**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

--⁎---⁂---⁎--

**MÔN: KHO DỮ LIỆU**

**ĐỀ TÀI:**

**POLICE VIOLENCE IN THE US**

GVHD: ThS. Quách Đình Hoàng

SVTH: Nhóm 6

Phùng Huy Hoàng 17133024

Nguyễn Thị Thu Thủy 17133063

Phạm Xuân Hà 17133016

Đỗ Lê Quí 17133052

Phạm Hồng Ý 17133075

TP.HCM, tháng 6 năm 2020

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

*TP. HCM, ngày .... tháng .... năm ....*

**GIÁO VIÊN HƯỚNG DẪN***(Ký và ghi rõ họ tên)*

**ThS. Quách Đình Hoàng**

**LỜI CÁM ƠN**

🙢🕮🙠

Nhóm chúng em xin chân thành cảm ơn Thầy đã tận tình giảng dạy, chỉ dẫn, góp ý và giúp đỡ nhóm em trong suốt quá trình học tập, nghiên cứu và làm đề tài. Những kiến thức mà thầy truyền dạy thực sự giúp chúng em rất nhiều, được hiểu biết thêm về những hệ thống nhúng trong cuộc sống, cách xây dựng, chế tạo ra những hệ thống đó.

Trong đề tài này, những kiến thức chỉ dừng lại nghiên cứu và tìm hiểu không chuyên sâu nên chắc chắn bài báo cáo của nhóm em không tránh khỏi sai sót. Rất mong được sự góp ý và chỉ dẫn thêm từ Thầy. Mọi góp ý em xin chân thành cảm ơn.

**MỤC LỤC**

[MỤC LỤC 3](#_Toc27926507)

[1. GIỚI THIỆU 4](#_Toc27926508)

[2. TÌM HIỂU XÂY DỰNG KHO DỮ LIỆU IMDb 5](#_Toc27926509)

[**2.1. Kho dữ liệu là gì? 5**](#_Toc27926510)

[**2.2. Quá trình Extract-Transform-Load 6**](#_Toc27926511)

[**2.2.1. BI là gì? 6**](#_Toc27926512)

[**2.2.2. ETL là gì? 7**](#_Toc27926513)

[**2.3. Tầm quan trọng xây dựng kho dữ liệu cho dữ IMDb 9**](#_Toc27926514)

[**2.3.1. Lý do chọn đề tài 9**](#_Toc27926515)

[**2.3.2. Tầm quan trọng 9**](#_Toc27926516)

[3. MỤC ĐÍCH 9](#_Toc27926517)

[PHẦN 2: NỘI DUNG 10](#_Toc27926518)

[1. THIẾT KẾ 10](#_Toc27926519)

[**1.1 Giới thiệu dataset dữ liệu điện ảnh IMDB 10**](#_Toc27926520)

[**1.1. Tiền xử lý dữ liệu 11**](#_Toc27926521)

[**1.2. Thiết kế data mart 11**](#_Toc27926522)

[2. TẠO CÁC BÁO CÁO TỔNG HỢP 12](#_Toc27926523)

[TÀI LIỆU THAM KHẢO 12](#_Toc27926524)

**PHẦN 1: MỞ ĐẦU**

1. **GIỚI THIỆU**

Một báo cáo gần đây của Cục Thống kê Tư pháp ước tính khoảng 1.100 người đã bị cảnh sát giết mỗi năm. Trong đó nhiều cáo buộc có liên quan tới phân biệt chủng tộc. Tuy nhiên những cáo buộc trước đây thường không có bằng chứng rõ ràng. Cho tới khi xảy ra cái chết của người đàn ông Mỹ gốc Phi George Floyd, làn sóng phản đối phân biệt chủng tộc mới bùng lên mạnh mẽ. Nhằm tìm hiểu vấn nạn phân biệt chủng tộc ở Mỹ qua việc cảnh sát ra tay trấn áp tội phạm, nhóm chúng em đã tìm đến bộ cơ sở dữ liệu “Police violence in the US”.

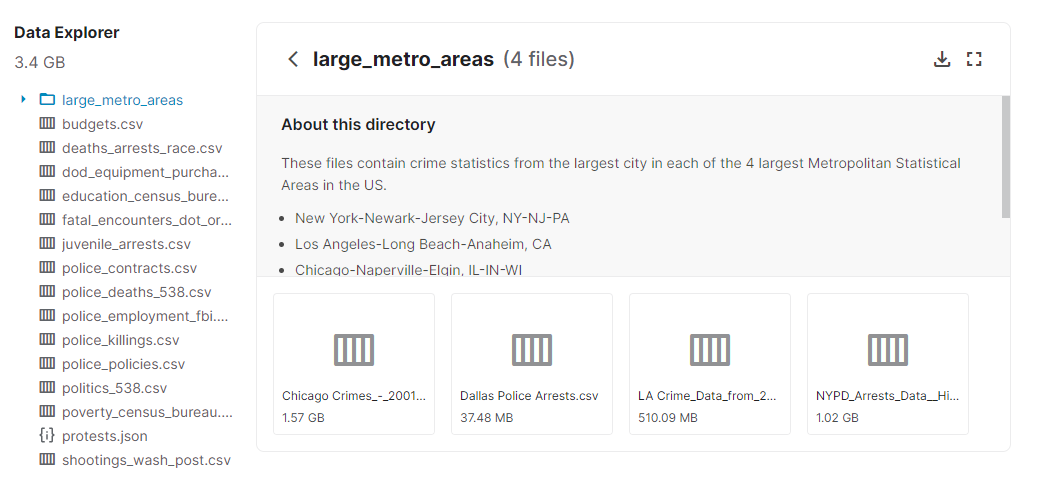
Cơ sở dữ liệu trên bao gồm nhiều mặt vấn đề ở Mỹ, bao gồm các thông tin về số vụ giết người do cảnh sát, những cái chết của cảnh sát, các điều luật về việc nổ súng của từng liên bang, trình độ học vấn của người dân, đóng góp thuế của người dân các bang...Mặc dù chắc chắn có những vụ cảnh sát giết người không có trong cơ sở dữ liệu (cụ thể là những vụ giết người không được báo chí đưa ra), những ước tính này cho thấy cơ sở dữ liệu nắm bắt được 92% tổng số vụ đã xảy ra kể từ năm 2013.

Thông tin này được tổng hợp từ ba cơ sở dữ liệu đám đông lớn nhất, toàn diện nhất và công bằng nhất về các vụ cảnh sát giết người ở nước Mỹ: FatalEncounters.org, Cơ sở dữ liệu của Cảnh sát Hoa Kỳ và KillsbyPolice.net



1. **MÔ TẢ DỮ LIỆU**

* Tên đề tài: Police violence in the US
* Đường link đến dataset:
* <https://www.kaggle.com/jpmiller/police-violence-in-the-us?fbclid=IwAR01IvDW9OxTXJ0BOxX8uXwcyJZNb-9q1UTiQ5csErw5t4A6l8r74lpQnEM&select=police_deaths_538.csv>
* Mô tải ngắn gọn về đề tài:



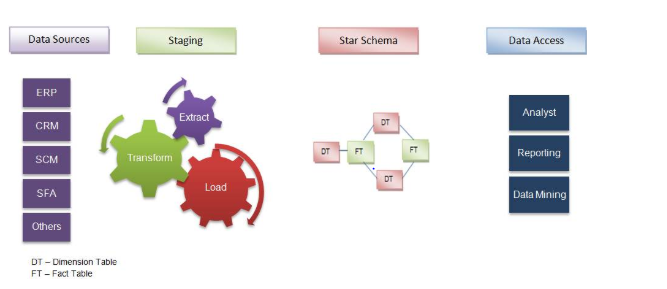
Bộ dữ liệu này tập hợp dữ liệu từ nhiều nguồn khác nhau về bạo lực liên quan đến cảnh sát ở Hoa Kỳ, bao gồm nhiều file CSV liên quan tới nhiều vấn đề. Tuy nhiên để phù hợp với một bài đồ án nhỏ, nhóm chúng em chọn 2 database sau:

- Police\_killings: số lượng tử vong của công dân liên quan đến cảnh sát

- Police\_deaths: số lượng cảnh sát tử vong trong nhiệm vụ

1. **TÌM HIỂU XÂY DỰNG KHO DỮ LIỆU IMDb**
   1. **Kho dữ liệu là gì?**

Kimball định nghĩa kho dữ liệu là một bản sao của dữ liệu giao dịch được cấu trúc cụ thể cho truy vấn và phân tích.



Inmon định nghĩa kho dữ liệu theo các thuật ngữ sau:

Hướng theo chủ đề **(subject – oriented)**: Dữ liệu trong kho dữ liệu được sắp xếp sao cho tất cả các yếu tố dữ liệu liên quan đến cùng một sự kiện hoặc đối tượng trong thế giới thực được liên kết với nhau. (ví dụ: khách hàng, sản phẩm và nhân viên thay vì quy trình kinh doanh.)

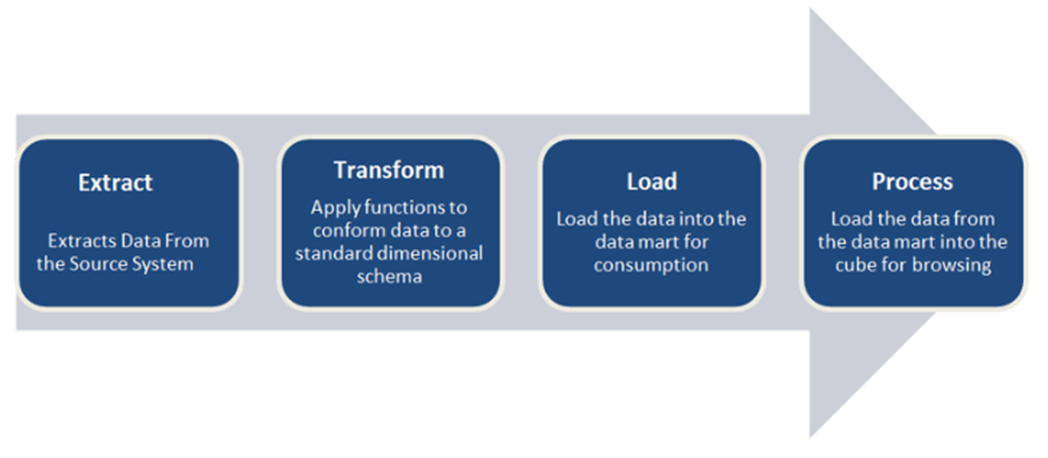
Giá trị lịch sử **(time – varying)**: Các thay đổi đối với dữ liệu trong cơ sở dữ liệu được theo dõi và ghi lại để các báo cáo có thể được tạo ra cho thấy các thay đổi theo thời gian (Ví dụ thông tin địa chỉ, email, số điện thoại của khách hàng có thể thay đổi. Nhưng việc thay đổi đó không được phép tác động đến giá trị báo cáo, phân tích thực hiện trước khi sự thay đổi xảy ra).

Tính toàn vẹn **(integrated):** Cơ sở dữ liệu chứa dữ liệu từ hầu hết hoặc tất cả các ứng dụng hoạt động của tổ chức và dữ liệu này được thực hiện nhất quán (đơn vị đo, định dạng dữ liệu, quy ước đặt tên).

Bất biến **(nonvolatile)**: Dữ liệu trong kho dữ liệu không bao giờ bị ghi đè hoặc bị xóa. Sau khi cam kết, dữ liệu là tĩnh, chỉ đọc và được giữ lại để báo cáo trong tương lai.

* 1. **Quá trình Extract-Transform-Load**
     1. **BI là gì?**

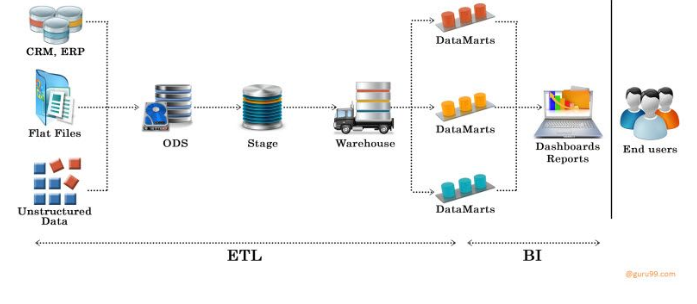
  Business Intelligence là quá trình thu thập dữ liệu thô hoặc dữ liệu kinh doanh và biến nó thành thông tin hữu ích và có ý nghĩa hơn. Các dữ liệu thô là các hồ sơ của các giao dịch hàng ngày của một tổ chức như tương tác với khách hàng, quản lý tài chính và quản lý nhân viên, .... Những dữ liệu này sẽ được sử dụng cho “Báo cáo, phân tích, khai thác dữ liệu, chất lượng dữ liệu và Giải thích, phân tích tiên đoán”. Data Warehouse là gì? Một kho dữ liệu là một cơ sở dữ liệu được thiết kế cho truy vấn và phân tích hơn là cho xử lý giao dịch. Các kho dữ liệu được xây dựng bằng cách tích hợp dữ liệu từ nhiều nguồn. Nó cho phép công ty hoặc tổ chức hợp nhất dữ liệu từ nhiều nguồn và tách khối lượng công việc phân tích từ khối lượng công việc giao dịch. Dữ liệu được chuyển thành thông tin chất lượng cao để đáp ứng tất cả các yêu cầu báo cáo doanh nghiệp cho tất cả các cấp của người dùng.



* + 1. **ETL là gì?**

ETL là viết tắt của Extract-Transform-Load và nó là quá trình làm thế nào dữ liệu được tải từ hệ thống nguồn vào kho dữ liệu. Dữ liệu được trích xuất từ cơ sở dữ liệu OLTP, được chuyển đổi để phù hợp với lược đồ kho dữ liệu và được nạp vào cơ sở dữ liệu kho dữ liệu. Nhiều kho dữ liệu cũng kết hợp dữ liệu từ các hệ thống không phải là OLTP như các tệp văn bản, các hệ thống kế thừa và bảng tính.

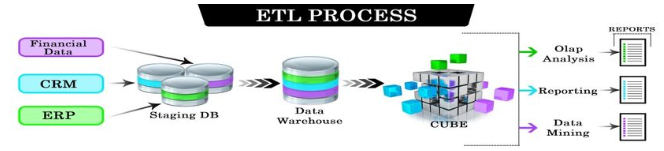
Các giải pháp là sử dụng một Datawarehouse để lưu trữ thông tin từ các nguồn khác nhau trong một cấu trúc thống nhất sử dụng ETL. ETL có thể biến đổi các bộ dữ liệu không giống nhau thành một cấu trúc thống nhất. Sau đó sử dụng các công cụ BI để thu thập các thông tin chi tiết và báo cáo có ý nghĩa từ dữ liệu này Sơ đồ dưới đây cho bạn biết MAP ROAD của quá trình ETL



**Extract:** Trích xuất dữ liệu có liên quan

**Transform:** Chuyển đổi dữ liệu sang định dạng DW (Data Warehouse) Khóa xây dựng - Khóa chính là một hoặc nhiều thuộc tính dữ liệu nhận dạng duy nhất một thực thể. Các loại khóa chính là chìa khóa chính, phím xen kẽ, khoá nước ngoài, phím ghép, phím đại diện. Kho dữ liệu sở hữu các phím này và không bao giờ cho phép bất kỳ thực thể khác gán cho chúng. Rửa sạch dữ liệu: Sau khi dữ liệu được chiết xuất, nó sẽ di chuyển vào giai đoạn tiếp theo, làm sạch và phù hợp của dữ liệu. Việc làm sạch sẽ làm mất dữ liệu cũng như xác định và sửa lỗi. Việc tuân thủ có nghĩa là giải quyết mâu thuẫn giữa những dữ liệu đó không tương thích để chúng có thể được sử dụng trong kho dữ liệu doanh nghiệp. Ngoài ra, hệ thống này tạo ra siêu dữ liệu được sử dụng để chẩn đoán các vấn đề hệ thống nguồn và nâng cao chất lượng dữ liệu.

**Load:** Tải dữ liệu vào DW (Data Warehouse) Build aggregates - Tạo tổng hợp là tổng hợp và lưu trữ dữ liệu có sẵn trong bảng sự kiện để cải thiện hiệu suất của các truy vấn của người dùng cuối.



* 1. **Tầm quan trọng xây dựng kho dữ liệu cho police violence in the US**
     1. **Lý do chọn đề tài**

Dựa vào dữ liệu được dùng để cung cấp tính minh bạch và trách nhiệm cao hơn cho các sở cảnh sát như là một phần của chiến dịch đang diễn ra nhằm chấm dứt bạo lực cảnh sát trong cộng đồng.

* + 1. **Tầm quan trọng**

Thống kê một cách toàn diện nhất về những người bị cảnh sát giết từ năm 2013. Một báo cáo gần đây của Cục Thống kê Tư pháp Mĩ ước tính khoảng 1.200 người đã bị cảnh sát giết từ tháng 6 năm 2015 đến tháng 5 năm 2016. Cơ sở dữ liệu này xác định 1.106 người bị cảnh sát giết trong khoảng thời gian này.

1. **MỤC ĐÍCH**

- Hiểu các khái niệm kho dữ liệu

- Áp dụng các phương pháp thiết kế kho dữ liệu để tạo bản thiết kế data warehouse/data mart dựa trên dataset đã chọn

- Tạo quy trình tích hợp dữ liệu

- Hỗ trợ cơ sở dữ liệu quan hệ cho kho dữ liệu

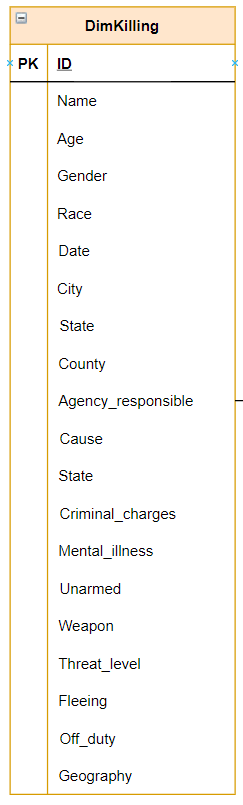
- Xây dựng truy vấn bằng SQL phân tích

- Thiết kế và sử dụng các khung nhìn cụ thể hóa

# PHẦN 2: NỘI DUNG

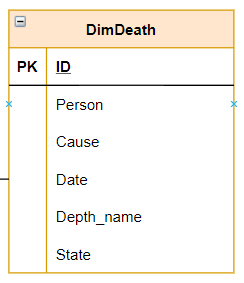
# THIẾT KẾ

* 1. **Giới thiệu dataset dữ liệu** 
     1. **DimKilling**



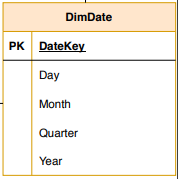
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Thuộc tính** | | **Ý nghĩa** | |
| 1 | Name | | Tên | |
| 2 | Age | | Tuổi | |
| 3 | Gender | | Giới tính | |
| 4 | Race | | Chủng tộc | |
| 5 | Date | | Thời gian tử vong | |
| 6 | City | | Thành phố | |
| 7 | State | | Bang | |
| 8 | | Country | | Quận | |
| 9 | | Agency\_responsible | | Sở cảnh sát tiếp nhận vụ án | |
| 10 | | Cause | | Nguyên nhân chết | |
| 11 | | Status | | Tình trạng xử lý vụ án | |
| 12 | | Criminal\_charges | | Truy tố trách nghiệm hình sự | |
| 13 | | Mental\_illness | | Bệnh tâm thần. Khi gây án thì có sử dụng chất kích thích, hay có bệnh lý gì không | |
| 14 | | Unarmed | | Có bị cáo buộc vũ trang hay không | |
| 15 | | Weapon | | Vũ khí sử dụng | |
| 16 | | Threat\_level | | Mức độ đe dọa | |
| 17 | | Fleeing | | Hình thức chạy trốn | |
| 18 | | Off\_duty | | Nạn nhân bị hạ trong hay ngoài nhiệm vụ | |
| 19 | | Geography | | Khu vực xảy ra | |

* + 1. **DimDeath**



|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa** |
| 1 | Person | Tên và chức vụ |
| 2 | Cause | Nguyên nhân mất |
| 3 | Date | Ngày mất |
| 4 | Depth\_name | Tên cơ quan làm việc |
| 5 | State | Bang |

* + 1. **DimDate**



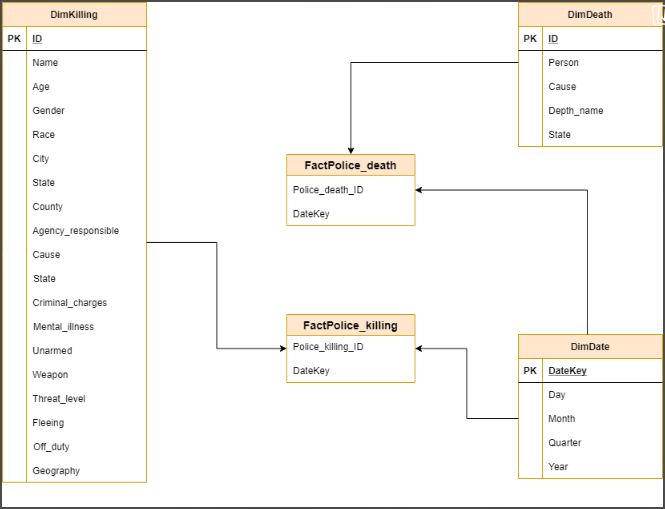
|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa** |
| 1 | Day | Ngày |
| 2 | Month | Tháng |
| 3 | Quarter | Quý |
| 4 | Year | Năm |

## Tiền xử lý dữ liệu

Do cơ sở dữ liệu nhận được có nhiều thuộc tính chưa được chuẩn hóa (tên các bang còn viết tắt, chưa thống nhất về định dạng cho ngày tháng), đồng thời nó chứa những cột không cần thiết nên ta tiến hành chuẩn hóa như sau:

|  |
| --- |
| import pandas as pd  file\_name = 'police\_killings.csv'  file\_export\_name = 'police\_killings2.csv'  coll\_names = [  'Name', 'Age', 'Gender', 'URL\_image', 'Date', 'Street',  'City', 'State', 'Zipcode', 'County', 'Agency\_responsible',  'Cause', 'Description', 'Status', 'Criminal\_charges', 'URL\_news',  'Mental\_illness', 'Unarmed', 'Weapon', 'Threat\_level', 'Fleeing',  'Body\_camera', 'WaPo\_ID', 'Off\_duty', 'Geography', 'ID'  ]  removed\_column = [  'URL\_image', 'Street', 'Zipcode', 'Description', 'URL\_news',  'Body\_camera', 'WaPo\_ID', 'ID'  ]  replace\_list = {  'AL':'Alabama',  'AK':'Alaska',  'AZ':'Arizona',  'AR':'Arkansas',  'CA':'California',  'CO':'Colorado',  'CT':'Connecticut',  'DE':'Delaware',  'DC':'Washington D.C.',  'FL':'Florida',  'GA':'Georgia',  'HI':'Hawaii',  'ID':'Idaho',  'IL':'Illinois',  'IN':'Indiana',  'IA':'Iowa',  'KS':'Kansas',  'KY':'Kentucky',  'LA':'Louisiana',  'ME':'Maine',  'MD':'Maryland',  'MA':'Massachusetts',  'MI':'Michigan',  'MN':'Minnesota',  'MS':'Mississippi',  'MO':'Missouri',  'MT':'Montana',  'NE':'Nebraska',  'NV':'Nevada',  'NH':'New Hampshire',  'NJ':'New Jersey',  'NM':'New Mexico',  'NY':'New York',  'NC':'North Carolina',  'ND':'North Dakota',  'OH':'Ohio',  'OK':'Oklahoma',  'OR':'Oregon',  'PA':'Pennsylvania',  'RI':'Rhode Island',  'SC':'South Carolina',  'SD':'South Dakota',  'TN':'Tennessee',  'TX':'Texas',  'US':'United States',  'UT':'Utah',  'VT':'Vermont',  'VA':'Virginia',  'WA':'Washington',  'WV':'Virginia',  'WI':'Wisconsin',  'WY':'Wyoming',  'AS':'American Samoa',  'GU':'Guam',  'MP':'Northern Mariana Islands',  'PR':'Puerto Rico',  'VI':'Virgin Islands'  }    # Đọc file csv  df = pd.read\_csv(file\_name)  # Đổi tên các cột  df.columns = coll\_names  # Loại bỏ những cột không cần thiết  df = df.drop(removed\_column, axis = 1)  # Thay thế tên bang viết tắt thành tên đầy đủ  df = df.replace(replace\_list)  # Chuẩn hóa cột Date  date = []  # Ta tách Date thành các phần, ngày và tháng ta kiểm tra  # xem có đủ 2 ký tự không  # Sau đó ghép chúng lại theo định dạng YearMonthDay  # (ví dụ 20200101 chính là ngày 1/1/2020)  for i in range(len(df)):  text = df.Date[i].split('/')  if (len(text[0]) != 2):  text[0] = '0' + text[0]  if (len(text[1]) != 2):  text[1] = '0' + text[1]  date.append(text[2] + text[0] + text[1])  df.Date = date  # Xuất file CSV  df.to\_csv(file\_export\_name, index = False) |

Ta làm tương tự với file police\_deaths.csv. File CSV sau khi tiền xử lý đã sẵn sàng để đưa vào SQL Server.



Star schema của Data warehouse

## 1.3. Thiết kế data mart

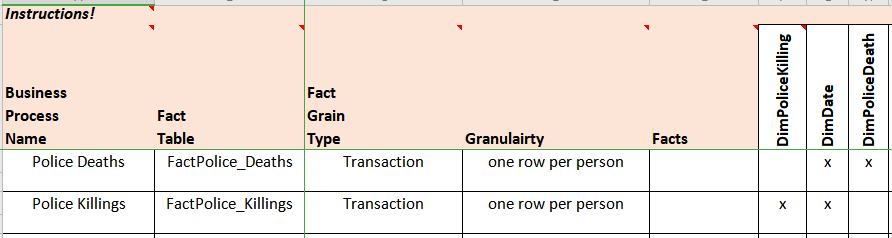
**Bảng fact mức chi tiết Police Deaths (Transaction grain fact table)**

Từ cơ sở dữ liệu đã cho, phân tích số lượng cảnh sát mất trong khoảng thời gian từ năm 2013-2016 theo chức vụ, thời gian và nguyên nhân.

**Bảng fact mức chi tiết Police Killing (Transaction grain fact table)**

Từ cơ sở dữ liệu đã cho, phân tích số lượng người dân bị cảnh sát giết trong khoảng thời gian từ năm 2013-2019 theo thời gian và nguyên nhân.

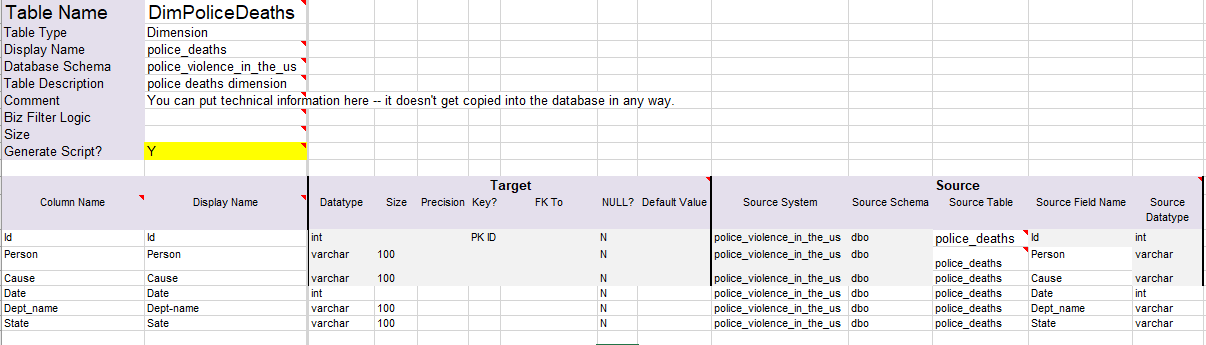
**High-Level-Dimensional-Modeling-Workbook**



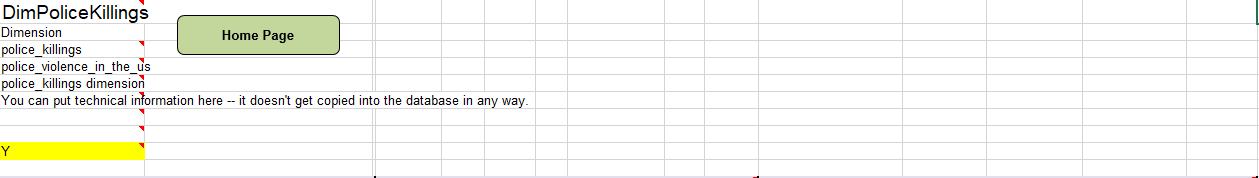
* Mỗi fact là mỗi dòng trong FactPolice\_Deaths mô tả về nguyên nhân chết, thời gian của một vị cảnh sát đảm nhiệm chức vụ gì, thuộc cơ quan nào ở đâu.
* Mỗi fact là mỗi dòng trong FactPolice\_Killing mô tả thông tin về một người bị cảnh sát giết (tên, tuổi, địa chỉ, giới tính, ….), nguyên nhân bị giết, thời gian bị giết …

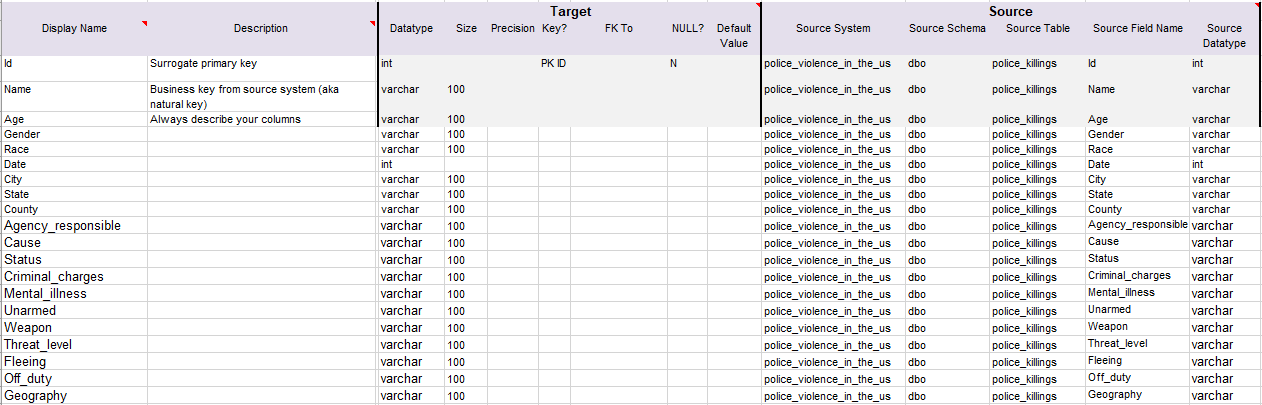
**Các Dimention**

* **Dimention DimPoliceKillings**



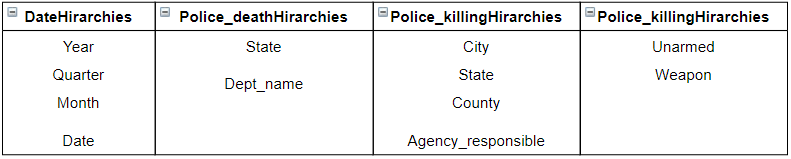
**- Dimention DimPoliceDeaths**





* + 1. **Bảng fact mức chi tiết (Transaction grain fact table): release reporting**

Khởi tạo các Hirarchies: DateHirarchies, Police\_deathHirarchies, Police\_killingHirarchies.



Data cubes:

Nhận xét:

* Số người bị cảnh sát giết phần lớn là 3 chủng tộc: da trắng, da đen và người Hispanic
* Những người bị giết thuộc các chủng tộc: người châu Á, người Mỹ bản địa, người dân đảo Thái Bình Dương chiếm tỷ lệ rất nhỏ

Nhận xét:

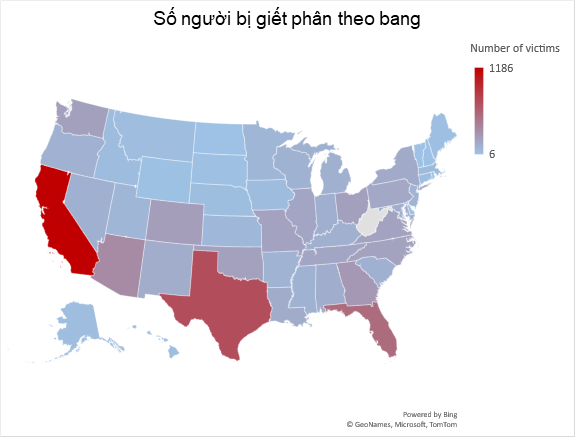
* Khi xét theo tỷ lệ người bị bắn, có 45% là người da trắng, 26% là người da đen, trong khi đó tỷ lệ dân số Mỹ chỉ có khoảng 13% là người da đen

Nhận xét:

* Đa số người bị giết là đàn ông, chiếm khoảng 94.88%
* Theo đó ta có tỷ lệ: cứ 19 người đàn ông thì sẽ có một người phụ nữ bị cảnh sát giết

Nhận xét:

* Các bang có nhiều vụ cảnh sát giết người nhất là California, Texas và Florida, kế đến là Arizona, Georgia.



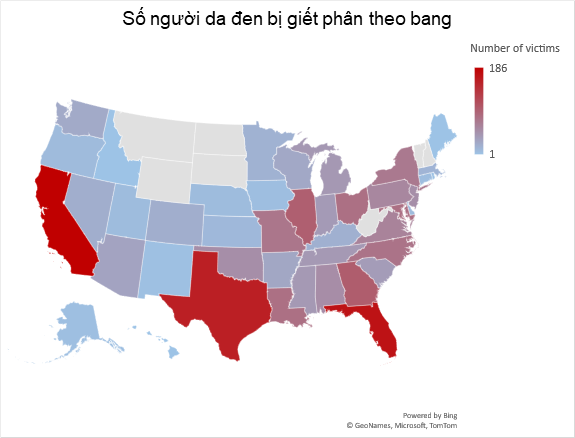
* Dựa theo bản đồ, ta có thể thấy ngoài 3 bang là California, Texas và Florida, số người bị cảnh sát giết phần lớn phân bố ở các bang phía đông nước Mỹ

Nhận xét:

* Bang California có nhiều người châu Á chết do cảnh sát nhất, đồng thời cũng là bang có nhiều người Châu Á nhất. Những bang còn lại số vụ việc xảy ra không đáng kể.

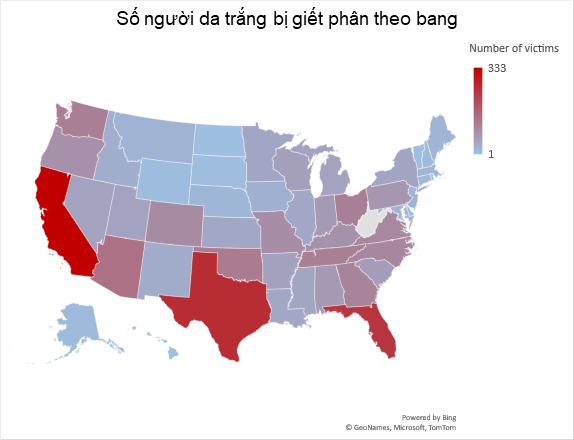
Nhận xét:

* California, Florida và Texas là ba bang có nhiều người da đen bị cảnh sát giết nhất. Các bang còn lại có số vụ xảy ra tương đối nhiều.



Nhận xét:

* Không khác mấy so với biểu đồ của người da đen, người da trắng cũng bị sát hại nhiều ở 3 bang California, Texas và Florida. Những bang còn lại có số vụ khá nhiều.



* Tuy nhiên theo bản đồ trên, ta thấy các vụ việc phân bố tập trung ở phía đông và phía tây nước Mỹ

Nhận xét:

* Về tổng quan, có khoảng 68% số vụ nạn nhân không hề bỏ chạy khi bị bắt, chiếm một tỷ lệ khá lớn
* Người châu Á có tỷ lệ không bỏ chạy lớn nhất (chiếm 76% số vụ)
* Người da đen có tỷ lệ bỏ chạy cao nhất, chiếm 38% số vụ, bao gồm các hình thức như chạy bằng ô tô, chạy bộ hoặc dùng các phương tiện khác

Nhận xét:

* Về tổng quan, có 71% số vụ người bị cảnh sát sát hại bị cáo buộc có sử dụng vũ trang
* Người châu Á tuy chỉ chiếm 1% số trường hợp bị sát hại, tuy nhiên tỷ lệ bị cáo buộc có vũ trang lại cao nhất, chiếm 79%
* Người da đen có tỷ lệ cáo buộc sử dụng vũ trang ít nhất, chiếm 67% số vụ

Nhận xét:

* 51% số vụ người châu Á bị cảnh sát giết trong trạng thái tỉnh táo, trong đó 25% trường hợp có vấn đề về tâm thần và 4% có sử dụng đồ uống có cồn
* Đối với người da trắng cũng có tỷ lệ gần giống với người châu Á
* Trong khi đó 71% số vụ người da đen bị giết trong tình trạng tỉnh táo

**KẾT LUẬN TỔNG QUAN**

* Qua những số liệu trên, ta có một cái nhìn sơ lược về vấn đề cảnh sát giết người tại Mỹ. Các quyết định nổ súng của cảnh sát tùy thuộc vào các yếu tố tại thời điểm đó như: đối tượng có bỏ chạy hay không, đối tượng có sử dụng vũ trang hay không, đối tượng có dấu hiệu về thần kinh hay say xỉn không… Ngoài ra còn có các yếu tố phụ như: tỷ lệ tội phạm người da màu tại một bang, các băng đảng tội phạm tại khu vực, số vụ đối tượng tội phạm chống trả lại cảnh sát (hoặc sát hại lại cảnh sát), trình độ học vấn, thu nhập, tỷ lệ thất nghiệp của khu vực,….
* Bởi vì vấn đề phân biệt chủng tộc là rất rộng, không chỉ trong việc cảnh sát trấn áp tội phạm mà còn trải dài trong các vấn đề như: việc làm, nhà ở, giáo dục, cho vay, tiếp cận y tế…. Nên bài phân tích trên không thể khái quát hóa hết vấn nạn này được.

# PHẦN 3: KẾT LUẬN

1. **KHÓ KHĂN KHI LÀM ĐỀ TÀI**

Do tập dữ liệu trước đó không đủ lớn để phân tích nên nhóm đã phải chọn lại dữ liệu, mất nhiều thời gian để tìm dữ liệu mới. Thời gian hạn hẹp nên không thể hoàn thành tốt nhất. Khá mất công trong việc tìm tập dữ liệu phù hợp.

Các bảng quá nhiều thuộc tính làm khó phán đoán và thực hiện chính xác nhất.

Dữ liệu tải về chưa được chuẩn hoá, quá trình chuẩn hoá và tiền xử lý dữ liệu mất nhiều thời gian.

1. **NHỮNG GÌ ĐÃ ĐẠT ĐƯỢC**

* Hiểu được Data Warehouse, và quy trình thực hiện nó.
* Áp dụng những kiến thức đã học vào tệp dữ liệu bất kì.
* Rèn luyện khả năng tư duy trong mỗi tệp dữ liệu khác nhau.
* Khả năng đọc hiểu tài liệu tiếng Anh.
* Chỉnh sửa với file dữ liệu lớn

1. **HƯỚNG PHÁT TRIỂN**

* Thực hiện cho toàn bộ tệp dữ liệu, xây dựng nhiều Dimention và Fact.

**TÀI LIỆU THAM KHẢO**

[1] Alejandro Vaisman Esteban Zimanyi, Data Warehouse Systems - Design and Implementation, Springer, 2014

[2] https://www.db-book.com/db7/university-lab-dir/sample\_tables-dir/index.html

[3] <https://www.computerweekly.com/tip/Inmon-or-Kimball-Which-approach-is-suitable-for-your-data-warehouse>

[4] <https://viblo.asia/p/kho-du-lieu-va-khai-pha-du-lieu-bJzKmX7B59N>

[5] https://viblo.asia/p/etl-testing-or-data-warehouse-testing-tutorial-phan-i-1Je5EM7m5nL