1: Datasets

- All of the datasets are available at: https://www.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page
- This is the data about stop, question, and frisk.
- We will be using datasets between years: 2013 – 2016
- Every year has a different dataset file (in total 4 datasets).
- Datasets between 2013 and 2016 are in .csv format and have similar column layout.

Year:	2013	2014	2015	2016	TOTAL:
Rows:	191851	45787	22563	12401	272602
Size:	61,3MB	15MB	7,2MB	4MB	87,5MB

In the first step, I am going to merge data from 2013 to 2016 into one file.

```
ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016$ cat *.csv > 2013-2016.csv
ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016$ ls
2013-2016.csv  sqf-2013.csv  sqf-2014.csv  sqf-2015.csv  sqf-2016.csv
ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016$ hadoop fs -rm  Project/*.csv
Deleted Project/sqf-2013.csv
Deleted Project/sqf-2014.csv
Deleted Project/sqf-2015.csv
Deleted Project/sqf-2016.csv
ss16249_nyu_edu@nyu-dataproc-m:~$ hadoop fs -put SubwayProject/2013-2016/data.csv Project
ss16249_nyu_edu@nyu-dataproc-m:~$ hadoop fs -ls Project
Found 1 items
-rw-r--r--   1 ss16249_nyu_edu ss16249_nyu_edu   87701814 2022-11-29 09:26 Project/data.csv
```

This merged dataset has now 272, 602 rows and 112 columns in total. Then I put it into HDFS cluster.

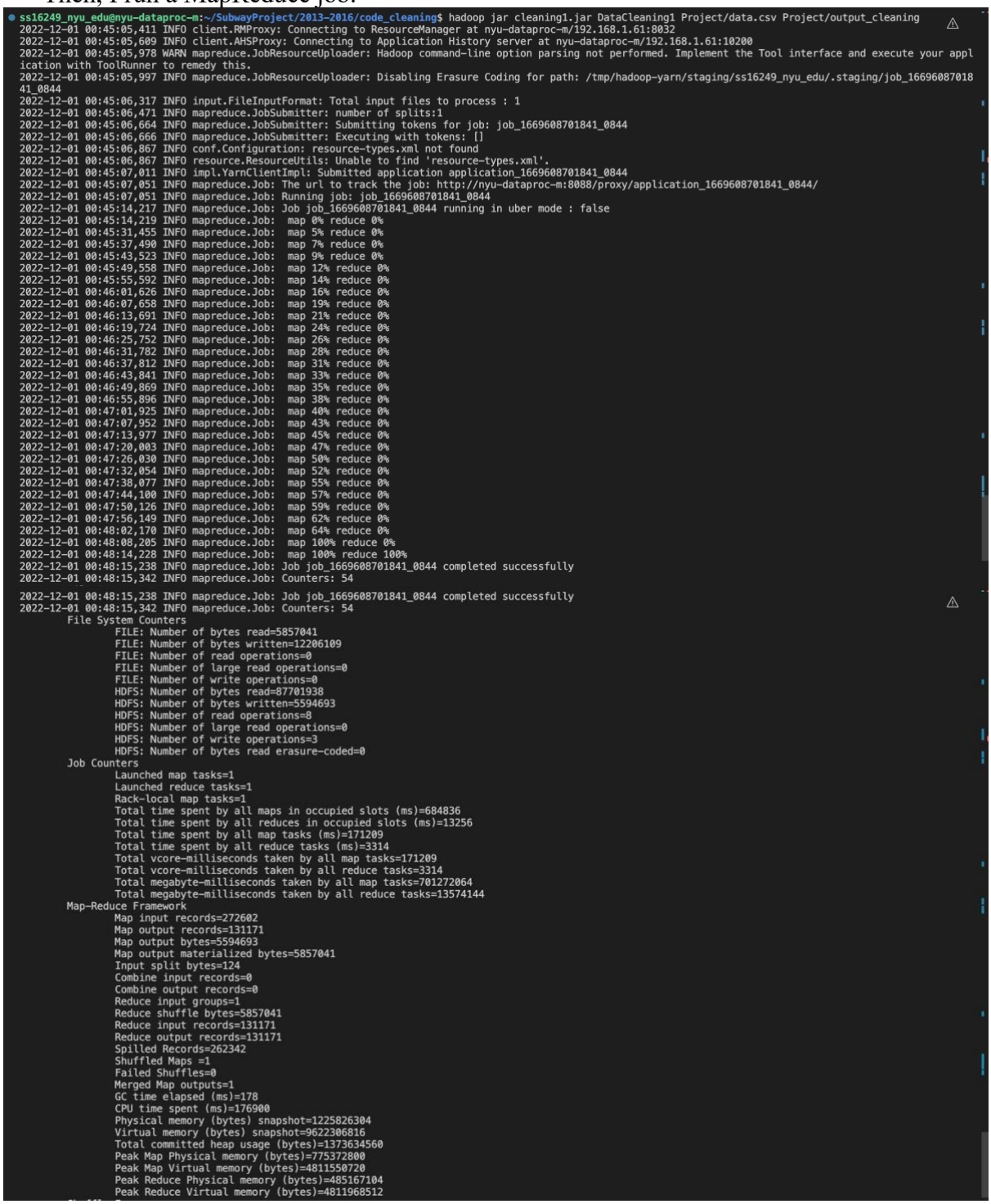
- I have manually deleted the 1 row, which is description of the columns in each of the four dataset files (sqf-2013.csv - sqf-2016.csv).
- As the data have 112 columns, I will not provide the data here in the report. (The short sample of the data will be provided in the .zip file).

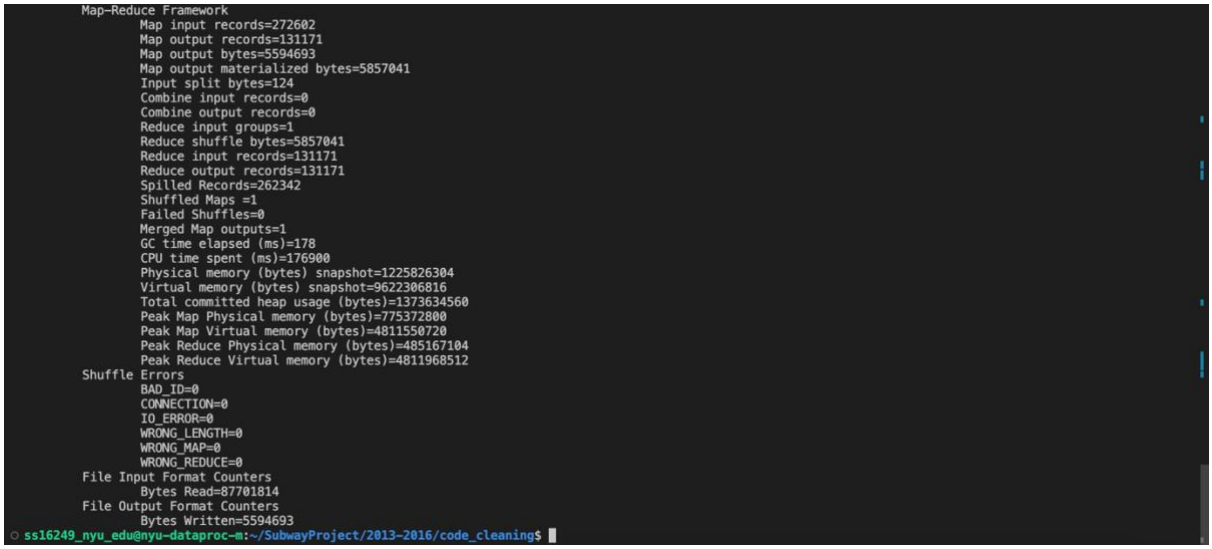
2: Data Cleaning

- The most important columns that I need from the dates are:
  - DATE OF STOP
  - TIME OF STOP
  - CRIME DESCRIPTION
  - X COORDINATES OF STOP
  - Y COORDINATES OF STOP
- I have filtered the data on the following criteria:
  - DATE: I checked if the date is not empty, if it is I dropped the row.
  - TIME: I checked if the time is in correct format (HHmm). If the time, was in format (Hmm), I added "0" before it. Otherwise, I dropped the row.
  - X&Y COORDINATES: I checked if the X Coordinate and Y Coordinate is not empty. If it was, I dropped the row.
  - VIOLENT CRIME:
    - I have considered something violent crime based on the following criteria:
      - If any weapon was found on the person.
      - If there was necessary physical force used by NYPD.
      - If the REASON FOR FRISK - VIOLENT CRIME SUSPECTED was "Y".

- Based on the chosen HashMap of Violent Crimes discussed with my teammates.
- If the CRIME DESCRIPTION was empty, I assumed it violent. (We agreed this way with my teammates.)

- Then, I run a MapReduce job.





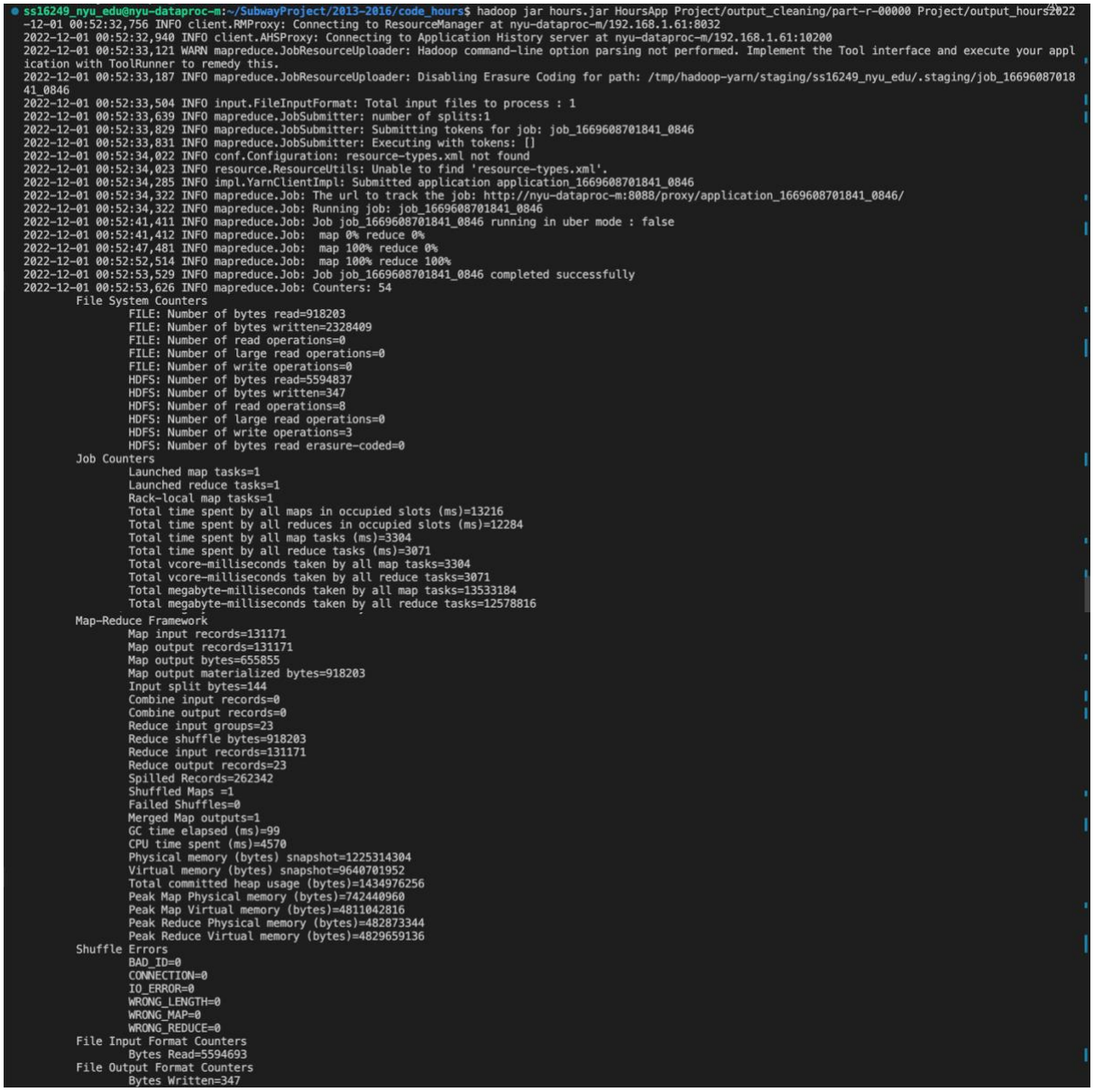
- This is sample cleaned data:

SubwayProject > 2013-2016 > code_cleaning > output > part-r-00000

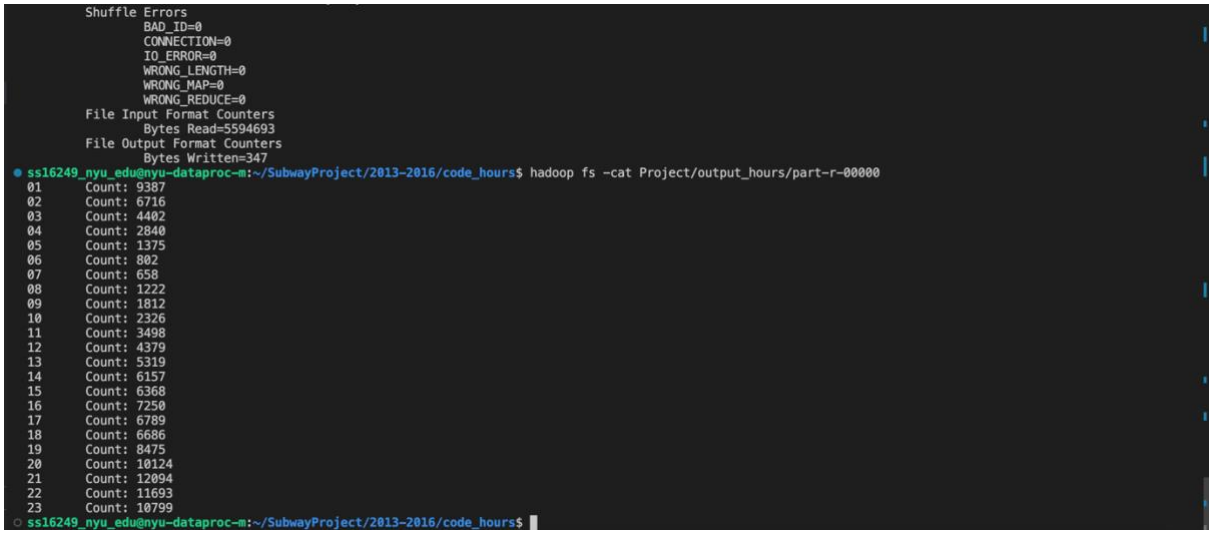
```
1   11222016,2212,CPW,939042,132823
2   12022016,1847,ASSAULT,935054,136263
3   11032016,1725,VIOLENT_CRIME_SUSPECTED,922098,135797
4   11032016,1035,ASSAULT,917675,129284
5   10162016,1951,CPW,918014,123544
6   10162016,1951,CPW,918014,123544
7   10162016,1951,CPW,918014,123544
8   10132016,1320,CPW,923947,129507
9   10112016,0120,PHYSICAL_FORCE_USED,925065,140569
10  12242016,1935,PHYSICAL_FORCE_USED,961179,159021
11  12222016,0304,VIOLENT_CRIME_SUSPECTED,958629,154949
12  12222016,0304,VIOLENT_CRIME_SUSPECTED,958629,154949
13  12172016,1404,CRIMINAL SALE OF CONTROLLED SUBSTANCE,942326,140590
14  12172016,1404,CRIMINAL SALE OF CONTROLLED SUBSTANCE,942326,140590
15  12112016,1454,PHYSICAL_FORCE_USED,951527,146987
16  12112016,0425,VIOLENT_CRIME_SUSPECTED,954598,143146
17  11242016,0315,PHYSICAL_FORCE_USED,938281,137671
18  11242016,0910,PHYSICAL_FORCE_USED,965461,154802
19  11242016,0910,PHYSICAL_FORCE_USED,965461,154802
20  11222016,0530,PHYSICAL_FORCE_USED,957236,161624
```

3: Data Profiling

- In the profiling, I wrote 2 MapReduce jobs. One for summarizing what are the total crimes per each hour of the day and another one, which summarizes the frequency of each crime description.
- Here is the MapReduce job log for the total crimes per each hour:

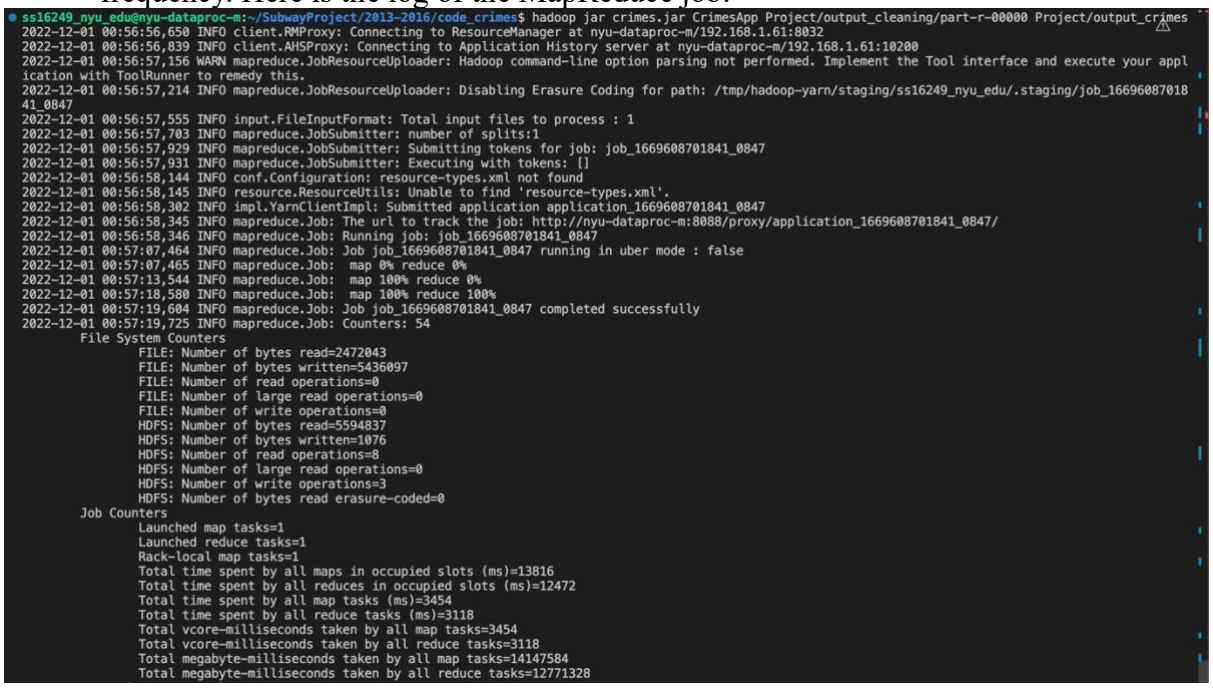


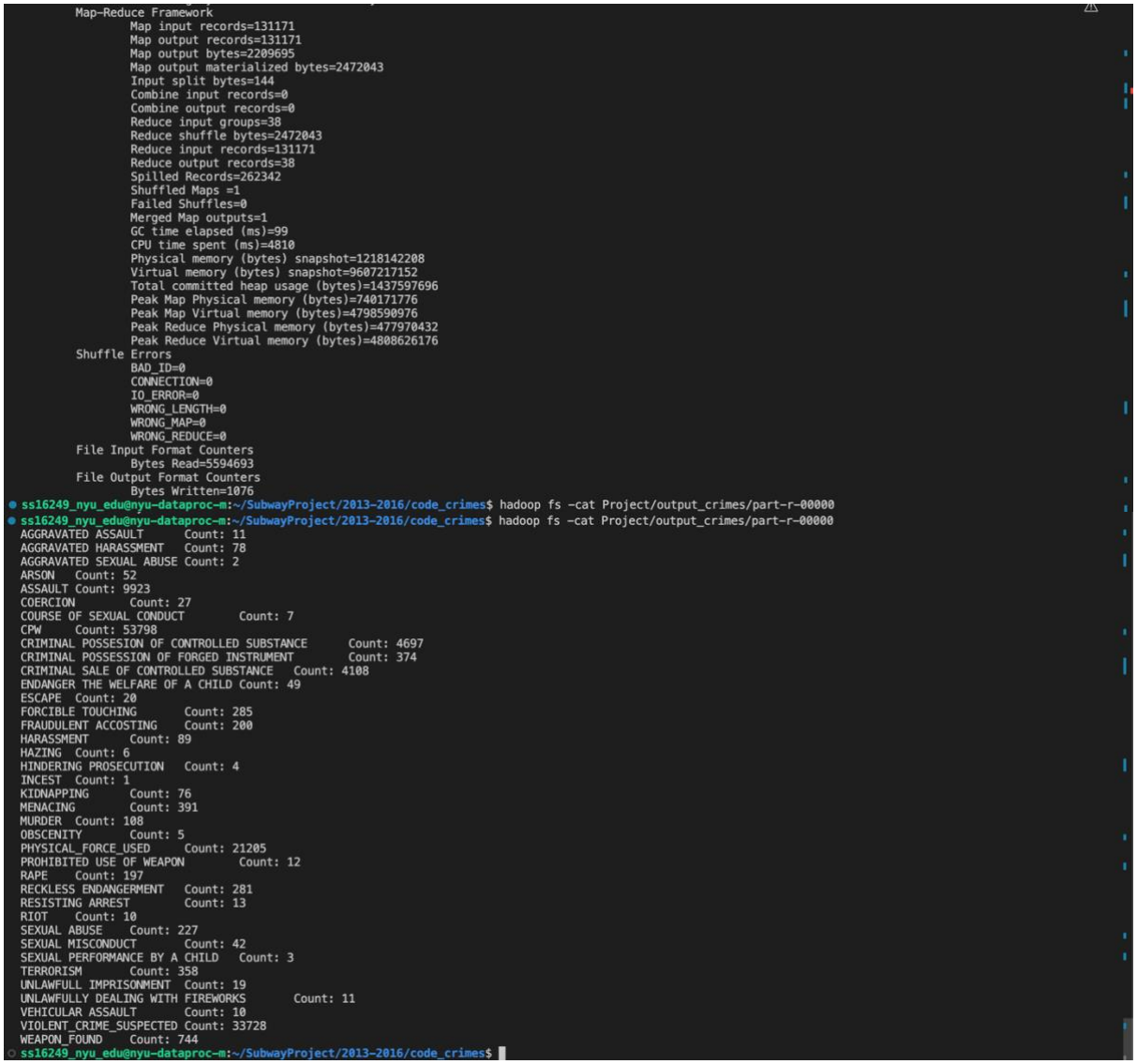
```
ss16249_nyu_edu@nyu-dataproc-m:~/SubwayProject/2013-2016/code_hours$ hadoop jar hours.jar HoursApp Project/output_cleaning/part-r-00000 Project/output_hours2022
-12-01 00:52:32,756 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.61:8032
2022-12-01 00:52:32,940 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.61:10200
2022-12-01 00:52:33,121 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your appl
ication with ToolRunner to remedy this.
2022-12-01 00:52:33,187 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/ss16249_nyu_edu/.staging/job_16696087018
41_0846
2022-12-01 00:52:33,504 INFO input.FileInputFormat: Total input files to process : 1
2022-12-01 00:52:33,639 INFO mapreduce.JobSubmitter: number of splits:1
2022-12-01 00:52:33,829 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1669608701841_0846
2022-12-01 00:52:33,831 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-12-01 00:52:34,022 INFO conf.Configuration: resource-types.xml not found
2022-12-01 00:52:34,023 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-12-01 00:52:34,285 INFO impl.YarnClientImpl: Submitted application application_1669608701841_0846
2022-12-01 00:52:34,322 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1669608701841_0846/
2022-12-01 00:52:34,322 INFO mapreduce.Job: Running job: job_1669608701841_0846
2022-12-01 00:52:41,411 INFO mapreduce.Job: Job job_1669608701841_0846 running in uber mode : false
2022-12-01 00:52:41,412 INFO mapreduce.Job:  map 0% reduce 0%
2022-12-01 00:52:47,481 INFO mapreduce.Job:  map 100% reduce 0%
2022-12-01 00:52:52,514 INFO mapreduce.Job:  map 100% reduce 100%
2022-12-01 00:52:53,529 INFO mapreduce.Job: Job job_1669608701841_0846 completed successfully
2022-12-01 00:52:53,626 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=918203
                FILE: Number of bytes written=2328409
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5594837
                HDFS: Number of bytes written=347
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=13216
                Total time spent by all reduces in occupied slots (ms)=12284
                Total time spent by all map tasks (ms)=3304
                Total time spent by all reduce tasks (ms)=3071
                Total vcore-milliseconds taken by all map tasks=3304
                Total vcore-milliseconds taken by all reduce tasks=3071
                Total megabyte-milliseconds taken by all map tasks=13533184
                Total megabyte-milliseconds taken by all reduce tasks=12578816
        Map-Reduce Framework
                Map input records=131171
                Map output records=131171
                Map output bytes=655855
                Map output materialized bytes=918203
                Input split bytes=144
                Combine input records=0
                Combine output records=0
                Reduce input groups=23
                Reduce shuffle bytes=918203
                Reduce input records=131171
                Reduce output records=23
                Spilled Records=262342
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=99
                CPU time spent (ms)=4570
                Physical memory (bytes) snapshot=1225314304
                Virtual memory (bytes) snapshot=9640701952
                Total committed heap usage (bytes)=1434976256
                Peak Map Physical memory (bytes)=742440960
                Peak Map Virtual memory (bytes)=4811042816
                Peak Reduce Physical memory (bytes)=482873344
                Peak Reduce Virtual memory (bytes)=4829659136
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=5594693
        File Output Format Counters
                Bytes Written=347
```



- From the results, we can see that the crime in the NYC is highest from 20:00 to 01:00 in the night. The least violent crime is between 06:00 to 07:00.

- Let's move to the second MapReduce job, where we are finding the violent crime frequency. Here is the log of the MapReduce job:





- As we can see the most crimes were with concealed prohibited weapon (CPW).

4: Future Work

- As there is still more data to clean and profile, it is obvious that I would like to add more data to this dataset.
- To do this, I would have to write another DataCleaning Mapper program, as the column layout and column data type changed from year to year before 2013 and after 2016. This will require to write a new DataCleaning program for every year separately and concatenate the result to the results that I have already produced.