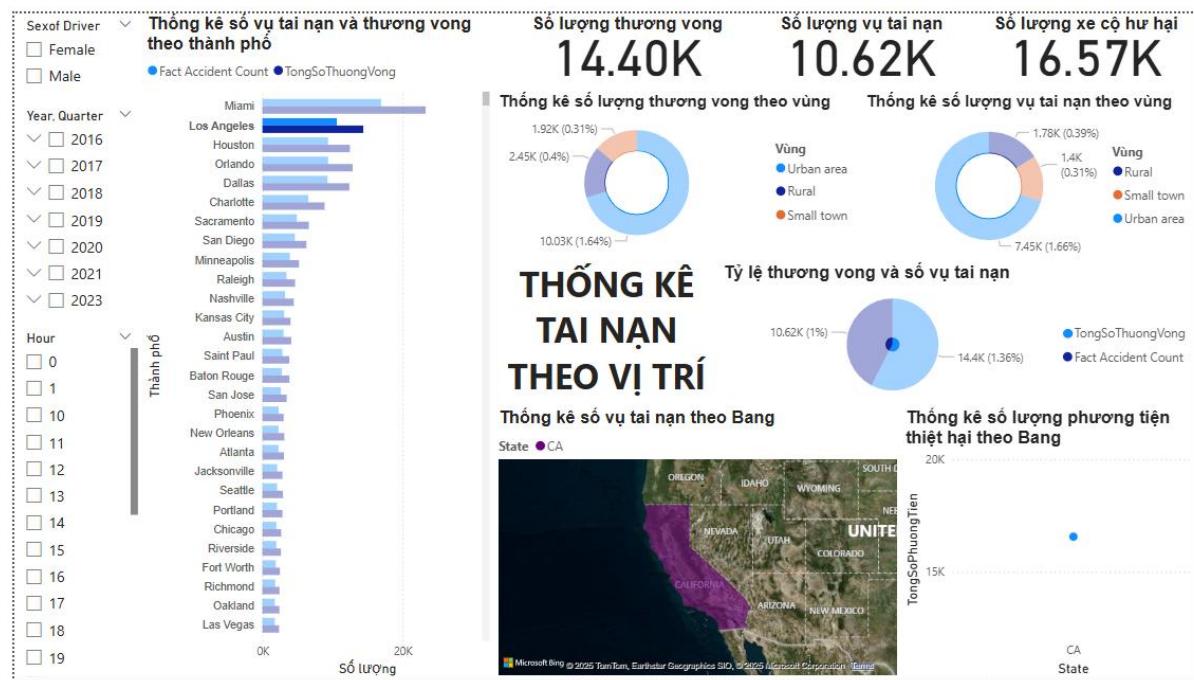
### 5.1. Khởi đầu hiển thị bức tranh toàn cảnh về tai nạn giao thông theo vị trí

Hãy bắt đầu hành trình của chúng ta với dashboard "Thống kê tai nạn theo vị trí". Tại đây, chúng ta có cái nhìn tổng quan về số vụ tai nạn, thương vong và thiệt hại phương tiện ở Mỹ.



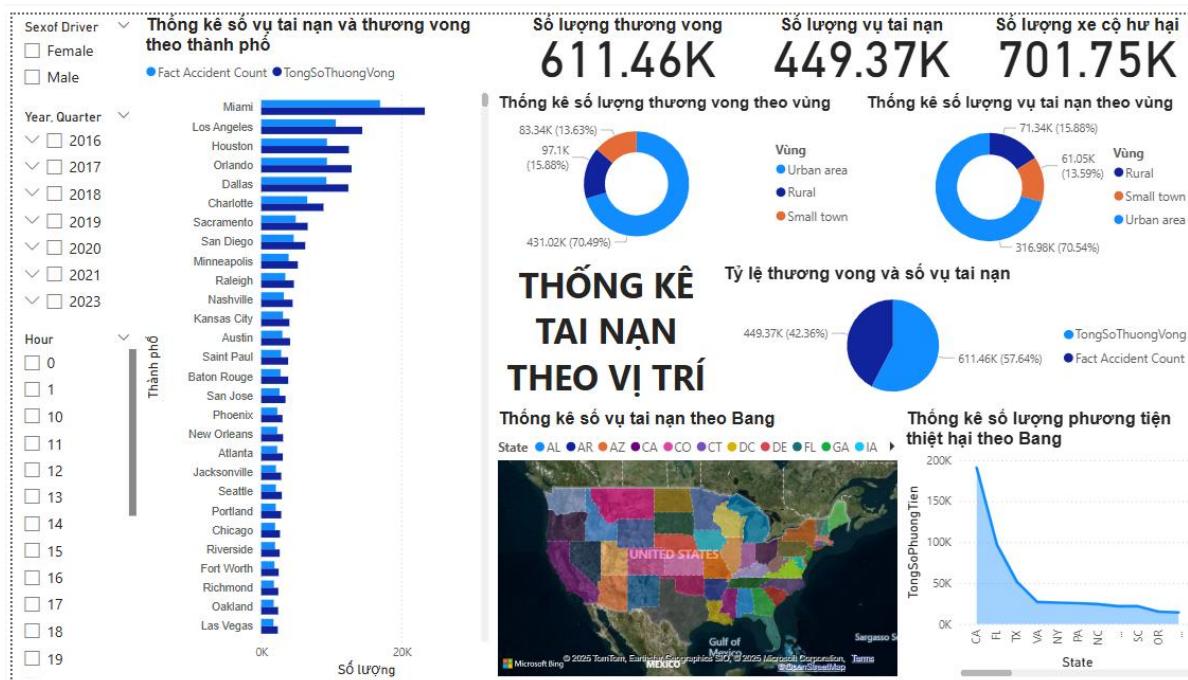
Thành phố nào có nhiều tai nạn nhất? Biểu đồ "Thống kê số vụ tai nạn và thương vong theo thành phố" cho thấy Miami đứng đầu với hơn 16.9K vụ tai nạn, tiếp theo là Los Angeles và Houston. Điều này không quá bất ngờ vì đây đều là những thành phố lớn, đông dân và có mật độ giao thông cao.



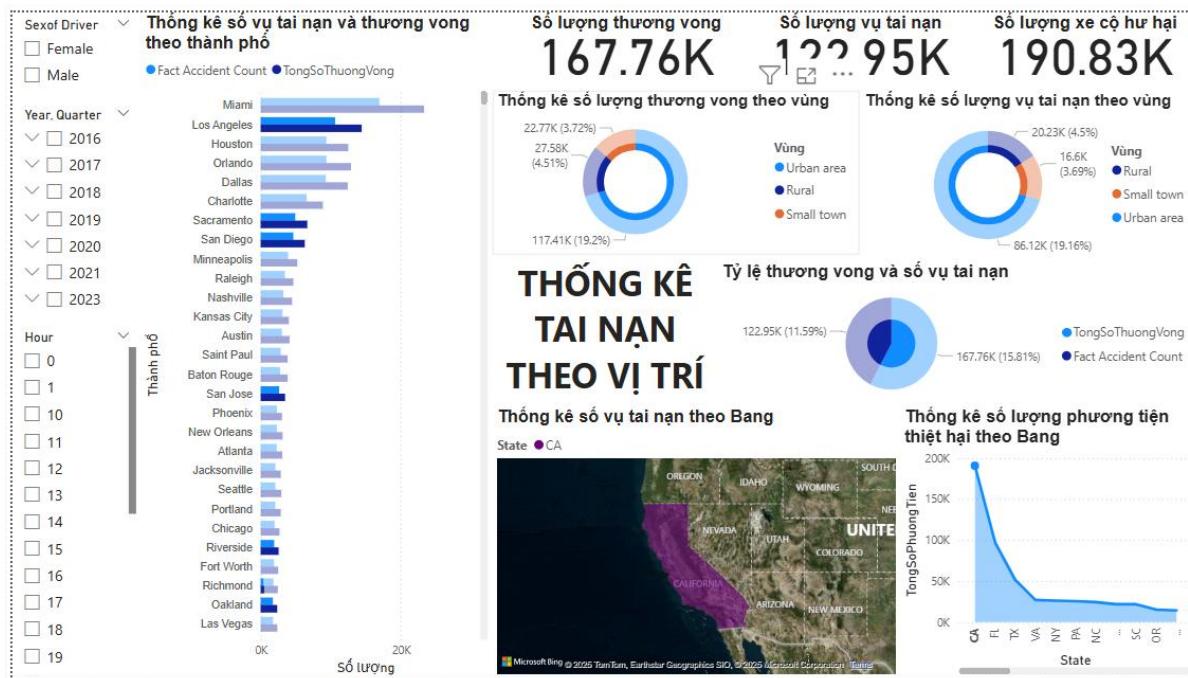


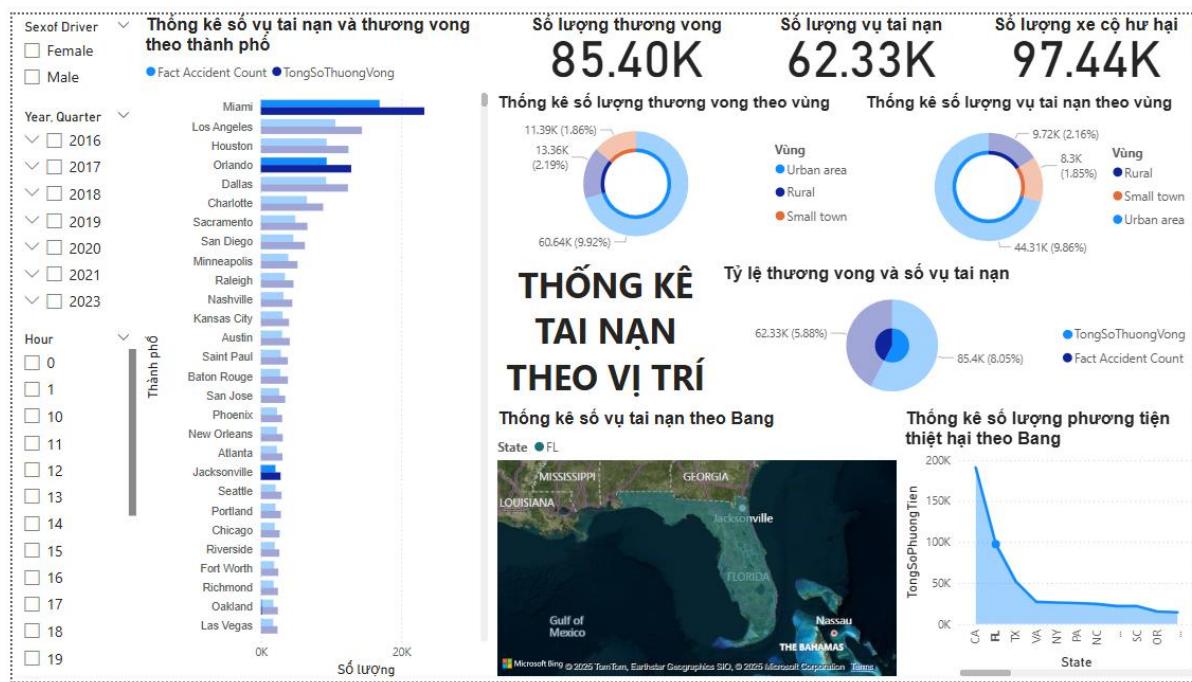
Tuy nhiên, điều đáng chú ý là số lượng thương vong tại các thành phố này cũng rất lớn, với Miami ghi nhận số thương vong cao nhất.

Tỷ lệ thương vong và số vụ tai nạn cho thấy cứ 100 vụ tai nạn thì có khoảng 61.05% dẫn đến thương vong, một con số đáng báo động. Điều này đặt ra câu hỏi: Liệu có yếu tố nào liên quan đến vị trí ảnh hưởng đến mức độ nghiêm trọng của tai nạn?



Thiệt hại phương tiện theo Bang và Số vụ tai nạn theo Bang cho thấy California (CA) và Florida (FL) là hai bang có số lượng phương tiện thiệt hại và tai nạn cao nhất. Điều này có thể liên quan đến mật độ dân số và lượng phương tiện giao thông tại đây.





Insight đầu tiên: Tai nạn giao thông tập trung nhiều nhất tại các thành phố lớn như Miami (hơn 16.9K vụ), Los Angeles, và Houston, nơi có mật độ dân số và giao thông cao. Tuy nhiên, tỷ lệ thương vong (61.05% trong 100 vụ tai nạn) đặc biệt đáng lo ngại, đặc biệt tại Miami, cho thấy các thành phố lớn không chỉ có số vụ cao mà còn tiềm ẩn rủi ro nghiêm trọng. Các bang như California (CA) và Florida (FL) dẫn đầu về số vụ tai nạn và thiệt hại phương tiện, có thể do lượng phương tiện và mật độ giao thông lớn. Điều này đặt ra nhu cầu cải thiện quản lý giao thông đô thị và cơ sở hạ tầng ở các khu vực trọng điểm.

## 5.2. Tai nạn giao thông khi có các điều kiện tác động

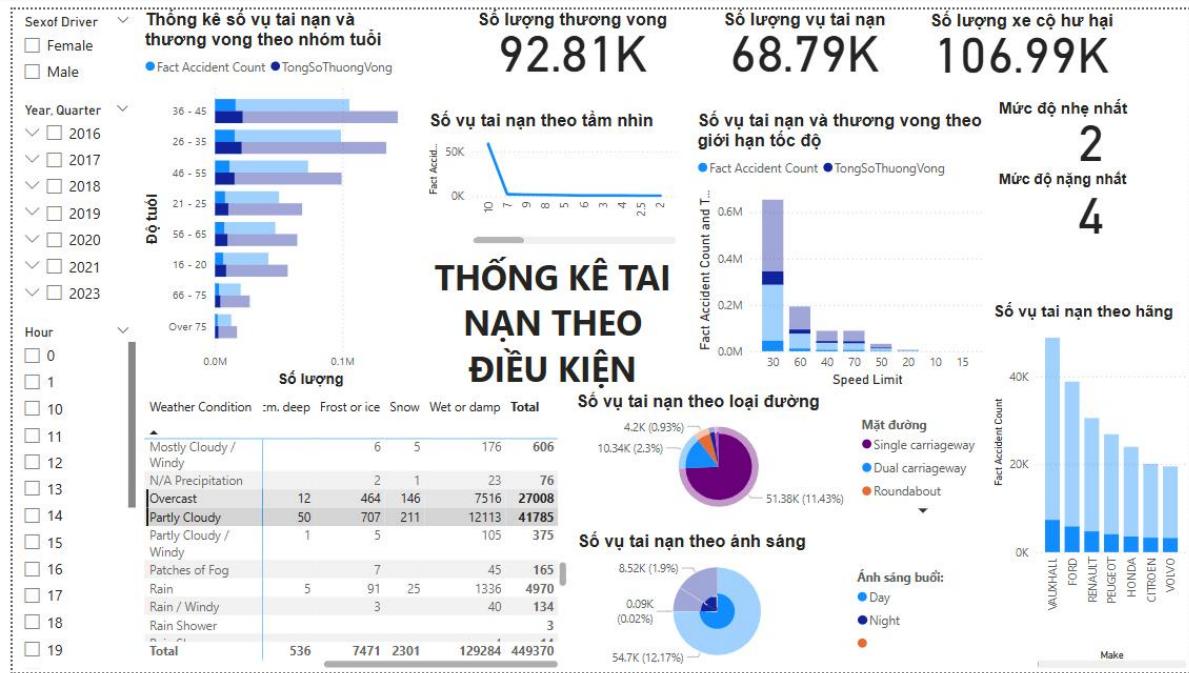
Chuyển sang dashboard "Thống kê tai nạn theo điều kiện", chúng ta sẽ khám phá các yếu tố liên quan đến môi trường và con người ảnh hưởng đến tai nạn.

Nhóm tuổi nào dễ gặp tai nạn nhất? Biểu đồ "Thống kê số vụ tai nạn và thương vong theo nhóm tuổi" cho thấy nhóm tuổi 36-45 có số vụ tai nạn cao nhất, tiếp theo là nhóm 26-35.



Điều này có thể phản ánh rằng những người trong độ tuổi lao động, thường xuyên di chuyển, có nguy cơ gặp tai nạn cao hơn. Nhóm trên 75 tuổi có số vụ tai nạn thấp nhất, nhưng tỷ lệ thương vong của họ lại cao, có thể do sức khỏe yếu hơn.

Biểu đồ "Thống kê số vụ tai nạn trong điều kiện thời tiết và mặt đường" cho thấy điều kiện "Partly Cloudy" (có mây 1 phần) và "Overcast" (u ám) là những yếu tố thời tiết ghi nhận nhiều ca tai nạn nhất.



Tuy nhiên, phần lớn tai nạn (68.83%) xảy ra trong điều kiện mặt đường khô (Dry), cho thấy thời tiết không phải lúc nào cũng là nguyên nhân chính.



"Số vụ tai nạn theo ánh sáng" chỉ ra rằng 74.93% tai nạn xảy ra vào ban ngày (Day), nhưng ban đêm (Night) lại có tỷ lệ thương vong cao hơn (15.89%).



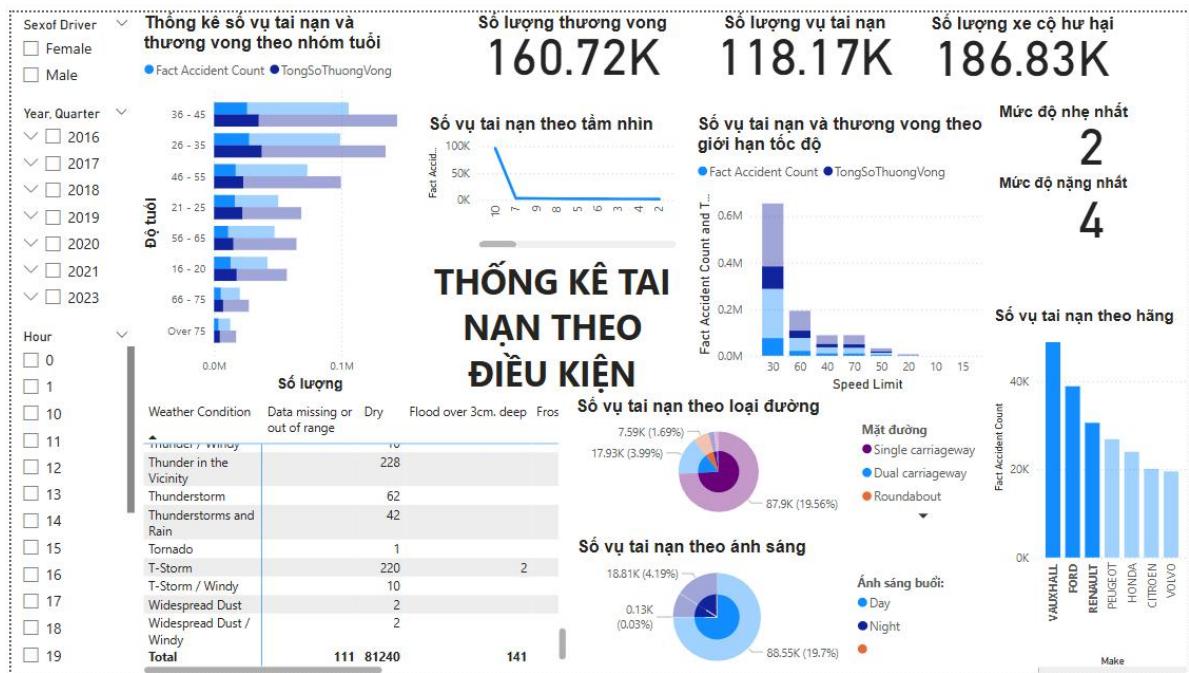
Điều này có thể liên quan đến tầm nhìn giảm và tốc độ phản ứng chậm hơn trong điều kiện ánh sáng yếu.

Biểu đồ "Số vụ tai nạn và thương vong theo giới hạn tốc độ" cho thấy giới hạn tốc độ 30 mph có số vụ tai nạn cao nhất (287K vụ).



Theo logic thông thường, giới hạn tốc độ cao sẽ có tỷ lệ tử vong cao nhưng giới hạn tốc độ cao hơn (70 mph) lại có tỷ lệ thương vong thấp hơn, phản ánh rằng giới hạn tốc độ cao có thể là các vụ tai nạn trên cao tốc, có lưu lượng xe thông thoáng nên người lái có thể xử lý hiệu quả hơn khi sự cố xảy ra.

Về hãng xe, chúng tôi muốn biết hãng xe nào dễ gặp tai nạn để một phần suy luận ra được có sự ảnh hưởng của phương tiện.

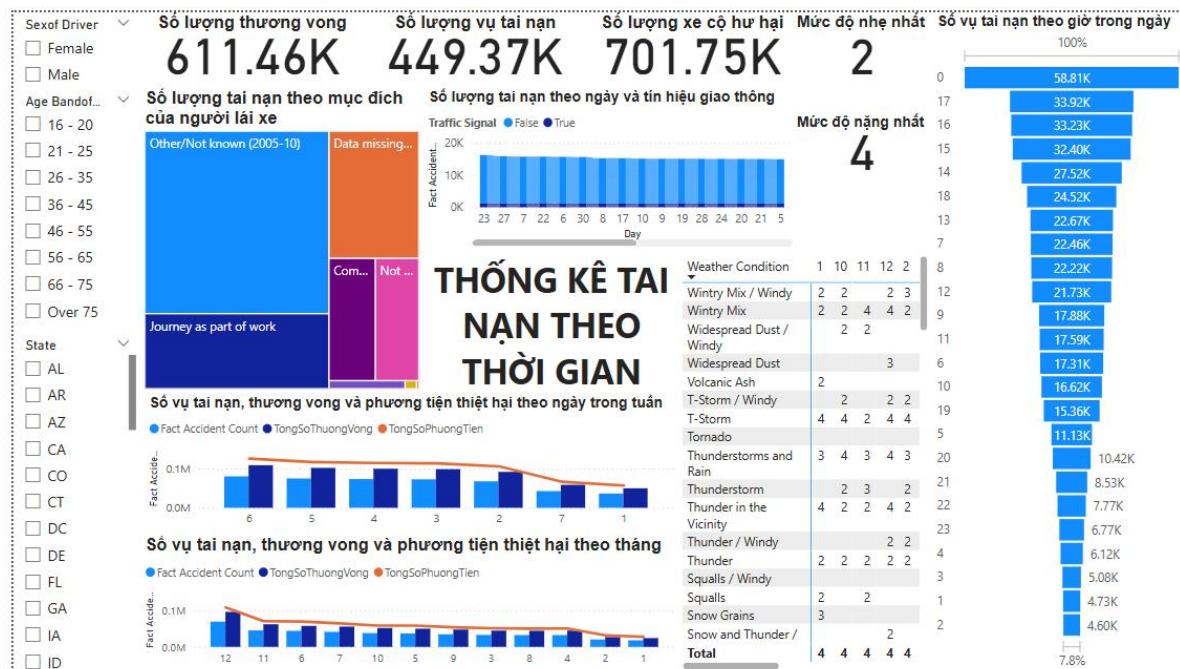


Bảng biểu đồ thống kê "Số vụ tai nạn theo hãng" chỉ ra Vauxhall, Ford, và Renault là ba hãng xe có số vụ tai nạn cao nhất, lần lượt là 66.3K, 52.85K và 41.57K vụ. Đối với Ford nguyên nhân khiến con số này cao đột biến có thể là do đây là một dòng xe phổ biến, nhưng Vauxhall và Renault là các dòng xe không được quan tâm chuộng nhưng lại có số lượng vụ tai nạn cao, điều này có thể liên quan đến chất lượng xe có vấn đề và thực sự ảnh hưởng đến tỷ lệ tai nạn.

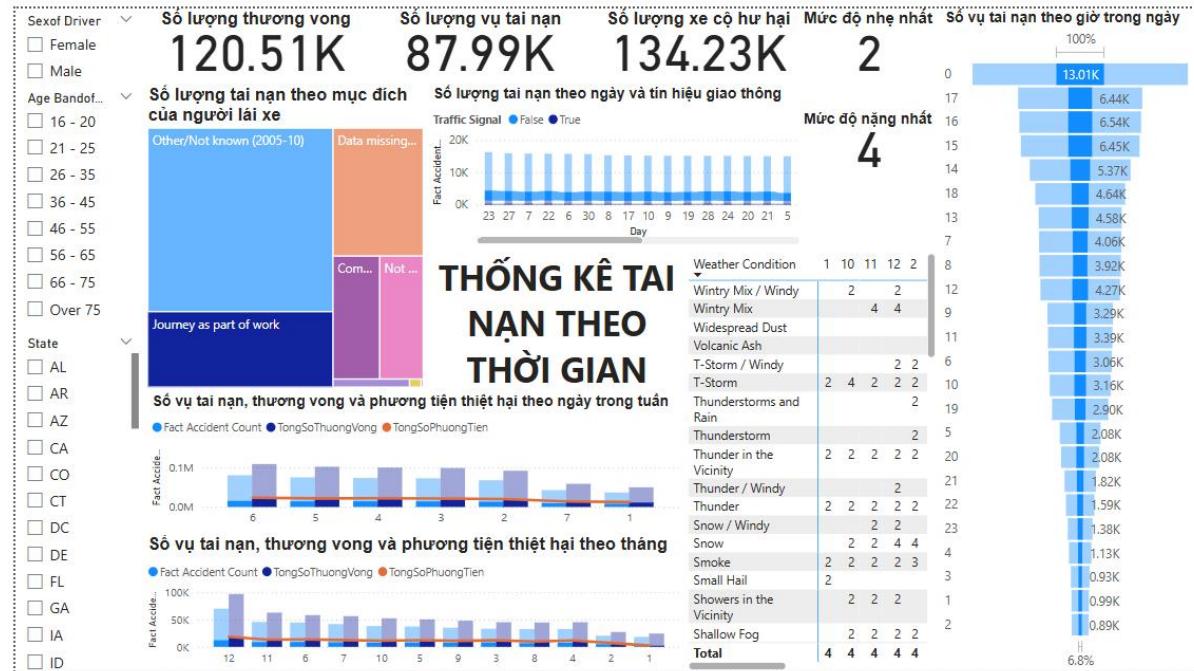
Insight thứ hai: Nhóm tuổi lao động (36-45 và 26-35) có nguy cơ tai nạn cao nhất do tần suất di chuyển nhiều, trong khi nhóm trên 75 tuổi có tỷ lệ thương vong cao hơn do sức khỏe yếu. Điều kiện thời tiết như "Partly Cloudy" và "Overcast" gia tăng số vụ tai nạn, nhưng mặt đường khô (68.83%) là nguyên nhân chính, cho thấy yếu tố con người có thể đóng vai trò lớn hơn thời tiết. Ban đêm (15.89% thương vong) và giới hạn tốc độ 30 mph (287K vụ) là những thời điểm và điều kiện nguy hiểm, dù tốc độ cao (70 mph) có thương vong thấp hơn do lưu lượng xe thông thoáng trên cao tốc. Về hãng xe, Vauxhall (66.3K vụ), Ford (52.85K vụ), và Renault (41.57K vụ) dẫn đầu, với Vauxhall và Renault có thể liên quan đến chất lượng xe, trong khi Ford phổ biến hơn. Đề xuất: Tăng cường giáo dục an toàn giao thông cho nhóm lao động, cải thiện tầm nhìn ban đêm, kiểm tra chất lượng xe của Vauxhall và Renault, và quản lý tốc độ hợp lý trên các tuyến đường khác nhau.

### 5.3. Thời gian kể câu chuyện gì? Dashboard "Thống kê tai nạn theo thời gian"

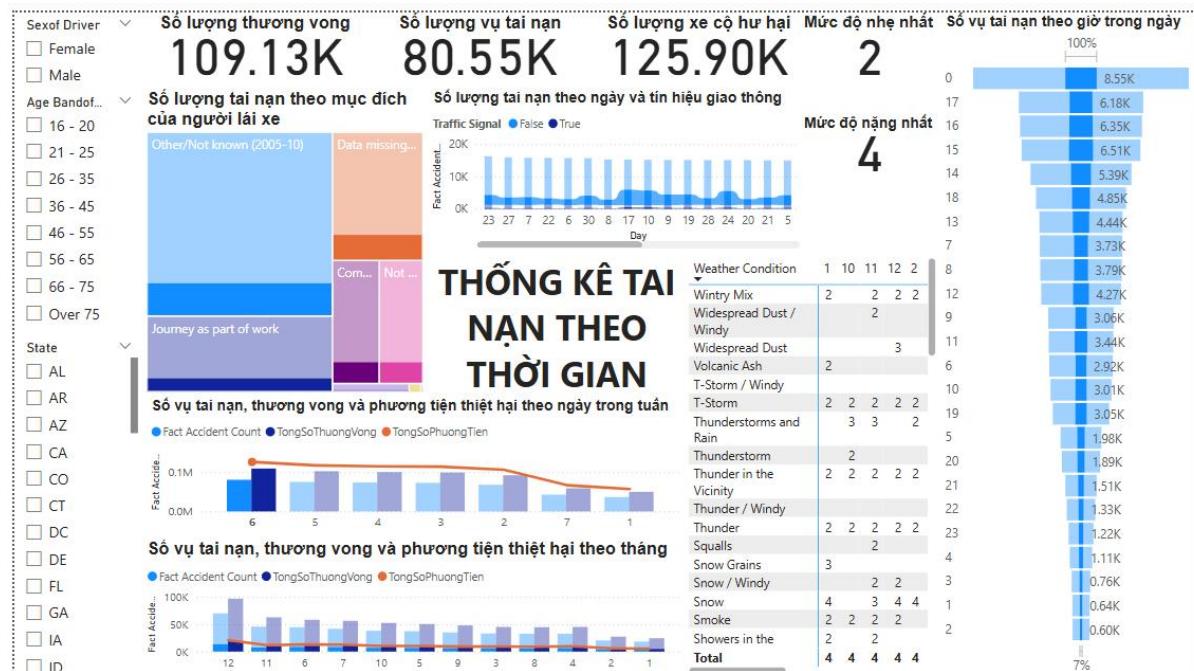
Cuối cùng, chúng ta đến với dashboard "Thống kê tai nạn theo thời gian", nơi thời gian trở thành yếu tố chính để phân tích.



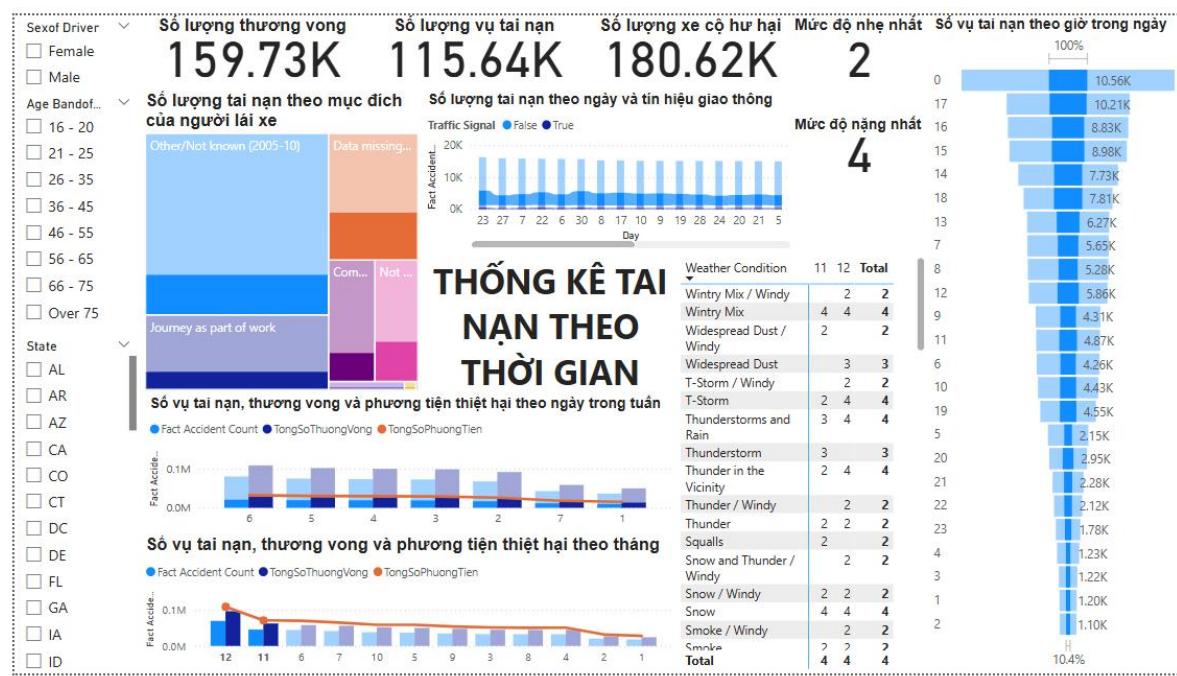
Biểu đồ "Số lượng tai nạn theo mục đích của người lái xe" cho thấy phần lớn tai nạn xảy ra trong các chuyến đi liên quan đến công việc (Journey as part of work), chiếm phần lớn trong tổng số 611.46K vụ tai nạn. Điều này cho thấy giờ cao điểm có thể là thời điểm nguy hiểm nhất.



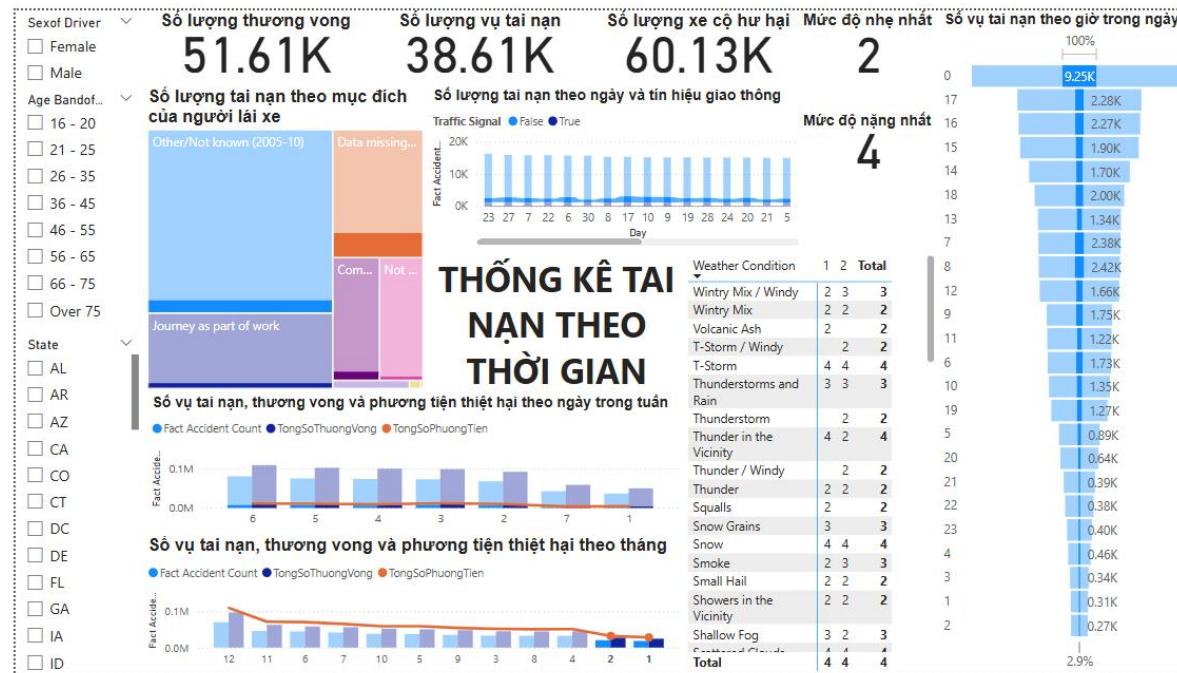
Ngày nào trong tuần nguy hiểm nhất? "Số vụ tai nạn, thương vong và phương tiện thiệt hại theo ngày trong tuần" cho thấy ngày thứ 6 (Friday) có số vụ tai nạn cao nhất (hơn 20K vụ), có thể do lượng giao thông tăng vào cuối tuần. Chủ nhật (Sunday) lại có số vụ thấp nhất.



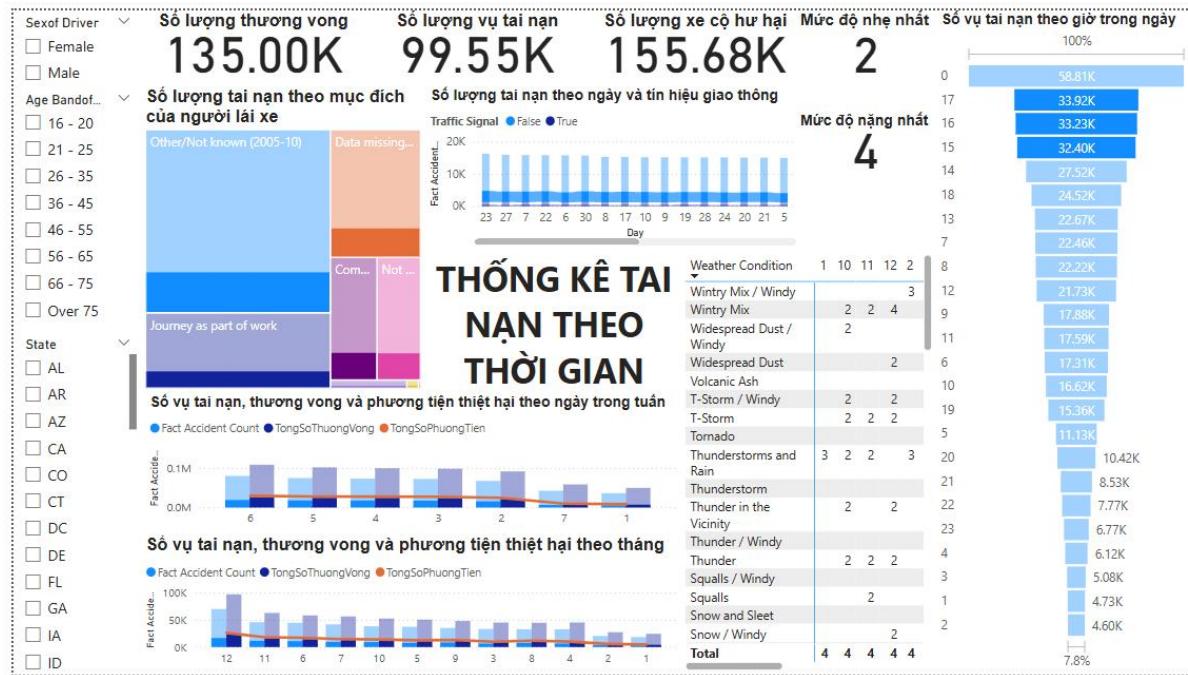
Tháng nào đáng lo ngại? "Số vụ tai nạn, thương vong và phương tiện thiệt hại theo tháng" cho thấy tháng 12 có số vụ tai nạn cao nhất, có thể liên quan đến thời tiết mùa thu kéo theo các hoạt động diễn ra sôi nổi ở những ngày gần cuối năm khiến người dân phải thường xuyên di chuyển trên đường và sự ảnh hưởng của thời tiết ở các tháng cuối năm như 11, 12 dẫn đến việc tăng tỷ lệ tai nạn.



Tháng 1, 2 có số vụ thấp nhất, có lẽ do đây là những tháng Mỹ bị bao phủ bởi tuyết, người dân hạn chế ra ngoài, do đó tỷ lệ tai nạn xảy ra cũng ít đi đáng kể (chỉ bằng  $\frac{1}{3}$  so với những tháng cuối năm).



Giờ nào nguy hiểm nhất? "Số vụ tai nạn theo giờ trong ngày" chỉ ra khoảng 15h-18h (3 PM - 6 PM) là khung giờ có nhiều tai nạn nhất, trùng với giờ tan làm. Đây là thời điểm giao thông đông đúc, dễ dẫn đến va chạm.



Matrix điều kiện thời tiết và tháng đối với sự tác động của mức độ vụ tai nạn : Matrix cho thấy các hiện tượng thời tiết càng cực đoan, mức độ nghiêm trọng của vụ tai nạn càng lớn, với các tháng giữa năm là khoảng thời gian thường xuyên xảy ra yếu tố thời tiết nguy hiểm nhất, trùng khớp với dữ liệu từ dashboard điều kiện.

Insight thứ ba: Tai nạn chủ yếu xảy ra trong các chuyến đi công việc (chiếm phần lớn trong 611.46K vụ), với giờ cao điểm (15h-18h) và ngày thứ 6 (hơn 20K vụ) là thời điểm nguy hiểm nhất do giao thông đông đúc. Tháng 12 có số vụ cao nhất do hoạt động cuối năm và thời tiết xấu, trong khi tháng 1, 2 thấp nhất nhờ thời tiết lạnh hạn chế di chuyển. Matrix điều kiện thời tiết cho thấy các tháng giữa năm, với thời tiết cực đoan, làm tăng mức độ nghiêm trọng của tai nạn. Đề xuất: Tăng cường kiểm soát giao thông vào giờ cao điểm và ngày thứ 6, nâng cao nhận thức an toàn trong tháng 12, và cải thiện cơ sở hạ tầng để ứng phó với thời tiết cực đoan giữa năm.

#### 5.4. Kết luận: Hành động từ dữ liệu

Từ hành trình khám phá này, chúng ta rút ra được những điểm chính:

Các thành phố lớn như Miami, Los Angeles, Houston và bang California, Florida cần cải thiện hệ thống giao thông đô thị và cơ sở hạ tầng để giảm thiểu tai nạn và thương vong, đặc biệt ở những khu vực có mật độ cao.

Tập trung nâng cao nhận thức an toàn cho nhóm tuổi lao động (26-45), cải thiện điều kiện ánh sáng ban đêm, kiểm tra chất lượng xe của Vauxhall và Renault, và điều chỉnh quản lý tốc độ phù hợp với từng loại đường.

Tăng cường giám sát giao thông vào giờ cao điểm (15h-18h), ngày thứ 6, và tháng 12; đồng thời đầu tư vào cơ sở hạ tầng để ứng phó với thời tiết cực đoan giữa năm.

Kết hợp các biện pháp giáo dục, kiểm tra phương tiện, và cải thiện hạ tầng giao thông dựa trên phân tích dữ liệu để giảm thiểu tai nạn, bảo vệ tính mạng và tài sản, đặc biệt trong các điều kiện và thời điểm nhạy cảm đã xác định.

### III. Kết luận

Báo cáo này được thực hiện với mục tiêu xây dựng một kho dữ liệu hiệu quả nhằm quản lý và phân tích dữ liệu các ca tai nạn giao thông tại Hoa Kỳ trong giai đoạn từ năm 2016 đến năm 2023. Trong bối cảnh an toàn giao thông đường bộ là một ưu tiên quan trọng của xã hội hiện đại, kho dữ liệu được thiết kế không chỉ để thu thập và lưu trữ thông tin mà còn hỗ trợ các cơ quan quản lý nhà nước, nhà nghiên cứu, và các tổ chức liên quan trong việc đưa ra các chính sách phù hợp. Hệ thống này cho phép phân tích xu hướng tai nạn, xác định các khu vực nguy cơ cao, và đề xuất biện pháp phòng ngừa hiệu quả nhằm giảm thiểu thiệt hại về người và tài sản. Với mục tiêu trở thành một hệ thống phân tán, kho dữ liệu đảm bảo xử lý nhanh chóng, chính xác các truy vấn phức tạp, đáp ứng nhu cầu phân tích dữ liệu lớn trong thời đại Big Data. Tính bảo mật, độ chính xác, và khả năng mở rộng được ưu tiên, hỗ trợ tích hợp với công nghệ mới, từ đó phục vụ tốt nhất các mục tiêu chiến lược của chính phủ Mỹ và các tổ chức liên quan đến an toàn giao thông.

Dự án đã thành công trong việc xây dựng kho dữ liệu dựa trên mô hình Kimball, thông qua việc thu thập dữ liệu từ tập US-Accidents trên Kaggle, tiền xử lý để đảm bảo chất lượng, và tích hợp thành một data warehouse hoàn chỉnh. Quá trình triển khai sử dụng SQL Server 2022 cho lưu trữ, SSIS cho quy trình ETL, SSAS cho phân tích đa chiều, và Power BI cho trực quan hóa dữ liệu qua các dashboard, tạo điều kiện thuận lợi cho việc xử lý truy vấn phức tạp và cung cấp các insight giá trị. Kho dữ liệu không chỉ đáp ứng nhu cầu phân tích hiện tại mà còn có tiềm năng mở rộng, tích hợp với các công nghệ tiên tiến, hỗ trợ các nhà hoạch định chính sách trong việc giảm thiểu tai nạn giao thông. Dù tồn tại một số hạn chế như xử lý giá trị thiếu bằng phương pháp ngẫu nhiên và thiếu dữ liệu ở một số ngày, dự án đã đặt nền tảng vững chắc cho các nghiên cứu sâu hơn trong tương lai. Với tính ứng dụng cao và ý nghĩa xã hội to lớn, kết quả này khẳng định vai trò quan trọng của công nghệ kho dữ liệu trong việc

nâng cao an toàn giao thông, bảo vệ tính mạng và tài sản, góp phần giải quyết các vấn đề thực tiễn tại Mỹ.

## TÀI LIỆU THAM KHẢO

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

Dataset: [https://www.kaggle.com/datasets/tsiaras/uk-road-safety-accidents-and-vehicles?select=Vehicle\\_Information.csv](https://www.kaggle.com/datasets/tsiaras/uk-road-safety-accidents-and-vehicles?select=Vehicle_Information.csv)