**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BÁO CÁO ĐỒ ÁN**

Dữ liệu phạm tội tại thành phố Chicago

*giai đoạn 2012-2017*



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

*Nhóm 4 thực hiện*

GVHD: Ths Hồ Thị Hoàng Vy

**THÔNG TIN NHÓM**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Mã nhóm*** | ***MSSV*** | ***Họ và tên*** | ***Ghi chú*** |
| 04 | 1753135 | Lý Thanh Long | Nhóm trưởng |
| 1753094 | Vũ Phùng Quang |  |
| 1753089 | Nguyễn Lý Nhật Phương |  |

**Phân công và đánh giá thành viên trong nhóm**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên | Phân công | Đánh giá | Ngày |
| Lý Thanh Long | Thiết kế NDS | Hoàn thành | 08/11/2020 |
| Vũ Phùng Quang | Thiết kế DDS | Hoàn thành | 08/11/2020 |
| Nguyễn Lý Nhật Phương | Chọn lựa thuộc tính cần thiết, để thiết kế NDS, DDS | Hoàn thành | 08/11/2020 |
| Lý Thanh Long | ETL, OLAP | Hoàn thành | 29/12/2020 |
| Vũ Phùng Quang | ETL, mining | Hoàn thành | 29/12/2020 |
| Nguyễn Lý Nhật Phương | Data Cleaning, ETL, Data Mining. | Hoàn thành | 29/12/2020 |
| Lý Thanh Long | Lập lịch định kỳ | Hoàn thành | 5/1/2021 |
| Vũ Phùng Quang | KPI | Hoàn thành | 5/1/2021 |
| Nguyễn Lý Nhật Phương | Mining | Hoàn thành | 5/1/2021 |

Mục lục

[I.](#_heading=h.30j0zll) Giới thiệu tổng quan về đồ án 5

[II.](#_heading=h.1fob9te) Các công cụ thực hiện 11

[III.](#_heading=h.3znysh7) Quy trình thực hiện 11

[1)](#_heading=h.2et92p0) Quy trình thiết kế kho dữ liệu 12

[2)](#_heading=h.tyjcwt) Quy trình xây dựng cube và mining 12

[3)](#_heading=h.3dy6vkm) Quy trình khai thác dữ liệu 14

[IV.](#_heading=h.1t3h5sf) Mô tả dữ liệu 14

[V.](#_heading=h.4d34og8) Thiết kế dữ liệu 17

[1)](#_heading=h.2s8eyo1) Thiết kế NDS 17

[2)](#_heading=h.17dp8vu) Thiết kế DDS 18

[VI.](#_heading=h.3rdcrjn) Nạp dữ liệu: 19

[1)](#_heading=h.26in1rg) Cách thức lưu trữ 19

[2)](#_heading=h.lnxbz9) Chuyển dữ liệu từ file .csv vào source trong SQL Server 19

[3)](#_heading=h.35nkun2) Chuyển dữ liệu từ nguồn vào stage 20

[4)](#_heading=h.1ksv4uv) Ứng dụng tool Openfire trong giai đoạn nhất quán dữ liệu 20

[5)](#_heading=h.44sinio) Làm sạch dữ liệu null 29

[VII.](#_heading=h.z337ya) Quy trình nạp dữ liệu từ Source -> Stage -> NDS -> DDS 30

[VIII.](#_heading=h.3j2qqm3) Lập lịch định kỳ cho ETL bằng cách deploy packages 38

[IX.](#_heading=h.1y810tw) Khai thác dữ liệu 43

[1)](#_heading=h.4i7ojhp) Report: 43

[a)](#_heading=h.2xcytpi) Thống kê sự phân phối của các vụ trộm theo thời gian tháng, năm. Nhận xét khoảng thời gian xảy ra nhiều nhất, ít nhất các vụ trộm… 43

[b)](#_heading=h.1ci93xb) Thống kê trong tất cả các năm/từng năm, các trường hợp trộm cắp mà không bị bắt giữ, hoặc bị bắt giữ 44

[c)](#_heading=h.3whwml4) Thống kê tỷ lệ trộm, tỷ lệ phạm tội khác theo từng địa điểm. Nhận xét 45

[d)](#_heading=h.2bn6wsx) Vẽ biểu đồ thể hiện tỷ lệ phạm tội nội địa 47

[2)](#_heading=h.qsh70q) OLAP 47

[a)](#_heading=h.3as4poj) Thống kê tần suất các loại tội phạm theo từng năm 47

[b)](#_heading=h.1pxezwc) Thống kê tần suất tội phạm theo thời gian và địa điểm 49

[c)](#_heading=h.49x2ik5) Thống kê tần suất theo các loại tội phạm 52

[3)](#_heading=h.2p2csry) Mining 53

[4)](#_heading=h.147n2zr) KPI 59

[5)](#_heading=h.3o7alnk) Kết luận chung 60

[X.](#_heading=h.23ckvvd) Tham khảo 61

# Giới thiệu tổng quan về đồ án

* Đồ án được thực hiện dựa trên bộ dữ liệu tội phạm tại thành phố Chicago từ năm 2012 đến năm 2017. Dữ liệu được trích xuất từ hệ thống Báo cáo và Phân tích Thi hành Luật Công dân của Sở Cảnh sát Chicago.

<https://www.kaggle.com/currie32/crimes-in-chicago>

* Đồng thời kết hợp thêm với dữ liệu điều tra dân số để kết hợp các yếu tố kinh tế xã hội

<https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-inC/kn9c-c2s2>

* Các thuộc tính dùng trong đồ án

Gồm 2 file CSV

Các thuộc tính sử dụng bao gồm: Case Number, Date,IUCR, Description, Arrest , Domestic, Year, Community Area, FBICode, Primary Type, Location Description

* Dữ liệu được lấy trong file Chicago\_Crimes\_2012\_to\_2017.csv (có đính kèm trong folder DataSet) hoặc có thể download tại đây

<https://www.kaggle.com/currie32/crimes-in-chicago?select=Chicago_Crimes_2012_to_2017.csv>

Ngoài ra kết hợp thêm dữ liệu về kinh tế xã hội. Các thuộc tính bao gồm: Community Area Number, Community Area Name, Percentage of housing crowed, Percentage households below poverty, Percentage aged 16+ unemployed, Percentage aged 25+ without high school diploma, percentage aged under 18 or over 64, per captita income, hardship index.

* Dữ liệu được lấy trong file Census\_Data\_\_Selected\_socioeconomic\_indicators\_in\_Chicago\_\_2008\_\_\_2012.csv (có đính kèm trong folder DataSet) hoặc có thể download tại đây

<https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-inC/kn9c-c2s2>

Ý nghĩa của các thuộc tính được mô tả cụ thể ở phần III Mô tả dữ liệu

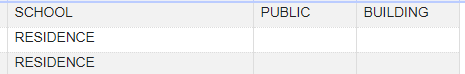
* Trong bộ dữ liệu này có một vài cột bị thiếu dữ liệu với số lượng dữ liệu bị thiếu như sau:
  + Case Number: 1
  + Location Description: 1658
  + District: 1
  + Ward: 14
  + Community Area: 40
  + X Coordinate: 37083
  + Y Coordinate: 37083
  + Latitude: 37083
  + Longitude: 37083
  + Location: 37083
* Một số vấn đề dữ liệu không nhất quán

Dữ liệu trong cột Location Description có những vấn đề sau (dùng công cụ OpenRefine để xem)

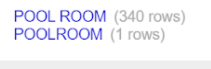
Các thuộc tính đa giá trị



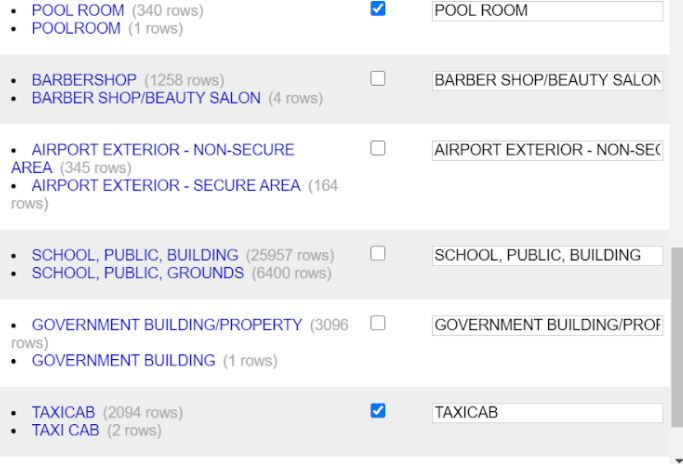
Các dòng có dấu “,” sẽ trở thành như hình



Cùng là pool room nhưng có sự khác biệt giữa 2 dòng dữ liệu



Cùng là ý nghĩa như nhau nhưng được thể hiện khác nhau



Về giải pháp xử lý được mô tả ở phần VI 4) Ứng dụng tool Openfire trong giai đoạn nhất quán dữ liệu

**Thống kê giá trị FBI code và ý nghĩa.**

Lưu trong cơ sở dữ liệu Stage\_ChicagoCrime ở bảng FBICode\_Meaning.

|  |  |  |
| --- | --- | --- |
| **Code** | **Sum** | **Meaning** |
| 01A | 2637 | The killing of one human being by another |
| 01B | 12 | The killing of another person through negligence |
| 02 | 7418 | Any sexual act directed against another person, forcibly and/or against that person's will or not forcibly or against the person's will in instances where the victim is incapable of giving consent |
| 03 | 57313 | The taking or attempting to take anything of value under confrontational circumstances from the control, custody, or care of another person by force or threat of force or violence and/or by putting the victim in fear of immediate harm |
| 04A | 23927 | An unlawful attack by one person upon another wherein the offender displays a weapon in a threatening manner. Placing someone in reasonable apprehension of receiving a battery |
| 04B | 36618 | An unlawful attack by one person upon another wherein the offender uses a weapon or the victim suffers obvious severe or aggravated bodily injury involving apparent broken bones, loss of teeth, possible internal injury, severe laceration, or loss of consciousness |
| 05 | 83397 | The unlawful entry into a building or other structure with the intent to commit a felony or a theft |
| 06 | 329460 | The unlawful taking, carrying, leading, or riding away of property from the possession or constructive possession of another person |
| 07 | 61138 | The theft of a motor vehicle |
| 08A | 68076 | An unlawful physical attack by one person upon another where neither the offender displays a weapon, nor the victim suffers obvious severe or aggravated bodily injury involving apparent broken bones, loss of teeth, possible internal injury, severe laceration, or loss of consciousness |
| 08B | 227082 | A person commits battery if he intentionally or knowingly without legal justification and by any means, (1) causes bodily harm to an individual or (2) makes physical contact of an insulting or provoking nature with an individual |
| 09 | 2200 | To unlawfully and intentionally damage or attempt to damage any real or personal property by fire or incendiary device |
| 10 | 8267 | The altering, copying, or imitation of something, without authority or right, with the intent to deceive or defraud by passing the copy or thing altered or imitated as that which is original or genuine or the selling, buying, or possession of an altered, copied, or imitated thing with the intent to deceive or defraud |
| 11 | 66547 | The intentional perversion of the truth for the purpose of inducing another person or other entity in reliance upon it to part with something of value or to surrender a legal right |
| 12 | 268 | The unlawful misappropriation by an offender to his/her own use or purpose of money, property, or some other thing of value entrusted to his/her care, custody, or control |
| 13 | 413 | Receiving, buying, selling, possessing, concealing, or transporting any property with the knowledge that it has been unlawfully taken, as by Burglary, Embezzlement, Fraud, Larceny, Robbery, etc |
| 14 | 155455 | To willfully or maliciously destroy, damage, deface, or otherwise injure real or personal property without the consent of the owner or the person having custody or control of it |
| 15 | 17326 | The violation of laws or ordinances prohibiting the manufacture, sale, purchase, transportation, possession, concealment, or use of firearms, cutting instruments, explosives, incendiary devices, or other deadly weapons |
| 16 | 7654 | To unlawfully engage in or promote sexual activities for profit |
| 17 | 5912 | The violation of laws prohibiting offenses against chastity, common decency, morals, and the like such as: adultery and fornication; bigamy; indecent exposure; and indecent liberties |
| 18 | 129796 | The violation of laws prohibiting the production, distribution, and/or use of certain controlled substances and the equipment or devices utilized in their preparation and/or use |
| 19 | 2215 | To unlawfully bet or wager money or something else of value; assist, promote, or operate a game of chance for money or some other stake; possess or transmit wagering information; manufacture, sell, purchase, possess, or transport gambling equipment, devices, or goods; or tamper with the outcome of a sporting event or contest to gain a gambling advantage |
| 20 | 6829 | Unlawful, nonviolent acts by a family member (or legal guardian) that threaten the physical, mental, or economic well-being or morals of another family member and that are not classifiable as other offenses, such as Assault, Incest, Statutory Rape, etc |
| 22 | 1953 | The violation of laws or ordinances prohibiting the manufacture, sale, purchase, transportation, possession, or use of alcoholic beverages |
| 24 | 17204 | Any behavior that tends to disturb the public peace or decorum, scandalize the community, or shock the public sense of morality |
| 26 | 137597 | The violation of miscellaneous laws or ordinances |

# Các công cụ thực hiện

* Microsoft SQL management Studio phiên bản 18 gồm Database Engine, Analysis Services
* Visual Studio 19 có cài thêm sql data tool tích hợp(Integration Services, Analysis Service Multidimensional and Data Mining) có thể tham khảo để tải tại đây: <https://www.enhansoft.com/how-do-you-install-sql-server-data-tools/> hoặc <https://docs.microsoft.com/en-us/sql/ssdt/download-sql-server-data-tools-ssdt?view=sql-server-ver15> .

**Lưu ý sử dụng phiên bản Develop**

* Công cụ hỗ trợ xử lý dữ liệu có áp dụng trong đồ án trong quá trình ETL: Open Refine

Có thể tham khảo bài viết để hiểu rõ hơn về công cụ

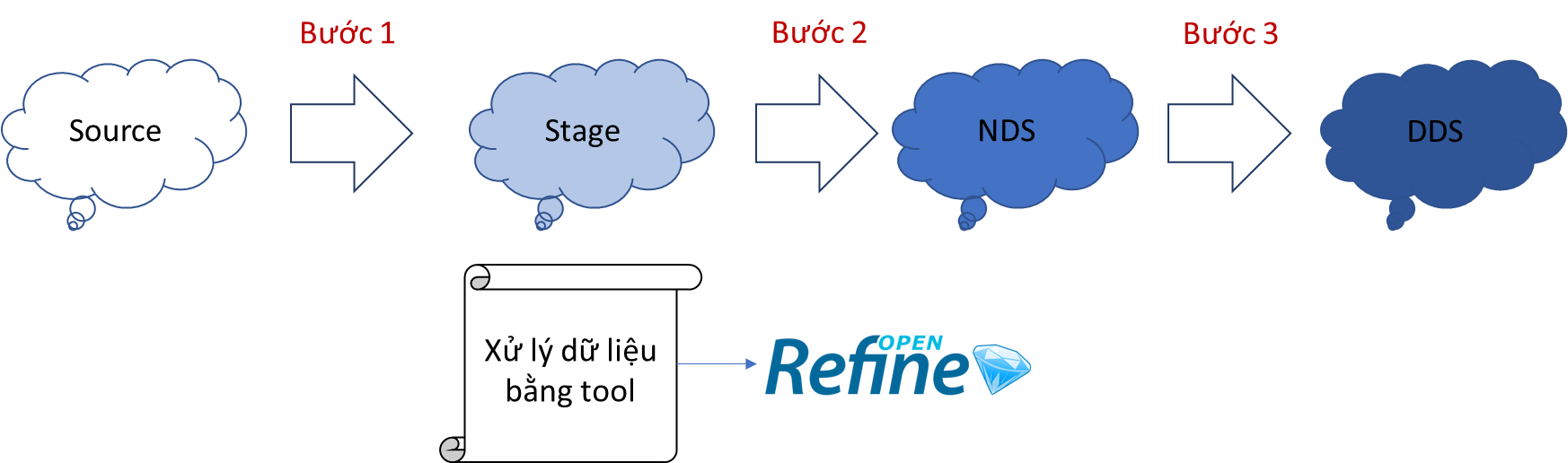
<https://drive.google.com/file/d/1Ic8hEnjfqVJ53Td3wnJOPuk0sl5afylB/view?usp=sharing>.

* Dùng Excel để thể hiện báo cáo , vẽ biểu đồ côt, biểu đồ tròn.

# Quy trình thực hiện

**Phương pháp thực hiện trong quá trình ETL** là kỹ thuật whole table. Ứng với mỗi mỗi giai đoạn nạp dữ liệu vào Stage -> NDS -> DDS sẽ tiến hành truncate dữ liệu trước khi nạp dữ liệu mới vào

## Quy trình thiết kế kho dữ liệu



Ứng với dữ liệu ở souce được mô tả ở phần I

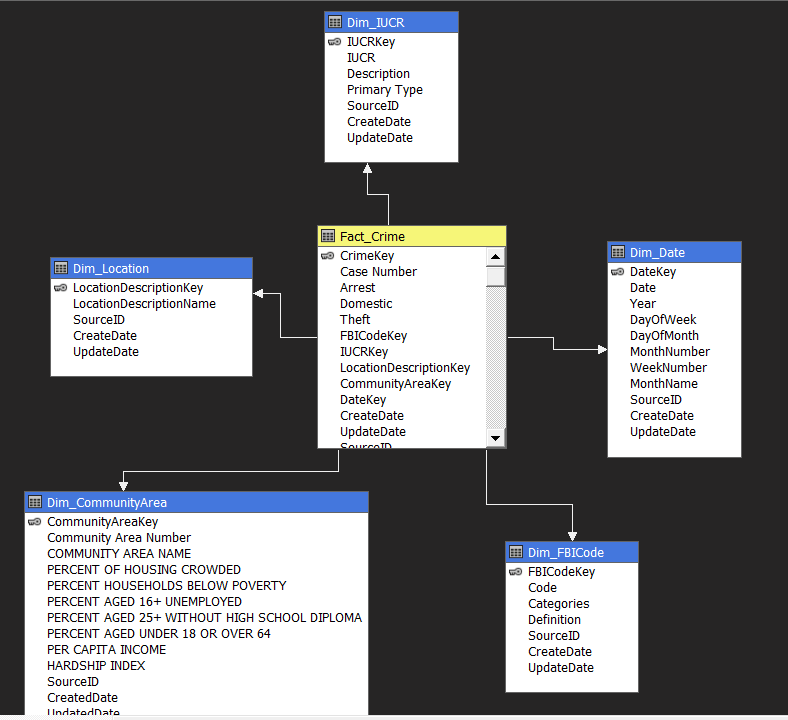
**Bước 1** tiến hành ETL sử dụng công cụ OpenRefine để xử lý dữ liệu để chuyển dữ liệu từ source vào stage.

**Bước 2** chuẩn hóa dữ liệu từ Stage lưu vào NDS.

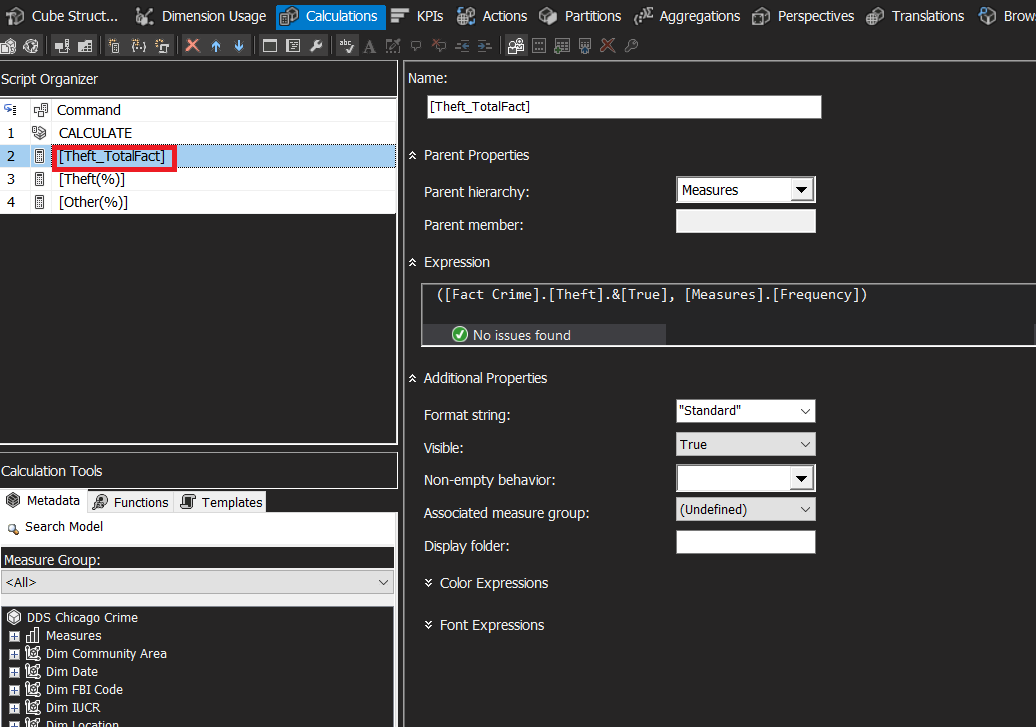
**Bước 3** dữ liệu sẽ được nạp vào DDS với chiều, fact, độ đo tương ứng (phần này sẽ được mô tả rõ hơn ở phần VI, 2 Thiết kế dữ liệu DDS.

## Quy trình xây dựng cube và mining

Sau khi có dữ liệu và lưu trong DDS ta sẽ tiến hành tạo cube



Trong cube ta sẽ tiến hành tạo Calculation và KPI sẽ được trình bày ở phần



## Quy trình khai thác dữ liệu

* Ta sẽ tiến hành tạo view vDM (chuyển các giá trị số của các thuộc tính chỉ số cuộc sống trong bảng community area thành các nhãn ‘high’, ‘medium’, ‘low’) bằng cách thực hiện kết các bảng lại với nhau
* Với hai nhu cầu khai thác dữ liệu
* Dự đoán tội phạm là tội phạm nội địa.
* Dự đoán tội phạm là trộm.

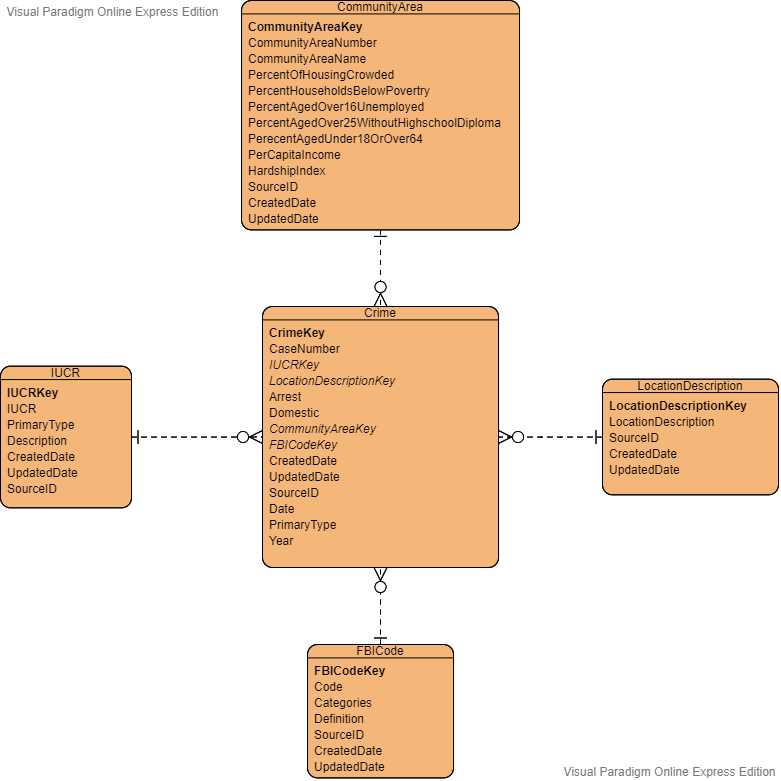
# Mô tả dữ liệu

Bộ dữ liệu trong đề tài bao gồm dữ liệu ghi chép các vụ án xảy ra tại thành phố Chicago từ năm 2012 đến năm 2017, kết hợp với một số bộ dữ liệu khác (dữ liệu mã IUCR, dữ liệu phạm vi tuần tra của các beat cảnh sát (police beat), dữ liệu phạm vi tuần tra của cảnh sát quận (police district), dữ liệu phạm vi của các khu vực (ward), dữ liệu phạm vi của các khu vực cộng đồng (community area), dữ liệu các chỉ số kinh tế xã hội, dữ liệu phân loại phạm tội) để có thêm các dữ liệu chi tiết về phạm vi của các beat, quận, khu vực, khu vực cộng đồng.

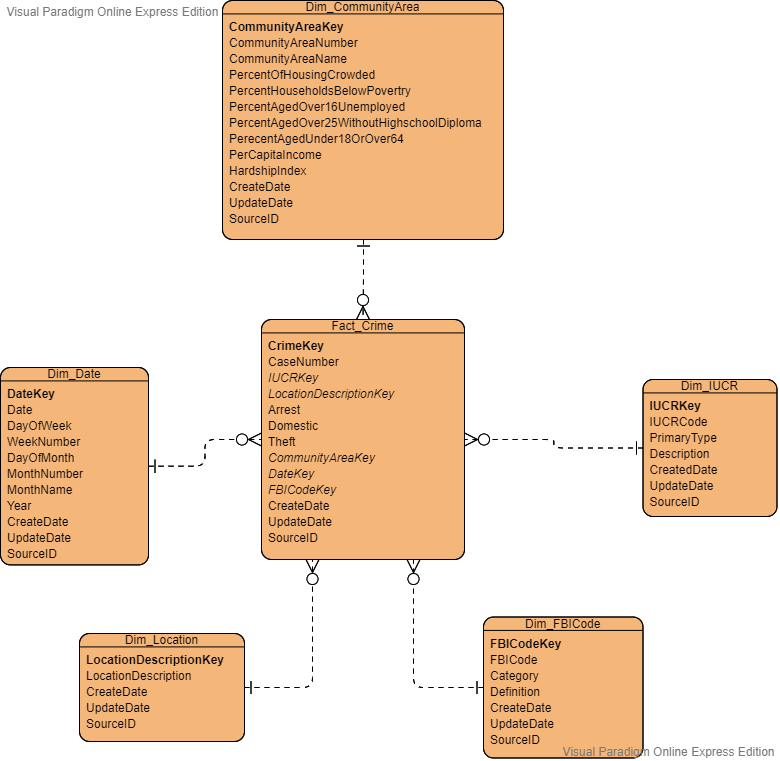
* ID: Định danh duy nhất cho mỗi ghi chéo.
* Case Number(Số hồ sơ): Số RD (Records Division Number) của sở cảnh sát Chicago.
* Date: Ngày xảy ra sự cố (đôi khi lấy ước tính tốt nhất)
* Block: Địa chỉ nơi xảy ra sự cố được biên tập lại một phần, nằm trên cùng khu nhà với địa chỉ thực .
* IUCR: Mã báo cáo tội phạm thống nhất của Illinois
* Primary Type: Mô tả chính cho mã IUCR.
* Description: Mô tả phụ cho mã IUCR, một danh mục phụ của mô tả chính.
* Location Description: Mô tả vị trí nơi xảy ra sự cố.
* Arrest (bắt giữ): Cho biết liệu đã bắt được người gây án hay chưa.
* Domestic (Gia đình): Cho biết liệu sự việc có liên quan đến gia đình theo định nghĩa của đạo luật bạo lực gia đình Illinois .
* Beat: Cho biết beat nơi xảy ra sự cố (beat là khu vực địa lý cảnh sát nhỏ nhất).
* District: Cho biết quận xảy ra sự cố.
* Ward: Cho biết khu vực nơi xảy ra sự cố.
* Community Area: Cho biết khu vực cộng đồng nơi xảy ra sự cố
* FBI Code: Cho biết phân loại phạm tội như được nêu trong hệ thống báo cáo của FBI (National Incident-Based Reporting System).
* Community Area Name: Tên khu vực cộng đồng (cần tích hợp thêm vào)
* FBI Code Description: Mô tả cho FBI Code (cần tích hợp thêm vào)
* X Coordinate: Tọa độ x của vị trí xảy ra sự cố theo phép chiếu State Plane Illinois East NAD 1983. Vị trí này được thay đổi từ vị trí thực tế để xử lý lại một phần nhưng vẫn nằm trên cùng một khu nhà.
* Y Coordinate: Tọa độ y của vị trí xảy ra sự cố theo phép chiếu State Plane Illinois East NAD 1983. Vị trí này được thay đổi từ vị trí thực tế để xử lý lại một phần nhưng vẫn nằm trên cùng một khu nhà.
* Year: Năm xảy ra sự cố.
* Updated On: Ngày và giờ được cập nhật lần cuối.
* Latitude: Vĩ độ của vị trí xảy ra sự cố. Vị trí này được thay đổi từ vị trí thực tế để xử lý lại một phần nhưng vẫn nằm trên cùng một khối giống nhau.
* Longitude: Kinh độ của vị trí xảy ra sự cố. Vị trí này được thay đổi từ vị trí thực tế để xử lý lại một phần nhưng nằm trên cùng một khối giống nhau.
* Location: Vị trí nơi xảy ra sự cố ở định dạng cho phép tạo bản đồ và các hoạt động địa lý khác nhau trên cổng dữ liệu. Vị trí này được thay đổi từ vị trí thực tế để xử lý lại một phần nhưng nằm trên cùng một khối giống nhau.

# Thiết kế dữ liệu

## Thiết kế NDS



## Thiết kế DDS



Với thiết kế DDS sẽ bao gồm 5 chiều: Dim\_Date, Dim\_FBICode, Dim\_IUCR, Dim\_Location, Dim\_CommunityArea và 1 fact là Fact\_Crime

Mức độ chi tiết dữ liệu trong bảng Dim\_Date

DayofMonth -> DayofWeek -> MonthNumber-> WeekNumber-> Year.

Ở bảng Fact\_Crime có thêm cột Theft là thuộc tính suy diễn kiểu bool từ bảng FBICode. Nếu FBICode là 8 hoặc 9 thì tương ứng True là trộm ngược lại False không phải trộm.

# Nạp dữ liệu:

## Cách thức lưu trữ

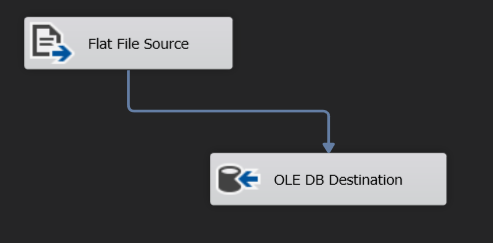
Load dữ liệu từ các file csv (Census\_Data\_-\_Selected\_socioeconomic\_indicators\_in\_Chicago\_\_2008\_\_\_2012.csv, Chicago\_Crimes\_2012\_to\_2017.csv, FBI-Categories-Code.csv) lưu vào cơ sở dữ liệu Chicago\_Crime.

Sau đó tạo một cơ sở dữ liệu Stage\_ChicagoCrime chứa các bảng Census\_Stage, Crime\_Stage, FBICode\_Stage chuyển dữ liệu từ Soure ( cơ sở dữ liệu Chicago\_Crime sang Stage).

Ngoài ra bảng DataNull chứa dữ liệu các cột có giá trị NULL (bị thiếu).

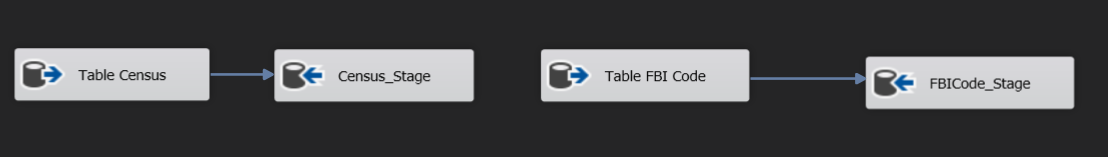
Tạo thêm 2 cơ sở dữ liệu là DDS\_ChicagoCrime và NDS\_ChicagoCrime.

## Chuyển dữ liệu từ file .csv vào source trong SQL Server



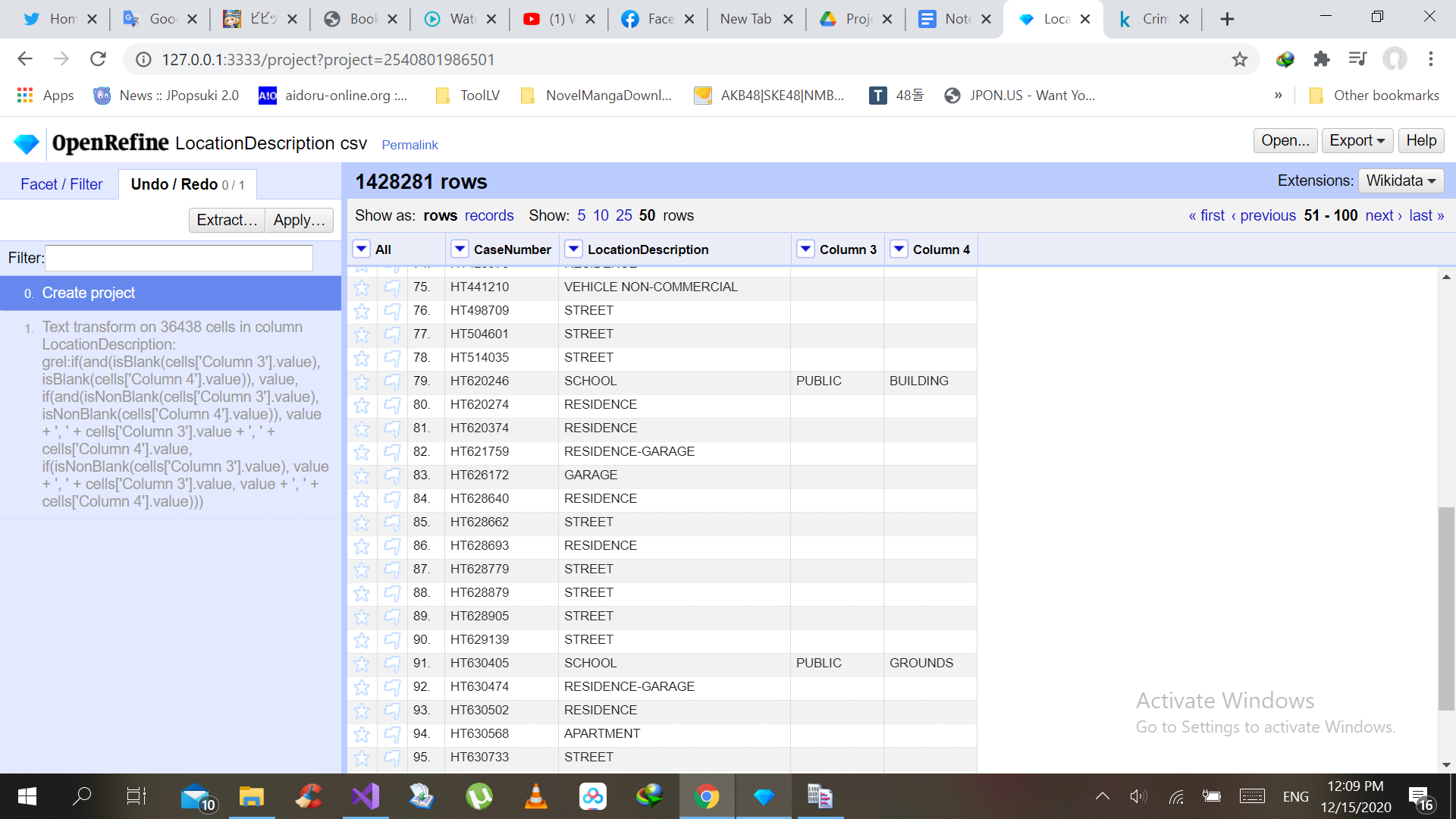
## Chuyển dữ liệu từ nguồn vào stage

Chuyển dữ liệu các bảng Census , FBI Code vào stage



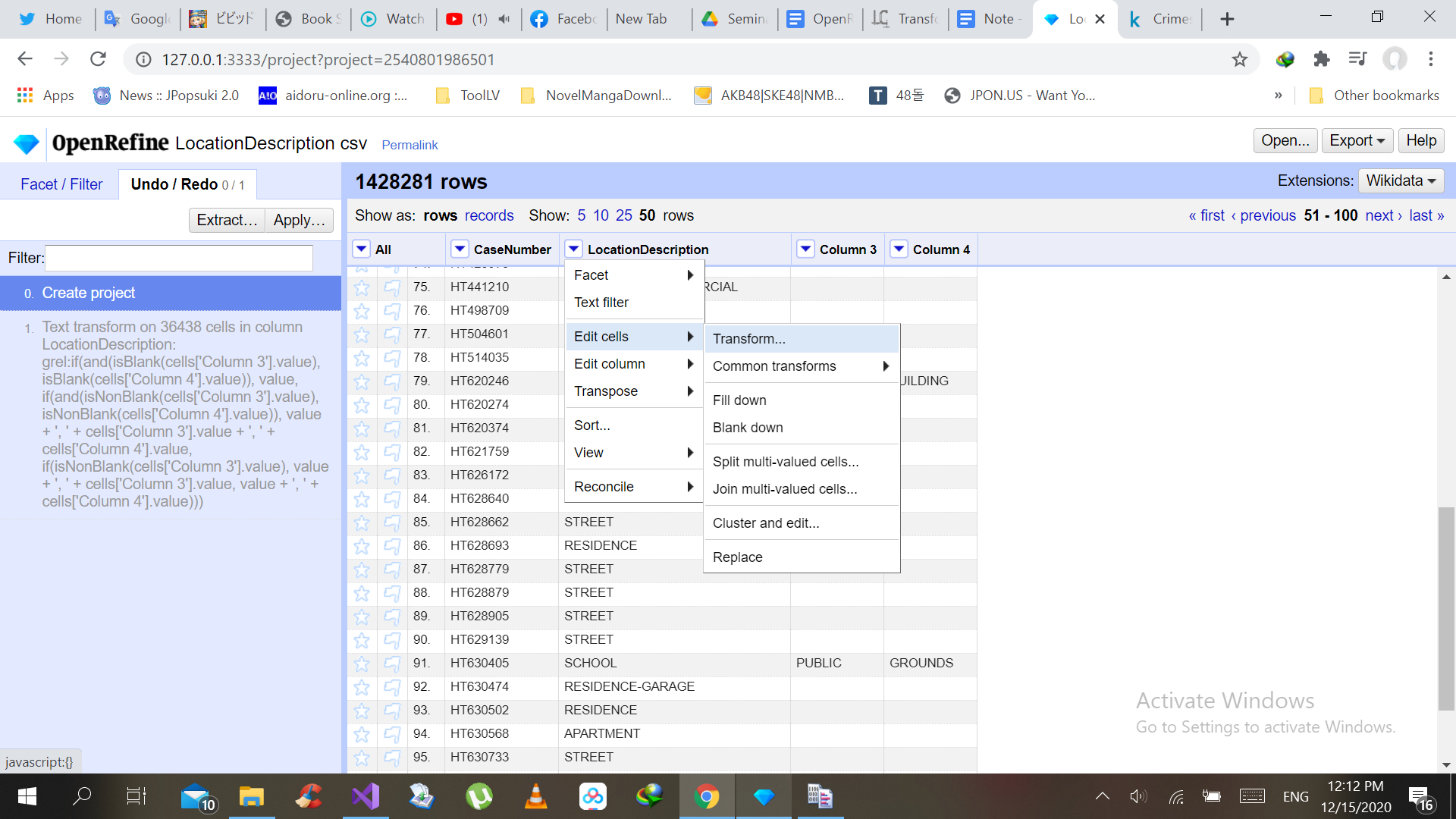
## Ứng dụng tool Openfire trong giai đoạn nhất quán dữ liệu

* Ở cột Location Description khi đọc file: Đọc dấu “,” như một phân cách cho cột mới.



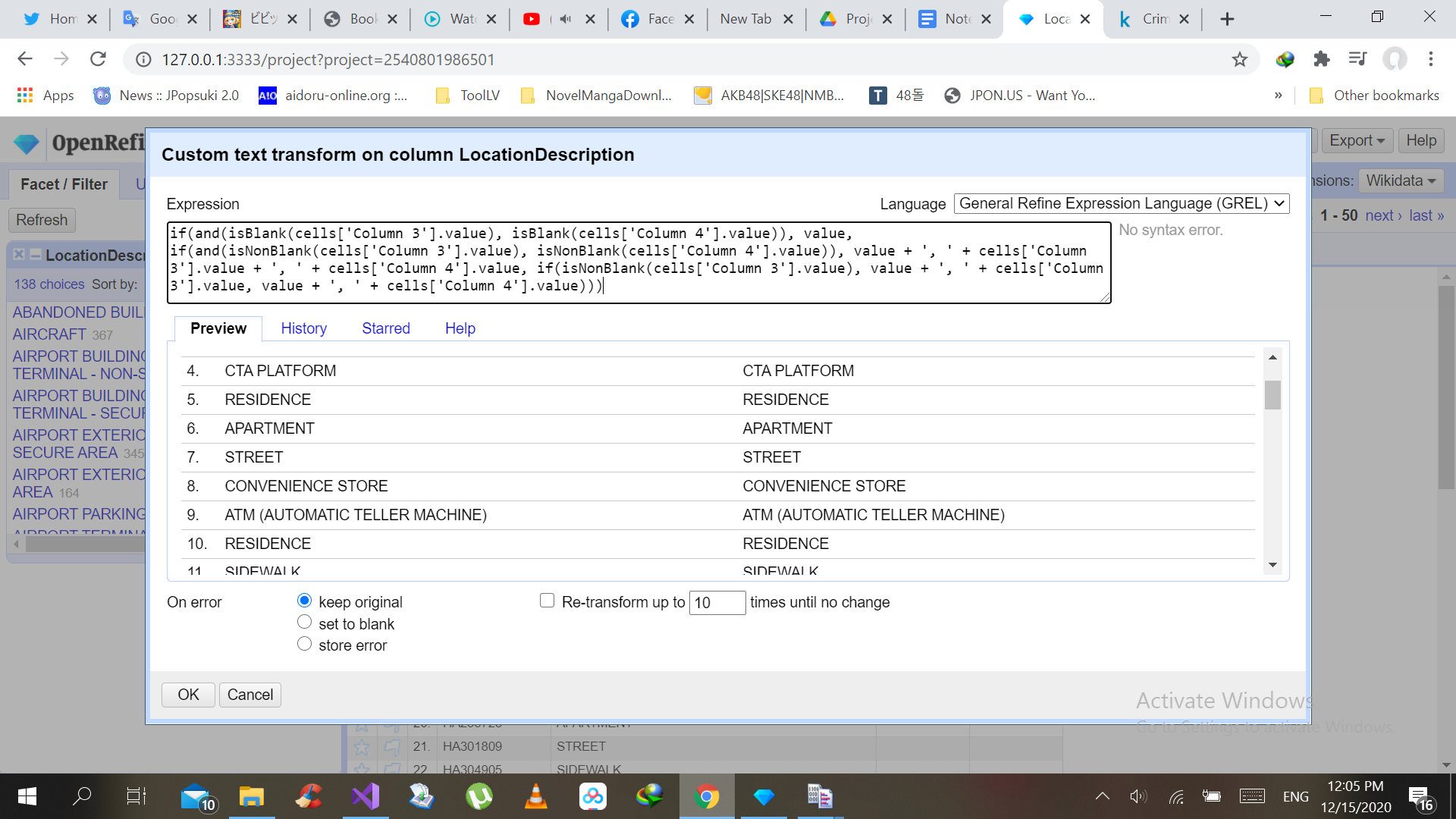
Giải pháp:

**Vào Edit cells -> Transform… của cột LocationDescription**

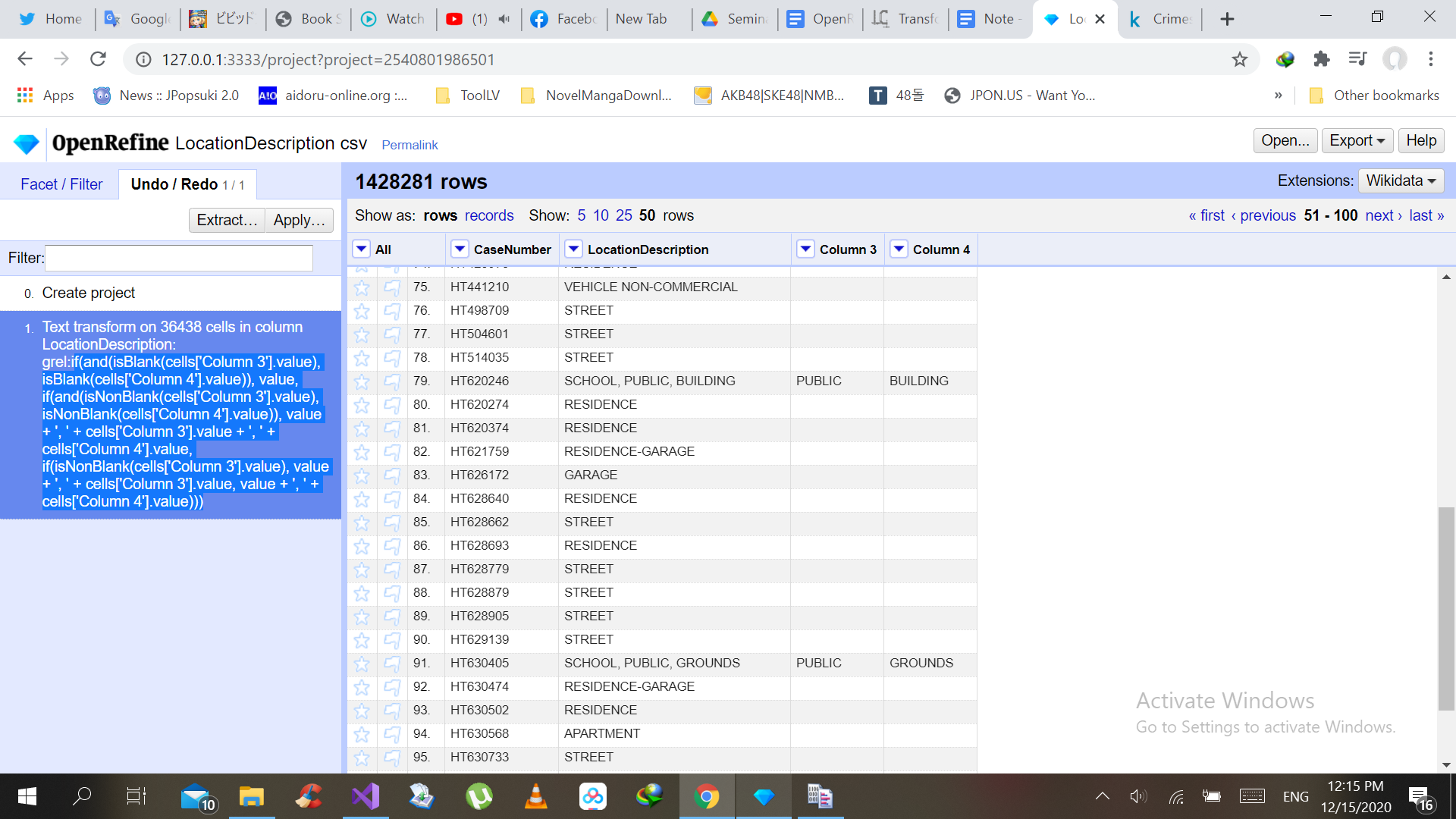


**Dùng câu lệnh để chuyển đổi như bên dưới**

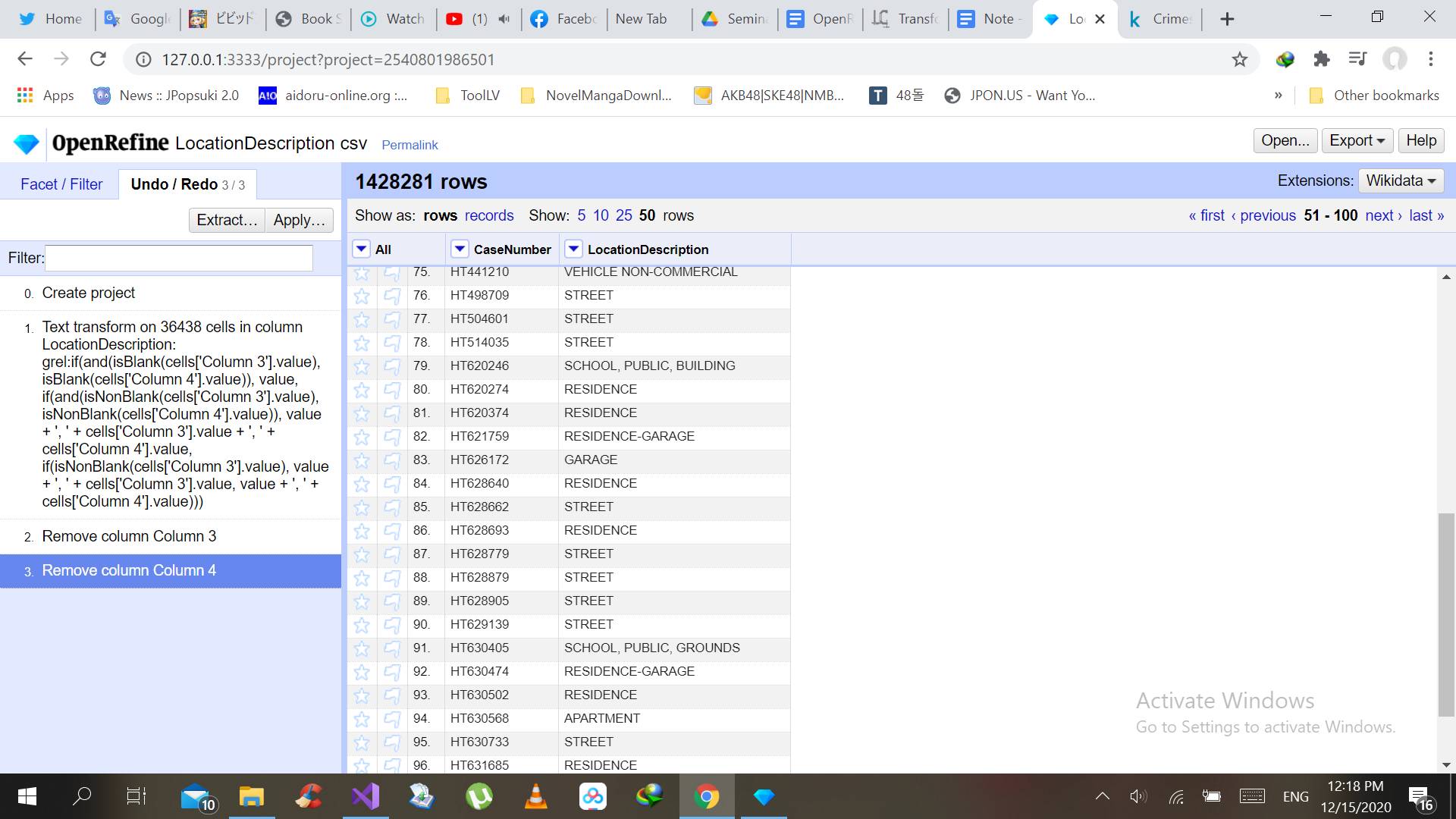
if(and(isBlank(cells['Column 3'].value), isBlank(cells['Column 4'].value)), value, if(and(isNonBlank(cells['Column 3'].value), isNonBlank(cells['Column 4'].value)), value + ', ' + cells['Column 3'].value + ', ' + cells['Column 4'].value, if(isNonBlank(cells['Column 3'].value), value + ', ' + cells['Column 3'].value, value + ', ' + cells['Column 4'].value)))



Các dòng có dấu “,” sẽ trở thành như hình

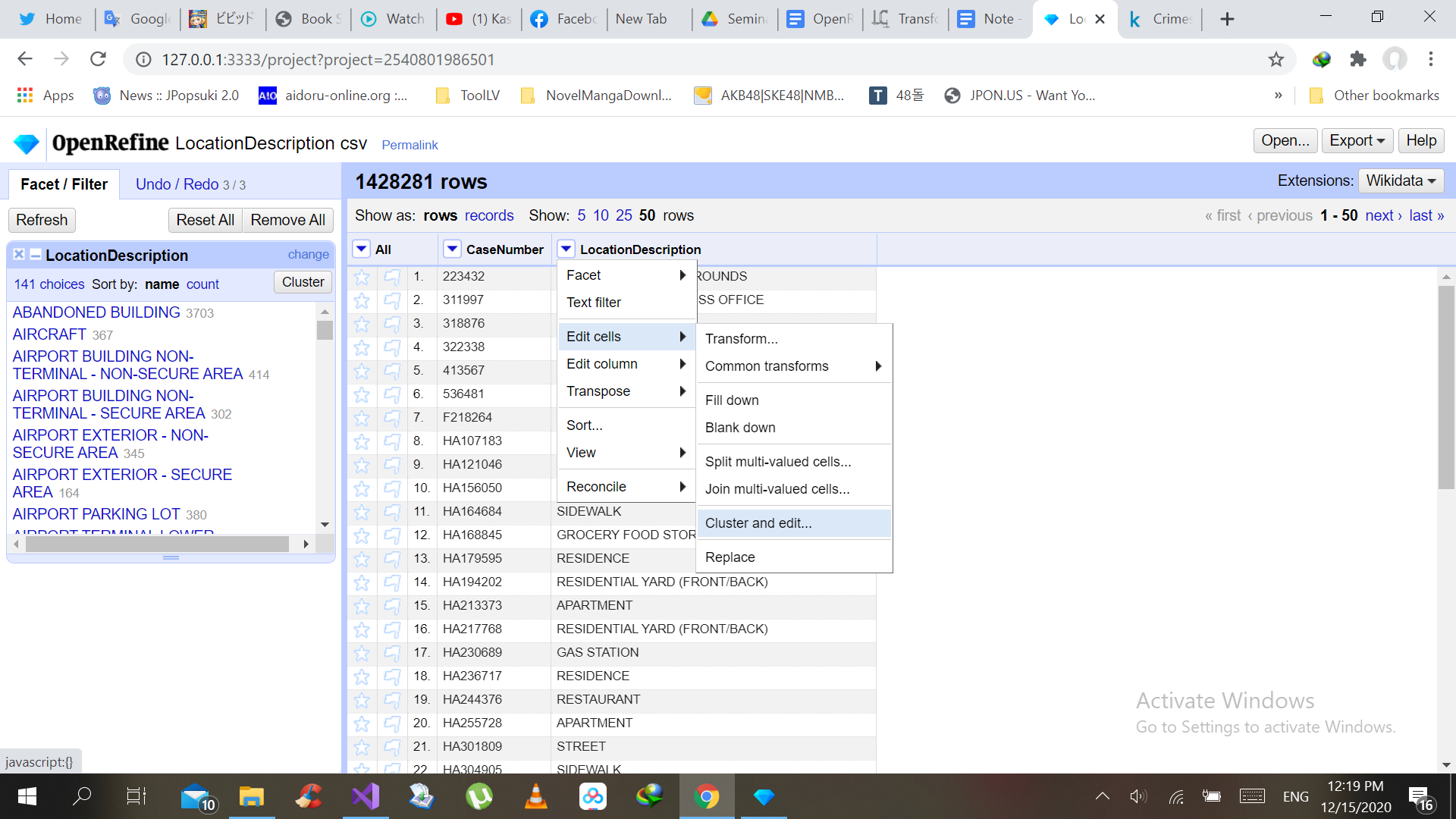


Xóa bỏ đi hai cột Column 3 và Column 4



Gộp những thuộc tính trong LocationDescription

Vào Edit cells -> Cluster and edit



**Các dòng dữ liệu sau đây sẽ được gộp lại với nhau**

POOL ROOM

POOLROOM **----->gộp**

TAXI CAB

TAXICAB **----->gộp**

GOVERMENT BUIDING

GOVERMENT BUIDING /PROPERTY  **----->không gộp**

**Lý do**: Bên mỹ property kiểu đất hay tài sản ấy, còn building thì chỉ là công trình hoặc nhà không bao gồm đất. Cái này là hành văn của người Mĩ rồi

BARERSHOP

BARERSHOP / BEATUTY SALON **----->gộp lấy  babershop/ beauty salon**

**Lý do** giống nhau

HOSPITAL

HOSPITAL  BUIDING/GROUND **-----> không gộp** chưa tìm ra lý do

HOTEL,

HOTEL /MOTEL với

MOTEL **-----> không gộp**

**Lý do** mỗi cái có ý nghĩa riêng, motel là ks tình yêu

VACANT LOT

VACANT LOT /land **----->gộp lấy vacant lot/land**

**Lý do** vacant lot/land phủ cái vacant lot

CHA PARKING LOT

CHA PARKING LOT/GROUND **----->gộp lấy CHA PARKING LOT/GROUND**

CHURCH PROPERTY

CHURCH/SYNAGOGUE/PLACE OF WORSHIP **-----> không gộp**  
  
DELIVERY TRUCK

TRUCK                     **-----> không gộp**

**Lý do**: Delivery chỉ rõ là xe tài giao đồ, Còn truck không thì rộng quá.

GAS STATION

GAS STATION DRIVE/PROP **-----> không gộp**

CLEANERS/LAUNDROMAT,

CLEANING STORE,

LAUNDRY ROOM **-----> không gộp**

**Lý do**:  Cả 3 đều là phòng giặt nhưng khác ngữ cảnh

NURSING HOME/RETIREMENT HOME

NURSING HOME  **----->gộp lấy NURSING HOME/RETIREMENT HOME**

PARK PROPERTY,

PARKING LOT,

 PARKING LOT/GARAGE(NON.RESID.),

POLICE FACILITY/VEH PARKING LOT **-----> không gộp**

**Lý do:** Cái thứ 3 bao gồm cả gara không dành cho dân cư

RESIDENCE PORCH/HALLWAY với

PORCH **-----> không gộp**

SMALL RETAIL STORE với

RETAIL STORE **-----> không gộp**

**Lý do**: Small thì giống cửa hàng gia đình chăng, Còn retail thì lớn hơn chút

TAVERN/LIQUOR STORE,

 TAVERN,

 LIQUOR STORE **-----> không gộp**

**Lý do**: Cái TaVERN/LiQUOR chiếm đa số, Mấy cái kia nhỏ quá nên gộp ... chắc không sao

PUBLIC HIGH SCHOOL 1,

 SCHOOL YARD 2,

 [SCHOOL, PRIVATE, BUILDING] 3057,

[SCHOOL, PRIVATE, GROUNDS] 1024,

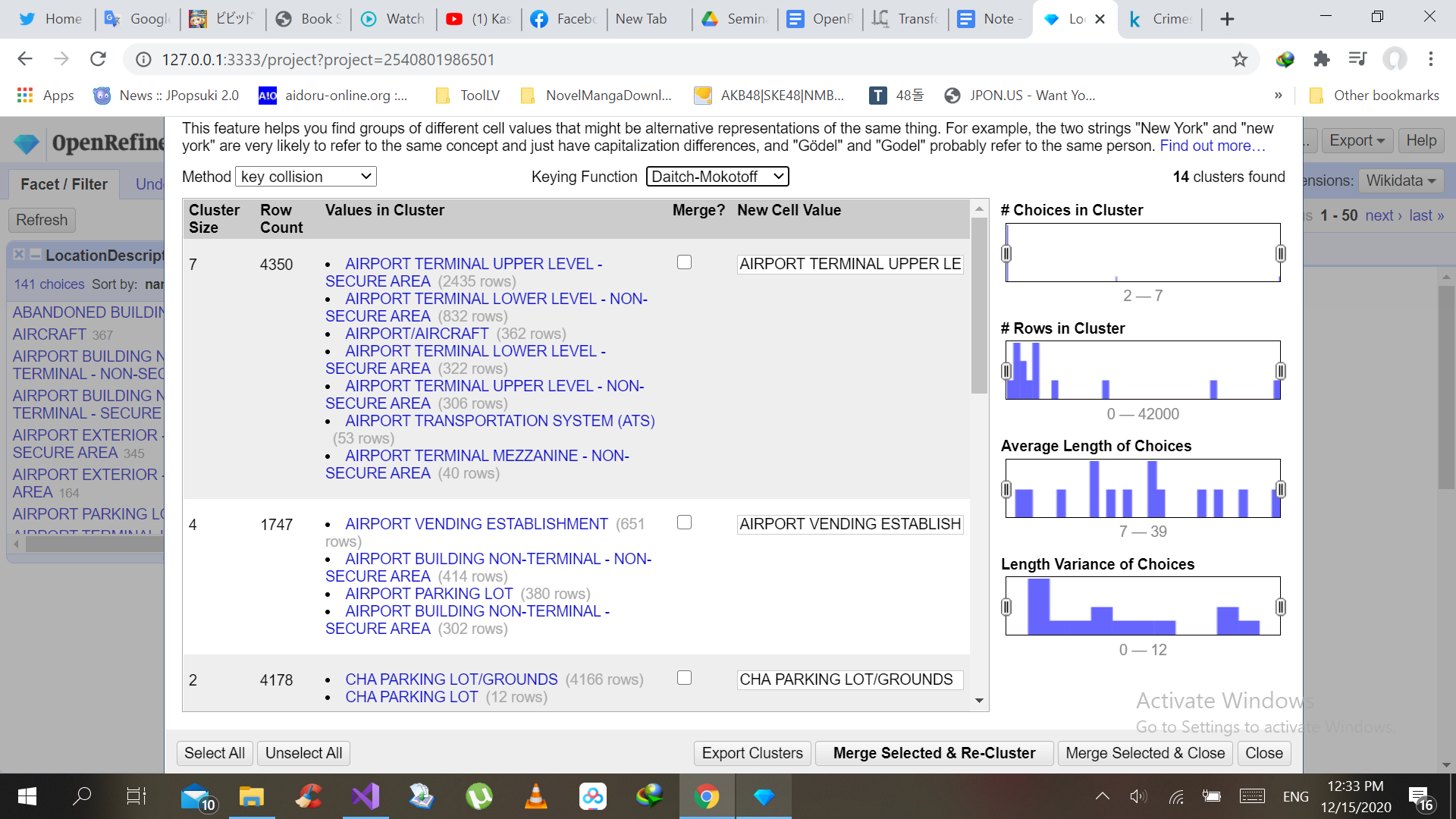
[SCHOOL, PUBLIC, BUILDING] 25957,

[SCHOOL, PUBLIC, GROUNDS] 6400

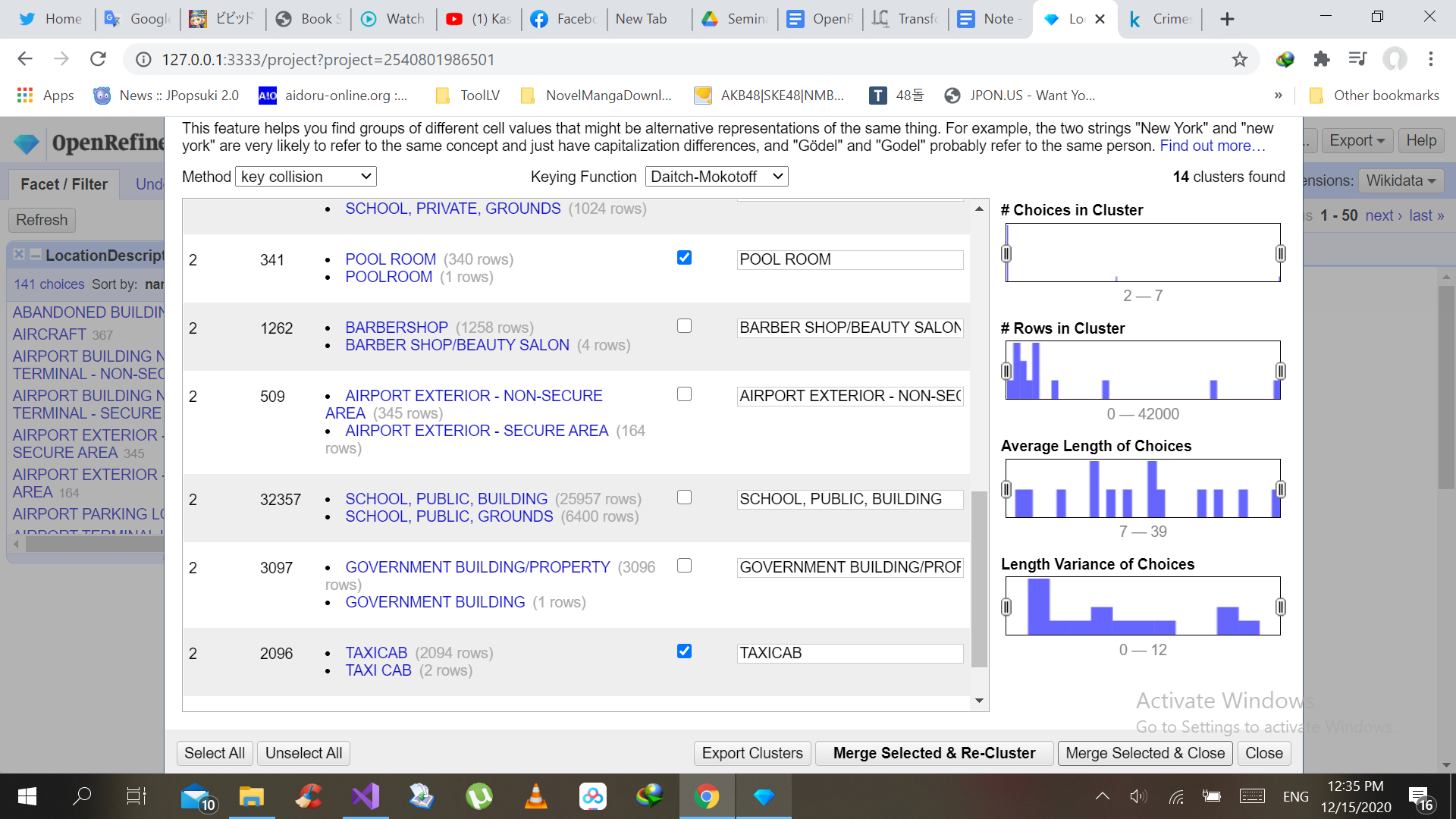
**-----> không gộp**

**Các lý do gộp hay không gộp ở đây là giải pháp của nhóm đề ra.Được tìm hiểu trên các bài báo, các nguồn tham khảo.**

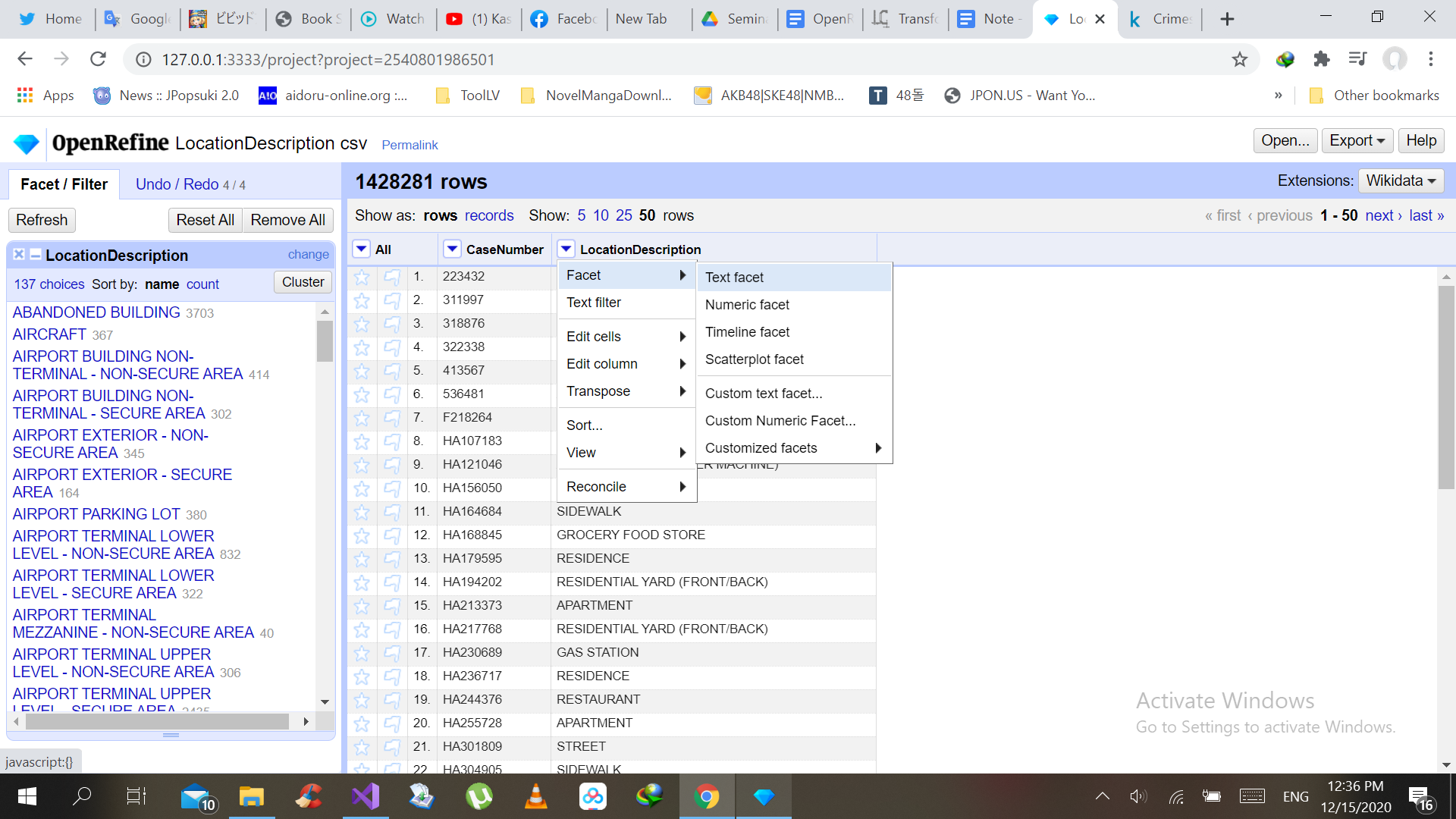
Sau khi quyết gộp hay không gộp dữ liệu ta sẽ chọn Method key collision và Key Function Daitch-Mokotoff.

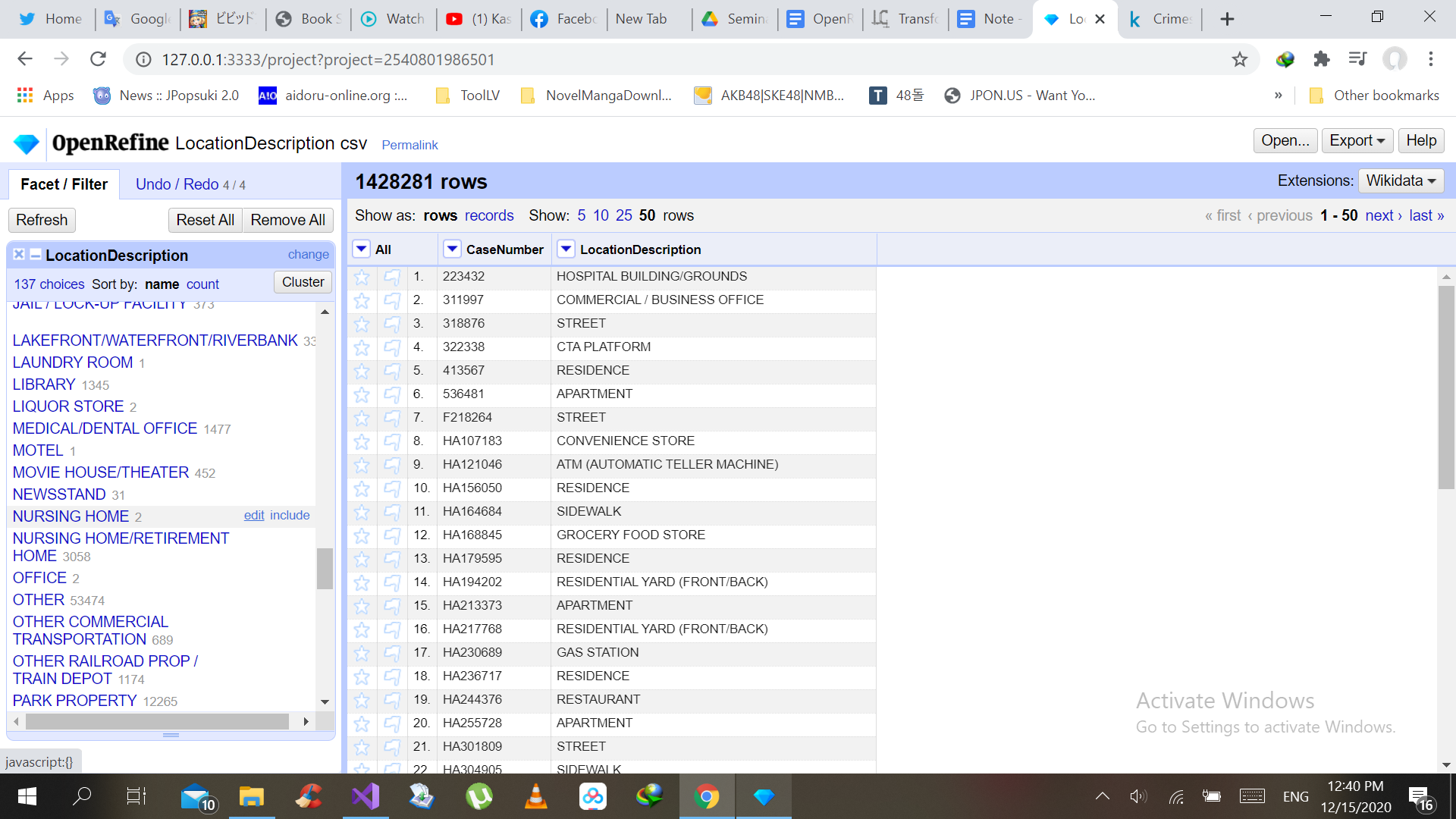


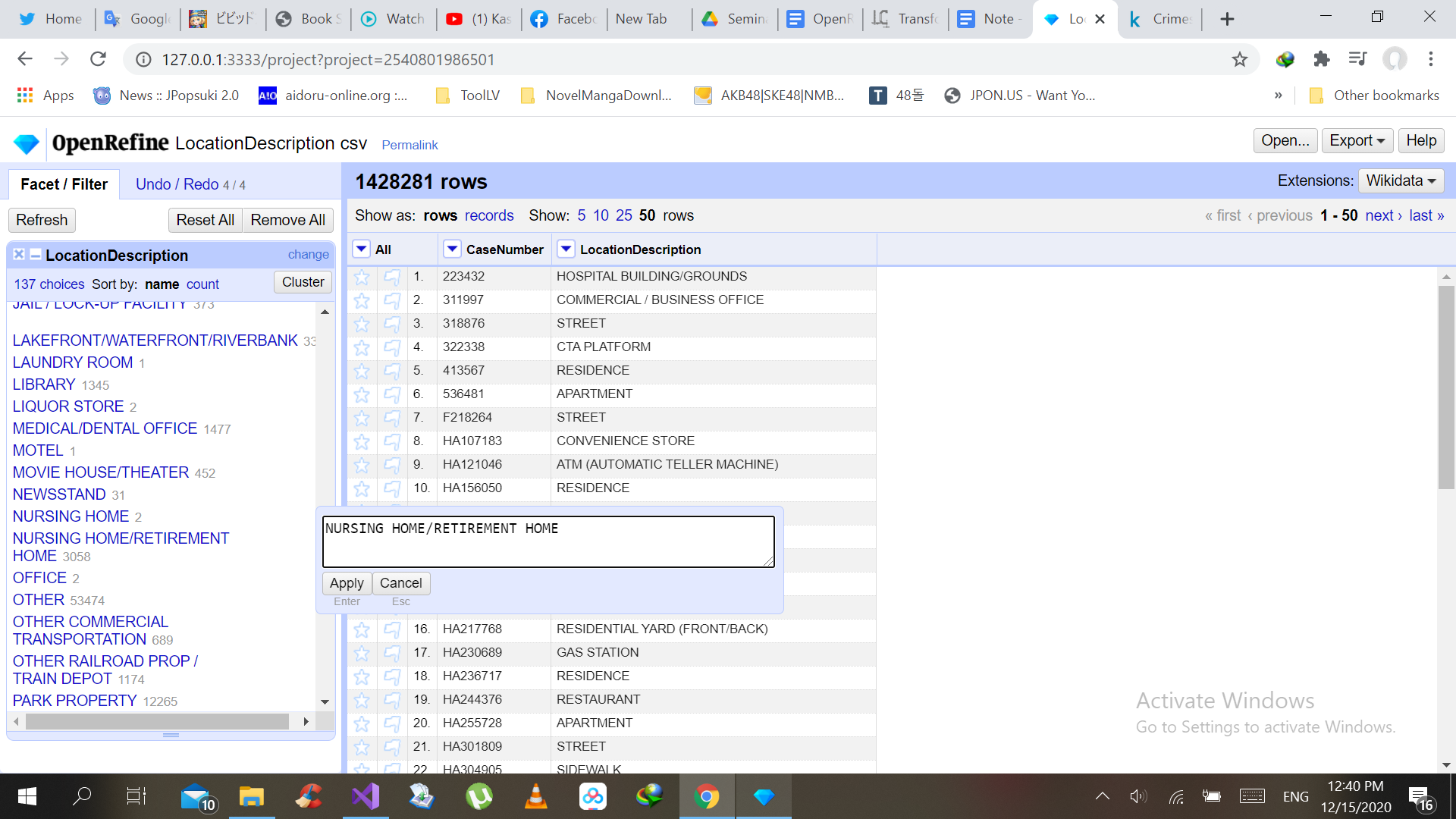
Lựa chọn các cụm muốn gộp lại với nhau và xác định giá trị sau khi gộp và chọn Merge Selected & Close.

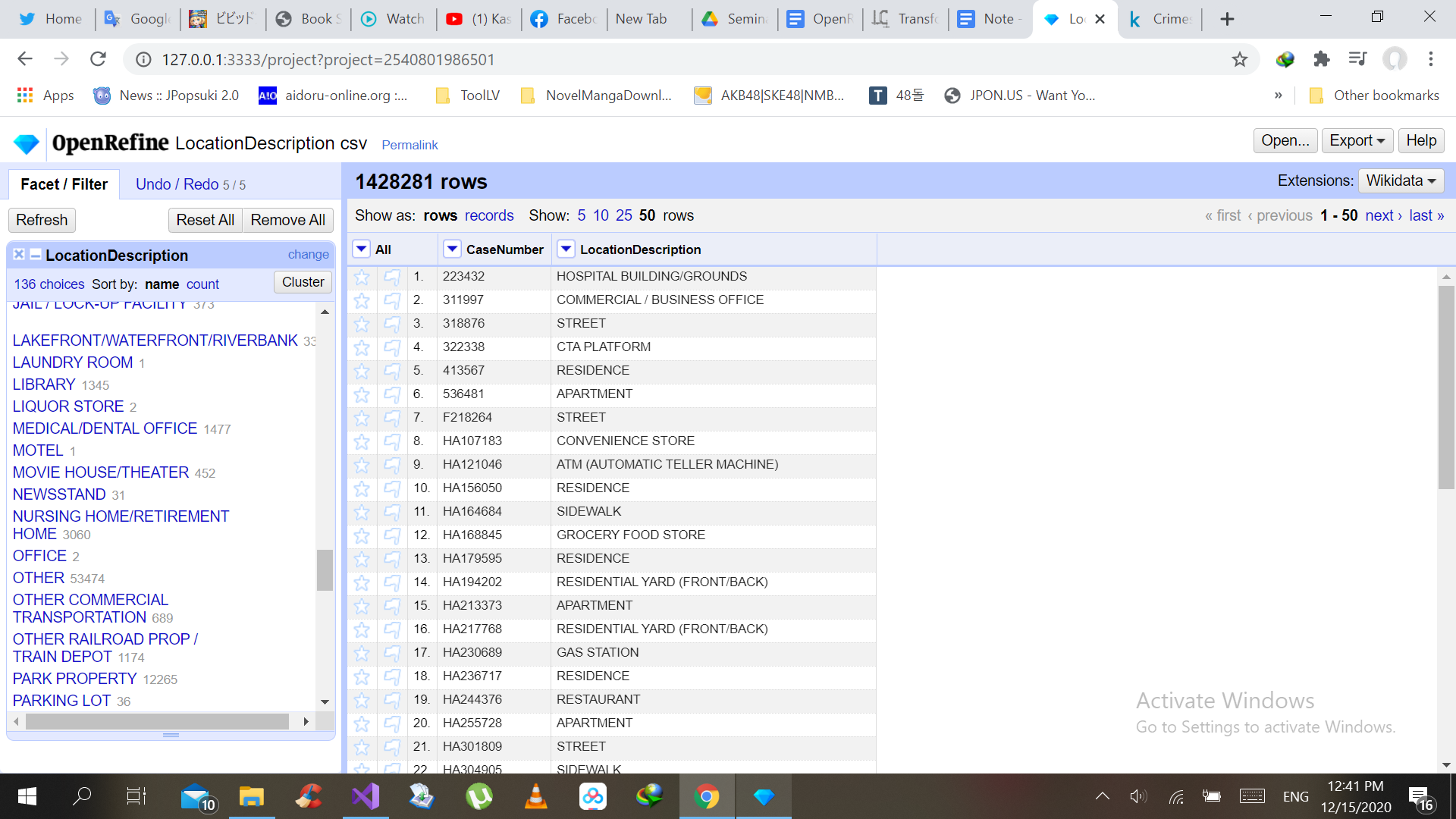


Có thể gộp thủ công bằng cách sử dụng Facet -> Text facet tại cột Location Description

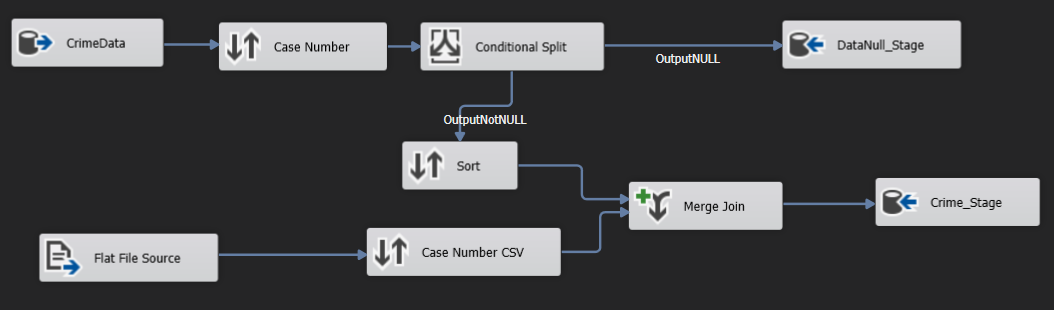








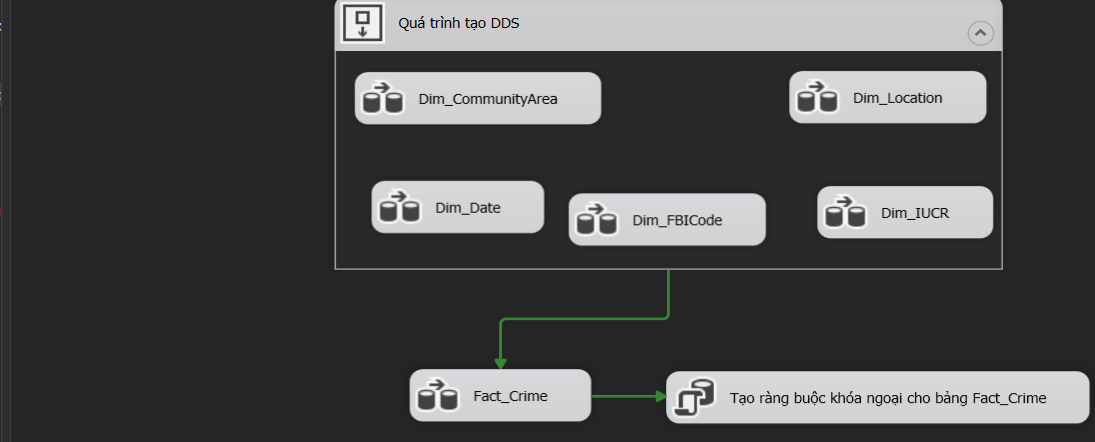
## Làm sạch dữ liệu null



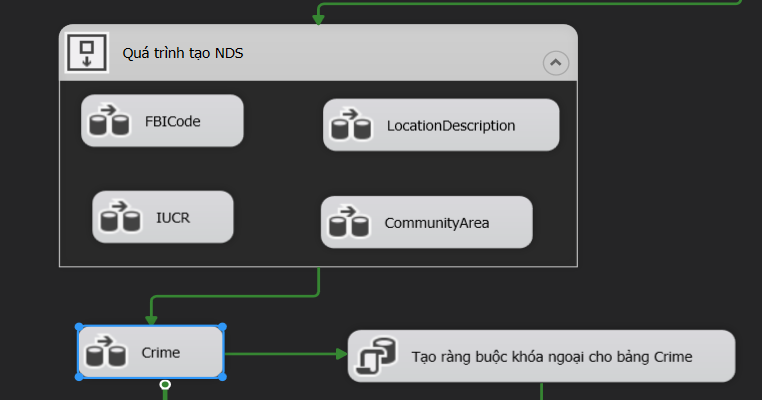
Đầu tiên ta sẽ load dữ liệu từ nguồn (cơ sở dữ liệu Chicago\_Crime), sau đó ta sẽ sort để loại bỏ các thuộc tính trong Case Number mà trùng lắp dữ liệu. Tiếp đó ta sẽ tách dữ giá trị bị thiếu (có giá trị NULL) vào bảng DataNull\_Stage (trong cơ sở dữ liệu Stage\_ChicagoCrime).Với các giá trị khác NULL ta sẽ tiến hành sort các giá trị theo thứ thự tăng dần. Sau đó ta load file CSV đã được xử lý ở bên trên và merge join lại với nhau và lưu vào Crime\_Stage.

# Quy trình nạp dữ liệu từ Source -> Stage -> NDS -> DDS

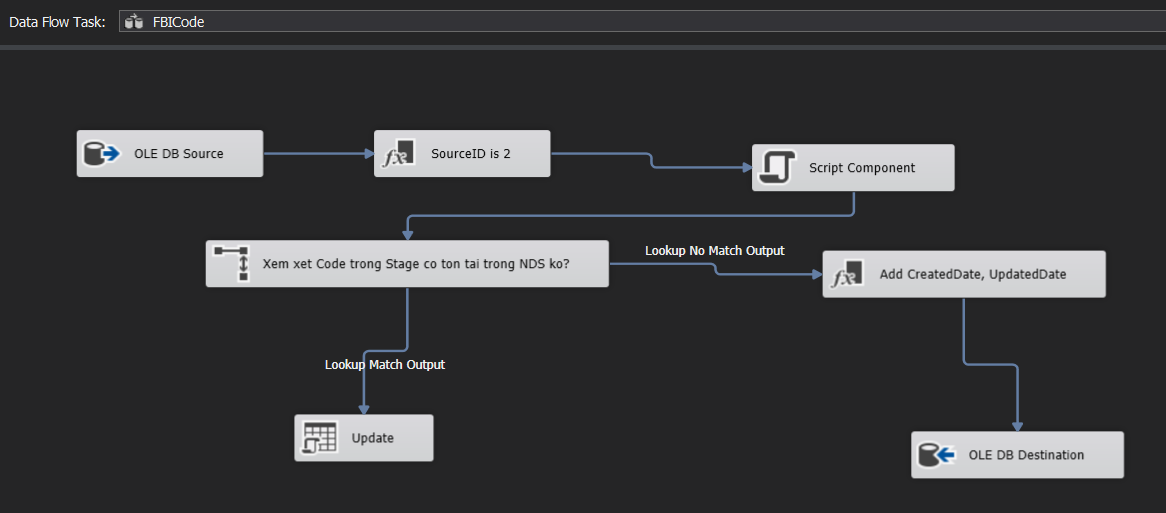




**Tại quá trình nạp dữ liệu vào NDS**

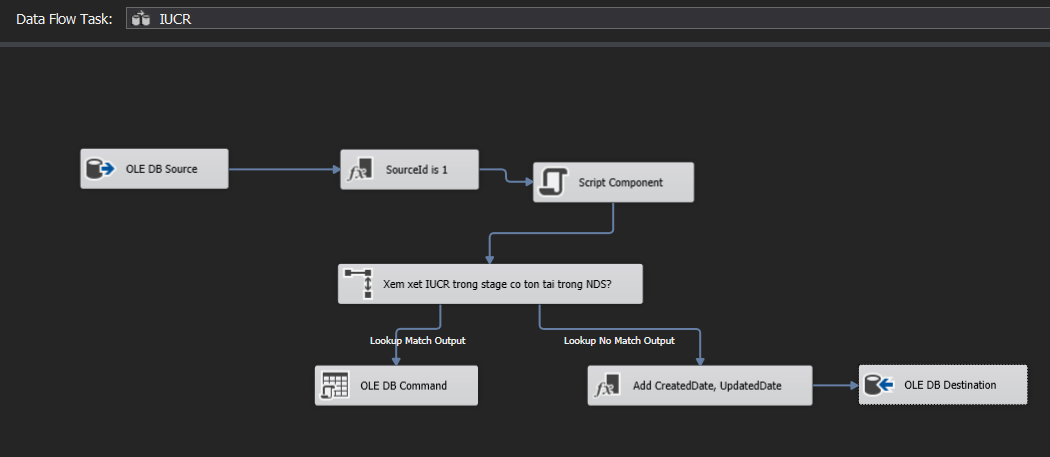


Ta sẽ nạp các chiều (gồm có FBICode, Community Area, IUCR, Location Description) vào trong Sequence Container. Ở mỗi Data Flow Task ta sẽ thực hiện chèn dữ liệu vào NDS như sau:

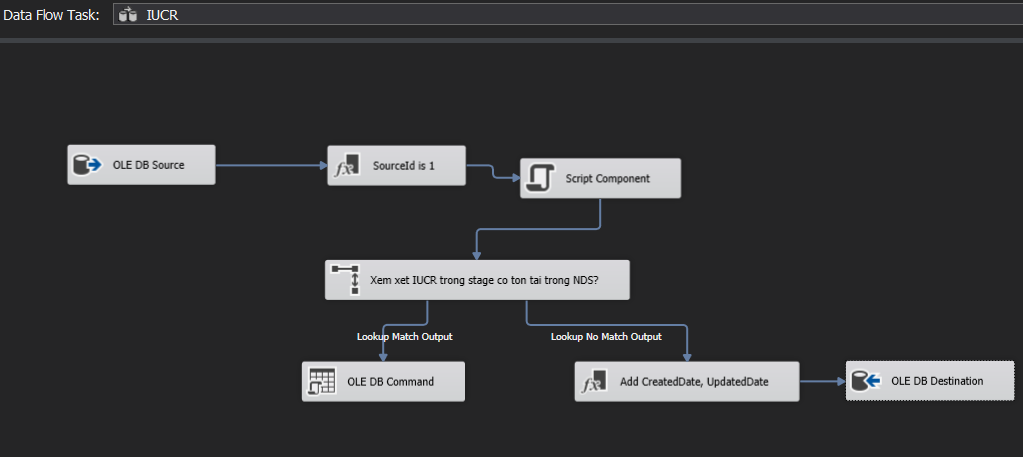


Tại bảng chiều FBI Code ta sẽ thêm một ghi nguồn được quy định trong cơ sở dữ liệu meta data trong bảng data flow. Tiếp theo ta sẽ thêm 1 cột chứa giá trị tự tăng trong component Script Component.Sau đó ta sẽ kiểm tra ứng Code có tồn tại trong NDS chưa. Nếu chưa tồn tại thì ta chèn vào bảng FBICode trong NDS. Nếu đã tồn tại thì sẽ tiến hành cập nhật nếu dữ liệu có thay đổi.

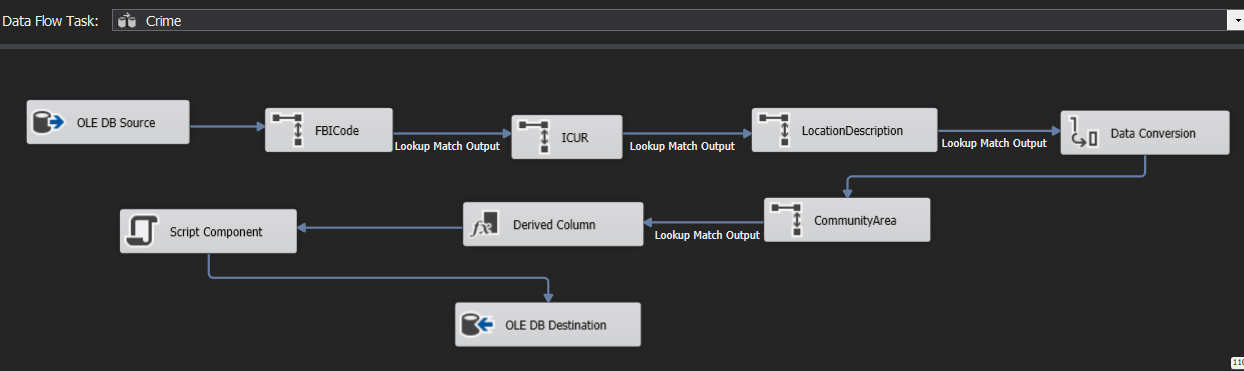
Làm tương tự cho các bảng chiều.





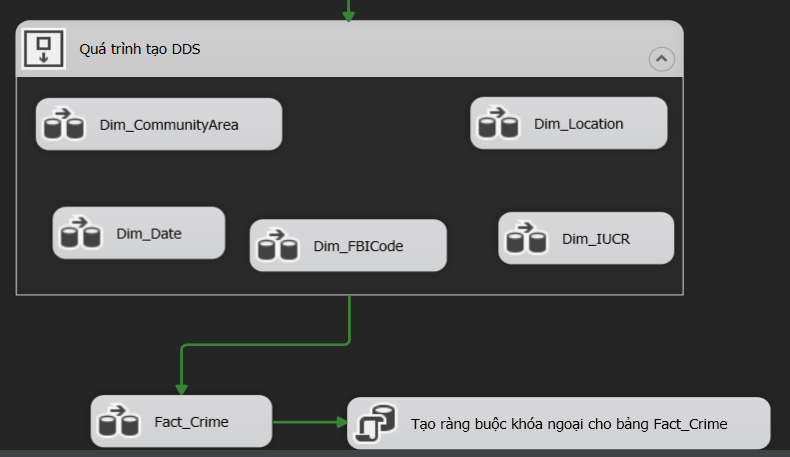


Khi các bảng chiều đã được nạp dữ liệu, ta sẽ nạp dữ liệu vào bảng fact (bảng Crime trong NDS).



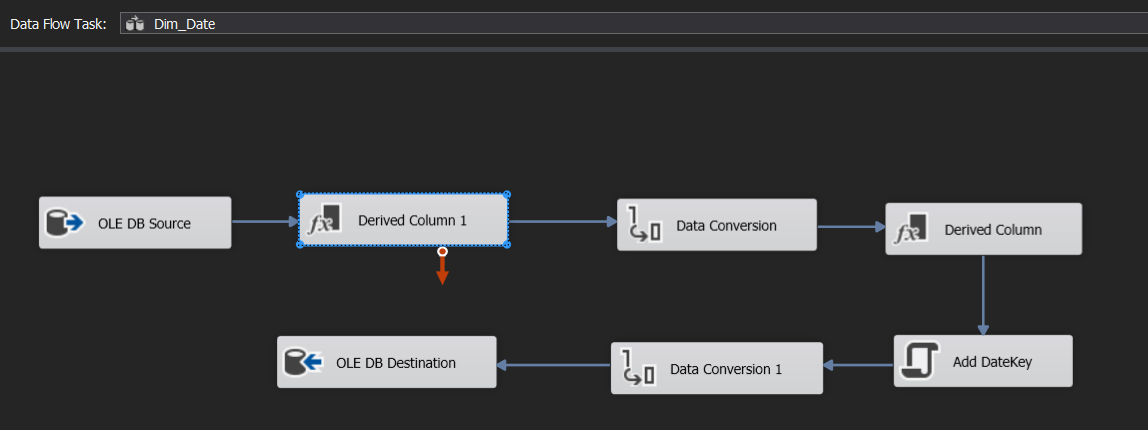
Ta sẽ nạp dữ liệu từ stage, sau đó sẽ lookup trên các bảng chiều để tham chiếu khóa ngoại. Cuối cùng sẽ tạo ràng buộc khóa ngoại cho bảng fact.

**Tại quá trình nạp dữ liệu vào DDS**

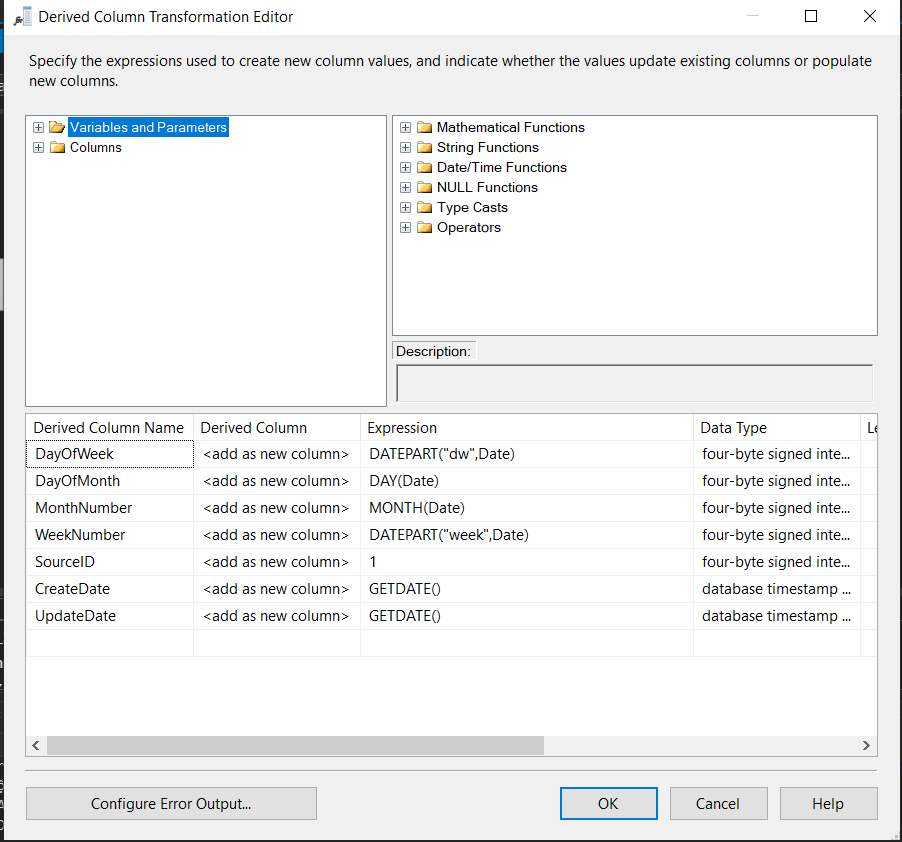
****

Ta sẽ nạp bảng chiều (gồm Dim\_CommunityArea, Dim\_Location, Dim\_Date, Dim\_FBICode, Dim\_IUCR) và bảng fact (Fact\_Crime)

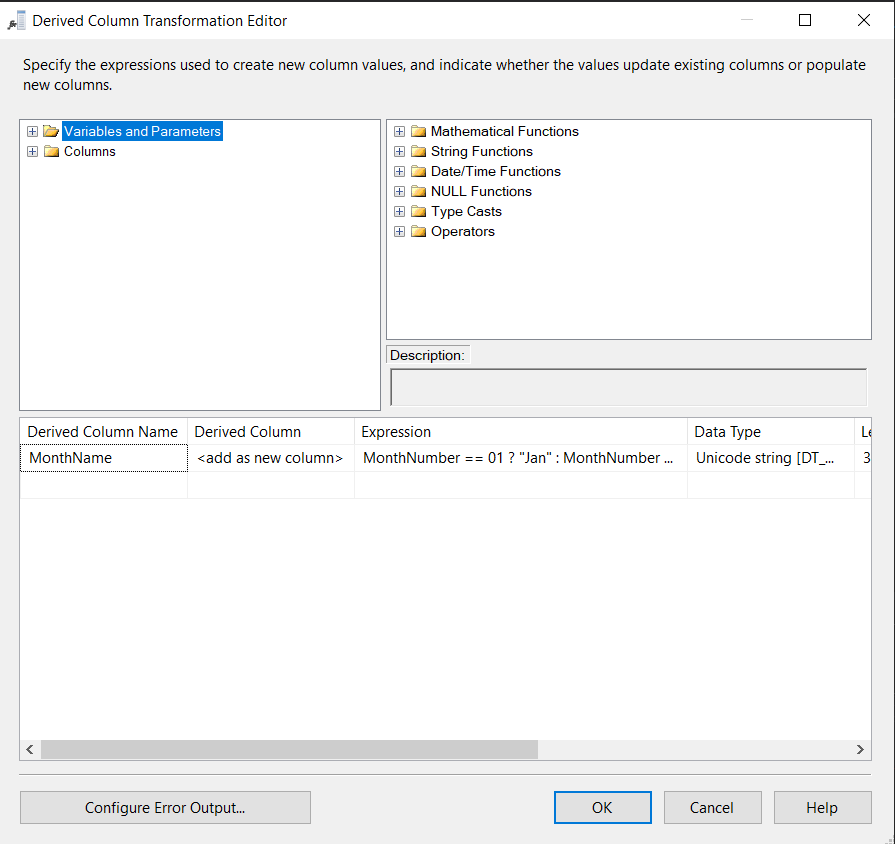
Đối với bảng Dim\_Date



Đầu tiên ta sẽ nạp dữ liệu từ stage, sau đó tạo ra các thuộc tính suy diễn như DayOfMonth, DayOfWeek, MonthNumber, WeekNumber.



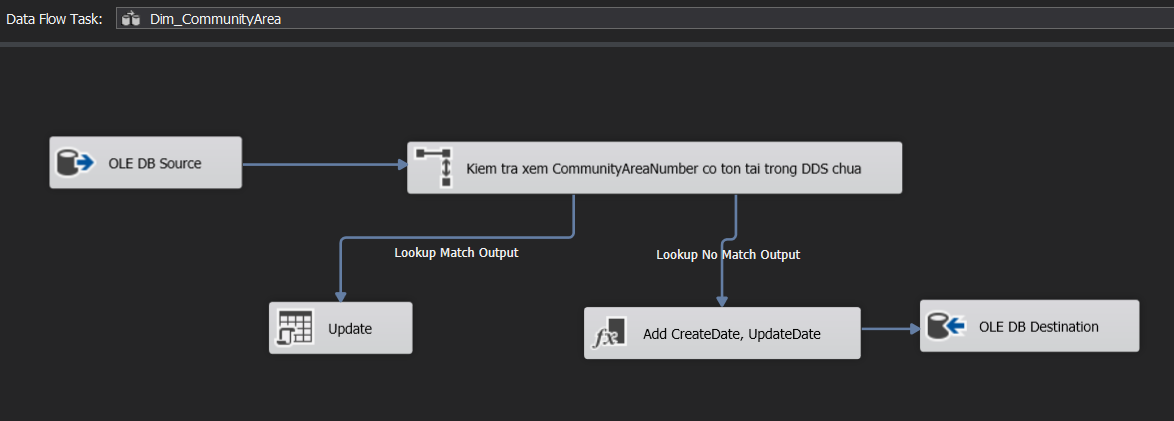
Tiếp theo ta sẽ suy diễn tên của tháng theo MonthNumber



**Cột Expression:**

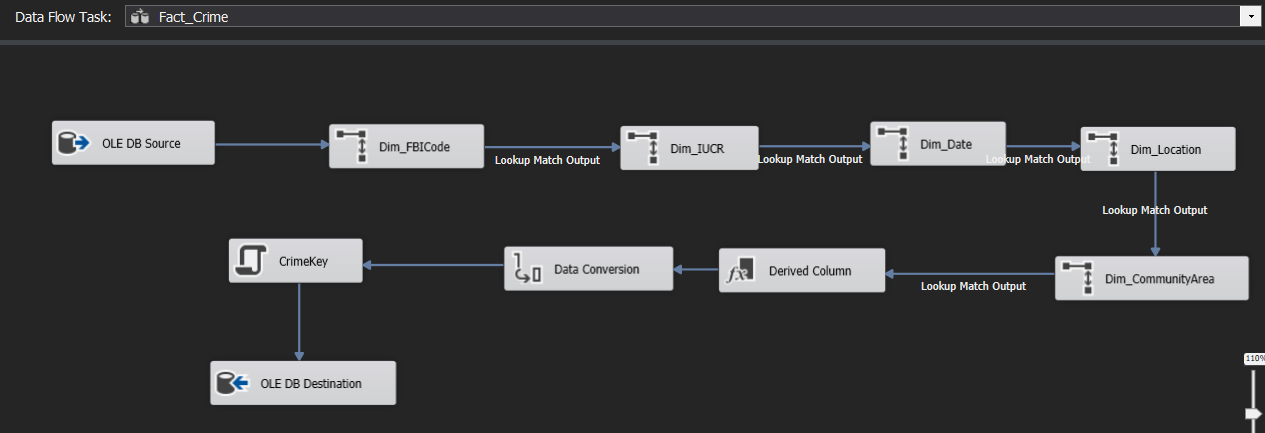
MonthNumber == "01" ? "Jan" : MonthNumber == "02" ? "Feb" : MonthNumber == "03" ? "Mar" : MonthNumber == "04" ? "Apr" : MonthNumber == "05" ? "May" : MonthNumber == "06" ? "Jun" : MonthNumber == "07" ? "Jul" : MonthNumber == "08" ? "Aug" : MonthNumber == "09" ? "Sep" : MonthNumber == "10" ? "Oct" : MonthNumber == "11" ? "Nov" : "Dec"

**Đối với bảng Dim\_CommunityArea**



Tương tự như ở NDS , ta sẽ kiểm tra có tồn tại trong DDS chưa.Nếu chưa tồn tại thì ta chèn vào bảng Dim\_CommunityArea trong DDS. Nếu đã tồn tại thì sẽ tiến hành cập nhật nếu dữ liệu có thay đổi.

**Đối với bảng Fact\_Crime**

****

Ta sẽ lookup trên các bảng chiều để tham chiếu khóa ngoại.

Cuối cùng sẽ tạo ràng buộc khóa ngoại cho bảng fact.

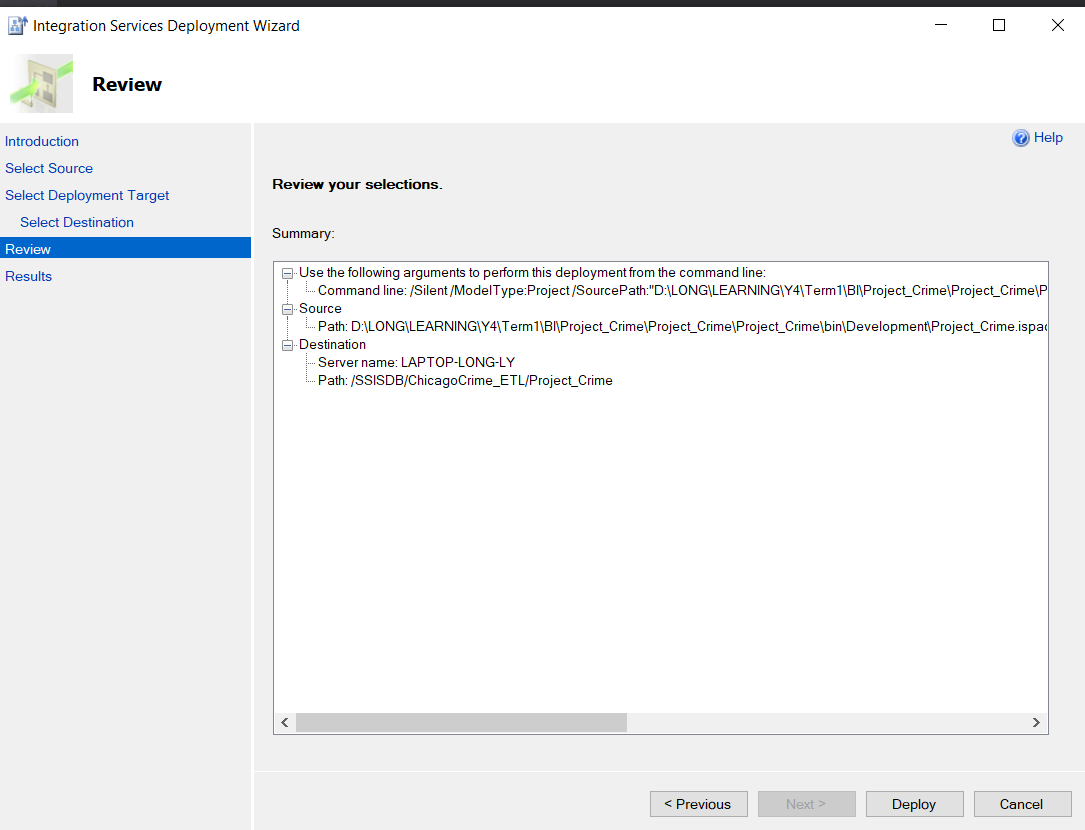
**Link video demo quá trình** ETL từ Stage->NDS->DDS

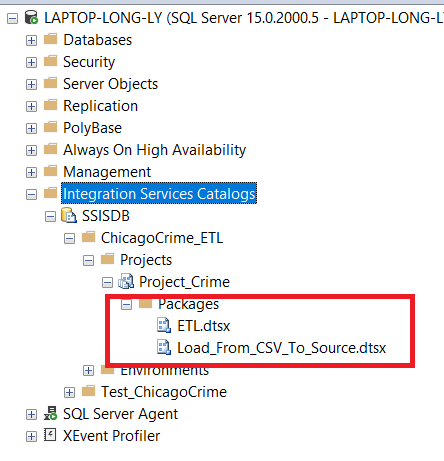
https://youtu.be/P\_kNraX-bmQ

# Lập lịch định kỳ cho ETL bằng cách deploy packages

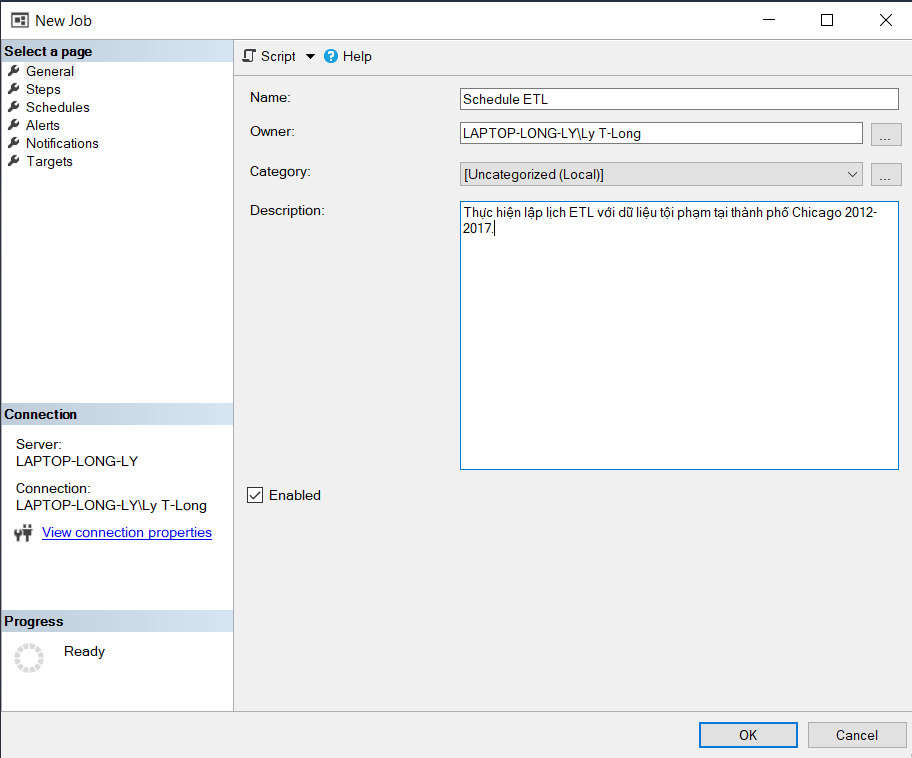
Các bước thực hiện như sau:

Đầu tiên ta sẽ deploy package ETL trong visual studio

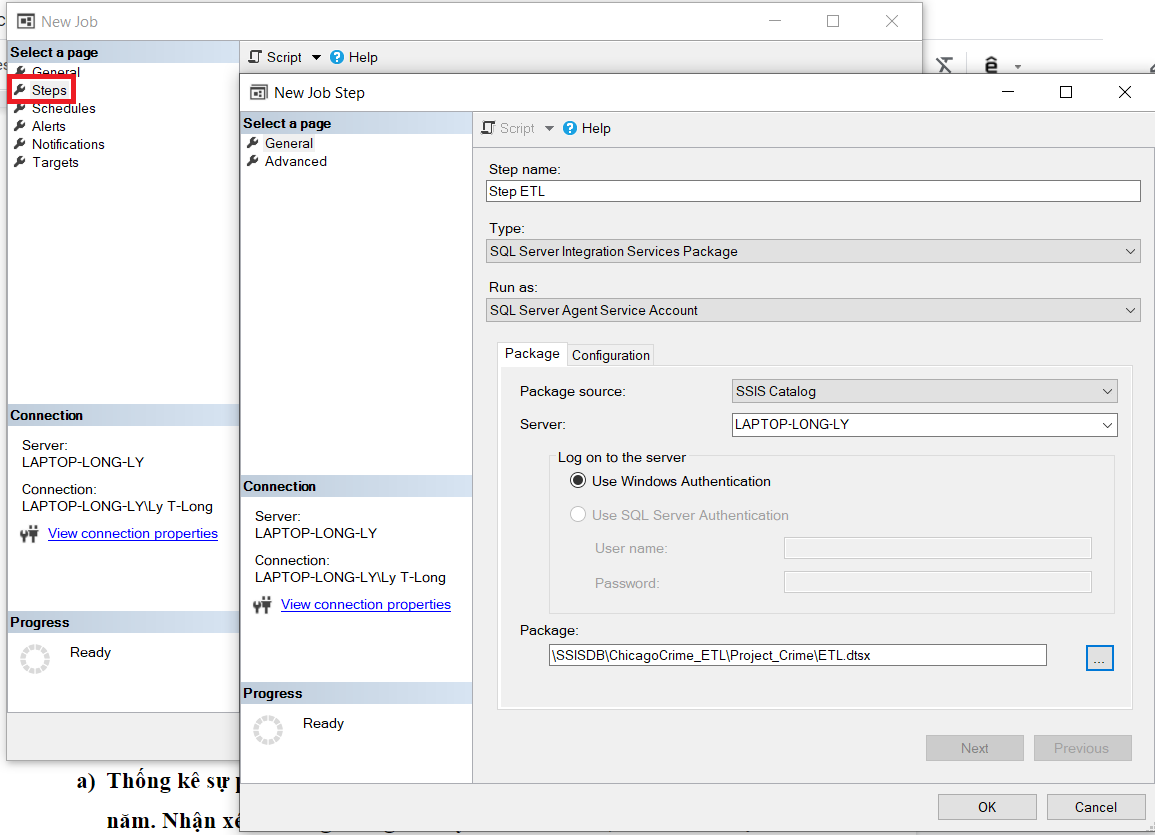
Sau đó kiểm tra trong SQL Server trong Database Engine trong Intergration Services Catalogs bên trong SSISDB trong đã deploy thành công chưa.



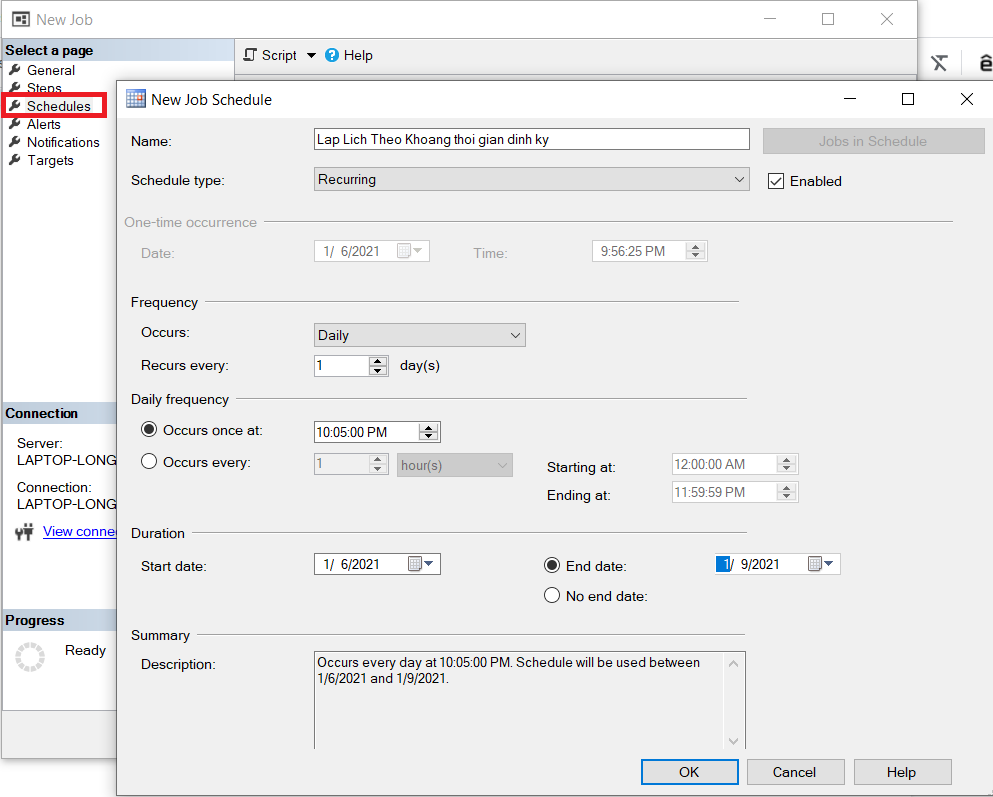
Tiếp đó ta vào SQL Server Agent để folder Jobs để tiến hành lập lịch ETL bằng cách nhấn vào “New Job”



Tiếp sau đó sẽ thêm cách thực hiện lập lịch (nhấn vào Steps bên tay phải). Đồng thời nhập tên các bước thực hiện và chọn loại Package Source, tên Server, chọn Package cần thực hiện lập lịch



Tiếp theo ta sẽ tiến hành lập lịch tại Schedules



**Link video demo quá trình định kỳ** (<https://youtu.be/-xXuF4C96qo> )

# Khai thác dữ liệu

## Report:

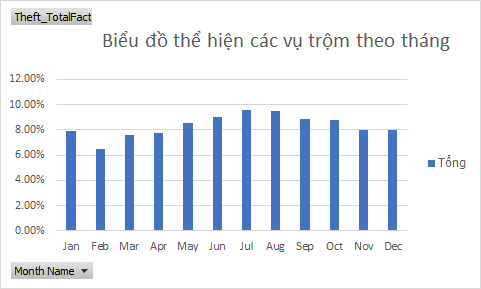
### Thống kê sự phân phối của các vụ trộm theo thời gian tháng, năm. Nhận xét khoảng thời gian xảy ra nhiều nhất, ít nhất các vụ trộm…

Tạo Calculate: [Theft\_TotalFact]

Expression: ([Fact Crime].[Theft].&[True], [Measures].[Frequency])

Format string: Standard

Ta chọn thuộc tính MonthName trong Dim\_Date và measure là [Theft\_TotalFact]

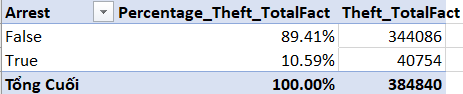


**Nhận xét:**

Nhìn chung các vụ trộm xảy ra đều giữa các tháng. Đỉnh điểm tháng 7 xảy ra nhiều vụ trộm nhất và tháng 2 ít xảy ra vụ trộm. Có thể thấy rằng từ tháng 2 đến 7 thì số vụ trộm xảy ra nhiều dần. Từ tháng 8 đến tháng 10 thì số vụ trộm giảm dần. Điều đáng nói ở đây là tháng 11 và tháng 12 thì các vụ trộm xảy ra bằng nhau

### Thống kê trong tất cả các năm/từng năm, các trường hợp trộm cắp mà không bị bắt giữ, hoặc bị bắt giữ

* Dùng Calculate: [Theft\_TotalFact] đã được tạo ở câu a
* Sau đó ta sẽ chọn thuộc tính Arrest trong trong bảng fact\_Crime cùng với measure [Theft\_TotalFact]



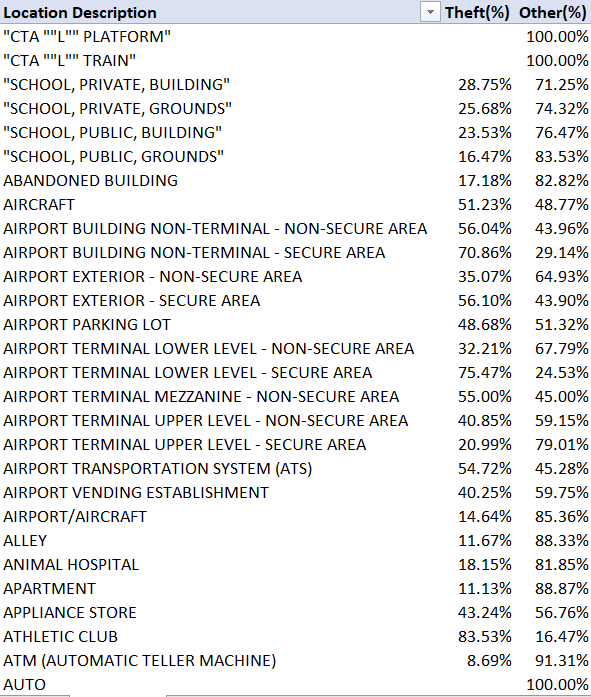
### Thống kê tỷ lệ trộm, tỷ lệ phạm tội khác theo từng địa điểm. Nhận xét

Ở đây ta sẽ xây dựng Calculate là [Theft(%)] để tính tỷ lệ là trộm

Với Expression:

([Dim IUCR].[Primary Type].&[THEFT],[Measures].[Frequency]) / [Measures].[Frequency]

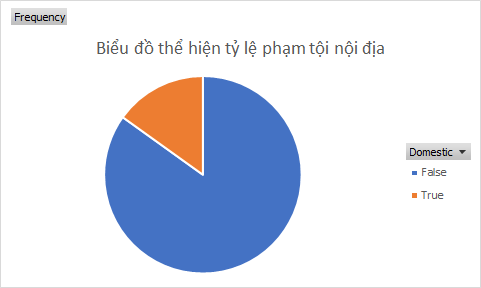
Còn tỉ lệ phạm tội khác thì 1 - Theft(%)



**Nhận xét**

Nhìn chung tỷ lệ các tội phạm khác(ngoại trừ trộm cắp) thì chiếm tỷ lệ cao hơn so với tội phạm trộm cắp ở từng địa điểm. Điều đáng nói ở đây là có một số địa điểm như Office, Goverment Building, Hallway, … tỷ lệ tội phạm khác chiếm 100%. Ngược lại cũng có một vài địa điểm như Vehicle - Delivery Truck, Athletic Club , Department Store, … tỉ lệ trộm cắp chiếm gấp ¾ so với tội phạm khác.

### Vẽ biểu đồ thể hiện tỷ lệ phạm tội nội địa



**Nhận xét**

Biểu đồ tròn thể hiện tội phạm là nội địa chiếm gần ¼ so với tội phạm không phải nội địa.

## OLAP

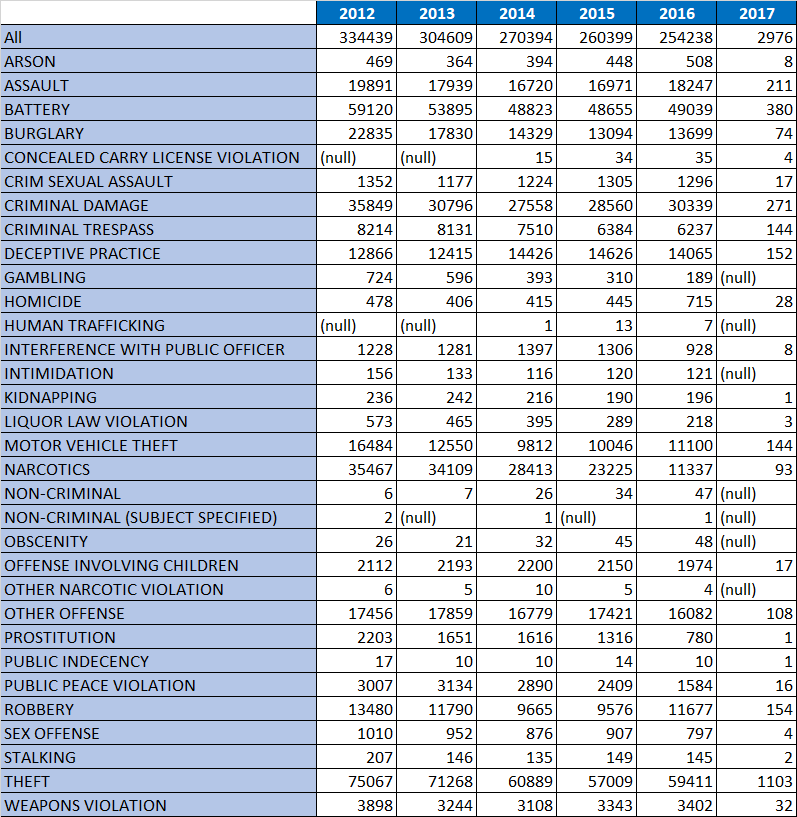
### Thống kê tần suất các loại tội phạm theo từng năm

* Query MDX:

select non empty [Dim Date].[Hierarchy].[Year] on columns,

non empty [Dim IUCR].[Primary Type].members on rows

from [DDS Chicago Crime]



**Nhận xét:**

Nhìn chung cho ta thấy số lượng tội phạm giảm dần theo từng năm. Điều đó cho ta kết luận rằng, mức độ an ninh được thực hiện nghiêm ngặt hơn, kiểm soát được tội phạm.

Vào năm 2017 thì số lượng tội phạm giảm một cách rõ rệt so với các năm trước.

Loại ***tội phạm trộm*** chiếm số lượng nhiều nhất tại các năm.

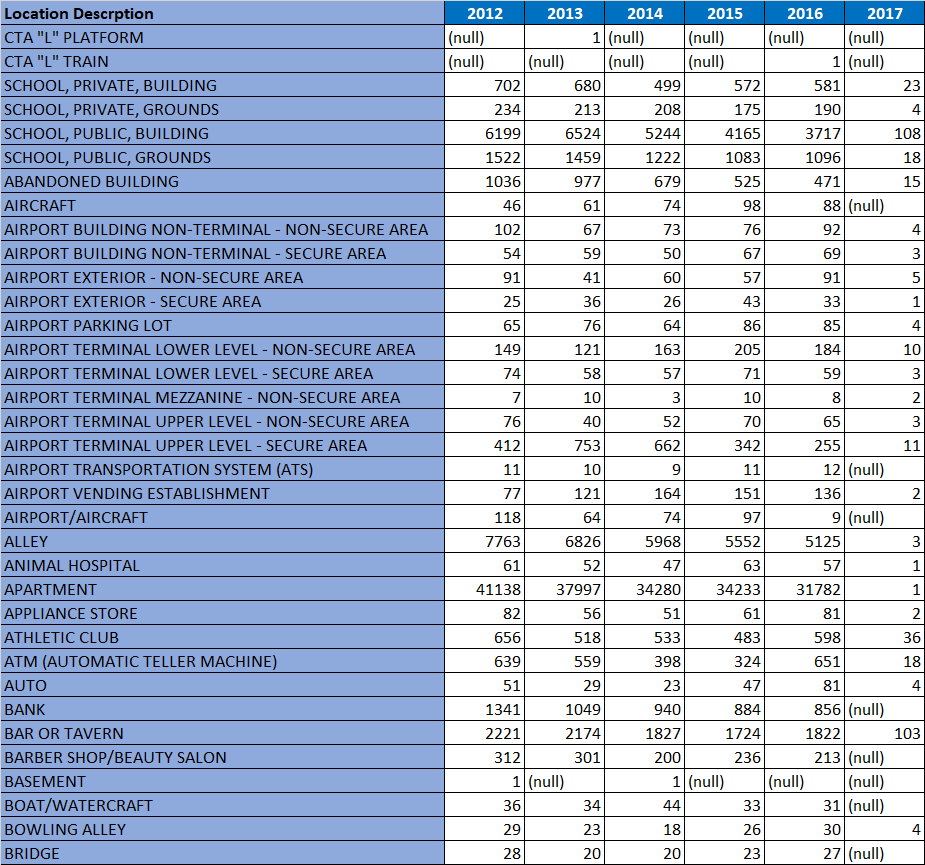
### Thống kê tần suất tội phạm theo thời gian và địa điểm

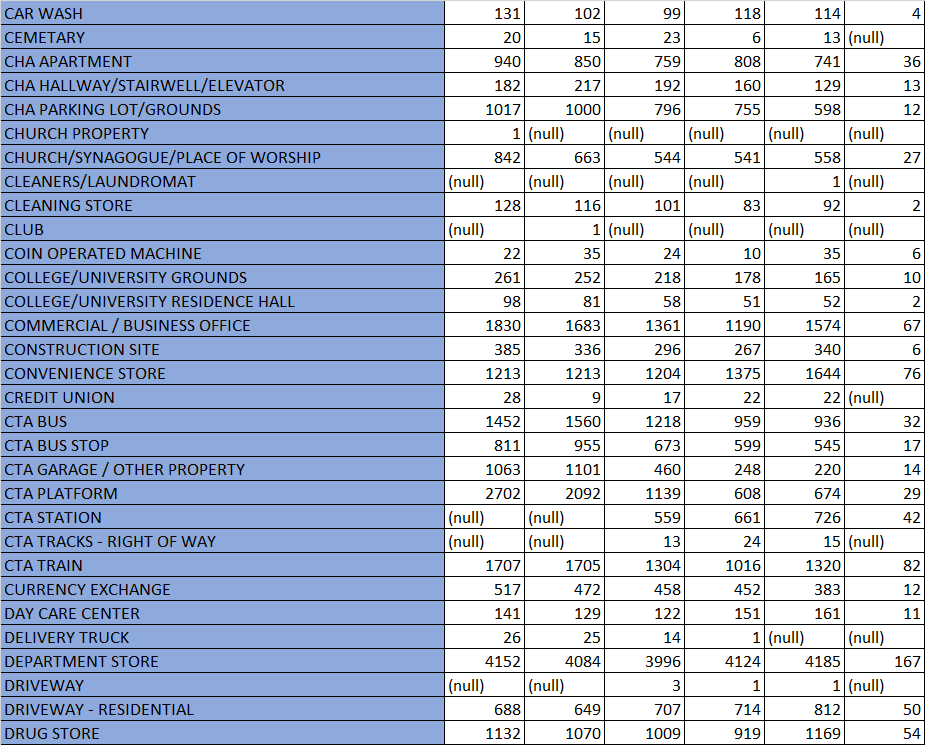
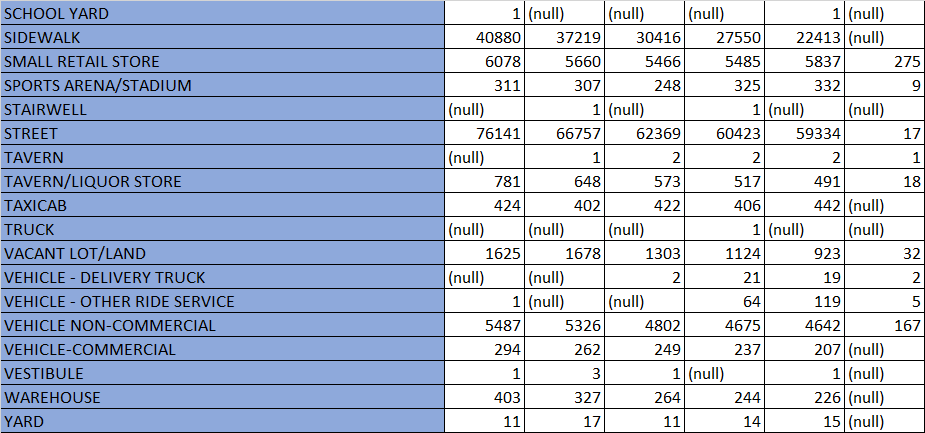
* Query MDX:

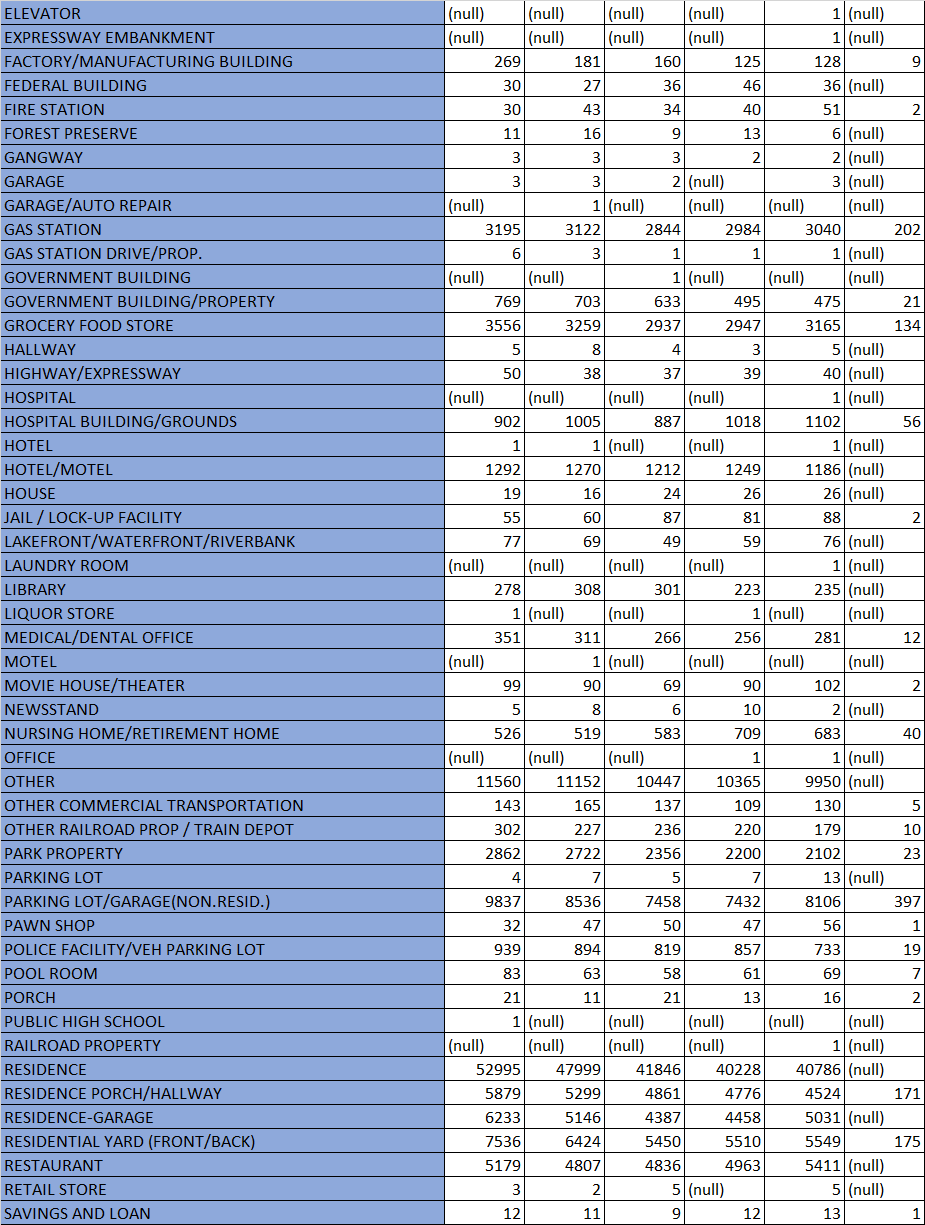
select non empty [Dim Date].[Hierarchy].[Year] on columns,

non empty [Dim Location].[Location Description Name].children on rows

from [DDS Chicago Crime]







### Thống kê tần suất theo các loại tội phạm

**Nhận xét:**

Với thống kê này cho ta thấy rằng loại tội phạm là trộm cắp thì tần suất xảy ra nhiều nhất. Ngược lại thì loại tội phạm là kẻ buôn người xảy ra ít nhất.

## Mining

* Tạo view vDM để thực hiện data mining (chuyển các giá trị số của các thuộc tính chỉ số cuộc sống trong bảng community area thành các nhãn ‘high’, ‘medium’, ‘low’) bằng cách thực hiện kết các bảng lại với nhau

create view vDM as

select f.[Case Number], f.Arrest, f.Theft, f.Domestic, fbi.Code, fbi.Categories,

d.DayOfMonth, d.MonthNumber, d.WeekNumber, d.Year,

c.[COMMUNITY AREA NAME],

case when c.[HARDSHIP INDEX] < 30 then 'low'

when c.[HARDSHIP INDEX] >= 30 and c.[HARDSHIP INDEX] <= 70 then 'medium'

else 'high' end as [Hardship Index],

case when c.[PER CAPITA INCOME] < 30000 then 'low'

when c.[PER CAPITA INCOME] >= 30000 and c.[PER CAPITA INCOME] <= 50000 then 'medium'

else 'high' end as [Income],

case when c.[PERCENT AGED 16+ UNEMPLOYED] < 15 then 'low'

when c.[PERCENT AGED 16+ UNEMPLOYED] >= 15 and c.[PERCENT AGED 16+ UNEMPLOYED] <= 25 then 'medium'

else 'high' end as [16+ UNEMPLOYED],

case when c.[PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA] < 25 then 'low'

else 'high' end as [25+ WITHOUT HIGH SCHOOL DIPLOMA],

case when c.[PERCENT AGED UNDER 18 OR OVER 64] < 30 then 'low'

else 'high' end as [UNDER 18 OR OVER 64],

case when c.[PERCENT HOUSEHOLDS BELOW POVERTY] < 20 then 'low'

when c.[PERCENT HOUSEHOLDS BELOW POVERTY] >= 20 and c.[PERCENT HOUSEHOLDS BELOW POVERTY] <= 35 then 'medium'

else 'high' end as [HOUSEHOLDS BELOW POVERTY],

case when c.[PERCENT OF HOUSING CROWDED] < 6 then 'low'

when c.[PERCENT OF HOUSING CROWDED] >= 6 and c.[PERCENT OF HOUSING CROWDED] <= 10 then 'medium'

else 'high' end as [HOUSING CROWDED],

iucr.IUCR, iucr.[Primary Type], iucr.Description,

l.LocationDescriptionName

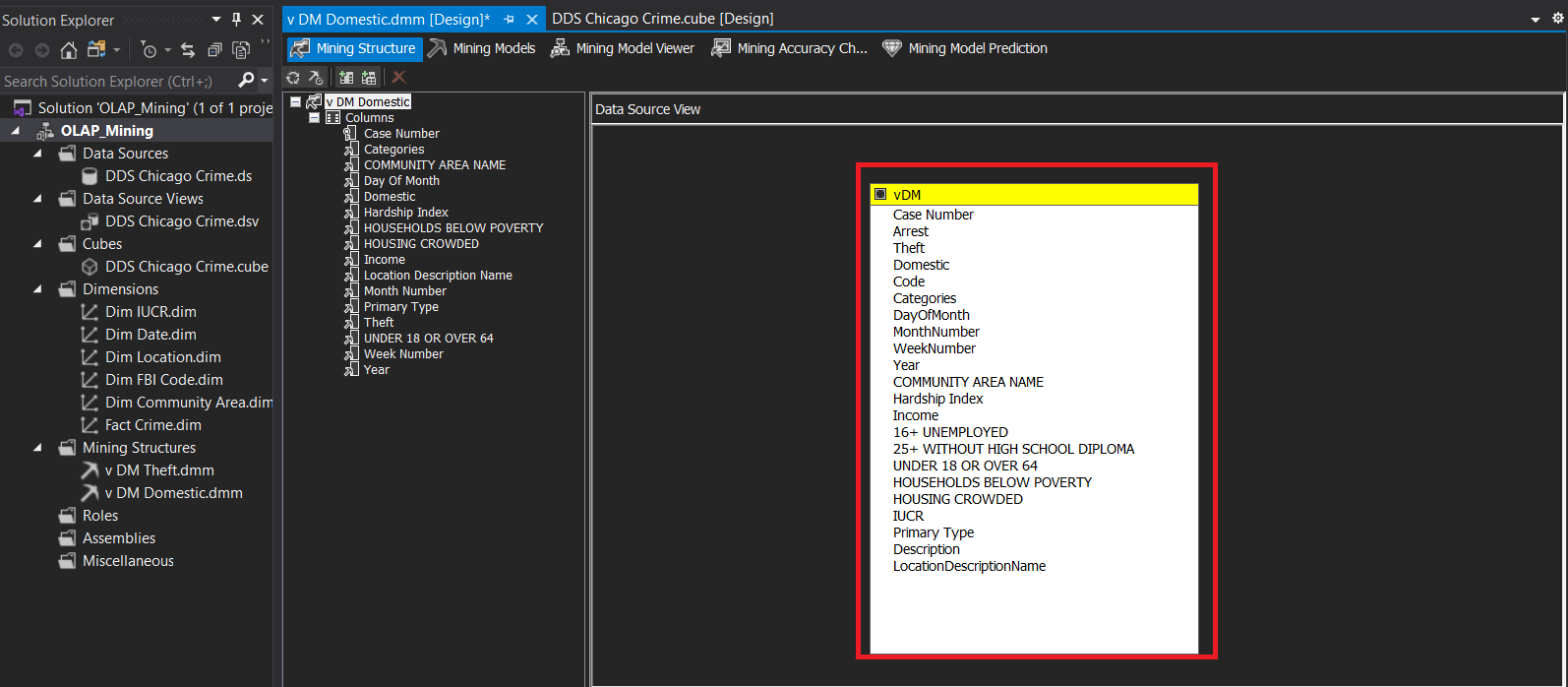
from Fact\_Crime f inner join Dim\_Date d on f.DateKey = d.DateKey

inner join Dim\_CommunityArea c on c.CommunityAreaKey = f.CommunityAreaKey

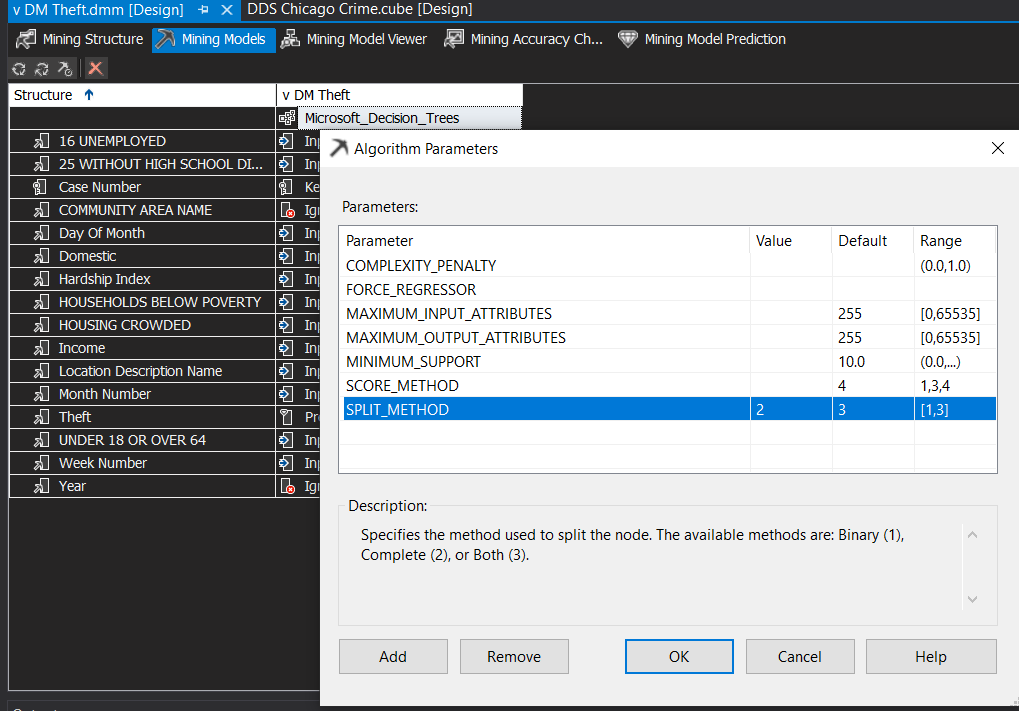
inner join Dim\_FBICode fbi on fbi.FBICodeKey = f.FBICodeKey

inner join Dim\_IUCR iucr on iucr.IUCRKey = f.IUCRKey

inner join Dim\_Location l on l.LocationDescriptionKey = f.LocationDescriptionKey;



* **Xây dựng mô hình decision tree cho hai nhu cầu**:
* Dự đoán tội phạm là trộm? (input gồm các thuộc tính chỉ số sinh sống của community area, tội phạm nội địa, mô tả địa điểm và các thuộc tính thời gian (không có năm); sử dụng split\_method complete)
* Dự đoán tội phạm là tội phạm nội địa? (input gồm các thuộc tính chỉ số sinh sống của community area, các thuộc tính phân loại tội phạm, mô tả địa điểm và các thuộc tính thời gian (không có năm); sử dụng split\_method complete)
* Ta sẽ chọn Split\_method “value = 2” trong Algorithms Parameters



Ở đây chọn Split\_Method value = 2 vì 1: xuống chi tiết quá, 3: rối quá nên nhóm chọn 2

Kết khi chọn split\_method có value = 1 bằng các thực thi vDM Domestic



Kết khi chọn split\_method có value = 2 bằng các thực thi vDM Domestic



Kết khi chọn split\_method có value = 3 bằng các thực thi vDM Domestic



**Kết quả chạy mining để dự đoán tội phạm là trộm**

****

**Kết quả chạy mining để dự đoán xem tội phạm là tội phạm nội địa**



Kết quả thực thi được quay lại và đăng trên youtube (<https://youtu.be/5gLr6C2LfdY>)

## KPI

* **Name:** TinhTrangToiPham
* **Value Expression**: [Measures].[Frequency]
* **Goal Expression**:

Case

When [Fact Crime].[Theft].CurrentMember Is

[Fact Crime].[Theft].[True]

Then 70000

Else 40000

End

* **Status**: Status indicator: Cylinder

Status Expression:

Case

When KpiValue("TinhTrangToiPham") / KpiGoal("TinhTrangToiPham") > 1.1

Then -1

When KpiValue("TinhTrangToiPham") / KpiGoal("TinhTrangToiPham") <= 1.1

And

KpiValue("TinhTrangToiPham") / KpiGoal("TinhTrangToiPham") > 1

Then 0

Else 1

End

## Kết luận chung

* Với bộ dữ liệu tội phạm ở thành phố Chicago giai đoạn 2012-2017 cho ta kết luận rằng số lượng tội phạm giảm dần theo từng năm. Điều đó cho ta kết luận rằng, mức độ an ninh được thực hiện nghiêm ngặt hơn.
* Vào năm 2017 thì số lượng tội phạm giảm một cách rõ rệt so với các năm trước.Loại ***tội phạm trộm*** chiếm số lượng nhiều nhất tại các năm.
* Với tội phạm là nội địa chiếm gần ¼ so với tội phạm không phải nội địa. Điều đó có ý nghĩa là tội phạm đến từ bên ngoài thành phố Chicago.
* Các vụ trộm xảy ra tập trung vào giữa năm. Đỉnh điểm các vụ trộm xảy ra vào tháng 6, 7, 8. Ngược lại tháng 2 thì số vụ trộm xảy ra ít nhất.
* Các vụ trộm cắp, các loại tội phạm khác thường tập trung vào các tòa nhà bỏ hoang với tần suất là 3703 vụ.

# Tham khảo

[1] [Get started with OpenRefine](http://miriamposner.com/classes/dh101f17/tutorials-guides/data-manipulation/get-started-with-openrefine/)

[2] <https://www.enhansoft.com/how-do-you-install-sql-server-data-tools/>

[3] <https://docs.microsoft.com/en-us/sql/ssdt/download-sql-server-data-tools-ssdt?view=sql-server-ver15>

[4] <https://docs.microsoft.com/en-us/analysis-services/multidimensional-tutorial/lesson-7-1-defining-and-browsing-kpis?view=asallproducts-allversions>

[5] <https://giaphiep.com/blog/phan-tich-kho-du-lieu-datawarehouse-bang-olap-26957>