Data Warehousing and Data Mining
Project Phase 1

Thapaswini Chowdary - 20161066
Jayitha .C - 20171401

August 24, 2018

Abstract

The objective of the project was to select a dataset and make a data warehouse using the STAR schema that when queried on yielded interesting results

1 Data Set - Homicide Data

This data set was chosen because it was easy to understand, the number of attributes weren't too overwhelming but just enough to enable us to run some fun queries to yield some interesting results and the number of records were good enough to derive some concrete practical inferences.

1.1 Properties of Data Set

- Number of Attributes - 12
- Number of Records - 52179
- Attributes are as listed below
  - Uid - Unique Identifier for each Victim
  - Reported_date - Date victim reported to have died in yyyymmdd format
  - victim_last - Victim's last name
  - victim_first - Victim's first name
  - victim_race - Victim's race that can take values - Asian, Black, Hispanic, Other, Unknown, White
  - victim_age - Age of victim during time of demise
  - victim_sex - Sex of victim that can take values - Female, Male, Unknown
  - city - City where victim's demise occurred (or where it was filed)
  - state - State where victims's demise occurred (or where it was filed)
  - lat - latitude coordinates of Victim last seen
  - lon - longitude coordinates of Victim last seen
  - disposition - State of Victim's case

2 DBML Chosen

Chose to complete the task using MySQL as we are pretty familiar with it and also because MySQL supports data cube queries such as GROUP BY WITH ROLL UP. Also the data set given can be easily represented using the Relational Database Model, and since MySQL uses a relational database model it can easily and efficiently represent and store the data for efficient query processing

3 Data Preprocessing - Data Cleaning

The following measures were taken to clean the data

- The date attributes provided with the data set were in the yyyymmdd format. It needed to be converted into a format that MySQL accepted. The date was converted to a 3 attribute format, year, month, day directly in the excel provided
- The data set provided used many non-ASCII characters to represent the name and other strings. All these strings had to converted to ASCII acceptable strings for MySQL to be able to load the data into the database
- The dataset contained many "unknown" values for the attributes - victim_age, victim_sex, victim_race, lat and lon. Considered replacing these unknowns with either the mean, median or mode of the data given contained within domains for example, maybe within a city or a state. But there are too many unknowns and replacing that many may cause the final result to be misleading. And hence unknowns were retained and contributed to the count of attributes whose values were known.
- The latitude and longitude information was bypassed, the values were very close to get any observations. We limit the scope of our observations to the other dimensions alone and discard these values. They are still present in the raw dataset but will not be a part of the warehouse.

4 STAR Schema

Fig(1) shows the STAR Schema representation for the Homicide dataset.
The dimension tables are

- AGE - ID(PK), LOW, HIGH
- RACE - ID(PK), Race
- STATE - ID(PK), State
- CITY - ID(PK), City
- LOCATION - ID(PK), StateID(FK), City(FK)
- DISPOSITION - ID(PK), Disposition
- SEX - ID(PK), Sex

- YEAR - ID(PK), Year
- MONTH - ID(PK), Month
- DAY - ID(PK), Day
- TIME - ID(PK), YearID(FK), MonthID(FK), DayID(FK)

FACT_TABLE - TimeID(FK), AgeID(FK), SexID(FK), RaceID(FK), LocationID(FK), DispositionID(FK), Count

DATA_CUBE - TimeID, AgeID, SexID, RaceID, LocationID, DispositionID, Count

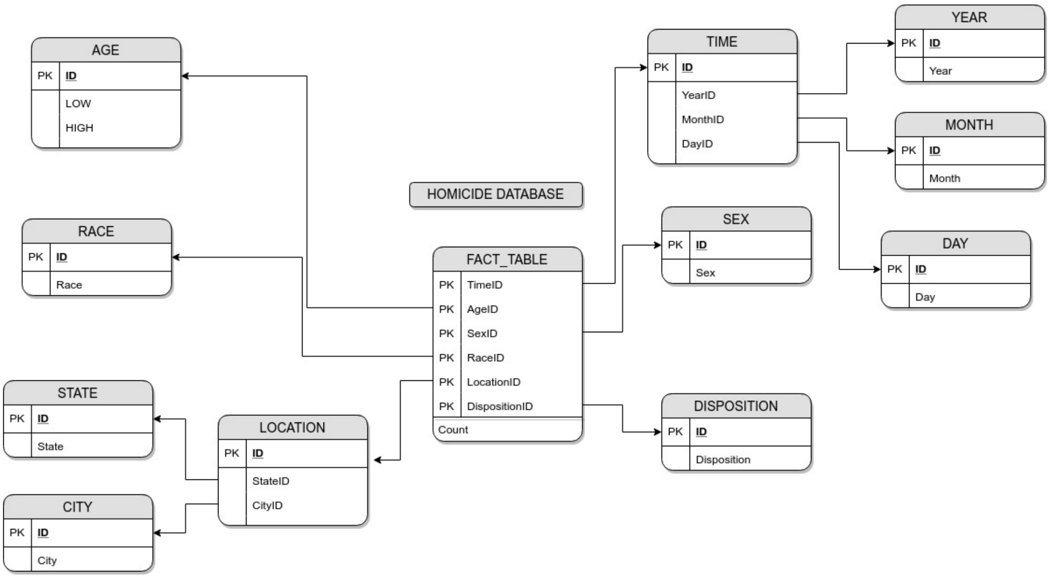
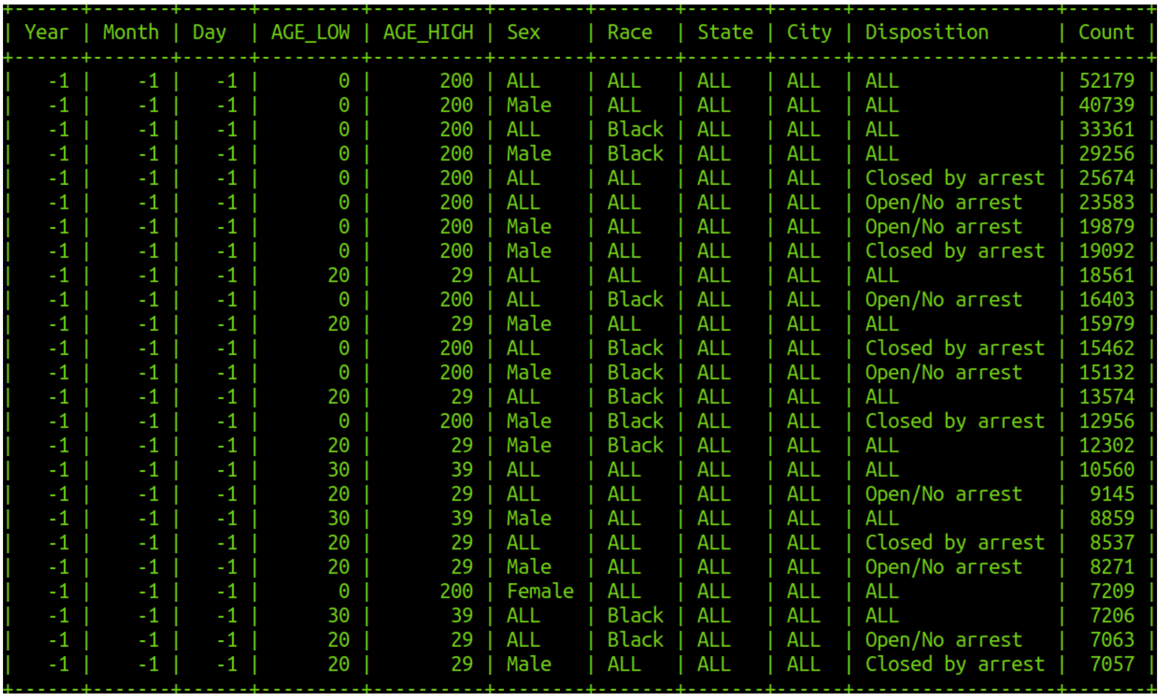


Figure 1: STAR Schema Diagram for Homicide Dataset

5 Interesting Observations

1. SELECT YEAR.Year, MONTH.Month, DAY.Day, AGE.LOW AS AGE_LOW, AGE.HIGH AS AGE_HIGH, SEX.Sex, RACE.Race, STATE.State, CITY.City, DISPOSITION.Disposition, Count FROM DATA_CUBE, YEAR, MONTH, DAY, AGE, SEX, RACE, STATE, CITY, DISPOSITION, TIME, LOCATION WHERE TIME.ID = TimeID AND TIME.YearID = YEAR.ID AND TIME.MonthID = MONTH.ID AND TIME.DayID = DAY.ID AND AGE.ID = AgeID AND SEX.ID = SexID AND RACE.ID = RaceID AND LOCATION.ID = LocationID AND STATE.ID = LOCATION.StateID AND CITY.ID = LOCATION.CityID AND DISPOSITION.ID = DispositionID ORDER BY Count DESC LIMIT 25



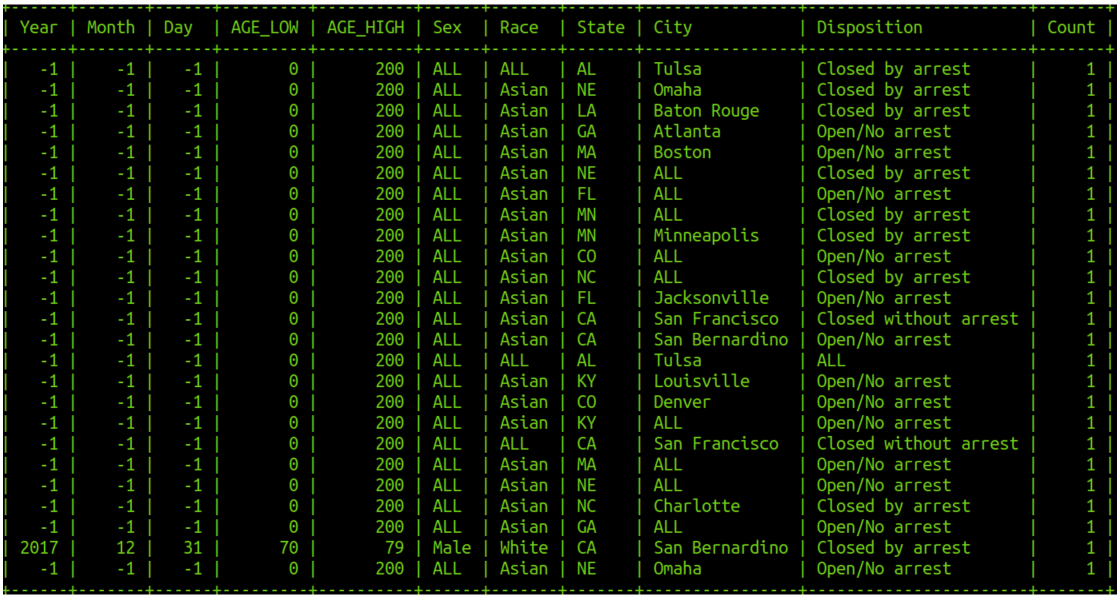
```
+------+-------+------+---------+----------+--------+-------+-------+------+------------------+-------+
| Year | Month | Day  | AGE_LOW | AGE_HIGH | Sex    | Race  | State | City | Disposition      | Count |
+------+-------+------+---------+----------+--------+-------+-------+------+------------------+-------+
|   -1 |    -1 |   -1 |       0 |      200 | ALL    | ALL   | ALL   | ALL  | ALL              | 52179 |
|   -1 |    -1 |   -1 |       0 |      200 | Male   | ALL   | ALL   | ALL  | ALL              | 40739 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL    | Black | ALL   | ALL  | ALL              | 33361 |
|   -1 |    -1 |   -1 |       0 |      200 | Male   | Black | ALL   | ALL  | ALL              | 29256 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL    | ALL   | ALL   | ALL  | Closed by arrest | 25674 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL    | ALL   | ALL   | ALL  | Open/No arrest   | 23583 |
|   -1 |    -1 |   -1 |       0 |      200 | Male   | ALL   | ALL   | ALL  | Open/No arrest   | 19879 |
|   -1 |    -1 |   -1 |       0 |      200 | Male   | ALL   | ALL   | ALL  | Closed by arrest | 19092 |
|   -1 |    -1 |   -1 |      20 |       29 | ALL    | ALL   | ALL   | ALL  | ALL              | 18561 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL    | Black | ALL   | ALL  | Open/No arrest   | 16403 |
|   -1 |    -1 |   -1 |      20 |       29 | Male   | ALL   | ALL   | ALL  | ALL              | 15979 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL    | Black | ALL   | ALL  | Closed by arrest | 15462 |
|   -1 |    -1 |   -1 |       0 |      200 | Male   | Black | ALL   | ALL  | Open/No arrest   | 15132 |
|   -1 |    -1 |   -1 |      20 |       29 | ALL    | Black | ALL   | ALL  | ALL              | 13574 |
|   -1 |    -1 |   -1 |       0 |      200 | Male   | Black | ALL   | ALL  | Closed by arrest | 12956 |
|   -1 |    -1 |   -1 |      20 |       29 | Male   | Black | ALL   | ALL  | ALL              | 12302 |
|   -1 |    -1 |   -1 |      30 |       39 | ALL    | ALL   | ALL   | ALL  | ALL              | 10560 |
|   -1 |    -1 |   -1 |      20 |       29 | ALL    | ALL   | ALL   | ALL  | Open/No arrest   |  9145 |
|   -1 |    -1 |   -1 |      30 |       39 | Male   | ALL   | ALL   | ALL  | ALL              |  8859 |
|   -1 |    -1 |   -1 |      20 |       29 | ALL    | ALL   | ALL   | ALL  | Closed by arrest |  8537 |
|   -1 |    -1 |   -1 |      20 |       29 | Male   | ALL   | ALL   | ALL  | Open/No arrest   |  8271 |
|   -1 |    -1 |   -1 |       0 |      200 | Female | ALL   | ALL   | ALL  | ALL              |  7209 |
|   -1 |    -1 |   -1 |      30 |       39 | ALL    | Black | ALL   | ALL  | ALL              |  7206 |
|   -1 |    -1 |   -1 |      20 |       29 | ALL    | Black | ALL   | ALL  | Open/No arrest   |  7063 |
|   -1 |    -1 |   -1 |      20 |       29 | Male   | ALL   | ALL   | ALL  | Closed by arrest |  7057 |
+------+-------+------+---------+----------+--------+-------+-------+------+------------------+-------+
```

Figure 2: Results of Query 1

Fig(2) shows results of Query 1. It shows all possible combinations of attributes that yield the highest 25 death counts, from this we can make the following observations

- Out of 52179 victims, 40739 are male. This could be attributed to the actual Male/Female ratio as well
- More than half the victims are Black people
- The number of clases that have closed down with an arrest is almost equal to the number of open/no arrest cases
- The number of cases of black victims that have been closed and left open are the same
- Out of 52K, 16K belong to the age group (20-30) making it the most prominent age group

2. SELECT YEAR.Year, MONTH.Month, DAY.Day, AGE.LOW AS AGE_LOW, AGE.HIGH AS AGE_HIGH, SEX.Sex, RACE.Race, STATE.State, CITY.City, DISPOSITION.Disposition, Count FROM DATA_CUBE, YEAR, MONTH, DAY, AGE, SEX, RACE, STATE, CITY, DISPOSITION, TIME, LOCATION WHERE TIME.ID = TimeID AND TIME.YearID = YEAR.ID AND TIME.MonthID = MONTH.ID AND TIME.DayID = DAY.ID AND AGE.ID = AgeID AND SEX.ID = SexID AND RACE.ID = RaceID AND LOCATION.ID = LocationID AND STATE.ID = LOCATION.StateID AND CITY.ID = LOCATION.CityID AND DISPOSITION.ID = DispositionID ORDER BY Count LIMIT 25;



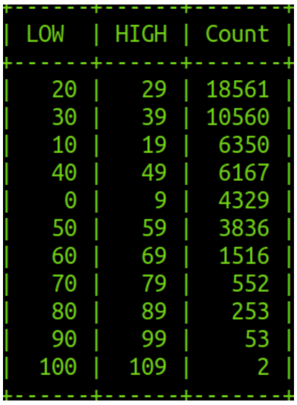
```
+------+-------+------+---------+----------+------+-------+-------+----------------+-----------------------+-------+
| Year | Month | Day  | AGE_LOW | AGE_HIGH | Sex  | Race  | State | City           | Disposition           | Count |
+------+-------+------+---------+----------+------+-------+-------+----------------+-----------------------+-------+
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | ALL   | AL    | Tulsa          | Closed by arrest      |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | NE    | Omaha          | Closed by arrest      |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | LA    | Baton Rouge    | Closed by arrest      |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | GA    | Atlanta        | Open/No arrest        |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | MA    | Boston         | Open/No arrest        |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | NE    | ALL            | Closed by arrest      |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | FL    | ALL            | Open/No arrest        |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | MN    | ALL            | Closed by arrest      |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | MN    | Minneapolis    | Closed by arrest      |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | CO    | ALL            | Open/No arrest        |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | NC    | ALL            | Closed by arrest      |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | FL    | Jacksonville   | Open/No arrest        |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | CA    | San Francisco  | Closed without arrest |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | CA    | San Bernardino | Open/No arrest        |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | ALL   | AL    | Tulsa          | ALL                   |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | KY    | Louisville     | Open/No arrest        |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | CO    | Denver         | Open/No arrest        |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | KY    | ALL            | Open/No arrest        |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | ALL   | CA    | San Francisco  | Closed without arrest |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | MA    | ALL            | Open/No arrest        |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | NE    | ALL            | Open/No arrest        |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | NC    | Charlotte      | Closed by arrest      |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | GA    | ALL            | Open/No arrest        |     1 |
| 2017 |    12 |   31 |      70 |       79 | Male | White | CA    | San Bernardino | Closed by arrest      |     1 |
|   -1 |    -1 |   -1 |       0 |      200 | ALL  | Asian | NE    | Omaha          | Open/No arrest        |     1 |
+------+-------+------+---------+----------+------+-------+-------+----------------+-----------------------+-------+
```

Figure 3: Results of Query 2

Fig(3) shows results of Query 2. It shows all possible combinations of attributes that yield the lowest 25 death counts, from this we can make the following observations

- Tulsa, AL is probably the city with the least crime rate. Of course this is if and only if the date provided here is accurate enough.
- The Asians don't seem to be very in number.

3. SELECT AGE.LOW, AGE.HIGH, a.Count FROM (SELECT AgeID, SUM(Count) AS Count FROM FACT_TABLE GROUP BY AgeID WITH ROLLUP) a, AGE WHERE AGE.ID = a.AgeID ORDER BY a.Count DESC;



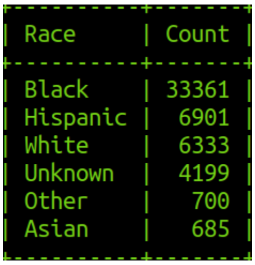
```
+------+------+-------+
| LOW  | HIGH | Count |
+------+------+-------+
|   20 |   29 | 18561 |
|   30 |   39 | 10560 |
|   10 |   19 |  6350 |
|   40 |   49 |  6167 |
|    0 |    9 |  4329 |
|   50 |   59 |  3836 |
|   60 |   69 |  1516 |
|   70 |   79 |   552 |
|   80 |   89 |   253 |
|   90 |   99 |    53 |
|  100 |  109 |     2 |
+------+------+-------+
```

Figure 4: Results of Query 3

Fig(4) shows results of Query 3. It shows total deaths per age group, from this we can make the following observations

- Majority of the victims belong to age groups between 10 and 50. This makes a little bit of sense, as people from other age groups do not go out much.

4. SELECT RACE.Race, a.Count FROM (SELECT RaceID, SUM(Count) AS Count FROM FACT_TABLE GROUP BY RaceID WITH ROLLUP) a, RACE WHERE RACE.ID = RaceID ORDER BY a.Count DESC;



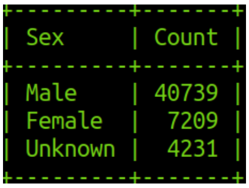
```
+----------+-------+
| Race     | Count |
+----------+-------+
| Black    | 33361 |
| Hispanic |  6901 |
| White    |  6333 |
| Unknown  |  4199 |
| Other    |   700 |
| Asian    |   685 |
+----------+-------+
```

Figure 5: Results of Query 4

Fig(5) shows results of Query 4. It shows total deaths per RACE, from this we can make the following observations

- It re establishes what we've already observed before, that the number of black victims is significantly large and that the number of Asians is significantly small

5. SELECT SEX.Sex, a.Count FROM (SELECT SexID, SUM(Count) AS Count FROM FACT_TABLE GROUP BY SexID WITH ROLLUP) a, SEX WHERE SEX.ID = SexID ORDER BY a.Count DESC;



```
+---------+-------+
| Sex     | Count |
+---------+-------+
| Male    | 40739 |
| Female  |  7209 |
| Unknown |  4231 |
+---------+-------+
```

Figure 6: Results of Query 5

Fig(6) shows results of Query 5. It shows total deaths per RACE, from this we can make the following observations

- It re establishes what we've already observed before, that the number of male victims is significantly large

6. SELECT YEAR.Year, MONTH.Month, DAY.Day, STATE.State, CITY.City, SUM(Count) AS COUNT FROM DATA_CUBE, TIME, LOCATION, YEAR, MONTH, DAY, STATE, CITY WHERE TIME.YearID = YEAR.ID AND TIME.MonthID = MONTH.ID AND TIME.DayID = DAY.ID AND LOCATION.StateID = STATE.ID AND LOCATION.CityID = CITY.ID

```
GROUP BY YEAR.Year, MONTH.Month, DAY.Day, STATE.State, CITY.City
ORDER BY COUNT DESC;
```

For each day, for each time how many people died in each location. Following observations made

- on 1st Oct 2017, at LV a group of 60 people were killed - historically we can see that some event has taken place here and similarly later on