**ĐẠI HỌC QUỐC GIA TP.HCM**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**ĐỒ ÁN THỰC HÀNH**

**HTTT PHỤC VỤ TRÍ TUỆ KINH DOANH**

****

**Lớp: CQ2018/1**

Giáo viên lý thuyết: Hồ Thị Hoàng Vy

Giáo viên thực hành: Tiết Gia Hồng - Nguyễn Thị Như Anh

**Mã nhóm thực hiện: TTKD-16**

**NỘP BÁO CÁO LẦN 3**

1. **Thông tin nhóm**

| **Mã nhóm** | **MSSV** | **Họ và tên** | **Email** | **SĐT** |
| --- | --- | --- | --- | --- |
| TTKD-16 | 1712819 | Nguyễn Thế Tình | 1712819@student.hcmus.edu.vn | 0376846295 |
| 1712847 | Nguyễn Trí Trung | 1712847@student.hcmus.edu.vn | 0827609705 |
| 1712898 | Trần Việt Văn | 1712898@student.hcmus.edu.vn | 0376387012 |
| 1712929 | Nguyễn Phượng Vỹ | 1712929@student.hcmus.edu.vn | 0385092206 |

1. **Thông tin đồ án thực hành:**

Link đồ án: [**https://drive.google.com/drive/folders/1URCW59T1mn7bTv-Bo3WfW1eGF1VUG-NO**](https://drive.google.com/drive/folders/1URCW59T1mn7bTv-Bo3WfW1eGF1VUG-NO)

1. **Kết quả bài làm**

**Mô tả dữ liệu**

**Bảng 1. Thông tin các trường hợp tai nạn**

| **Accident Circumstances** | **Data type** | **Description** |
| --- | --- | --- |
| Accident Index | Varchar(50) | Số hiệu của vụ tai nạn |
| Accident Severity | int | Mức độ nghiêm trọng của vụ tai nạn  1 = đặc biệt nghiêm trọng  2 = nghiêm trọng  3 = nhẹ |
| Number of Vehicles | int | Số lượng phương tiện bị tai nạn |
| Number of Casualties | int | Số lượng người thương vong |
| Date (DD/MM/YYYY) | Varchar(50) | Ngày tháng năm xảy ra tai nạn |
| Day of Week | int | Ngày xảy ra tai nạn trong tuần:   | 1 | chủ nhật | | --- | --- | | 2 | thứ hai | | 3 | thứ ba | | 4 | thứ tư | | 5 | thứ 5 | | 6 | thứ 6 | | 7 | thứ 7 | |
| Time (HH:MM) | Varchar(50) | Thời gian diễn ra tai nạn |
| Local Authority (District) | int | Mã quận, thành phố nơi xảy ra vụ tai nạn |
| Local Authority (Highway Authority - ONS code) | Varchar(50) | Mã đường cao tốc nơi xảy ra vụ tai nạn |
| Road Type | int | Loại đường, cụ thể:   | 1 | Bùng binh | | --- | --- | | 2 | Đường một chiều | | 3 | Làn kép | | 6 | Đường đơn | | 7 | Đường trượt | | 9 | không xác định | | 12 | Đường một chiều / Đường trượt | | -1 | Dữ liệu bị thiếu hoặc nằm ngoài phạm vi | |
| Speed limit | int | Tốc độ cho phép tối đa trên đoạn đường xảy ra tai nạn |
| Road Surface Conditions | int | Điều kiện của mặt đường  1 - khô  2 - ẩm ướt  3 - có tuyết  4 - sương giá hoặc đóng băng  5 - ngập sâu trên 3cm  6 - dầu nhớt  7 - sình lầy  -1 - dữ liệu thiếu hoặc nằm ngoài phạm vi |
| Special Conditions at Site | int | Những điều kiện ảnh hưởng đến tai nạn khác  0 - không có  1 - tín hiệu giao thông bị mất  2 - phần tín hiệu tự động bị lỗi  3 - biển báo hoặc vạch kẻ đường bị hư hoặc bị che khuất  4 - đường đang thi công  5 - mặt đường hỏng  6 - đường trơn  7 - đường sình lầy  -1 - dữ liệu thiếu hoặc ngoài phạm vi |
| Carriageway Hazards | int | Các mối nguy hiểm khác  0 - không có  1 - phương tiện mắc kẹt  2 - vật thể lạ trên đường  3 - tai nạn trước  4 - chó chạy trên đường  5 - động vật khác trên đường  6 - người qua đường ( không bị thương vong)  7 - động vật qua đường (không bị thương vong)  -1 - dữ liệu thiếu hoặc ngoài phạm vi |
| Urban or Rural Area | int | Khu vực thành thị hay nông thôn:  1 - thành thị  2 - nông thôn  3 - chưa xác định |
| LSOA\_of\_Accident\_Location | Varchar(50) | Khu vực địa lý LSOA của địa điểm tại nạn |

**Bảng 2. Thông tin phương tiện giao thông**

| **Vehicle** | **Data type** | **Description** |
| --- | --- | --- |
| Accident\_Index | Varchar(50) | Mã số tai nạn. Dùng để phân biệt từng vụ tai nạn |
| Vehicle\_Type | Int | Cho biết loại xe nào được sử dụng |
| Journey Purpose of Driver | Int | Mục đích di chuyển của người lái xe 1 = công việc lái xe 2 = Đi làm đến / từ nơi làm việc 3 = Đưa học sinh đến / từ trường 4 = lái xe đến / từ trường 5 = Khác 6 = Không biết  15 = Khác / Không biết -1 = Dữ liệu bị thiếu hoặc nằm ngoài phạm vi |
| Age of Driver | Int | Tuổi của người lái xe -1 = Dữ liệu bị thiếu hoặc nằm ngoài phạm vi |
| Age of Vehicle (manufacture) | Int | Tuổi của xe: -1 = Dữ liệu bị thiếu hoặc nằm ngoài phạm vi |

**Bảng 3. Thông tin người thương vong trong các vụ tai nạn**

| **Casualty** | **Data type** | **Description** |
| --- | --- | --- |
| Accident Index | Varchar(50) | Mã số tai nạn. Dùng để phân biệt từng vụ tai nạn |
| Sex of Casualty | Int | Giới tính người thương vong 1 = Nam 2 = Nữ -1= dữ liệ bị mất hoặc ngoài phạm vi |
| Age of Casualty | Int | Tuổi người thương vong 0-120: tuổi hợp lệ -1 : tuổi không hợp lệ hoặc mất mát dữ liệu |
| Age Band of Casualty | Int | Thuộc nhóm tuổi |
| Casualty Severity | Int | Mức độ nghiêm trong 1 = Gây tử vong 2 = Nguy kịch 3 = Nhẹ |
| Casualty Type | Int | Loại thương vong tương ứng với loại phương tiện di chuyển 0 người đi bộ 1 = người đi xe đạp 2 = người lái hoặc chơi xe mô tô 50 cc trở lên 3 = người lái hoặc chở xe mô tô 125cc trở xuống 4 = người lái hoặc chở xe mô tô 125 đến 500 cc 5 = người lái hoặc chở xe mô tô trên 500 cc 8 = taxi hoặc xe thuê 9 = người ngồi trên ô tô 10 = xe bus 8-10 chỗ 11 = xe bus hoặc xe khách từ 17 chỗ trở lên 16 = người đi người 17 = xe công nông 18 = người lái xe điện 19 = vận chuyển hàng hóa,xe chở hàng dưới 3,5 tấn 20 = xe chở hàng 3,5 đến 7,5 tấn 21 = xe chở hàng trên 7,5 tấn 22 = người lái tay ga tự động 23 = người lái và hành khách mô tô điện 90 = người ngồi trên xe khác 97 = người lái và hành khách xe máy không xác định phân khối 98 = xe chở hàng không xác định phân khối |

**Bảng 4. Dữ liệu LSOA-Postcode mapping.**

| **Thuộc tính** | **Data type** | **Mô tả** |
| --- | --- | --- |
| PCD7 | Varchar(50) | Mã bưu chính của quận và phần đơn vị ngành |
| LSOA11CD | Varchar(50) | Mã khu vực siêu đầu ra lớp thấp hơn (Lower layer Super Output Area) |

**Bảng 5. Dữ liệu UK-Postcodes.**

| **Thuộc tính** | **Data type** | **Mô tả** |
| --- | --- | --- |
| postcode | Varchar(50) | Mã vùng |
| city | Varchar(50) | Thành phố |
| county | Varchar(50) | Quận |
| country\_name | Varchar(50) | Tên quốc gia |
| region\_name | Varchar(50) | Nằm ở vùng nào trên trên thế giới |

Từ các thuộc tính được tô màu phía trên, để có thể thuận tiện cho việc thiết kế Geo dim, ta có thể xử lí sơ bộ và gom nhóm.

Đối với England và xứ Wales, có thể tiến hành so sánh mã LSOA (thuộc tính cuối trong bảng accident) để tìm ra địa điểm. Vì bảng mã LSOA không có chứa thông tin về địa điểm cụ thể (thành phố, quận, …) nhưng có chứa trường pcd7 với 3 hoặc 4 kí tự đầu tiên tượng trưng cho mã postcode. Đối với những vùng thuộc Scotland và Northern Ireland (không có mã LSOA) vì dữ liệu không có null ở thuộc tính Local\_Authority (District) nên ta có thể sử dựng thuộc tính này, ghép phần description cho thuộc tính này vào chung với bảng postcode, ta được bảng dùng để tham chiếu địa điểm cho những tai nạn không có mã LSOA. Tổng hợp 2 bảng ta được bảng dành cho địa điểm của từng vụ tai nạn

**Bảng 6. Địa điểm của từng vụ tai nạn**

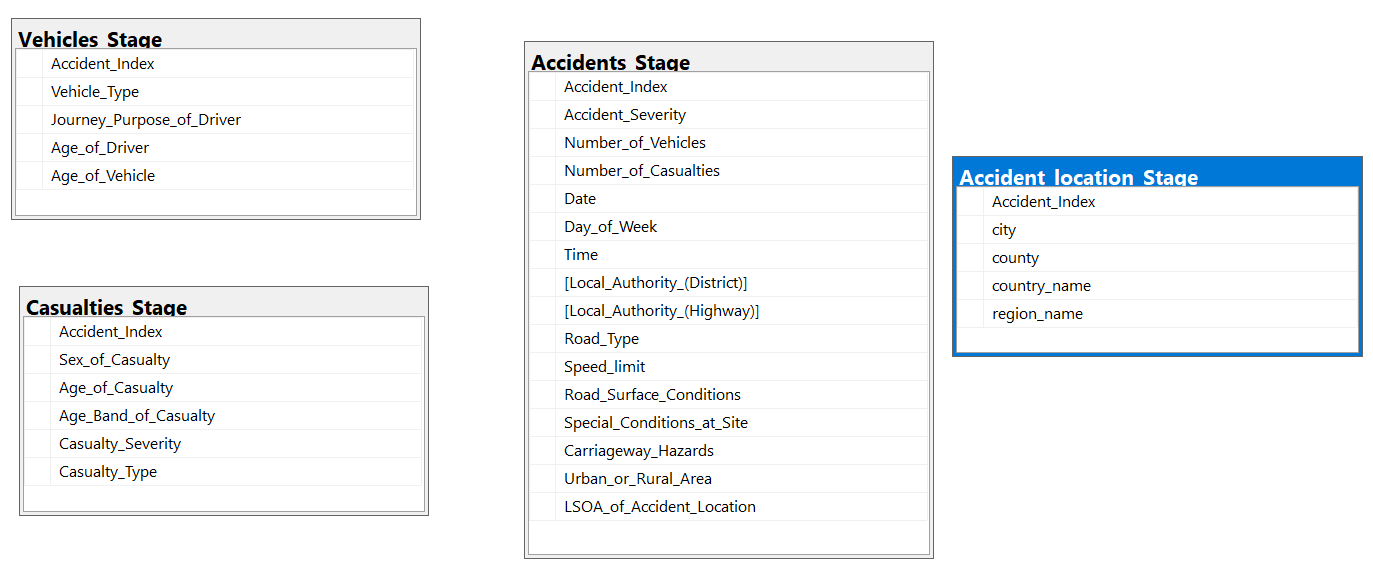
| **Column** | **Data type** | **Description** |
| --- | --- | --- |
| Accident\_Index | Varchar(50) | Mã vụ tai nạn |
| city | Varchar(50) | Thành phố |
| county | Varchar(50) | Quận |
| country\_name | Varchar(50) | Quốc gia |
| region\_name | Varchar(50) | Vùng |

**Mô tả quá trình nạp dữ liệu từ Source vào Stage**

**1. Tổng quát:**

Sau khi phân tích thì nhóm sẽ nạp 5 bảng dữ liệu chính đó là: Accidents\_Stage, Vehicles\_Stage, Casualties\_Stage, LDA\_to\_dim\_Scotland\_NIreland\_Stage và LSOA\_to\_dim\_Eng\_Wales\_Stage.

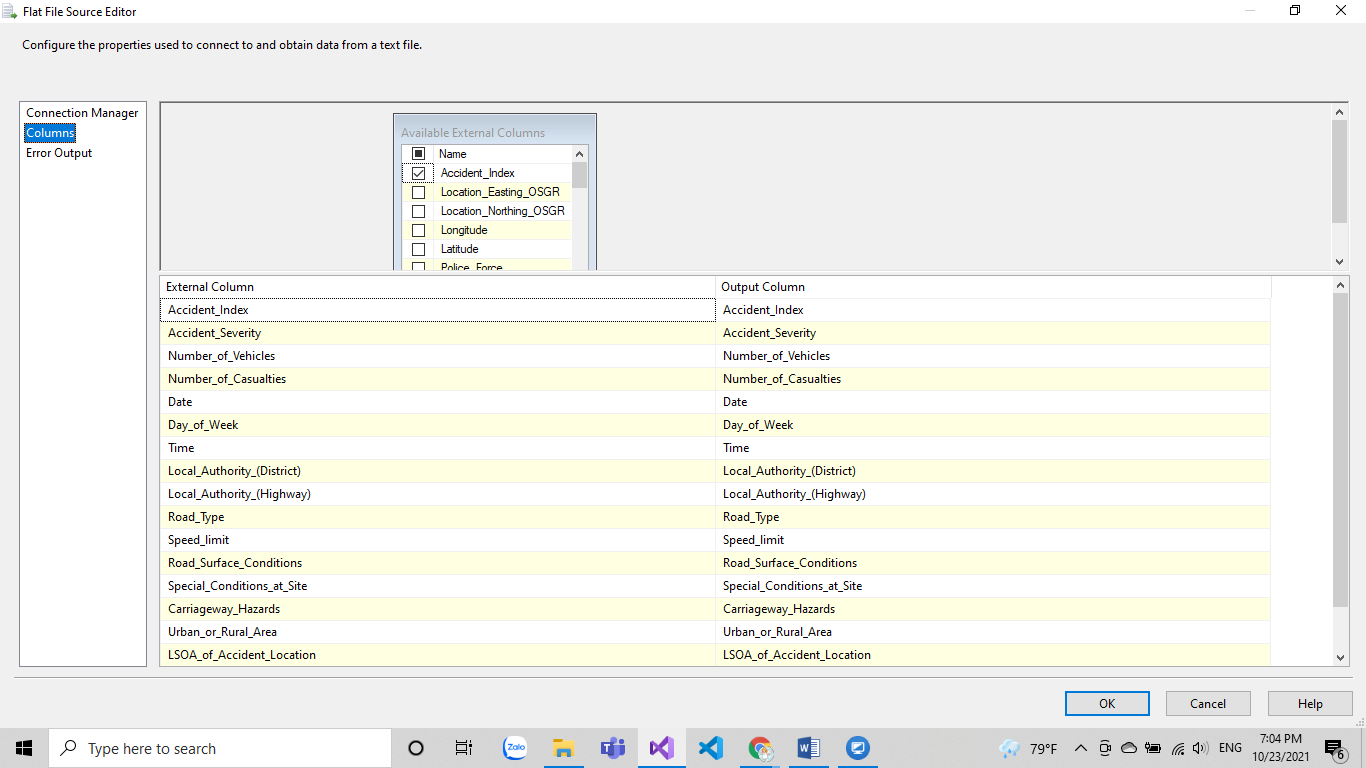
**2. Stage Design:**



**3. Mô tả:**

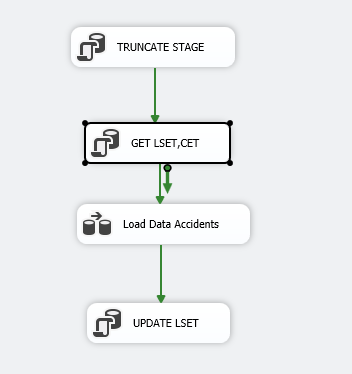
Đổ 5 bảng Accidents1114, Casualties1114, Vehicles1114, LSOA\_to\_dim\_Eng\_Wales và LDA\_to\_dim\_Scotland\_NIreland FlatFileToSource vào tương ứng với 5 bảng trong CSDL **UCA\_Stage:**

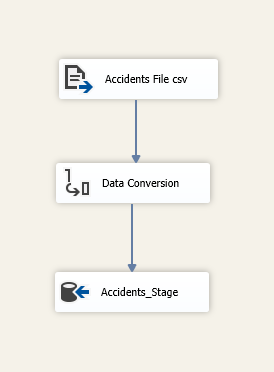
* **Bước 1**: Đầu tiên, trước khi đổ dữ liệu từ nguồn vào Stage, ta cần truncate Stage trước để chắc chắn dữ liệu đổ vào không bị lặp
* **Bước 2**: update cột CET = ngày giờ hiện tại và lấy dữ liệu LSET và CET trong bảng Data\_Flow trong CSDL UCA\_LSET\_CET để gán làm cột mốc cho các dữ liệu được đổ vào.
* **Bước 3:** Đổ dữ liệu từ nguồn vào Stage. Tại đây, ta chỉ chọn những dữ liệu cần nạp vào Stage.



* **Bước 4**: Cập nhật lại dữ liệu LSET = CET trong bảng Data\_Flow trong CSDL UCA\_LSET\_CET

Quy trình nạp:

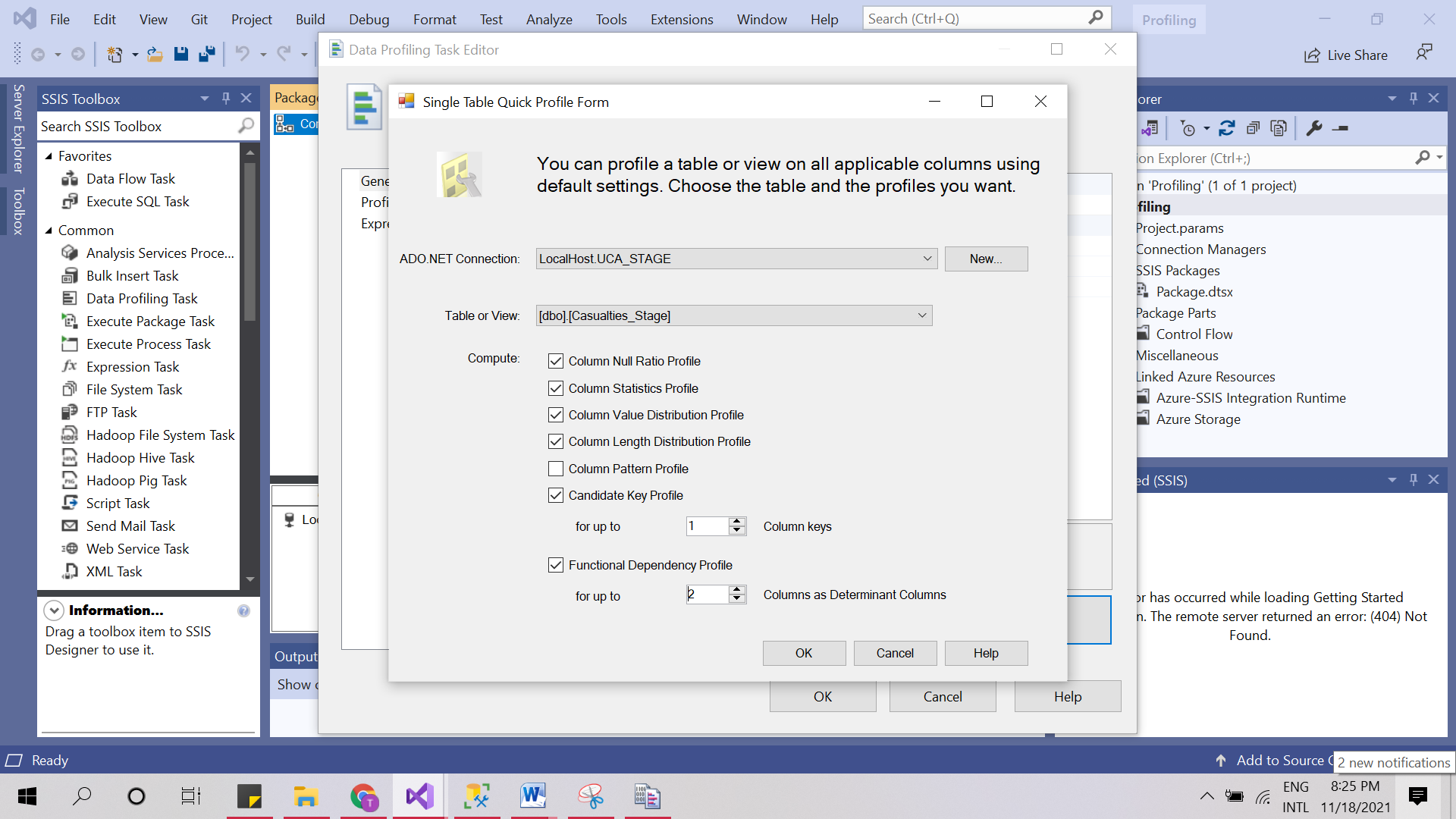




**Data Profiling quá trình nạp dữ liệu từ Source vào Stage**

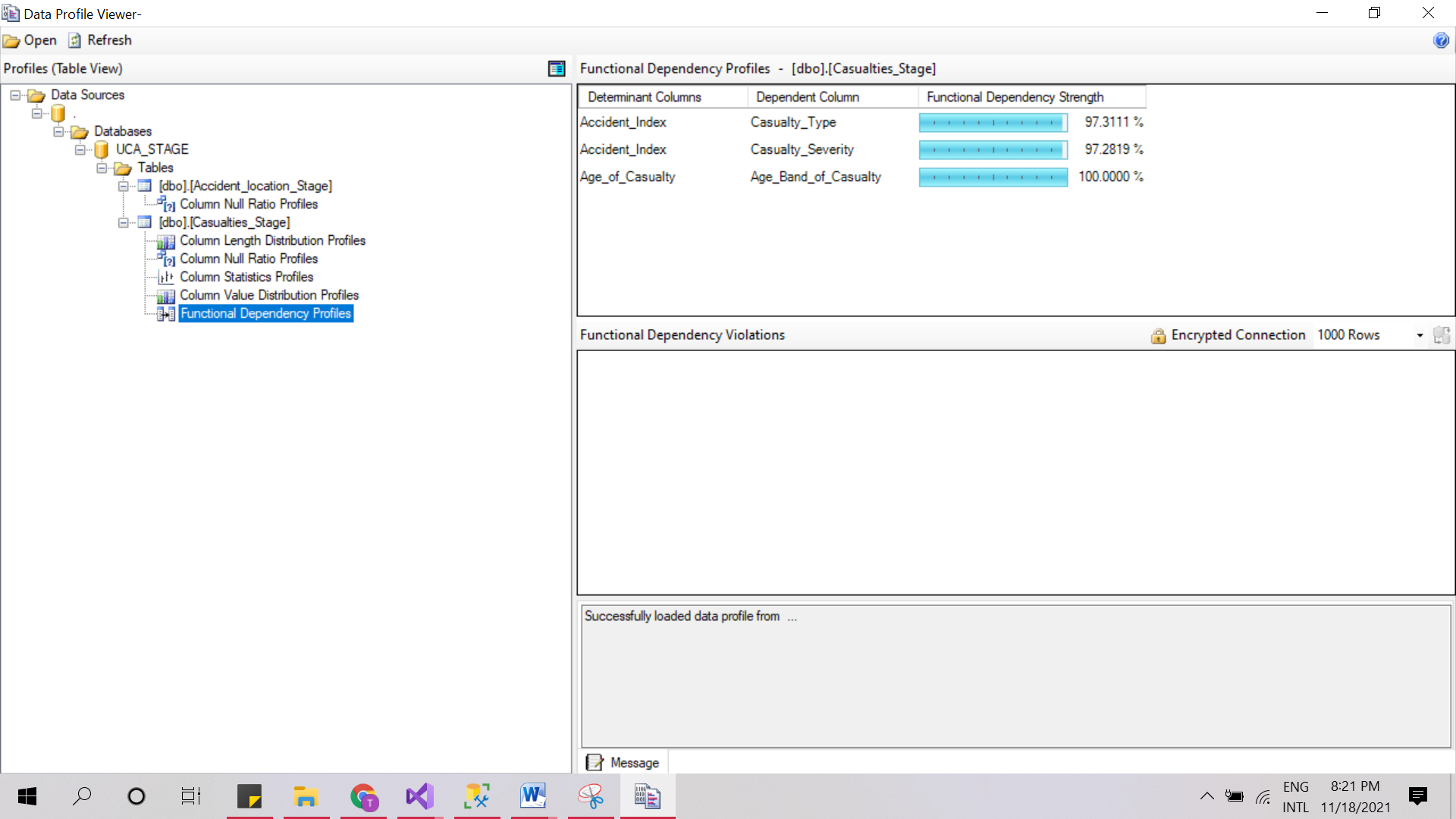
Để thiết kế lược đồ NDS:

**Bước 1:** Chạy Data Profiling task để kiểm tra sự phụ thuộc của dữ liệu



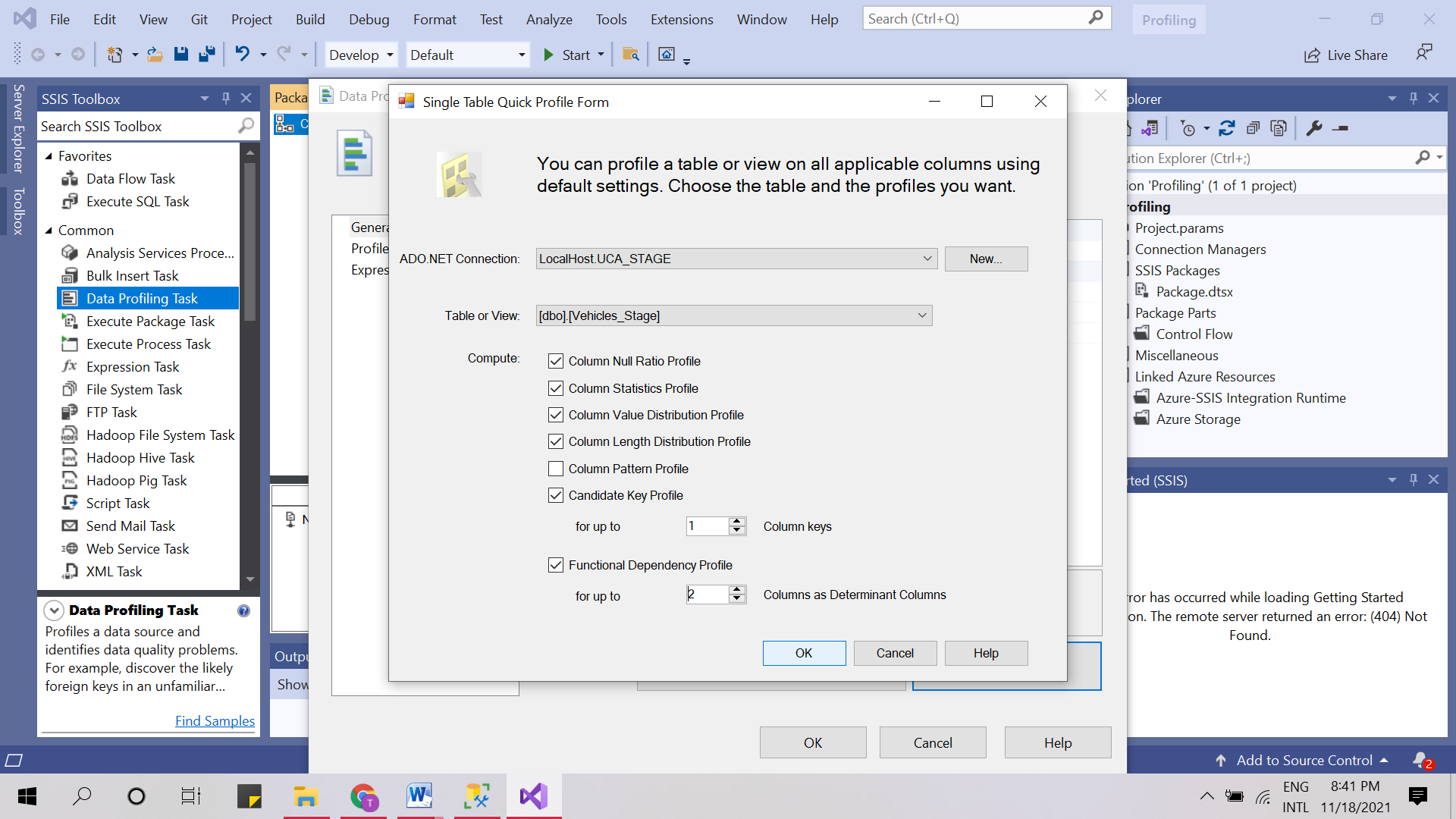
Ta chú ý ớ click vào Functional Dependency Profile và chọn số thuộc tính để xem sự phụ thuộc giữa chúng

**Bước 2**: Phân tích dữ liệu của Data Profiling Task

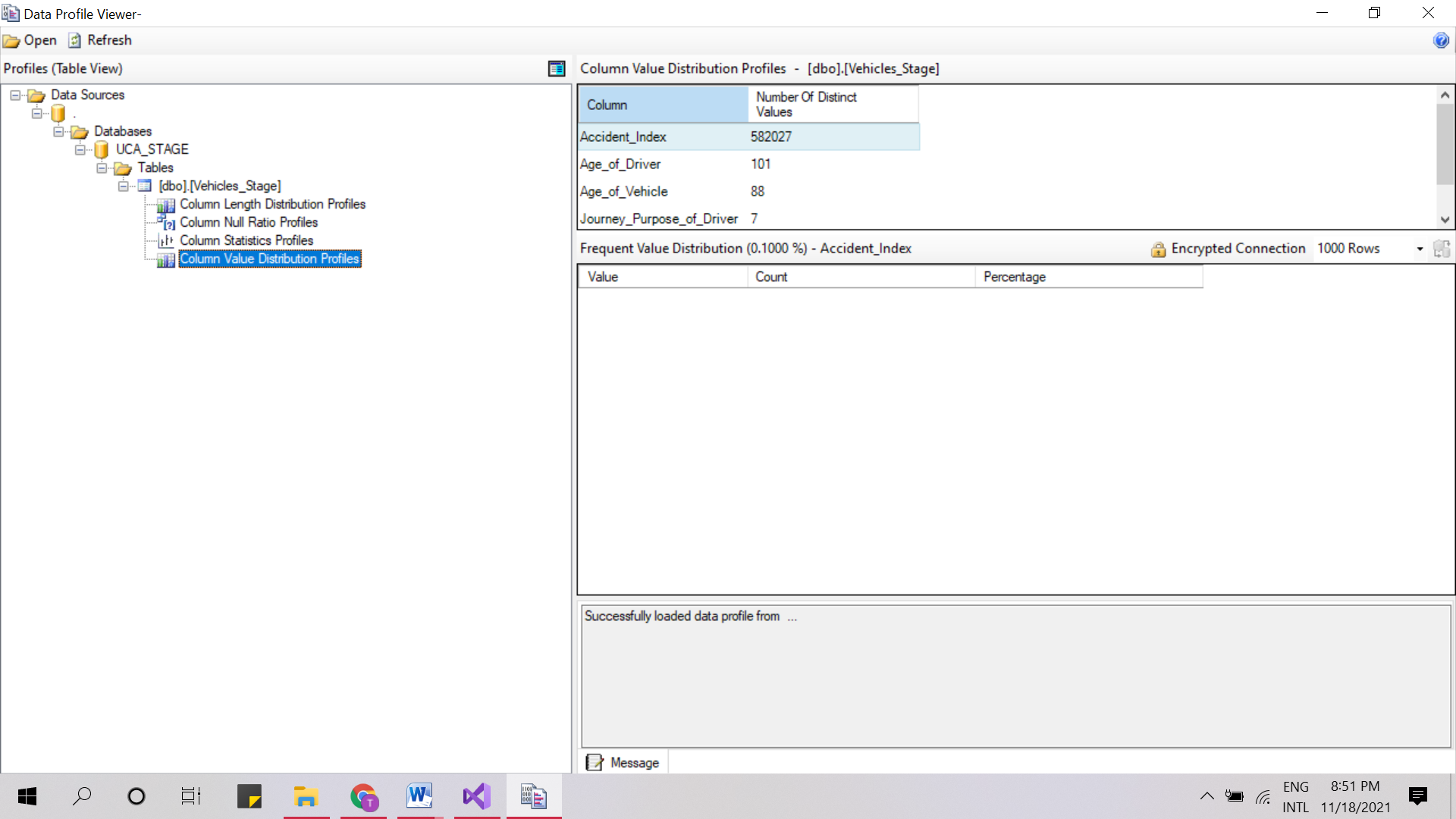


Ta thấy [Age\_of\_Casualty] sẽ suy ra được [Age\_Band\_of\_Casualty] với 100% sự phụ thuộc, còn 2 cái ở trên chỉ gần đúng nên nếu ta sử dụng sẽ có khi mất dữ liệu do không suy ra được. Nên từ đó ta sẽ đưa ra những giải pháp để có thể nâng dạng chuẩn để thiết kế NDS từ Stage

Làm tương tự với các bảng khác ta sẽ tìm được sự phụ thuộc và từ đó tách bảng để có lược đồ NDS

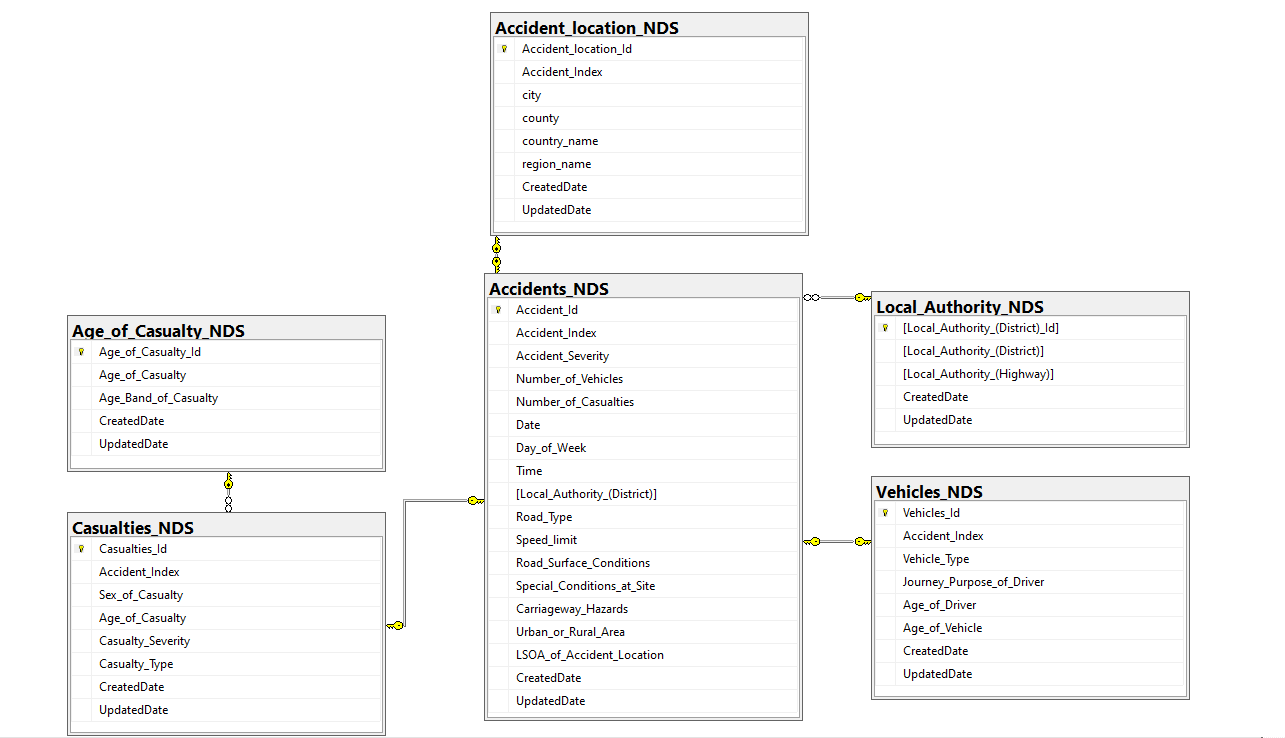


Ta thử làm phân tích với bảng Vehicles



Không có Functional Dependency Profiles nên ta không thể tách bảng được, hay nói các khách không có sự phụ thuộc vào các thuộc tính khác(ngoài khóa)

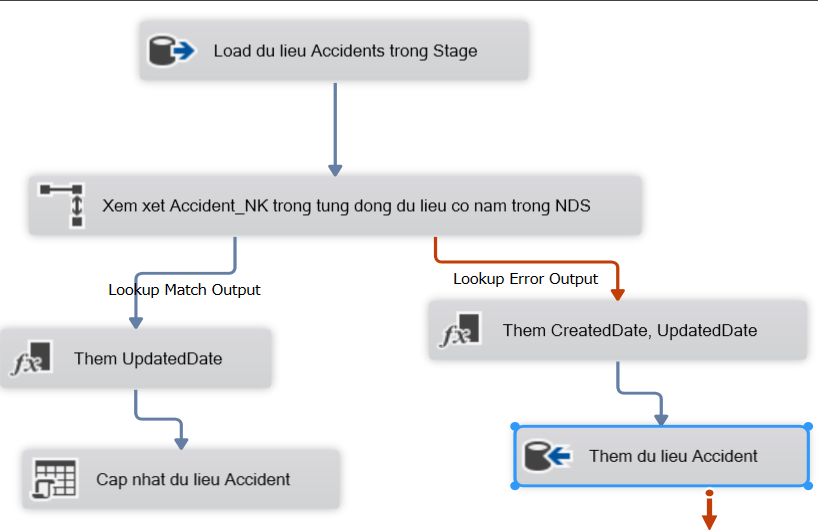
**NDS Design**



Quá trình nạp dữ liệu từ Stage vào NDS:

Đầu tiền ta phải xác định các bảng nào sẽ nạp trước hay nói các khác là thứ tự của nó. Tổng cộng ta có 6 bảng nên ta sẽ có 6 lần nạp dữ liệu vào NDS tương ứng của nó

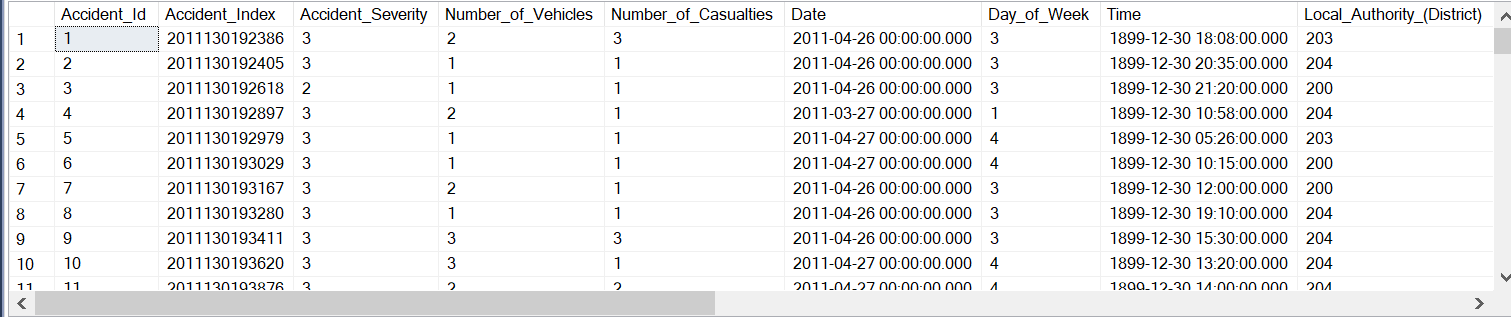
1. **Bảng Accidents\_Stage vào Accidents\_NDS Stage**



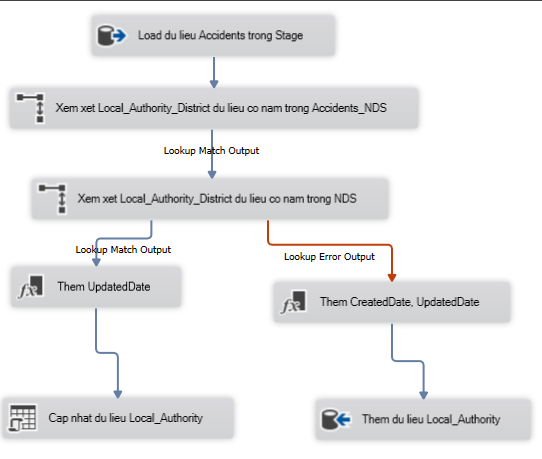
Ta sẽ load dữ liệu và kiểm tra xem Accident\_NK(Accident\_Index) có nằm trong NDS chưa

* Nếu có rồi thì chạy update bên trái
* Nếu chưa có thì sẽ chạy thêm dữ liệu bên phải

Kết quả sau khi chạy:



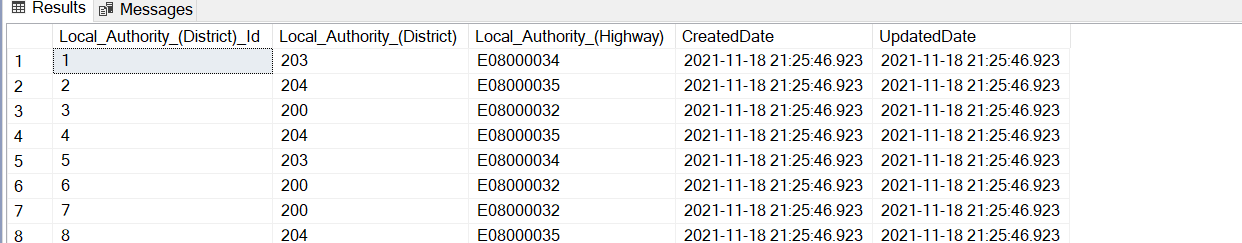
1. **Bảng Accidents\_Stage vào Local\_Authority\_NDS**



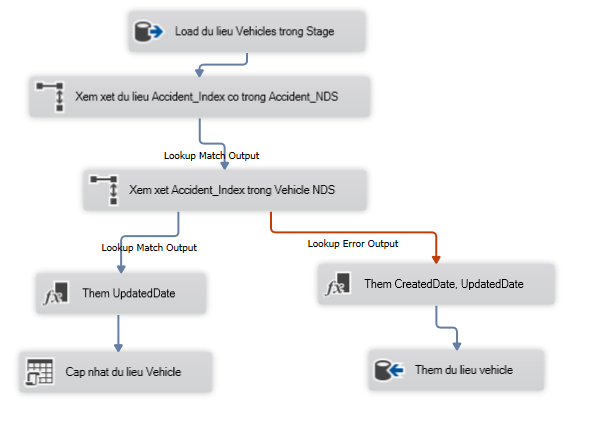
Ta sẽ load dữ liệu và kiểm tra xem [Local\_Authority\_(District)] có nằm trong Accidents\_NDS chưa và xem [Local\_Authority\_(District)] có nằm trong Local\_Authority\_NDS

* Nếu có rồi thì chạy update bên trái
* Nếu chưa có thì sẽ chạy thêm dữ liệu bên phải

Kết quả sau khi chạy:



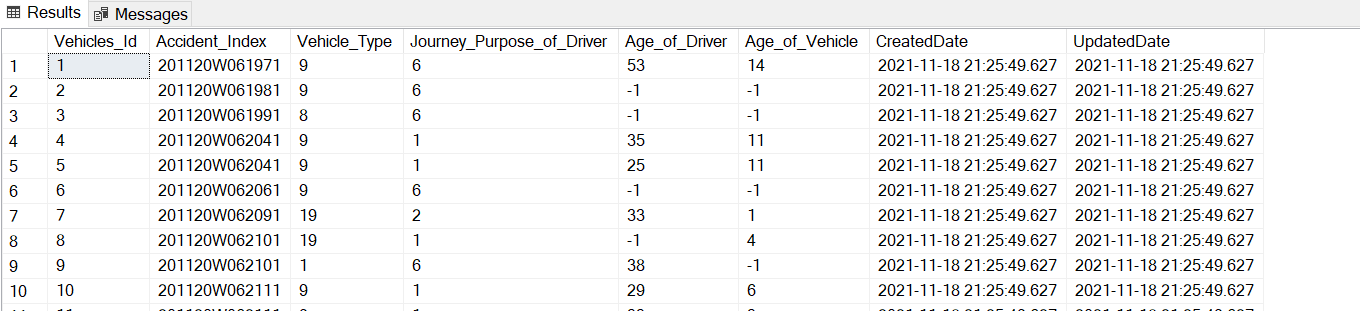
1. **Bảng Vehicles\_Stage vào Vehicles\_NDS**



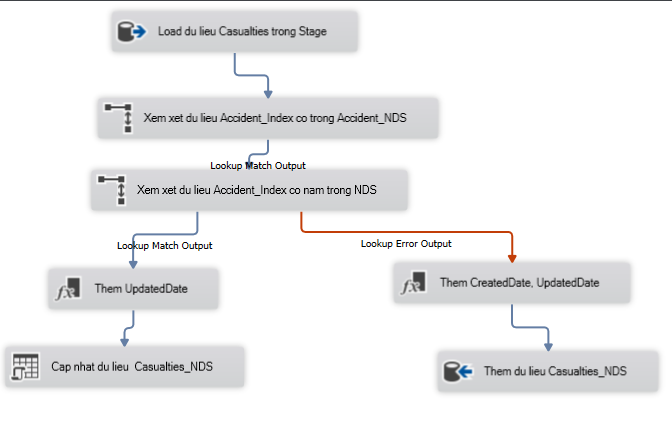
Ta sẽ load dữ liệu và kiểm tra xem [Accident\_Index] có nằm trong Accidents\_NDS chưa và xem [Accident\_Index] có nằm trong Vehicles\_NDS.

* Nếu có rồi thì chạy update bên trái
* Nếu chưa có thì sẽ chạy thêm dữ liệu bên phải

Kết quả sau khi chạy:



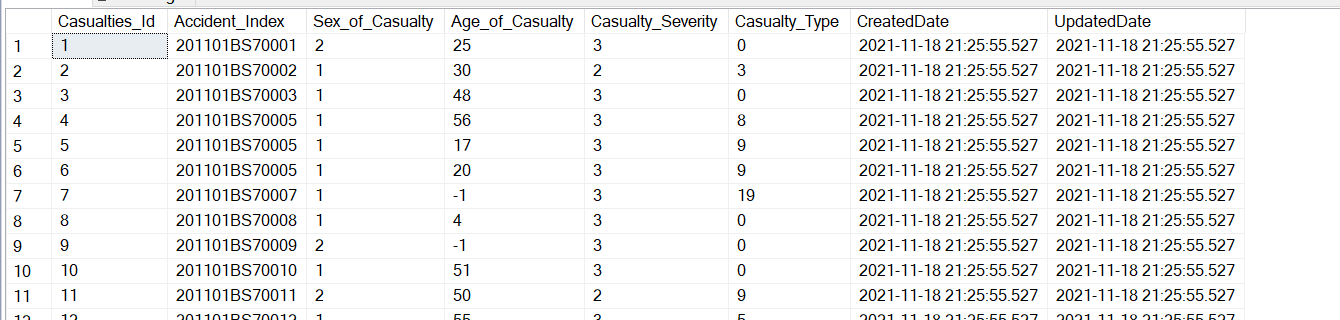
1. **Bảng Casualties\_Stage vào Casualties \_NDS**



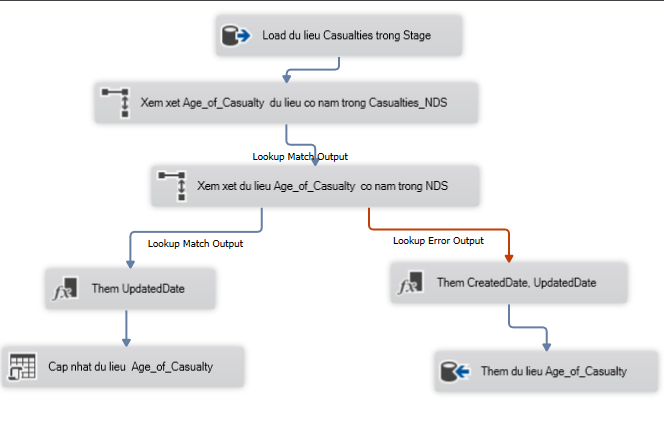
Ta sẽ load dữ liệu và kiểm tra xem [Accident\_Index] có nằm trong Accidents\_NDS chưa và xem [Accident\_Index] có nằm trong Vehicles\_NDS.

* Nếu có rồi thì chạy update bên trái
* Nếu chưa có thì sẽ chạy thêm dữ liệu bên phải

Kết quả sau khi chạy:



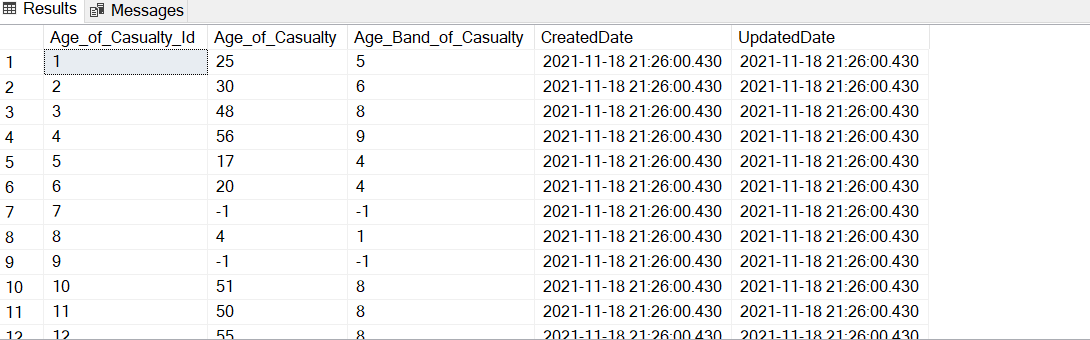
1. **Bảng Casualties\_Stage vào Age\_of\_Casualties\_NDS**



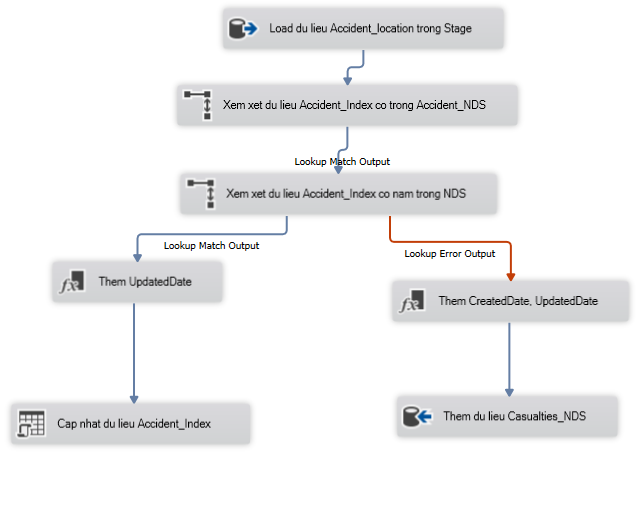
Ta sẽ load dữ liệu và kiểm tra xem [Age\_of\_Casualty] có nằm trong Casualties\_NDS chưa và xem [Age\_of\_Casualty] có nằm trong Age\_of\_Casualty\_NDS.

* Nếu có rồi thì chạy update bên trái
* Nếu chưa có thì sẽ chạy thêm dữ liệu bên phải

Kết quả sau khi chạy:



1. **Bảng Accident\_location\_Stage vào Accident\_location\_NDS**



Ta sẽ load dữ liệu và kiểm tra xem [Accident\_Index] có nằm trong Accidents\_NDS chưa và xem [Accident\_Index] có nằm trong Accident\_location\_NDS.

* Nếu có rồi thì chạy update bên trái
* Nếu chưa có thì sẽ chạy thêm dữ liệu bên phải

Kết quả sau khi chạy:



Sau đó ta nối các Data Flow Task chạy theo thứ tự:

