**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

****

**PROJECT SUBMISSION 2**

Phạm Ngọc Thùy Trang 18127022

Ngô Thị Thanh Thảo 18127053

Hoàng Nguyên Trúc 18127055

**Môn học: Hệ thống thông tin phục vụ trí tuệ kinh doanh**

**Thành phố Hồ Chí Minh**

**Ộ MÔN HỆ**

# Mục lục

[I. Mục lục 1](#_Toc86524250)

[II. Bảng thông tin chi tiết nhóm 2](#_Toc86524251)

[III. Phân công công việc 3](#_Toc86524252)

[IV. Mô tả ý nghĩa các thuộc tính của nguồn dữ liệu 4](#_Toc86524253)

[1. Postcode 4](#_Toc86524254)

[2. LSOA 5](#_Toc86524255)

[3. Bảng Casualties: 11](#_Toc86524256)

[4. Bảng Accidents: 15](#_Toc86524257)

[5. Bảng Vehicles: 18](#_Toc86524258)

[V. ETL từ Source sang Stage 21](#_Toc86524259)

[VI. References 27](#_Toc86524260)

# Bảng thông tin chi tiết nhóm

|  |  |  |  |
| --- | --- | --- | --- |
| **Mã nhóm:** | Nhóm 2 | | |
| **Số lượng:** | 3 | | |
| **MSSV** | **Họ tên** | **Email** | **SĐT** |
| 18127022 | Phạm Ngọc Thùy Trang | 18127022@  student.hcmus.edu.vn |  |
| 18127053 | Ngô Thị Thanh Thảo | 18127053@  student.hcmus.edu.vn |  |
| 18127055 | Hoàng Nguyên Trúc | 18127055@  student.hcmus.edu.vn |  |

# Phân công công việc

|  |  |  |
| --- | --- | --- |
| **Họ tên thành viên** | **Nội dung công việc** | **Phần trăm hoàn thành (%)** |
| Phạm Ngọc Thùy Trang |  |  |
| Ngô Thị Thanh Thảo |  |  |
| Hoàng Nguyên Trúc |  |  |

# Mô tả ý nghĩa các thuộc tính của nguồn dữ liệu

## Postcode

**Các thuộc tính sử dụng:**

* **Postcode (Mã bưu chính):** là mã được quy định bởi liên hiệp bưu chính toàn cầu giúp định vị khi chuyển thư, bưu phẩm. Hiểu đơn giản, mã bưu chính là một “tọa độ” của địa chỉ trên “bản đồ bưu chính”. Mỗi quốc gia có một hệ thống ký hiệu mã bưu chính riêng. Mỗi mã bưu chính trong một quốc gia, ký hiệu cho một địa phận bưu điện duy nhất. Địa phận bưu điện có thể là một làng nhỏ, một thành phố nhỏ, một quận của một thành phố lớn.
* **Latitude:** là vĩ độ - một giá trị xác định vị trí của một điểm trên bề mặt Trái Đất ở phía bắc hay phía nam của xích đạo.
* **Longitude:** là kinh độ - một giá trị tọa độ địa lý theo hướng đông-tây, được sử dụng phổ biến nhất trong bản đồ học và hoa tiêu toàn cầu
* **City:** Thành phố
* **County (Hạt):** Là một khu vực nằm trong một bang và lớn hơn thành phố, có chính quyền riêng để xử lý các công việc địa phương. Ví dụ, Hạt El Paso nằm trong bang Colorado, có thành phố Colorado Springs và các cộng đồng dân cư khác.
* **Region\_code:** mã vùng (mã khu vực địa lý). Thông thường, một khu vực địa lý có thể bao gồm nhiều quốc gia khác nhau.
* **Region name:** Tên khu vực địa lý tương ứng với region\_code
* **Country\_code:** Mã quốc gia (một loại mã địa lý bằng chữ hoặc bằng số được dùng để đại diện cho một quốc gia hoặc một khu vực phụ thuộc)
* **Country\_name:** Tên quốc gia tương ứng với country\_code

## LSOA

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên thuộc tính** | **Ý nghĩa/Mô tả thuộc tính** | **Phạm vi giá trị của mã** | **Độ dài thuộc tính** |
| **pcd7** | Còn được gọi là unit postcode (mã bưu chính độc nhất) phiên bản dài 7 ký tự. Pdc7 chính là những mã bưu chính còn hoạt động cho đến thời điểm hiện tại nằm trong vương quốc Anh, Channel Islands và Isle of Man có nhận thư hàng tháng từ Royal Mail - công ty dịch vụ bưu chính và chuyển phát nhanh đa quốc gia của Anh, cũng như tất cả những mã bưu chính bị đóng và không được sử dụng lại bởi Royal Mail trong vương quốc Anh và bởi các cơ quan hành chính của Channel Island và Isle of Man | AB1Δ1AA-ZE999ZZ (trong đó Δ đại diện cho ký tự khoảng trắng) | 7 ký tự (bao gồm ký tự khoảng trắng) |
| **pcd8** | Như trên, chỉ khác ở điểm đây là mã bưu chính độc nhất phiên bản 8 ký tự | AB1ΔΔ1AAZE99Δ9ZZ (trong đó Δ là ký tự khoảng trắng) | 8 ký tự (bao gồm ký tự khoảng trắng) |
| **oac11cd** | Viết tắt của cụm từ Output Area và có tên gọi đầy đủ là 2011 Census Output Area (hoặc là SA - Small Area - những khu vực nhỏ trong vương quốc Anh). Trong đó, pseudo oac11cd code sẽ được dùng cho Channel Islands và Isle of Man. oac11cd code sẽ để trống đối với các mã bưu chính không nằm trong hệ thống lưới. Oac11d được dùng để định nghĩa một cấp độ địa lý có cấp cao hơn unit postcode. Một output area sẽ có thể có hoặc không có khu vực kế vị (cũng là một output area), nhưng sẽ phải có một khu vực cha - hay còn gọi là SOA (Lower Super Output Area) | E00XXXXXX, W00XXXXXX, S00XXXXXX, 95AA010001- 95ZZ519999, pseudo hoặc null (không có thông tin).  Cụ thể hơn:  E00000001 - E00165665 = England; W00000001 - W00009769 = Wales; S00000001 - S00042604 = Scotland; 95AA010001 - 95ZZ519999 = Northern Ireland; L99999999 (pseudo code) = Channel Islands; M99999999 (pseudo code) = Isle of Man | 9 - 10 ký tự (hoặc chỉ có 1 ký tự nào trong trường hợp không có thông tin) |
| **ladnmw** | Tên khu vực tương ứng với mã oac11cd, nếu không có thì để trống |  |  |
| **lsoa11cd** | Viết tắt của cụm từ Lower Super Output Area và có tên gọi đầy đủ là 2011 Census Lower Layer Super Output Area, hoặc là Data Zone (DZ). **lsoa11cd** đại diện cho mã 2011 LSOA của khu vực England và xứ Wales, đại diện cho mã SOA của khu vực Northern Ireland and DZ code cho Scotland. Trong đó, pseudo **lsoa11cd** code sẽ được dùng cho Channel Islands và Isle of Man. **lsoa11cd** code sẽ để trống đối với các mã bưu chính không có mã OA. Một LSOA có thể có nhiều hoặc không có các output area. Thông thường, một LSOA sẽ phải có một MSOA (middle super output area). **lsoa11cd** sẽ có tên gọi là DZONE1 dùng để chỉ những mã bưu chính nằm trong Scotland, và sẽ được gọi là SOA1 khi nó đại diện cho những mã bưu chính nằm trong England và Wales. | E01XXXXXX  S01XXXXXX  W01XXXXXX  Null (nếu không có thông tin hiển thị)  Cụ thể hơn:  E01000001- E01032482 = England;  W01000001- W01001896 = Wales;  S99999999 (pseudo) = Scotland;  N99999999 (pseudo) = Northern Ireland;  L99999999 (pseudo) = Channel Islands;  M99999999 (pseudo) = Isle of Man | 9 ký tự (hoặc chỉ 1 ký tự null nếu không có thông tin) |
| **lsoa11nm** | Tên khu vực tương ứng với mã **lsoa11cd**, nếu không có thì để trống |  |  |
| **msoa11cd** | Viết tắt của cụm từ Middle Super Output Area và có tên gọi đầy đủ là 2011 Middle Lower Layer Super Output Area đối với khu vực England và Wales hoặc có tên là Intermediate Zone(IZ) đối với khu vực Scotland. Trong đó, pseudo **msoa11cd** code sẽ được dùng cho Channel Islands và Isle of Man. **msoa11cd** code sẽ để trống đối với các mã bưu chính nằm trong hệ thống lưới. Một MSOA có thể có nhiều hoặc không có LSOA cũng như có thể có hoặc không có các workplace zone. Thông thường, một MSOA sẽ phải có một khu vực cha gọi là Local Authority - LA (Chính quyền địa phương). Tuy nhiên, nó có thể có hoặc không có LSOA cũng như workplace zone. | * **E02XXXXXX** * **S02XXXXXX** * **W02XXXXXX**   Cụ thể là:  E02000001- E02006781 = England; W02000001- W02000413 = Wales; S99999999 (pseudo) = Scotland; N99999999 (pseudo) = Northern Ireland; L99999999 (pseudo) = Channel Islands; M99999999 (pseudo) = Isle of Man | 9 ký tự (hoặc chỉ 1 ký tự null nếu không có thông tin) |
| **msoa11nm** | Tên khu vực tương ứng với mã **msoa11cd**, nếu không có thì để trống. |  |  |
| **ladcd** | Viết tắt của cụm từ Local Authority, **ladcd** chính là  Quận hiện tại mà mã bưu chính đã được chỉ định. LA có rất nhiều tên đầy đủ khác nhau, trong đó có Local Authority District (LAD), hoặc là unitary authority (UA), hoặc là metropolitan district (MD), hoặc London borough (LB), hoặc council area (CA) hoặc district council area (DCA). Trong đó, pseudo **ladcd** code sẽ được dùng cho Channel Islands và Isle of Man. **ladcd** code sẽ để trống đối với các mã bưu chính không nằm trong hệ thống lưới. Một LA có thể có nhiều hoặc không có các phường cũng như có thể có hoặc không có các khu vực cộng đồng. Tuy nhiên, một LA cần phải thuộc về một quốc gia nào đó. | E06XXXXXX  E07XXXXXX  E08XXXXXX  E09XXXXXX  N09XXXXXX  S12XXXXXX  W06XXXXXX  Cụ thể hơn:  E06000001 - E06000056 = England (UA); E07000001 - E07000239 = England (LAD); E08000001 - E08000036 = England (MD); E09000001 - E09000033 = England (LB); W06000001 - W06000024 = Wales (UA); S12000005 - S12000044 = Scotland (CA); 95A – 95Z = Northern Ireland (DCA); L99999999 (pseudo) = Channel Islands; M99999999 (pseudo) = Isle of Man |  |
| **ladnm** | Tên chính quyền địa phương tương ứng với mã ladcd, nếu không có thì để trống |  |  |

## Bảng Casualties:

**Các thuộc tính sử dụng:**

* **Accident\_Index:** mã số tai nạn
* **Sex\_of\_Casualty:** giới tính nạn nhân. Các giá trị được quy định như sau

|  |  |
| --- | --- |
| Code | Label |
| 1 | Nam |
| 2 | Nữ |
| 3 | Không rõ |
| -1 | Dữ liệu bị thiếu hoặc nằm ngoài phạm vi dữ liệu có sẵn |

* **Age\_Band\_of\_Casualty:** nhóm độ tuổi thương vong (0-5, 6-10, 11-15,...hơn 75 tuổi). Các giá trị được quy định như sau

|  |  |
| --- | --- |
| Code | Label |
| 1 | 0-5 |
| 2 | 6-10 |
| 3 | 11-15 |
| ... | ... |

* **Casualty\_Severity:** mức độ nghiêm trọng gồm ba mức độ :gây tử vong,

nghiêm trọng, nhẹ. Các giá trị được quy định như sau

|  |  |
| --- | --- |
| Code | Label |
| 1 | Gây tử vong |
| 2 | Nghiêm trọng |
| 3 | Nhẹ |

* **Casualty\_Type:** loại nạn nhân. Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 0 | Người dân |
| 1 | Người đi xe đạp |
| 2 | Hành khách hoặc những người đi xe moto có phân khối <= 50cc |
| ... | ... |

* **Age\_of\_Casualty:** độ tuổi của nạn nhân bị thương vong.
* **Age\_Band\_of\_Casualty**: các nhóm độ tuổi của nạn nhân bị thương vong. Các giá trị được quy định như sau

|  |  |
| --- | --- |
| Code | Label |
| 1 | 0-5 |
| 2 | 6-10 |
| 3 | 11-15 |

* **Casualty\_Home\_Area\_Type**: loại khu vực nạn nhân sinh sống. Các giá trị được quy định như sau

|  |  |
| --- | --- |
| Code | Label |
| 1 | Khu vực thành thị |
| 2 | Thị trấn nhỏ |
| 3 | Nông thôn |
| -1 | Dữ liệu bị thiếu hoặc nằm ngoài phạm vi dữ liệu có sẵn |

* **Casualty\_Class**: lớp nạn nhân (gồm 3 loại: người lái xe hoặc tài xế, cư dân, hành khách). Các giá trị được quy định như sau

|  |  |
| --- | --- |
| Code | Label |
| 1 | Tài xế hoặc người lái xe |
| 2 | Hành khách |
| 3 | Cư dân |

* **Car\_passenger:** Vị trí của hành khách bị xảy ra tai nạn nếu có đi ô tô. Các giá trị được quy định như sau

|  |  |
| --- | --- |
| Code | Label |
| 0 | Không phải hành khách đi ô tô |
| 1 | Hành khách ngồi ghế trước |
| 2 | Hành khách ngồi ghế sau |
| -1 | Dữ liệu bị thiếu hoặc nằm ngoài phạm vi dữ liệu có sẵn |

* **Bus\_passenger:** Vị trí của hành khách bị xảy ra tai nạn nếu có đi xe buýt. Các giá trị được quy định như sau

|  |  |
| --- | --- |
| Code | Label |
| 0 | Không phải là hành khách đi xe buýt hoặc xe khách |
| 1 | Lên xe |
| 2 | Xuống xe |
| 3 | Hành khách đứng trên xe |
| 4 | Hành khách ngồi trên xe |
| -1 | Dữ liệu bị thiếu hoặc nằm ngoài phạm vi dữ liệu có sẵn |

* **Pedestrian\_Location**: Vị trí của người dân bị xảy ra tai nạn. Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 0 | Không phải là khu dân cư |
| 1 | Đi qua dải đi bộ |
| 2 | Đi qua đường zig zag |
| ... | ... |

* **Pedestrian\_Movement:** Chuyển động của người dân. Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 0 | Không phải là người đi |
| 1 | Băng qua từ phía gần người lái |
| 2 | Băng qua đường gần - bị che bởi xe đang đỗ hoặc đang dừng |
| ... | ... |

* **Vehicle\_Reference:** mã phương tiện

## Bảng Accidents:

**Các thuộc tính sử dụng:**

* **Accident\_Index:** mã số của tai nạn
* **Latitude:** là vĩ độ - một giá trị xác định vị trí của một điểm trên bề mặt Trái Đất ở phía bắc hay phía nam của xích đạo.
* **Longitude:** là kinh độ - một giá trị tọa độ địa lý theo hướng đông-tây, được sử dụng phổ biến nhất trong bản đồ học và hoa tiêu toàn cầu
* **Accident\_Severity:** mức độ nghiêm trọng. Các giá trị được quy định như sau

|  |  |
| --- | --- |
| Code | Label |
| 1 | Gây tử vong |
| 2 | Nghiêm trọng |
| 3 | Nhẹ |

* **Local\_Authority(District):** địa phương xảy ra tai nạn

|  |  |
| --- | --- |
| Code | Label |
| 1 | Westminster |
| 2 | Camden |
| 3 | Islington |
| ... | ... |

* **Date:** Ngày/tháng/năm xảy ra tai nạn
* **Number\_of\_Vehicles**: Số lượng phương tiện tham gia
* **Number\_of\_Casualties**: Số lượng thương vong
* **Speed\_limit:** tốc độ tối đa cho phép
* **Local\_Authority\_(District)** **(tính theo quận):** Các mã LA (chính quyền địa phương) tại các quận. Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 1 | Westminster |
| 2 | Camden |
| 3 | Islington |
| …. | …. |

* **Local\_Authority\_(Highway):** Các mã LA (chính quyền địa phương) tại các đường cao tốc. Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| S12000033 | Aberdeen City |
| S12000034 | Aberdeenshire |
| S12000041 | Angus |
| ... | ... |

* **Day\_of\_Week**: thứ (buổi) xảy ra tai nạn. Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 1 | Chủ nhật |
| 2 | Thứ hai |
| 3 | Thứ bai |
| ... | ... |
| 7 | Thứ bảy |

* **Time:** Thời điểm xảy ra tai nạn (giờ:phút)
* **Weather\_Conditions**: Điều kiện thời tiết. Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 1 | Thời tiết ổn và không có gió mạnh, gió giật |
| 2 | Trời mưa nhưng không có gió mạnh, gió giật |
| 3 | Tuyết rơi nhưng không có gió mạnh, gió giật |
| ... | …. |

* **Light\_Conditions:** Điều kiện ánh sáng. Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 1 | Ánh sáng ban ngày |
| 2 | Trời tối, đèn sáng |
| 3 | Trời tối, đèn không sáng |
| ... | ... |

* + **Road\_Surface**: Bề mặt đường. Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 1 | Khô ráo |
| 2 | Ẩm ướt |
| 3 | Tuyết phủ |
| ... | ... |

* + **Road\_Type:** Loại đường. Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 1 | Vòng xoay |
| 2 | Đường một chiều |
| 3 | Xa lộ hai chiều |
| ... | ... |

* + **Urban\_or\_Rural\_Area:** Loại khu vực. Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 1 | Đô thị |
| 2 | Nông thôn |
| 3 | Chưa phân bổ |

* **LSOA\_of\_Accident\_Location**: Mã vị trí nơi xảy ra tai nạn

## Bảng Vehicles:

**Các thuộc tính sử dụng:**

* + **Accident\_Index:** Mã số tai nạn
  + **Vehicle\_Reference:** Mã phương tiện
  + **Driver\_Home\_Area\_Type:** Loại khu vực nhà ở của tài xế. Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 1 | Khu đô thị |
| 2 | Thị trấn nhỏ |
| 3 | Nông thôn |
| -1 | Không có dữ liệu |

* + **Sex\_of\_Driver:** Giới tính của tài xế. Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 1 | Nam |
| 2 | Nữ |
| 3 | Không rõ |
| -1 | Dữ liệu bị thiếu hoặc nằm ngoài phạm vi dữ liệu có sẵn |

* + **Age\_Band\_of\_Driver:** Nhóm độ tuổi của tài xế (0-5, 6-10, 11-15,...hơn 75 tuổi). Các giá trị được quy định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 1 | 0-5 |
| 2 | 6-10 |
| 3 | 11-15 |
| ... | ... |

* **Journey\_Purpose\_of\_Driver:** Mục đích hành trình. Các giá trị được quy

định như sau:

|  |  |
| --- | --- |
| Code | Label |
| 1 | Hành trình như một phần công việc |
| 2 | Đi làm/Tan làm |
| 3 | Đưa/đón học sinh đi học |
| ... | ... |

* + **Was\_Vehicle\_Left\_Hand\_Drive:** Tay thuận của tài xế có là tay trái không

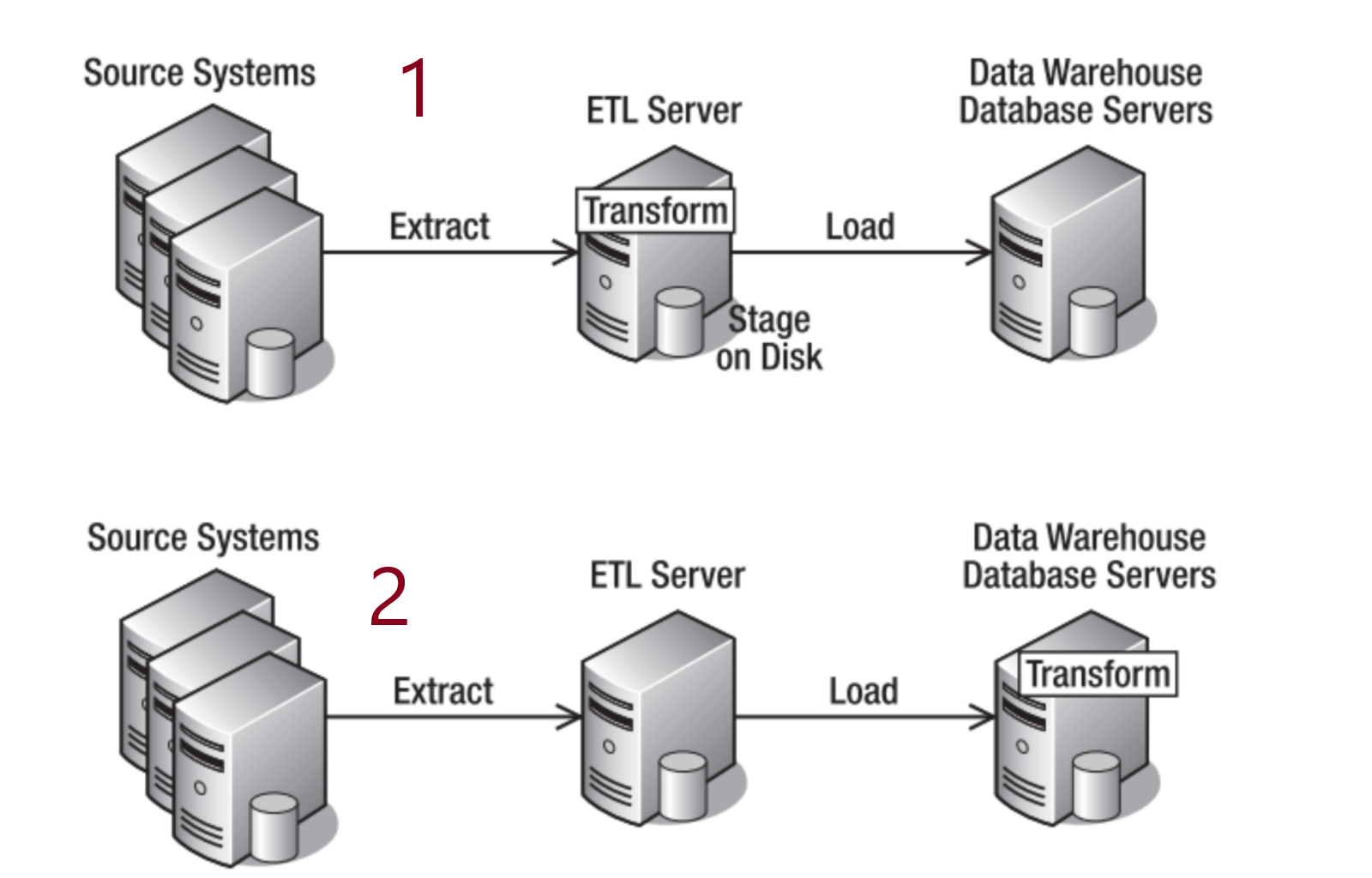
|  |  |
| --- | --- |
| Code | Label |
| 1 | Không |
| 2 | Có |
| -1 | Không có dữ liệu |

* **Vehicle\_Type:** Loại phương tiện. Các giá trị được quy định như sau:

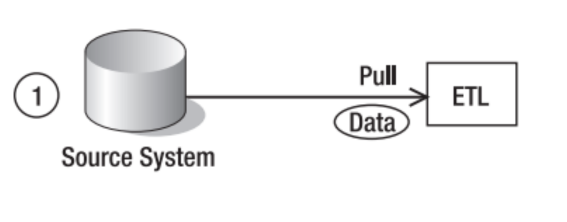
|  |  |
| --- | --- |
| Code | Label |
| 1 | Xe đạp |
| 2 | Xe mô tô dưới 50 phân khối |
| 3 | Xe mô tô dưới 125 phân khối |
| ... | ... |

# ETL từ Source sang Stage

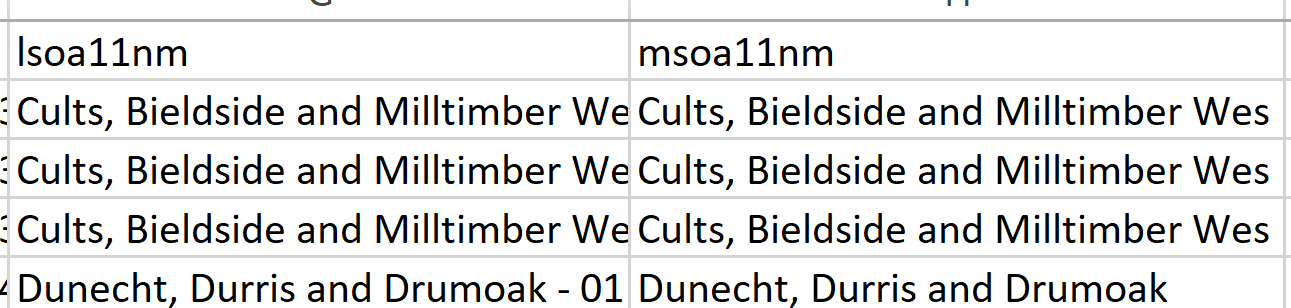
* Kiến trúc mà nhóm tiếp cận là kiến trúc số 1 vì dữ liệu sử dụng trên đồ án khá nhỏ gồm 5 file CSV có dung lượng nhỏ hơn … Vì vậy thực hiện transform (xử lý) ở memory nhanh hơn sau đó load vào Data Warehouse luôn. Độ lớn bộ nhớ của ETL server đủ để có thể hỗ trợ cho việc transform ở memory. Data WareHouse chỉ có nhiệm vụ nhận dữ liệu đã qua transform chứ không transform tại Data WareHouse như kiến trúc số 2 vì nhóm không có Data Warehouse đủ mạnh để vừa chứa dữ liệu vừa transform, có khả xử lý song song đồng thời



Ở đây đối tượng chủ động để lấy dữ liệu từ source system là ETL Server viết những thủ tục, chương trình để lấy dữ liệu từ source system một cách đều đặn. Source System đóng vai trò bị động. Hình ảnh dưới đây minh họa thể hiện rõ đối tượng lấy dữ liệu từ source system:



* Theo đồ án, quá trình ETL dữ liệu từ Source sang State sử dụng 5 Source csv là:
  + Accidents1114.csv
  + Casualty1114.csv
  + PCD\_OA\_LSOA\_MSOA\_LAD\_AUG21\_UK\_LU.csv
  + postcodes.csv
  + Vehicles1114.csv
* Với các nguồn như Accident1114, Casualty1114, postcodes, Vehicle1114 nhóm sử dụng trực tiếp lấy từ nguồn là flat file (file csv), đồng thời tạo thêm 2 trường đại diện cho nhãn đại diện cho thời gian (create\_date và update\_date) và 1 nhãn đại diện cho trạng thái “Dòng dữ liệu đã được xóa hay chưa” có tên là TrangThai (ở đây đang xét ngữ cảnh là dữ liệu bị xóa “soft deleted”) thủ công trên các file csv. Nhóm cũng đã chủ động bỏ những cột không cần thiết (không phục vụ quá trình phân tích và làm report) trong quá trình ETL.
* Với nguồn PCD\_OA\_LSOA\_MSOA\_LAD\_AUG21\_UK\_LU thì nhóm quyết định sử dụng cơ sở dữ liệu quan hệ để lưu dữ liệu nguồn và từ bây giờ, bất kỳ mọi thay đổi nào hoặc thêm mới dữ liệu sẽ được thực hiện dựa trên cơ sở dữ liệu quan hệ này thay vì file flat file như các bảng khác. Bên cạnh đó, việc tạo thêm 2 trường đại diện cho nhãn thời gian (createdate và updatedate) và TrangThai đại điện cho việc dòng dữ liệu đã bị xóa hay chưa. Nhóm cũng đã chủ động bỏ những cột không cần thiết (không phục vụ quá trình phân tích và làm report) trong quá trình ETL.
* Lý do việc nguồn PCD\_OA\_LSOA\_MSOA\_LAD\_AUG21\_UK\_LU thì nhóm quyết định sử dụng cơ sở dữ liệu quan hệ để lưu dữ liệu nguồn chứ không trực tiếp lấy từ nguồn là flat file (file csv) vì gặp một số khó khăn trong quá trình nhóm thử nghiệm trực tiếp lấy nguồn PCD\_OA\_LSOA\_MSOA\_LAD\_AUG21\_UK\_LU là flat file. Như hình minh họa ở dưới, khi xử lý dữ liệu, delimiter của SSIS là dấu phẩy thì lsoa11nm sẽ có giá trị là “Cults, Bieldside and Milltimber Wes” và msoa11nm là “Cults, Bieldside and Milltimber Wess”, nhưng delimiter của SSIS nhận diện sai trong phẩy trong csv vì dữ liệu cột lsoa11nm có dấu phẩy giống với dấu phẩy ngăn cách các cột dữ liệu trong csv dẫn đến việc mapping việc lỗi. Ví dụ như sau quá trình mapping thì lsoa11nm sẽ là “Cults” và msoa11nm là “Bieldside and Milltimber Wes” sai với kết quả mong muốn ban đầu. Nên nhóm quyết định với nguồn PCD\_OA\_LSOA\_MSOA\_LAD\_AUG21\_UK\_LU thì sử dụng cơ sở dữ liệu quan hệ để lưu dữ liệu nguồn .



* Đối với các nhãn đại diện cho thời gian là create\_date và update\_date, nhóm quyết định sử dụng kiểu dữ liệu là string thay cho timestamp hay datetime là vì khi thiết lập kết nối cho flat file cũng như thực hiện mapping column giữa dữ liệu nguồn và dữ liệu đích. Nhóm nhận thấy rằng luôn có lỗi xảy ra khi người dùng chọn kiểu dữ liệu là datetime hoặc timestamp khi đọc dữ liệu từ file csv lên, kể cả khi dùng script task để custom kiểu dữ liệu. SSIS không thể chuyển đổi được dữ liệu có được ở trường create\_date và update\_date thành kiểu dữ liệu ở dạng datetime.

“Error 0xc02020a1: Data Flow Task 1: Data conversion failed. The data conversion for column "create\_date" returned status value 2 and status text "The value could not be converted because of a potential loss of data.".

Chính là lỗi mà nhóm gặp phải. Để giải quyết vấn đề này, ban đầu nhóm có thử sử dụng derived column và dùng expression để chuyển đổi kiểu dữ liệu cho các cột này, nhưng vẫn không thành công. Cuối cùng, nhóm quyết định dùng kiểu dữ liệu

là chuỗi thay cho kiểu dữ liệu datetime. Điều này áp dụng cho cả các bảng ở cơ sở dữ liệu Stage và Temp. Các data flow vẫn chạy đúng và ổn định.

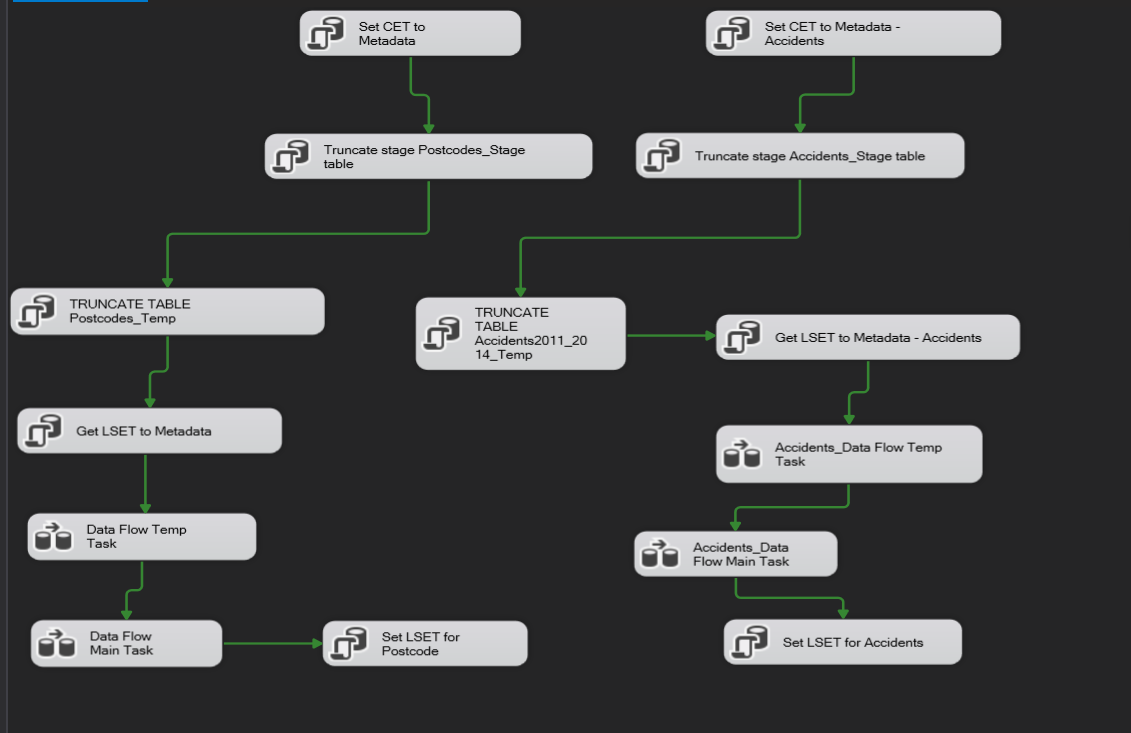
* Nhóm quyết định tạo thêm 1 cơ sở dữ liệu tạm gọi là Temp bên cạnh cơ sở dữ liệu nguồn là Stage và Metadata. Lý do là vì nếu load trực tiếp dữ liệu từ flat file csv thì không thể dùng cách tiếp cận “Incremental Extract” như làm việc với nguồn là cơ sở dữ liệu do ta không thể sử dụng SQL Command trên file CSV. Vì lẽ đó mà nhóm cần đổ dữ liệu vào cơ sở dữ liệu tạm trước, sau đó từ cơ sở dữ liệu tạm này sẽ dùng SQL Command để thực hiện “Incremental Extract” vào cơ sở dữ liệu Stage. Những dữ liệu được lấy sẽ là những dữ liệu có thời gian cập nhật hoặc thời gian được tạo lớn hơn hoặc bằng thời điểm LSET và nhỏ hơn thời điểm CET, đồng thời dữ liệu cũng phải có trạng thái “FALSE” (tức chưa bị xóa soft deleted).
* Tương ứng với 5 file csv ta sẽ có 5 bảng là:
  + Postcodes
  + Accidents2011\_2014
  + Casualty2011\_2014
  + Vehicles2011\_2014
  + UK\_Area\_Information
* Chiến lược rút trích dữ liệu từ cơ sở dữ liệu: Incremental extract
  + Cơ bản về đặc điểm của chiến lược Incremental extract là chỉ rút trích những thay đổi so với lần rút trích gần đây nhất (Dòng mới thêm, dòng mới xoá, dòng được cập nhật) dựa vào thuộc tính nhãn thời gian và những dòng có trạng thái dữ liệu cho biết dữ liệu chưa bị xóa (soft deleted) hay chưa.
  + Trong đồ án này sử dụng hai nhãn thời gian là create\_date (ngày tạo) và update\_date (ngày cập nhật) và nhãn đại diện cho Trạng thái của dữ liệu. Giá trị ban đầu của create\_date là thời điểm tạo ra/thêm mới dòng dữ liệu, giá trị ban đầu của update\_date là NULL, giá trị ban đầu của trạng thái là FALSE (tức là dữ liệu chưa có bị xóa - soft deleted). Mỗi khi có một dòng thay đổi thì cập nhật lại nhãn thời gian và trạng thái dữ liệu.
* Khi đổ dữ liệu vào nguồn tất cả các bảng có tên giống với tên đặt của csv ngoại trừ csv có tên là PCD\_OA\_LSOA\_MSOA\_LAD\_AUG21\_UK\_LU khi đổ vào nguồn đổi tên bảng thành UK\_Area\_Information
* Tạo một cơ sở dữ liệu metadata gồm bảng có các cột Name\_DataFlow (Tên bảng), LSET, CET. Trong đó:
  + LSET: Thời điểm rút trích hiện tại
  + CET: Thời điểm rút trích thành công gần đây nhất
* Tạo một cơ sở dữ liệu Stage có cấu trúc và các bảng giống với cơ sở dữ liệu Temp và cơ sở dữ liệu nguồn.
* Tạo một cơ sở dữ liệu Temp có cấu trúc và các bảng giống với cơ sở dữ liệu nguồn.
* Các bước thực hiện đổ từ bảng nguồn vào stage (đối với PCD\_OA\_LSOA\_MSOA\_LAD\_AUG21\_UK\_LU):

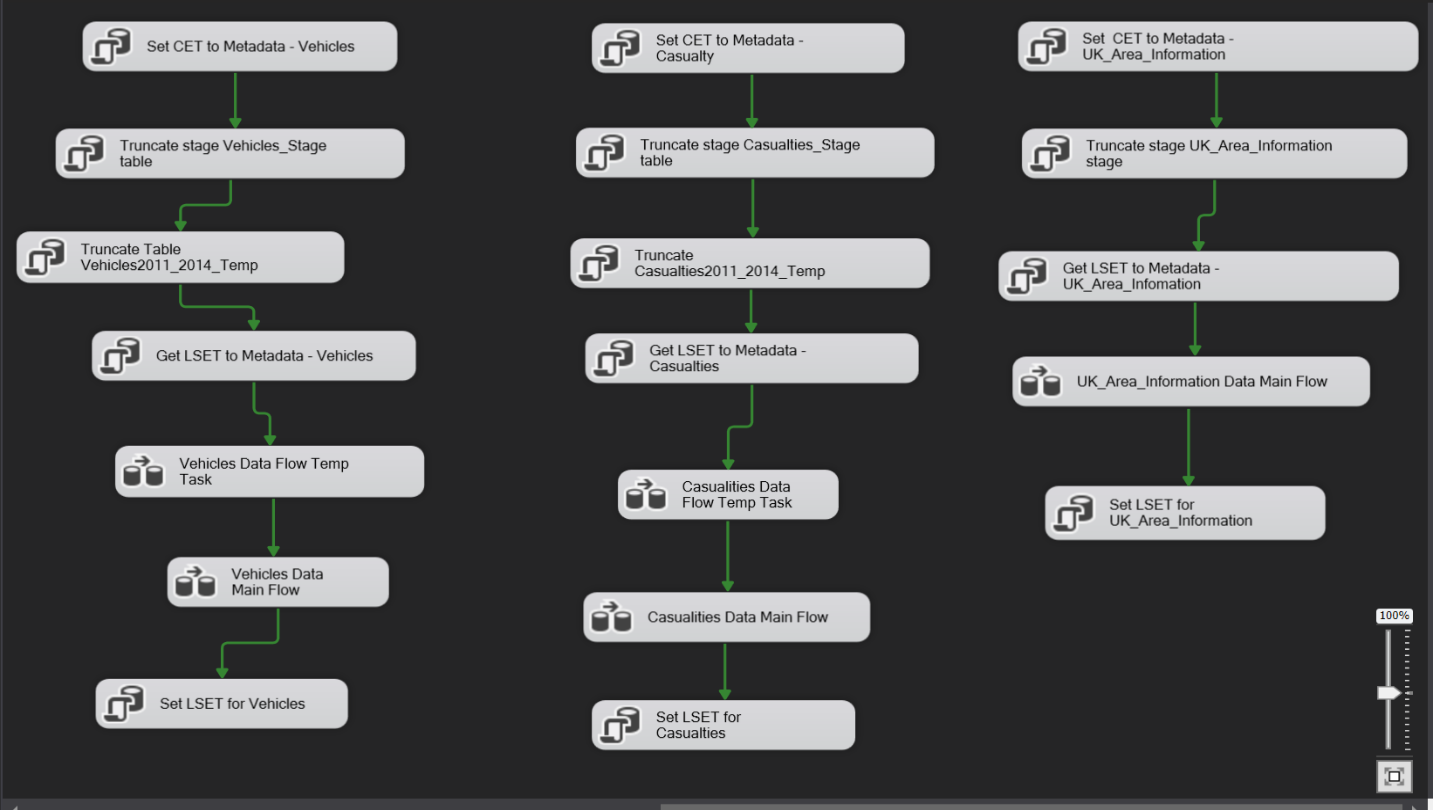
1. Lấy thông tin LSET được lưu trong metadata
2. Lấy CET bằng thời gian khởi động package
3. Truncate bảng Stage hiện tại: TRUNCATE TABLE <BangStage>
4. Rút trích dữ liệu từ bảng Source sang bảng Stage: select \* from BangSource where (createdate >= LSET and createdate < CET) or (updatedate >= LSET and updatedate < CET) AND (TrangThai = ‘FALSE’)
5. Cập nhật LSET = CET

* Các bước thực hiện đổ từ bảng Temp vào stage (đối với các nguồn còn lại):

1. Lấy thông tin LSET được lưu trong metadata
2. Lấy CET bằng thời gian khởi động package
3. Truncate bảng Stage hiện tại: TRUNCATE TABLE <BangStage>
4. Truncate bảng Temp hiện tại: TRUNCATE TABLE <BangTemp>
5. Rút trích dữ liệu từ flat file csv cụ thể vào trong bảng Temp
6. Rút trích dữ liệu từ bảng Temp sang bảng Stage: SELECT \* FROM <Tên bảng TEMP> WHERE (create\_date >= CAST(? AS VARCHAR)) AND (create\_date < CAST(? AS VARCHAR)) OR (update\_date >= CAST(? AS VARCHAR)) AND (update\_date < CAST(? AS VARCHAR)) AND (trangthai = 'FALSE').
7. Cập nhật LSET = CET.

* Flow khi thực hiện trên Visual Studio Code





# References

<http://doc.ukdataservice.ac.uk/doc/7961/mrdoc/pdf/7961_nspl_user_guide_2011.pdf>

[Lower Super Output Area - Area (findthatpostcode.uk)](https://findthatpostcode.uk/areatypes/lsoa11.html) (lsoa)

[Output area - Area (findthatpostcode.uk)](https://findthatpostcode.uk/areatypes/oa11.html) (oa)

[Middle Super Output Area - Area (findthatpostcode.uk)](https://findthatpostcode.uk/areatypes/msoa11.html) (msoa)

[Local Authority - Area (findthatpostcode.uk)](https://findthatpostcode.uk/areatypes/laua.html?p=45) (ladcd)

[Postcode to Output Area to Lower Layer Super Output Area to Middle Layer Super Output Area to Local Authority District (February 2018) Lookup in the UK - data.gov.uk](https://data.gov.uk/dataset/9b090605-9861-4bb4-9fa4-6845daa2de9b/postcode-to-output-area-to-lower-layer-super-output-area-to-middle-layer-super-output-area-to-local-authority-district-february-2018-lookup-in-the-uk)