Advanced Data Management Project 2017/2018

Fighting Crime With Big Data

Contributors: Viktoriya Bodnar, Yasser Abou Haila

PART I

INSPIRED BY

https://datafloq.com/read/los-angeles-police-department-predicts-fights-crim/279

OVERVIEW

This project relates to very important issue nowadays: crime. It represents one of the top problem that need to be inspected and measured on different granularity levels. We are highly interested in creating a database of crime acts that can be used in order to develop an application that will track all crime details in US (or even can be extended to the worldwide system). This application can be a clear example of effective E-Governance, and the users of it will be mainly police departments, law enforcement agencies; federal, state, and local policy makers; state statistical analysis centers. The main purpose of this application is to gather relevant information and based on that to develop more powerful strategies, to create new laws and to change existing policies aiming at crime reduction.

In the application the information, as we already mentioned above, will concern the crime scene: people that have committed the crime, victims, investigatory police departments, weapons used in the crime, places where the crime actually happened and the correspondent dates.

Having analyzed possible details about the application necessities and currently used database technologies, we concluded that the best alternative would be a graph database which will cope with entities and relationships between them. Therefore, we decided to use Neo4j as reference technology and Cypher for our workload implementation.

APPLICATION REQUIREMENTS

The E-Governance application requires a database containing information about criminals, crimes, victims, police departments, weapons and cities so that (for a sample of data see Appendix 1):

CRIMES

The crime act is registered by a database administrator through the application. Each act of crime is described by the following properties:

1. crimeid (should be unique)
2. crimecode (standard crime code that corresponds to the crime type like 'Robbery','Vandalism', 'Criminal Homicide' etc…)
3. crimedesc (description of the committed crime)
4. datereport (date of being reported)
5. address
6. city
7. victimid (the id of the victim)
8. criminalid (id of the person who committed the crime or the person who is suspected)
9. pdid (id of the police department that is investigating the case)
10. status (status of the investigation: 'Arrest','Invest Cont' that stands for Investigation continues and 'Other'

When the crime has happened the police department that has registered the criminal behavior have to insert the crime and its details into the system having unique crimeid. When the case first was registered by the called department and then was passed to the department responsible for that area, the new department is added without overwriting the previous (to keep track of all departments that worked for the particular case). For one crime scene there can be more than one victim and more than one criminal. The criminalid can be updated in order to pass from suspected person to the arrested one. In this case, we don't keep track of all suspected people. The status is another information that is being changed during the investigation.

Possible data analysis can be done based on to the number of crimes in particular city, most criminal city, the number of crimes of particular type, analysis based on

single victim/criminal information (how many times the person was involved into a crime)

CRIMINAL

Each criminal registration contains the information about:

1. criminalid (should be unique)
2. firstname
3. lastname
4. age
5. address
6. city

Here, all the information is inserted once and is not updated after. The age, the address and the city of the criminal will be registered due to the time when the crime was committed.

Criminals' information is queried using criminalid. The average age of the criminals can be computed for further analysis.

VICTIM

The following table has information about victims:

1. victimid (should be unique)
2. firstname
3. lastname
4. age
5. sex (M or F)
6. address
7. city

The table is not updated after insertion and the age, address, city attributes are related to the period of time of the crime.

The victim info may be inspected by querying the victimid. Statistics can be done on questions like: if there are more female or male victims, what is the most frequent victims age etc.

WEAPON

The weapon table represents the following fields:

1. weaponcode (should be unique)
2. weapondesc (description about the weapon ,e.g. ‘knife’,’revolver’…)

The codes and correspondent descriptions can be easily queried or added to the table.

CITY

The city table is used to have extended information about the city where the people live, where the police departments are located and where the crime was committed.

1. city (should be unique)
2. state
3. country

Since we consider only US as a target country, we will have city names as unique values(otherwise the combination of city and the country needs to represent the key). Cities and correspondent states and countries can be inserted once without updating, taking the official gov database of all cities and states in the country.

PD

Each row that represents the police department contains some attributes such as:

1. pdid (should be unique)
2. city
3. location (geographical location given by longitude and latitude)

Here the possible updates can be if the department is moved the location need to be changed. The rare deletion operation also possible in case of closing of the department for some reason.

Besides the entities we also have some relationships between them. In the graph model they are presented as directional archs between nodes.

COMMITED BY

The relationship models the connection between criminal and its crime. The added property is crime date (dateoccurred) which describes when exactly the act occurred.

USING

This relationship is created between the crime the weapon used in that crime.

Can be used in order to compute the most used weapons during crimes.

INVESTIGATED BY

The crime is investigated by the police department and the datereport value of the relationship is used to indicate when the process has started. The delay between the crime occurred and the investigation started can be measured comparing these dates.

The rank of police departments based on number of investigated crimes can be calculated based on this relationship.

COMMITTED ON

The committed on relationship aims at connecting the crime and its victims.

LIVES IN

This connection is related to the people and their place of living at the moment of the inserting their information to the database.

LOCATED IN

Each police department is located in some city and this relationship also needs to be included.

COMMITED IN

The city where the crime was committed and the crime itself are also connected by the arch to represent the relationship.

Cross analysis can be done using this relationship in order to understand the correlation between states and crimes.

Based on the above information, we can see that the typical workload for the application is a mixture of read and write operations.

POSSIBLE EXTENSION IN THE FUTURE

As possible extensions to our database we can consider:

- adding witnesses, evidences, responsible policemen, suspected people as new nodes
- the relationships between criminals and victims like family/friends…
- additional attributes such as fingerprints for criminals/victims that can act as unique keys.
- Examination of other higher entities that control the police departments.
- Dates as relationship properties: lives in from/to, investigation starts/ends

…

SYSTEM REQUIREMENTS

Taking into consideration the fact that in graph databases it is really difficult to have a data distribution, in our scenario we are considering the data related to only one month (December 2015).Furthermore, the other reason why we reduced the original dataset is to have better performance and better visualization (see Appendix 2). The other drawback of graph databases is costly update. Since we don't have a lot of update operations on our data, most of the times the new information will be just added, so the cost of updates will not impact the application performance. However, in real life, as we can imagine, typos and other reasons for updates and deletion can occur, the crimes happen every day across the country and the database is growing very fast. As a highly challenging but a possible alternative to cope with these issues can be clustering but this is out of the scope of this project.

TECHNOLOGY

Having considered the application requirements, it follows that the workload will be consisted of adding new information and querying the existed one. The updates and deletion are rarely performed, so the Neo4j is a good option for data storage and management. In addition, cypher language is quite simple for querying and the results can be visualized in the graph context. That is a big advantage when we want to understand better the relationships between the entities.

DATASET

The dataset considered in our project is semi-simulated. In the sense that the dataset of real crimes was taken from the U.S. Government's open data website. The names, addresses and other confidential information was simulated in order to have more detailed description of the crime scenes.

RESOURCES

https://www.data.gov/

https://neo4j.com/

https://neo4j.com/developer/cypher-query-language/

https://www.forbes.com/sites/emc/2014/06/03/data-analysis-helps-police-departments-fight-crime/#63f95e543952

APPENDIX 1

DR_Number	Date_Reported	Date_Occurred	Crime_Code	Crime_descr	Weapon_U	Address	City	VictimId	CriminalId	PdId	Status_Description
150128751	12/31/2015	12/30/2015	510	VEHICLE - STOLEN		700 N GrandAv	Arlington	1	101	1.5E+13	Invest Cont
150718981	12/30/2015	12/29/2015	510	VEHICLE - STOLEN		Saint Elmo	Flushing	2	102	1.5E+13	Invest Cont
150101677	12/2/2015	12/2/2015	230	ASSAULT WITH DE	400	700 Yale St	Pasadena	3	103	1.5E+13	Adult Arrest
150101678	12/2/2015	12/2/2015	230	ASSAULT WITH DE	400	600 S San Pedro St	Philadelphia	4	104	1.5E+13	Adult Arrest
150101679	12/3/2015	12/3/2015	740	VANDALISM - FELONY ($400 &	400	Cottage Home St	Shawnee Missio	5	105	1.5E+13	Invest Cont
150101681	12/1/2015	12/1/2015	421	THEFT FROM MOTOR VEHICL	1300	S OliveSt	Lexington	6	106	1.5E+13	Invest Cont
150101683	12/5/2015	12/5/2015	626	INTIMATE PARTNE	400	600 S Spring St	Honolulu	7	107	1.5E+13	Invest Cont
150101684	12/5/2015	12/5/2015	320	BURGLARY, ATTEM	400	500 S Alameda St	Saint Paul	8	108	1.5E+13	Adult Arrest
150101686	12/6/2015	12/6/2015	626	INTIMATE PARTNE	400	500 Wall St	Akron	9	109	1.5E+13	Adult Other
150101695	12/9/2015	12/9/2015	626	INTIMATE PARTNE	400	500 Wall St	Newport Beach	10	110	1.5E+13	Adult Other

CriminalId	Last Name	First Name	Age	Street	City
101	Sandimani	Boima	30		Jacksonville
102	Moore	Donnell	42		Van Nuys
103	Lewis	Jonathan	40	400 Block	Henderson
104	Adams	Rolando	24	19800 Blo	Springfield
105	Bean	Ronnie	24	0 Block Of	Olympia
106	Leonard	Anthony	18	15300 Blo	Huntington
107	Battle	Malcolm	26	10000 Blo	Denver
108	Nti	Joseph	62	11500 Blo	Jamaica
109	Lesmez-Sa	Sandra	32	7900 Bloc	Lake Charles
110	Makela	Cedric	33	20100 Blo	Atlanta
111	Yenchi	Frankline	35		Seminole
112	Aregahegn	Eliyas	22		Greensboro
113	Thompson	Zak	26	5400 Bloc	Knoxville

Victimid	Victim_Age	Victim_Sex	first_name	last_name	address	city
1	14		Elbert	Scoyles	36 Hollow Ridge Point	Jacksonville
2	14		Andrei	Brandone	63457 Corry Avenue	Van Nuys
3	78	F	Alasteir	McCusker	593 Elgar Parkway	Henderson
4	32	F	Mehetabel	Michieli	634 Melrose Park	Springfield
5	30	M	Jabez	Hedgeman	6 Graceland Pass	Olympia
6	29	F	Roseanne	Spong	50 International Parkv	Huntington
7	41	F	Cayla	Fowley	9205 Sundown Drive	Denver
8	35	M	Bartel	Galpen	52 Stone Corner Road	Jamaica
9	43	F	Dewitt	Hayesman	2240 Banding Parkway	Lake Charles
10	42	F	Melva	Phebey	8 Drewry Court	Atlanta
11	37	F	Garrett	Klempke		Seminole
12	29	F	Ivonne	Wildblood	98854 Debra Court	Greensboro
13	27	M	Debra	Eddisforth	732 Stoughton Parkwa	Knoxville

city	state	stateName	Country
Anchorage	AK	Alaska	US
Fairbanks	AK	Alaska	US
Birmingham	AL	Alabama	US
Tuscaloosa	AL	Alabama	US
Mobile	AL	Alabama	US
Montgomery	AL	Alabama	US
Huntsville	AL	Alabama	US
Fort Smith	AR	Arkansas	US

PdId	City	Location
1.50E+13	Aiken	(37.7617007179518, -122.42158168137)
1.50E+13	Aiken	(37.7841907151119, -122.414406029855)
1.50E+13	Aiken	(37.7841907151119, -122.414406029855)
1.50E+13	Aiken	(37.7841907151119, -122.414406029855)
1.50E+13	Akron	(37.8004687042875, -122.431118543788)
1.50E+13	Akron	(37.7870853907529, -122.451781767894)
1.46E+13	Akron	(37.729203356539, -122.374019331833)
1.50E+13	Akron	(37.729203356539, -122.374019331833)
1.50E+13	Akron	(37.7878092959561, -122.40656817787)
1.50E+13	Akron	(37.7862578545865, -122.417295322526)
1.50E+13	Akron	(37.7690748003847, -122.413354187018)

Weapon_Used_C	Weapon_Description
101	REVOLVER
102	HAND GUN
103	RIFLE
104	SHOTGUN
105	SAWED OFF RIFLE/SHOTGUN
106	UNKNOWN FIREARM
107	OTHER FIREARM
108	AUTOMATIC WEAPON/SUB-M
109	SEMI-AUTOMATIC PISTOL
110	SEMI-AUTOMATIC RIFLE

APPENDIX 2

CRIME GRAPH