

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.

0836A	23102
1136A	22236
1140A	20265
0840A	19801
0806A	19685

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.

2017SUBN	294340
2018SUBN	259404
2016SUBN	232847
20174DSD	178758
20164DSD	160561
2015SUBN	160509
20154DSD	138100
20184DSD	123354
2014SUBN	109452
20144DSD	100052
2013SUBN	97171
2016VAN 95435	
20134DSD	94226
2017VAN 93550	
2011SUBN	83864
2012SUBN	82959
2008SUBN	82541
2015VAN 80749	
2005SUBN	79661
2007SUBN	78683
2006SUBN	78615
20124DSD	72480
2004SUBN	72069
20074DSD	65511
2010SUBN	65290
2013VAN 63533	
20104DSD	63500
20114DSD	61198
2018VAN 60942	
20084DSD	60255
2014VAN 58552	
2003SUBN	57109
20094DSD	56883
2009SUBN	53807
20064DSD	53120
2012VAN 52245	
2011VAN 47855	
20054DSD	47642
2007VAN 46899	
2006VAN 44203	
2002SUBN	44190
20044DSD	42812
2008VAN 39985	
20034DSD	39671

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.

```
root@instance-1: /project1/1-4
Map-Reduce Framework
  Map input records=6945029
  Map output records=6945028
  Map output bytes=41139907
  Map output materialized bytes=55030023
  Input split bytes=1370
  Combine input records=0
  Combine output records=0
  Reduce input groups=1353
  Reduce shuffle bytes=55030023
  Reduce input records=6945028
  Reduce output records=1350
  Spilled Records=13890056
  Shuffled Maps =10
  Failed Shuffles=0
  Merged Map outputs=10
  GC time elapsed (ms)=52956
  CPU time spent (ms)=115410
  Physical memory (bytes) snapshot=2261934080
  Virtual memory (bytes) snapshot=21311967232
  Total committed heap usage (bytes)=1439604736
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=1224827660
File Output Format Counters
  Bytes Written=9623
19/03/24 21:30:58 INFO streaming.StreamJob: Output directory: /proj1-4/output/
WH      1322719
GY      1163511
BK      1027129
WHITE   760800
BL      406526
BLACK   328096
RD      260231
GREY    219776
BROWN   192164
SILVE   139439
BLUE    128508
RED     114907
GR      111961
No Record      83607
TN         73650
OTHER        60537
```