Task 4

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

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

Code File reference: mrtask_a.py

Code execution screenshot:

```
[hadoop@ip-172-31-77-47 ~]$ python mrtask_a.py < yellow_tripdata_2017-03.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_a.hadoop.20230312.203403.902112
Running step 1 of 2...
reading from STDIN
Running step 2 of 2...
job output is in /tmp/mrtask_a.hadoop.20230312.203403.902112/output
Streaming final output from /tmp/mrtask_a.hadoop.20230312.203403.902112/output...
5583181 "VeriFone Inc."
Removing temp directory /tmp/mrtask_a.hadoop.20230312.203403.902112...
[hadoop@ip-172-31-77-47 ~]$
```

Comment: VeriFone Inc. holds the highest number of trips in the yellow_tripdata_2017-03.csv dataset, totaling **5,583,181 trips.**

Code File reference:

mrtask_a_TC.py Code execution screenshot:

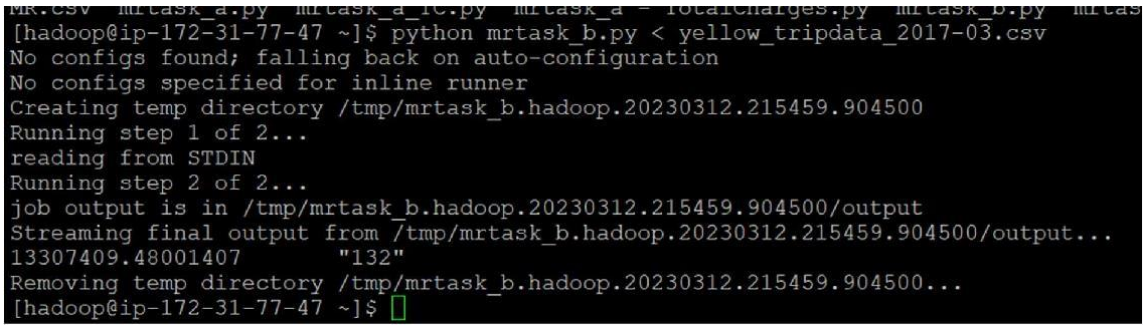
```
[hadoop@ip-172-31-77-47 ~]$ python mrtask_a_TC.py < yellow_tripdata_2017-03.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_a_TC.hadoop.20230312.211218.095998
Running step 1 of 1...
reading from STDIN
job output is in /tmp/mrtask_a_TC.hadoop.20230312.211218.095998/output
Streaming final output from /tmp/mrtask_a_TC.hadoop.20230312.211218.095998/output...
"Creative Mobile Technologies"  75347398.64700934
"VeriFone Inc." 91682368.32536966
Removing temp directory /tmp/mrtask_a_TC.hadoop.20230312.211218.095998...
[hadoop@ip-172-31-77-47 ~]$
```

Comment: VeriFone Inc. has generated a total revenue of **91,682,368.32.**

TASK 4

Code File reference: mrtask_b.py

Code execution screenshot:

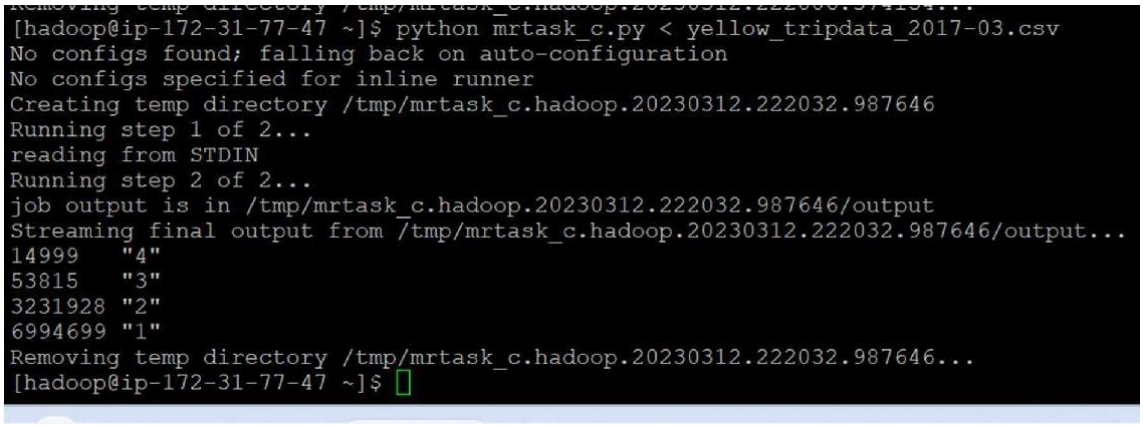


Remarks: Location 132 generated most revenue in the file tested.

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

Code File reference: mrtask_c.py

Code execution screenshot:



Remark: Credit card is mostly used followed by cash, then no charge and lowest one observed is dispute

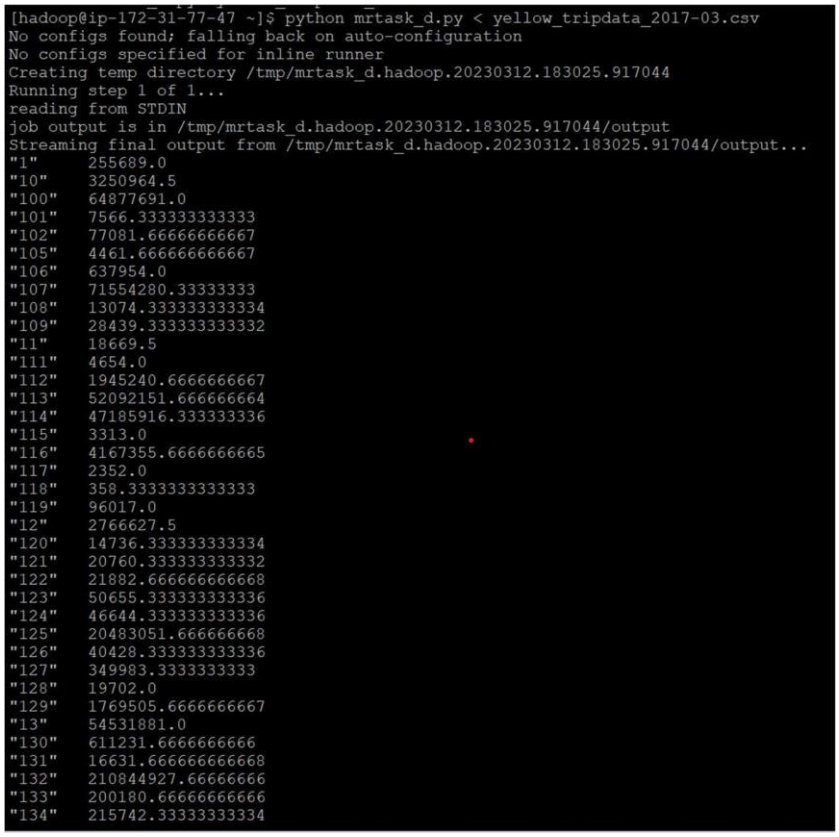
1= Credit card, 2= Cash, 3= No charge,4= Dispute, 5= Unknown, 6= Voided trip

TASK 4

What is the average trip time for different pickup locations?

Code File reference: mrtask_d.py

Code execution screenshot:



Remarks: Running map reduce job on yellow_tripdata_2017-03.csv. Above screenshot has average trip time in seconds for each location.

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

TASK 4

Code File reference: mrtask_e.py

```
[hadoop@ip-172-31-77-47 ~]$ python mrtask_e.py < yellow_tripdata_2017-03.csv
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_e.hadoop.20230312.214218.070937
Running step 1 of 2...
reading from STDIN
Running step 2 of 2...
job output is in /tmp/mrtask_e.hadoop.20230312.214218.070937/output
Streaming final output from /tmp/mrtask_e.hadoop.20230312.214218.070937/output...
0.0     "27"
0.0     "59"
0.026188979677351 77     "245"
0.05555555555555555      "118"
0.08324022346368715      "30"
0.08330668372241766      "44"
0.08333333333333333      "46"
0.08413395744299974      "117"
0.11013215859030838      "5"
0.13513927150290786      "214"
0.13664929969115955      "156"
0.15384615384615385      "187"
0.15791949398506777      "183"
0.16539883643787795      "221"
0.1664588528678304       "84"
0.18001762032128074      "206"
0.1966803775847734       "176"
0.199110563232387        "184"
0.2115120004910365       "139"
0.21793842034805888      "115"
0.23677482792527046      "204"
0.3277439150190755       "253"
0.3423711643513236       "111"
0.3727399722441718       "172"
0.3993444695654762       "150"
0.43193567854033804      "109"
0.45540935672514626      "105"
0.4941239458789766       "222"
0.4965950145681869       "58"
0.5082618573388116       "15"
0.552106455325866        "240"
0.6292976213757296       "122"
0.6713688303560592       "248"
0.6782093351324366       "154"
0.6834950879863939       "32"
0.688921042420514        "3"
0.7185910067954439       "254"
0.7281153939519145       "2"
0.7288072293656345       "64"
0.7310509617742412       "205"
```

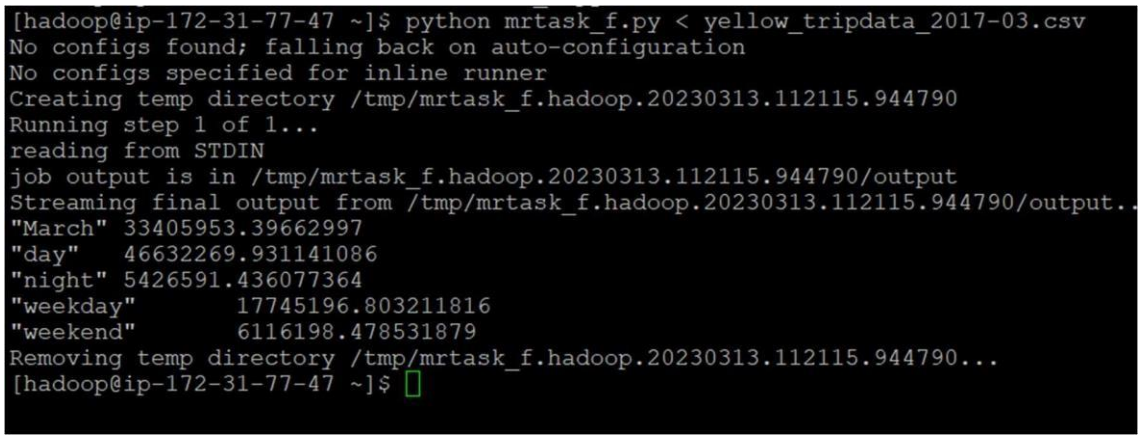
Comment: The screenshot above was taken during the testing of the code designed to calculate the average tip-to-revenue ratio for drivers in various locations.

Furthermore, the analysis aims to understand how revenue changes over time by calculating the average trip revenue per month. This analysis is carried out by considering the time of day (day versus night) and the day of the week (weekday versus weekend) as factors for comparison.

TASK 4

Code File reference: mrtask_f.py

Code execution screenshot:



Observations/Findings:

1. The dataset predominantly consists of data for the month of March.

2. The average revenue for trips in March can be calculated.

3. On average, trip revenue during daytime hours exceeds that of nighttime hours. We assume nighttime to be between 11 PM and 5 AM.

4. The dataset suggests that the average trip revenue during weekdays surpasses that of weekends, with Saturday and Sunday considered as weekends.