

Asad Ranavaya

23 November 2022

Real Time Big Data Analysis

Project Data Ingestion

Data Profiling

For this portion of the final project I profiled, cleaned, and ingested the following two datasets: NYPD crime data and the subway geo dataset. When writing the MapReduce programs, I wrote separate mapper and reducer classes and simply combined them in a single driver java class. There was a total of 6 MapReduce Jobs performed; to run the code in the driver class I just commented out the other jobs not being ran.

NYPD Arrest Data:

The first step to make this data useable was to filter out unnecessary data. The following salient data was kept in a csv format:

1. Arrest Identification Key (solely to gather other information, if necessary, later on)
2. Arrest Date
3. Arrest Classification code
4. Arrest Classification code description
5. Arrest borough
6. Latitude
7. Longitude

The data was also filtered by only keeping felony arrests. Misdemeanors and violations were removed. After filtering this, out of 5.31 million records, 226,235 were kept. The next step was to get a better understanding of the remaining data. I created a list of unique felony crimes contained within the filtered dataset by running a MapReduce count job. I also used a dynamic global counter in this job to see how many felony crimes were committed per borough:

Borough Count per Crime

Bronx: 46187

Brooklyn: 74470

Manhattan: 45661

Queens: 50787

Staten Island: 9130

There were 91 unique crime categories. This will help us by acting as a key for filtering out violent crimes from non-violent felonies such as the category "TAX LAW". A snippet of the key (code, description and count):

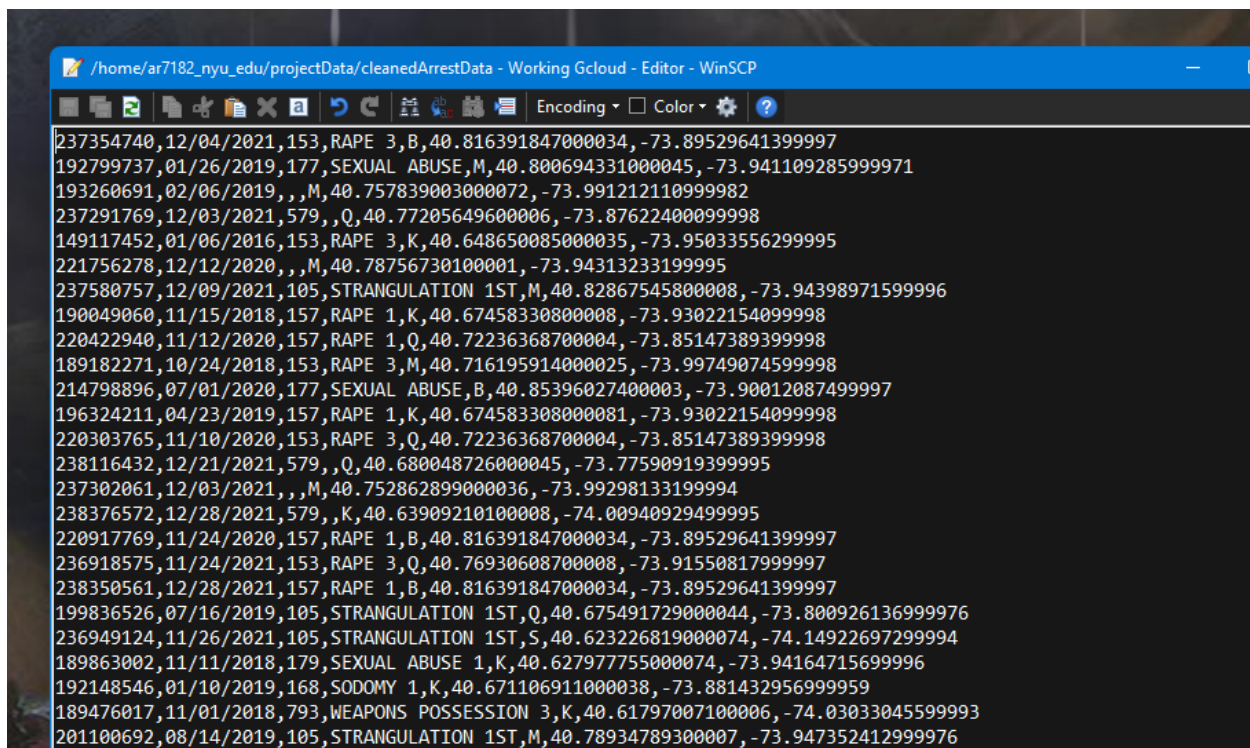
```

100 STALKING COMMIT SEX OFFENSE, Count: 2
104 VEHICULAR ASSAULT (INTOX DRIVER), Count: 1037
105 STRANGULATION 1ST, Count: 21222
106 ASSAULT POLICE/PEACE OFFICER, Count: 3246
107 ENDANGERING VULNERABLE ELDERLY, Count: 268
112 MENACING 1ST DEGREE (VICT NOT PEACE OFFICER), Count: 2599
117 RECKLESS ENDANGERMENT 1, Count: 14967
119 PROMOTING SUICIDE ATTEMPT, Count: 3
123 Not Provided, Count: 14
124 Not Provided, Count: 8
126 Not Provided, Count: 3
143 ABORTION 1, Count: 10
153 RAPE 3, Count: 3214
155 RAPE 2, Count: 1892
157 RAPE 1, Count: 7982
164 SODOMY 3, Count: 680
166 SODOMY 2, Count: 404
168 SODOMY 1, Count: 2783
176 Not Provided, Count: 162

```

Because the resulting key set is quite manageable at 91 records, I went through and removed non-violent crimes (kept keys with no description because we could not ascertain whether they were violent or not). This index was used to further filter out the arrest data to provide the most salient arrests. Over 23 thousand records were removed after this process.

A snippet of the cleaned arrest data of violent crimes:

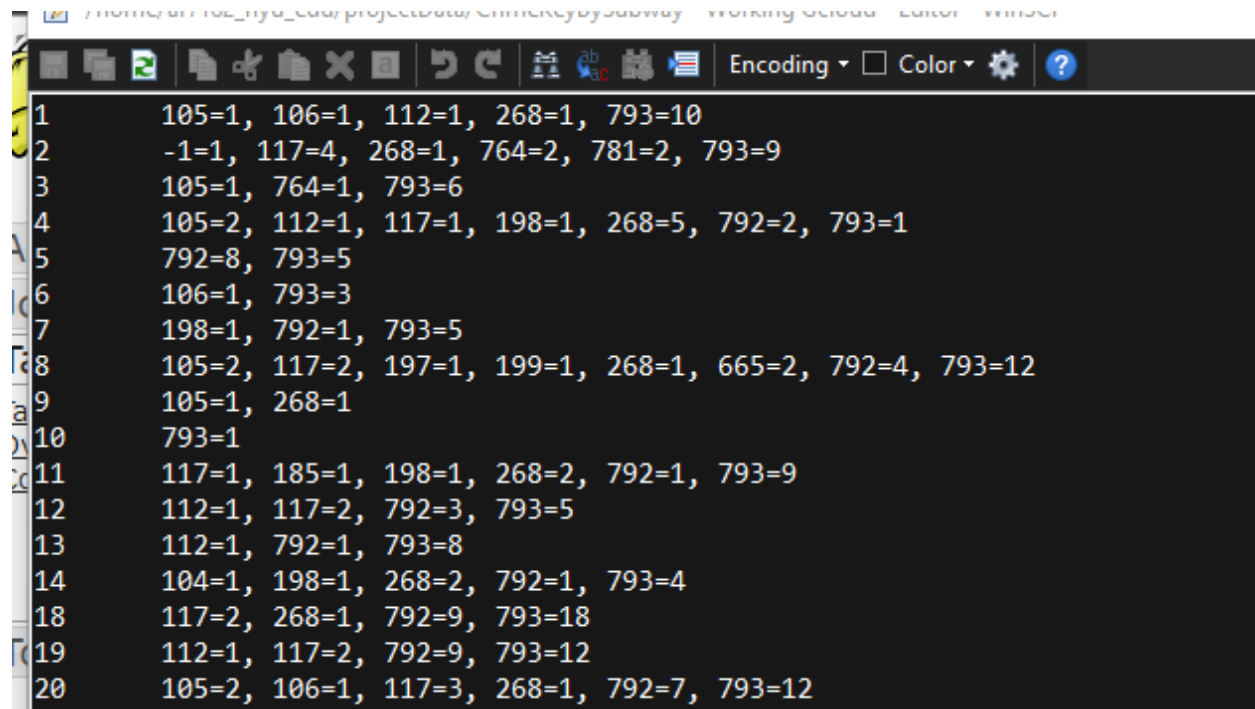


```

/home/ar7182_nyu_edu/projectData/cleanedArrestData - Working Gcloud - Editor - WinSCP
Encoding Color
237354740,12/04/2021,153,RAPE 3,B,40.816391847000034,-73.89529641399997
192799737,01/26/2019,177,SEXUAL ABUSE,M,40.800694331000045,-73.941109285999971
193260691,02/06/2019,,M,40.757839003000072,-73.991212110999982
237291769,12/03/2021,579,,Q,40.772056496000006,-73.87622400099998
149117452,01/06/2016,153,RAPE 3,K,40.648650085000035,-73.95033556299995
221756278,12/12/2020,,M,40.78756730100001,-73.94313233199995
237580757,12/09/2021,105,STRANGULATION 1ST,M,40.82867545800008,-73.94398971599996
190049060,11/15/2018,157,RAPE 1,K,40.67458330800008,-73.93022154099998
220422940,11/12/2020,157,RAPE 1,Q,40.72236368700004,-73.85147389399998
189182271,10/24/2018,153,RAPE 3,M,40.716195914000025,-73.99749074599998
214798896,07/01/2020,177,SEXUAL ABUSE,B,40.85396027400003,-73.90012087499997
196324211,04/23/2019,157,RAPE 1,K,40.674583308000081,-73.93022154099998
220303765,11/10/2020,153,RAPE 3,Q,40.72236368700004,-73.85147389399998
238116432,12/21/2021,579,,Q,40.680048726000045,-73.77590919399995
237302061,12/03/2021,,M,40.752862899000036,-73.99298133199994
238376572,12/28/2021,579,,K,40.63909210100008,-74.00940929499995
220917769,11/24/2020,157,RAPE 1,B,40.816391847000034,-73.89529641399997
236918575,11/24/2021,153,RAPE 3,Q,40.76930608700008,-73.91550817999997
238350561,12/28/2021,157,RAPE 1,B,40.816391847000034,-73.89529641399997
199836526,07/16/2019,105,STRANGULATION 1ST,Q,40.675491729000044,-73.800926136999976
236949124,11/26/2021,105,STRANGULATION 1ST,S,40.623226819000074,-74.14922697299994
189863002,11/11/2018,179,SEXUAL ABUSE 1,K,40.627977755000074,-73.94164715699996
192148546,01/10/2019,168,SODOMY 1,K,40.671106911000038,-73.881432956999959
189476017,11/01/2018,793,WEAPONS POSSESSION 3,K,40.61797007100006,-74.03033045599993
201100692,08/14/2019,105,STRANGULATION 1ST,M,40.78934789300007,-73.947352412999976

```

Using the previous information, I implemented a replicated join on the Mapper side and combined the violent crime data with the cleaned subway geo information data. To do so I cached the subwayGeoInformation file to be read by each mapper. In the setup process, the mapper node read the subway ID number along with its' latitude and longitude into a global HashMap. In the map function, the location of the committed felony was compared within a given tolerance of 0.0005 to all subway's latitude and longitudes. The reducer simply aggregated all of the values to create a list of subway IDs with the keycode of crimes committed in a close proximity with their related crime counts. Sample of the output data (-1 for crimes entered into NYPD dataset without a keycode):



```
1 105=1, 106=1, 112=1, 268=1, 793=10
2 -1=1, 117=4, 268=1, 764=2, 781=2, 793=9
3 105=1, 764=1, 793=6
4 105=2, 112=1, 117=1, 198=1, 268=5, 792=2, 793=1
5 792=8, 793=5
6 106=1, 793=3
7 198=1, 792=1, 793=5
8 105=2, 117=2, 197=1, 199=1, 268=1, 665=2, 792=4, 793=12
9 105=1, 268=1
10 793=1
11 117=1, 185=1, 198=1, 268=2, 792=1, 793=9
12 112=1, 117=2, 792=3, 793=5
13 112=1, 792=1, 793=8
14 104=1, 198=1, 268=2, 792=1, 793=4
18 117=2, 268=1, 792=9, 793=18
19 112=1, 117=2, 792=9, 793=12
20 105=2, 106=1, 117=3, 268=1, 792=7, 793=12
```

There was a total of 6439 crimes recorded in close proximity to subway stations (from map output records since each output was considered as one unique count of a crime).

This culmination of arrest data will be compared with the NYPD complaint dataset and Stop and Frisk dataset to see if there is a correlation between the three NYPD interactions. Ultimately, the goal is to recommend which stations should be the most heavily patrolled and what types of crimes should be expected.

The final MapReduce job was basically the same as the previous job, except instead of aggregating crime by station, I added the station ID to the filtered cleanedArrestData. I did this so we can later use HiveQL to write queries with time information. Sample from the dataset:

192389746,01/16/2019,,M,40.752198698000029,-73.993465044999937,437
188126065,09/28/2018,,M,40.79032242500005,-73.94768749799994,458
187490113,09/12/2018,157,RAPE 1,K,40.576157300000034,-73.97598379999994,198
214204908,06/15/2020,157,RAPE 1,K,40.576157300000034,-73.97598379999994,198
220803428,11/21/2020,792,WEAPONS POSSESSION 1 & 2,K,40.66256275400008,-73.90892111899994,218
220672343,11/18/2020,106,ASSAULT POLICE/PEACE OFFICER,M,40.82643872800003,-73.95045219099995,156
221024262,11/27/2020,106,ASSAULT POLICE/PEACE OFFICER,M,40.74978011300004,-73.98778087399995,366
221767153,12/13/2020,106,ASSAULT POLICE/PEACE OFFICER,B,40.90348934600007,-73.85034201899998,280
220970485,11/25/2020,268,CRIMINAL MIS 2 & 3,M,40.749157106000034,-73.98827182799994,145
221024260,11/27/2020,106,ASSAULT POLICE/PEACE OFFICER,M,40.74978011300004,-73.98778087399995,366
222099711,12/21/2020,268,CRIMINAL MIS 2 & 3,M,40.72063684600005,-74.00521077799993,410
222107788,12/21/2020,793,WEAPONS POSSESSION 3,B,40.83592508100002,-73.92183088599995,291
221569415,12/09/2020,268,CRIMINAL MIS 2 & 3,M,40.750664208000046,-73.99086657499998,358
221795661,12/14/2020,268,CRIMINAL MIS 2 & 3,K,40.682666375000046,-73.91002579499997,35
220190966,11/07/2020,792,WEAPONS POSSESSION 1 & 2,K,40.661440353000046,-73.91643415899993,221