NYC Parking Violation

1. When are tickets most likely to be issued?

Analysis

Among all those attributes which are relevant, we use the 'Violation Time' to continue. Because we think it's more important to know the daily time than which month.

The sample of violation time is like: 0912A, 0548P.

Design

- Mapper

Input: Source data.

Output:

<key, value> // Key is the time of records.

// Value is the count, which should '1' for each one.

- Reducer

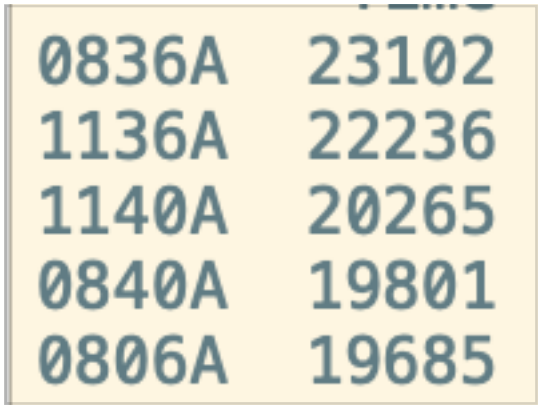
Input: Result of mapper

Output:

It collects and counts the frequency of each unique time first. Then find the record with largest count numbers, which should be the most likely time for tickets to be issued.

<key, value> // Key is the time of records.

// Value is the count of that record.



The most likely time for tickets to be issued: 08:36 am.

Review

I count the time accurate to the minute. Because I think it's easy to guess the peak hour would be the morning, which is the time period for people to office and find a place to park.

2. What are the most common years and types of cars to be ticketed?

Design

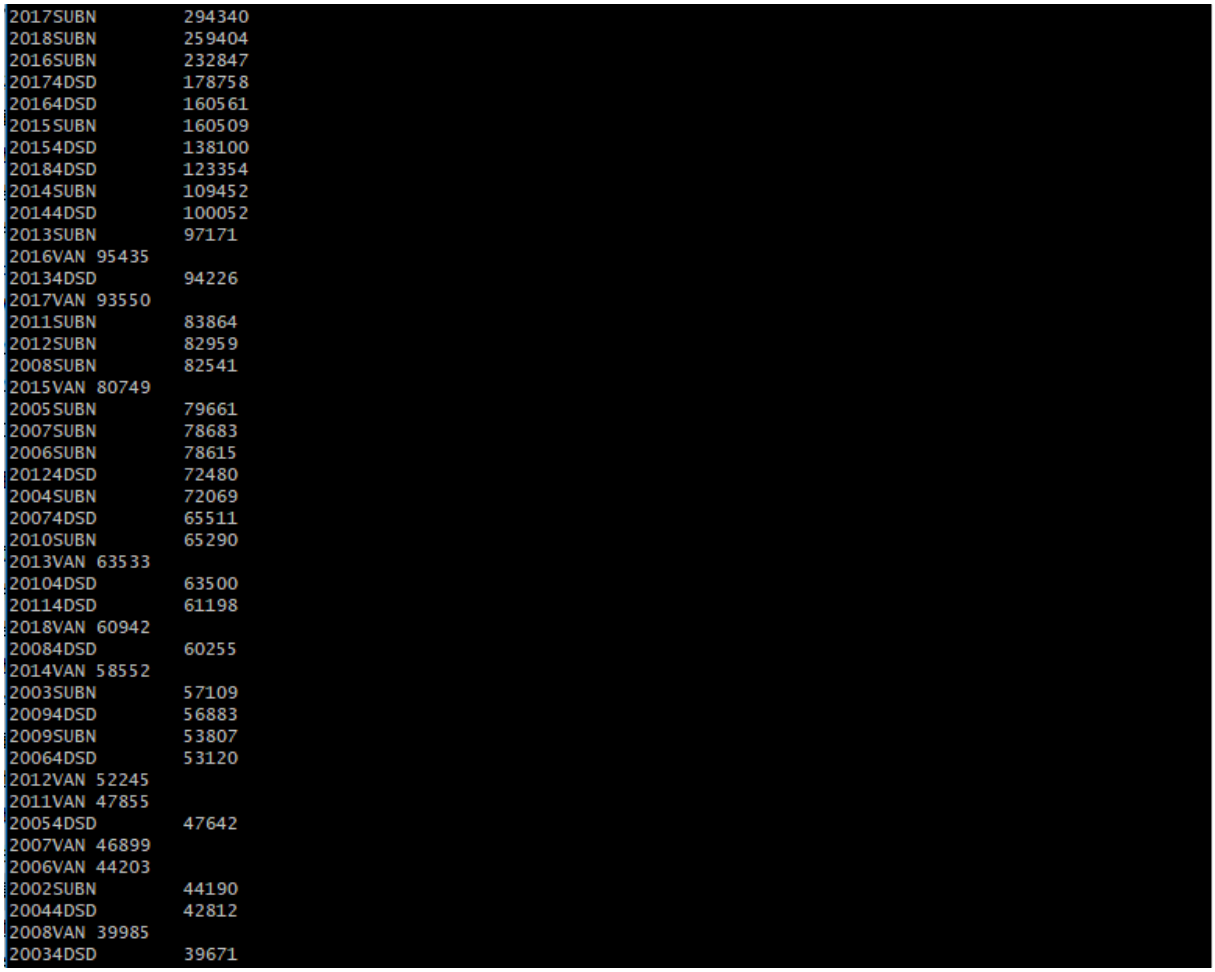
- Mapper

We extract the year and type information from the column Vehicle Year and Vehicle Body

Type. the output key is a combined string of Vehicle Year and Vehicle Body Type, the value is 1 count.

- Reducer

  It counts the frequency for each key taken from the mapper and sort the key by the cumulative counts descendingly. The most common years and types of cars are shown below with the highest frequency on top.



3. Where are tickets most commonly issued?

Design

- Mapper

  We extract the code information from “Violation Location” column from each line, and pass it to the variable “location”. The output key is the string of location, the output value is 1 count.

- Reducer

  For each new location, we record its accumulated counts. Then we sort the locations by their counts in the descending order. The top one is where the tickets are most commonly issued.

```
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1224827660
        File Output Format Counters
                Bytes Written=1519
19/03/28 20:50:51 INFO streaming.StreamJob: Output directory: /1-3/output/
No Record       1028148
0019    310779
0014    235934
0018    230483
0001    212499
0013    196377
114     188587
109     141030
115     127978
0006    124429
0020    116952
0070    116263
112     114190
0017    110621
0084    110337
103     109881
0052    104867
108     104592
0010    102665
0061    99744
104     96004
0066    93560
0046    93333
0009    87516
0024    85540
0062    84049
0005    81835
0090    81502
110     80601
0043    79056
0034    77731
102     77599
0068    72586
0067    71978
0094    71345
0075    70956
0049    69299
0023    67964
0079    66490
0078    65790
107     64860
0047    64837
0044    63596
0048    61556
0072    58269
0083    56380
0033    54482
```

4. Which color of the vehicle is most likely to get a ticket?

- Mapper

  We extract color from “Vehicle Color” column line by line. Empty values are converted to

"No Record". Filter values which are not English Letters. The output key is a string of Vehicle Color, the value is 1 count.

- Reducer

  It counts the frequency of the Vehicle Color and sort descendingly. The most common color of cars are shown below with the highest frequency on top. After we run each of them successfully, we put them into YARN with Fair Scheduler to manage resource efficiently.