Bonus Question 1:

- Given a Black vehicle parking illegally at 34510, 10030, 34050 (street codes). What is the probability that it will get a ticket? (very rough prediction).

Conclusion: Probability of black vehicle parking illegally at 34510, 10030, and 34050: 19.11%

Mapper 1: We will be filtering data with black vehicles, street codes 34510, 10030, and 34050, and violation descriptions that relate to illegal parking, which will be part of parking infractions. We will end up having street code as a key and parking infractions as a value.

As our probability space we take every black vehicle at the given street codes and we use any combination of these points, then consider as a parking violation any appearance of the word stand or the fragment “park” the violation description field.

```
import sys
import re
black_color = re.compile('[a-zA-Z]*[B|b][a-zA-Z]*[c|k]')
street_codes = ['34510', '10030', '34050']
parking_tickets =
re.compile('[A-Za-z0-9]*stand+[A-Za-z0-9]*|[A-Za-z0-9]*park+[A-Za-z0-9]*')
for line in sys.stdin:
  line = line.strip(',').split(',')
  line_len = len(line)
  color = line[33].strip().strip('.'); color = ''.join(color.split(' ')).lower()
  color_val = black_color.search(color)
  street_code_1 = line[9]; street_code_2 = line[10]; street_code_3 = line[11]
  if color_val and line_len == 43 and (street_code_1 in street_codes or street_code_2 in
street_codes or street_code_3 in street_codes):
    violation_description = re.sub('\W+','',line[39]).lower()
    parking = parking_tickets.search(violation_description)
    parking_infraction = 1 if parking else 0
    if street_code_1 in street_codes:
      street_code = street_code_1
    elif street_code_2 in street_codes:
```

```
            street_code = street_code_2
        else:
            street_code = street_code_3
        print('{}\t{}'.format(street_code, parking_infraction))
    else:
        continue
```

Reducer 1: Using mapper 1's output, it will split it into a key and value that will be converted to a dictionary in order to call its items and do a global aggregation.

```
import sys
dict_parking_violations = {}
for line in sys.stdin:
    line = line.strip().split('\t')
    key, value = line[0], line[1]
    try:
        value = int(value)
        dict_parking_violations[key] = [dict_parking_violations.get(key, [0,0])[0] + value,
dict_parking_violations.get(key, [0,0])[1]+1]
    except ValueError:
        pass
for key, value in dict_parking_violations.items():
    print(key+'\t'+str(value))
```

Mapper 2: Mapper 2 will be taking the results from the previous reducer into the following two splits: street_codes and street_violation_info. "street_violation_info" is a string data type so we must use the function eval() which evaluates Python expressions from any input that comes as a string or a compiled code object. From there, we will call parts of the previously evaluated string to divide them and store it in "street_violation_info" again to update the street_code key's value.

```
import sys
for line in sys.stdin:
    street_codes, street_violation_info = line.strip().split('\t')
```

```
  try:
    street_violation_info = eval(street_violation_info)
    street_violation_info = street_violation_info[0]/street_violation_info[1]
    print(street_codes+'\t'+str(street_violation_info))
  except:
    continue
```

Reducer 2: This reducer's input is street_codes and the newer version of street_violation_info as a key and value, respectively. The purpose is to purely focus on the value now and transform it into a percentage value from illegally parked tickets out of the total tickets given in the three street code areas, which will be our final output for this exercise.

```
import sys
stored_values = 1
for line in sys.stdin:
  key, value = line.strip().split('\t')
  stored_values = stored_values * eval(value)
stored_values = round(stored_values* 100, 2)
print('Probability of black vehicle parking illegally at 34510, 10030, and 34050: '+
str(stored_values)+'%')
```

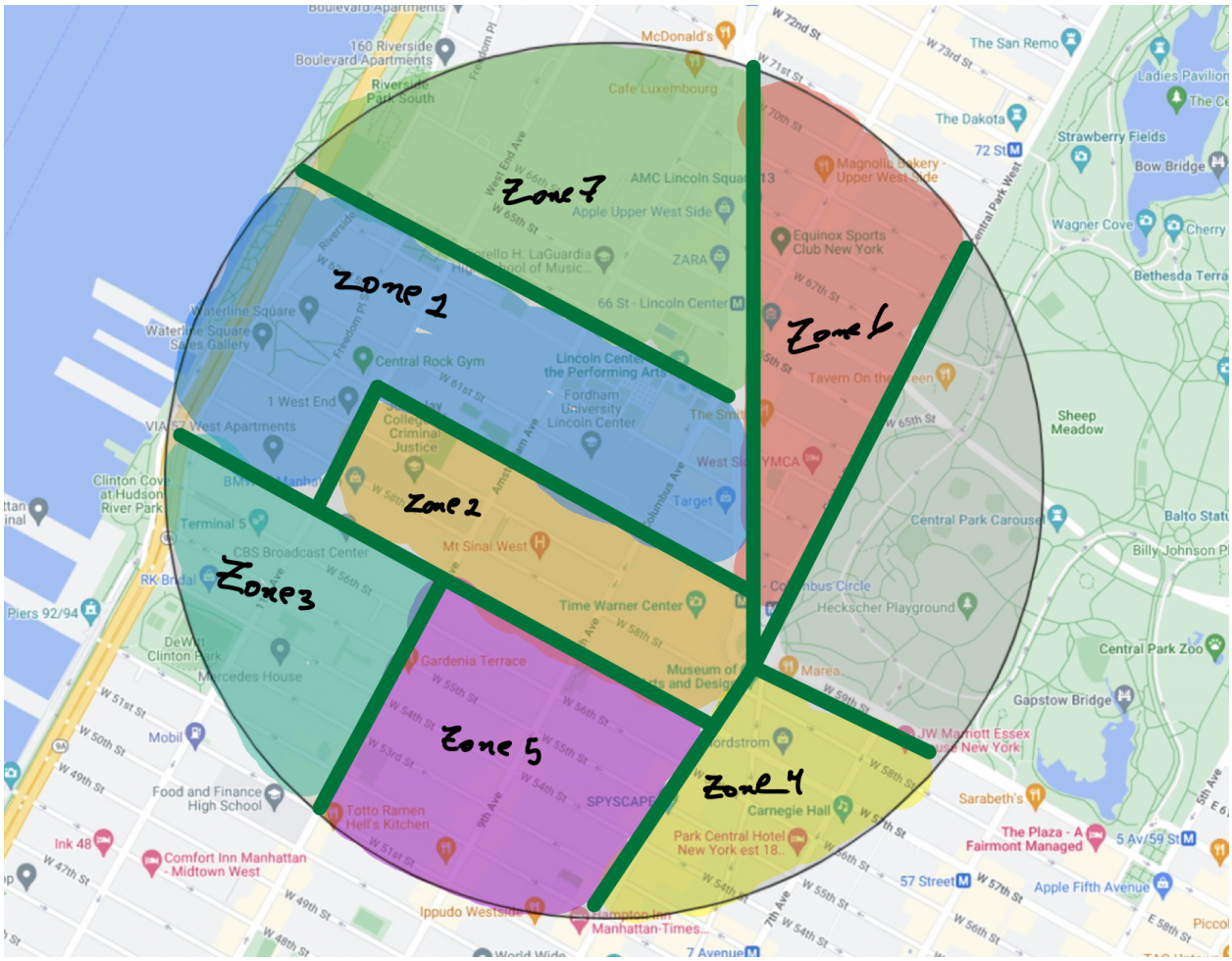
```
2022-04-06 06:24:50,610 INFO mapreduce.Job: Running job: job_1649226253333_0001
2022-04-06 06:25:01,878 INFO mapreduce.Job: Job job_1649226253333_0001 running in uber mode : false
2022-04-06 06:25:01,879 INFO mapreduce.Job:  map 0% reduce 0%
2022-04-06 06:25:42,328 INFO mapreduce.Job:  map 1% reduce 0%
2022-04-06 06:25:44,339 INFO mapreduce.Job:  map 8% reduce 0%
2022-04-06 06:25:49,372 INFO mapreduce.Job:  map 9% reduce 0%
2022-04-06 06:25:50,379 INFO mapreduce.Job:  map 17% reduce 0%
2022-04-06 06:25:51,386 INFO mapreduce.Job:  map 19% reduce 0%
2022-04-06 06:25:52,392 INFO mapreduce.Job:  map 23% reduce 0%
2022-04-06 06:25:55,411 INFO mapreduce.Job:  map 24% reduce 0%
2022-04-06 06:25:56,418 INFO mapreduce.Job:  map 34% reduce 0%
2022-04-06 06:25:57,446 INFO mapreduce.Job:  map 42% reduce 0%
2022-04-06 06:25:58,455 INFO mapreduce.Job:  map 55% reduce 0%
2022-04-06 06:25:59,461 INFO mapreduce.Job:  map 58% reduce 0%
2022-04-06 06:26:02,481 INFO mapreduce.Job:  map 59% reduce 0%
2022-04-06 06:26:03,486 INFO mapreduce.Job:  map 62% reduce 0%
2022-04-06 06:26:04,492 INFO mapreduce.Job:  map 67% reduce 0%
2022-04-06 06:26:08,517 INFO mapreduce.Job:  map 68% reduce 0%
2022-04-06 06:26:09,530 INFO mapreduce.Job:  map 72% reduce 0%
2022-04-06 06:26:10,538 INFO mapreduce.Job:  map 79% reduce 0%
2022-04-06 06:26:12,549 INFO mapreduce.Job:  map 85% reduce 0%
2022-04-06 06:26:13,555 INFO mapreduce.Job:  map 100% reduce 0%
2022-04-06 06:26:15,567 INFO mapreduce.Job:  map 100% reduce 100%
2022-04-06 06:26:15,583 INFO mapreduce.Job: Job job_1649226253333_0001 completed successfully
2022-04-06 06:26:15,689 INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=54696
                FILE: Number of bytes written=4258179
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1884594574
                HDFS: Number of bytes written=54
                HDFS: Number of read operations=47
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=2
                Launched map tasks=15
                Launched reduce tasks=1
                Data-local map tasks=14
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=873803
                Total time spent by all reduces in occupied slots (ms)=14686
                Total time spent by all map tasks (ms)=873803
                Total time spent by all reduce tasks (ms)=14686
                Total vcore-milliseconds taken by all map tasks=873803
                Total vcore-milliseconds taken by all reduce tasks=14686
                Total megabyte-milliseconds taken by all map tasks=894774272
                Total megabyte-milliseconds taken by all reduce tasks=15038464
        Map-Reduce Framework
                Map input records=9980450
                Map output records=5469
                Map output bytes=43752
                Map output materialized bytes=54774
                Input split bytes=1372
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
```

```
                Reduce input groups=3
                Reduce shuffle bytes=54774
                Reduce input records=5469
                Reduce output records=3
                Spilled Records=10938
                Shuffled Maps =14
                Failed Shuffles=0
                Merged Map outputs=14
                GC time elapsed (ms)=5430
                CPU time spent (ms)=98090
                Physical memory (bytes) snapshot=4404121600
                Virtual memory (bytes) snapshot=41842327552
                Total committed heap usage (bytes)=3312451584
                Peak Map Physical memory (bytes)=323321856
                Peak Map Virtual memory (bytes)=2819928064
                Peak Reduce Physical memory (bytes)=221556736
                Peak Reduce Virtual memory (bytes)=2799640576
        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=1884593202
        File Output Format Counters
                Bytes Written=54
2022-04-06 06:26:15,690 INFO streaming.StreamJob: Output directory: /Part1/output/
2022-04-06 06:26:17,227 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files ins
tead.
packageJobJar: [../../BonusQ/Part1/mapper2.py, ../../BonusQ/Part1/reducer2.py, /tmp/hadoop-unjar826317654866223996
9/] [] /tmp/streamjob1002892576395370835.jar tmpDir=null
2022-04-06 06:26:18,662 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /10.128.0
.5:8032
2022-04-06 06:26:18,909 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /10.128.0
.5:8032
2022-04-06 06:26:19,234 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/st
aging/root/.staging/job_1649226253333_0002
2022-04-06 06:26:19,696 INFO mapred.FileInputFormat: Total input files to process : 1
2022-04-06 06:26:19,805 INFO mapreduce.JobSubmitter: number of splits:2
2022-04-06 06:26:20,115 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1649226253333_0002
2022-04-06 06:26:20,115 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-04-06 06:26:20,350 INFO conf.Configuration: resource-types.xml not found
2022-04-06 06:26:20,351 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-04-06 06:26:20,470 INFO impl.YarnClientImpl: Submitted application application_1649226253333_0002
2022-04-06 06:26:20,527 INFO mapreduce.Job: The url to track the job: http://instance-0:8088/proxy/application_164
9226253333_0002/
2022-04-06 06:26:20,530 INFO mapreduce.Job: Running job: job_1649226253333_0002
2022-04-06 06:26:30,709 INFO mapreduce.Job: Job job_1649226253333_0002 running in uber mode : false
2022-04-06 06:26:30,710 INFO mapreduce.Job:  map 0% reduce 0%
2022-04-06 06:26:40,822 INFO mapreduce.Job:  map 100% reduce 0%
2022-04-06 06:26:47,891 INFO mapreduce.Job:  map 100% reduce 100%
2022-04-06 06:26:47,904 INFO mapreduce.Job: Job job_1649226253333_0002 completed successfully
2022-04-06 06:26:48,043 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=75
                FILE: Number of bytes written=829853
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
```

```
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=16060
                Total time spent by all reduces in occupied slots (ms)=4121
                Total time spent by all map tasks (ms)=16060
                Total time spent by all reduce tasks (ms)=4121
                Total vcore-milliseconds taken by all map tasks=16060
                Total vcore-milliseconds taken by all reduce tasks=4121
                Total megabyte-milliseconds taken by all map tasks=16445440
                Total megabyte-milliseconds taken by all reduce tasks=4219904
        Map-Reduce Framework
                Map input records=3
                Map output records=3
                Map output bytes=63
                Map output materialized bytes=81
                Input split bytes=196
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=81
                Reduce input records=3
                Reduce output records=1
                Spilled Records=6
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=210
                CPU time spent (ms)=2880
                Physical memory (bytes) snapshot=780419072
                Virtual memory (bytes) snapshot=8370348032
                Total committed heap usage (bytes)=581959680
                Peak Map Physical memory (bytes)=292589568
                Peak Map Virtual memory (bytes)=2788462592
                Peak Reduce Physical memory (bytes)=211656704
                Peak Reduce Virtual memory (bytes)=2794471424
        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=81
        File Output Format Counters
                Bytes Written=83
2022-04-06 06:26:48,050 INFO streaming.StreamJob: Output directory: /Part1-2/output/
Probability of black vehicle parking illegally at 34510, 10030, and 34050: 19.11%
Deleted /Part1/input
Deleted /Part1/output
Deleted /Part1-2/output
Stopping namenodes on [instance-0.c.big-data-339500.internal]
Stopping datanodes
Stopping secondary namenodes [instance-0]
Stopping nodemanagers
10.128.0.3: WARNING: nodemanager did not stop gracefully after 5 seconds: Trying to kill with kill -9
10.128.0.4: WARNING: nodemanager did not stop gracefully after 5 seconds: Trying to kill with kill -9
Stopping resourcemanager
WARNING: Use of this script to stop the MR JobHistory daemon is deprecated.
WARNING: Attempting to execute replacement "mapred --daemon stop" instead.
root@instance-0:/BigData_Project/BonusQ/Part1#
```

Bonus Question 2:

At 10 am, I want to go to Lincoln Center and I just want to walk within 0.5 mile. Where should I park? (Divided into zones).



Conclusion: At 10 am, we should park at zone 1: 48.2%

Mapper 1: We will be filtering data within the W51 street (southest point) up to W71 street (northest point) based on a 0.5 radius, and using the precinct 20. Then we create a random cluster and then calculate the euclidean distance of each data point from each of the clusters and assign it the key of the cluster with minimum distance. By using K -means clustering, we are finding 7 clusters for the matrix {STREET CODE 1, STREET CODE 2, STREET CODE 3}. We defined 7 clusters as it converged faster and are representative zones with a high density transit.

```
import sys
zones_mapper = sys.argv[1]
zones_mapper = [eval(dp) for dp in zones_mapper.split('Z')[1].strip('_').split('_')]
zones_mapper = {1:[zones_mapper[0],zones_mapper[1],zones_mapper[2]],
                2:[zones_mapper[3],zones_mapper[4],zones_mapper[5]],
                3:[zones_mapper[6],zones_mapper[7],zones_mapper[8]],
                4:[zones_mapper[9],zones_mapper[10],zones_mapper[11]]}
def euclidean_distance(A, B):
 return sum((a-b)**2 for a, b in zip(A[:], B[:])) ** (1/2)
mapper_1_output = {1:[0,[0,0,0]],
                   2:[0,[0,0,0]],
                   3:[0,[0,0,0]],
                   4:[0,[0,0,0]]}
for line in sys.stdin:
 line = line.strip(',').split(',')
 line_len = len(line)
 if line_len == 23:
   try:
     SHOT_DIST = float(line[12].strip(""))
     CLOSE_DEF_DIST= float(line[-5].strip(""))
     SHOT_CLOCK = float(line[9].strip(""))
     if SHOT_DIST < 29 and CLOSE_DEF_DIST < 11 and SHOT_CLOCK < 26: # removing
outliers
       data = [SHOT_DIST, CLOSE_DEF_DIST, SHOT_CLOCK]
       data_centroids_distances = {1: euclidean_distance(data, zones_mapper[1]),
                                   2: euclidean_distance(data, zones_mapper[2]),
                                   3: euclidean_distance(data, zones_mapper[3]),
                                   4: euclidean_distance(data, zones_mapper[4])}
       data_cluster_key = min(data_centroids_distances, key = data_centroids_distances.get)
#argmin
       mapper_1_output[data_cluster_key][0] += 1 # counter
```

```
        mapper_1_output[data_cluster_key][1][0] += data[0] #sum all SHOT_DIST
        mapper_1_output[data_cluster_key][1][1] += data[1] #sum all CLOSE_DEF_DIST
        mapper_1_output[data_cluster_key][1][2] += data[2] #sum all SHOT_CLOCK

    except:
      continue

combiner_1_input = mapper_1_output
for key, values in combiner_1_input.items():
 print('{key}\t{values}'.format(key=key, values=values))
```

Reducer 1: Finding the centroid of each cluster to update the new centroid in the next for loop which is implemented in the test.sh file.

The for loop inside the **test.sh** will stop when it finds that the centroids coming from Reducer 1 converge.

```
import sys
reducer_1_output = {}
for line in sys.stdin:
   key, values = line.split('\t')
   values = eval(values)
   count = values[0]
   sum_SHOT_DIST = values[1][0]
   sum_CLOSE_DEF_DIST = values[1][1]
   sum_SHOT_CLOCK = values[1][2]
      reducer_1_output[int(key)] = [sum_SHOT_DIST/count, sum_CLOSE_DEF_DIST/count, sum_SHOT_CLOCK/count]
output = ''
for key, values in reducer_1_output.items():
```

```
 for value in values:
   output = output + '_' +str(value)
output = 'ClusterZ' + output[1:]
print(output)
```

Mapper 2 : Outside the for loop to get the centroids, Mapper 2 filters the dataset just for the required street codes, calculating the euclidean distance of each data point from each of the 7 clusters (zones) and assigning it the key of the cluster with minimum distance. Then we filter our results by the parking time (specifically any time at 10AM) being our probability space and considering as a parking violation any appearance of the word stand or the fragment “park” the violation description field within those zones.

```
import sys

zones_mapper = sys.argv[1]
zones_mapper = [eval(dp) for dp in zones_mapper.split('Z')[1].strip('_').split('_')]
zones_mapper = {1:[zones_mapper[0],zones_mapper[1],zones_mapper[2]],
            2:[zones_mapper[3],zones_mapper[4],zones_mapper[5]],
            3:[zones_mapper[6],zones_mapper[7],zones_mapper[8]],
            4:[zones_mapper[9],zones_mapper[10],zones_mapper[11]]}
def euclidean_distance(A, B):
   return sum((a-b)**2 for a, b in zip(A[:], B[:])) ** (1/2)
players = ['stephen curry', 'james harden', 'chris paul','lebron james']

for line in sys.stdin:
   line = line.strip(',').split(',')
```

```
    line_len = len(line)
    player = line[-2]
    if line_len == 23 and player in players:
        try:
            player = player.split(' ')
            player = player[0]+player[1]
            shot = 1 if line[14] == 'made' else 0
            SHOT_DIST = float(line[12].strip(""))
            CLOSE_DEF_DIST = float(line[-5].strip(""))
            SHOT_CLOCK = float(line[9].strip(""))
            data = [SHOT_DIST, CLOSE_DEF_DIST, SHOT_CLOCK]
            data_centroids_distances = {1: euclidean_distance(data, zones_mapper[1]),
                                        2: euclidean_distance(data, zones_mapper[2]),
                                        3: euclidean_distance(data, zones_mapper[3]),
                                        4: euclidean_distance(data, zones_mapper[4])}
                data_cluster_key = min(data_centroids_distances, key = data_centroids_distances.get)
#argmin
            print(player +"_"+str(data_cluster_key)+"\t"+str(shot))
        except:
            continue
```

Reducer 2: Converting our parking ticket information into a grouped information by clusters/zones as it groups all “parking tickets” and “not parking tickets” into a structure of [“sum of parking ticket” / (“sum of parking tickets” + “not parking ticket”)].

```
import sys

dict_score_count = {}

for line in sys.stdin:

  record = line.split('\t')

  data,count = record[0], record[1]

  try:

    count = int(count)

                dict_score_count[data] = [dict_score_count.get(data, [0,0])[0] + count,
dict_score_count.get(data, [0,0])[1]+1]

  except ValueError:

    pass

for key, value in dict_score_count.items():

  print(key+'\t'+str(value)
```

Mapper 3: Turns our grouped information by Reducer 2 into probability information dividing the sum of parking tickets by the total number of tickets issued in each zone/cluster.

```
import sys

for entry in sys.stdin:
```

```
    player_cluster, result_shots = entry.split('\t')
    try:
        result_shots = eval(result_shots)
        if float(result_shots[1]) == 1 and float(result_shots[0]) == 0:
            continue
        else:
            hit_rate = float(result_shots[0])/float(result_shots[1])
    except:
        pass
    print(player_cluster+'\t'+str(hit_rate))
```

Reducer 3: Extract the lowest probability among the defined zones. As the final output of it, we can see that it will be the lowest probability to get a parking ticket at a 0.5 radius of Lincoln Center or the best place to park around 10AM.

```
import sys
reducer_3_output = {}
for entry in sys.stdin:
    player_cluster, hit_rate = entry.split('\t')
    player, cluster = player_cluster.split('_')
    hit_rate = float(hit_rate)
    if player not in reducer_3_output:
        reducer_3_output[player] = [cluster, hit_rate]
    elif hit_rate > float(reducer_3_output[player][1]):
        reducer_3_output[player] = [cluster, hit_rate]
    else:
```

```
        continue
for key, value in reducer_3_output.items():
    print(key+'\t'+'Cluster: '+str(value[0])+' | Hit Rate: '+str(value[1]))
```

Test. sh :

```
starting_zones=ClusterZ6_3_26_15_8_1_2_5_10_26_4_5
zones="0"
new_zones=""
for i in {0..50}; do
  if [[ $zones == $new_zones ]]; then
     break
  elif [[ $new_zones != "" ]]; then
     zones=$new_zones
  else
     zones=$starting_zones
  fi
                    new_zones=`cat /BigData_Project/test-data\/shot_logs.csv | python3 /BigData_Project/part2/Part2/mapper1.py "$zones" | python3 /BigData_Project/part2/Part2/reducer1.py`
done
../../start.sh
/usr/local/hadoop/bin/hdfs dfs -rm -r /Part2/input/
/usr/local/hadoop/bin/hdfs dfs -rm -r /Part2/output/
/usr/local/hadoop/bin/hdfs dfs -mkdir -p /Part2/input/
/usr/local/hadoop/bin/hdfs dfs -copyFromLocal ../../test-data/shot_logs.csv /Part2/input/
/usr/local/hadoop/bin/hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar \
-file ../../part2/Part2/mapper2.py -mapper "../../part2/Part2/mapper2.py $zones" \
-file ../../part2/Part2/reducer2.py -reducer ../../part2/Part2/reducer2.py \
-input /Part2/input/* -output /Part2/output/
```

**/usr/local/hadoop/bin/hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.1.jar **

**-file ../../part2/Part2/mapper3.py -mapper ../../part2/Part2/mapper3.py **

**-file ../../part2/Part2/reducer3.py -reducer ../../part2/Part2/reducer3.py **

-input /Part2/output/* -output /Part2-2/output/

/usr/local/hadoop/bin/hdfs dfs -cat /Part2-2/output/part-00000

/usr/local/hadoop/bin/hdfs dfs -rm -r /Part2/input/

- **/usr/local/hadoop/bin/hdfs dfs -rm -r /Part2/output/**
- **/usr/local/hadoop/bin/hdfs dfs -rm -r /Part2-2/output/**
- **../../stop.sh**

```
root@instance-0:/BigData_Project/BonusQ/Part2# bash test.sh
Starting namenodes on [instance-0.c.big-data-339500.internal]
Starting datanodes
Starting secondary namenodes [instance-0]
Starting resourcemanager
Starting nodemanagers
WARNING: Use of this script to start the MR JobHistory daemon is deprecated.
WARNING: Attempting to execute replacement "mapred --daemon start" instead.
Safe mode is OFF
rm: `/Part2/input/': No such file or directory
rm: `/Part2/output/': No such file or directory
2022-04-06 07:01:27,939 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [../../BonusQ/Part2/mapper2.py, ../../BonusQ/Part2/reducer2.py, /tmp/hadoop-unjar8623618970463185421/] [] /tmp/stre
amjob1971367323916670754.jar tmpDir=null
2022-04-06 07:01:29,365 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /10.128.0.5:8032
2022-04-06 07:01:29,598 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /10.128.0.5:8032
2022-04-06 07:01:29,943 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.sta
ing/job_1649228454833_0001
2022-04-06 07:01:30,450 INFO mapred.FileInputFormat: Total input files to process : 1
2022-04-06 07:01:30,496 INFO net.NetworkTopology: Adding a new node: /default-rack/10.128.0.3:9866
2022-04-06 07:01:30,497 INFO net.NetworkTopology: Adding a new node: /default-rack/10.128.0.4:9866
2022-04-06 07:01:30,629 INFO mapreduce.JobSubmitter: number of splits:14
2022-04-06 07:01:31,039 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1649228454833_0001
2022-04-06 07:01:31,039 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-04-06 07:01:31,285 INFO conf.Configuration: resource-types.xml not found
2022-04-06 07:01:31,286 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-04-06 07:01:31,599 INFO impl.YarnClientImpl: Submitted application application_1649228454833_0001
2022-04-06 07:01:31,751 INFO mapreduce.Job: The url to track the job: http://instance-0:8088/proxy/application_1649228454833_0001/
2022-04-06 07:01:31,769 INFO mapreduce.Job: Running job: job_1649228454833_0001
2022-04-06 07:01:43,081 INFO mapreduce.Job: Job job_1649228454833_0001 running in uber mode : false
2022-04-06 07:01:43,082 INFO mapreduce.Job:  map 0% reduce 0%
2022-04-06 07:02:20,414 INFO mapreduce.Job:  map 2% reduce 0%
2022-04-06 07:02:21,425 INFO mapreduce.Job:  map 4% reduce 0%
2022-04-06 07:02:26,460 INFO mapreduce.Job:  map 5% reduce 0%
2022-04-06 07:02:27,466 INFO mapreduce.Job:  map 6% reduce 0%
2022-04-06 07:02:29,479 INFO mapreduce.Job:  map 9% reduce 0%
2022-04-06 07:02:32,499 INFO mapreduce.Job:  map 10% reduce 0%
2022-04-06 07:02:33,505 INFO mapreduce.Job:  map 12% reduce 0%
2022-04-06 07:02:35,523 INFO mapreduce.Job:  map 14% reduce 0%
2022-04-06 07:02:36,528 INFO mapreduce.Job:  map 15% reduce 0%
2022-04-06 07:02:39,559 INFO mapreduce.Job:  map 18% reduce 0%
2022-04-06 07:02:41,570 INFO mapreduce.Job:  map 20% reduce 0%
2022-04-06 07:02:42,582 INFO mapreduce.Job:  map 21% reduce 0%
2022-04-06 07:02:45,611 INFO mapreduce.Job:  map 24% reduce 0%
2022-04-06 07:02:47,624 INFO mapreduce.Job:  map 26% reduce 0%
2022-04-06 07:02:48,630 INFO mapreduce.Job:  map 27% reduce 0%
2022-04-06 07:02:50,642 INFO mapreduce.Job:  map 28% reduce 0%
2022-04-06 07:02:51,648 INFO mapreduce.Job:  map 31% reduce 0%
2022-04-06 07:02:53,664 INFO mapreduce.Job:  map 32% reduce 0%
2022-04-06 07:02:54,669 INFO mapreduce.Job:  map 34% reduce 0%
2022-04-06 07:02:57,686 INFO mapreduce.Job:  map 37% reduce 0%
2022-04-06 07:02:59,698 INFO mapreduce.Job:  map 39% reduce 0%
2022-04-06 07:03:00,708 INFO mapreduce.Job:  map 40% reduce 0%
2022-04-06 07:03:02,722 INFO mapreduce.Job:  map 41% reduce 0%
2022-04-06 07:03:03,728 INFO mapreduce.Job:  map 43% reduce 0%
2022-04-06 07:03:05,755 INFO mapreduce.Job:  map 48% reduce 0%
2022-04-06 07:03:06,760 INFO mapreduce.Job:  map 49% reduce 0%
2022-04-06 07:03:07,812 INFO mapreduce.Job:  map 55% reduce 0%
2022-04-06 07:03:08,818 INFO mapreduce.Job:  map 56% reduce 0%
2022-04-06 07:03:09,824 INFO mapreduce.Job:  map 59% reduce 0%
2022-04-06 07:03:10,829 INFO mapreduce.Job:  map 61% reduce 0%
2022-04-06 07:03:11,836 INFO mapreduce.Job:  map 66% reduce 0%
2022-04-06 07:03:12,843 INFO mapreduce.Job:  map 67% reduce 0%
2022-04-06 07:03:17,872 INFO mapreduce.Job:  map 69% reduce 0%
2022-04-06 07:03:18,880 INFO mapreduce.Job:  map 71% reduce 0%
2022-04-06 07:03:24,920 INFO mapreduce.Job:  map 74% reduce 0%
2022-04-06 07:03:28,944 INFO mapreduce.Job:  map 74% reduce 14%
2022-04-06 07:03:29,952 INFO mapreduce.Job:  map 75% reduce 14%
2022-04-06 07:03:30,957 INFO mapreduce.Job:  map 77% reduce 14%
2022-04-06 07:03:31,965 INFO mapreduce.Job:  map 80% reduce 14%
2022-04-06 07:03:33,975 INFO mapreduce.Job:  map 83% reduce 14%
2022-04-06 07:03:34,981 INFO mapreduce.Job:  map 86% reduce 17%
2022-04-06 07:03:35,986 INFO mapreduce.Job:  map 92% reduce 17%
2022-04-06 07:03:36,992 INFO mapreduce.Job:  map 97% reduce 17%
2022-04-06 07:03:39,000 INFO mapreduce.Job:  map 100% reduce 17%
2022-04-06 07:03:40,006 INFO mapreduce.Job:  map 100% reduce 100%
2022-04-06 07:03:40,014 INFO mapreduce.Job: Job job_1649228454833_0001 completed successfully
2022-04-06 07:03:40,152 INFO mapreduce.Job: Counters: 56
```

```
2022-04-06 07:03:40,152 INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=14808
                FILE: Number of bytes written=4184088
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1884594574
                HDFS: Number of bytes written=15
                HDFS: Number of read operations=47
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=2
                Launched map tasks=16
                Launched reduce tasks=1
                Data-local map tasks=13
                Rack-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=1415681
                Total time spent by all reduces in occupied slots (ms)=31410
                Total time spent by all map tasks (ms)=1415681
                Total time spent by all reduce tasks (ms)=31410
                Total vcore-milliseconds taken by all map tasks=1415681
                Total vcore-milliseconds taken by all reduce tasks=31410
                Total megabyte-milliseconds taken by all map tasks=1449657344
                Total megabyte-milliseconds taken by all reduce tasks=32163840
        Map-Reduce Framework
                Map input records=9980450
                Map output records=2467
                Map output bytes=9868
                Map output materialized bytes=14886
                Input split bytes=1372
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=14886
                Reduce input records=2467
                Reduce output records=1
                Spilled Records=4934
                Shuffled Maps =14
                Failed Shuffles=0
                Merged Map outputs=14
                GC time elapsed (ms)=4069
                CPU time spent (ms)=244320
                Physical memory (bytes) snapshot=4513222656
                Virtual memory (bytes) snapshot=41833009152
                Total committed heap usage (bytes)=3304062976
                Peak Map Physical memory (bytes)=333037568
                Peak Map Virtual memory (bytes)=2816352256
                Peak Reduce Physical memory (bytes)=215097344
                Peak Reduce Virtual memory (bytes)=2795261952
        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=1884593202
        File Output Format Counters
                Bytes Written=15
2022-04-06 07:03:40,155 INFO streaming.StreamJob: Output directory: /Part2/output/
2022-04-06 07:03:41,728 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [../../BonusQ/Part2/mapper3.py, ../../BonusQ/Part2/reducer3.py, /tmp/hadoop-unjar1046751970086926208l/] [] /tmp/str
eamjob7169607482531423305.jar tmpDir=null
2022-04-06 07:03:43,179 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /10.128.0.5:8032
2022-04-06 07:03:43,413 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /10.128.0.5:8032
2022-04-06 07:03:43,895 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.stag
ing/job_1649228454833_0002
2022-04-06 07:03:44,860 INFO mapred.FileInputFormat: Total input files to process : 1
2022-04-06 07:03:44,976 INFO mapreduce.JobSubmitter: number of splits:3
2022-04-06 07:03:45,455 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1649228454833_0002
2022-04-06 07:03:45,455 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-04-06 07:03:45,705 INFO conf.Configuration: resource-types.xml not found
2022-04-06 07:03:45,705 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-04-06 07:03:45,778 INFO impl.YarnClientImpl: Submitted application application_1649228454833_0002
2022-04-06 07:03:45,872 INFO mapreduce.Job: The url to track the job: http://instance-0:8088/proxy/application_1649228454833_0002/
2022-04-06 07:03:45,874 INFO mapreduce.Job: Running job: job_1649228454833_0002
2022-04-06 07:03:55,044 INFO mapreduce.Job: Job job_1649228454833_0002 running in uber mode : false
```

```
2022-04-06 07:03:55,045 INFO mapreduce.Job:  map 0% reduce 0%
2022-04-06 07:04:08,177 INFO mapreduce.Job:  map 33% reduce 0%
2022-04-06 07:04:09,185 INFO mapreduce.Job:  map 100% reduce 0%
2022-04-06 07:04:15,234 INFO mapreduce.Job:  map 100% reduce 100%
2022-04-06 07:04:15,249 INFO mapreduce.Job: Job job_1649228454833_0002 completed successfully
2022-04-06 07:04:15,364 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=25
                FILE: Number of bytes written=1106343
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=318
                HDFS: Number of bytes written=41
                HDFS: Number of read operations=14
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=34844
                Total time spent by all reduces in occupied slots (ms)=3990
                Total time spent by all map tasks (ms)=34844
                Total time spent by all reduce tasks (ms)=3990
                Total vcore-milliseconds taken by all map tasks=34844
                Total vcore-milliseconds taken by all reduce tasks=3990
                Total megabyte-milliseconds taken by all map tasks=35680256
                Total megabyte-milliseconds taken by all reduce tasks=4085760
        Map-Reduce Framework
                Map input records=1
                Map output records=1
                Map output bytes=17
                Map output materialized bytes=37
                Input split bytes=294
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=37
                Reduce input records=1
                Reduce output records=1
                Spilled Records=2
                Shuffled Maps =3
                Failed Shuffles=0
                Merged Map outputs=3
                GC time elapsed (ms)=328
                CPU time spent (ms)=3610
                Physical memory (bytes) snapshot=1093591040
                Virtual memory (bytes) snapshot=11155910656
                Total committed heap usage (bytes)=914358272
                Peak Map Physical memory (bytes)=301768704
                Peak Map Virtual memory (bytes)=2793541632
                Peak Reduce Physical memory (bytes)=207511552
                Peak Reduce Virtual memory (bytes)=2791514112
        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=24
        File Output Format Counters
                Bytes Written=41
2022-04-06 07:04:15,370 INFO streaming.StreamJob: Output directory: /Part2-2/output/
At 10 am, we should park at zone 1      48.2%
Deleted /Part2/input
Deleted /Part2/output
Deleted /Part2-2/output
Stopping namenodes on [instance-0.c.big-data-339500.internal]
Stopping datanodes
Stopping secondary namenodes [instance-0]
Stopping nodemanagers
10.128.0.3: WARNING: nodemanager did not stop gracefully after 5 seconds: Trying to kill with kill -9
10.128.0.4: WARNING: nodemanager did not stop gracefully after 5 seconds: Trying to kill with kill -9
Stopping resourcemanager
WARNING: Use of this script to stop the MR JobHistory daemon is deprecated.
WARNING: Attempting to execute replacement "mapred --daemon stop" instead.
root@instance-0:/BigData_Project/BonusQ/Part2#
```