-

**Trường Đại học Khoa học Tự nhiên, ĐHQG – HCM**

**Mã nhóm:** TTKD-04

**Tên thành viên**:

* Lâm Xương Đức 18120163
* Phạm Đức Huy 18120399
* Lâm Ngọc Anh Khoa 18120414
* Ngô Nhật Tân 18120547

Tp. HCM, ngày 19 tháng 10 năm 2021

**HỆ THỐNG THÔNG TIN PHỤC VỤ TRÍ TUỆ KINH DOANH**

Báo cáo đồ án thực hành

**Khoa Công nghệ thông tin**

**MỤC LỤC**

[**I.** **Mô tả dữ liệu** 4](#_Toc92050555)

[**1.** **Dữ liệu UK Accident** 4](#_Toc92050556)

[**1.1.** **Accidents** 4](#_Toc92050557)

[**1.2.** **Casualties** 7](#_Toc92050558)

[**1.3.** **Vehicles** 8](#_Toc92050559)

[**2.** **Dữ liệu LSOA- Postcode Mapping** 10](#_Toc92050560)

[**3.** **Dữ liệu về UK Postcode** 11](#_Toc92050561)

[**3.1.** **Dữ liệu UK Postcodes** 11](#_Toc92050562)

[**3.2.** **Dữ liệu Wiki Postcode** 13](#_Toc92050563)

[**3.3.** **Dữ liệu Postcode districts** 13](#_Toc92050564)

[**II.** **Thiết kế cấu trúc cài đặt** 14](#_Toc92050565)

[**III.** **Thiết kế cấu trúc các thành phần** 14](#_Toc92050566)

[**1.** **Phân tích yêu cầu phân tích** 14](#_Toc92050567)

[**2.** **Thiết kế Stage** 16](#_Toc92050568)

[**3.** **Thiết kế NDS** 17](#_Toc92050569)

[**3.1.** **Bảng Vehicle** 17](#_Toc92050570)

[**3.2.** **Bảng Vehicle Type** 17](#_Toc92050571)

[**3.3.** **Bảng Journey Purpose** 18](#_Toc92050572)

[**3.4.** **Bảng Ageband** 18](#_Toc92050573)

[**3.5.** **Bảng Casual Type** 18](#_Toc92050574)

[**3.6.** **Bảng Accident** 19](#_Toc92050575)

[**3.7.** **Bảng Police Force** 20](#_Toc92050576)

[**3.8.** **Bảng Road Type** 20](#_Toc92050577)

[**3.9.** **Bảng Severity** 21](#_Toc92050578)

[**3.10.** **Bảng Source\_ID** 21](#_Toc92050579)

[**3.11.** **Bảng Local Authority Highway** 21](#_Toc92050580)

[**3.12.** **Bảng Casualties Class** 21](#_Toc92050581)

[**3.13.** **Bảng Casualties** 22](#_Toc92050582)

[**3.14.** **Bảng Region** 23](#_Toc92050583)

[**3.15.** **Bảng LSOA\_Mapping** 23](#_Toc92050584)

[**3.16.** **Bảng Country** 23](#_Toc92050585)

[**3.17.** **Bảng County** 24](#_Toc92050586)

[**3.18.** **Bảng Town** 24](#_Toc92050587)

[**3.19.** **Bảng PostCode** 25](#_Toc92050588)

[**4.** **Thiết kế DDS** 26](#_Toc92050589)

[**IV.** **Giai đoạn Staging** 27](#_Toc92050590)

[**1.** **Data flow** 27](#_Toc92050591)

[**2.** **Profiling dữ liệu vừa rút trích** 27](#_Toc92050592)

[**2.1.** **Bảng Accidents** 27](#_Toc92050593)

[**2.2.** **Bảng Casualties** 27](#_Toc92050594)

[**2.3.** **Bảng Vehicles** 27](#_Toc92050595)

[**2.4.** **Bảng LSOA Mapping** 27](#_Toc92050596)

[**2.5.** **Bảng Postcode** 27](#_Toc92050597)

[**2.6.** **Bảng Wiki Postcode** 28](#_Toc92050598)

[**2.7.** **Bảng Postcode District** 28](#_Toc92050599)

[**V.** **Đổ dữ liệu từ Stage vào NDS** 28](#_Toc92050600)

[**VI.** **Hướng dẫn cài đặt** 28](#_Toc92050601)

[**VII.** **Phân công** 28](#_Toc92050602)

[**VIII.** **Tham khảo** 28](#_Toc92050603)

[**1.** **Nguồn của các tập dữ liệu bên ngoài** 28](#_Toc92050604)

[**2.** **Nguồn tài liệu hỗ trợ nghiên cứu** 29](#_Toc92050605)

* 1. **Mô tả dữ liệu**

1. **Dữ liệu UK Accident**
   1. **Accidents**

Chứa dữ liệu về các vụ tai nạn được trình báo với cảnh sát Anh từ năm 2005 đến năm 2015

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên cột** | **Kiểu giá trị** | **Ý nghĩa** | **Ghi chú** |
| Accident\_Index | Chuỗi | Mã định danh vụ tai nạn |  |
| Location\_Easting | Số | Độ lệch về phía đông | Thuộc tính của hệ quy chiếu lưới. Tượng trưng cho hoành độ của một điểm |
| Location\_Northing | Số | Độ lệch về phía bắc | Thuộc tính của hệ quy chiếu lưới. Tượng trưng cho tung độ của một điểm |
| Longtitude | Số | Kinh độ |  |
| Latitude | Số | Vĩ độ |  |
| Police\_Force | Số | Mã sở cảnh sát | 1: Metropolitan Police  3: Cumbria  4: Lancashire  …  Xem thêm trong bảng Police\_Force.csv trong thư mục contextCSVs |
| Accident\_Severity | Số | Mức độ nghiêm trọng | 1: Chết người  2: Nghiêm trọng  3: Thương nhẹ |
| Number\_of\_Vehicles | Số | Lượng phương tiện bị ảnh hưởng |  |
| Number\_of\_Casualties | Số | Số người bị thương |  |
| Date | Chuỗi | Ngày xảy ra vụ việc |  |
| Day\_of\_Week | Số | Thứ trong tuần |  |
| Time | DateTime | Giờ xảy ra |  |
| Local\_Authority\_(District) | Số | Mã quận | Quận nơi xảy ra tai nạn |
| Local\_Authority\_(Highway) | Chuỗi | Mã tổ chức giao thông vận tải | Mã tổ chức giao thông UK nơi xảy ra tai nạn  [Highway authority - Wikipedia](https://en.wikipedia.org/wiki/Highway_authority) |
| 1st\_Road\_Class | Số | Phân loại đường thứ nhất | 0: Không ở gần nút giao  1: Xa lộ 2: A(M)  …  Xem thêm trong bảng Road\_Class.csv |
| 1st\_Road\_Number | Số | Số đường | Dùng kết hợp với Road Class  Ví dụ: A 2314, M 4563  [Great Britain road numbering scheme - Wikipedia](https://en.wikipedia.org/wiki/Great_Britain_road_numbering_scheme) |
| Road\_Type | Số | Loại đường | 1: Vòng xuyến  2: Đường một chiều  3: Xa lộ có dải phân cách  …  Xem thêm trong bảng Road\_Type.csv |
| Speed\_limit | Số | Tốc độ tối đa cho phép |  |
| Junction\_Detail | Số | Đặc điểm nút giao | 0: Không có trong phạm vi 20m  1: Vòng xuyến  2: Vòng xuyến mini  …  Xem thêm trong bảng Junction\_Detail.csv |
| Junction\_Control | Số | Người điều khiển nút giao | 0: Không có trong phạm vi 20m 1: Người có thẩm quyền  2: Đèn tín hiệu  …  Xem thêm trong bảng Junction\_Control.csv |
| 2nd\_Road\_Class | Số | Phân loại đường thứ Hai | 0: Không có trong phạm vi 20m  1: Xa lộ 2: A(M)  …  Xem thêm trong bảng Road\_Class.csv |
| 2nd\_Road\_Number | Số | Số đường | Dùng kết hợp với Road\_Class  Ví dụ: A 2314, M 4563 |
| Pedestrian\_Crossing-Human\_Control | Số | Người điều khiển người đi bộ | 0: Không có trong phạm vi 50m 1: Điều khiển bởi người được phân công  2: Điều khiển bởi người có thẩm quyền khác  …  Xem thêm trong bảng Ped\_Cross\_Human.csv |
| Pedestrian\_Crossing-Physical\_Facilities | Số | Cơ sở hạ tầng dành cho người đi bộ | 0: Không có trong phạm vi 50m  1: Vạch kẻ đường dành cho người đi bộ  4: Đèn báo hiệu người đi bộ  …  Xem thêm trong Ped\_Cross\_Physical.csv |
| Light\_Conditions | Số | Điều kiện ánh sáng | 1: Ban ngày 2: Buổi tối – có đèn đường  3: Buổi tối – không có đèn đường  …  Xem thêm trong Light\_Conditions.csv |
| Weather\_Conditions | Số | Điều kiện thời tiết | Không có bảng tra cứu cụ thể |
| Road\_Surface\_Conditions | Số | Trạng thái mặt đường | Không có bảng tra cứu cụ thể |
| Special\_Conditions\_at\_Site | Số | Các điều kiện đặc biệt ở hiện trường | Không có bảng tra cứu cụ thể |
| Carriageway\_Hazards | Số | Các mối nguy khác trên đường | Không có bảng tra cứu cụ thể |
| Urban\_or\_Rural\_Area | Số | Thành thị hoặc nông thôn | 1: Thành thị  2: Nông thôn |
| Did\_Police\_Officer\_Attend  \_Scene\_of\_Accident | Số | Cảnh sát có đến hiện trường sau tai nạn | 1: Có  2: Không |
| LSOA\_of\_Accident\_Location | Chuỗi | Mã LSOA của nơi xảy ra tai nạn | [Lower Layer Super Output Area (datadictionary.nhs.uk)](https://datadictionary.nhs.uk/nhs_business_definitions/lower_layer_super_output_area.html) |

* **Nhận xét**
* Có 4 cột không có bảng tra cứu ý nghĩa kèm theo.
* Một số bảng tra cứu ý nghĩa có số đánh không theo thứ tự (Ped\_Cross\_Physical.csv, Police\_Force.csv).
  1. **Casualties**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên cột** | **Kiểu giá trị** | **Ý nghĩa** | **Ghi chú** |
| Accident\_Index | Chuỗi | Mã định danh vụ tai nạn |  |
| Vehicle\_Reference | Số |  |  |
| Casualty\_Reference | Số |  |  |
| Casualty\_Class | Số | Phân loại người bị thương | 1: Người điều khiển  2: Hành khách  3: Người đi bộ |
| Sex\_of\_Casualty | Số | Giới tính người bị thương | 1: Nam  2: Nữ,  3: Không xác định được  -1: Dữ liệu bị mất hoặc ngoài miền giá trị |
| Age\_of\_Casualty | Số | Tuổi người bị thương |  |
| Age\_Band\_of\_Casualty | Số | Độ tuổi người bị thương | -1: Không biết  1: 0-5  2: 6-10  …  Xem thêm trong bảng Age\_Band.csv |
| Casualty\_Severity | Số | Mức độ nghiêm trọng | 1: Chết người  2: Nghiêm trọng  3: Thương nhẹ |
| Pedestrian\_Location | Số | Vị trí người đi bộ | 0: Không phải người đi bộ  1: Băng qua đường trên vạch kẻ dành cho người đi bộ  …  Xem thêm trong bảng Ped\_Location.csv |
| Pedestrian\_Movement | Số | Hành động của người đi bộ | 0: Không phải người đi bộ  1: Băng qua đường từ góc chuẩn của người lái  …  Xem thêm trong bảng Ped\_Movement.csv |
| Car\_Passenger | Số | Vị trí người trên xe ô tô |  |
| Bus\_or\_Coach\_Passenger | Số | Vị trí người trên xe buýt |  |
| Pedestrian\_Road\_Maintenance\_Worker | Số |  |  |
| Casualty\_Type | Số | Loại thương vong | 0: Người đi bộ  1: Người đi xe đạp  2: Người lái hoặc khách xe mô tô dưới 50 phân khối  …  Xem thêm trong bảng Casualty\_Type.csv |
| Casualty\_Home\_Area\_Type | Số | Loại khu vực dân cư nơi xảy ra tai nạn |  |

* 1. **Vehicles**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên cột** | **Kiểu giá trị** | **Ý nghĩa** | **Ghi chú** |
| Accident\_Index | Chuỗi | Mã định danh vụ tai nạn |  |
| Vehicle\_Reference | Số |  |  |
| Vehicle\_Type | Số | Loại phương tiện | 1: Xe đạp  2: Mô tô 50cc trở xuống  3: Mô tô 125cc trở xuống  …  Xem thêm trong bảng Vehicle\_Type.csv |
| Towing\_and\_Articulation | Số | Xe có móc/kéo xe khác | 0: Không 1: Có |
| Vehicle\_Manoeuvre | Số | Hành động của phương tiẹn | 1: Lùi 2: Đang đỗ  3: Dừng  …  Xem thêm trong bảng Vehicle\_Manoeuvre.csv |
| Vehicle\_Location-Restricted\_Lane | Số | Phương tiện có đang đi trên làn bị cấm | 1: Có  2: Không |
| Junction\_Location | Số | Vị trí của phương tiện so với nút giao gần nhất | 0: Không có nút giao trong phạm vị 20m  1: Đang tiến lại gần hoặc đang đỗ ở nút giao  …  Xem thêm trong bảng Junction\_Location.csv |
| Skidding\_and\_Overturning | Số | Quãng đường phương tiện trượt khi gặp tai nạn? |  |
| Hit\_Object\_in\_Carriageway | Số | Số vật va chạm khi gặp tai nạn? |  |
| Vehicle\_Leaving\_Carriageway | Số | Số xe đang rời khỏi đường đi |  |
| Hit\_Object\_off\_Carriageway | Số | Số vật bị văng khỏi đường đi |  |
| 1st\_Point\_of\_Impact | Số | Điểm va chạm đầu tiên |  |
| Was\_Vehicle\_Left\_Hand\_Drive? | Số | Xe đặt tay lái trái hay phải |  |
| Journey\_Purpose\_of\_Driver | Số | Mục đích chuyến đi | 1: Đi làm  2: Đi đến/Rời khỏi nơi làm việc  3: Đưa con đi học/về nhà  …  Xem thêm trong bảng Journey\_Purpose.csv |
| Sex\_of\_Driver | Số | Giới tính người lái | 1: Nam  2: Nữ,  3: Không xác định được  -1: Dữ liệu bị mất hoặc ngoài miền giá trị |
| Age\_of\_Driver | Số | Tuổi người lái |  |
| Age\_Band\_of\_Driver | Số | Độ tuổi người lái | -1: Không biết  1: 0-5  2: 6-10  …  Xem thêm trong bảng Age\_Band.csv |
| Engine\_Capacity\_(CC) | Số | Phân khối xe (đơn vị cc) |  |
| Propulsion\_Code | Số |  |  |
| Age\_of\_Vehicle | Số | Tuổi xe |  |
| Driver\_IMD\_Decile | Số | Chỉ số phân bổ khu vực nơi người lái sống | Thước đo độ nghèo khó của 1 vùng |
| Driver\_Home\_Area\_Type | Số | Loại khu vực dân cư nơi người lái sống |  |

1. **Dữ liệu LSOA- Postcode Mapping**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên cột** | **Kiểu giá trị** | **Ý nghĩa** | **Ghi chú** |
| pcd7 | chuỗi | Phiên bản 7. Của mã postcode | Mã bưu điện 7 ký tự |
| pcd8 | chuỗi | Phiên bản 8. Của mã postcode | Mã bưu điện 8 ký tự |
| pcds | chuỗi | Phiên bản S | Mã bưu điện có khoảng cách giữa 2 phần outward và inward |
| dointr | Số | Ngày mã bưu điện được giới thiệu |  |
| doterm | Số | Ngày mã bưu điện bị chấm dứt | (ví dụ: 199304)  Điều này sẽ chỉ áp dụng cho các mã bưu điện đã bị hủy và không được sử dụng lại. |
| usertype | bool | Loại Người dùng | 0 = người dùng nhỏ  1 = người dùng lớn. |
| oa11cd | chuỗi | Mã Code | Mã 9 ký tự oa được định nghĩa vào năm 2011 |
| lsoa11cd | chuỗi | Mã Code |  |
| msoa11cd | chuỗi | Mã Code |  |
| ladcd | chuỗi | Mã Code |  |
| lsoa11nm | chuỗi | Mã Code |  |
| msoa11nm | chuỗi | Mã Code |  |
| ladnm | chuỗi | Tên thủ đô, quận, địa điểm cụ thể(như trường học, công ty, …) |  |
| ladnmw | chuỗi | Tên thủ đô, quận, địa điểm cụ thể(như trường học, công ty, …) |  |

1. **Dữ liệu về UK Postcode**
   1. **Dữ liệu UK Postcodes**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên cột** | **Kiểu giá trị** | **Ý nghĩa** | **Ghi chú** |
| Postcode | Chuỗi | Mã bưu chính |  |
| Easting | Số | Độ lệch về phía đông | Thuộc tính của hệ quy chiếu lưới. Tượng trưng cho hoành độ của một điểm |
| Northing | Số | Độ lệch về phía bắc | Thuộc tính của hệ quy chiếu lưới. Tượng trưng cho tung độ của một điểm |
| Latitude | Số | Vĩ độ |  |
| Longtitude | Số | Kinh độ |  |
| City | Chuỗi | Tên thành phố |  |
| County | Chuỗi | Tên quận, hạt |  |
| Country\_code | Chuỗi | Mã quốc gia |  |
| Country\_name | Chuỗi | Tên quốc gia |  |
| ISO3166-2 | Chuỗi | Mã ISO 3166-2 | Dùng để định danh các vùng tại UK   |  |  |  | | --- | --- | --- | | Code | Tên vùng | Loại vùng | | GB - ENG | England | Thành phố | | GB-NIR | Nothern Ireland | Tỉnh | | GB-SCT | Scotland | Thành phố | | GB-WLS | Wales | Thành phố | | GG | Guernsey |  | | IM | Isle of Man |  | | JE | Jersey |  |   Tham khảo: <https://en.wikipedia.org/wiki/ISO_3166-2:GB> |
| Region\_code | Chuỗi | Mã vùng |  |
| Region\_name | Chuỗi | Tên vùng |  |

* 1. **Dữ liệu Wiki Postcode**

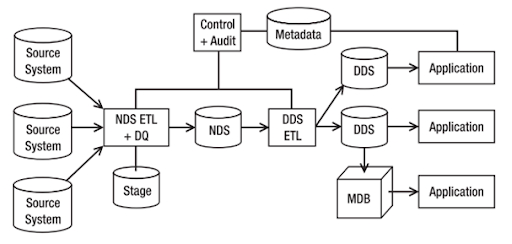
|  |  |  |  |
| --- | --- | --- | --- |
| **Tên cột** | **Kiểu giá trị** | **Ý nghĩa** | **Ghi chú** |
| Postcode area | Chuỗi | Mã khu vực |  |
| Postcode district | Chuỗi | Mã outward postcode |  |
| Postown | Chuỗi | Tên thị trấn |  |
| Former postal county | Chuỗi | Tên county |  |

* 1. **Dữ liệu Postcode districts**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên cột** | **Kiểu giá trị** | **Ý nghĩa** | **Ghi chú** |
| Postcode | Chuỗi | Mã bưu chính |  |
| Easting | Số | Độ lệch về phía đông | Thuộc tính của hệ quy chiếu lưới. Tượng trưng cho hoành độ của một điểm |
| Northing | Số | Độ lệch về phía bắc | Thuộc tính của hệ quy chiếu lưới. Tượng trưng cho tung độ của một điểm |
| Latitude | Số | Vĩ độ |  |
| Longtitude | Số | Kinh độ |  |
| Grid reference | Chuỗi |  |  |
| Town/Area | Chuỗi | Tên thị trấn |  |
| Region | Chuỗi | Tên vùng | Borough |
| Postcodes | Số | Postcodes |  |
| Active postcodes | Số | Postcodes |  |
| Population | Số | Dân số |  |
| Household | Số | Hộ gia đình |  |
| Nearby district | Chuỗi | Các quận lân cận |  |
| UK Region | Chuỗi | Thuộc vùng nào của nước Anh |  |

* 1. **Thiết kế cấu trúc cài đặt**

Với yêu cầu phân tích của bài toán, ta lựa chọn kiến trúc cài đặt NDS + DDS, cụ thể như sau:



*Mô hình kiến trúc cài đặt*

Trong đó dữ liệu từ các nguồn sẽ được lấy từ nguồn và đổ vào stage trước khi được đổ vào NDS. Phương pháp rút trích từ nguồn vào Stage sẽ được thực hiện theo hướng tiếp cận Incremental.

Sau khi dữ liệu từ các nguồn đã được đổ vào stage, ta sẽ tiến hành Profiling dữ liệu để kiểm tra tình trạng của dữ liệu trước được đổ vào NDS. Trong quá trình profiling ta cũng sẽ thực hiện các công đoạn làm sạch hoặc biến đổi dữ liệu nếu cần thiết.

NDS là một cơ sở dữ liệu đã được chuẩn hóa dùng để lưu trữ tất cả dữ liệu. Tùy thuộc vào các nhu cầu phân tích ta thực hiện rút trích dữ liệu từ NDS thành các DDS, Data mart để sử dụng.

* 1. **Thiết kế cấu trúc các thành phần**

1. **Phân tích yêu cầu phân tích**

* Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa phương trong tất cả các năm. Đối với yêu cầu này, ta cần các dữ liệu:
* Accident\_Severity, Local\_Authority từ bảng Accidents
* Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa Phương theo các Quý trong từng năm. Đối với yêu cầu này, ta cần các dữ liệu:
* Accident\_Severity, Local\_Authority, Date
* Thống kê số lượng người tử vong theo Giới Tính, Loại Nạn Nhân và Nhóm Tuổi theo các năm. Đối với yêu cầu này, ta cần các dữ liệu:
* Date, Sex\_of\_Casualty, Casualty\_Type, Age\_Band\_of\_Casualty
* Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng và Thời Điểm Trong Ngày trong các năm. Đối với yêu cầu này, ta cần các dữ liệu:
* Accident\_Severity, Date, Time
* Thống kê số lượng TNGT theo Mức Độ Nghiêm Trọng, Vùng và Kiểu Đường trong các năm. Đối với yêu cầu này, ta cần các dữ liệu:
* Accident\_Severity, Date, Urban\_or\_Rural\_Area, Road\_Type
* Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng, Loại Nạn Nhân và Độ Tuổi trong các năm.Đối với yêu cầu này, ta cần các dữ liệu:
* Accident\_Severity, Date, Casualty\_Type, Age\_of\_Casualty
* Tổng hợp số lượng tai nạn theo Mục Đích Hành Trình và Loại Phương Tiện. Đối với yêu cầu này, ta cần các dữ liệu:
* Journey\_Purpose, Vehicle\_Type
* Thống kê số lượng tai nạn theo Mức Độ Nghiêm Trọng, Loại Phương Tiện, Built-up Road trong các năm. Đối với yêu cầu này, ta cần các dữ liệu:
  + Accident\_Severity, Date, Built-up\_Road, Vehicle\_Type

1. **Thiết kế DDS**
   1. **Thiết kế Stage**

Do tính chất của việc rút trích từ nguồn sang Stage phải được thực hiện một cách nhanh chóng và không gián đoạn cơ sở dữ liệu nguồn quá lâu nên cấu trúc Stage sẽ tương đồng với cơ sở dữ liệu nguồn và được lọc lại các cột có giá trị sử dụng



*Cấu trúc các bảng trong Stage*

* 1. **Giai đoạn Staging**

1. **Data flow**

Sau khi đã xác định được những dữ liệu cần thiết, ta tiến hành các bước để đổ dữ liệu từ nguồn vào Stage.

* Đầu tiên, dữ liệu từ các file csv vào từng database riêng, đồng thời tạo nên 2 cột Create\_time và Update\_time để theo dõi thời điểm dữ liệu được tạo/cập nhật.
* Tiếp theo ta load dữ liệu từ các database trung gian qua Stage:
  + Ở Metadata, ghi lại thời điểm hiện tại CET khi bắt đầu ETL dữ liệu từ nguồn vào Stage
  + Ở Metadata, lấy ra giá trị LSET của bảng đang chuẩn bị nạp vào stage
  + Ở nguồn ta sẽ chọn ra những dòng dữ liệu được tạo ra hoặc được cập nhật trong khoảng từ LSET đến CET để đổ vào Stage
  + Dọn dẹp dữ liệu cũ trong stage
  + Tiến hành load dữ liệu mới vào Stage
  + Profiling dữ liệu mới vừa được load vào
  + Cập nhật LSET bằng giá trị CET hiện tại vào Metadata

1. **Profiling dữ liệu vừa rút trích**
   1. **Bảng Accidents**

* Dữ liệu missing data khoảng 6%
* Khóa chính rơi vào cột accident\_Index
  1. **Bảng Casualties**
* Khóa chính rơi vào accident\_Index
* Dữ liệu không có missing data
* Phân phối giá trị
* Age\_of\_Casualty: 19 18 là nhiều
* Giới tính 1 (nam 58%)
  1. **Bảng Vehicles**
* Age\_of \_Driver: nhiều nhất có giá trị là 30
  1. **Bảng LSOA Mapping**
* Dữ liệu không có missing data
* Đối với cột pcds thì cột này được đánh giá có khả năng làm khóa chính cho bảng này
* Về chiều dài ký tự:
  + Cột pcds: độ dài chuỗi thay đổi trong khoảng từ 8 -10 ký tự. Trong đó 9 và 10 chiếm tỉ lệ phần lớn
  + Cột LSOA: độ dài chuỗi dao động trong các giá trị 2, 10 và 11. Trong đó 11 chiếm tỉ lệ lên tới 97%
* Về pattern của từng cột: giá trị có nghĩa của 2 cột đều được đặt trong cặp dấu **“ ”**. Từ đó có thể thấy những giá trị của cột LSOA với độ dài chuỗi là 2 là những chuỗi rỗng. Thông qua thống kê ta đếm được có tổng cộng 10332 hàng có giá trị như vậy chiếm tỉ lệ 4%.
  1. **Bảng Postcode**
* Phần lớn các cột không có dữ liệu rỗng trừ cột City với tỉ lệ rỗng là 19% (chuỗi có độ dài bằng 0)
* Cột Postcode có khả năng là khóa chính của bảng này
* Đối với cột Country ta có 8 giá trị tượng trưng cho 8 nước của vương quốc Anh
* Đối với cột Region name thì có 6 dòng mang giá trị “#N/A”
  1. **Bảng Wiki Postcode**
* Không có dữ liệu null
* Đối với cột Postcode, các giá trị postcode được nối với nhau bởi dấu phẩy
* Ở cột county, một số giá trị sẽ được đặt trong cặp dấu “( )”
  1. **Bảng Postcode District**
* Dữ liệu tương đối gọn gàng
* Chỉ có duy nhất 1 ô dữ liệu null
  1. **Thiết kế NDS**



*Cấu trúc các bảng trong NDS*

* 1. **Bảng Vehicle**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| Vehicle\_ID | Khóa chính tự tăng |  |  |
| Accident\_ID | Mã vụ tai nạn | Bảng Accidents |  |
| VehicleType | Mã loại phương tiện | Bảng VehicleType |  |
| JourneyPurpose | Mã mục đích chuyến đi | Bảng JourneyPurpose |  |
| DriverSex | Giới tính tài xế | Vehicles | Thay 1, 2 thành M, F |
| AgeBand | Độ tuổi của tài xế | Bảng AgeBand |  |
| EngineCapacity | Phân khối của động cơ | Vehicles |  |
| VehicleAge | Tuổi của phương tiện | Vehicles |  |
| Source\_ID | Lưu giữ nguồn của dữ liệu |  | Tự định nghĩa |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Vehicle Type**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| VehicleType\_ID | Khóa chính tự tăng |  |  |
| Code\_NK | Khóa tự nhiên | VehicleType - Codebook |  |
| Label | Loại phương tiện | VehicleType - Codebook |  |
| Source\_ID | Lưu giữ nguồn của dữ liệu |  | Tự định nghĩa |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Journey Purpose**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| JourneyPurpose\_ID | Khóa chính tự tăng |  |  |
| Code\_NK | Khóa tự nhiên | JourneyPurpose - Codebook |  |
| Label | Mục đích chuyến đi | JourneyPurpose - Codebook |  |
| Source\_ID | Lưu giữ nguồn của dữ liệu |  | Tự định nghĩa |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Ageband**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| AgeBand\_ID | Khóa chính tự tăng |  |  |
| Code\_NK | Độ tuổi | Ageband - Codebook |  |
| Label | Tên độ tuổi | Ageband - Codebook |  |
| Source\_ID | Lưu giữ nguồn của dữ liệu |  | Tự định nghĩa |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Casual Type**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| CasualType\_ID | Khóa chính tự tăng |  |  |
| Code\_NK | Khóa tự nhiên | CasualType - Codebook |  |
| Label | Loại nạn nhân | CasualType - Codebook |  |
| Source\_ID | Lưu giữ nguồn của dữ liệu |  | Tự định nghĩa |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Accident**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| Accident\_ID | Khóa đại diện |  |  |
| Accident\_NK | Khóa tự nhiên | Accident\_Index  - dbo.Accidents (Stage) |  |
| Accident\_Severity | Mức độ nghiêm trọng | Accident\_Severity  - dbo.Accidents (Stage) | Từ stage, lookup dbo.Severity (NDS) để ra khóa đại diện |
| Date | Ngày xảy ra | Date  - dbo.Accidents (Stage) |  |
| Time | Thời gian xảy ra | Time  - dbo.Accidents (Stage) |  |
| Local\_Authority | Chính quyền địa phương | Local\_Authority\_(Highway)  - dbo.Accidents (Stage) | Từ stage, lookup dbo.Local\_Authority\_Highway (NDS) để ra khóa đại diện |
| RoadType | Loại đường | Road\_Type  - dbo.Accidents (Stage) | Từ stage, lookup dbo.RoadType (NDS) để ra khóa đại diện |
| SpeedLimit | Giới hạn tốc độ | Speed\_limit  - dbo.Accidents (Stage) |  |
| Build\_Up\_Road |  |  |  |
| Location\_ID | Địa điểm | Easting, Northing, Latitude, Longtitude  - dbo.Accidents (Stage) | Từ stage, lookup dbo.Location (NDS) để ra khóa đại diện |
| Police\_ID | Lực lượng cảnh sát | Police\_Force  - dbo.Accidents (Stage) | Từ stage, lookup dbo.PoliceForce (NDS) để ra khóa đại diện |
| County\_ID | Mã hạt | Bảng County |  |
| Town\_ID | Mã thị trấn | Bảng Town |  |
| Source\_ID | Nguồn dữ liệu | Tự thêm |  |
| Location\_Easting | Vị trí phía đông | Cột Postcode  Bảng dbo.accidents  Database Stage | Từ stage, lookup dbo.Region (NDS) để ra khóa đại diện |
| Location\_Northing | Vị trí phía bắc | Cột Postcode  Bảng dbo. accidents  Database Stage |  |
| Location\_Longtitude | Kinh độ | Cột Postcode  Bảng dbo. accidents  Database Stage |  |
| Location\_Latitude | Vĩ độ | Cột Postcode  Bảng dbo. accidents  Database Stage |  |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Police Force**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| PoliceForce\_ID | Khóa đại diện |  |  |
| Code\_NK | Khóa tự nhiên | Codebook.csv– PoliceForce sheet |  |
| Label | Tên lực lượng cảnh sát | Codebook.csv– PoliceForce sheet |  |
| SourceID | Nguồn dữ liệu | Tự thêm |  |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Road Type**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| RoadType\_ID | Khóa đại diện |  |  |
| Code\_NK | Khóa tự nhiên | Codebook.csv– RoadType sheet |  |
| Label | Tên loại đường | Codebook.csv– RoadType sheet |  |
| SourceID | Nguồn dữ liệu | Tự thêm |  |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Severity**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| Severity\_ID | Khóa đại diện |  |  |
| Code\_NK | Khóa tự nhiên | Codebook.csv– Severity sheet |  |
| Label | Tên mức độ | Codebook.csv– Severity sheet |  |
| SourceID | Nguồn dữ liệu | Tự thêm |  |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Source\_ID**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| Source\_ID | Khóa đại diện | Khóa tự tăng |  |
| Source\_Name | Khóa tự nhiên | Tùy theo loại data mà có giá trị khác nhau |  |

* 1. **Bảng Local Authority Highway**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| Local\_Authority\_Highway\_ID | Khóa đại diện | Khóa tự tăng |  |
| Code\_NK | Khóa tự nhiên | Codebook.csv – Local Authority (Highway) sheet |  |
| Label | Tên tổ chức giao thông vận tải |  |  |
| SourceID | Nguồn dữ liệu | Tự thêm |  |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Casualties Class**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| CasualtiesClass\_ID | Khóa đại diện | Khóa tự tăng |  |
| Code\_NK | Khóa tự nhiên | Codebook.csv – Casualty Class sheet |  |
| Label | Tên tổ chức giao thông vận tải |  |  |
| SourceID | Nguồn dữ liệu | Tự thêm |  |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Casualties**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| Casualties\_ID | Khóa đại diện | Khóa tự tăng |  |
| Accident\_ID |  | Accident\_Index (Stage) – dbo.Casualties | Từ stage, lookup Accident (NDS) để lấy khóa |
| Casualty\_Class |  | Casualty\_Class (Stage) – dbo.Casualties | Từ stage, lookup CasualtiesClass (NDS) để lấy khóa |
| Sex | Giới tính | Sex\_Of\_Casualty (Stage) – dbo.Casualties |  |
| Age | Tuổi | Age\_Of\_Casualty (Stage) – dbo.Casualties |  |
| AgeBand | Độ tuổi | Age\_Band\_Of\_Casualty (Stage) – dbo.Casualties | Từ stage, lookup AgeBand (NDS) để lấy khóa |
| CasualtyType |  | Casualty\_Type (Stage) – dbo.Casualties | Từ stage, lookup CasualtyType (NDS) để lấy khóa |
| SourceID | Nguồn dữ liệu | Tự thêm |  |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Region**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| Region\_ID | Khóa tự tăng |  |  |
| RegionName | Tên Khu Vực | Cột region\_name Bảng dbo.postcodes  Database Stage |  |
| Source\_ID | Mã Nguồn | Bảng Source\_ID |  |
| Region\_NK | Khóa của nguồn | Postcode | Từ stage, lookup dbo.Region (NDS) để ra khóa đại diện |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng LSOA\_Mapping**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| LSOA | Mã LSOA | Bảng Stage LSOA |  |
| Postcode\_ID | Mã postcode | Bảng dbo.postcodes | Từ stage, lookup dbo.postcode  (NDS) để ra khóa đại diện |
| Source\_ID | Mã Nguồn | Bảng Source\_ID |  |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Country**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| Country\_ID | Khóa Tự Tăng |  |  |
| Country\_NK | Khóa của nguồn | Cột country\_code Bảng dbo.postcodes  Database Stage | Từ stage, lookup dbo.Country (NDS) để ra khóa đại diện |
| CountryName | Tên quốc gia | Cột country\_name Bảng dbo.postcodes  Database Stage |  |
| Source\_ID | Mã Nguồn | Bảng Source\_ID |  |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng County**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| County\_ID | Khóa Tự Tăng |  |  |
| County\_NK | Khóa của nguồn | Cột Postcode\_district của Bảng dbo.wikipostcodes của Database Stage | Phải tách cách giá trị kép thành các dòng đơn |
| CountyName | Tên hạt | Cột county\_name Bảng dbo.wikipostcodes  Database Stage | Xóa cặp dấu ( ) |
| Source\_ID | Mã Nguồn | Bảng Source\_ID |  |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng Town**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| Town\_ID | Khóa Tự Tăng |  |  |
| Town \_NK | Khóa của nguồn | Cột postcode trong bảng Postcode districs ở db Stage |  |
| TownName | Tên thị trấn | Cột Town Bảng dbo.postcodesDistrict  Database Stage | Select distinc |
| Source\_ID | Mã Nguồn | Bảng Source\_ID |  |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Bảng PostCode**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Nguồn | Transformation |
| Postcode\_ID | Khóa Tự Tăng |  |  |
| Postcode | Khóa của nguồn |  |  |
| Latitude | Vĩ độ | Cột Latitude  Bảng dbo.postcodes  Database Stage |  |
| Longtitude | Kinh độ | Cột Longtitude  Bảng dbo.postcodes  Database Stage |  |
| Easting | Hướng Nam | Cột Easting  Bảng dbo.postcodes  Database Stage |  |
| Northing | Hướng Bắc | Cột Northing  Bảng dbo.postcodes  Database Stage |  |
| TownID | Thị trấn | NDS\_Town | Look up bằng Postcode\_NK |
| CountyID | Quận | NDS\_County | Look up bằng Postcode\_NK |
| CountryID | Mã Quốc gia | Cột CountryID  Bảng dbo.postcodes  Database Stage |  |
| RegionID | Mã khu vực | Cột RegionID  Bảng dbo.postcodes  Database Stage |  |
| Source\_ID | Mã Nguồn |  |  |
| Create\_time | Thời gian tạo |  |  |
| Update\_time | Thời gian cập nhật |  |  |

* 1. **Đổ dữ liệu từ Stage vào NDS**

Căn cứ vào thiết kế của NDS ở trên ta thực hiện các phép rút trích dữ liệu để đổ vào các bảng trong NDS. Song song trong quá trình đó ta cũng cần phải giải quyết những sai sót của dữ liệu từ Stage như dữ liệu rỗng, trùng lắp dữ liệu, các giá trị ngoại lai, …

* 1. **Hướng dẫn cài đặt**
  2. **Phân công**

Link github: <https://github.com/NhatTanNgo/UK-Car-Accidents-Data-warehouse>

|  |  |  |  |
| --- | --- | --- | --- |
| **Lần** | **Tên thành viên** | **Phân công** | **Mức độ hoàn thành** |
| 1 | Lâm Xương Đức | * Mô tả dữ liệu bảng Casualties * Mô tả dữ liệu bảng Vehicles | 100% |
| Phạm Đức Huy | * Mô tả dữ liệu bảng LSOA Mapping | 100% |
| Lâm Ngọc Anh Khoa | * Mô tả dữ liệu bảng Accidents * Mô tả dữ liệu bảng WikiPostcodes | 100% |
| Ngô Nhật Tân | * Mô tả dữ liệu bảng UK Postcodes * Mô tả dữ liệu bảng Postcode districts | 100% |
| 2 | Lâm Xương Đức | * Staging bảng Casualties * Staging bảng Vehicles | 100% |
| Phạm Đức Huy | * Staging bảng LSOA- Postcode Mapping | 100% |
| Lâm Ngọc Anh Khoa | * Staging bảng Accidents | 100% |
| Ngô Nhật Tân | * Staging bảng UK Postcodes * Staging bảng Codebook | 100% |
| 3 | Lâm Xương Đức | * ETL bảng Casualties vào NDS | 100% |
| Phạm Đức Huy | * ETL bảng Vehicles vào NDS | 100% |
| Lâm Ngọc Anh Khoa | * ETL bảng Accidents vào NDS | 100% |
| Ngô Nhật Tân | * ETL các bảng từ Codebook vào NDS | 100% |

* 1. **Tham khảo**

1. **Nguồn của các tập dữ liệu bên ngoài**

* <https://www.doogal.co.uk/PostcodeDistricts.php>
* <https://en.wikipedia.org/wiki/List_of_postcode_districts_in_the_United_Kingdom>

1. **Nguồn tài liệu hỗ trợ nghiên cứu**

* <https://www.getthedata.com/postcode>
* <https://www.mapmarketing.com/pages/custom-postcode-maps-site-centred-maps-to-suit-your-business>
* <https://traveltime.com/blog/postcode-map>
* <https://www.townscountiespostcodes.co.uk>
* <https://www.visitnorthwest.com/counties/>
* https://en.wikipedia.org/wiki/Regions\_of\_England