**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



**BÀI TẬP A2.1:**

**KHO DỮ LIỆU VÀ OLAP**

**Sinh viên thực hiện:** 19521783- Lê Văn Long

19520854- Phạm Hoàng Phúc

**Lớp:** IS217.M22.HTCL

**Giảng viên hướng dẫn:**TS Đỗ Thị Minh Phụng

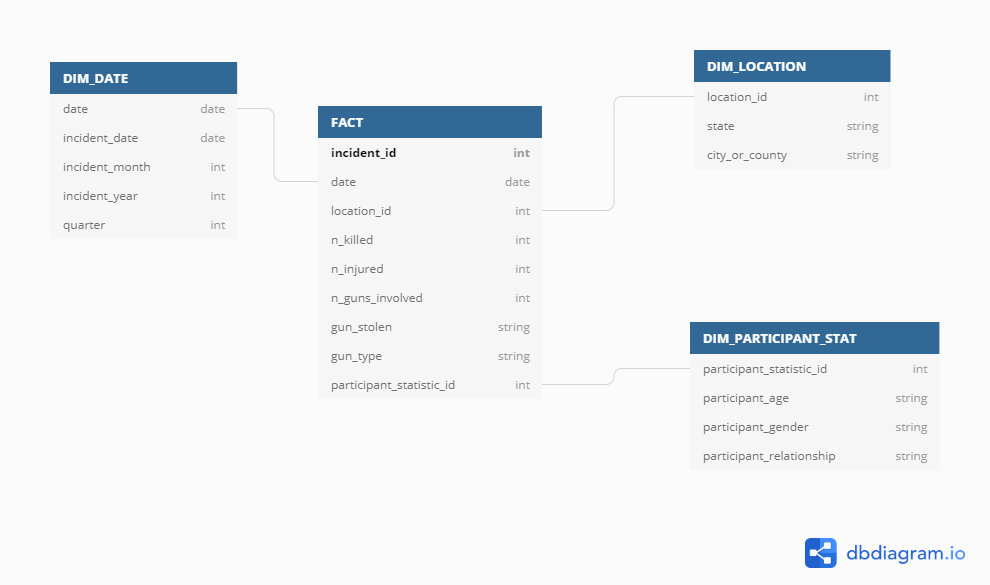
CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI

* 1. Lý do chọn đề tài
* Hiện nay, nạn xả súng tại Mỹ đang là một vấn nạn lớn, gây ra thiệt hại nghiêm trọng về tính mạng con người. Nguyên nhân xả súng đến từ nhiều lí do, trong đó đáng kể nhất là sự quản lý lỏng lẻo của chính quyền, nạn phân biệt chủng tộc tôn giáo, bắt nạt, tư thù cá nhân, mục đích chính trị,…
* Để hiểu rõ hơn về những lí do trên, nhóm chọn đề tài “Phân tích dữ liệu về nạn xả súng ở Mỹ từ 2013 – 2018”
  1. Giới thiệu về Dataset
     1. Giới thiệu về nguồn dữ liệu
        1. Tên dataset
* Tên dataset: Gun Violence Data
* Ngày cập nhật gần nhất: 2018-04-15
  + - 1. Đơn vị cung cấp
* Đây là bộ dữ liệu mô tả về hơn 260,000 vụ xả súng ở Mỹ từ năm 2013 tới 2018, được cung cấp bởi tài khoản Kaggle có tên James Ko. Hiện không có thông tin chi tiết về tác giả, nhưng dataset có 32,500 lượt download, 226,000 lượt xem và 858 lượt upvote cho thấy đây là một dataset có độ chính xác cao.
  + - 1. Mô tả số dòng, cột, thời gian thu thập
* Dataset gồm 29 cột và 239,675 dòng
* Data được thu thâp từ 2008 tới 2013
  + - 1. Nguồn tải dataset
* Link dataset: <https://www.kaggle.com/jameslko/gun-violence-data>
  + 1. Định hướng chủ đề của dataset
* Dataset định hướng theo chủ dề: Social science (khoa học xã hội)
  + 1. Mô tả chi tiết các cột thuộc tính dataset

Sau khi trích lọc các cột và dòng không cần thiết cho mục đích của đồ án, dataset còn 21 cột và 140179 dòng.

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tính | Ý nghĩa | Kiểu dữ liệu |
| 1 | incident\_id | Mã sự kiện | String |
| 2 | date | Ngày sự kiện | Date |
| 3 | state | Tiểu bang xảy ra sự kiện | String |
| 4 | city or country | Thành phố hoặc hạt xảy ra sự kiện | String |
| 5 | address | Địa chỉ xảy ra sự kiện | String |
| 6 | n\_killed | Số trường hợp thiệt mạng | Int |
| 7 | n\_injured | Số trường hợp bị thương | Int |
| 8 | congressional\_district | Khu vực quốc hội | Int |
| 9 | gun\_stolen | Súng bị trộm | Int |
| 10 | gun\_type | Loại súng | Int |
| 11 | incident\_characteristics | Hành động của mỗi cá thể liên quan | String |
| 12 | n\_guns\_involved | Số lượng súng trong sự kiện | Int |
| 13 | participant\_age | Tuổi của các cá thể liên quan | String |
| 14 | participant\_age\_group | Độ tuổi của các cá thể liên quan | String |
| 15 | participant\_gender | Giới tính của các cá thể liên quan | String |
| 16 | participant\_relationship | Quan hệ giữa các cá thể liên quan | String |
| 17 | participant\_status | Trạng thái của các cá thể sau sự kiện | String |
| 18 | participant\_type | Vai trò của các cá thể trong sự kiện | String |
| 19 | state\_house\_district | Khu tiểu bang quốc hội | Int |
| 20 | state\_senate\_district | Khu tiểu bang nghị viện | Int |

* 1. Thiết kế kho dữ liệu
     1. Lượt đồ kho dữ liệu (lượt đồ bông tuyết)



* + 1. Mô tả chi tiết bảng FACT

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Ý nghĩa | Ràng buộc | Kiểu dữ liệu |
| 1 | incident\_id | Mã sự kiện | PK | Int |
| 2 | date | Ngày sự xảy ra sự kiện | FK | Int |
| 3 | location\_id | Mã vị trí | FK | Int |
| 4 | n\_killed | Số tử vong |  | Int |
| 5 | n\_injured | Số bị thương |  | Int |
| 6 | n\_guns\_involved | Số lượng súng |  | Int |
| 7 | gun\_stolen | Súng bị trộm |  | String |
| 8 | gun\_type | Loại súng |  | String |
| 9 | participant\_statistic\_id | Mã số liệu của các cá thể liên quan | FK | Int |

* + 1. Mô tả chi tiết các bảng dimension
       1. DIM\_DATE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Ý nghĩa | Ràng buộc | Kiểu dữ liệu |
| 1 | date | ngày xảy ra sự kiện | PK | Date |
| 2 | incident\_month | Tháng xảy ra sự kiện |  | Int |
| 3 | incident\_year | Năm xảy ra sự kiện |  | Int |
| 4 | incident\_quarter | Quý xảy ra sự kiện |  | Int |

* + - 1. DIM\_PARTICIPANT\_STAT

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Ý nghĩa | Ràng buộc | Kiểu dữ liệu |
| 1 | participant\_statistic\_id | Mã số liệu về người liên quan | PK | Int |
| 2 | participant\_age | Độ tuổi của người liên quan |  | String |
| 3 | male | Số lượng nam trong vụ án |  | int |
| 4 | female | Số lượng nữ trong vụ án |  | int |

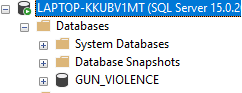
* 1. Các câu truy vấn:
* Thống kê số vụ xả súng xảy ra tại từng tiểu bang của Mỹ trong năm 2013
* Thống kê tháng trong năm xảy ra số vụ xả súng nhiều nhất
* Thống kê độ tuổi chiếm nhiều nhất của nạn nhân các vụ xả súng
* Thống kê tiểu bang có số vụ xả súng nhiều nhất trong 5 năm
* Thống kê tiểu bang có số vụ xả súng ít nhất trong 5 năm
* Thống kê số lượng thương vong xảy ra bởi các vụ xả súng trong từng năm
* Thống kê số lượng súng được sử dụng trong mỗi vụ xả súng trong từng năm
* Thống kê tổng các vụ xả súng theo từng năm để xem xét mức độ tăng giảm của số vụ xả súng
* Thống kê tổng các vụ xả súng theo từng năm ở từng tiểu bang để xem xét mức độ tăng giảm của số vụ xả súng
* Thống kê về số lượng các vụ xả súng xảy ra theo từng buổi trong ngày
* Thống kê về quan hệ giữa nạn nhân và tội phạm xả súng
* Thống kê về loại súng được sử dụng nhiều nhất trong các vụ xả súng
* Thống kê về các vụ xả súng là sự thanh trừng giữa các băng đảng
* Thống kê về các vụ sả súng hướng tới các thành viên trong gia đình

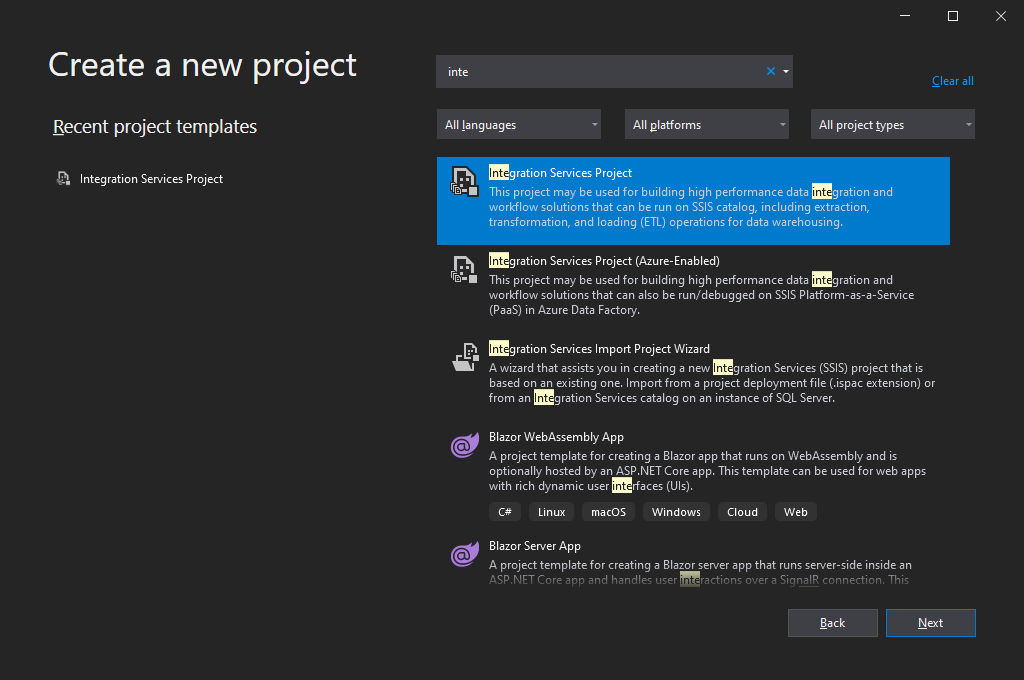
CHƯƠNG 2: QUÁ TRÌNH TẠO PROJECT VÀ THAO TÁC DỮ LIỆU

* 1. SSIS:

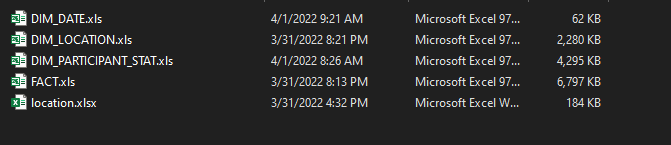
Bước 1: Clean data, xóa bỏ data lỗi và thừa bằng Excel

Bước 2: Tạo Project Integration Services Projects trong Visual Studio và database rỗng trên SQL Server

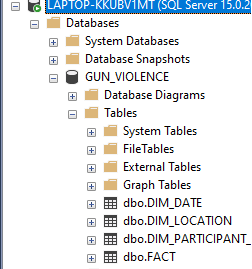




Bước 3: Tiến hành chia dữ liệu trong Excel ra bảng Fact và các bảng Dim



Bước 4: Tạo các bảng trống trong Database đã tạo với cấu trúc giống bước 3

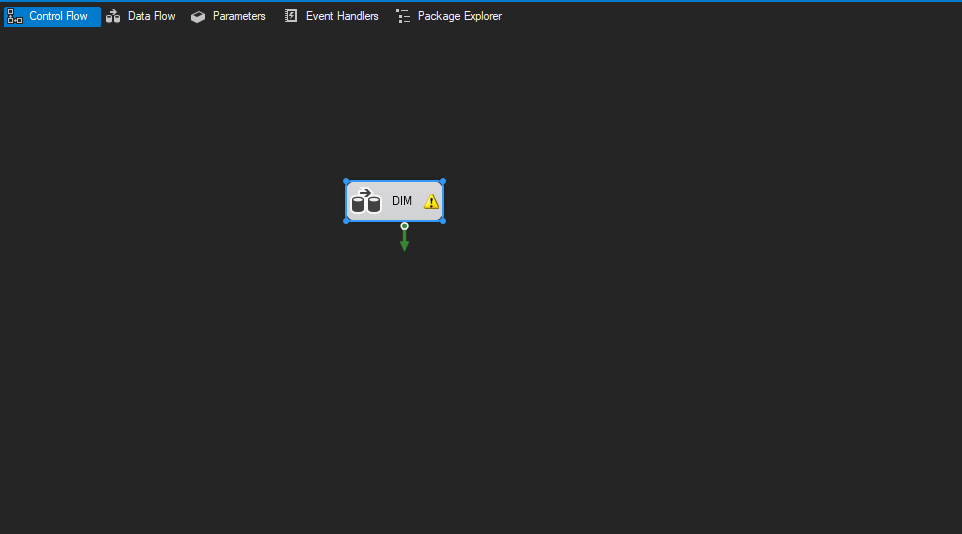


Bước 5: Tạo 2 package trong Visual Studio để chứa các thao tác đổ data vào

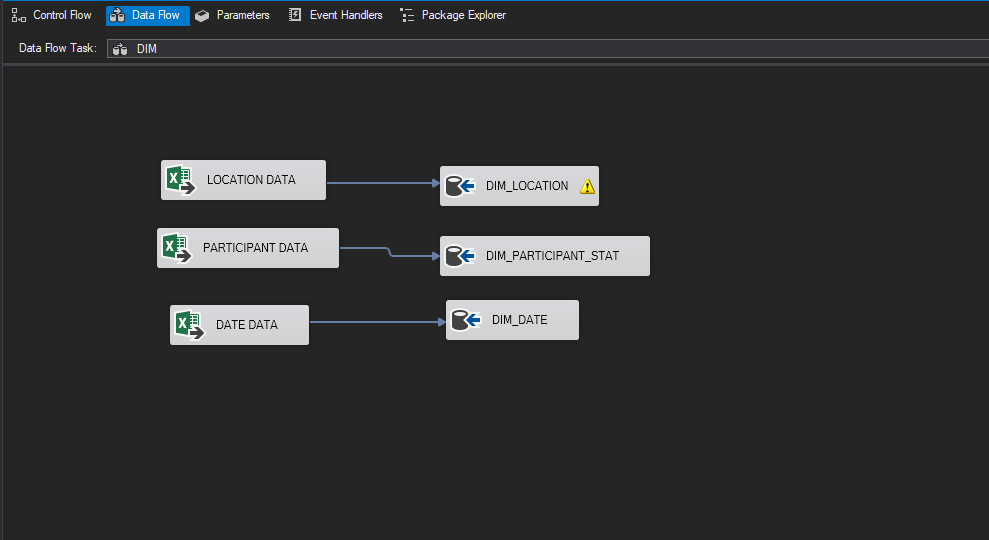
các bảng Dim và bảng Fact.



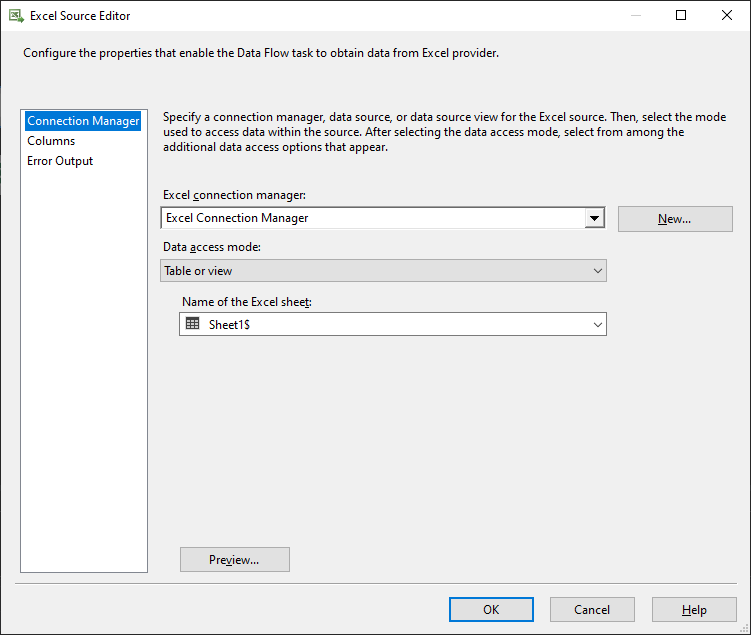
Bước 6: Tại Packet Dim, thiết lập các Data Source và Destination



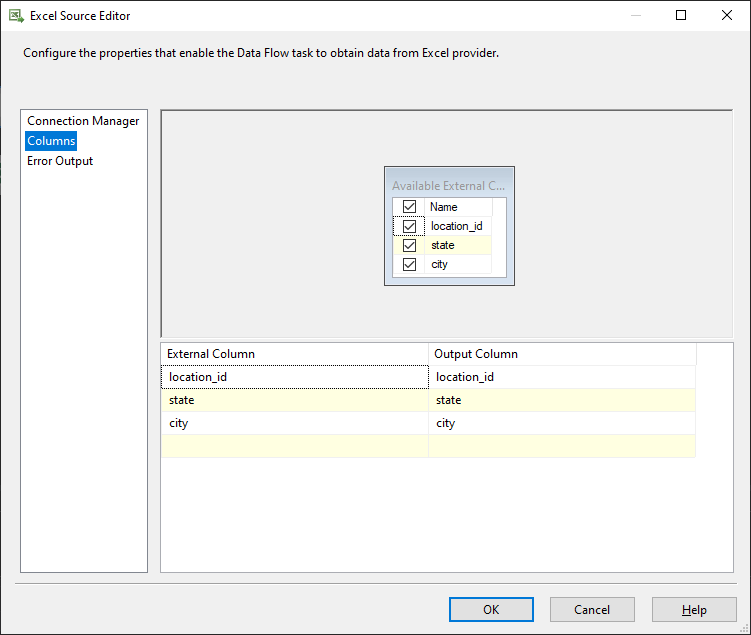
Tạo Data Flow Task



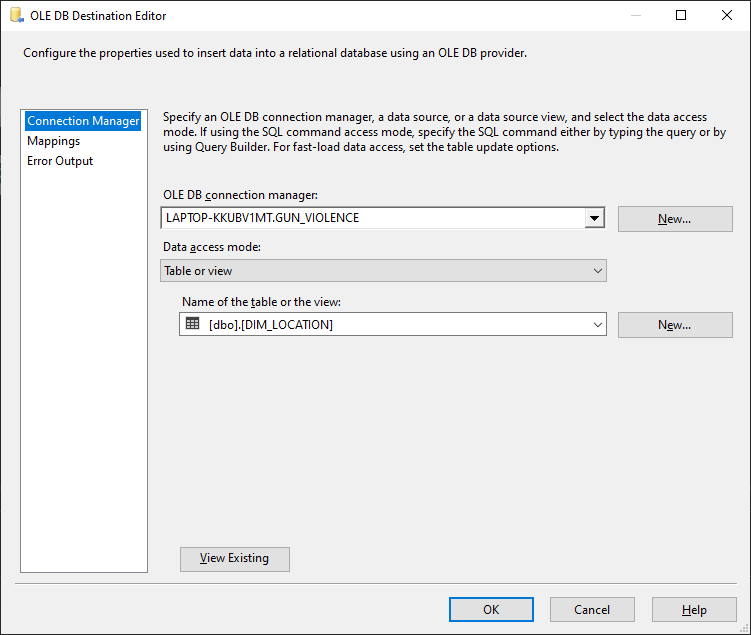
Thiết lập Data Source và Destination



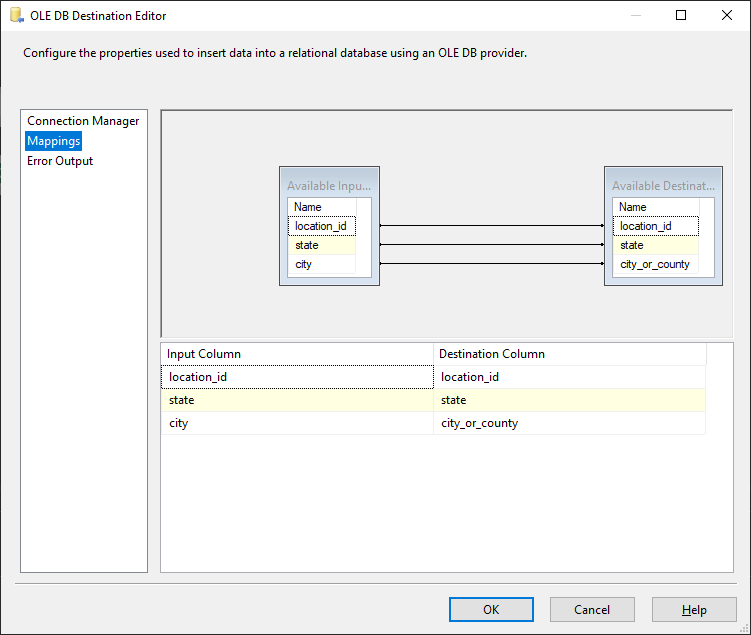
Ở mỗi Data source, browse tới file excel



Chỉ giữ lại các cột cần dùng

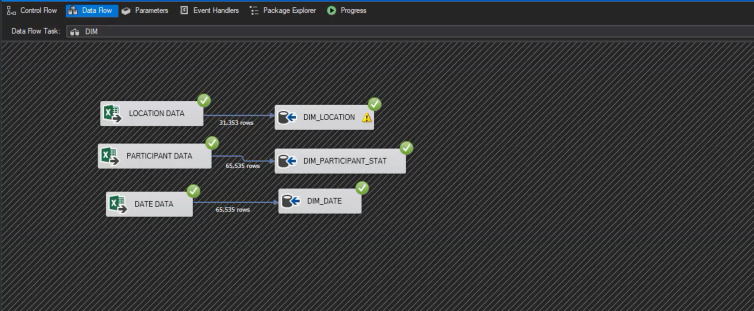


Ở mỗi Data Souce thì set up tên server và bảng sẽ đổ data vào như hình trên



Map các cột từ file Excel với các cột trong database

Bước 7 Run Project để đổ data từ file excel vào database



Bước 8: Làm tương tự cho Fact packet

Bước 9: Chạy query để tạo khóa chính và khóa ngoại



Link sản phẩm:

<https://drive.google.com/drive/folders/1io7yIEiK61xupEZkSw56CyEH0E1REenL?usp=sharing>