MapReduce Tasks

a. Which vendors have the most trips, and what is the total revenue generated by that vendor?

Answer: Vendor 2 has the most number of trips with a total revenue of 525037658.13717655

```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/a.ec2-user.20240409.112416.608715
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/a.ec2-user.20240409.112416.608715/output
Streaming final output from /tmp/a.ec2-user.20240409.112416.608715/output...
"2"     525037658.13737655
Removing temp directory /tmp/a.ec2-user.20240409.112416.608715...
```

b. Which pickup location generates the most revenue?

Answer: Pick up location 132 generates the most revenue of 77196812.23977433

```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/b.ec2-user.20240409.112916.307257
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/b.ec2-user.20240409.112916.307257/output
Streaming final output from /tmp/b.ec2-user.20240409.112916.307257/output...
"132"   77196812.23977433
Removing temp directory /tmp/b.ec2-user.20240409.112916.307257...
```

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

```
[hadoop@ip-172-31-63-115 ~]$ python mrtask_c.py input > output.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_c.hadoop.20240409.070532.176989
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_c.hadoop.20240409.070532.176989/output
Streaming final output from /tmp/mrtask_c.hadoop.20240409.070532.176989/output...
Removing temp directory /tmp/mrtask_c.hadoop.20240409.070532.176989...
[hadoop@ip-172-31-63-115 ~]$ ls
input  input_sample  mrtask_c.py  output.txt  output_sample.txt
[hadoop@ip-172-31-63-115 ~]$ vi output.txt
[hadoop@ip-172-31-63-115 ~]$ cat output.txt
"1"     39754212
"2"     18832370
"3"     306912
"4"     88794
"5"     3
```

Answer: The payment types used by the customers in descending order of their counts are – 1,2,3,4 and 5.

d. What is the average trip time for different pickup locations?

```
[hadoop@ip-172-31-63-115 ~]$ cat output_d.txt
"1"     "0.0hours 8.0minutes 11.18705207835643seconds"
"10"    "0.0hours 49.0minutes 46.1142231548215seconds"
"100"   "0.0hours 15.0minutes 27.10363236544333seconds"
"101"   "0.0hours 18.0minutes 31.171568627451O5seconds"
"102"   "0.0hours 22.0minutes 6.7421875seconds"
"104"   "0.0hours 23.0minutes 37.0seconds"
"105"   "0.0hours 19.0minutes 58.66666666666674seconds"
"106"   "0.0hours 14.0minutes 8.54320246871248seconds"
"107"   "0.0hours 14.0minutes 10.20619699486872seconds"
"108"   "0.0hours 14.0minutes 12.504854368932001seconds"
"109"   "1.0hours 0.0minutes 46.57692307692287seconds"
"11"    "0.0hours 17.0minutes 10.152439024390333seconds"
"110"   "0.0hours 3.0minutes 11.0seconds"
"111"   "0.0hours 11.0minutes 38.079470198675494seconds"
"112"   "0.0hours 14.0minutes 26.766771653543287seconds"
"113"   "0.0hours 14.0minutes 59.99097191589476seconds"
"114"   "0.0hours 15.0minutes 55.27778777868252seconds"
"115"   "0.0hours 14.0minutes 6.008298755186729seconds"
"116"   "0.0hours 15.0minutes 38.26993039957847seconds"
"117"   "0.0hours 19.0minutes 19.05813953488382seconds"
"118"   "0.0hours 12.0minutes 15.590476190476238seconds"
"119"   "0.0hours 16.0minutes 19.25931842385512seconds"
"12"    "0.0hours 24.0minutes 21.522806414169736seconds"
"120"   "0.0hours 13.0minutes 48.48725212464592seconds"
"121"   "0.0hours 14.0minutes 8.666666666666629seconds"
"122"   "0.0hours 27.0minutes 9.769999999999982seconds"
"123"   "0.0hours 15.0minutes 30.87990762124707seconds"
"124"   "0.0hours 28.0minutes 55.628432956381175seconds"
"125"   "0.0hours 15.0minutes 53.96649617934008seconds"
"126"   "0.0hours 18.0minutes 33.706723891273214seconds"
"127"   "0.0hours 15.0minutes 33.96583996256436seconds"
"128"   "0.0hours 14.0minutes 53.876574307304736seconds"
"129"   "0.0hours 14.0minutes 15.521798520204925seconds"
"13"    "0.0hours 19.0minutes 30.120009577578458seconds"
"130"   "0.0hours 35.0minutes 10.004253413924289seconds"
"131"   "0.0hours 14.0minutes 6.705234159779593seconds"
"132"   "0.0hours 43.0minutes 46.21732317850319seconds"
```

Answer: The average trip time for different pickup locations are as follows. Full output is in attached file output_d.txt

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

```
[hadoop@ip-172-31-63-115 ~]$ python mrtask_e.py input > output_e.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_e.hadoop.20240409.094337.190240
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_e.hadoop.20240409.094337.190240/output
Streaming final output from /tmp/mrtask_e.hadoop.20240409.094337.190240/output...
Removing temp directory /tmp/mrtask_e.hadoop.20240409.094337.190240...
[hadoop@ip-172-31-63-115 ~]$ cat output_e.txt
"30"    0.25610758750293633
"104"   0.2000665778961385
"187"   0.17978913134704155
"109"   0.17861970356364315
"5"     0.173567645642754
"172"   0.17223686242471645
"117"   0.1654611532154412
"176"   0.15889955267208697
"201"   0.15164728494049584
"58"    0.14391826665236512
"199"   0.14024694882592245
"122"   0.1324025567728717
"138"   0.13191299589451488
"52"    0.12903601946205775
"175"   0.1277230776008282
"210"   0.12696291028237974
"191"   0.12487168062809832
"87"    0.1247167308613737
```

Answer: The average tips to revenue ratio for different pickup locations are as follows. Full output is in attached file output_e.txt

Pick Up Location "30" has the highest tips to revenue ratio of 0.2561.

f. 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).

```
[root@ip-172-31-48-217 ~]# python mrtask_f_hour.py input > output_f_hour.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_f_hour.root.20240409.113351.567781
Running step 1 of 1...
job output is in /tmp/mrtask_f_hour.root.20240409.113351.567781/output
Streaming final output from /tmp/mrtask_f_hour.root.20240409.113351.567781/output...
Removing temp directory /tmp/mrtask_f_hour.root.20240409.113351.567781...
[root@ip-172-31-48-217 ~]# ls
input  mrtask_f_day.py  mrtask_f_hour.py  output_e_day.txt  output_f_day.txt  output_f_hour.txt
[root@ip-172-31-48-217 ~]# cat output_f_hour.txt
"Hour 0"        16.911903083006774
"Hour 1"        16.203888442736385
"Hour 10"       15.610222704522403
"Hour 11"       15.571164563395133
"Hour 12"       15.674443398415205
"Hour 13"       16.214996511293894
"Hour 14"       16.64061771359595
"Hour 15"       16.535629057645693
"Hour 16"       17.805189024367923
"Hour 17"       16.93630905349087
"Hour 18"       15.957381758597501
"Hour 19"       15.742812257388346
"Hour 2"        16.271619479811246
"Hour 20"       15.731142849516877
"Hour 21"       16.16968223699619
"Hour 22"       16.884520665560537
"Hour 23"       17.13291665957615
"Hour 3"        16.243783546663856
"Hour 4"        18.605361944275
"Hour 5"        20.06167337315054
"Hour 6"        15.937381145994852
"Hour 7"        14.750753078438422
"Hour 8"        14.817895492275344
"Hour 9"        15.226427545790274
[root@ip-172-31-48-217 ~]#
```

Answer:

Analyzing by hour of the day

Hour 5 (5AM to 5.59AM) has the highest revenue average of 20.06, Hour 7 (7AM to 7.59AM) has the lowest revenue average of 14.75

Answer - Analyzing by day of the week: Thursday has the highest average revenue of 16.73 while Saturday has the lowest average revenue of 15.07. Weekdays have higher revenue than weekends.

```
[root@ip-172-31-48-217 ~]# cat output_f_day.txt
"Friday"        16.453621777861922
"Monday"        16.340551651245836
"Saturday"      15.070632379167373
"Sunday"        16.08719346754698
"Thursday"      16.73224467327292
"Tuesday"       16.126401800151292
"Wednesday"     16.586508769673927
```