

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

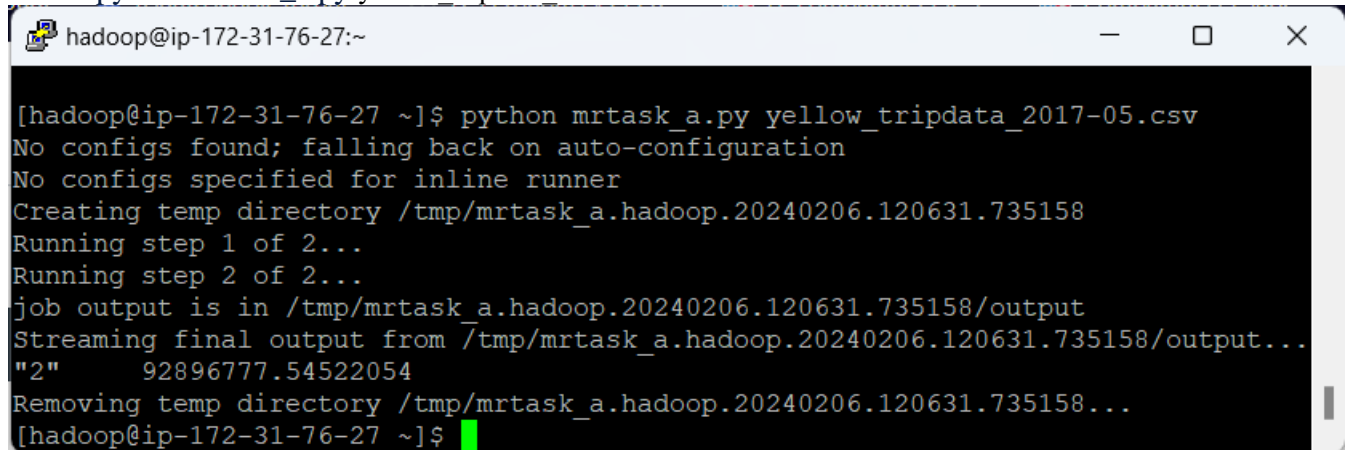
The dataset used to perform tasks: 'yellow_tripdata_2017-05.csv'.

Pre-requisite for these tasks:

- a) **Install GCC:** yum install gcc
- b) **Install Python:** sudo yum install python3-devel
- c) **Install Happybase:** pip install happybase
- d) **Start ThriftServer:** hbase thrift start
- e) **Import Happybase to python:** python -c 'import happybase'
- f) **Install MRJob:** pip install mrjob
- g) **Install pandas for python:** pip install pandas

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

Code: `python mrtask_a.py yellow_tripdata_2017-05.csv`

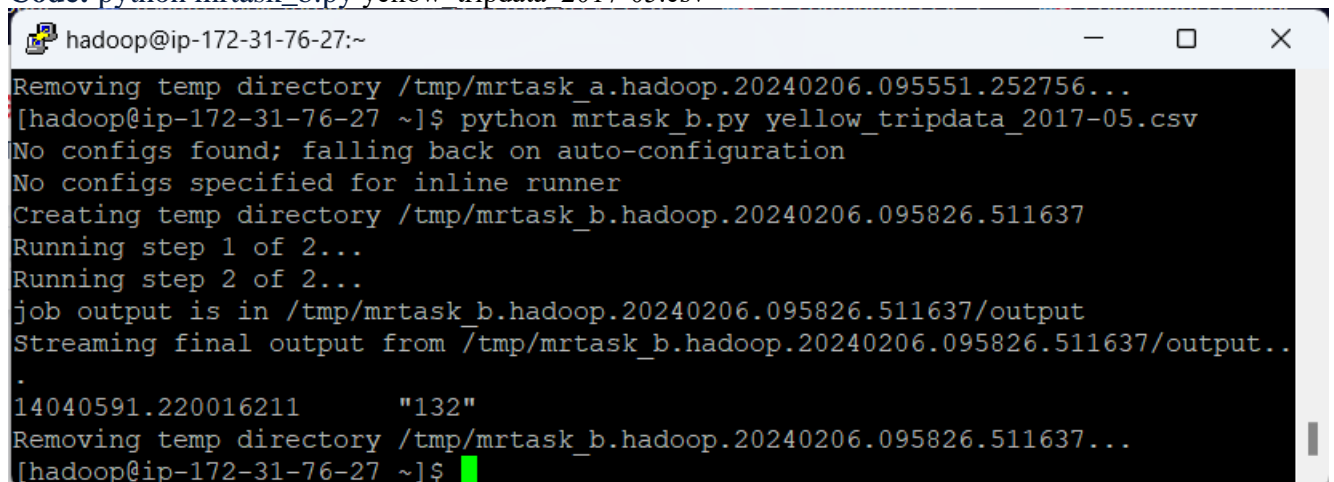


```
hadoop@ip-172-31-76-27:~  
[hadoop@ip-172-31-76-27 ~]$ python mrtask_a.py yellow_tripdata_2017-05.csv  
No configs found; falling back on auto-configuration  
No configs specified for inline runner  
Creating temp directory /tmp/mrtask_a.hadoop.20240206.120631.735158  
Running step 1 of 2...  
Running step 2 of 2...  
job output is in /tmp/mrtask_a.hadoop.20240206.120631.735158/output  
Streaming final output from /tmp/mrtask_a.hadoop.20240206.120631.735158/output...  
"2"          92896777.54522054  
Removing temp directory /tmp/mrtask_a.hadoop.20240206.120631.735158...  
[hadoop@ip-172-31-76-27 ~]$
```

Output: Vendor 2, i.e. “VeriFone Inc.”, has most trips and the total revenue generated is 92896777.54522054.

2. Which pickup location generates the most revenue?

Code: `python mrtask_b.py yellow_tripdata_2017-05.csv`



```
hadoop@ip-172-31-76-27:~  
Removing temp directory /tmp/mrtask_a.hadoop.20240206.095551.252756...  
[hadoop@ip-172-31-76-27 ~]$ python mrtask_b.py yellow_tripdata_2017-05.csv  
No configs found; falling back on auto-configuration  
No configs specified for inline runner  
Creating temp directory /tmp/mrtask_b.hadoop.20240206.095826.511637  
Running step 1 of 2...  
Running step 2 of 2...  
job output is in /tmp/mrtask_b.hadoop.20240206.095826.511637/output  
Streaming final output from /tmp/mrtask_b.hadoop.20240206.095826.511637/output..  
.  
14040591.220016211      "132"  
Removing temp directory /tmp/mrtask_b.hadoop.20240206.095826.511637...  
[hadoop@ip-172-31-76-27 ~]$
```

Output: Location ‘132’ generated most revenue in the dataset, i.e. 14040591.220016211.

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

Code: `python mrtask_c.py yellow_tripdata_2017-05.csv`

```
hadoop@ip-172-31-76-27:~  
[hadoop@ip-172-31-76-27 ~]$ python mrtask_c.py yellow_tripdata_2017-05.csv  
No configs found; falling back on auto-configuration  
No configs specified for inline runner  
Creating temp directory /tmp/mrtask_c.hadoop.20240206.100257.500211  
Running step 1 of 2...  
Running step 2 of 2...  
job output is in /tmp/mrtask_c.hadoop.20240206.100257.500211/output  
Streaming final output from /tmp/mrtask_c.hadoop.20240206.100257.500211/output..  
.  
15791    "4"  
55027    "3"  
3250362  "2"  
6780947  "1"  
Removing temp directory /tmp/mrtask_c.hadoop.20240206.100257.500211...  
[hadoop@ip-172-31-76-27 ~]$
```

Output: The predominant payment method is credit card, followed by cash, with no charge being the next common, and dispute observed least frequently.
1= Credit card, 2= Cash, 3= No charge, 4= Dispute, 5= Unknown, 6= Voided trip

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

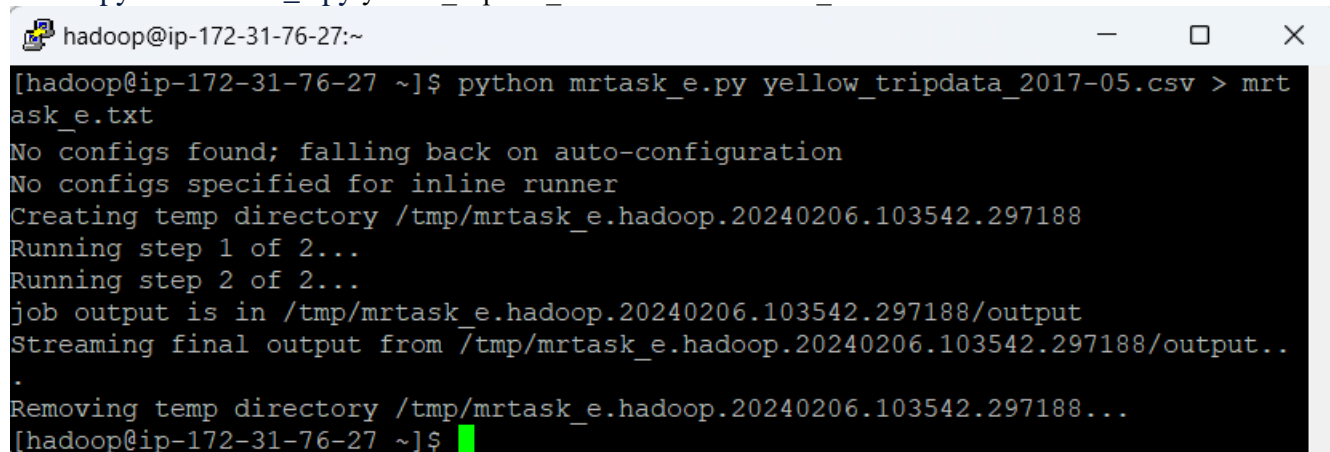
Code: `python mrtask_d.py yellow_tripdata_2017-05.csv > mrtask_d.txt`

```
hadoop@ip-172-31-76-27:~  
[hadoop@ip-172-31-76-27 ~]$ python mrtask_d.py yellow_tripdata_2017-05.csv > mrtask_d.txt  
No configs found; falling back on auto-configuration  
No configs specified for inline runner  
Creating temp directory /tmp/mrtask_d.hadoop.20240206.121831.210498  
Running step 1 of 1...  
job output is in /tmp/mrtask_d.hadoop.20240206.121831.210498/output  
Streaming final output from /tmp/mrtask_d.hadoop.20240206.121831.210498/output...  
Removing temp directory /tmp/mrtask_d.hadoop.20240206.121831.210498...  
[hadoop@ip-172-31-76-27 ~]$
```

Output: The output is generated in the file 'mrtask_d.txt'

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

Code: `python mrtask_e.py yellow_tripdata_2017-05.csv > mrtask_e.txt`

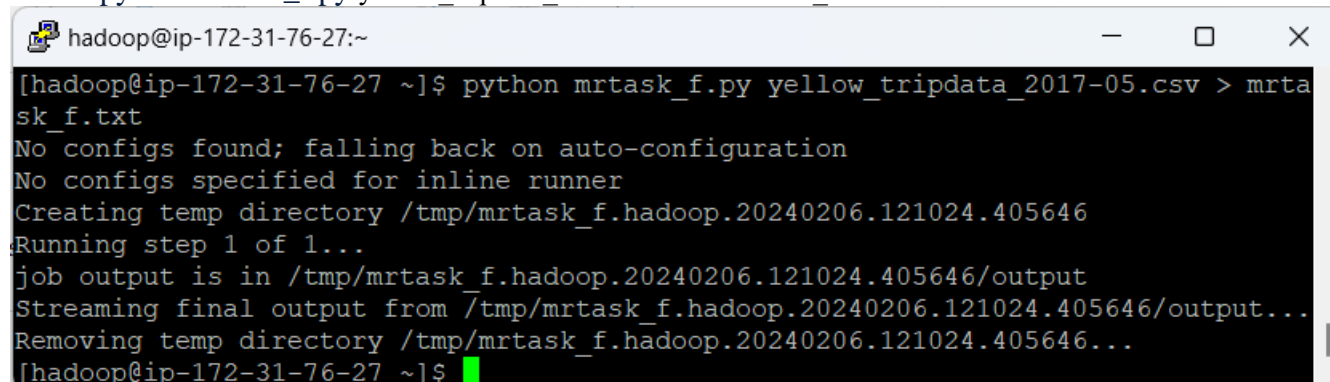
A terminal window titled 'hadoop@ip-172-31-76-27:~' showing the execution of the command 'python mrtask_e.py yellow_tripdata_2017-05.csv > mrtask_e.txt'. The output shows that no configurations were found, a temporary directory was created, and the job ran successfully in two steps. The final output is streamed to the file 'mrtask_e.txt' and the temporary directory is removed.

```
hadoop@ip-172-31-76-27:~$ python mrtask_e.py yellow_tripdata_2017-05.csv > mrtask_e.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_e.hadoop.20240206.103542.297188
Running step 1 of 2...
Running step 2 of 2...
job output is in /tmp/mrtask_e.hadoop.20240206.103542.297188/output
Streaming final output from /tmp/mrtask_e.hadoop.20240206.103542.297188/output...
.
Removing temp directory /tmp/mrtask_e.hadoop.20240206.103542.297188...
hadoop@ip-172-31-76-27:~$
```

Output: The output is generated in the file 'mrtask_e.txt'

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).

Code: `python mrtask_f.py yellow_tripdata_2017-05.csv > mrtask_f.txt`

A terminal window titled 'hadoop@ip-172-31-76-27:~' showing the execution of the command 'python mrtask_f.py yellow_tripdata_2017-05.csv > mrtask_f.txt'. The output shows that no configurations were found, a temporary directory was created, and the job ran successfully in one step. The final output is streamed to the file 'mrtask_f.txt' and the temporary directory is removed.

```
hadoop@ip-172-31-76-27:~$ python mrtask_f.py yellow_tripdata_2017-05.csv > mrtask_f.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/mrtask_f.hadoop.20240206.121024.405646
Running step 1 of 1...
job output is in /tmp/mrtask_f.hadoop.20240206.121024.405646/output
Streaming final output from /tmp/mrtask_f.hadoop.20240206.121024.405646/output...
Removing temp directory /tmp/mrtask_f.hadoop.20240206.121024.405646...
hadoop@ip-172-31-76-27:~$
```

Output: The output is generated in the file 'mrtask_f.txt'