**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**



**BÁO CÁO ĐỒ ÁN MÔN HỌC**

**KHO DỮ LIỆU VÀ OLAP**

**ĐỀ TÀI:**

**XÂY DỰNG KHO DỮ LIỆU PHÂN TÍCH**

**TAI NẠN GIAO THÔNG NƯỚC MỸ**

**GIAI ĐOẠN 2018 - 2023**

**Lớp**: IS207.Q13

**GVHD**: ThS. Đỗ Thị Minh Phụng

**Thành viên nhóm**: Ngô Tiến Sỹ - 23521367

Nguyễn Văn Nam – 23520982

**TP. HỒ CHÍ MINH, NĂM 2025**

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

*……., ngày……...tháng……năm 2025*

**Người nhận xét**

*(Ký tên và ghi rõ họ tên****)***

**LỜI CẢM ƠN**

Kính gửi Cô Đỗ Thị Minh Phụng - giảng viên môn Kho Dữ liệu và OLAP,

Lời đầu tiên, nhóm thực hiện đề tài xin bày tỏ lòng biết ơn sâu sắc đến Cô. Chúng em vô cùng trân trọng những kiến thức quý báu mà Cô đã tận tâm truyền đạt trong suốt thời gian qua. Sự nhiệt huyết và phương pháp giảng dạy lôi cuốn của Cô đã khơi gợi niềm đam mê học hỏi trong chúng em.

Chúng em xin chân thành cảm ơn Cô vì những lời khuyên, định hướng và góp ý quý báu. Nhờ có sự chỉ dẫn tận tình của Cô, chúng em đã có thể xác định được hướng đi đúng đắn và hoàn thành đề tài này.

Trong quá trình thực hiện đồ án, chúng em đã cố gắng vận dụng tối đa những kiến thức được Cô trang bị trên lớp, đồng thời tích cực tìm tòi, nghiên cứu thêm từ nhiều nguồn tài liệu khác nhau. Mặc dù đã nỗ lực hết mình, do còn hạn chế về kinh nghiệm thực tế, đồ án của chúng em chắc chắn không tránh khỏi những thiếu sót.

Vì vậy, chúng em kính mong nhận được những nhận xét, đánh giá và góp ý chân thành từ Cô. Những ý kiến đóng góp của Cô sẽ là nguồn động lực to lớn giúp chúng em củng cố kiến thức, trau dồi kinh nghiệm và hoàn thiện hơn trong các dự án tương lai.

Xin chân thành cảm ơn Cô!

# LÝ DO CHỌN ĐỀ TÀI

Trong bối cảnh các vấn đề an toàn giao thông ngày càng được quan tâm, việc nghiên cứu và phân tích tình hình tai nạn giao thông tại Hoa Kỳ trong giai đoạn 2018 –2023 trở nên đặc biệt cần thiết. Đề tài này không chỉ cung cấp cái nhìn toàn diện về thực trạng tai nạn giao thông mà còn mang lại nguồn dữ liệu có giá trị cao cho các bài toán phân tích và hỗ trợ ra quyết định trong lĩnh vực quản lý giao thông.

Dataset được lựa chọn sở hữu cấu trúc đa chiều rõ ràng và phong phú, bao gồm các thông tin về nguồn thu thập dữ liệu, thời gian, địa điểm, điều kiện thời tiết cùng nhiều yếu tố môi trường khác. Đặc điểm này tạo điều kiện lý tưởng để xây dựng các mô hình kho dữ liệu theo Star Schema hoặc Snowflake Schema, từ đó triển khai hiệu quả các phép phân tích OLAP. Thông qua việc thực hiện các kỹ thuật phân tích như drill-down, roll-up và slice/dice, nhóm có thể khám phá sâu sắc các yếu tố tác động đến tai nạn giao thông, nhận diện các xu hướng tiềm ẩn và đưa ra những giải pháp khoa học để cải thiện an toàn giao thông.

Bên cạnh đó, khả năng tích hợp dataset với các nguồn dữ liệu bổ sung như thông tin dân số, mật độ giao thông và dữ liệu khí tượng sẽ mở rộng đáng kể phạm vi nghiên cứu và nâng cao chất lượng phân tích. Sự kết hợp đa chiều này không chỉ giúp làm sáng tỏ các mối quan hệ phức tạp giữa các yếu tố ảnh hưởng mà còn tạo ra nền tảng vững chắc cho việc đề xuất các chính sách và biện pháp cải thiện an toàn giao thông có tính khả thi cao.

Như vậy, việc lựa chọn đề tài này vừa thể hiện tính cấp thiết của vấn đề thực tiễn, vừa mang ý nghĩa quan trọng trong việc ứng dụng và phát triển các phương pháp phân tích dữ liệu hiện đại trong lĩnh vực giao thông, đồng thời phù hợp với mục tiêu học tập và nghiên cứu của môn Kho Dữ liệu và OLAP.

# MỤC LỤC

[LÝ DO CHỌN ĐỀ TÀI 4](#_Toc209519098)

[MỤC LỤC 5](#_Toc209519099)

[CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ DỮ LIỆU 6](#_Toc209519100)

[1.1 PHÁT BIỂU VỀ DỮ LIỆU 6](#_Toc209519101)

[1.1.1 Mô tả dữ liệu 6](#_Toc209519102)

[1.1.2 Tiền xử lý 7](#_Toc209519103)

[1.1.3 Thuộc tính của dữ liệu 7](#_Toc209519104)

[1.1.4 Hướng chủ đề 7](#_Toc209519105)

[1.2 XÂY DỰNG KHO DỮ LIỆU 7](#_Toc209519106)

[1.2.1 Thiết kế lược đồ 7](#_Toc209519107)

[1.2.2 Các bảng chiều 7](#_Toc209519108)

[1.3 CÁC CÂU TRUY VẤN 7](#_Toc209519109)

[CHƯƠNG 2: QUÁ TRÌNH XÂY DỰNG KHO DỮ LIỆU (SSIS) 8](#_Toc209519110)

# CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ DỮ LIỆU

## PHÁT BIỂU VỀ DỮ LIỆU

### 1.1.1 Mô tả dữ liệu

* Tên bộ dữ liệu: US Accidents (2016 – 2023).
* Tác giả: Sobhan Moosavi.
* Đây là Dataset về tai nạn xe hơi trên toàn nước Mỹ, bao phủ **49 bang**. Dữ liệu tai nạn được thu thập từ **tháng 2/2016 đến tháng 3/2023** cho toàn bộ các bang thuộc lục địa Hoa Kỳ, sử dụng nhiều API phát trực tuyến dữ liệu sự cố (hoặc sự kiện) giao thông. Các API này truyền dữ liệu giao thông được ghi nhận từ nhiều nguồn khác nhau, bao gồm:
  + Bộ Giao thông Vận tải Hoa Kỳ và các bang.
  + Cơ quan thực thi pháp luật.
  + Camera giao thông.
  + Cảm biến giao thông trong mạng lưới đường bộ.
* Kho dữ liệu gồm 7,728,394 dòng và 46 thuộc tính.
* Link dataset: [US Accidents (2016 - 2023)](https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents).

A screenshot of a web page

AI-generated content may be incorrect.

### 1.1.2 Tiền xử lý

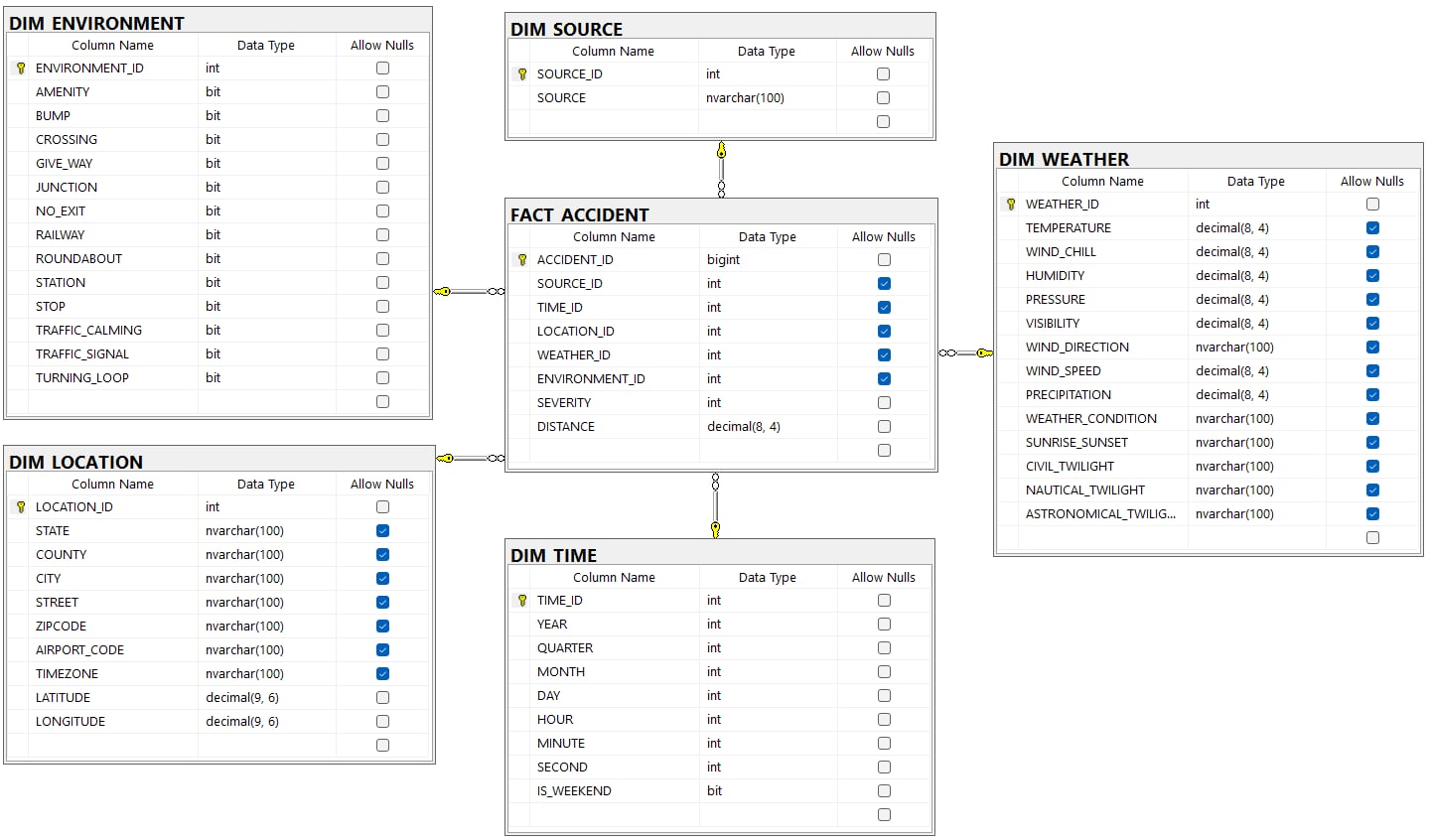
### 1.1.3 Thuộc tính của dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Ý nghĩa** |
| 1 | SOURCE | VARCHAR(50) | Nguồn dữ liệu tai nạn thô (Source1, Source2, Other). |
| 2 | SEVERITY | INT | Mức độ nghiêm trọng tai nạn (1-4), 1 = ít ảnh hưởng đến giao thông, 4 = ảnh hưởng nghiêm trọng. |
| 3 | LATITUDE | FLOAT | Vĩ độ GPS của điểm bắt đầu tai nạn |
| 4 | LONGITUDE | FLOAT | Kinh độ GPS của điểm bắt đầu tai nạn |
| 5 | DISTANCE | FLOAT | Độ dài đoạn đường bị ảnh hưởng bởi tai nạn (đơn vị: dặm) |
| 6 | SECOND | INT | Thời điểm xảy ra tai nạn (giây) |
| 7 | MINUTE | INT | Thời điểm xảy ra tai nạn (phút) |
| 8 | HOUR | INT | Thời điểm xảy ra tai nạn (giờ) |
| 9 | DAY | INT | Thời điểm xảy ra tai nạn (ngày) |
| 10 | MONTH | INT | Thời điểm xảy ra tai nạn (tháng) |
| 11 | QUARTER | INT | Thời điểm xảy ra tai nạn (quý) |
| 12 | YEAR | INT | Thời điểm xảy ra tai nạn (năm) |
| 13 | IS\_WEEKEND | BIT | Có phải cuối tuần không. |
| 14 | STREET | VARCHAR(255) | Tên đường trong trường địa chỉ |
| 15 | CITY | VARCHAR(100) | Tên thành phố trong trường địa chỉ |
| 16 | COUNTY | VARCHAR(100) | Tên quận/hạt trong trường địa chỉ |
| 17 | STATE | VARCHAR(10) | Tên tiểu bang trong trường địa chỉ |
| 18 | ZIPCODE | VARCHAR(20) | Mã bưu chính trong trường địa chỉ |
| 19 | TIMEZONE | VARCHAR(50) | Múi giờ dựa trên vị trí tai nạn (US/Eastern, US/Pacific, etc.) |
| 20 | AIRPORT\_CODE | VARCHAR(10) | Mã sân bay của trạm quan trắc thời tiết gần nhất |
| 21 | TEMPERATURE | FLOAT | Nhiệt độ (độ Fahrenheit) |
| 22 | WIND\_CHILL | FLOAT | Chỉ số gió lạnh (độ Fahrenheit) |
| 23 | HUMIDITY | FLOAT | Độ ẩm (phần trăm) |
| 24 | PRESSURE | FLOAT | Áp suất khí quyển (inch) |
| 25 | VISIBILITY | FLOAT | Tầm nhìn (dặm) |
| 26 | WIND\_DIRECTION | VARCHAR(10) | Hướng gió (CALM, S, N, etc.) |
| 27 | WIND\_SPEED | FLOAT | Tốc độ gió (dặm/giờ) |
| 28 | PRECIPITATION | FLOAT | Lượng mưa (inch, nếu có) |
| 29 | WEATHER\_CONDITION | VARCHAR(100) | Điều kiện thời tiết (Fair, Mostly Cloudy, rain, snow, etc.) |
| 30 | AMENITY | BIT | Có tiện ích gần đó hay không (POI annotation) |
| 31 | BUMP | BIT | Có gờ giảm tốc gần đó hay không |
| 32 | CROSSING | BIT | Có lối băng qua đường gần đó hay không |
| 33 | GIVE\_WAY | BIT | Có biển nhường đường gần đó hay không |
| 34 | JUNCTION | BIT | Có ngã tư/giao lộ gần đó hay không |
| 35 | NO\_EXIT | BIT | Có biển cấm ra gần đó hay không |
| 36 | RAILWAY | BIT | Có đường sắt gần đó hay không |
| 37 | ROUNDABOUT | BIT | Có bùng binh gần đó hay không |
| 38 | STATION | BIT | Có trạm (xe buýt, tàu, etc.) gần đó hay không |
| 39 | STOP | BIT | Có biển dừng gần đó hay không |
| 40 | TRAFFIC\_CALMING | BIT | Có biện pháp làm chậm giao thông gần đó hay không |
| 41 | TRAFFIC\_SIGNAL | BIT | Có đèn giao thông gần đó hay không |
| 42 | TURNING\_LOOP | BIT | Có vòng xuyến gần đó hay không (không có giá trị true nào) |
| 43 | SUNRISE\_SUNSET | VARCHAR(10) | Thời điểm trong ngày dựa trên mặt trời mọc/lặn (Day/Night) |
| 44 | CIVIL\_TWILIGHT | VARCHAR(10) | Thời điểm trong ngày dựa trên hoàng hôn dân dụng (Day/Night) |
| 45 | NAUTICAL\_TWILIGHT | VARCHAR(10) | Thời điểm trong ngày dựa trên hoàng hôn hàng hải (Day/Night) |
| 46 | ASTRONOMICAL\_TWILIGHT | VARCHAR(10) | Thời điểm trong ngày dựa trên hoàng hôn thiên văn (Day/Night) |

### 1.1.4 Hướng chủ đề

## 1.2 XÂY DỰNG KHO DỮ LIỆU

### 1.2.1 Thiết kế lược đồ



### 1.2.2 Các bảng chiều

## 1.3 CÁC CÂU TRUY VẤN

1. Danh sách tai nạn theo trạng thái có tín hiệu giao thông hay không.
2. Khoảng cách trung bình của vụ tai nạn theo tiểu bang.
3. Phân tích tai nạn theo múi giờ.
4. Ảnh hưởng của tầm nhìn (VISIBILITY) đến độ nghiêm trọng.
5. Phân bố tai nạn theo quý và năm.
6. Phân tích mức độ nghiêm trọng tai nạn theo từng giai đoạn ánh sáng (dựa trên twilight).
7. Danh sách tai nạn gần đường sắt hoặc trạm xe lửa theo từng năm.
8. Tần suất và mức độ nghiêm trọng của tai nạn gần các tiện ích.
9. Top 10 con đường nguy hiểm nhất và các yếu tố môi trường liên quan.
10. Top 5 địa điểm tai nạn theo độ nghiêm trọng trung bình.
11. Phân tích mối liên hệ giữa điều kiện thời tiết khắc nghiệt và số lượng/mức độ tai nạn.
12. Xếp hạng 5 thành phố có số vụ tai nạn cao nhất trong mỗi tiểu bang.
13. Phân tích tốc độ tăng trưởng số vụ tai nạn qua từng năm theo tiểu bang.
14. Top 5 thành phố có tốc độ tăng trưởng tai nạn trung bình cao nhất giai đoạn 2019–2022.
15. So sánh mức độ nghiêm trọng trung bình của tai nạn giữa cuối tuần và ngày thường theo từng giờ trong ngày.

# CHƯƠNG 2: QUÁ TRÌNH XÂY DỰNG KHO DỮ LIỆU (SSIS)