

Task 4. Write MapReduce codes to perform the tasks using the files you've downloaded on your EMR Instance:

We downloaded all the 6 csv files into our EMR cluster and also pushed the respective python files for the analytical queries below.

```
[hadoop@ip-172-31-1-58 ~]$ ls
mrtask_a.py mrtask_c.py mrtask_e.py yellow_tripdata_2017-01.csv yellow_tripdata_2017-03.csv yellow_tripdata_2017-05.csv
mrtask_b.py mrtask_d.py mrtask_f.py yellow_tripdata_2017-02.csv yellow_tripdata_2017-04.csv yellow_tripdata_2017-06.csv
```

Then we created a directory for storing these csv files for which we ran the command **hdfs dfs -mkdir -p /user/Hadoop/assignment/input** and moved all the csv files into this directory by running the command **hdfs dfs -put /home/hadoop/*.csv /user/Hadoop/assignment/input**.

```
[hadoop@ip-172-31-1-58 ~]$ hdfs dfs -mkdir -p /user/hadoop/assignment/input
[hadoop@ip-172-31-1-58 ~]$ ls
mrtask_a.py mrtask_c.py mrtask_e.py yellow_tripdata_2017-01.csv yellow_tripdata_2017-03.csv yellow_tripdata_2017-05.csv
mrtask_b.py mrtask_d.py mrtask_f.py yellow_tripdata_2017-02.csv yellow_tripdata_2017-04.csv yellow_tripdata_2017-06.csv
[hadoop@ip-172-31-1-58 ~]$ pwd
/home/hadoop
[hadoop@ip-172-31-1-58 ~]$ hdfs dfs -put /home/hadoop/*.csv /user/hadoop/assignment/input
```

Question 1: Which vendors have the most trips, and what is the total revenue generated by that vendor?

We used the command **python mrtask_a.py -r Hadoop hdfs:///user/Hadoop/assignment/input/*.csv > output_a.txt** for running the code on all the csv files and then we used the command **cat output_a.txt** for seeing the content i.e. the solution of the query.

```
[hadoop@ip-172-31-1-58 ~]$ python mrtask_a.py -r hadoop hdfs:///user/hadoop/assignment/input/*.csv > output_a.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.8.5
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/mrtask_a.hadoop.20240408.101626.084502
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/mrtask_a.hadoop.20240408.101626.084502/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/mrtask_a.hadoop.20240408.101626.084502/files/
Running step 1 of 2...
packageJobJar: [ [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-6.jar] /tmp/streamjob8367823691772134333.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-1-58.ec2.internal/172.31.1.58:8032
Connecting to ResourceManager at ip-172-31-1-58.ec2.internal/172.31.1.58:8032
Loaded native gpl library
Successfully loaded & initialized native-lzo library [hadoop-lzo rev ff8f5709577defb6b78cdc1f98cfe129c4b6fe46]
Total input files to process : 6
Adding a new node: /default-rack/172.31.1.58:50010
number of splits:43
Submitting tokens for job: job_1712568903374_0002
Submitted application application_1712568903374_0002
The url to track the job: http://ip-172-31-1-58.ec2.internal:20888/proxy/application_1712568903374_0002/
Running job: job_1712568903374_0002
```

```
[hadoop@ip-172-31-1-58 ~]$ cat output_a.txt
"2" 525037658.13640213
```

Answer: Vendor 2 (VeriFone Inc.) has the most number of trips and the total revenue generated by them is 525037658.13640213.

Question 2: Which pickup location generates the most revenue?

We used the command **python mrtask_b.py -r Hadoop**

hdfs:///user/Hadoop/assignment/input/*.csv > output_b.txt for running the code on all the csv files and then we used the command **cat output_b.txt** for seeing the content i.e. the solution of the query.

```
[hadoop@ip-172-31-1-58 ~]$ python mrtask_b.py -r hadoop hdfs:///user/hadoop/assignment/input/*.csv > output_b.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.8.5
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/mrtask_b.hadoop.20240408.103440.515710
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/mrtask_b.hadoop.20240408.103440.515710/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/mrtask_b.hadoop.20240408.103440.515710/files/
Running step 1 of 2...
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-6.jar] /tmp/streamjob2841570728038678982.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-1-58.ec2.internal/172.31.1.58:8032
Connecting to ResourceManager at ip-172-31-1-58.ec2.internal/172.31.1.58:8032
Loaded native gpl library
Successfully loaded & initialized native-lzo library [hadoop-lzo rev ff8f5709577defb6b78cdc1f98cfe129c4b6fe46]
Total input files to process : 6
Adding a new node: /default-rack/172.31.1.58:50010
number of splits:43
Submitting tokens for job: job_1712568903374_0004
Submitted application application_1712568903374_0004
The url to track the job: http://ip-172-31-1-58.ec2.internal:20888/proxy/application_1712568903374_0004/
Running job: job_1712568903374_0004
Job job_1712568903374_0004 running in uber mode : false
  map 0% reduce 0%
  map 1% reduce 0%
  map 2% reduce 0%
  map 3% reduce 0%
  map 4% reduce 0%
  map 5% reduce 0%
```

```
[hadoop@ip-172-31-1-58 ~]$ cat output_b.txt
"132" 77196812.23975265
```

Answer: Pickup location 132 has generated the most revenue 77196812.23975265.

Question 3: What are the different payment types used by customers and their count? The final results should be in a sorted format.

We used the command **python mrtask_c.py -r Hadoop**

hdfs:///user/Hadoop/assignment/input/*.csv > output_c.txt for running the code on all the csv files and then we used the command **cat output_c.txt** for seeing the content i.e. the solution of the query.

```
[hadoop@ip-172-31-1-58 ~]$ python mrtask_c.py -r hadoop hdfs:///user/hadoop/assignment/input/*.csv > output_c.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.8.5
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/mrtask_c.hadoop.20240408.105114.957772
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/mrtask_c.hadoop.20240408.105114.957772/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/mrtask_c.hadoop.20240408.105114.957772/files/
Running step 1 of 3...
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-6.jar] /tmp/streamjob6035324153864795024.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-1-58.ec2.internal/172.31.1.58:8032
Connecting to ResourceManager at ip-172-31-1-58.ec2.internal/172.31.1.58:8032
Loaded native gpl library
Successfully loaded & initialized native-lzo library [hadoop-lzo rev ff8f5709577defb6b78cdc1f98cfe129c4b6fe46]
Total input files to process : 6
Adding a new node: /default-rack/172.31.1.58:50010
number of splits:43
Submitting tokens for job: job_1712568903374_0006
Submitted application application_1712568903374_0006
The url to track the job: http://ip-172-31-1-58.ec2.internal:20888/proxy/application_1712568903374_0006/
Running job: job_1712568903374_0006
Job job_1712568903374_0006 running in uber mode : false
  map 0% reduce 0%
  map 1% reduce 0%
  map 2% reduce 0%
  map 3% reduce 0%
  map 4% reduce 0%
  map 5% reduce 0%
  map 6% reduce 0%
  map 7% reduce 0%
  map 8% reduce 0%
```

```
[hadoop@ip-172-31-1-58 ~]$ cat output_c.txt
"1"      39754212
"2"      18832370
"3"      306912
"4"      88794
"5"      3
```

Answer: Different payment types used by customers and their counts are as follows:

- 1 -> Credit card -> 39754212
- 2 -> Cash -> 18832370
- 3 -> No charge -> 306912
- 4 -> Dispute -> 88794
- 5 -> Unknown -> 3

Question 4: What is the average trip time for different pickup locations?

We used the command **python mrtask_d.py -r Hadoop**


hdfs:///user/Hadoop/assignment/input/*.csv > output_d.txt for running the code on all the csv files and then we used the command **cat output_d.txt** for seeing the content i.e. the solution of the query.

```
[hadoop@ip-172-31-1-58 ~]$ python mrtask_d.py -r hadoop hdfs:///user/hadoop/assignment/input/*.csv > output_d.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.8.5
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/mrtask_d.hadoop.20240408.110712.544314
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/mrtask_d.hadoop.20240408.110712.544314/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/mrtask_d.hadoop.20240408.110712.544314/files/
Running step 1 of 1...
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-6.jar] /tmp/streamjob707217505596945385.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-1-58.ec2.internal/172.31.1.58:8032
Connecting to ResourceManager at ip-172-31-1-58.ec2.internal/172.31.1.58:8032
Loaded native gpl library
Successfully loaded & initialized native-lzo library [hadoop-lzo rev ff8f5709577defb6b78cdc1f98cfe129c4b6fe46]
Total input files to process : 6
Adding a new node: /default-rack/172.31.1.58:50010
number of splits:43
Submitting tokens for job: job_1712568903374_0009
Submitted application application_1712568903374_0009
The url to track the job: http://ip-172-31-1-58.ec2.internal:20888/proxy/application_1712568903374_0009/
Running job: job_1712568903374_0009
Job job_1712568903374_0009 running in uber mode : false
  map 0% reduce 0%
  map 1% reduce 0%
  map 2% reduce 0%
```


```
[hadoop@ip-172-31-1-58 ~]$ cat output_d.txt
```

```
"102" 22.112369791666666  
"105" 19.977777777777778  
"108" 14.208414239482202  
"111" 11.63465783664459  
"114" 15.921296462978038  
"117" 19.317635658914725  
"12" 24.358713440236155  
"120" 13.8081208687441  
"123" 15.514665127020782  
"126" 18.56177873152122  
"129" 14.258696642003411  
"132" 43.7702887196418  
"135" 18.103793774319065  
"138" 37.32248703573805  
"141" 12.262026243285483  
"144" 16.80377873996948  
"147" 13.154071246819338  
"15" 14.54189602446483  
"150" 18.492307692307694  
"153" 13.562989801395593  
"156" 19.404368932038835  
"159" 14.204843982169384  
"162" 15.103685552025429  
"165" 18.458268590455052  
"168" 12.855600663523658  
"171" 12.873132183908044  
"174" 13.067625899280578  
"177" 19.16066056245912  
"18" 14.326474820143885  
"180" 30.037775246772977  
"183" 12.080208333333333  
"186" 16.811887227994642  
"189" 15.16632127050247  
"192" 18.249265905383357  
"195" 20.830252339098386  
"198" 13.249704098239382  
"201" 9.882394366197184  
"204" 3.5511904761904773  
"207" 8.466067108446406  
"21" 20.498177083333333
```


"210"	17.948028169014083
"213"	16.38755186721992
"216"	28.418558295765934
"219"	45.452684525827955
"222"	29.71230158730159
"225"	14.926513473606494
"228"	15.833887297039157
"231"	16.844082243497855
"234"	15.002941362166112
"237"	12.365686895410613
"24"	13.555459891591278
"240"	14.076029962546821
"243"	16.346816612643945
"246"	15.718104183838525
"249"	14.940339887463704
"252"	19.827374999999996
"255"	16.32139020236068
"258"	20.483721431651574
"261"	21.84984265746389
"264"	14.775037341482173
"27"	8.77878787878788
"3"	21.66262626262626
"30"	25.10158730158731
"33"	18.095158681576805
"36"	16.87346516927083
"39"	18.234363575938854
"42"	12.959889655721879
"45"	18.46525038820368
"48"	14.690364201807599
"51"	16.277023809523808
"54"	17.032464712269277
"57"	14.428294573643413
"6"	7.501923076923076
"60"	12.572296296296297
"63"	19.20178025338576
"66"	17.815385709115354
"69"	16.73716407355021
"72"	19.675488721804513
"75"	12.912284473764666
"78"	12.958470353761832
"81"	13.033249581239536
"84"	11.774166666666667
"87"	20.44241956084242
"9"	84.49022346368714

 hadoop@ip-172-31-1-58:~


```
"90"      14.177380536854745
"93"      33.98838046893884
"96"      15.732733812949643
"99"      12.726666666666668
"1"       8.186450867972605
"10"      49.7685703859137
"100"     15.4517272060907
"106"     14.142386707811873
"109"     60.77628205128205
"112"     14.44611286089239
"115"     14.100138312586447
"118"     12.259841269841266
"121"     14.144444444444442
"124"     28.927140549273023
"127"     15.566097332709408
"13"      19.502000159626302
"130"     35.166737556898745
"133"     16.60591997376189
"136"     11.392984869325996
"139"     59.65673076923076
"142"     13.862745217439608
"145"     12.773806451317668
"148"     16.668522984065728
"151"     13.21462795906497
"154"     24.45357142857143
"157"     20.50430167597765
"16"      21.921666666666667
"160"     20.065231092436967
"163"     15.8627499990011332
"166"     14.844402353615042
"169"     15.798503166378815
"172"     15.978240740740741
"175"     12.795833333333333
"178"     6.8091676040494935
"181"     15.84955056973453
"184"     16.714705882352945
"187"     38.99130434782609
"19"      12.450746268656715
"190"     19.28594566353187
"193"     14.370130099617871
"196"     17.658003230790836
"199"     17.451587301587306
"202"     16.211688073394495
"205"     25.10547839506173
```

 hadoop@ip-172-31-1-58:~

```
"208" 26.57637860082305
"211" 16.764861618388338
"214" 9.997727272727275
"217" 17.581798105576837
"22" 21.920180061484405
"220" 12.070778210116734
"223" 14.933757307290453
"226" 16.553316516925054
"229" 13.281866325145375
"232" 17.19366355955486
"235" 16.920605573419078
"238" 12.907283813308853
"241" 15.209362987369575
"244" 16.833093464363078
"247" 16.078691837963365
"25" 15.483484374354518
"250" 16.762733529990175
"253" 22.27765957446809
"256" 16.111507431291688
"259" 15.158308605341245
"262" 12.843559332956236
"265" 9.502508024311965
"28" 28.173525921869295
"31" 24.915330788804074
"34" 15.864830368906453
"37" 16.128120727549618
"4" 14.994658890266352
"40" 16.212705440212662
"43" 15.613800432104453
"46" 17.604385964912286
"49" 13.98831907671419
"52" 17.332517617264916
"55" 24.754869281045757
"58" 6.46923076923077
"61" 15.336157115469021
"64" 13.288849765258217
"67" 19.378268398268403
"7" 13.264168029334517
"70" 25.207598586309523
"73" 20.80793103448275
"76" 21.72324318787134
"79" 15.19048047461381
"82" 14.64116856693155
"85" 19.751419667590028
```

 hadoop@ip-172-31-1-58:~

```
"88"      21.651543974983273
"91"      20.863813989839784
"94"      19.335945273631843
"97"      15.708167304502073
"101"     18.519526143790845
"104"     23.616666666666667
"107"     14.170103283247789
"11"      17.169207317073177
"110"     3.1833333333333333
"113"     14.999849531931579
"116"     15.63783217332631
"119"     16.320988640397587
"122"     27.162833333333342
"125"     15.899441602988999
"128"     14.897942905121742
"131"     14.111753902662992
"134"     15.84854374498337
"137"     13.785309907042645
"14"      16.86699357650837
"140"     13.92865655046817
"143"     13.114742778188864
"146"     15.048244012458383
"149"     16.72864406779661
"152"     14.159447307099024
"155"     27.028838383838384
"158"     16.619863399832717
"161"     15.588043428280558
"164"     15.424552738208508
"167"     14.993433980349867
"17"      14.029984432134967
"170"     14.685756051764045
"173"     14.65303605313093
"176"     30.772619047619052
"179"     14.502859519376642
"182"     13.937760814249364
"185"     18.767688679245282
"188"     16.133771704730133
"191"     16.42855450236967
"194"     24.988386546429446
"197"     15.662662013958126
"2"       38.16594202898551
"20"      14.108603491271824
"200"     12.831729264475747
"203"     27.164993306559573
```


 hadoop@ip-172-31-1-58:~

```
"206" 13.607482993197276
"209" 19.135244233080925
"212" 21.123603839441543
"215" 48.54615384615384
"218" 20.67351945854484
"221" 14.373676680972823
"224" 13.713107552798089
"227" 14.55116129032258
"23" 12.765217391304349
"230" 16.930732667330382
"233" 15.269118834293334
"236" 12.747239751606356
"239" 13.126088201984295
"242" 14.231421677802523
"245" 8.677222222222222
"248" 15.719244734931014
"251" 11.8752688172043
"254" 19.044134078212284
"257" 15.420086767895883
"26" 11.983088803088805
"260" 15.507379827068222
"263" 12.159215838037262
"29" 21.90633270321361
"32" 20.141068580542257
"35" 19.435623073535883
"38" 50.98867102396513
"41" 12.956678529226924
"44" 11.74047619047619
"47" 16.59013605442177
"5" 11.444871794871794
"50" 14.86682709472455
"53" 17.147445887445887
"56" 20.112871287128716
"59" 16.816666666666667
"62" 15.30254922644163
"65" 16.960852573062777
"68" 15.32582473334189
"71" 16.314197530864202
"74" 12.6286895569543
"77" 19.922931442080372
"8" 17.490176908752332
"80" 15.157747494884847
"83" 15.877727292119967
"86" 14.451282051282048
```

```
"89" 15.994494320263835
"92" 17.523183493875607
"95" 17.861217802164305
"98" 12.256024096385541
```

Question 5: Calculate the average tips to revenue ratio of the drivers for different pickup locations in sorted format.


We used the command **python mrtask_e.py -r Hadoop**

hdfs:///user/Hadoop/assignment/input/*.csv > output_e.txt for running the code on all the csv files and then we used the command **cat output_e.txt** for seeing the content i.e. the solution of the query.


```
[hadoop@ip-172-31-1-58 ~]$ python mrtask_e.py -r hadoop hdfs:///user/hadoop/assignment/input/*.csv > output_e.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.8.5
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/mrtask_e.hadoop.20240408.121002.970948
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/mrtask_e.hadoop.20240408.121002.970948/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/mrtask_e.hadoop.20240408.121002.970948/files/
Running step 1 of 1...
packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-6.jar] /tmp/streamjob2626659764121334840.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-1-58.ec2.internal/172.31.1.58:8032
Connecting to ResourceManager at ip-172-31-1-58.ec2.internal/172.31.1.58:8032
Loaded native gpl library
Successfully loaded & initialized native-lzo library [hadoop-lzo rev ff8f5709577defb6b78cdc1f98cfe129c4b6fe46]
Total input files to process : 6
Adding a new node: /default-rack/172.31.1.58:50010
number of splits:43
Submitting tokens for job: job_1712568903374_0010
Submitted application application_1712568903374_0010
The url to track the job: http://ip-172-31-1-58.ec2.internal:20888/proxy/application_1712568903374_0010/
Running job: job_1712568903374_0010
Job job_1712568903374_0010 running in uber mode : false
  map 0% reduce 0%
  map 1% reduce 0%
  map 2% reduce 0%
  map 3% reduce 0%
  map 4% reduce 0%
  map 5% reduce 0%
  map 6% reduce 0%
  map 7% reduce 0%
  map 8% reduce 0%
  map 9% reduce 0%
  map 10% reduce 0%
```

```
[hadoop@ip-172-31-1-58 ~]$ cat output_e.txt
```


```
"102" 0.09286325213174113
"105" 0.08289275441720316
"108" 0.08561073480894167
"111" 0.10175458286569401
"114" 0.11737171385745776
"117" 0.1654611532154413
"12" 0.08816912666410082
"120" 0.08303252748719829
"123" 0.11055898094072128
"126" 0.07223106849013392
"129" 0.06331017333926042
"132" 0.10404480517960506
"135" 0.0911299209238622
"138" 0.13191299589436906
"141" 0.10768193607242953
"144" 0.1128301352168931
"147" 0.044604149528370926
"15" 0.11600468842514833
"150" 0.10406209924668991
"153" 0.0605637325357794
"156" 0.08792537666968177
"159" 0.04663800562314162
"162" 0.12066632396973818
"165" 0.06797416237704265
"168" 0.049620759878179044
"171" 0.0805299331453422
"174" 0.0627814335084243
"177" 0.0764169635288451
"18" 0.052419716500855626
"180" 0.1052369675037582
"183" 0.05853409985743475
"186" 0.10977067019782542
"189" 0.11197515428640377
"192" 0.07124383780662925
"195" 0.10463665842928402
"198" 0.08806313330584344
"201" 0.15164728494049579
"204" 0.06247952716655987
"207" 0.07416425281278811
"21" 0.0829740468340519
"210" 0.1269629102823797
```

 hadoop@ip-172-31-1-58:~


```
"213" 0.09930916492713443
"216" 0.10479778546735057
"219" 0.10013130970242759
"222" 0.09461551127943356
"225" 0.08262074885724496
"228" 0.0972540407536762
"231" 0.1188786215234706
"234" 0.12013283079536198
"237" 0.10990442459596406
"24" 0.10590891911111384
"240" 0.07105544591130968
"243" 0.0964975181793712
"246" 0.11961317882575757
"249" 0.11990213646712744
"252" 0.11791966686185029
"255" 0.11685699515470141
"258" 0.11397415106035945
"261" 0.10766619243050964
"264" 0.11265242617792516
"27" 0.0880988704482764
"3" 0.06985350809560525
"30" 0.2561075875029364
"33" 0.12071128564971682
"36" 0.1058048905047359
"39" 0.09651707726856311
"42" 0.07036241083616263
"45" 0.0993427451689516
"48" 0.10627040728725275
"51" 0.08952152479566311
"54" 0.12031650579198637
"57" 0.09639556671648823
"6" 0.0483181888584852
"60" 0.04087201889401171
"63" 0.10608349633094599
"66" 0.11716659766185668
"69" 0.047707317902002305
"72" 0.08865379692027035
"75" 0.09013747412603465
"78" 0.049456821210188295
"81" 0.08090785981257781
"84" 0.12333321182517408
"87" 0.12471673086142322
"9" 0.09026115108356467
"90" 0.11445563465274146
```

 hadoop@ip-172-31-1-58:~


```
"93"      0.11009789332101659
"96"      0.112153063815818
"99"      0.06008865539320309
"1"       0.11931462821703112
"10"      0.10761668826560952
"100"     0.10176112040929118
"106"     0.11182299192052228
"109"     0.17861970356364318
"112"     0.10939659242523393
"115"     0.09805645544907124
"118"     0.117289635611784
"121"     0.09104834906722972
"124"     0.08450445139029464
"127"     0.0883881479101126
"13"      0.12263264632357321
"130"     0.10459971957672475
"133"     0.09428655857404536
"136"     0.040904378620177845
"139"     0.09244706142211939
"142"     0.11242578567788385
"145"     0.09530064643710734
"148"     0.11466547302905533
"151"     0.10736778554968535
"154"     0.09108097210043588
"157"     0.10078849415880729
"16"      0.123473429808929
"160"     0.09647705819251776
"163"     0.11222164046426629
"166"     0.11278675284128852
"169"     0.06451074143663851
"172"     0.17223686242471645
"175"     0.1277230776008282
"178"     0.12194781013939289
"181"     0.11303400885963028
"184"     0.11633468693321353
"187"     0.17978913134704158
"19"      0.07943494943114209
"190"     0.10147578909371881
"193"     0.04602694182348803
"196"     0.08616922680964281
"199"     0.1402469488259225
"202"     0.09722093931447474
"205"     0.0830228408954791
"208"     0.07718716885014357
```

 hadoop@ip-172-31-1-58:~

```
"211" 0.11399355397890307
"214" 0.07745375602878152
"217" 0.07433903922530723
"22" 0.06303381261429236
"220" 0.07548055451382421
"223" 0.09657319511944477
"226" 0.08554438430856706
"229" 0.11256386704368666
"232" 0.0836417852316917
"235" 0.051485492996097205
"238" 0.11044770763095635
"241" 0.061427235636888386
"244" 0.09413161238217443
"247" 0.07294934365907735
"25" 0.10931197115095471
"250" 0.057044008706736905
"253" 0.0944494192326738
"256" 0.10853673778737989
"259" 0.08717567117240324
"262" 0.11338221978049443
"265" 0.11224718082319106
"28" 0.10225113334394038
"31" 0.09122118342582707
"34" 0.10658808178106197
"37" 0.09921178378979158
"4" 0.1068487320092819
"40" 0.12031726919563414
"43" 0.10492270664853011
"46" 0.08234332088746263
"49" 0.09943300939914614
"52" 0.12903601946206247
"55" 0.0947055133274003
"58" 0.14391826665236512
"61" 0.09216038290043962
"64" 0.11356366975963511
"67" 0.10270469503747046
"7" 0.07831986411658208
"70" 0.10788380927811847
"73" 0.07567600767754318
"76" 0.11266074016140312
"79" 0.1180040681982511
"82" 0.06474358825816975
"85" 0.06602854511265933
"88" 0.11647722217028368
```

 hadoop@ip-172-31-1-58:~

```
"91"      0.0777068998470948
"94"      0.03261240109944186
"97"      0.09860378966753362
"101"     0.10391200650224514
"104"     0.2000665778961385
"107"     0.11974753027366687
"11"      0.06611923513377356
"110"     0.0
"113"     0.11854772693416192
"116"     0.09022031542011356
"119"     0.06778234295101979
"122"     0.1324025567728717
"125"     0.12389238738089078
"128"     0.10195419121098445
"131"     0.08338783482815319
"134"     0.10749874267013129
"137"     0.1127747256591832
"14"      0.09044594848522673
"140"     0.11118444211894547
"143"     0.11408622408797313
"146"     0.08853658145675754
"149"     0.08233859572151551
"152"     0.08319834814829787
"155"     0.09633375755470426
"158"     0.11639038989044094
"161"     0.1135904923441109
"164"     0.11217684253880635
"167"     0.061222523830325414
"17"      0.08317536901455541
"170"     0.11808315282274244
"173"     0.06008145804370089
"176"     0.15889955267208694
"179"     0.08477276212070424
"182"     0.04956118360347682
"185"     0.07530969983194838
"188"     0.07438531115226625
"191"     0.1248716806280983
"194"     0.1212934343540359
"197"     0.09057550505118195
"2"       0.10551996408687951
"20"      0.07539779360644339
"200"     0.10311919266153374
"203"     0.08922376129095635
"206"     0.08357280929682974
```

 hadoop@ip-172-31-1-58:~

```
"209" 0.11438950713881713
"212" 0.07459991612034522
"215" 0.1082922167109166
"218" 0.09787185841442556
"221" 0.10296610645201126
"224" 0.11526880032062242
"227" 0.08060571925596914
"23" 0.11893642866783635
"230" 0.1076097076744632
"233" 0.11587691604712344
"236" 0.11108195526453167
"239" 0.11135965083851737
"242" 0.07134362315468343
"245" 0.08812313347116932
"248" 0.05406105607423323
"251" 0.08299641950457239
"254" 0.045390942684416416
"257" 0.1131136567021167
"26" 0.06577859806429147
"260" 0.07000191673138242
"263" 0.11388129538454757
"29" 0.06573652675700856
"32" 0.06109787327761728
"35" 0.10022554243576677
"38" 0.05791427287264455
"41" 0.09013710776489915
"44" 0.06402542001826438
"47" 0.03560717179556587
"5" 0.173567645642754
"50" 0.10916652254418897
"53" 0.07740291031997901
"56" 0.09883778991554348
"59" 0.011954905847373635
"62" 0.07661527162660449
"65" 0.10543895761087184
"68" 0.11443748108215406
"71" 0.09976058424195855
"74" 0.07135080378568034
"77" 0.09702969688085886
"8" 0.1086282452986951
"80" 0.1102718770090173
"83" 0.05736768211329472
"86" 0.10935811216524559
"89" 0.08958574112049193
```


```
"92" 0.06480851959104653
"95" 0.0833569398816454
"98" 0.08788658752859359
```


Question 6: How does revenue vary over time? Calculate the average trip revenue per month - analysing it by hour of the day (day vs night) and the day of the week (weekday vs weekend).

We used the command **python mrtask_f.py -r Hadoop**


hdfs:///user/Hadoop/assignment/input/*.csv > output_f.txt for running the code on all the csv files and then we used the command **cat output_f.txt** for seeing the content i.e. the solution of the query.

```
hadoop@ip-172-31-1-58:~$ python mrtask_f.py -r hadoop hdfs:///user/hadoop/assignment/input/*.csv > output_f.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.8.5
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar
Creating temp directory /tmp/mrtask_f.hadoop.20240408.122825.715298
uploading working dir files to hdfs:///user/hadoop/tmp/mrjob/mrtask_f.hadoop.20240408.122825.715298/files/wd...
Copying other local files to hdfs:///user/hadoop/tmp/mrjob/mrtask_f.hadoop.20240408.122825.715298/files/
Running step 1 of 1...
packageJobJar: [ [/usr/lib/hadoop/hadoop-streaming-2.8.5-amzn-6.jar] /tmp/streamjob330379128793805024.jar tmpDir=null
Connecting to ResourceManager at ip-172-31-1-58.ec2.internal/172.31.1.58:8032
Connecting to ResourceManager at ip-172-31-1-58.ec2.internal/172.31.1.58:8032
Loaded native gpl library
Successfully loaded & initialized native-lzo library [hadoop-lzo rev ff8f5709577defb6b78cdc1f98cfe129c4b6fe46]
Total input files to process : 6
Adding a new node: /default-rack/172.31.1.58:50010
number of splits:43
Submitting tokens for job: job_1712568903374_0011
Submitted application application_1712568903374_0011
The url to track the job: http://ip-172-31-1-58.ec2.internal:20888/proxy/application_1712568903374_0011/
Running job: job_1712568903374_0011
Job job_1712568903374_0011 running in uber mode : false
  map 0% reduce 0%
  map 1% reduce 0%
  map 2% reduce 0%
  map 3% reduce 0%
  map 4% reduce 0%
  map 5% reduce 0%
  map 6% reduce 0%
  map 7% reduce 0%
  map 8% reduce 0%
  map 9% reduce 0%
  map 10% reduce 0%
  map 11% reduce 0%
  map 12% reduce 0%
  map 13% reduce 0%
  map 14% reduce 0%
  map 15% reduce 0%
  map 16% reduce 0%
  map 17% reduce 0%
```


 hadoop@ip-172-31-1-58:~

```
[hadoop@ip-172-31-1-58 ~]$ cat output_f.txt
```


```
[1, 0, 0]      17.519291696330246
[1, 0, 3]      16.626480239594823
[1, 0, 6]      15.44823421897282
[1, 1, 2]      16.55492861952621
[1, 1, 5]      14.981404492508563
[1, 10, 1]     15.127706071055114
[1, 10, 4]     15.164584937260672
[1, 11, 0]     14.480895520162157
[1, 11, 3]     15.248743102814249
[1, 11, 6]     13.477203683198983
[1, 12, 2]     15.195805597340291
[1, 12, 5]     13.17964793309867
[1, 13, 1]     15.129944553176106
[1, 13, 4]     15.731715024471264
[1, 14, 0]     15.17123758065054
[1, 14, 3]     16.288452460943006
[1, 14, 6]     15.353656693544734
[1, 15, 2]     15.808623447320434
[1, 15, 5]     14.286627006407732
[1, 16, 1]     16.79477246941084
[1, 16, 4]     17.140632285850383
[1, 17, 0]     15.603011225898907
[1, 17, 3]     16.493862512031395
[1, 17, 6]     15.842124390133883
[1, 18, 2]     15.838856512260497
[1, 18, 5]     13.799004771741297
[1, 19, 1]     15.184492458519966
[1, 19, 4]     15.181602676051245
[1, 2, 1]      16.653553023303136
[1, 2, 4]      15.440661491541219
[1, 20, 0]     15.381471427086986
[1, 20, 3]     15.787915693741633
[1, 20, 6]     16.05652368409678
[1, 21, 2]     15.75965917212238
[1, 21, 5]     14.782438900603976
[1, 22, 1]     16.13572880016946
[1, 22, 4]     15.62656435800056
[1, 23, 0]     17.64719986594598
[1, 23, 3]     16.713724633164848
[1, 23, 6]     17.937987911386074
[1, 3, 0]      18.7976142942689
[1, 3, 3]      17.382108248058906
[1, 3, 6]      16.042656562921923
```

 hadoop@ip-172-31-1-58:~


```
[1, 4, 2]      19.310694200352604
[1, 4, 5]      16.99999958379708
[1, 5, 1]      18.400929206212094
[1, 5, 4]      20.039551492738905
[1, 6, 0]      16.266336098291532
[1, 6, 3]      14.47549190956348
[1, 6, 6]      19.54508136254779
[1, 7, 2]      14.02216718071056
[1, 7, 5]      15.439742067086193
[1, 8, 1]      14.314401389019036
[1, 8, 4]      14.534665978680467
[1, 9, 0]      14.765277183661
[1, 9, 3]      23.709058874286395
[1, 9, 6]      13.630402302883965
[2, 0, 1]      17.435129216127834
[2, 0, 4]      16.750568706440365
[2, 1, 0]      15.836377001999157
[2, 1, 3]      15.593291943538407
[2, 1, 6]      15.080928222376645
[2, 10, 2]     15.584856505534193
[2, 10, 5]     13.235825072012585
[2, 11, 1]     15.137241340754445
[2, 11, 4]     15.81282871425917
[2, 12, 0]     14.997769908453845
[2, 12, 3]     15.09902583793382
[2, 12, 6]     13.83217228096712
[2, 13, 2]     16.214876939591328
[2, 13, 5]     14.139899300861922
[2, 14, 1]     15.76594127130329
[2, 14, 4]     16.988947496203075
[2, 15, 0]     15.391670236531361
[2, 15, 3]     15.68904171508541
[2, 15, 6]     15.989882115791604
[2, 16, 2]     17.706524452030504
[2, 16, 5]     14.637107029630501
[2, 17, 1]     16.018190446543887
[2, 17, 4]     16.885100841332783
[2, 18, 0]     15.211415337181817
[2, 18, 3]     15.424042083941371
[2, 18, 6]     15.117485791174863
[2, 19, 2]     15.4460226975043
[2, 19, 5]     14.126359078349951
[2, 2, 2]      15.231154903757654
[2, 2, 5]      15.293780018084675
```

 hadoop@ip-172-31-1-58:~


```
[2, 20, 1] 15.249453542939467
[2, 20, 4] 15.531191802088282
[2, 21, 0] 16.226778214033512
[2, 21, 3] 16.01842724605061
[2, 21, 6] 16.746228025406182
[2, 22, 2] 20.630915118155208
[2, 22, 5] 15.260177546685307
[2, 23, 1] 16.808227004006397
[2, 23, 4] 16.582257244731913
[2, 3, 1] 16.94913790051442
[2, 3, 4] 16.41850477558805
[2, 4, 0] 20.0647689099957
[2, 4, 3] 18.939269769650853
[2, 4, 6] 16.763769939732974
[2, 5, 2] 19.832279648292477
[2, 5, 5] 20.540294614213263
[2, 6, 1] 14.334615571706394
[2, 6, 4] 16.161704803588766
[2, 7, 0] 14.546294714961382
[2, 7, 3] 13.693460850038047
[2, 7, 6] 17.079383247104946
[2, 8, 2] 14.14771303439259
[2, 8, 5] 13.672305677349025
[2, 9, 1] 14.75983417427716
[2, 9, 4] 15.184932077007824
[3, 0, 2] 17.60184462105493
[3, 0, 5] 16.53061715387153
[3, 1, 1] 16.412711989776323
[3, 1, 4] 16.16010741918623
[3, 10, 0] 15.792245838245725
[3, 10, 3] 16.30214075876189
[3, 10, 6] 13.905262614980394
[3, 11, 2] 16.3419719938788
[3, 11, 5] 13.390678702178555
[3, 12, 1] 15.43057074350426
[3, 12, 4] 16.617459412199377
[3, 13, 0] 16.386373148835407
[3, 13, 3] 17.851531498396284
[3, 13, 6] 15.151229027649938
[3, 14, 2] 17.23241996833708
[3, 14, 5] 14.679936774741646
[3, 15, 1] 15.852790671648803
[3, 15, 4] 17.562402247233944
[3, 16, 0] 18.042703708594036
```

 hadoop@ip-172-31-1-58:~


```
[3, 16, 3]      19.769173255703908
[3, 16, 6]      16.793010721934042
[3, 17, 2]      17.585606638175708
[3, 17, 5]      14.502202293992896
[3, 18, 1]      17.727856811937453
[3, 18, 4]      16.31471042780239
[3, 19, 0]      15.74178967979457
[3, 19, 3]      16.543771928306818
[3, 19, 6]      15.698528578631938
[3, 2, 0]       18.220575718169883
[3, 2, 3]       15.890491018705799
[3, 2, 6]       15.354934237852813
[3, 20, 2]      16.057102211972925
[3, 20, 5]      14.780932733913394
[3, 21, 1]      16.068186873652795
[3, 21, 4]      15.868536399795511
[3, 22, 0]      16.649822266774535
[3, 22, 3]      17.34474521935216
[3, 22, 6]      18.42269414747654
[3, 23, 2]      17.39454532968209
[3, 23, 5]      16.039129766204066
[3, 3, 2]       17.506017974835537
[3, 3, 5]       15.996357103092546
[3, 4, 1]       18.381255108993038
[3, 4, 4]       20.2244169797604
[3, 5, 0]       21.591044198517025
[3, 5, 3]       20.327765055595027
[3, 5, 6]       20.056067382880784
[3, 6, 2]       15.13516179216114
[3, 6, 5]       19.677382655728316
[3, 7, 1]       14.135772450457434
[3, 7, 4]       14.53246586138814
[3, 8, 0]       15.08263894319259
[3, 8, 3]       15.118939288942213
[3, 8, 6]       15.442457006755744
[3, 9, 2]       15.382085430660505
[3, 9, 5]       13.149422371698428
[4, 0, 0]       19.806474700536782
[4, 0, 3]       17.856533620807415
[4, 0, 6]       15.542700442550043
[4, 1, 2]       18.574224053254436
[4, 1, 5]       16.25433095878499
[4, 10, 1]      15.571743585666471
[4, 10, 4]      15.974719159315567
```

 hadoop@ip-172-31-1-58:~

```
[4, 11, 0]      19.70811885114484
[4, 11, 3]      16.619808877658688
[4, 11, 6]      14.445806276315977
[4, 12, 2]      16.32499333063759
[4, 12, 5]      14.506856961057395
[4, 13, 1]      16.01739194415448
[4, 13, 4]      17.114006931653943
[4, 14, 0]      16.603916121847025
[4, 14, 3]      17.886031287752214
[4, 14, 6]      16.840753872341526
[4, 15, 2]      16.831821811911965
[4, 15, 5]      15.83041643239797
[4, 16, 1]      17.303590020325895
[4, 16, 4]      18.66117987598206
[4, 17, 0]      16.935342234490466
[4, 17, 3]      18.247976802512632
[4, 17, 6]      17.193937591282953
[4, 18, 2]      16.369772137047704
[4, 18, 5]      14.624873769593135
[4, 19, 1]      15.497339795567646
[4, 19, 4]      15.762378440794592
[4, 2, 1]       16.34083949787082
[4, 2, 4]       18.132232268254125
[4, 20, 0]      15.806433938201556
[4, 20, 3]      16.359462410280326
[4, 20, 6]      16.98237080332982
[4, 21, 2]      16.449890988815383
[4, 21, 5]      15.33205354060174
[4, 22, 1]      16.76673100333906
[4, 22, 4]      16.520808734827668
[4, 23, 0]      17.74243445353905
[4, 23, 3]      18.103497884552784
[4, 23, 6]      19.53774978766057
[4, 3, 0]       17.035402713379195
[4, 3, 3]       17.008209039548387
[4, 3, 6]       15.842982230340015
[4, 4, 2]       20.618848994142695
[4, 4, 5]       17.593235454788704
[4, 5, 1]       18.900795007561847
[4, 5, 4]       21.08462074977878
[4, 6, 0]       16.050704128041698
[4, 6, 3]       15.132386678529734
[4, 6, 6]       20.668970097719686
[4, 7, 2]       14.375317088653782
```

 hadoop@ip-172-31-1-58:~

```
[4, 7, 2]      14.375317088653782
[4, 7, 5]      16.12946016701572
[4, 8, 1]      14.541790406913538
[4, 8, 4]      14.856661646321268
[4, 9, 0]      15.706627748472956
[4, 9, 3]      15.331071204911453
[4, 9, 6]      13.942830341050417
[5, 0, 2]      17.686158986638155
[5, 0, 5]      17.10512802611608
[5, 1, 1]      19.429909785929077
[5, 1, 4]      18.217479700824367
[5, 10, 2]     27.168365753879677
[5, 10, 5]     14.029692293163865
[5, 11, 1]     17.194492364611953
[5, 11, 4]     16.728098769456157
[5, 12, 0]     16.480360929345643
[5, 12, 3]     17.975730053716745
[5, 12, 6]     15.27335388620442
[5, 13, 2]     17.774827762048254
[5, 13, 5]     14.990348808989964
[5, 14, 1]     17.347854799758533
[5, 14, 4]     18.613387909145654
[5, 15, 0]     16.915001820668373
[5, 15, 3]     18.75078904235611
[5, 15, 6]     17.056592424578085
[5, 16, 2]     19.654270891695475
[5, 16, 5]     15.635747754242638
[5, 17, 1]     17.747241962602104
[5, 17, 4]     18.411963137349318
[5, 18, 0]     16.500203407323646
[5, 18, 3]     17.221604263570693
[5, 18, 6]     16.3638792079018
[5, 19, 2]     16.27062012585948
[5, 19, 5]     14.486719040663518
[5, 2, 0]      17.069302560401542
[5, 2, 3]      15.885527411272996
[5, 2, 6]      15.219828879149524
[5, 20, 1]     15.922293130064107
[5, 20, 4]     16.04136461456434
[5, 21, 0]     16.81774921469835
[5, 21, 3]     17.121041382564865
[5, 21, 6]     16.8340853226704
[5, 22, 2]     16.730750002516107
[5, 22, 5]     15.161717603507663
```

 hadoop@ip-172-31-1-58:~

```
[6, 14, 1]      17.266575733489276
[6, 14, 4]      17.678796007574075
[6, 15, 0]      16.774017164809443
[6, 15, 3]      18.218019954522337
[6, 15, 6]      18.202896681798887
[6, 16, 2]      23.566478451392054
[6, 16, 5]      15.709298124154508
[6, 17, 1]      17.859978338442936
[6, 17, 4]      18.004343607922287
[6, 18, 0]      16.199296073804263
[6, 18, 3]      17.252703605015164
[6, 18, 6]      17.152516511911525
[6, 19, 2]      16.433430441398777
[6, 19, 5]      14.558758718910452
[6, 2, 0]       17.455471156604666
[6, 2, 3]       16.19311378592417
[6, 2, 6]       15.19330634222936
[6, 20, 0]      15.53537526218704
[6, 20, 3]      15.991703056782695
[6, 20, 6]      17.398259868691976
[6, 21, 2]      16.31239022151239
[6, 21, 5]      14.878435517330166
[6, 22, 1]      25.084740807689542
[6, 22, 4]      16.1677463476754
[6, 23, 0]      18.15033067701675
[6, 23, 3]      17.608237395538673
[6, 23, 6]      19.025608108112273
[6, 3, 2]       16.555133350596815
[6, 3, 5]       15.795642122531195
[6, 4, 1]       20.348739593983524
[6, 4, 4]       19.93230409052122
[6, 5, 0]       21.28951760015899
[6, 5, 3]       19.57409796286547
[6, 5, 6]       20.694516962284194
[6, 6, 2]       14.786332033232819
[6, 6, 5]       18.14070623729714
[6, 7, 1]       14.369559163977886
[6, 7, 4]       14.77600703237791
[6, 8, 0]       15.437935202167196
[6, 8, 3]       15.091789660540183
[6, 8, 6]       15.88194598284995
[6, 9, 2]       15.763047256856884
[6, 9, 5]       13.891624417895832
```

[hadoop@ip-172-31-1-58 ~]\$ █