**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** 3](#_Toc85637937)

[**1.** **Dữ liệu UK Accident** 3](#_Toc85637938)

[**1.1.** **Accidents** 3](#_Toc85637939)

[**1.2.** **Casualties** 6](#_Toc85637940)

[**1.3.** **Vehicles** 7](#_Toc85637941)

[**2.** **Dữ liệu LSOA- Postcode Mapping** 9](#_Toc85637942)

[**3.** **Dữ liệu UK Postcodes** 11](#_Toc85637943)

[**II.** **Phân tích yêu cầu** 13](#_Toc85637944)

[**III.** **Giai đoạn Staging** 13](#_Toc85637945)

[**IV.** **Phân công** 14](#_Toc85637946)

* 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 | Số lượng người trên xe ô tô |  |
| Bus\_or\_Coach\_Passenger | Số | Số lượng 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**

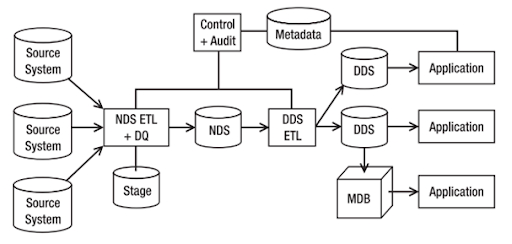
|  |  |  |  |
| --- | --- | --- | --- |
| **Tên cột** | **Kiểu giá trị** | **Ý nghĩa** | **Ghi chú** |
| pcd7 | chuỗi | Phiên bản 7. Của mã postcode | -Một khoảng cách giữa quận và phần đơn vị ngành của mã bưu điện.  Ví dụ:  'BT1 1AA',  'BT48 6PL'  - Có thể là định dạng phổ biến nhất của mã bưu điện. |
| pcd8 | chuỗi | Phiên bản 8. Của mã postcode | - Một khoảng cách giữa quận và phần đơn vị ngành của mã bưu điện. |
| pcds | chuỗi | Phiên bản S | - Một khoảng cách giữa quận và phần đơn vị ngành của mã bưu điện. |
| 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 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. **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 phương pháp 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.

* 1. **Phân tích yêu cầu**

Quá trình phân tích, xây dựng CSDL nguồn diễn ra như sau:

* Đầu tiên, cần tổng hợp dữ liệu từ các nguồn vào từng bảng riêng biệt để dễ dàng cập nhật, rút trích dữ liệu trong tương lai.
* Cập nhật thời gian dữ liệu được tạo ra hoặc được chỉnh sửa
* Sau khi đã hoàn thành đánh giá giá trị, ta cần xem xét các yêu cầu phân tích để rút ra những dữ liệu cần thiết lưu trữ trong Stage.

1. 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. Cả 2 dữ liệu này đều có bảng để tham chiếu giá trị nội dung.

1. 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

1. 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

1. 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

1. 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

1. 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

1. 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

1. 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. **Giai đoạn Staging**

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 ta load 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.
* Sau khi đã load đầy đủ và điều chỉnh các dữ liệu thời gian, ta tiến hành profiling để đánh giá dữ liệu.
* Tiếp theo ta load dữ liệu từ các database trung gian qua Stage
  + Ở Metadata, ta 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 thời điểm hiện tại 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 vào Metadata
  1. **Phân công**

|  |  |  |  |
| --- | --- | --- | --- |
| **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- Postcode Mapping | 100% |
| Lâm Ngọc Anh Khoa | * Mô tả dữ liệu bảng Accidents | 100% |
| Ngô Nhật Tân | * Mô tả dữ liệu bảng UK Postcodes | 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% |