Data Warehousing

I. Introduction

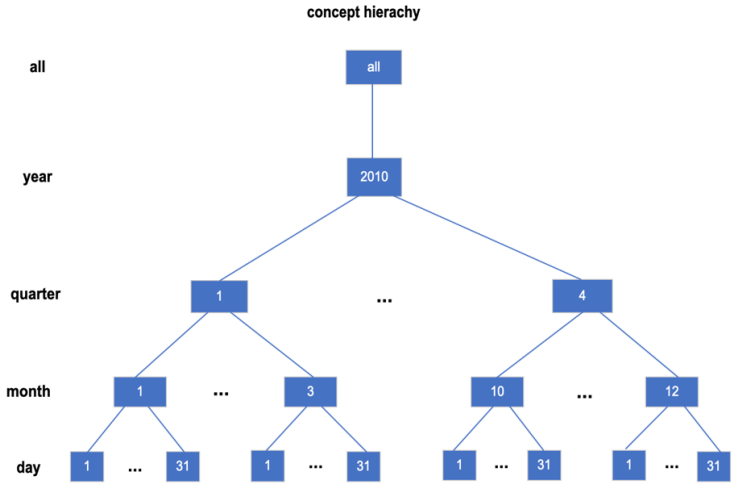
- This data warehouse can count the number of crime depending on different time, different area and different crime types.
- By data mining, we can see the relationship between beat and crime, and which crime will affect with each other in one beat. Besides, how different beats will affect each other.

II. The fact table and dimension tables design

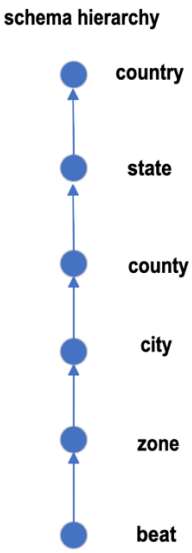
- Dimension tables include DimDate table (DateKey, Date, Day, Month, Quarter, Year), DimLocation table (Beat, Zone, City, County, State, Country) and DimCrime table (CrimeID, CrimeName, SeverityOfCrime).
  These three tables are made by using data from DimDate.csv, DimLocation.csv and DimCrime.csv. These three files' data are from all_data.csv.
  The file of all_data.csv collects all data for 2010 from all provided files in this project, and only keeps the following columns (crime, date, beat, city, county, state, country). In addition, zone column needs to be added in this file.
  - DimDate table: This table is used to store time information from all_data.csv
  - DimLocation table: This table is used to store location information from all_data.csv
  - DimCrime table: This table is used to store crime information from all_data.csv
- Fact table is designed to be an event-based fact table.
  Reason: There is no data related to numerical measures in the original files, so this table is used to record the occurrence of each even, and count the number of events.
- The reason of removing some columns (neighbourhood, neighbourhood-look, lat, long, npu, road)
  In location table, the smallest level is beat. Because I think beat is same level with other columns like npu. Besides, I think it is very meaningful to study beat, because the division of beat is based on the division of the police. This is directly related to the level of crime rate.

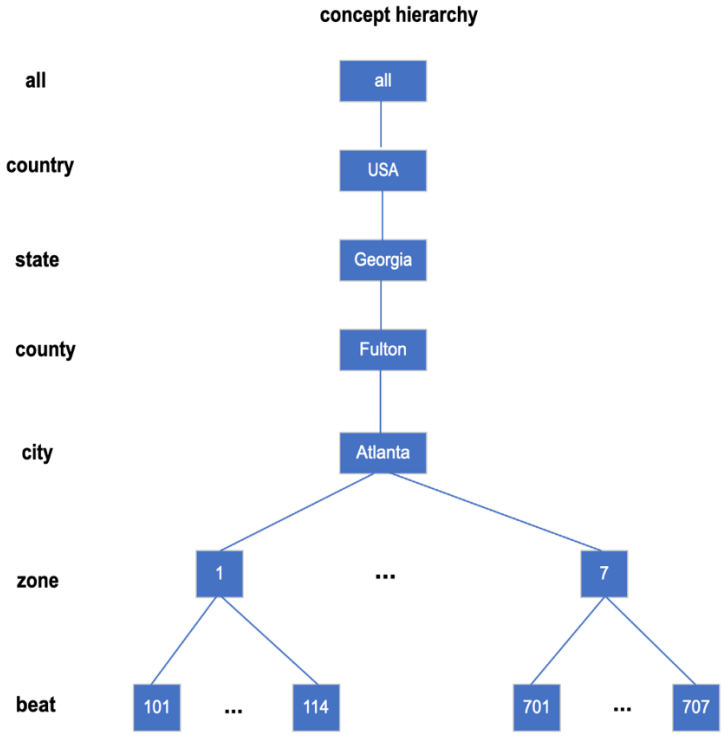
1. Schema and concept hierarchies of each dimension

- Schema and concept hierarchies of time dimension

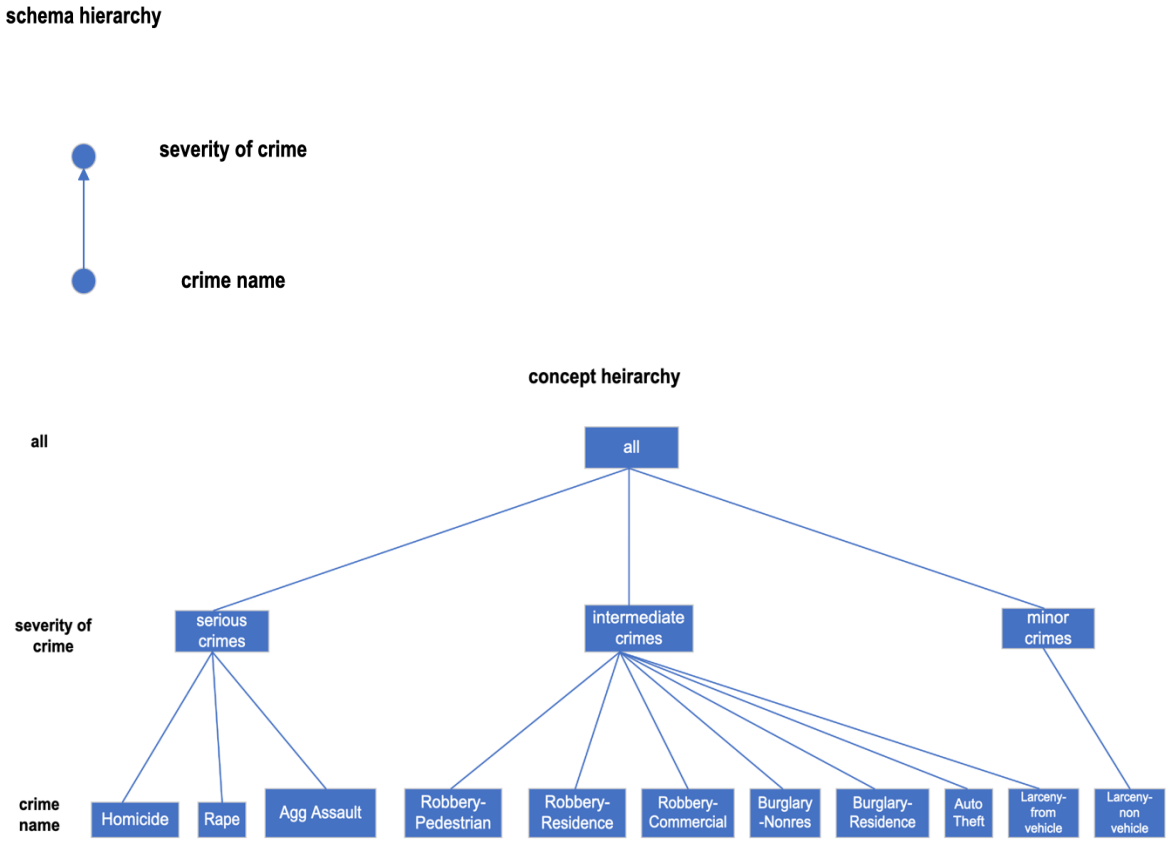


- Schema and concept hierarchies of location dimension

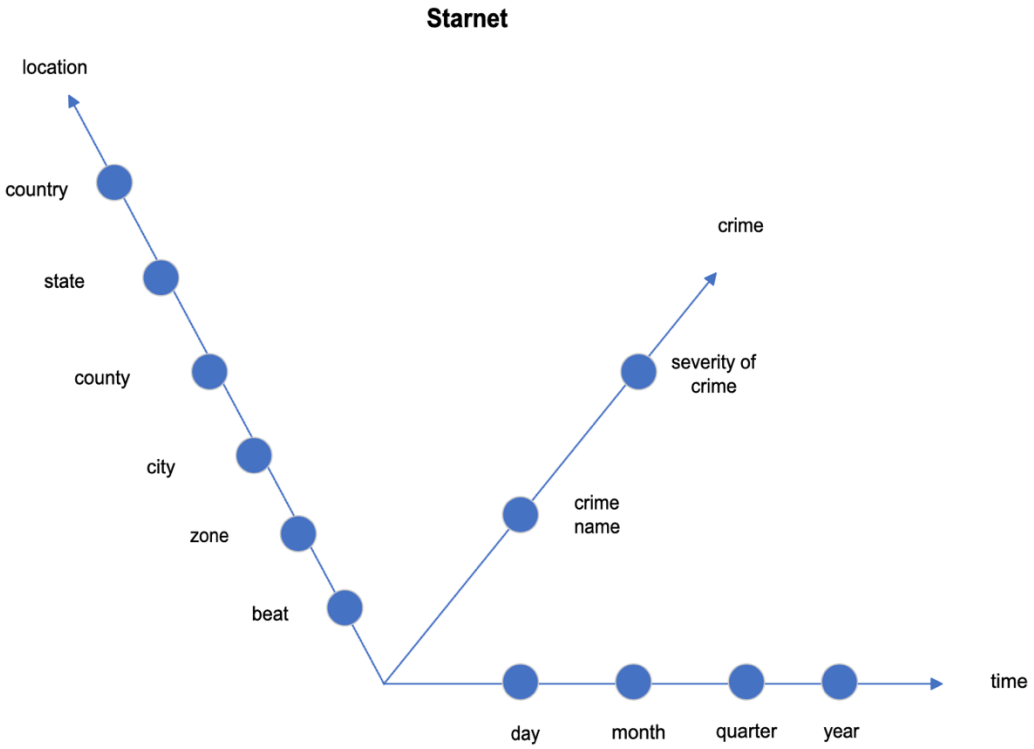




- Schema and concept hierarchies of crime dimension



2. Starnet



3. ETL process

- ETL process for DimDate.csv, the tools used are Excel and bash.
  1) The first step is to extract the date column from all_data.csv, then remove duplicate lines, display the date in the format of yyyy/mm/dd, and then arrange the date in order. (using bash command in terminal) The following is the code and screenshot.
     awk -F ',' '{print $4}' alldata.csv | sort -u | awk '{print $3"/"$1"/"$2}' | sort -n -t '/' -k3,3 -k1,1 -k2,2 > DimDate.csv

```
Last login: Tue Apr 18 18:25:23 on ttys000
rongchuansun@RongchuandeMacBook-Pro ~ % cd Desktop
rongchuansun@RongchuandeMacBook-Pro Desktop % cd project1
rongchuansun@RongchuandeMacBook-Pro project1 % awk -F ',' '{print $4}' all_data.
csv | sort -u | awk '{print $3"/"$1"/"$2}' | sort -n -t '/' -k3,3 -k1,1 -k2,2

/date/
/1/1/2010/
/2/1/2010/
/3/1/2010/
/4/1/2010/
/5/1/2010/
```

  2) The second step needs to extract the year, quarter, month and day corresponding to each row of data into a separate column. (Excel function)
     Year() function is to extract year from date column.
     Month() function is to extract month from date column.
     RONGDUP() function is to extract quarter from date column.
     DAY() function is to extract day from date column.
     Below is a screenshot of the completed.

DimDate

20100101	1/1/2010	1	1	1	2010
20100102	2/1/2010	2	1	1	2010
20100103	3/1/2010	3	1	1	2010
20100104	4/1/2010	4	1	1	2010
20100105	5/1/2010	5	1	1	2010
20100201	1/2/2010	1	2	1	2010
20100202	2/2/2010	2	2	1	2010
20100203	3/2/2010	3	2	1	2010
20100204	4/2/2010	4	2	1	2010
20100205	5/2/2010	5	2	1	2010
20100206	6/2/2010	6	2	1	2010
20100207	7/2/2010	7	2	1	2010
20100208	8/2/2010	8	2	1	2010
20100209	9/2/2010	9	2	1	2010

- ETL process for DimLocation.csv (using bash command and excel)
  1) The first step requires copying all_data.csv as DimLocation.csv and then keeping only required columns (country, state, county, city, beat). The result only shows the information after same beat row has been removed. Finally, it is to remove the data of other counties and only keep Fulton.
     The following is bash command and screenshot.
     cut -d "," -f 6,14,15,16,18 all_data.csv | awk '{if(NR>1)print}' | awk '!seen[$1]++' | sort -n -t ',' -k 1 > DimLocation.csv

DimLocation

101	Atlanta	Fulton County	Georgia	United States
102	Atlanta	Fulton County	Georgia	United States
103	Atlanta	Fulton County	Georgia	United States
104	Atlanta	Fulton County	Georgia	United States
105	Atlanta	Fulton County	Georgia	United States
106	Atlanta	Fulton County	Georgia	United States
107	Atlanta	Fulton County	Georgia	United States
108	Atlanta	Fulton County	Georgia	United States
109	Atlanta	Fulton County	Georgia	United States
110	Atlanta	Fulton County	Georgia	United States
111	Atlanta	Fulton County	Georgia	United States
112	Atlanta	Fulton County	Georgia	United States

  2) The second step needs to express the zone column according to beat column.
     Specific operation: select the beat data column, and then use Text to Columns function in data toolbar, and then the first digit can be separated out in beat column into a new column.
     Below is a screenshot of the completed.

DimLocation

101	1	Atlanta	Fulton County	Georgia	United States
102	1	Atlanta	Fulton County	Georgia	United States
103	1	Atlanta	Fulton County	Georgia	United States
104	1	Atlanta	Fulton County	Georgia	United States
105	1	Atlanta	Fulton County	Georgia	United States
106	1	Atlanta	Fulton County	Georgia	United States
107	1	Atlanta	Fulton County	Georgia	United States
108	1	Atlanta	Fulton County	Georgia	United States
109	1	Atlanta	Fulton County	Georgia	United States
110	1	Atlanta	Fulton County	Georgia	United States
111	1	Atlanta	Fulton County	Georgia	United States
112	1	Atlanta	Fulton County	Georgia	United States
113	1	Atlanta	Fulton County	Georgia	United States

- ETL process for DimCrime.csv (using excel)
    1) The first step requires copying all_data.csv as DimCrime.csv and then keeping only crime column.
       Specific operation: use excel to delete all other columns, and then remove the duplication of crime column data.
    2) The second step is to fill in the data in the corresponding row according to the severity of the crime.
    3) Below is a screenshot of the completed

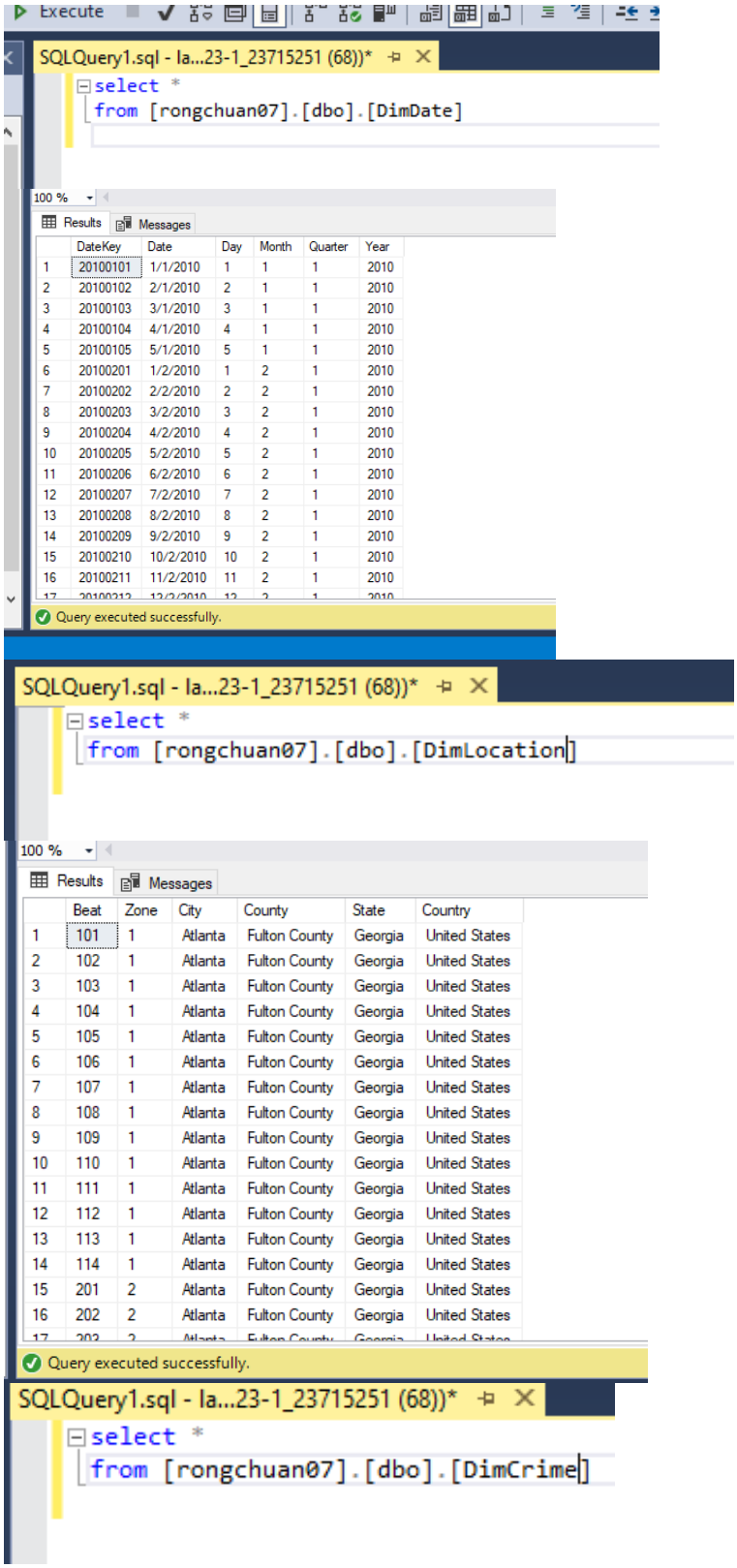
DimCrime

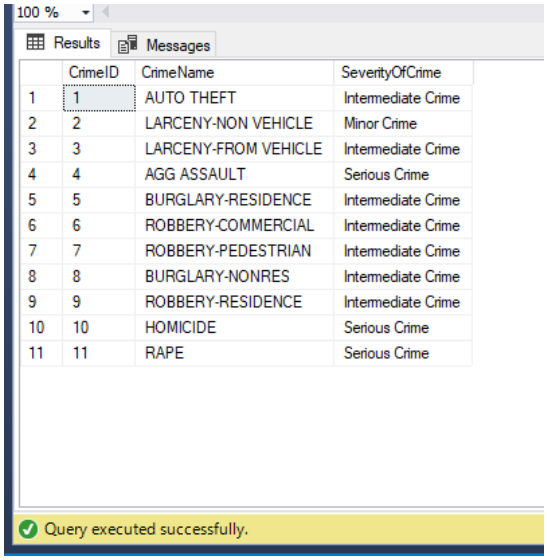
1	AUTO THEFT	Intermediate Crime
2	LARCENY-NON VEHICLE	Minor Crime
3	LARCENY-FROM VEHICLE	Intermediate Crime
4	AGG ASSAULT	Serious Crime
5	BURGLARY-RESIDENCE	Intermediate Crime
6	ROBBERY-COMMERCIAL	Intermediate Crime
7	ROBBERY-PEDESTRIAN	Intermediate Crime
8	BURGLARY-NONRES	Intermediate Crime
9	ROBBERY-RESIDENCE	Intermediate Crime
10	HOMICIDE	Serious Crime
11	RAPE	Serious Crime

4. Design dimensional tables and a fact table and import data in SSMS.

- CreateTables.sql is used to create three dimensional tables.
- InsertData.sql is used to upload data into dimensional tables from csv files.
- All_data.sql is used to create all_data table and import data from all_data.csv.
- FactCrime.sql is used to import data into the FactCrime table by using query statements.

- Here are some screenshots of the completed.



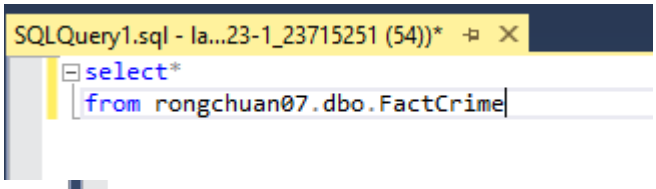


100 %

Results | Messages

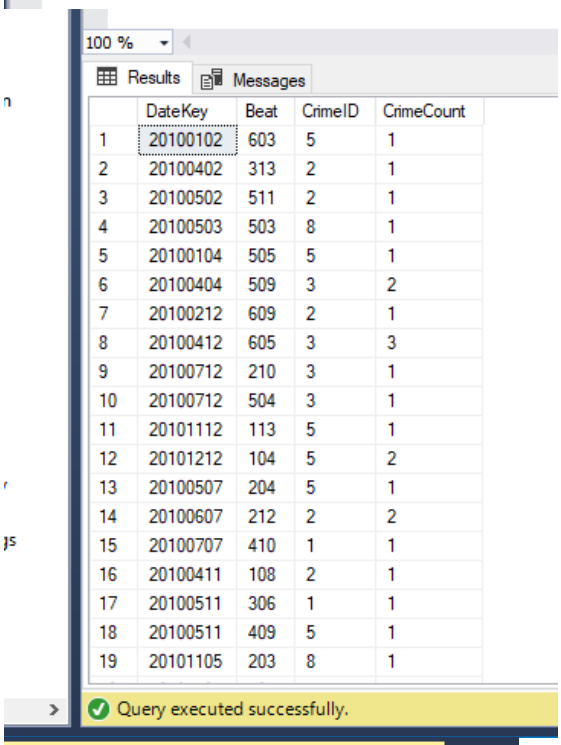
	CrimeID	CrimeName	SeverityOfCrime
1	1	AUTO THEFT	Intermediate Crime
2	2	LARCENY-NON VEHICLE	Minor Crime
3	3	LARCENY-FROM VEHICLE	Intermediate Crime
4	4	AGG ASSAULT	Serious Crime
5	5	BURGLARY-RESIDENCE	Intermediate Crime
6	6	ROBBERY-COMMERCIAL	Intermediate Crime
7	7	ROBBERY-PEDESTRIAN	Intermediate Crime
8	8	BURGLARY-NONRES	Intermediate Crime
9	9	ROBBERY-RESIDENCE	Intermediate Crime
10	10	HOMICIDE	Serious Crime
11	11	RAPE	Serious Crime

Query executed successfully.



SQLQuery1.sql - la...23-1_23715251 (54))*

```
select*
from rongchuan07.dbo.FactCrime
```

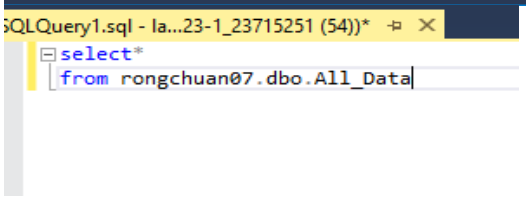


100 %

Results | Messages

	DateKey	Beat	CrimeID	CrimeCount
1	20100102	603	5	1
2	20100402	313	2	1
3	20100502	511	2	1
4	20100503	503	8	1
5	20100104	505	5	1
6	20100404	509	3	2
7	20100212	609	2	1
8	20100412	605	3	3
9	20100712	210	3	1
10	20100712	504	3	1
11	20101112	113	5	1
12	20101212	104	5	2
13	20100507	204	5	1
14	20100607	212	2	2
15	20100707	410	1	1
16	20100411	108	2	1
17	20100511	306	1	1
18	20100511	409	5	1
19	20101105	203	8	1

Query executed successfully.



SQLQuery1.sql - la...23-1_23715251 (54))*

```
select*
from rongchuan07.dbo.All_Data
```

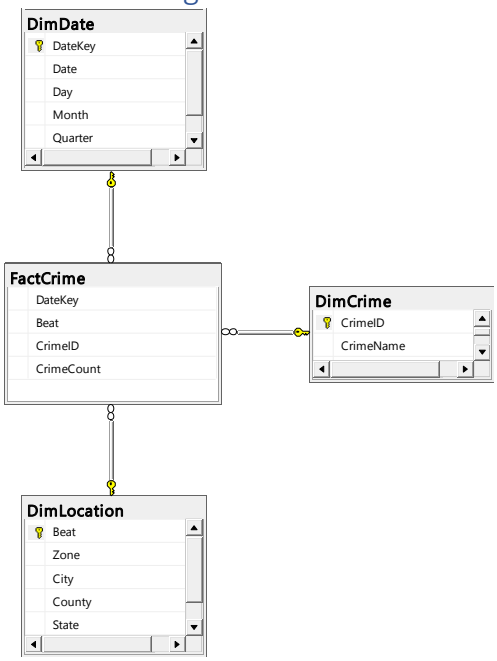
100 %

Results | Messages

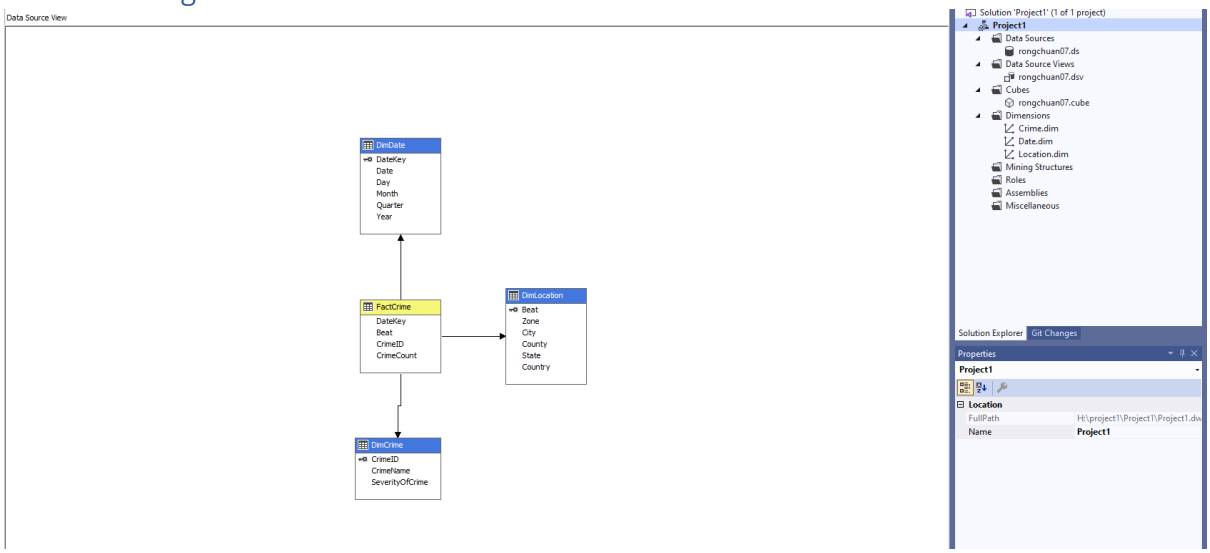
	crime	date	beat	city	county	state	country
1	LARCENY-FROM VEHICLE	1/1/2010	304	Atlanta	Fulton County	Georgia	United States
2	LARCENY-FROM VEHICLE	1/1/2010	208	Atlanta	Fulton County	Georgia	United States
3	BURGLARY-RESIDENCE	1/1/2010	204	Atlanta	Fulton County	Georgia	United States
4	LARCENY-NON VEHICLE	1/1/2010	401	Atlanta	Fulton County	Georgia	United States
5	LARCENY-NON VEHICLE	1/1/2010	511	Atlanta	Fulton County	Georgia	United States
6	ROBBERY-COMMERCIAL	1/1/2010	213	Atlanta	Fulton County	Georgia	United States
7	AUTO THEFT	1/1/2010	502	Atlanta	Fulton County	Georgia	United States
8	LARCENY-NON VEHICLE	1/1/2010	210	Atlanta	Fulton County	Georgia	United States
9	AUTO THEFT	1/1/2010	207	Atlanta	Fulton County	Georgia	United States
10	AUTO THEFT	1/1/2010	605	Atlanta	Fulton County	Georgia	United States
11	LARCENY-FROM VEHICLE	1/1/2010	511	Atlanta	Fulton County	Georgia	United States
12	LARCENY-NON VEHICLE	1/1/2010	210	Atlanta	Fulton County	Georgia	United States
13	BURGLARY-RESIDENCE	1/1/2010	409	Atlanta	Fulton County	Georgia	United States
14	LARCENY-NON VEHICLE	1/1/2010	602	Atlanta	Fulton County	Georgia	United States
15	AGG ASSAULT	1/1/2010	112	Atlanta	Fulton County	Georgia	United States
16	AGG ASSAULT	1/1/2010	301	Atlanta	Fulton County	Georgia	United States
17	LARCENY-FROM VEHICLE	1/1/2010	412	Atlanta	Fulton County	Georgia	United States
18	AUTO THEFT	1/1/2010	407	Atlanta	Fulton County	Georgia	United States
19	BURGLARY-RESIDENCE	1/1/2010	404	Atlanta	Fulton County	Georgia	United States

Query executed successfully.

5. Below is a database diagram in SSMS.

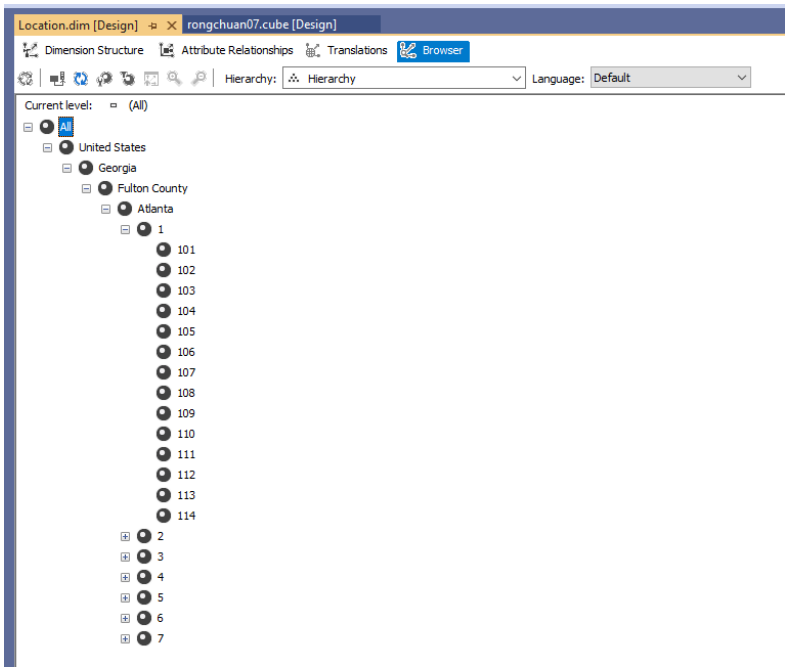


6. The cube diagram of SSDT.

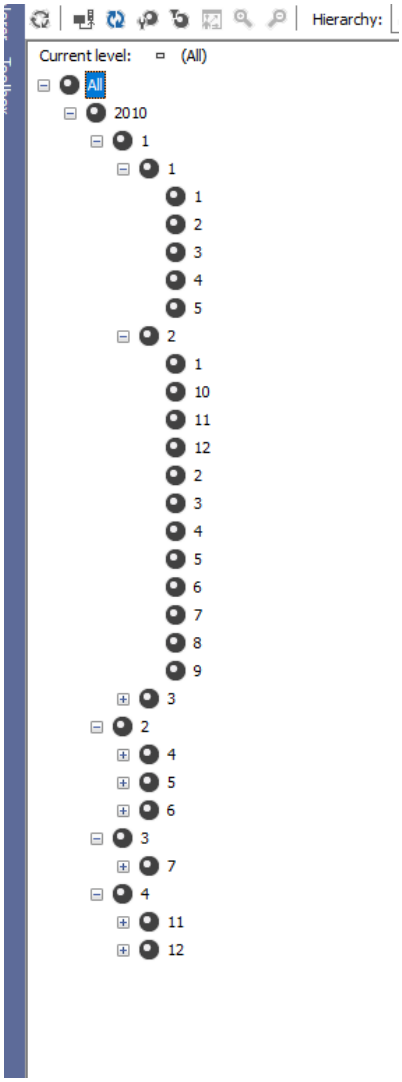


7. Concept hierarchies of three dimensions in SSDT.

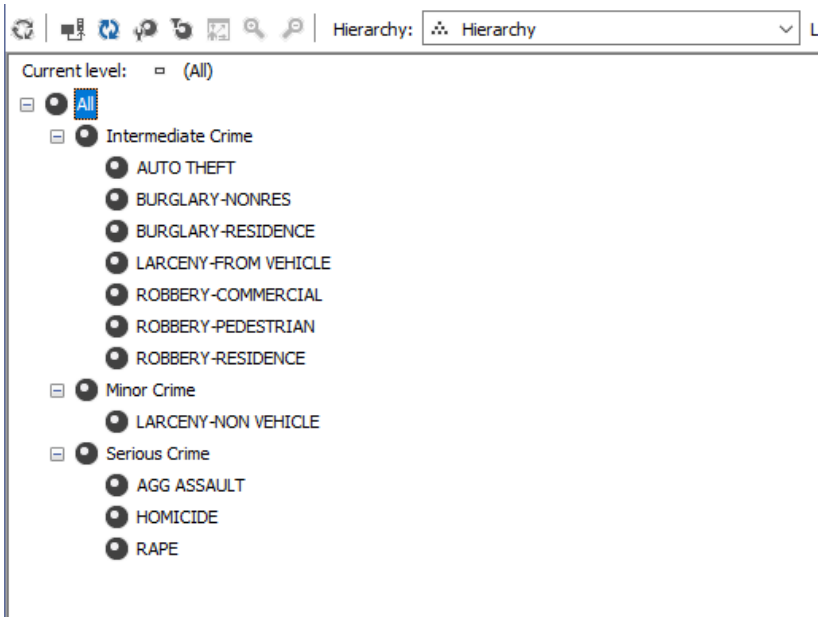
- Location



- Date


Hierarchy:
Current level: (All)
All
2010
1
1
1
2
3
4
5
2
1
10
11
12
2
3
4
5
6
7
8
9
3
2
4
5
6
3
7
4
11
12

- Crime

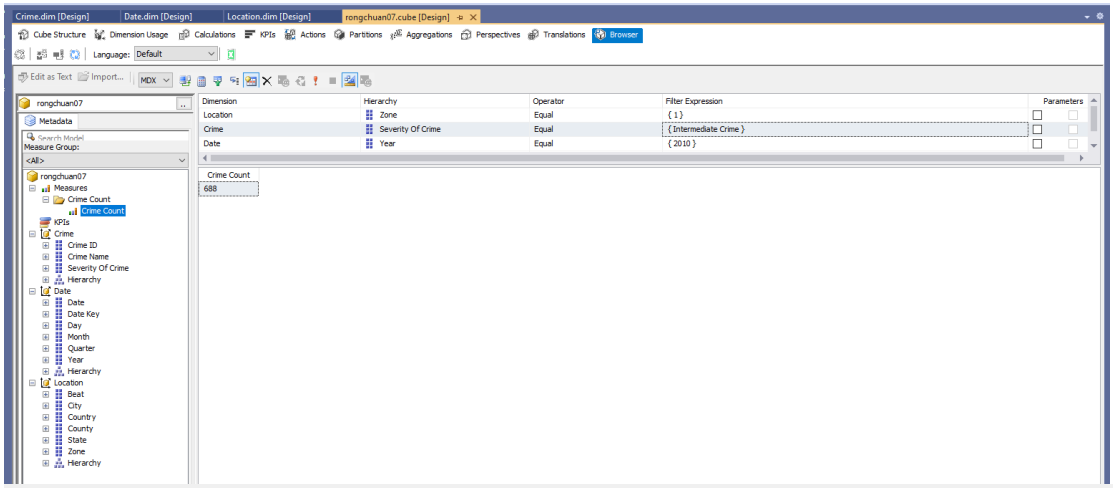

Hierarchy:
Hierarchy
Current level: (All)
All
Intermediate Crime
AUTO THEFT
BURGLARY-NONRES
BURGLARY-RESIDENCE
LARCENY-FROM VEHICLE
ROBBERY-COMMERCIAL
ROBBERY-PEDESTRIAN
ROBBERY-RESIDENCE
Minor Crime
LARCENY-NON VEHICLE
Serious Crime
AGG ASSAULT
HOMICIDE
RAPE

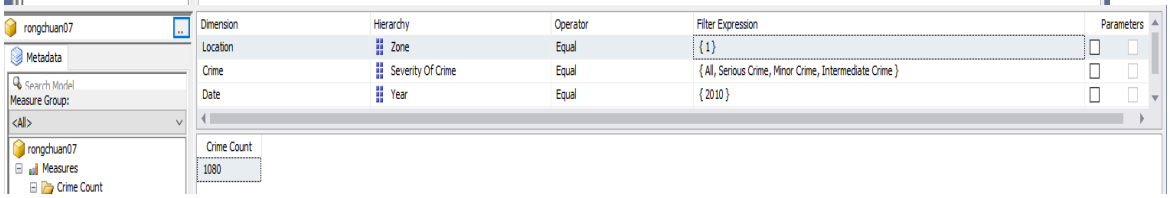
8. Multi-dimensional cubes are used to facilitate roll-up and drill down analysis

- Roll-up

  For example, I can query the count of intermediate crimes that occurred in different zones in 2010, also query the count of all crimes that occurred in different zones in 2010.

  Here are some screenshots.



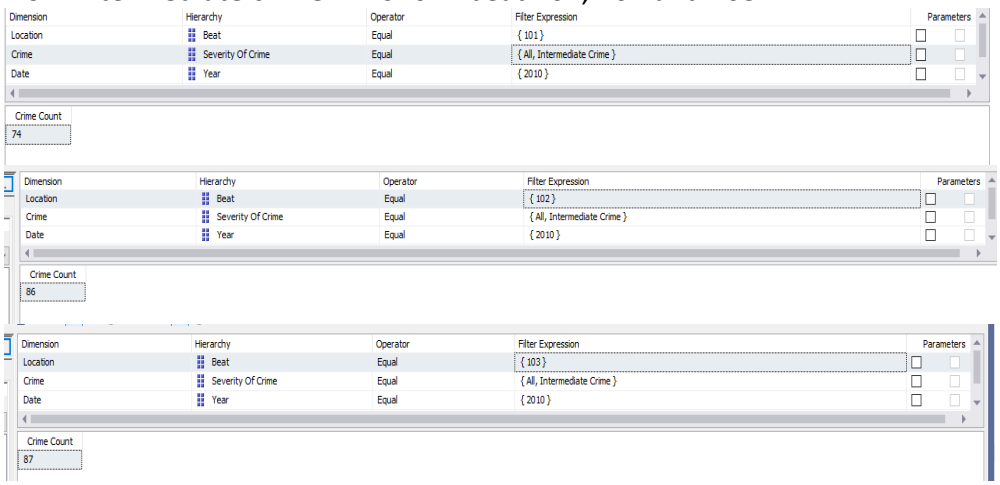


  The first screenshot shows the count of crime is 688 from intermediate crime in 2010 in zone 1.

  The second screenshot shows the count of crime is 1080 from all crimes in 2010 in zone 1.

- Drill down

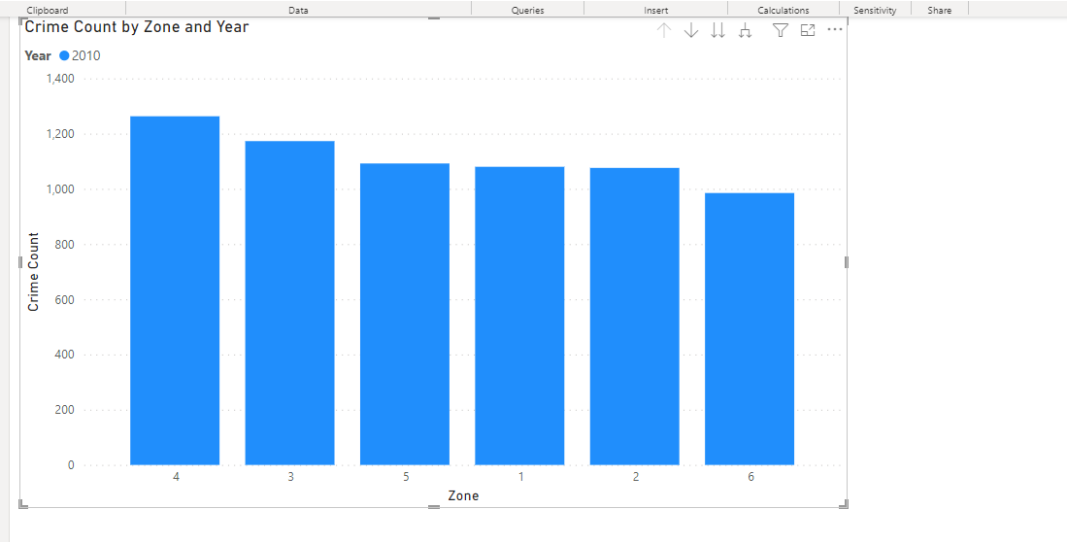
  The following three screenshots show that the crime count is 74, 86 and 87 from intermediate crime in 2010 in beat 101, 102 and 103.



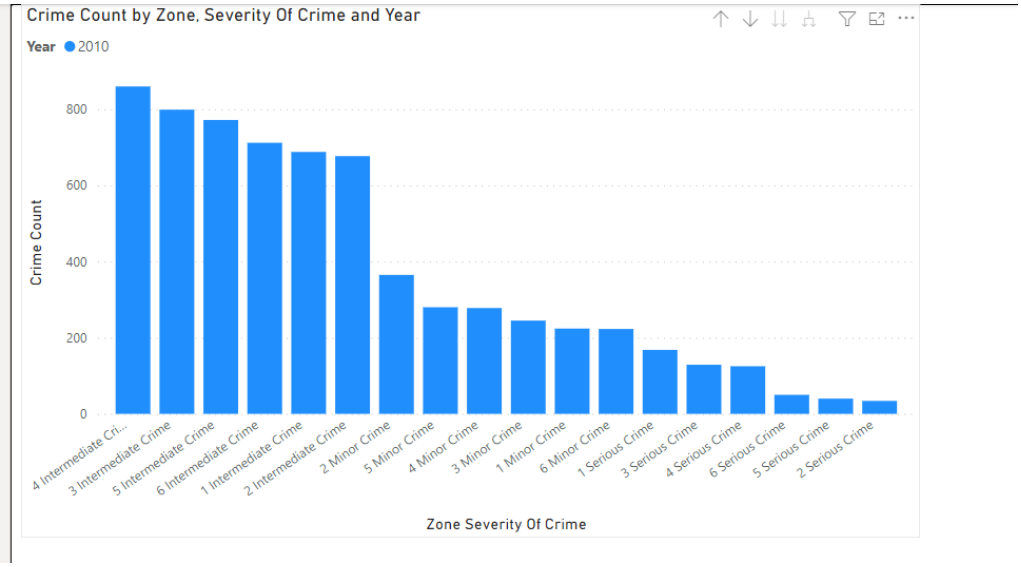
9. Five business queries and PowerBI visualization with the queries.

- Crime number of all crimes in different zone in 2010.

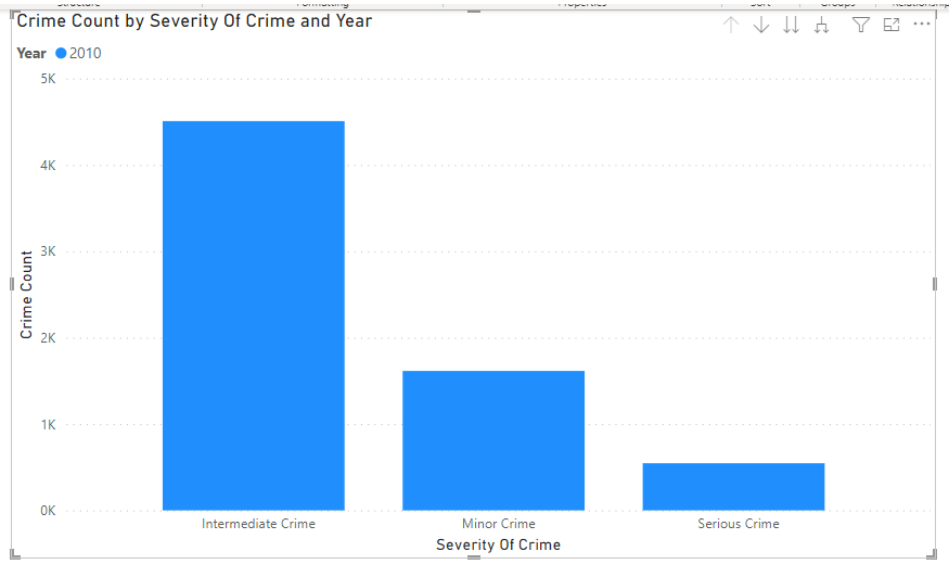
From the figure below, we can know that the area with the highest crime rate is zone 4, while the area with the lowest crime rate is zone 6.



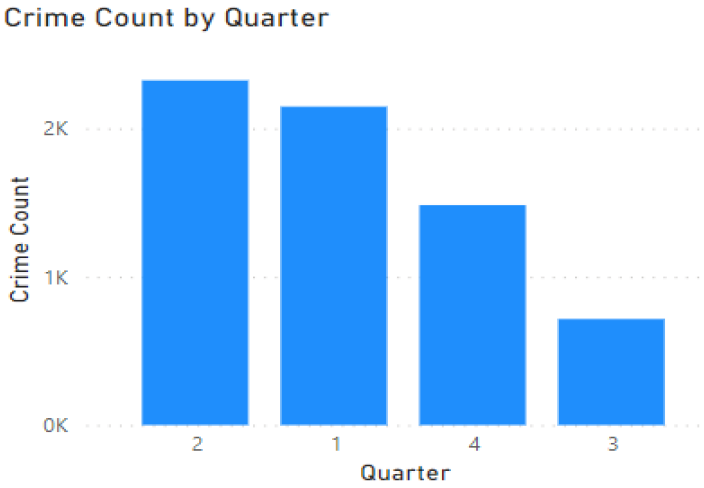
- Crime number of three different severities in different zone in 2010.
  From the figure below, we can know that intermediate crimes in zone 4 are the most, while serious crimes in zone 2 are the least.



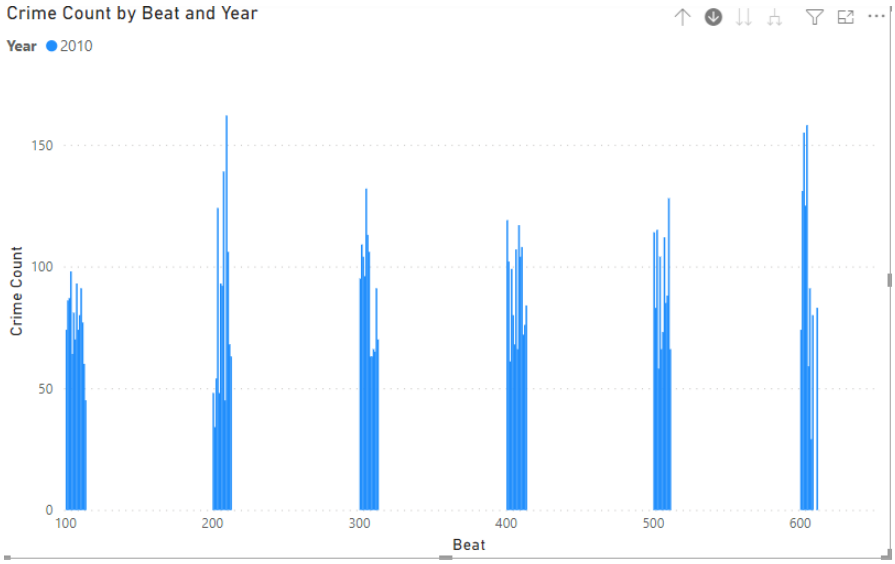
- Crime number of three different severities in 2010 in the whole area.
  From the figure below, we can know that intermediate crimes are the most and serious crimes are the least.



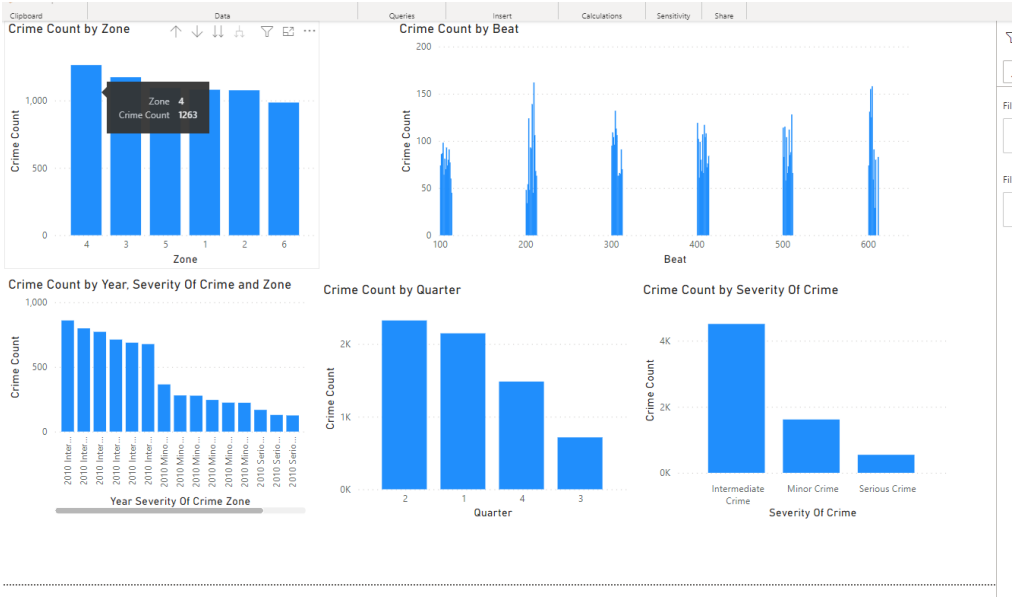
- Crime number of different quarter in 2010 in all area with all crimes.
  From the figure below, we can know that the crime number of the second quarter is the highest, while the crime number of the third quarter is the lowest.



- Crime number of different beat in 2010 with all crimes
  From below graph, we can know there is the highest crime count between beat 200 and beat 300. Besides, the crime count generally is lower than other beats between beat 100 and beat 200.



- The screenshot of PowerBI with the completed.



III. Association Rule Mining

1. Association rule mining process

1) Determine the research object

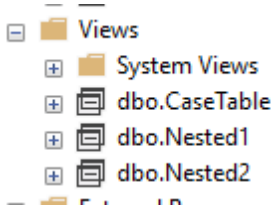
The relationship between beat and crime

We treat each beat and the crime instances of each day's records as a transaction. Each transaction contains multiple crime types. For example, the beat 511 in 2020-01-01 have multiple crime types, such as "Larceny-non vehicle" and "Larceny-from vehicle". I want to explore to see if there is a relationship between beat and crime type through various crime types.

2) Creating views in SSMS

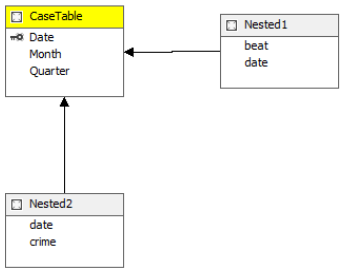
We need three views. One of views is to create cast table, reminding two views are to create Nested tables.

Below is a screenshot of views
CaseTable view includes Date, Month and Quarter columns, and Date is primary key.
Nested1 view contains beat and date columns
Nested2 view contains crime and date columns.



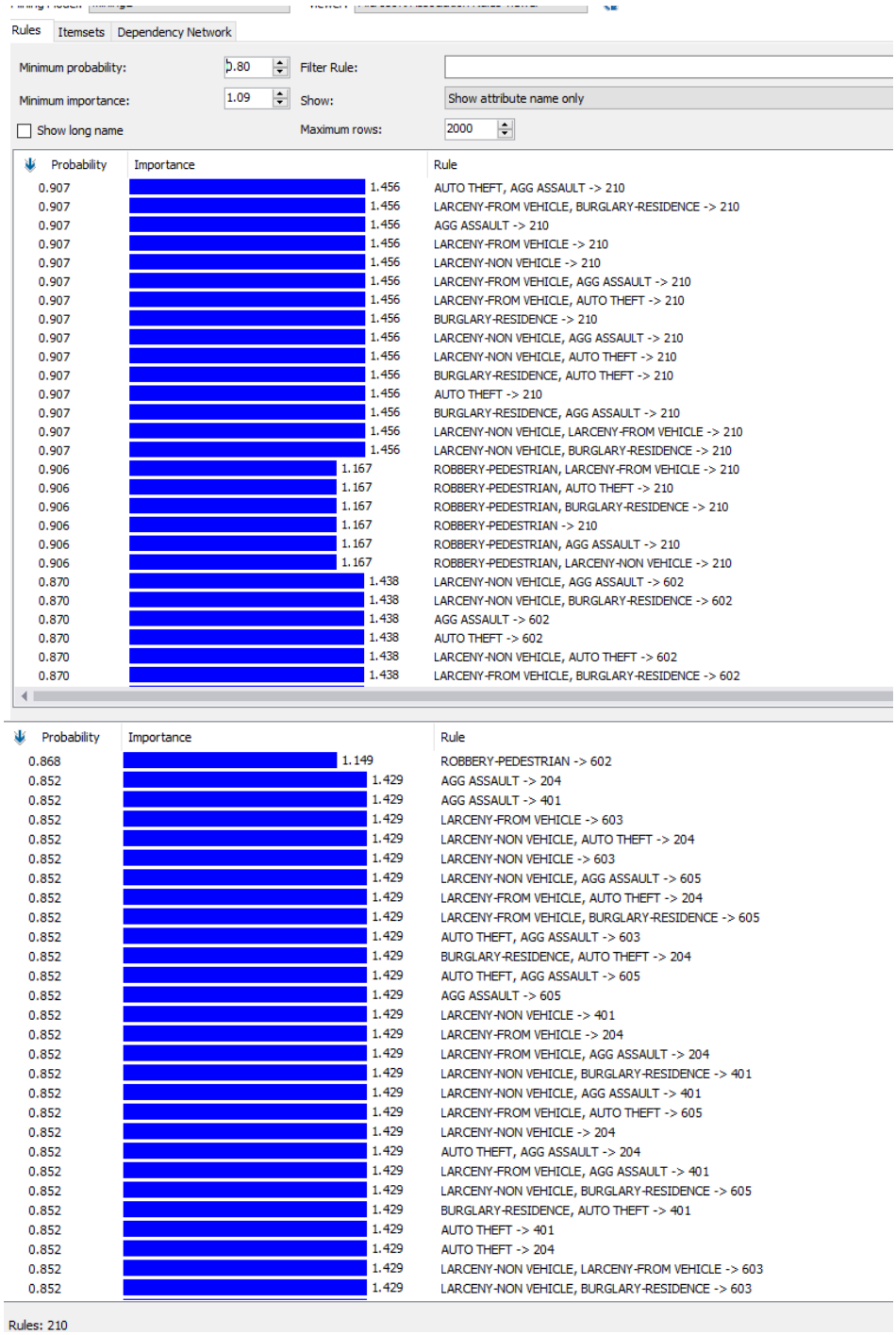
3) Creating relationship among these views in SSDT.

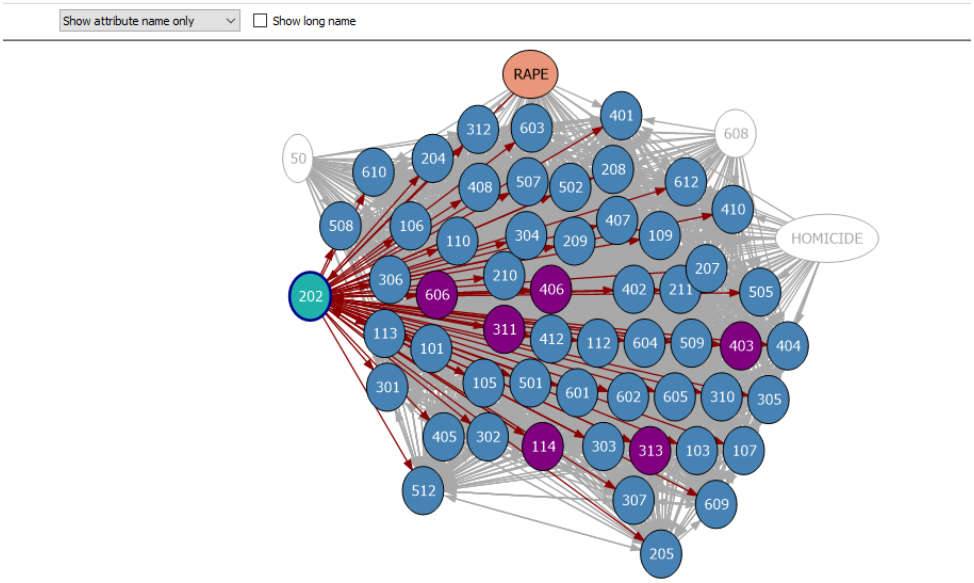
Firstly, we need to connect database of SSMS by using Data Sources.
Secondly, we need to import views into SSDT, and then create relationship among them.
Below is a screenshot of the relationship.



4) To data mining

We need to use mining structures to create data table and nested tables, and then choose key, input and predictable column. I choose crime and beat as input and predictable columns, and date as key column.
After that, we need to process in Mining Model Viewer, and then the result will show.
Below is a screenshot of rule view and dependent network.


Rules
Itemsets
Dependency Network
Minimum probability:
0.80
Filter Rule:
Minimum importance:
1.09
Show:
Show attribute name only
Show long name
Maximum rows:
2000
Probability
Importance
Rule
0.907 1.456 AUTO THEFT, AGG ASSAULT -> 210
0.907 1.456 LARCENY-FROM VEHICLE, BURGLARY-RESIDENCE -> 210
0.907 1.456 AGG ASSAULT -> 210
0.907 1.456 LARCENY-FROM VEHICLE -> 210
0.907 1.456 LARCENY-NON VEHICLE -> 210
0.907 1.456 LARCENY-FROM VEHICLE, AGG ASSAULT -> 210
0.907 1.456 LARCENY-FROM VEHICLE, AUTO THEFT -> 210
0.907 1.456 BURGLARY-RESIDENCE -> 210
0.907 1.456 LARCENY-NON VEHICLE, AGG ASSAULT -> 210
0.907 1.456 LARCENY-NON VEHICLE, AUTO THEFT -> 210
0.907 1.456 BURGLARY-RESIDENCE, AUTO THEFT -> 210
0.907 1.456 AUTO THEFT -> 210
0.907 1.456 BURGLARY-RESIDENCE, AGG ASSAULT -> 210
0.907 1.456 LARCENY-NON VEHICLE, LARCENY-FROM VEHICLE -> 210
0.907 1.456 LARCENY-NON VEHICLE, BURGLARY-RESIDENCE -> 210
0.906 1.167 ROBBERY-PEDESTRIAN, LARCENY-FROM VEHICLE -> 210
0.906 1.167 ROBBERY-PEDESTRIAN, AUTO THEFT -> 210
0.906 1.167 ROBBERY-PEDESTRIAN, BURGLARY-RESIDENCE -> 210
0.906 1.167 ROBBERY-PEDESTRIAN -> 210
0.906 1.167 ROBBERY-PEDESTRIAN, AGG ASSAULT -> 210
0.906 1.167 ROBBERY-PEDESTRIAN, LARCENY-NON VEHICLE -> 210
0.870 1.438 LARCENY-NON VEHICLE, AGG ASSAULT -> 602
0.870 1.438 LARCENY-NON VEHICLE, BURGLARY-RESIDENCE -> 602
0.870 1.438 AGG ASSAULT -> 602
0.870 1.438 AUTO THEFT -> 602
0.870 1.438 LARCENY-NON VEHICLE, AUTO THEFT -> 602
0.870 1.438 LARCENY-FROM VEHICLE, BURGLARY-RESIDENCE -> 602
Probability
Importance
Rule
0.868 1.149 ROBBERY-PEDESTRIAN -> 602
0.852 1.429 AGG ASSAULT -> 204
0.852 1.429 AGG ASSAULT -> 401
0.852 1.429 LARCENY-FROM VEHICLE -> 603
0.852 1.429 LARCENY-NON VEHICLE, AUTO THEFT -> 204
0.852 1.429 LARCENY-NON VEHICLE -> 603
0.852 1.429 LARCENY-NON VEHICLE, AGG ASSAULT -> 605
0.852 1.429 LARCENY-FROM VEHICLE, AUTO THEFT -> 204
0.852 1.429 LARCENY-FROM VEHICLE, BURGLARY-RESIDENCE -> 605
0.852 1.429 AUTO THEFT, AGG ASSAULT -> 603
0.852 1.429 BURGLARY-RESIDENCE, AUTO THEFT -> 204
0.852 1.429 AUTO THEFT, AGG ASSAULT -> 605
0.852 1.429 AGG ASSAULT -> 605
0.852 1.429 LARCENY-NON VEHICLE -> 401
0.852 1.429 LARCENY-FROM VEHICLE -> 204
0.852 1.429 LARCENY-FROM VEHICLE, AGG ASSAULT -> 204
0.852 1.429 LARCENY-NON VEHICLE, BURGLARY-RESIDENCE -> 401
0.852 1.429 LARCENY-NON VEHICLE, AGG ASSAULT -> 401
0.852 1.429 LARCENY-FROM VEHICLE, AUTO THEFT -> 605
0.852 1.429 LARCENY-NON VEHICLE -> 204
0.852 1.429 AUTO THEFT, AGG ASSAULT -> 204
0.852 1.429 LARCENY-FROM VEHICLE, AGG ASSAULT -> 401
0.852 1.429 LARCENY-NON VEHICLE, BURGLARY-RESIDENCE -> 605
0.852 1.429 BURGLARY-RESIDENCE, AUTO THEFT -> 401
0.852 1.429 AUTO THEFT -> 401
0.852 1.429 AUTO THEFT -> 204
0.852 1.429 LARCENY-NON VEHICLE, LARCENY-FROM VEHICLE -> 603
0.852 1.429 LARCENY-NON VEHICLE, BURGLARY-RESIDENCE -> 603
Rules: 210



2. Interpretation of top rules

1) Rule view

Minimum probability: 0.8

Minimum importance: 1.09

The graph shows the highest probability is 0.907, and corresponding importance value is 1.456.

By watching rule column, we know that there is a positive relationship between “auto theft” and “agg assault” in beat 210. This means when one of crimes occurs in beat 210, another one will most likely occur.

Another example, probability is 0.87, and importance is 1.438. By watching rule column, we know there is a positive relationship between “larceny-non vehicle” and “agg assault” in beat 602.

2) Dependent Network view

From this graph, we can see the relationship among different beats by different crimes. For example, the beat 202 has many different relationships with other beats. This beat can affect beat 508, beat 301, beat 512 and so on. This means others relevant beats might appear crime mostly if the beat has crime instances.