Data Profiling and Cleansing

Dhruv Shetty - Real Time and Big Data Analysis

Data Source : https://data.cityofnewyork.us/api/views/qgea-i56i/rows.csv
Size : 2391 MB ; 7.83 Million Rows and 35 Columns with each row being a complaint
Important Columns from the Dataset

Index	Name	Description
0	CMPLNT_NUM	Randomly generated persistent ID for each complaint
1	CMPLNT_FR_DT	Exact date of occurrence for the reported event
2	CMPLNT_FR_TM	Exact time of occurrence for the reported event
7	KY_CD	Three digit offense classification code
8	OFNS_DESC	Description of offense corresponding with key code
12	LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
15	PREM_TYP_DESC	Specific description of premises
29	Lat_Lon	Geospatial Location Point
31	STATION_NAME	Transit station name

Step 1

Used curl to download and store the dataset in HDFS for further processing.

a. Total time taken : 8 minutes 19 seconds

```
dks7920_nyu_edu@nyu-dataproc-m:~$ curl https://data.cityofnewyork.us/api/views/qgea-i56i/rows.csv | hadoop fs -put - project/crime.csv
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 2391M    0 2391M    0     0  4897k      0 --:--:--  0:08:19 --:--:-- 5184k
```

Step 2

Compiled a MapReduce based on filtering the columns as listed. The columns were also checked for empty or incorrect rows and then filtered by felony. Finally putting through an identity reducer to get the output. From 7.8 million rows, it came down to 2.4 Million rows

Mapper Logic

- Split input based on comma.

- Drop malformed rows
- Read indexes as noted in table above
- Check for empty values in ID, Date, Time, Offense Code, Level of Offense, LatLong as these are needed for further analysis
- If Level of Offense is felony, write all required columns with a comma delimiter to output

Reducer : Identity Reducer

```
dks7920_nyu_edu@nyu-dataproc-m:~$ javac -classpath `hadoop classpath` CleanCrimeDataMapper.java
dks7920_nyu_edu@nyu-dataproc-m:~$ javac -classpath `hadoop classpath` CleanCrimeDataReducer.java
dks7920_nyu_edu@nyu-dataproc-m:~$ javac -classpath `hadoop classpath`:. CleanCrimeData.java
dks7920_nyu_edu@nyu-dataproc-m:~$ jar cvf cleanCrimeData.jar *.class
added manifest
adding: CleanCrimeData.class(in = 1468) (out= 828)(deflated 43%)
adding: CleanCrimeDataMapper.class(in = 2122) (out= 934)(deflated 55%)
adding: CleanCrimeDataReducer.class(in = 1584) (out= 638)(deflated 59%)
dks7920_nyu_edu@nyu-dataproc-m:~$ hadoop jar cleanCrimeData.jar CleanCrimeData project/crime.csv project/crimeFilterBasic/output
2022-11-30 00:28:02,743 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.61:8032
2022-11-30 00:28:02,925 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.61:10200
2022-11-30 00:28:03,246 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2022-11-30 00:28:03,308 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/dks7920_nyu_edu/.staging/job_1669608701841_0466
2022-11-30 00:28:03,596 INFO input.FileInputFormat: Total input files to process : 1
2022-11-30 00:28:03,766 INFO mapreduce.JobSubmitter: number of splits:19
2022-11-30 00:28:03,952 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1669608701841_0466
2022-11-30 00:28:03,954 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-11-30 00:28:04,144 INFO conf.Configuration: resource-types.xml not found
2022-11-30 00:28:04,144 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-11-30 00:28:04,291 INFO impl.YarnClientImpl: Submitted application application_1669608701841_0466
2022-11-30 00:28:04,329 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1669608701841_0466/
2022-11-30 00:28:04,329 INFO mapreduce.Job: Running job: job_1669608701841_0466
2022-11-30 00:28:11,426 INFO mapreduce.Job: Job job_1669608701841_0466 running in uber mode : false
```

```
        Map-Reduce Framework
                Map input records=7825500
                Map output records=2417683
                Map output bytes=281880469
                Map output materialized bytes=286815520
                Input split bytes=2375
                Combine input records=0
                Combine output records=0
                Reduce input groups=2417302
                Reduce shuffle bytes=286815520
                Reduce input records=2417683
                Reduce output records=2417683
                Spilled Records=4835366
```

Step 3

Run MapReduce to count the number of unique felony crimes within the filtered dataset by running a count job. Got 42 different offense codes with count as shown below

Mapper Logic:

- Read previous output as input - filtered date
- Output offense code as key and offense description as value with a 1 to count in reducer

Reducer Logic:

- For all values under one offense code, count number of values

```
dks7920_nyu_edu@nyu-dataproc-m:~$ javac -classpath `hadoop classpath`:. UniqueCrime.java
dks7920_nyu_edu@nyu-dataproc-m:~$ jar cvf UniqueCrime.jar *.class
added manifest
adding: UniqueCrime.class(in = 1456) (out= 830)(deflated 42%)
adding: UniqueCrimeMapper.class(in = 1748) (out= 724)(deflated 58%)
adding: UniqueCrimeReducer.class(in = 1894) (out= 815)(deflated 56%)
```

```
dks7920_nyu_edu@nyu-dataproc-m:~$ jar cvf UniqueCrime.jar *.class
added manifest
adding: UniqueCrime.class(in = 1456) (out= 830)(deflated 42%)
adding: UniqueCrimeMapper.class(in = 1670) (out= 682)(deflated 59%)
adding: UniqueCrimeReducer.class(in = 1910) (out= 823)(deflated 56%)
dks7920_nyu_edu@nyu-dataproc-m:~$ hadoop jar UniqueCrime.jar UniqueCrime project/crimeFilterBasic/output/part-r-00000 project/uniqueCrime/output2
2022-11-30 02:41:02,551 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.61:8032
2022-11-30 02:41:02,752 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.61:10200
2022-11-30 02:41:03,060 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRun
ner to remedy this.
2022-11-30 02:41:03,080 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/dks7920_nyu_edu/.staging/job_1669608701841_0526
2022-11-30 02:41:03,369 INFO input.FileInputFormat: Total input files to process : 1
2022-11-30 02:41:03,527 INFO mapreduce.JobSubmitter: number of splits:2
2022-11-30 02:41:03,736 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1669608701841_0526
2022-11-30 02:41:03,738 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-11-30 02:41:03,942 INFO conf.Configuration: resource-types.xml not found
2022-11-30 02:41:03,943 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-11-30 02:41:04,087 INFO impl.YarnClientImpl: Submitted application application_1669608701841_0526
2022-11-30 02:41:04,129 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1669608701841_0526/
2022-11-30 02:41:04,130 INFO mapreduce.Job: Running job: job_1669608701841_0526
2022-11-30 02:41:11,224 INFO mapreduce.Job: Job job_1669608701841_0526 running in uber mode : false
2022-11-30 02:41:11,225 INFO mapreduce.Job:  map 0% reduce 0%
2022-11-30 02:41:28,352 INFO mapreduce.Job:  map 54% reduce 0%
```

```
-rw-r--r--   1 dks7920_nyu_edu dks7920_nyu_edu        846 2022-11-30 02:41 project/uniqueCrime/ou
dks7920_nyu_edu@nyu-dataproc-m:~$ hadoop fs -cat project/uniqueCrime/output2/part-r-00000
101     MURDER & NON-NEGL. MANSLAUGHTER:6306
102     HOMICIDE-NEGLIGENT-VEHICLE:109
103     "HOMICIDE-NEGLIGENT:179
104     RAPE:23073
105     ROBBERY:279673
106     FELONY ASSAULT:308361
107     BURGLARY:266342
109     GRAND LARCENY:678343
110     GRAND LARCENY OF MOTOR VEHICLE:143625
111     POSSESSION OF STOLEN PROPERTY:13849
112     THEFT-FRAUD:80324
113     FORGERY:76989
114     ARSON:18247
115     PROSTITUTION & RELATED OFFENSES:256
116     SEX CRIMES:19214
117     DANGEROUS DRUGS:88203
118     DANGEROUS WEAPONS:76796
119     INTOXICATED/IMPAIRED DRIVING:63
120     CHILD ABANDONMENT/NON SUPPORT:572
121     CRIMINAL MISCHIEF & RELATED OF:137517
122     GAMBLING:150
123     ABORTION:7
124     KIDNAPPING & RELATED OFFENSES:3049
125     NYS LAWS-UNCLASSIFIED FELONY:7455
126     MISCELLANEOUS PENAL LAW:188460
233     :10
235     :2
236     :3
340     :2
341     :1
343     :1
344     :11
347     :1
351     :3
352     :2
355     :1
359     :13
361     :8
364     :455
365     :1
366     :1
578     :6
dks7920_nyu_edu@nyu-dataproc-m:~$ rm UniqueCrimeMapper.java
```

Tried by using the offense code + description as the key. Got 56 different offenses

```
Map-Reduce Framework
        Map input records=2417683
        Map output records=2417683
        Map output bytes=57105282
        Map output materialized bytes=61940660
        Input split bytes=304
        Combine input records=0
        Combine output records=0
        Reduce input groups=56
        Reduce shuffle bytes=61940660
        Reduce input records=2417683
        Reduce output records=56
        Spilled Records=4835366
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=306
```

Analysis of the difference between the 56 and 42 lead me to understand that the difference was just because some of the key codes were missing descriptions and had slightly different names that led to the same offense. Therefore the 42 key codes were good enough to take into consideration.

Step 4

Run MapReduce to filter out only violent crimes and remove the non-violent crimes so that we can only use the most dangerous complaints.

Key Codes of Violent Crimes (Analyzed from previous output) :
101,102,103,104,105,106,114,115,116,117,118,124,125,233,235,236,340,341,343,344,347,351,352,355,359,361,364,365,366,578

```
Map-Reduce Framework
        Map input records=2417683
        Map output records=831442
        Map output bytes=93798190
        Map output materialized bytes=95471054
        Input split bytes=304
        Combine input records=0
        Combine output records=0
        Reduce input groups=831396
        Reduce shuffle bytes=95471054
        Reduce input records=831442
        Reduce output records=831442
        Spilled Records=1662884
```

2.4 Million Rows has been filtered down to 0.8 Million rows for complaints

Extra

Interested to analyze the number of violent crime complaints per hour of the day to understand if there is a skewing to a time of day.
Run MapReduce to count the number of violent crimes within the filtered dataset by running a count job using the hour as the key

```
Map-Reduce Framework
        Map input records=831442
        Map output records=831442
        Map output bytes=5820094
        Map output materialized bytes=7482984
        Input split bytes=154
        Combine input records=0
        Combine output records=0
        Reduce input groups=24
        Reduce shuffle bytes=7482984
        Reduce input records=831442
        Reduce output records=24
```

1 output per hour

```
00      51615
01      41698
02      35516
03      32294
04      29228
05      19111
06      18571
07      15191
08      19266
09      19208
10      20832
11      23392
12      29015
13      30100
14      36204
15      44420
16      42131
17      41737
18      42683
19      44600
20      47912
21      48829
22      49480
23      48409
dks7920_nyu_ed
```

Snippet of final output

```
118935871,03/11/2017,04:02:00,106,FELONY ASSAULT,FELONY,DEPARTMENT STORE,"(40.862707191, -73.902808351)",
118936158,11/13/2021,14:55:00,106,FELONY ASSAULT,FELONY,RESIDENCE-HOUSE,"(40.71091484500005, -73.78220073599994)",
118937538,08/27/2008,12:45:00,117,DANGEROUS DRUGS,FELONY,STREET,"(40.694996471, -73.907297305)",
118939300,02/10/2014,15:30:00,106,FELONY ASSAULT,FELONY,TRANSIT - NYC SUBWAY,"(40.683767303, -73.978742975)",PACIFIC STREET
118939888,03/23/2007,20:55:00,118,DANGEROUS WEAPONS,FELONY,STREET,"(40.850808226, -73.932650058)",
118940579,12/28/2008,18:00:00,105,ROBBERY,FELONY,STREET,"(40.667163928, -73.910093567)",
118941278,12/30/2009,22:30:00,117,DANGEROUS DRUGS,FELONY,STREET,"(40.769458934, -73.790186437)",
118942166,07/11/2008,20:00:00,106,FELONY ASSAULT,FELONY,TRANSIT - NYC SUBWAY,"(40.692241674, -73.987300996)",JAY STREET-BOROUGH HALL
118943225,12/01/2018,06:00:00,106,FELONY ASSAULT,FELONY,RESIDENCE - APT. HOUSE,"(40.721355481, -73.988146846)",
118943946,01/04/2017,12:15:00,105,ROBBERY,FELONY,STREET,"(40.656179569, -73.951609408)",
118944882,01/12/2010,15:50:00,106,FELONY ASSAULT,FELONY,FAST FOOD,"(40.632240547, -73.947624063)",
118950198,11/17/2011,00:05:00,106,FELONY ASSAULT,FELONY,BAR/NIGHT CLUB,"(40.750664208, -73.990866575)",
118950455,05/08/2016,19:50:00,106,FELONY ASSAULT,FELONY,RESIDENCE-HOUSE,"(40.890888801, -73.857079576)",
118950595,12/28/2014,02:57:00,114,ARSON,FELONY,STREET,"(40.816232675, -73.852630025)",
118952354,08/30/2019,02:55:00,106,FELONY ASSAULT,FELONY,GROCERY/BODEGA,"(40.84324043000004, -73.93635868099993)",
118953642,04/21/2015,20:25:00,118,DANGEROUS WEAPONS,FELONY,STREET,"(40.659345203, -73.92725676)",
118955649,05/14/2006,06:10:00,105,ROBBERY,FELONY,STREET,"(40.838512492, -73.905980617)",
118956082,07/13/2015,11:50:00,106,FELONY ASSAULT,FELONY,RESIDENCE - PUBLIC HOUSING,"(40.698312502, -73.953147596)",
118956317,06/15/2011,19:40:00,118,DANGEROUS WEAPONS,FELONY,RESIDENCE - APT. HOUSE,"(40.830641297, -73.874879189)",
118956987,12/15/2008,19:30:00,105,ROBBERY,FELONY,STREET,"(40.668806093, -73.931121567)",
118957782,03/26/2006,00:15:00,105,ROBBERY,FELONY,STREET,"(40.596601118, -73.934890539)",
118959953,05/11/2007,07:36:00,106,FELONY ASSAULT,FELONY,STREET,"(40.745939829, -73.883746644)",
118960505,12/05/2007,05:50:00,105,ROBBERY,FELONY,RESIDENCE - APT. HOUSE,"(40.672517206, -73.868333593)",
118961596,08/21/2013,05:00:00,105,ROBBERY,FELONY,STREET,"(40.822368382, -73.950614198)",
118964031,04/11/2010,21:30:00,105,ROBBERY,FELONY,TRANSIT - NYC SUBWAY,"(40.70777977, -74.01296455)",RECTOR STREET
118964139,07/12/2015,01:50:00,105,ROBBERY,FELONY,RESIDENCE - PUBLIC HOUSING,"(40.836652519, -73.907143295)",
118964960,11/12/2007,17:20:00,106,FELONY ASSAULT,FELONY,OTHER,"(40.725980641, -73.990801407)",
118965260,07/12/2012,19:50:00,106,FELONY ASSAULT,FELONY,RESIDENCE - PUBLIC HOUSING,"(40.696450014, -73.943606737)",
118965791,07/03/2007,17:00:00,118,DANGEROUS WEAPONS,FELONY,RESIDENCE - APT. HOUSE,"(40.726975307, -73.978281813)",
118967288,10/28/2016,15:20:00,106,FELONY ASSAULT,FELONY,VARIETY STORE,"(40.693702488, -73.754962204)",
118968593,01/12/2007,21:15:00,106,FELONY ASSAULT,FELONY,RESIDENCE - PUBLIC HOUSING,"(40.669132419, -73.86555017)",
118968960,09/03/2017,12:00:00,105,ROBBERY,FELONY,RESIDENCE - APT. HOUSE,"(40.831247096, -73.909232255)",
118969295,01/26/2014,02:50:00,106,FELONY ASSAULT,FELONY,RESIDENCE - APT. HOUSE,"(40.698969978, -73.916873982)",
118969470,01/04/2014,19:55:00,105,ROBBERY,FELONY,RESIDENCE - PUBLIC HOUSING,"(40.681803475, -73.923789039)",
118970044,10/21/2019,15:30:00,106,FELONY ASSAULT,FELONY,RESIDENCE - APT. HOUSE,"(40.744768383000064, -73.97280134899995)",
118971018,12/02/2010,00:45:00,117,DANGEROUS DRUGS,FELONY,STREET,"(40.68579842, -73.860692154)",
118972314,10/23/2007,20:45:00,105,ROBBERY,FELONY,STREET,"(40.596339568, -74.087034091)",
118974175,09/12/2007,00:30:00,105,ROBBERY,FELONY,OTHER,"(40.645138905, -73.948965248)",
```

831442 Rows