

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
THÀNH PHỐ HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐÖ 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.	Giới thiệu tổng quan về đồ án.....	5
II.	Các công cụ thực hiện	11
III.	Quy trình thực hiện.....	11
1)	Quy trình thiết kế kho dữ liệu.....	12
2)	Quy trình xây dựng cube và mining	12
3)	Quy trình khai thác dữ liệu	14
IV.	Mô tả dữ liệu	14
V.	Thiết kế dữ liệu	17
1)	Thiết kế NDS.....	17
2)	Thiết kế DDS	18
VI.	Nạp dữ liệu:	19
1)	Cách thức lưu trữ.....	19
2)	Chuyển dữ liệu từ file .csv vào source trong SQL Server.....	19
3)	Chuyển dữ liệu từ nguồn vào stage	20
4)	Ứng dụng tool Openfire trong giai đoạn nhất quán dữ liệu	20
5)	Làm sạch dữ liệu null	29
VII.	Quy trình nạp dữ liệu từ Source -> Stage -> NDS -> DDS	30
VIII.	Lập lịch định kỳ cho ETL bằng cách deploy packages	38
IX.	Khai thác dữ liệu	43
1)	Report:	43

a)	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)	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)	Thống kê tỷ lệ trộm, tỷ lệ phạm tội khác theo từng địa điểm. Nhận xét ...	45
d)	Vẽ biểu đồ thể hiện tỷ lệ phạm tội nội địa	47
2)	OLAP	47
a)	Thống kê tần suất các loại tội phạm theo từng năm.....	47
b)	Thống kê tần suất tội phạm theo thời gian và địa điểm	49
c)	Thống kê tần suất theo các loại tội phạm.....	52
3)	Mining	53
4)	KPI	59
5)	Kết luận chung	60
X.	Tham khảo.....	61

I. 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 crowded, Percentage households below poverty, Percentage aged 16+ unemployed, Percentage aged 25+ without high school

diploma, percentage aged under 18 or over 64, per capita 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

HOSPITAL BUILDING/GROUNDS
COMMERCIAL / BUSINESS OFFICE

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

SCHOOL	PUBLIC	BUILDING
RESIDENCE		
RESIDENCE		

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

POOL ROOM (340 rows)
POOLROOM (1 rows)

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

• POOL ROOM (340 rows)	<input checked="" type="checkbox"/>	POOL ROOM
• POOLROOM (1 rows)	<input type="checkbox"/>	
• BARBERSHOP (1258 rows)	<input type="checkbox"/>	BARBER SHOP/BEAUTY SALON
• BARBER SHOP/BEAUTY SALON (4 rows)	<input type="checkbox"/>	
• AIRPORT EXTERIOR - NON-SECURE AREA (345 rows)	<input type="checkbox"/>	AIRPORT EXTERIOR - NON-SECURE AREA
• AIRPORT EXTERIOR - SECURE AREA (164 rows)	<input type="checkbox"/>	
• SCHOOL, PUBLIC, BUILDING (25957 rows)	<input type="checkbox"/>	SCHOOL, PUBLIC, BUILDING
• SCHOOL, PUBLIC, GROUNDS (6400 rows)	<input type="checkbox"/>	
• GOVERNMENT BUILDING/PROPERTY (3096 rows)	<input type="checkbox"/>	GOVERNMENT BUILDING/PROPERTY
• GOVERNMENT BUILDING (1 rows)	<input type="checkbox"/>	
• TAXICAB (2094 rows)	<input checked="" type="checkbox"/>	TAXICAB
• TAXI CAB (2 rows)	<input type="checkbox"/>	

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

Thông 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

II. 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>.

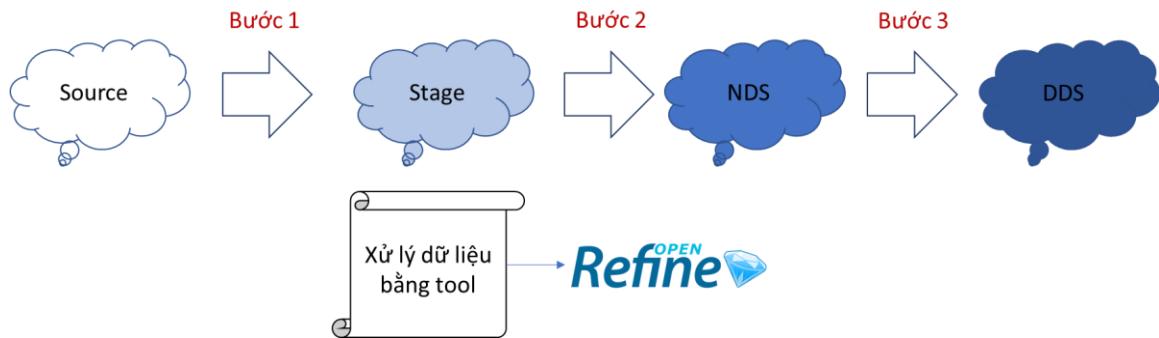
Lưu ý sử dụng phiên bản Develop

- 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/1Ic8hEnjfqVJ53Td3wnJOPuk0sl5afy1B/view?usp=sharing>.
- Dùng Excel để thể hiện báo cáo , vẽ biểu đồ cột, biểu đồ tròn.

III. 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

1) Quy trình thiết kế kho dữ liệu



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

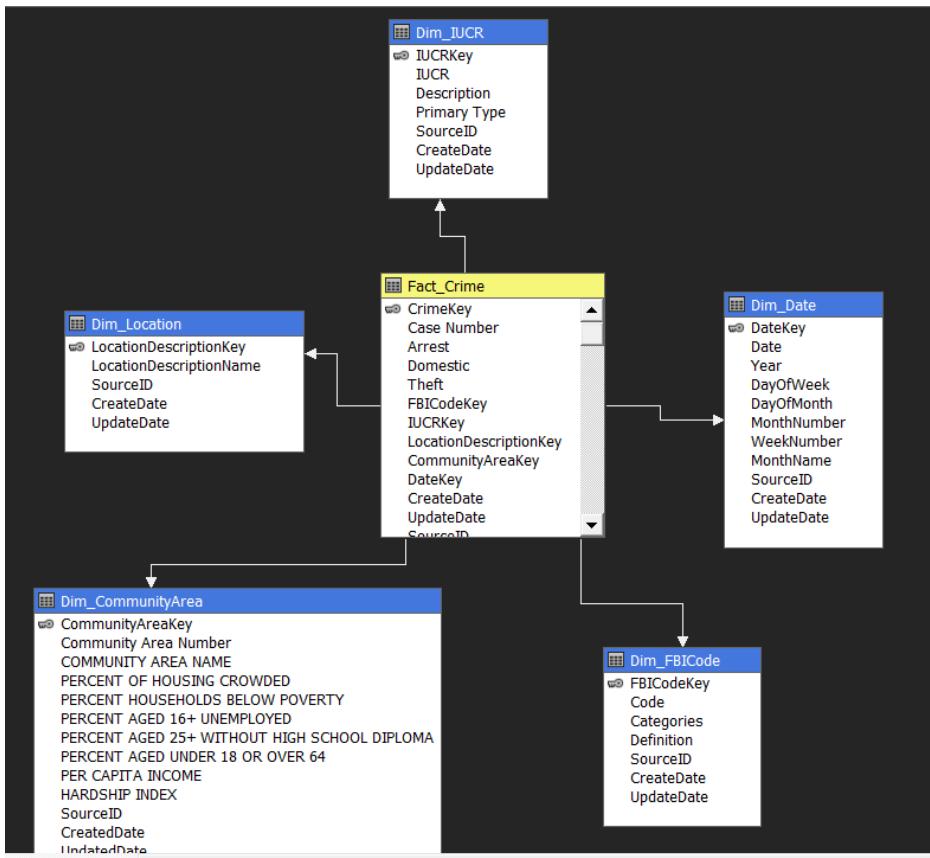
Bước 1 tiến hành ETL sử dụng công cụ OpenRefine để xử lý dữ liệu để chuyển dữ liệu từ source vào stage.

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

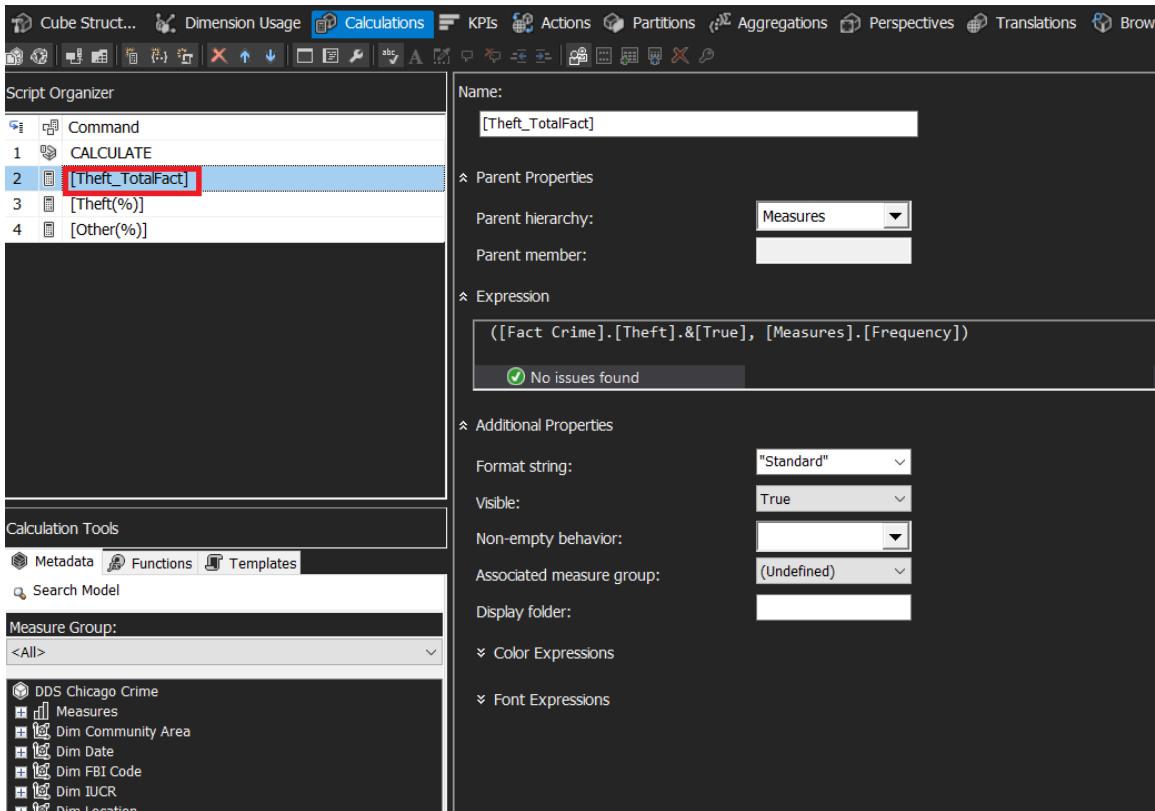
Bước 3 dữ liệu sẽ được nạp vào DDS với chiều, fact, độ đo tương ứng (phần này sẽ được mô tả rõ hơn ở phần VI, 2 Thiết kế dữ liệu DDS).

2) 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



3) 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.

IV. 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.

V. Thiết kế dữ liệu

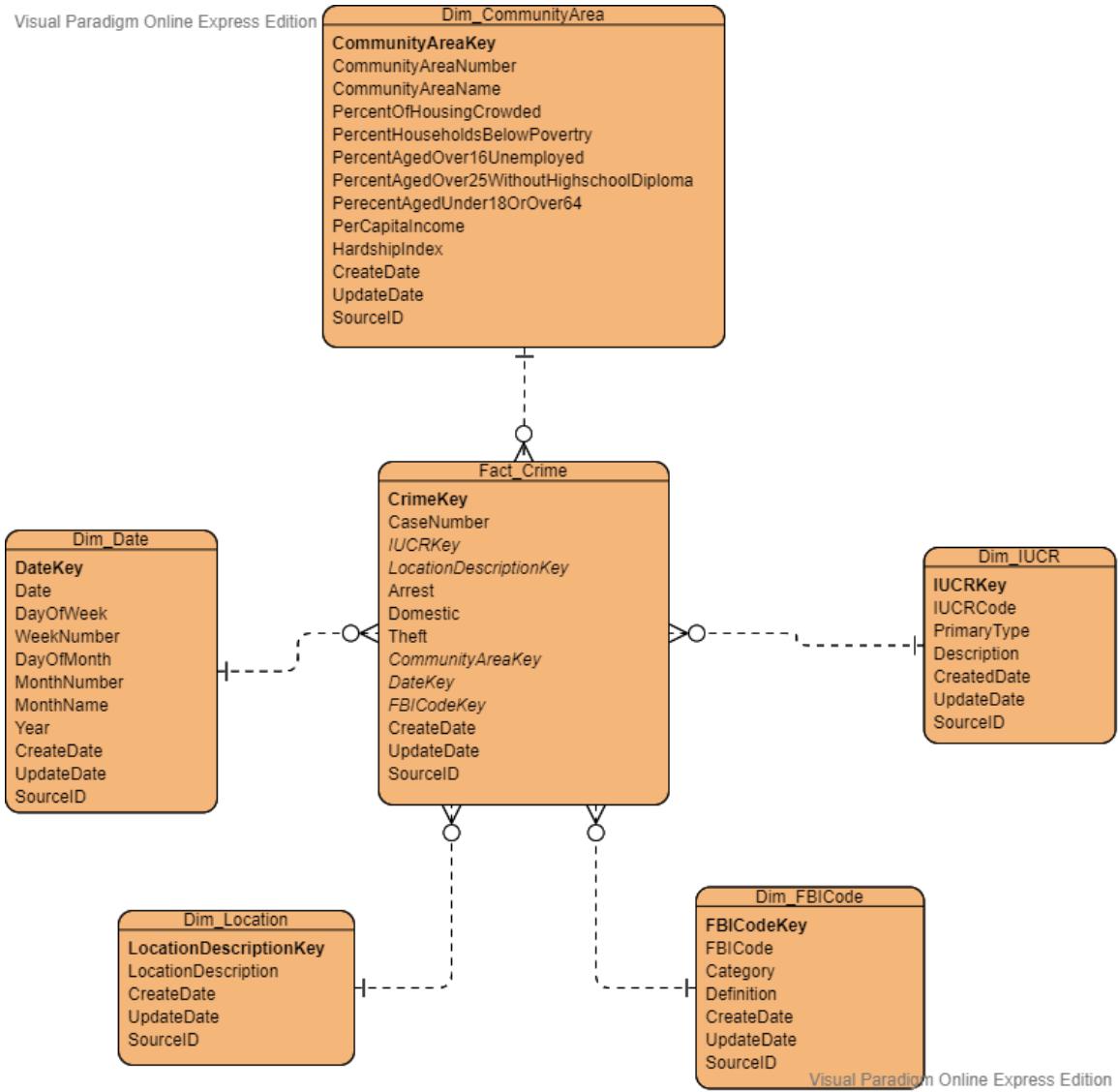
1) Thiết kế NDS

Visual Paradigm Online Express Edition



Visual Paradigm Online Express Edition

2) 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.

VI. Nạp dữ liệu:

1) 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.

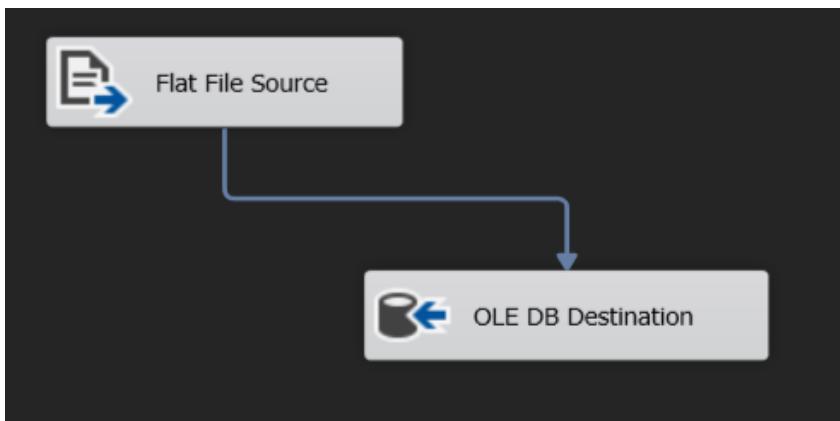
Sau đó tạo một cơ sở dữ liệu Stage_ChicagoCrime chứa các bảng
Census_Stage, Crime_Stage, FBICode_Stage chuyển dữ liệu từ Soure
(cơ sở dữ liệu Chicago_Crime sang Stage).

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.

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





3) Chuyển dữ liệu từ nguồn vào stage

Chuyển dữ liệu các bảng Census , FBI Code vào stage



4) 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.

CaseNumber	LocationDescription	Column 3	Column 4
75. HT441210	VEHICLE NON-COMMERCIAL		
76. HT498709	STREET		
77. HT504601	STREET		
78. HT514035	STREET		
79. HT620246	SCHOOL	PUBLIC	BUILDING
80. HT620274	RESIDENCE		
81. HT620374	RESIDENCE		
82. HT621759	RESIDENCE-GARAGE		
83. HT626172	GARAGE		
84. HT628640	RESIDENCE		
85. HT628662	STREET		
86. HT628693	RESIDENCE		
87. HT628779	STREET		
88. HT628879	STREET		
89. HT628905	STREET		
90. HT629139	STREET		
91. HT630405	SCHOOL	PUBLIC	GROUNDS
92. HT630474	RESIDENCE-GARAGE		
93. HT630502	RESIDENCE		
94. HT630568	APARTMENT		
95. HT630733	STREET		

Giải pháp:

Vào Edit cells -> Transform... của cột LocationDescription

The screenshot shows the OpenRefine interface with a project containing 142821 rows. A context menu is open over a row in the 'LocationDescription' column, specifically for row 75. The menu path 'Edit cells > Transform...' is highlighted. Other options visible in the menu include 'Edit column', 'Transpose', 'Sort...', 'View', 'Reconcile', and 'Replace'. The interface also features a 'Facet / Filter' bar at the top left, a toolbar with 'Extract...', 'Apply...', and 'Extensions' dropdowns, and a status bar at the bottom right showing system information like battery level and date/time.

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)))
```

Custom text transform on column LocationDescription

```
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)))
```

Original	Transformed
4. CTA PLATFORM	CTA PLATFORM
5. RESIDENCE	RESIDENCE
6. APARTMENT	APARTMENT
7. STREET	STREET
8. CONVENIENCE STORE	CONVENIENCE STORE
9. ATM (AUTOMATIC TELLER MACHINE)	ATM (AUTOMATIC TELLER MACHINE)
10. RESIDENCE	RESIDENCE
11. SIDEWALK	SIDEWALK

On error keep original set to blank store error

Re-transform up to 10 times until no change

OK Cancel

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

142821 rows

CaseNumber	LocationDescription	Column 3	Column 4
75. HT441210	VEHICLE NON-COMMERCIAL		
76. HT498709	STREET		
77. HT504601	STREET		
78. HT514035	STREET		
79. HT620246	SCHOOL, PUBLIC, BUILDING	PUBLIC	BUILDING
80. HT620274	RESIDENCE		
81. HT620374	RESIDENCE		
82. HT621759	RESIDENCE-GARAGE		
83. HT626172	GARAGE		
84. HT628640	RESIDENCE		
85. HT628662	STREET		
86. HT628693	RESIDENCE		
87. HT628779	STREET		
88. HT628879	STREET		
89. HT628905	STREET		
90. HT629139	STREET		
91. HT630405	SCHOOL, PUBLIC, GROUNDS	PUBLIC	GROUNDS
92. HT630474	RESIDENCE-GARAGE		
93. HT630502	RESIDENCE		
94. HT630568	APARTMENT		
95. HT630733	STREET		

Xóa bỏ đi hai cột Column 3 và Column 4

The screenshot shows the OpenRefine interface with a list of 142821 rows. The columns are CaseNumber and LocationDescription. The LocationDescription column contains values like 'VEHICLE NON-COMMERCIAL', 'STREET', 'SCHOOL, PUBLIC, BUILDING', etc. On the left, there's a facet for LocationDescription with 141 choices. A script editor on the left side contains a complex transformation script using Greplin syntax. The bottom status bar shows system information like battery level, time (12:18 PM), and date (12/15/2020).

Gộp những thuộc tính trong LocationDescription Vào Edit cells -> Cluster and edit

This screenshot shows the same OpenRefine interface as above, but with a context menu open over the LocationDescription column. The 'Cluster' option is highlighted in the menu. The menu also includes other options like 'Edit cells', 'Edit column', 'Transpose', 'Sort...', 'View', 'Reconcile', and 'Cluster and edit...'. The bottom status bar shows system information like battery level, time (12:19 PM), and date (12/15/2020).

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 / BEAUTY SALON ----->**gộp lấy barbershop/ 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 tải 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 chǎng, Còn retail thì lớn hơn chút

TAVERN/LIQUOR STORE,

TAVERN,

LIQUOR STORE -----> **không gộp**

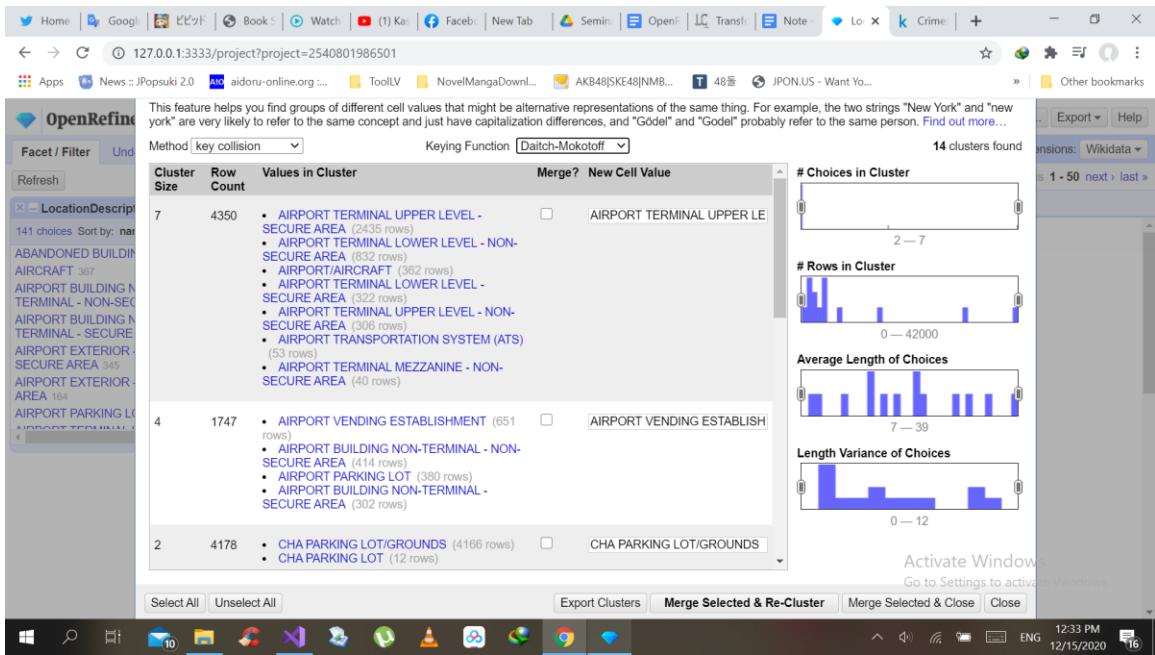
Lý do: Cái TAVERN/LIQUOR chiếm đa số, Mấy cái kia nhỏ quá nên gộp ...
chắc không sao

PUBLIC HIGH SCHOOL 1,
SCHOOL YARD 2,
[SCHOOL, PRIVATE, BUILDING] 3057,
[SCHOOL, PRIVATE, GROUNDS] 1024,
[SCHOOL, PUBLIC, BUILDING] 25957,
[SCHOOL, PUBLIC, GROUNDS] 6400

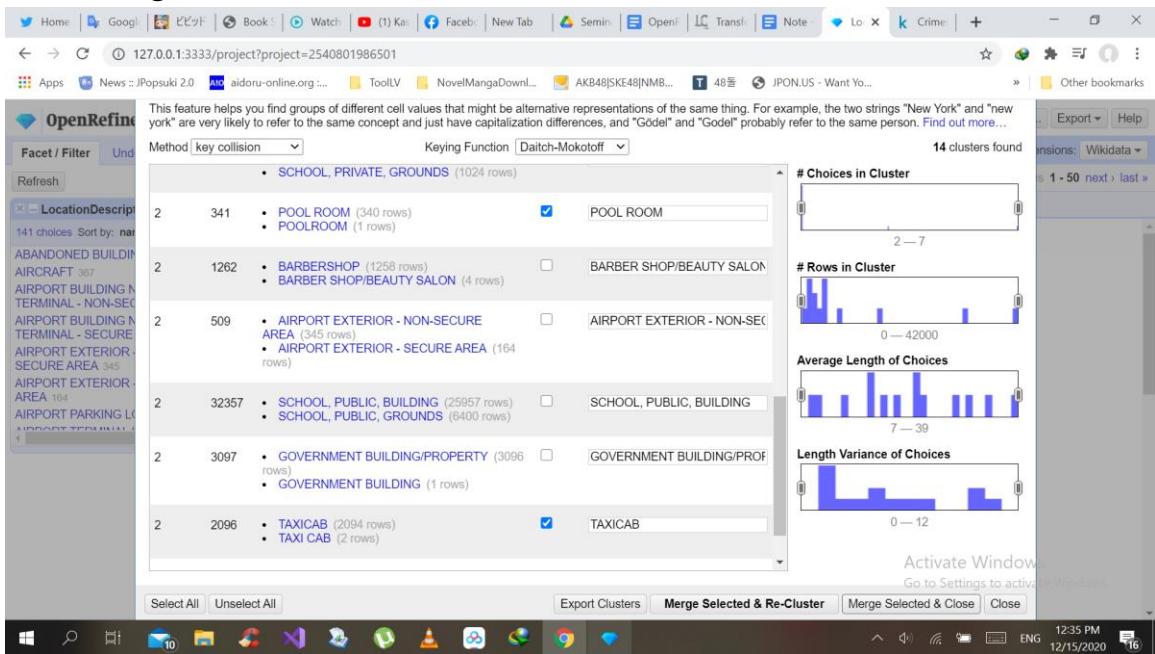
-----> **không gộp**

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 định 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

Facet / Filter Undo / Redo 4 / 4

Refresh Reset All Remove All

LocationDescription change

137 choices Sort by: name count Cluster

ABANDONED BUILDING 3703
AIRCRAFT 367
AIRPORT BUILDING NON-TERMINAL - NON-SECURE AREA 414
AIRPORT BUILDING NON-TERMINAL - SECURE AREA 302
AIRPORT EXTERIOR - NON-SECURE AREA 345
AIRPORT EXTERIOR - SECURE AREA 164
AIRPORT PARKING LOT 380
AIRPORT TERMINAL LOWER LEVEL - NON-SECURE AREA 832
AIRPORT TERMINAL LOWER LEVEL - SECURE AREA 322
AIRPORT TERMINAL MEZZANINE - NON-SECURE AREA 40
AIRPORT TERMINAL UPPER LEVEL - NON-SECURE AREA 306
AIRPORT TERMINAL UPPER LEVEL - SECURE AREA 114

Facet Text facet

Text filter Numeric facet

Edit cells Timeline facet

Edit column Scatterplot facet

Transpose Custom text facet...

Sort... Custom Numeric Facet...

View Customized facets

Reconcile

SIDEWALK
GROCERY FOOD STORE
RESIDENCE
RESIDENTIAL YARD (FRONT/BACK)
APARTMENT
RESIDENTIAL YARD (FRONT/BACK)
GAS STATION
RESIDENCE
RESTAURANT
APARTMENT
STREET
SIDEWALK

Activate Windows
Go to Settings to activate Windows.

Extensions: Wikidata ▾

1428281 rows

Show as: rows records Show: 5 10 25 50 rows

« first < previous 1 - 50 next > last »

Open... Export Help

12:36 PM 12/15/2020

Facet / Filter Undo / Redo 4 / 4

Refresh Reset All Remove All

LocationDescription change

137 choices Sort by: name count Cluster

JAIL / LAUNDRY FACILITY 373
LAKEFRONT/WATERFRONT/RIVERBANK 32
LAUNDRY ROOM 1
LIBRARY 1345
LIQUOR STORE 2
MEDICAL/DENTAL OFFICE 1477
MOTEL 1
MOVIE HOUSE/THEATER 452
NEWSSTAND 31
NURSING HOME 2 edit include
NURSING HOME/RETIREMENT HOME 3058
OFFICE 2
OTHER 53474
OTHER COMMERCIAL TRANSPORTATION 689
OTHER RAILROAD PROP / TRAIN DEPOT 1174
PARK PROPERTY 12265

HOSPITAL BUILDING/GROUNDS
COMMERCIAL / BUSINESS OFFICE
STREET
CTA PLATFORM
RESIDENCE
APARTMENT
STREET
CONVENIENCE STORE
ATM (AUTOMATIC TELLER MACHINE)
RESIDENCE
SIDEWALK
GROCERY FOOD STORE
RESIDENCE
RESIDENTIAL YARD (FRONT/BACK)
APARTMENT
RESIDENTIAL YARD (FRONT/BACK)
GAS STATION
RESIDENCE
RESTAURANT
APARTMENT
STREET
SIDEWALK

Activate Windows
Go to Settings to activate Windows.

Extensions: Wikidata ▾

1428281 rows

Show as: rows records Show: 5 10 25 50 rows

« first < previous 1 - 50 next > last »

Open... Export Help

12:40 PM 12/15/2020

Facet / Filter Undo / Redo 4 / 4

LocationDescription

137 choices Sort by: name count Cluster

JAIL / LOOK-UP FACILITY 373

LAKEFRONT/WATERFRONT/RIVERBANK 35

LAUNDRY ROOM 1

LIBRARY 1345

LIQUOR STORE 2

MEDICAL/DENTAL OFFICE 1477

MOTEL 1

MOVIE HOUSE/THEATER 452

NEWSSTAND 31

NURSING HOME 2

NURSING HOME/RETIREMENT HOME 3058

OFFICE 2

OTHER 53474

OTHER COMMERCIAL TRANSPORTATION 689

OTHER RAILROAD PROP / TRAIN DEPOT 1174

PARK PROPERTY 12265

Activate Windows
Go to Settings to activate Windows.

1428281 rows

Show as: rows records Show: 5 10 25 50 rows

All CaseNumber LocationDescription

CaseNumber	LocationDescription
1. 223432	HOSPITAL BUILDING/GROUNDS
2. 311997	COMMERCIAL / BUSINESS OFFICE
3. 318876	STREET
4. 322338	CTA PLATFORM
5. 413567	RESIDENCE
6. 536481	APARTMENT
7. F218264	STREET
8. HA107183	CONVENIENCE STORE
9. HA121046	ATM (AUTOMATIC TELLER MACHINE)
10. HA156050	RESIDENCE
11. HA164684	SIDEWALK
12. HA168845	GROCERY FOOD STORE
13. HA179595	RESIDENCE
14. HA194202	RESIDENTIAL YARD (FRONT/BACK)
15. HA213373	APARTMENT
16. HA217768	RESIDENTIAL YARD (FRONT/BACK)
17. HA230689	GAS STATION
18. HA236717	RESIDENCE
19. HA244376	RESTAURANT
20. HA255728	APARTMENT
21. HA301809	STREET
22. HA304905	SIDEWALK

Apply Cancel Enter Esc

12:40 PM 12/15/2020

Facet / Filter Undo / Redo 5 / 5

LocationDescription

136 choices Sort by: name count Cluster

JAIL / LOOK-UP FACILITY 373

LAKEFRONT/WATERFRONT/RIVERBANK 35

LAUNDRY ROOM 1

LIBRARY 1345

LIQUOR STORE 2

MEDICAL/DENTAL OFFICE 1477

MOTEL 1

MOVIE HOUSE/THEATER 452

NEWSSTAND 31

NURSING HOME/RETIREMENT HOME 3060

OFFICE 2

OTHER 53474

OTHER COMMERCIAL TRANSPORTATION 689

OTHER RAILROAD PROP / TRAIN DEPOT 1174

PARK PROPERTY 12265

PARKING LOT 36

Activate Windows
Go to Settings to activate Windows.

1428281 rows

Show as: rows records Show: 5 10 25 50 rows

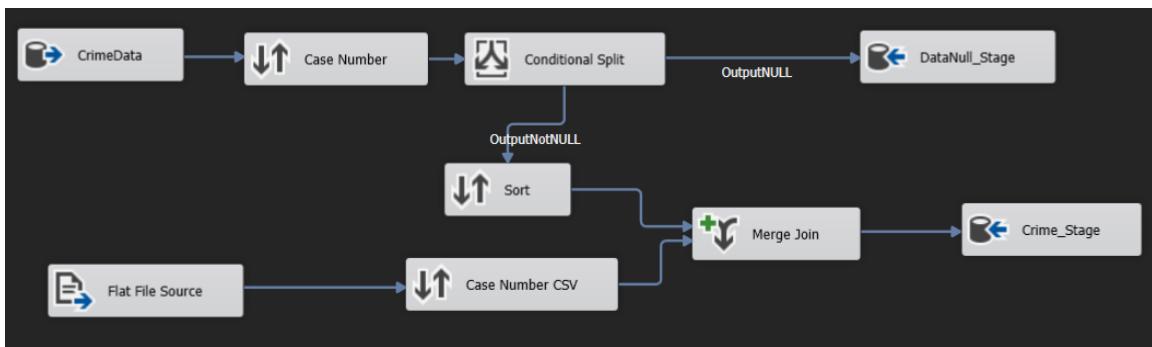
All CaseNumber LocationDescription

CaseNumber	LocationDescription
1. 223432	HOSPITAL BUILDING/GROUNDS
2. 311997	COMMERCIAL / BUSINESS OFFICE
3. 318876	STREET
4. 322338	CTA PLATFORM
5. 413567	RESIDENCE
6. 536481	APARTMENT
7. F218264	STREET
8. HA107183	CONVENIENCE STORE
9. HA121046	ATM (AUTOMATIC TELLER MACHINE)
10. HA156050	RESIDENCE
11. HA164684	SIDEWALK
12. HA168845	GROCERY FOOD STORE
13. HA179595	RESIDENCE
14. HA194202	RESIDENTIAL YARD (FRONT/BACK)
15. HA213373	APARTMENT
16. HA217768	RESIDENTIAL YARD (FRONT/BACK)
17. HA230689	GAS STATION
18. HA236717	RESIDENCE
19. HA244376	RESTAURANT
20. HA255728	APARTMENT
21. HA301809	STREET
22. HA304905	SIDEWALK

Apply Cancel Enter Esc

12:41 PM 12/15/2020

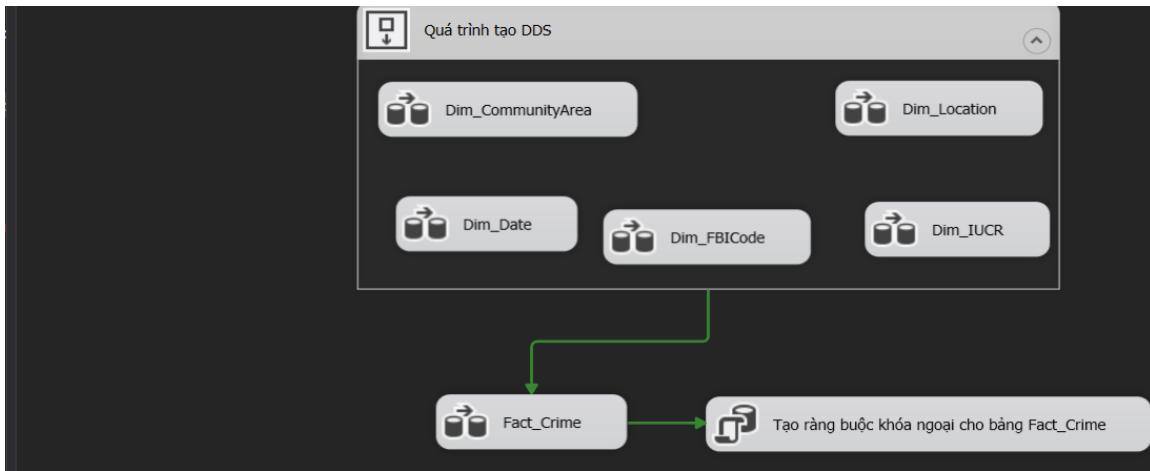
5) Làm sạch dữ liệu null



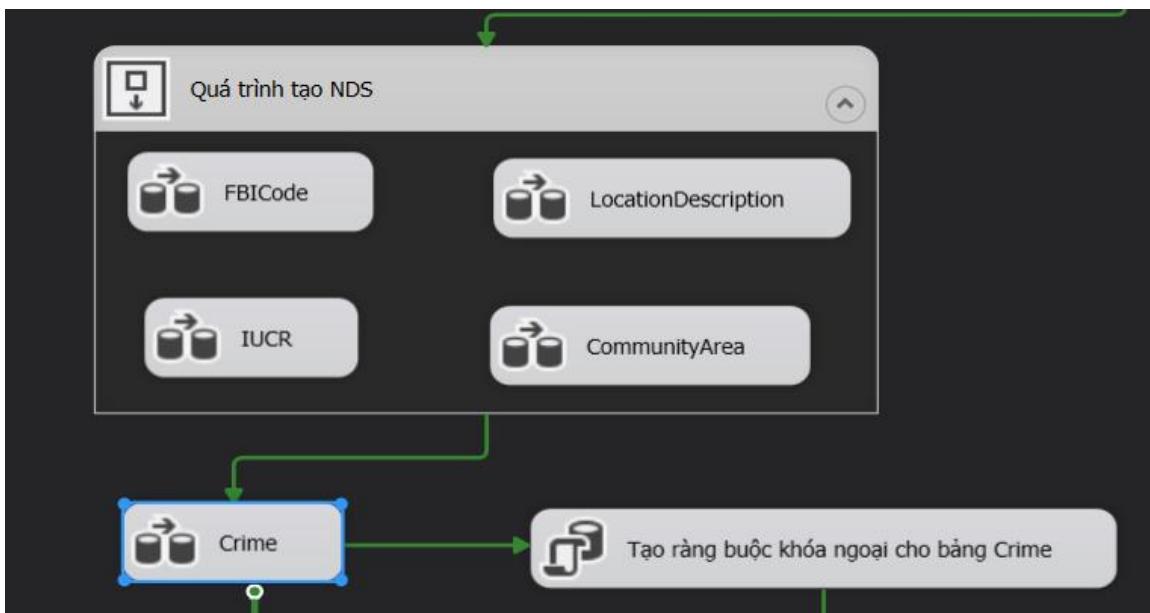
Đầu tiên ta sẽ load dữ liệu từ nguồn (cơ sở dữ liệu Chicago_Crime), sau đó ta sẽ sort để loại bỏ các thuộc tính trong Case Number mà trùng lắp dữ liệu. Tiếp đó ta sẽ tách dữ giá trị bị thiếu (có giá trị NULL) vào bảng DataNull_Stage (trong cơ sở dữ liệu Stage_ChicagoCrime). VỚI CÁC GIÁ TRỊ KHÁC NULL 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.

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

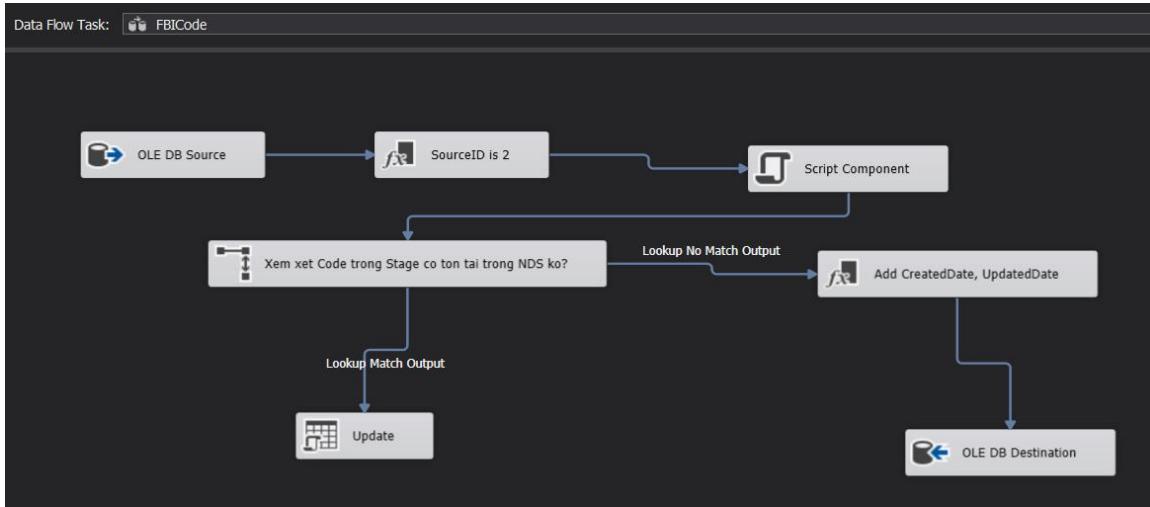




Tai quá trình nạp dữ liệu vào NDS

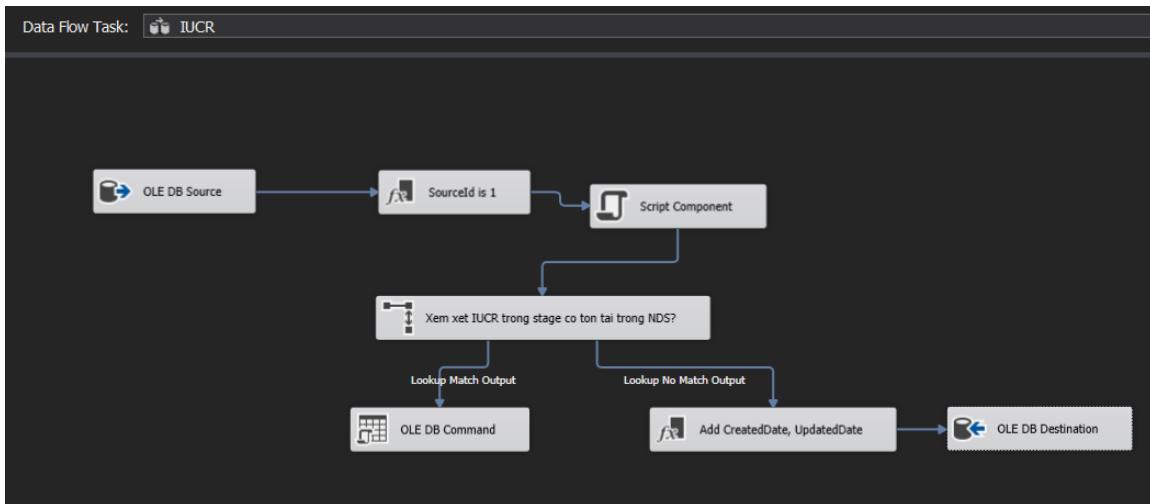


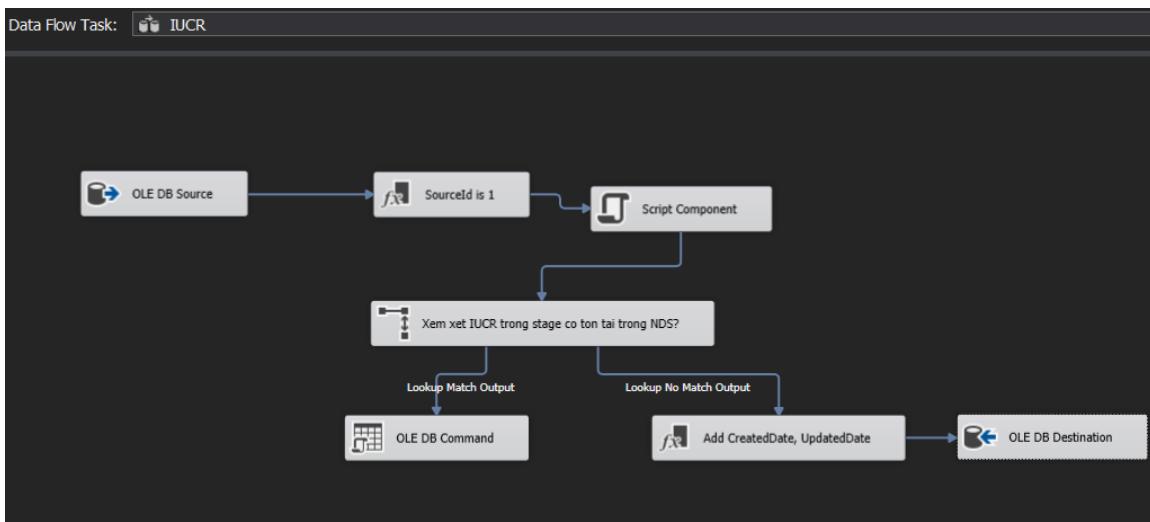
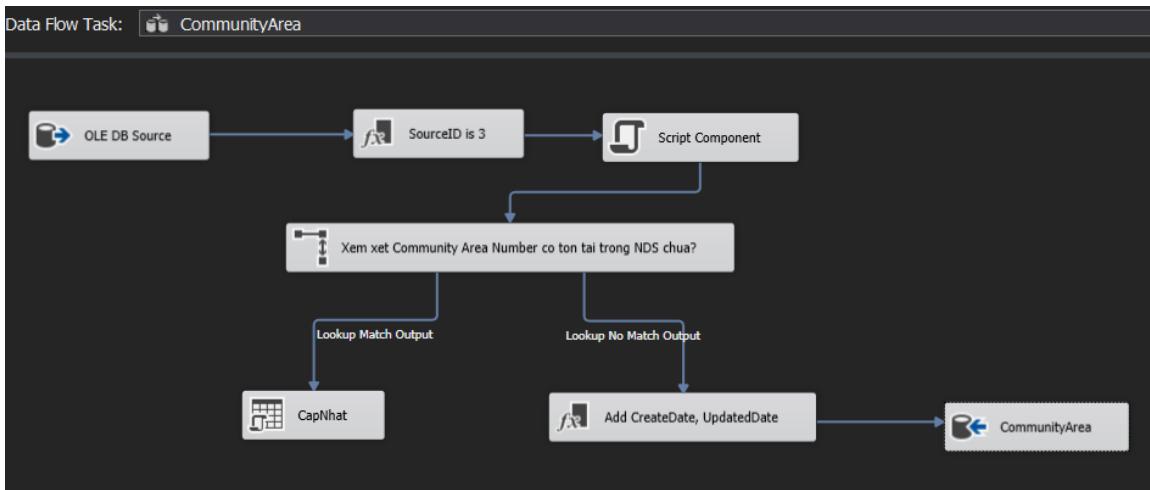
Ta sẽ nạp các chiề (gồm có FBICode, Community Area, IUCR, Location Description) vào trong Sequence Container. Ở mỗi Data Flow Task ta sẽ thực hiện chèn dữ liệu vào NDS như sau:



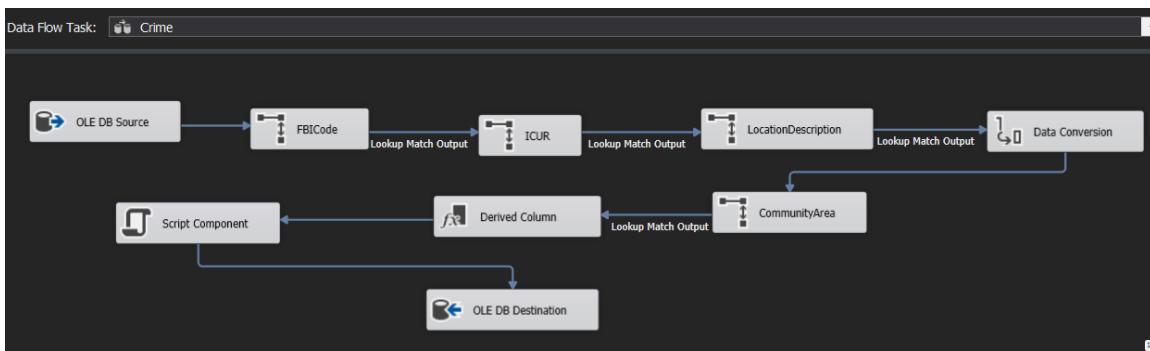
Tại bảng chiề FBI Code ta sẽ thêm một ghi nguồn được quy định trong cơ sở dữ liệu meta data trong bảng data flow. Tiếp theo ta sẽ thêm 1 cột chứa giá trị tự tăng trong component Script Component.Sau đó ta sẽ kiểm tra ứng Code có tồn tại trong NDS chưa. Nếu chưa tồn tại thì ta chèn vào bảng FBICode trong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

Làm tương tự cho các bảng chiề.



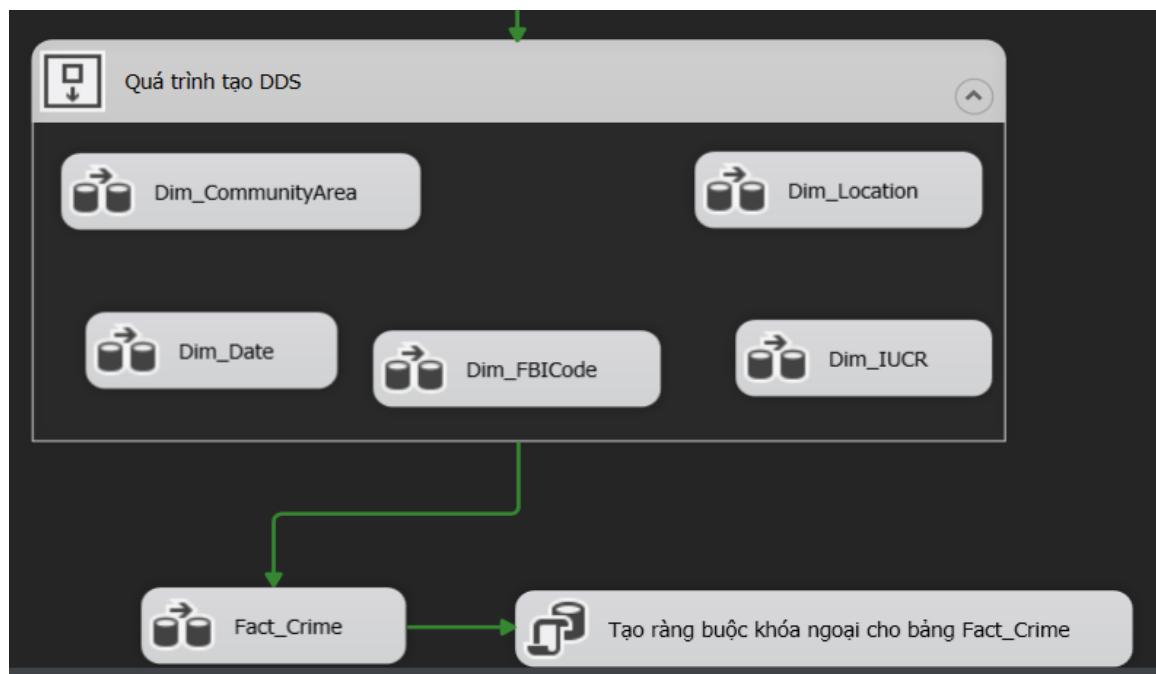


Khi các bảng chiêu đã được nạp dữ liệu, ta sẽ nạp dữ liệu vào bảng fact (bảng Crime trong NDS).

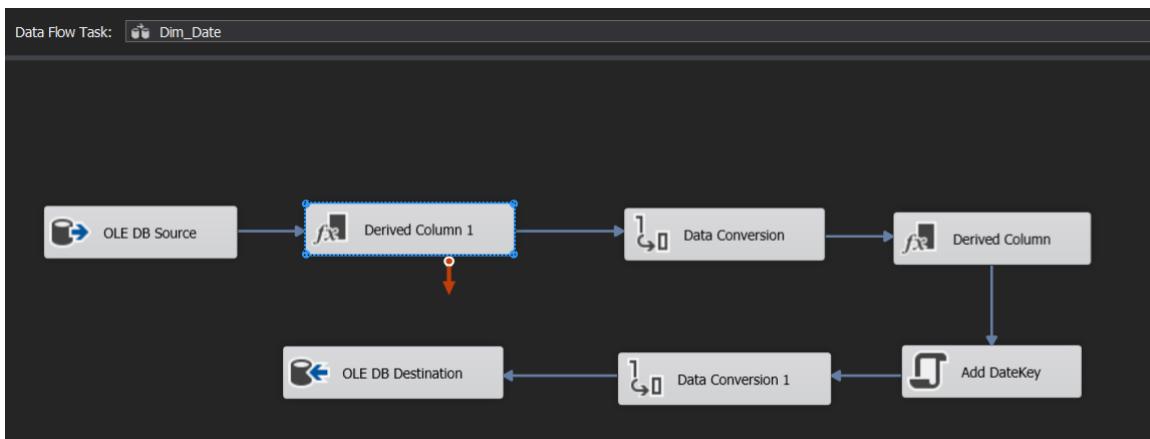


Ta sẽ nạp dữ liệu từ stage, sau đó sẽ lookup trên các bảng chiều để tham chiếu khóa ngoại. Cuối cùng sẽ tạo ràng buộc khóa ngoại cho bảng fact.

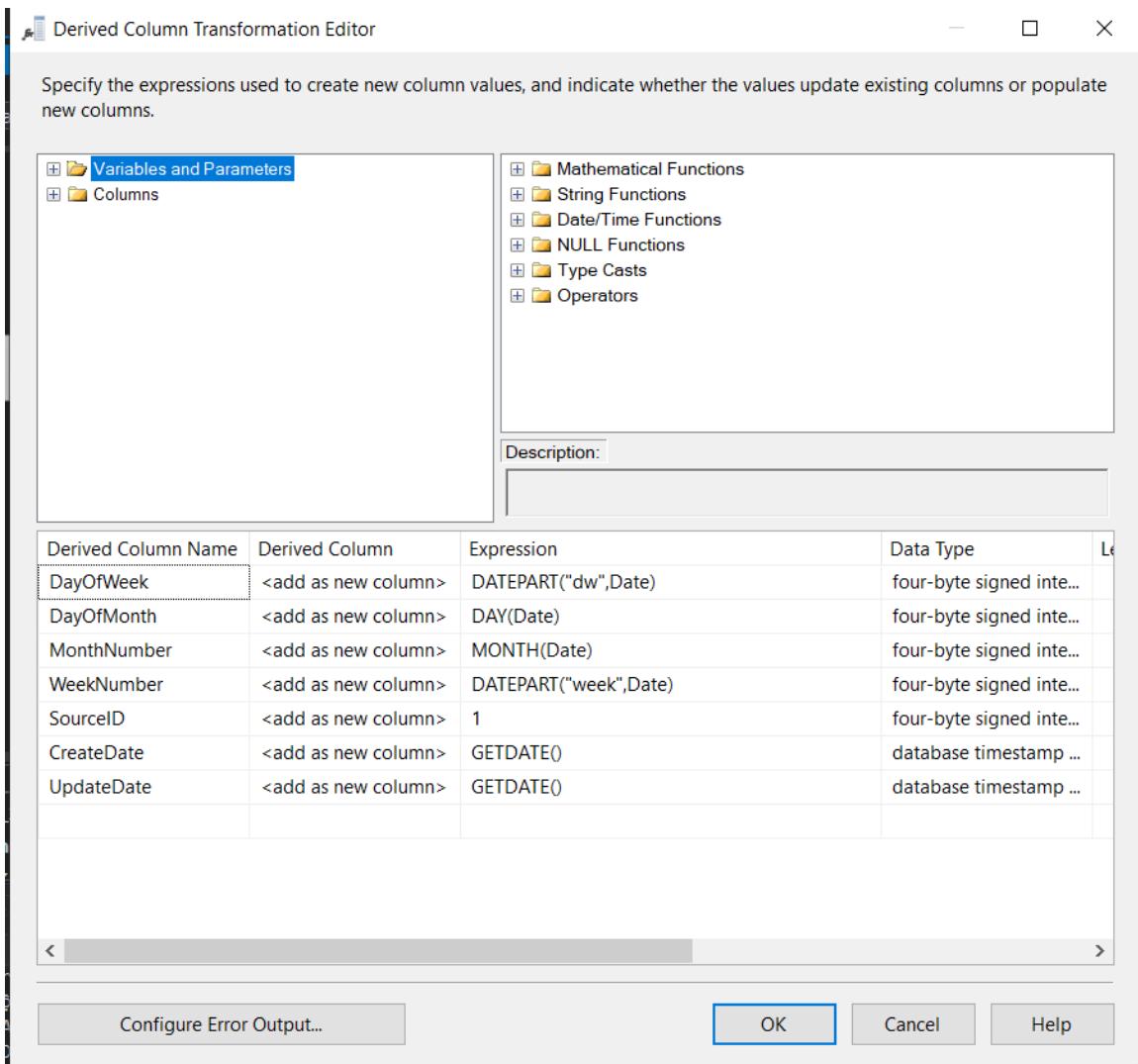
Tai quá trình nạp dữ liệu vào DDS



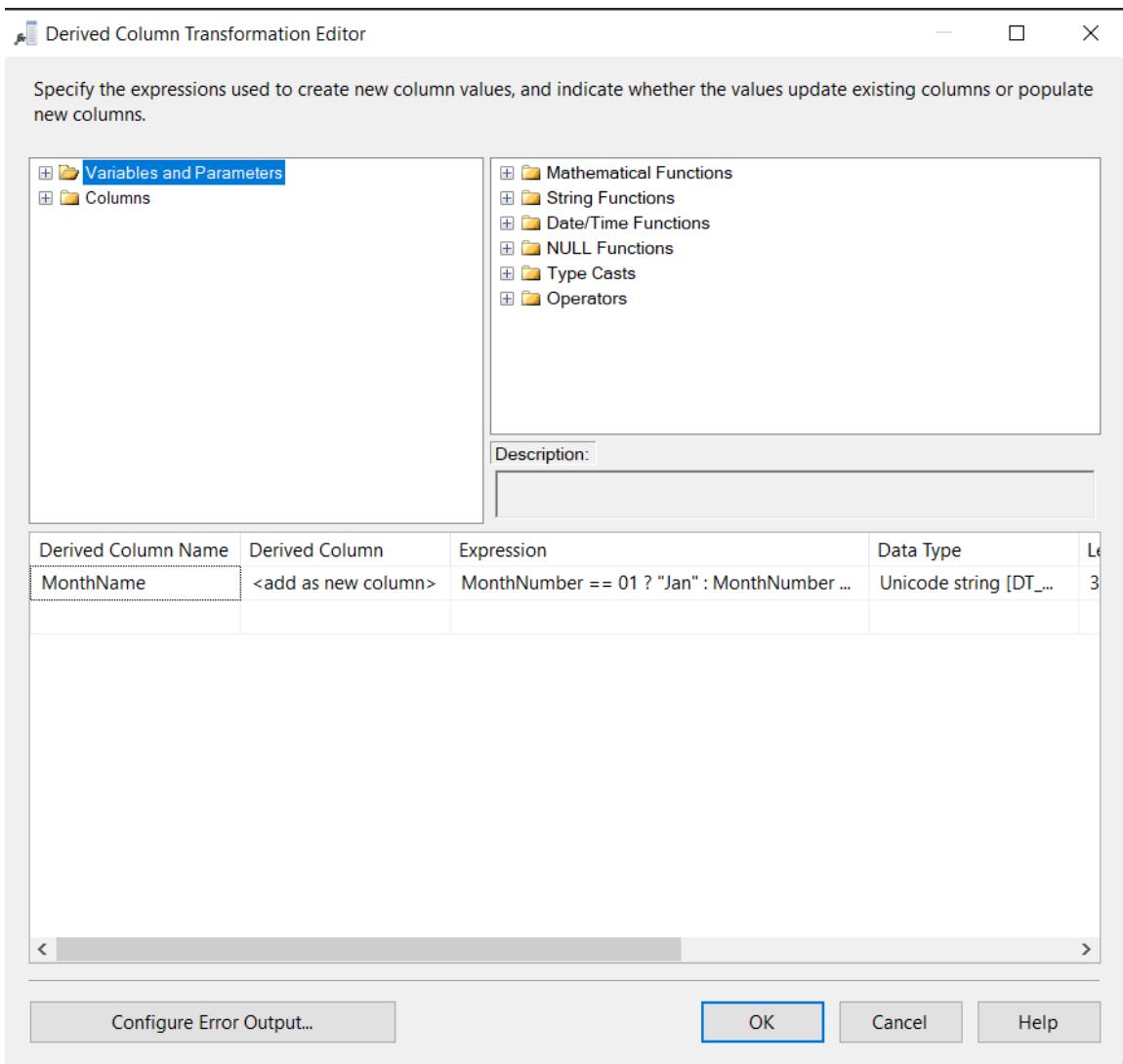
Ta sẽ nạp bảng chiều (gồm Dim_CommunityArea, Dim_Location, Dim_Date, Dim_FBICode, Dim_IUCR) và bảng fact (Fact_Crime)
Đối với bảng Dim_Date



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



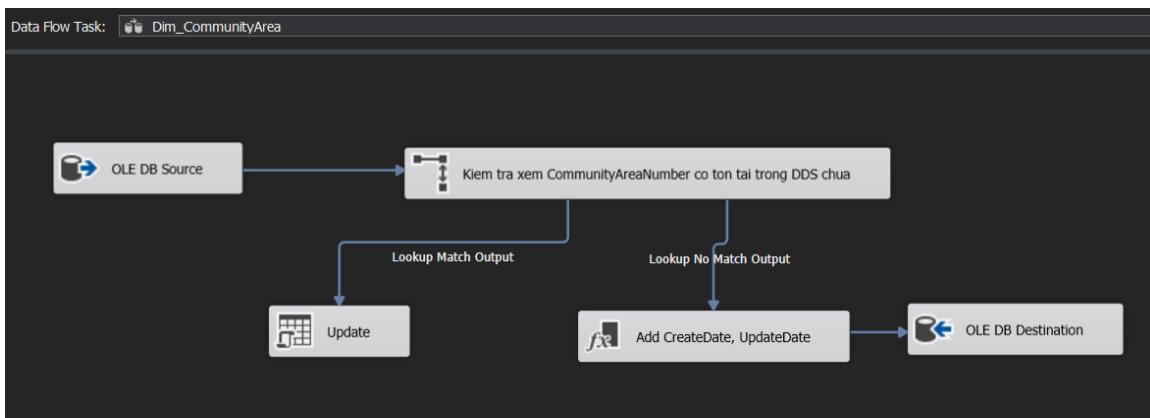
Cột Expression:

```

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"

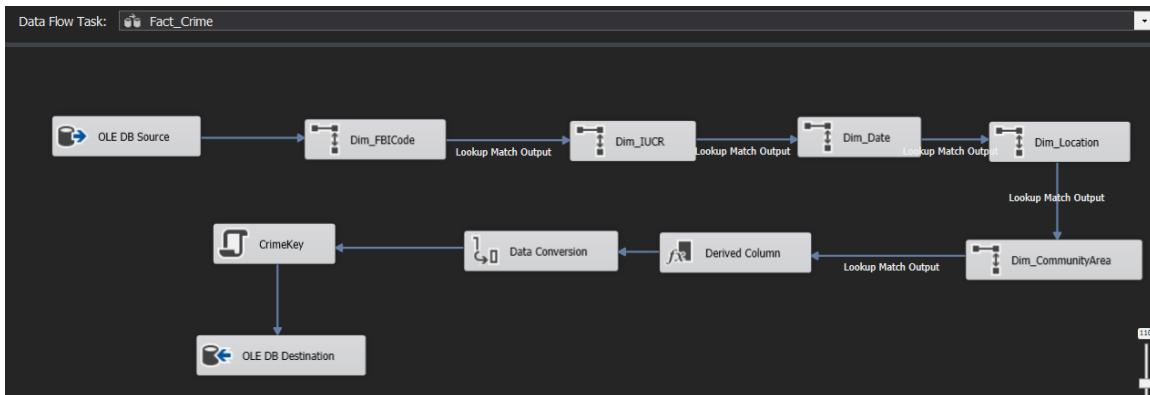
```

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.

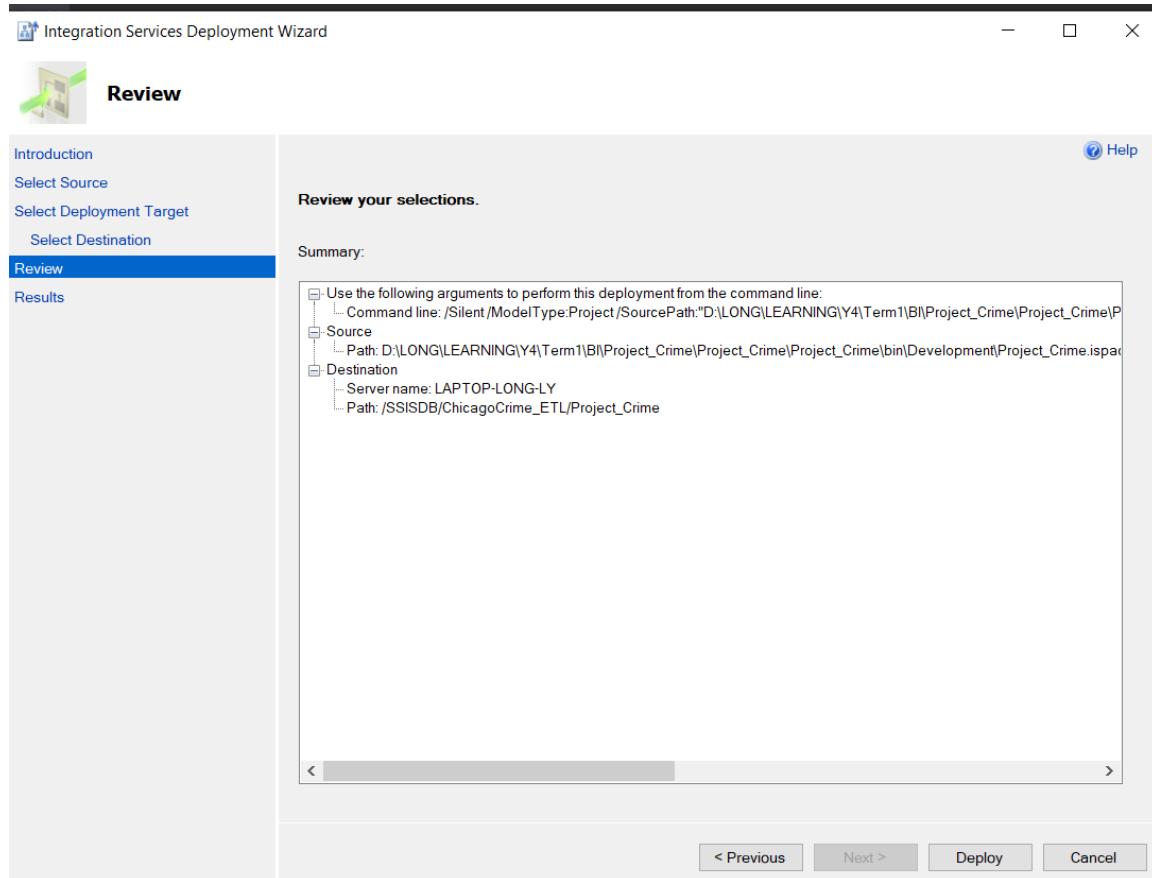
Link video demo quá trình ETL từ Stage->NDS->DDS

https://youtu.be/P_kNraX-bmQ

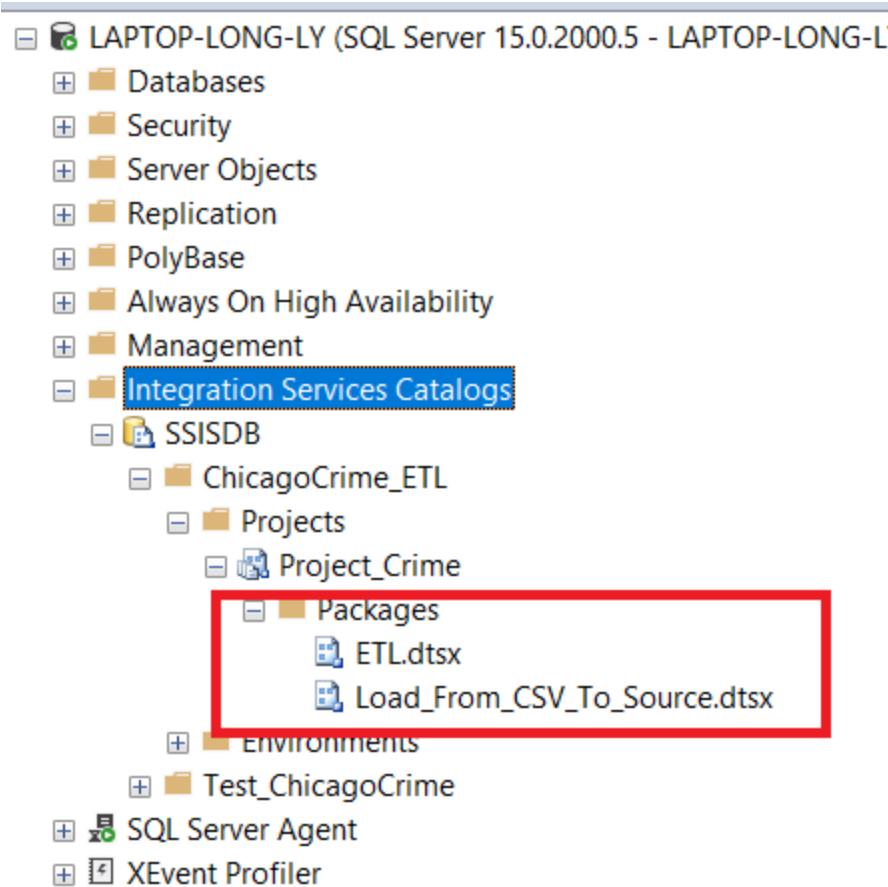
VIII. Lập lịch định kỳ cho ETL bằng cách deploy packages

Các bước thực hiện như sau:

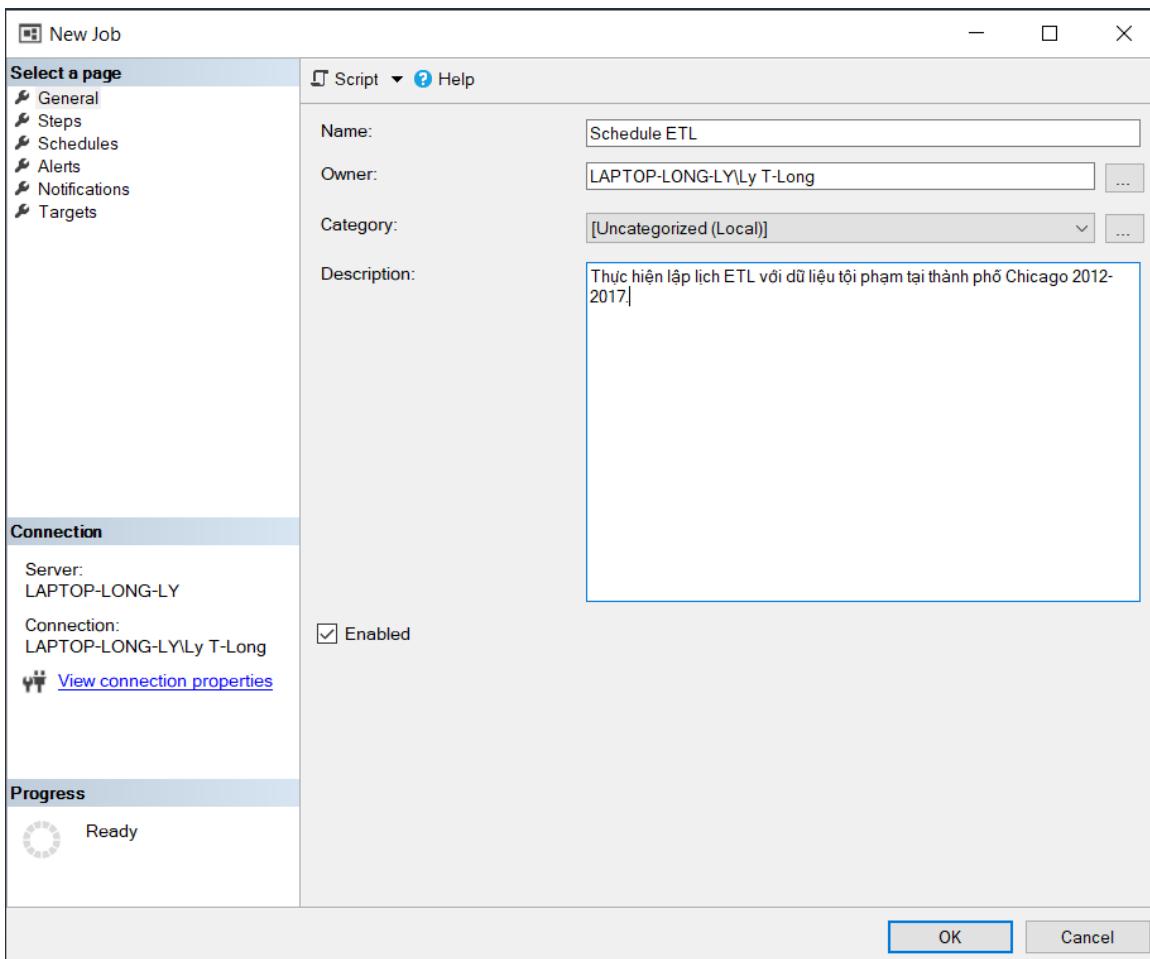
Đầu tiên ta sẽ deploy package ETL trong visual studio



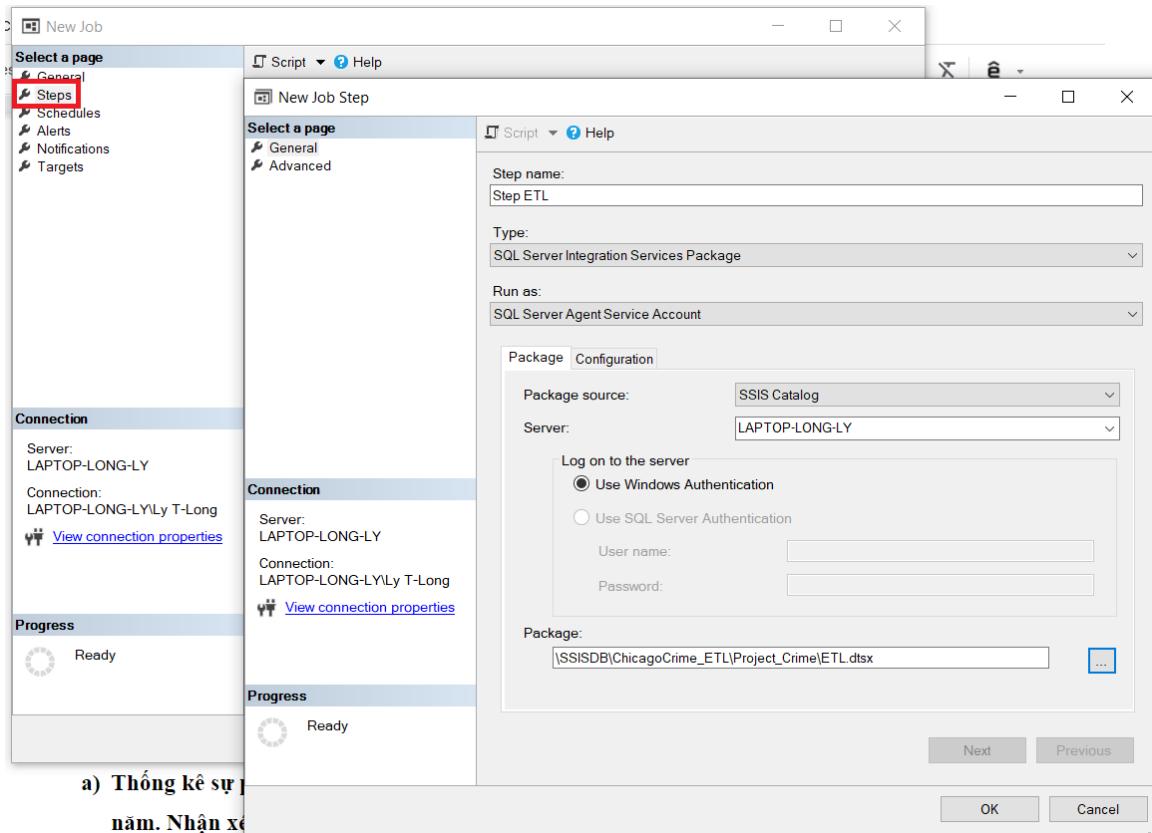
Sau đó kiểm tra trong SQL Server trong Database Engine trong Intergration Services Catalogs bên trong SSISDB trong đã deploy thành công chưa.



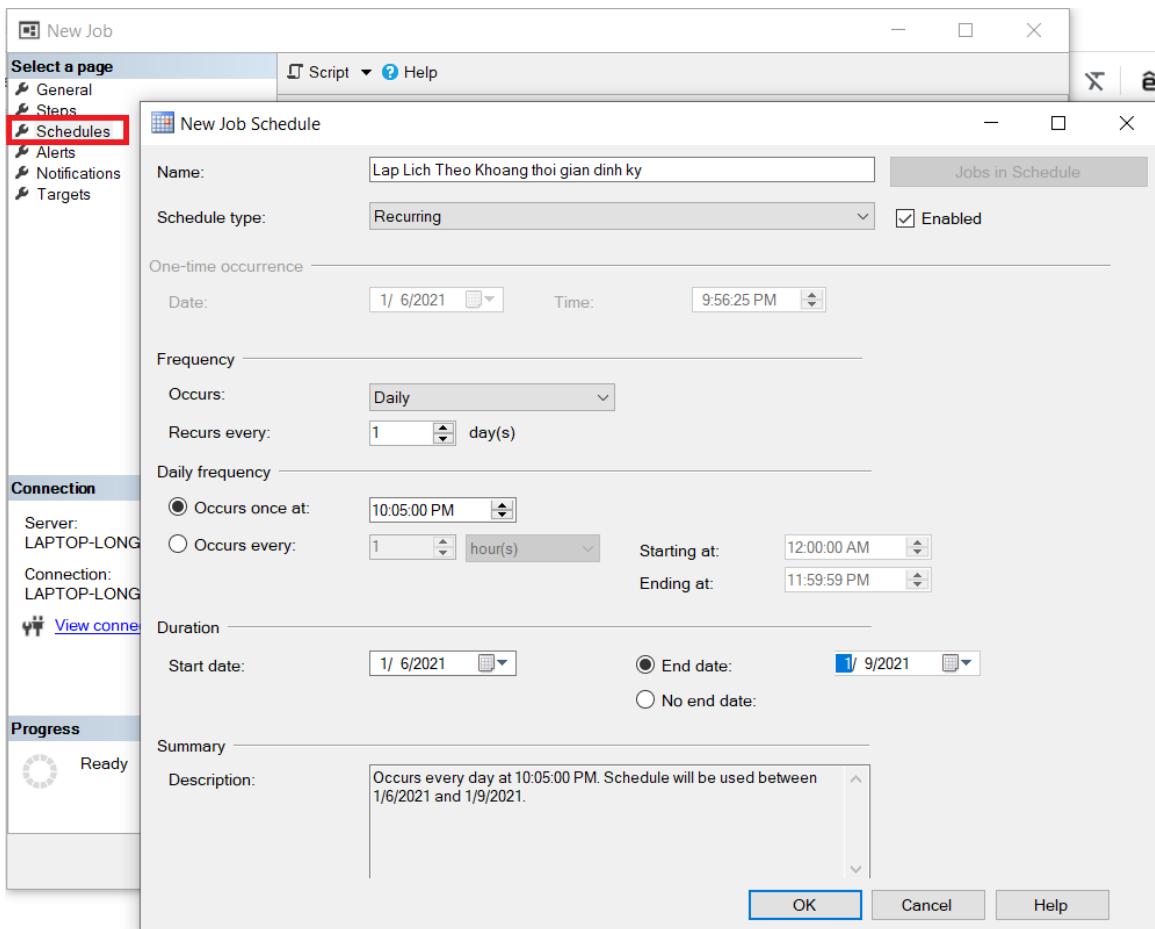
Tiếp đó ta vào SQL Server Agent để folder Jobs để tiến hành lập lịch ETL bằng cách nhấp vào “New Job”



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>)

IX. Khai thác dữ liệu

1) Report:

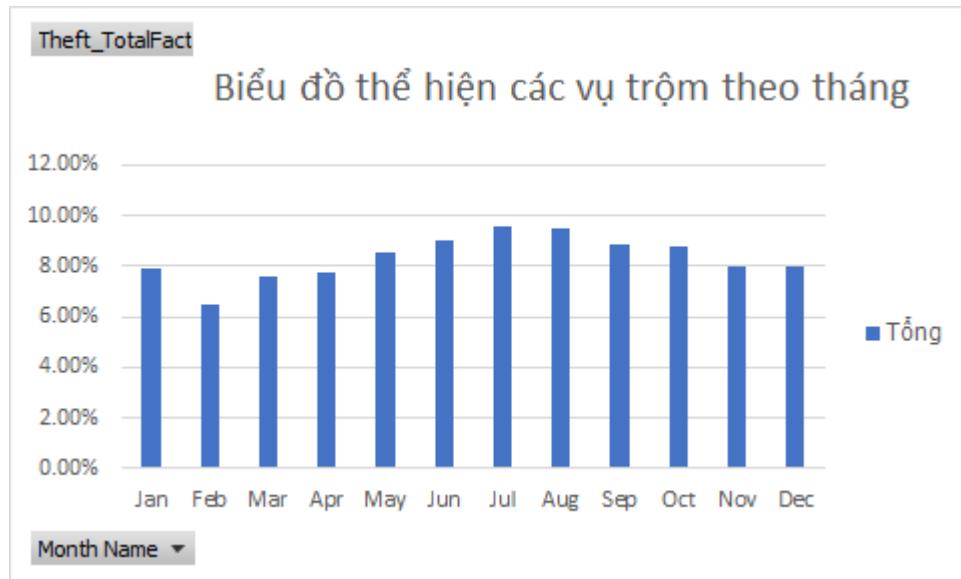
- a) **Thống kê sự phân phối của các vụ trộm theo thời gian tháng, năm. Nhận xét khoảng thời gian xảy ra nhiều nhất, ít nhất các vụ trộm...**

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. Điểm đỉnh 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

b) 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]

Arrest	Percentage_Theft_TotalFact	Theft_TotalFact
False	89.41%	344086
True	10.59%	40754
Tổng Cuối	100.00%	384840

c) **Thông 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:

$([\text{Dim IUCR}].[\text{Primary Type}].\&[\text{THEFT}],[\text{Measures}].[\text{Frequency}]) / [\text{Measures}].[\text{Frequency}]$

Còn tỉ lệ phạm tội khác thì $1 - \text{Theft}(\%)$

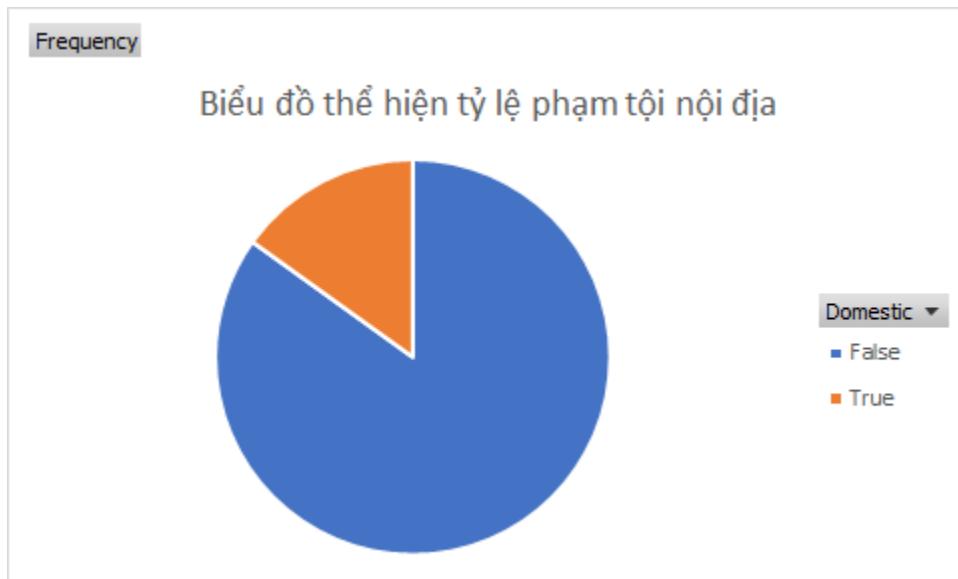
Location Description	Theft(%)	Other(%)
"CTA ""L"" PLATFORM"	100.00%	
"CTA ""L"" TRAIN"	100.00%	
"SCHOOL, PRIVATE, BUILDING"	28.75%	71.25%
"SCHOOL, PRIVATE, GROUNDS"	25.68%	74.32%
"SCHOOL, PUBLIC, BUILDING"	23.53%	76.47%
"SCHOOL, PUBLIC, GROUNDS"	16.47%	83.53%
ABANDONED BUILDING	17.18%	82.82%
AIRCRAFT	51.23%	48.77%
AIRPORT BUILDING NON-TERMINAL - NON-SECURE AREA	56.04%	43.96%
AIRPORT BUILDING NON-TERMINAL - SECURE AREA	70.86%	29.14%
AIRPORT EXTERIOR - NON-SECURE AREA	35.07%	64.93%
AIRPORT EXTERIOR - SECURE AREA	56.10%	43.90%
AIRPORT PARKING LOT	48.68%	51.32%
AIRPORT TERMINAL LOWER LEVEL - NON-SECURE AREA	32.21%	67.79%
AIRPORT TERMINAL LOWER LEVEL - SECURE AREA	75.47%	24.53%
AIRPORT TERMINAL MEZZANINE - NON-SECURE AREA	55.00%	45.00%
AIRPORT TERMINAL UPPER LEVEL - NON-SECURE AREA	40.85%	59.15%
AIRPORT TERMINAL UPPER LEVEL - SECURE AREA	20.99%	79.01%
AIRPORT TRANSPORTATION SYSTEM (ATS)	54.72%	45.28%
AIRPORT VENDING ESTABLISHMENT	40.25%	59.75%
AIRPORT/AIRCRAFT	14.64%	85.36%
ALLEY	11.67%	88.33%
ANIMAL HOSPITAL	18.15%	81.85%
APARTMENT	11.13%	88.87%
APPLIANCE STORE	43.24%	56.76%
ATHLETIC CLUB	83.53%	16.47%
ATM (AUTOMATIC TELLER MACHINE)	8.69%	91.31%
AUTO		100.00%

Nhân 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 $\frac{3}{4}$ so với tội phạm khác.

d) 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 $\frac{1}{4}$ so với tội phạm không phải nội địa.

2) OLAP

a) 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]
```

	2012	2013	2014	2015	2016	2017
All	334439	304609	270394	260399	254238	2976
ARSON	469	364	394	448	508	8
ASSAULT	19891	17939	16720	16971	18247	211
BATTERY	59120	53895	48823	48655	49039	380
BURGLARY	22835	17830	14329	13094	13699	74
CONCEALED CARRY LICENSE VIOLATION	(null)	(null)	15	34	35	4
CRIM SEXUAL ASSAULT	1352	1177	1224	1305	1296	17
CRIMINAL DAMAGE	35849	30796	27558	28560	30339	271
CRIMINAL TRESPASS	8214	8131	7510	6384	6237	144
DECEPTIVE PRACTICE	12866	12415	14426	14626	14065	152
GAMBLING	724	596	393	310	189	(null)
HOMICIDE	478	406	415	445	715	28
HUMAN TRAFFICKING	(null)	(null)	1	13	7	(null)
INTERFERENCE WITH PUBLIC OFFICER	1228	1281	1397	1306	928	8
INTIMIDATION	156	133	116	120	121	(null)
KIDNAPPING	236	242	216	190	196	1
LIQUOR LAW VIOLATION	573	465	395	289	218	3
MOTOR VEHICLE THEFT	16484	12550	9812	10046	11100	144
NARCOTICS	35467	34109	28413	23225	11337	93
NON-CRIMINAL	6	7	26	34	47	(null)
NON-CRIMINAL (SUBJECT SPECIFIED)	2	(null)	1	(null)	1	(null)
OBSCENITY	26	21	32	45	48	(null)
OFFENSE INVOLVING CHILDREN	2112	2193	2200	2150	1974	17
OTHER NARCOTIC VIOLATION	6	5	10	5	4	(null)
OTHER OFFENSE	17456	17859	16779	17421	16082	108
PROSTITUTION	2203	1651	1616	1316	780	1
PUBLIC INDECENCY	17	10	10	14	10	1
PUBLIC PEACE VIOLATION	3007	3134	2890	2409	1584	16
ROBBERY	13480	11790	9665	9576	11677	154
SEX OFFENSE	1010	952	876	907	797	4
STALKING	207	146	135	149	145	2
THEFT	75067	71268	60889	57009	59411	1103
WEAPONS VIOLATION	3898	3244	3108	3343	3402	32

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.

b) Thông 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]
```

Location Description	2012	2013	2014	2015	2016	2017
CTA "L" PLATFORM	(null)	1	(null)	(null)	(null)	(null)
CTA "L" TRAIN	(null)	(null)	(null)	(null)	1	(null)
SCHOOL, PRIVATE, BUILDING	702	680	499	572	581	23
SCHOOL, PRIVATE, GROUNDS	234	213	208	175	190	4
SCHOOL, PUBLIC, BUILDING	6199	6524	5244	4165	3717	108
SCHOOL, PUBLIC, GROUNDS	1522	1459	1222	1083	1096	18
ABANDONED BUILDING	1036	977	679	525	471	15
AIRCRAFT	46	61	74	98	88	(null)
AIRPORT BUILDING NON-TERMINAL - NON-SECURE AREA	102	67	73	76	92	4
AIRPORT BUILDING NON-TERMINAL - SECURE AREA	54	59	50	67	69	3
AIRPORT EXTERIOR - NON-SECURE AREA	91	41	60	57	91	5
AIRPORT EXTERIOR - SECURE AREA	25	36	26	43	33	1
AIRPORT PARKING LOT	65	76	64	86	85	4
AIRPORT TERMINAL LOWER LEVEL - NON-SECURE AREA	149	121	163	205	184	10
AIRPORT TERMINAL LOWER LEVEL - SECURE AREA	74	58	57	71	59	3
AIRPORT TERMINAL MEZZANINE - NON-SECURE AREA	7	10	3	10	8	2
AIRPORT TERMINAL UPPER LEVEL - NON-SECURE AREA	76	40	52	70	65	3
AIRPORT TERMINAL UPPER LEVEL - SECURE AREA	412	753	662	342	255	11
AIRPORT TRANSPORTATION SYSTEM (ATS)	11	10	9	11	12	(null)
AIRPORT VENDING ESTABLISHMENT	77	121	164	151	136	2
AIRPORT/AIRCRAFT	118	64	74	97	9	(null)
ALLEY	7763	6826	5968	5552	5125	3
ANIMAL HOSPITAL	61	52	47	63	57	1
APARTMENT	41138	37997	34280	34233	31782	1
APPLIANCE STORE	82	56	51	61	81	2
ATHLETIC CLUB	656	518	533	483	598	36
ATM (AUTOMATIC TELLER MACHINE)	639	559	398	324	651	18
AUTO	51	29	23	47	81	4
BANK	1341	1049	940	884	856	(null)
BAR OR TAVERN	2221	2174	1827	1724	1822	103
BARBER SHOP/BEAUTY SALON	312	301	200	236	213	(null)
BASEMENT	1	(null)	1	(null)	(null)	(null)
BOAT/WATERCRAFT	36	34	44	33	31	(null)
BOWLING ALLEY	29	23	18	26	30	4
BRIDGE	28	20	20	23	27	(null)

SCHOOL YARD	1	(null)	(null)	(null)	1	(null)
SIDEWALK	40880	37219	30416	27550	22413	(null)
SMALL RETAIL STORE	6078	5660	5466	5485	5837	275
SPORTS ARENA/STADIUM	311	307	248	325	332	9
STAIRWELL	(null)		1 (null)		1 (null)	(null)
STREET	76141	66757	62369	60423	59334	17
TAVERN	(null)		1	2	2	1
TAVERN/LIQUOR STORE	781	648	573	517	491	18
TAXICAB	424	402	422	406	442	(null)
TRUCK	(null)	(null)	(null)		1 (null)	(null)
VACANT LOT/LAND	1625	1678	1303	1124	923	32
VEHICLE - DELIVERY TRUCK	(null)	(null)		2	21	19
VEHICLE - OTHER RIDE SERVICE		1 (null)	(null)		64	119
VEHICLE NON-COMMERCIAL	5487	5326	4802	4675	4642	167
VEHICLE-COMMERCIAL	294	262	249	237	207	(null)
VESTIBULE		1	3	1 (null)		1 (null)
WAREHOUSE	403	327	264	244	226	(null)
YARD	11	17	11	14	15	(null)

CAR WASH	131	102	99	118	114	4
CEMETARY	20	15	23	6	13	(null)
CHA APARTMENT	940	850	759	808	741	36
CHA HALLWAY/STAIRWELL/ELEVATOR	182	217	192	160	129	13
CHA PARKING LOT/GROUNDS	1017	1000	796	755	598	12
CHURCH PROPERTY	1 (null)	(null)	(null)	(null)	(null)	(null)
CHURCH/SYNAGOGUE/PLACE OF WORSHIP	842	663	544	541	558	27
CLEANERS/LAUNDROMAT	(null)	(null)	(null)	(null)	1	(null)
CLEANING STORE	128	116	101	83	92	2
CLUB	(null)		1 (null)	(null)	(null)	(null)
COIN OPERATED MACHINE	22	35	24	10	35	6
COLLEGE/UNIVERSITY GROUNDS	261	252	218	178	165	10
COLLEGE/UNIVERSITY RESIDENCE HALL	98	81	58	51	52	2
COMMERCIAL / BUSINESS OFFICE	1830	1683	1361	1190	1574	67
CONSTRUCTION SITE	385	336	296	267	340	6
CONVENIENCE STORE	1213	1213	1204	1375	1644	76
CREDIT UNION	28	9	17	22	22	(null)
CTA BUS	1452	1560	1218	959	936	32
CTA BUS STOP	811	955	673	599	545	17
CTA GARAGE / OTHER PROPERTY	1063	1101	460	248	220	14
CTA PLATFORM	2702	2092	1139	608	674	29
CTA STATION	(null)	(null)	559	661	726	42
CTA TRACKS - RIGHT OF WAY	(null)	(null)		13	24	15 (null)
CTA TRAIN	1707	1705	1304	1016	1320	82
CURRENCY EXCHANGE	517	472	458	452	383	12
DAY CARE CENTER	141	129	122	151	161	11
DELIVERY TRUCK	26	25	14	1 (null)	(null)	
DEPARTMENT STORE	4152	4084	3996	4124	4185	167
DRIVEWAY	(null)	(null)		3	1	1 (null)
DRIVEWAY - RESIDENTIAL	688	649	707	714	812	50
DRUG STORE	1132	1070	1009	919	1169	54

CSC12107 – HTTT PV TRÍ TUỆ KINH DOANH

ELEVATOR	(null)	(null)	(null)	(null)	1	(null)
EXPRESSWAY EMBANKMENT	(null)	(null)	(null)	(null)	1	(null)
FACTORY/MANUFACTURING BUILDING	269	181	160	125	128	9
FEDERAL BUILDING	30	27	36	46	36	(null)
FIRE STATION	30	43	34	40	51	2
FOREST PRESERVE	11	16	9	13	6	(null)
GANGWAY	3	3	3	2	2	(null)
GARAGE	3	3	2	(null)	3	(null)
GARAGE/AUTO REPAIR	(null)	1	(null)	(null)	(null)	(null)
GAS STATION	3195	3122	2844	2984	3040	202
GAS STATION DRIVE/PROP.	6	3	1	1	1	(null)
GOVERNMENT BUILDING	(null)	(null)	1	(null)	(null)	(null)
GOVERNMENT BUILDING/PROPERTY	769	703	633	495	475	21
GROCERY FOOD STORE	3556	3259	2937	2947	3165	134
HALLWAY	5	8	4	3	5	(null)
HIGHWAY/EXPRESSWAY	50	38	37	39	40	(null)
HOSPITAL	(null)	(null)	(null)	(null)	1	(null)
HOSPITAL BUILDING/GROUNDS	902	1005	887	1018	1102	56
HOTEL	1	1	(null)	(null)	1	(null)
HOTEL/MOTEL	1292	1270	1212	1249	1186	(null)
HOUSE	19	16	24	26	26	(null)
JAIL / LOCK-UP FACILITY	55	60	87	81	88	2
LAKEFRONT/WATERFRONT/RIVERBANK	77	69	49	59	76	(null)
LAUNDRY ROOM	(null)	(null)	(null)	(null)	1	(null)
LIBRARY	278	308	301	223	235	(null)
LIQUOR STORE	1	(null)	(null)	1	(null)	(null)
MEDICAL/DENTAL OFFICE	351	311	266	256	281	12
MOTEL	(null)	1	(null)	(null)	(null)	(null)
MOVIE HOUSE/THEATER	99	90	69	90	102	2
NEWSSTAND	5	8	6	10	2	(null)
NURSING HOME/RETIREMENT HOME	526	519	583	709	683	40
OFFICE	(null)	(null)	(null)	1	1	(null)
OTHER	11560	11152	10447	10365	9950	(null)
OTHER COMMERCIAL TRANSPORTATION	143	165	137	109	130	5
OTHER RAILROAD PROP / TRAIN DEPOT	302	227	236	220	179	10
PARK PROPERTY	2862	2722	2356	2200	2102	23
PARKING LOT	4	7	5	7	13	(null)
PARKING LOT/GARAGE(NON.RESID.)	9837	8536	7458	7432	8106	397
PAWN SHOP	32	47	50	47	56	1
POLICE FACILITY/VEH PARKING LOT	939	894	819	857	733	19
POOL ROOM	83	63	58	61	69	7
PORCH	21	11	21	13	16	2
PUBLIC HIGH SCHOOL	1	(null)	(null)	(null)	(null)	(null)
RAILROAD PROPERTY	(null)	(null)	(null)	(null)	1	(null)
RESIDENCE	52995	47999	41846	40228	40786	(null)
RESIDENCE PORCH/HALLWAY	5879	5299	4861	4776	4524	171
RESIDENCE-GARAGE	6233	5146	4387	4458	5031	(null)
RESIDENTIAL YARD (FRONT/BACK)	7536	6424	5450	5510	5549	175
RESTAURANT	5179	4807	4836	4963	5411	(null)
RETAIL STORE	3	2	5	(null)	5	(null)
SAVINGS AND LOAN	12	11	9	12	13	1

c) Thống kê tần suất theo các loại tội phạm

Primary Type	Frequency
ARSON	2191
ASSAULT	89979
BATTERY	259912
BURGLARY	81861
CONCEALED CARRY LICENSE VIOLATION	88
CRIM SEXUAL ASSAULT	6371
CRIMINAL DAMAGE	153373
CRIMINAL TRESPASS	36620
DECEPTIVE PRACTICE	68550
GAMBLING	2212
HOMICIDE	2487
HUMAN TRAFFICKING	21
INTERFERENCE WITH PUBLIC OFFICER	6148
INTIMIDATION	646
KIDNAPPING	1081
LIQUOR LAW VIOLATION	1943
MOTOR VEHICLE THEFT	60136
NARCOTICS	132644
NON-CRIMINAL	120
NON-CRIMINAL (SUBJECT SPECIFIED)	4
OBSCENITY	172
OFFENSE INVOLVING CHILDREN	10646
OTHER NARCOTIC VIOLATION	30
OTHER OFFENSE	85705
PROSTITUTION	7567
PUBLIC INDECENCY	62
PUBLIC PEACE VIOLATION	13040
ROBBERY	56342
SEX OFFENSE	4546
STALKING	784
THEFT	324747
WEAPONS VIOLATION	17027
Tổng Cuối	1427055

Nhân 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.

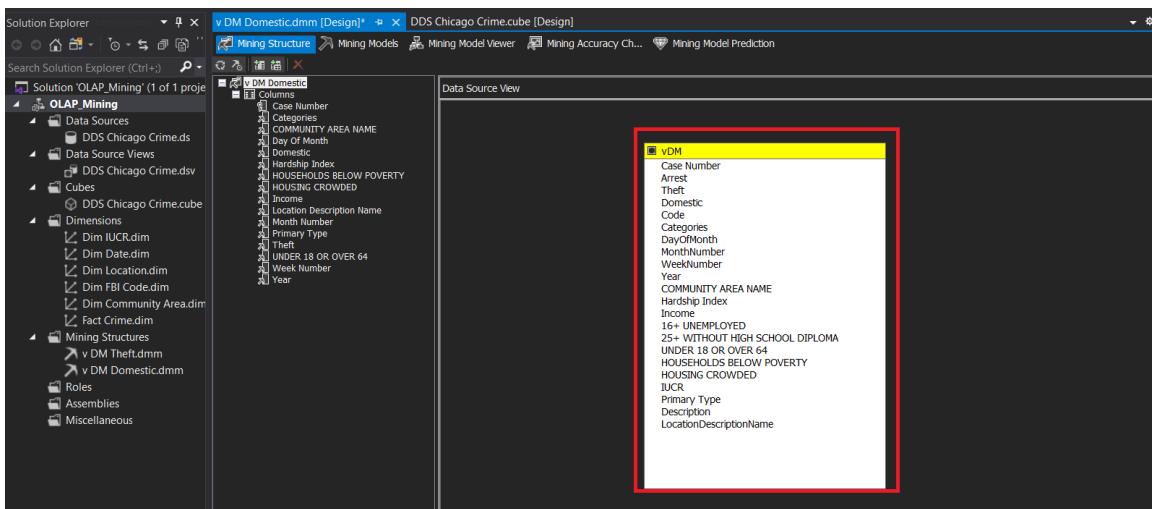
3) Mining

- 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

v DM Theft.dmm [Design] DDS Chicago Crime.cube [Design]

Mining Structure Mining Models Mining Model Viewer Mining Accuracy Ch... Mining Model Prediction

Structure

16 UNEMPLOYED	In
25 WITHOUT HIGH SCHOOL DI...	In
Case Number	Ke
COMMUNITY AREA NAME	Ig
Day Of Month	In
Domestic	In
Hardship Index	In
HOUSEHOLDS BELOW POVERTY	In
HOUSING CROWDED	In
Income	In
Location Description Name	In
Month Number	In
Theft	Pr
UNDER 18 OR OVER 64	In
Week Number	In
Year	Ig

Algorithm Parameters

Parameters:

Parameter	Value	Default	Range
COMPLEXITY_PENALTY			(0.0,1.0)
FORCE_REGRESSOR			
MAXIMUM_INPUT_ATTRIBUTES	255	255	[0,65535]
MAXIMUM_OUTPUT_ATTRIBUTES	255	255	[0,65535]
MINIMUM_SUPPORT	10.0	10.0	(0.0,...)
SCORE_METHOD	4	4	1,3,4
SPLIT_METHOD	2	3	[1,3]

Description:

Specifies the method used to split the node. The available methods are: Binary (1), Complete (2), or Both (3).

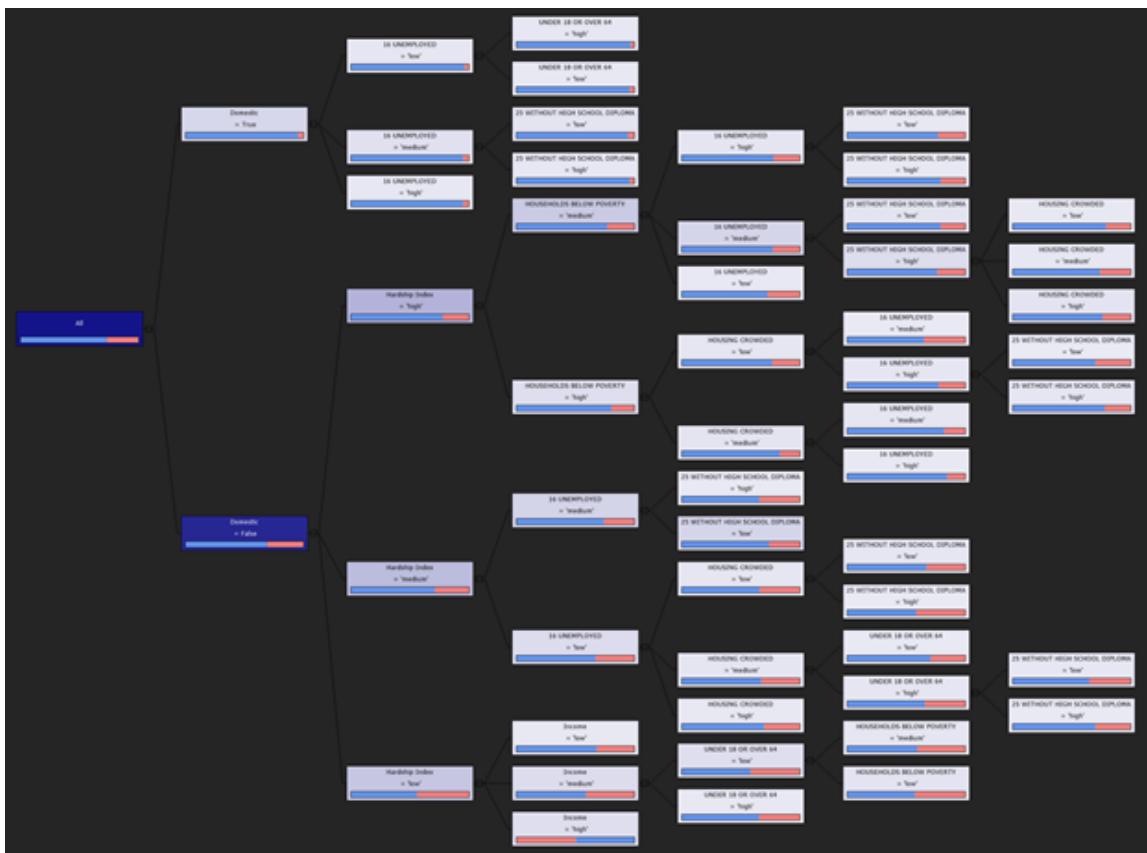
Add Remove OK Cancel Help

Ở đây chọn Split_Method value = 2 vì 1: xuông chi tiết quá, 3: rồi quá nên nhóm chọn 2

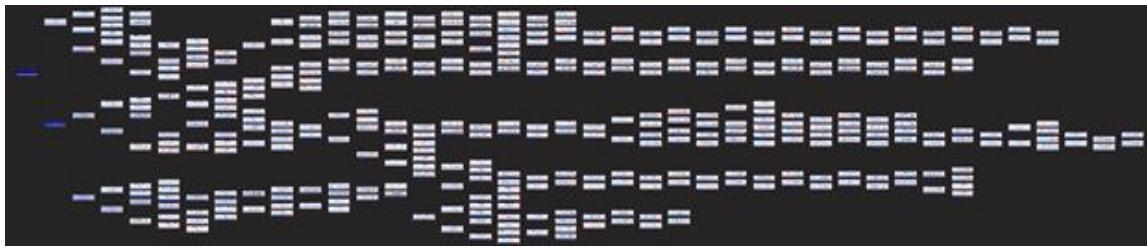
Kết khi chọn split_method có value = 1 bằng các thực thi vDM Domestic



Kết khi chọn split_method có value = 2 bằng các thực thi vDM Domestic



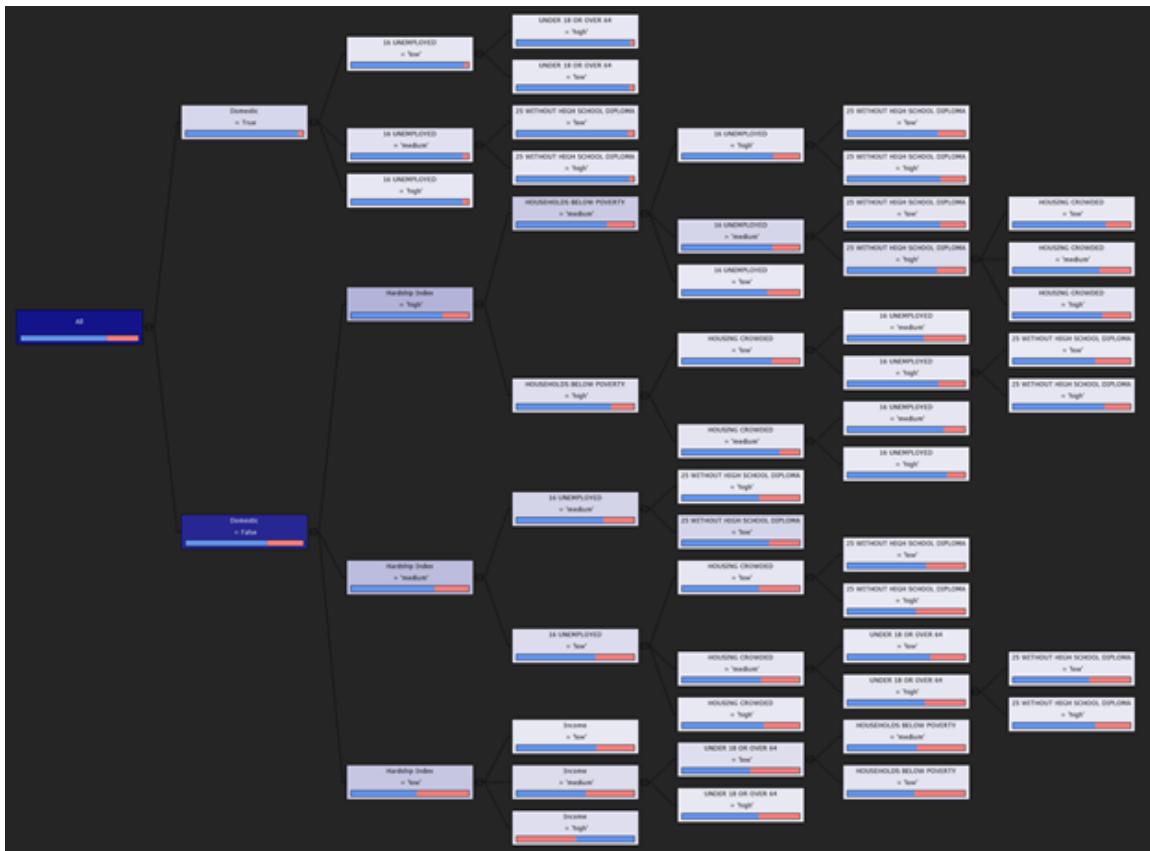
Kết khi chọn split_method có value = 3 bằng các thực thi vDM Domestic



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>)

4) 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
- Then 0
- Else 1
- And
- KpiValue("TinhTrangToiPham") /
KpiGoal("TinhTrangToiPham") > 1
- Then 0
- Else 1
- End
- 5) 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 $\frac{1}{4}$ 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.

X. Tham khảo

- [1] [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>