# MapReduce Tasks (Task 4)

## Table of Contents

# MapReduce Tasks (Task 4)

This task is independent to Task 1, 2, and 3. However, It is assumed that before beginning of this Task, following common steps are already performed (common to Task 1, 2, and 3):

- Dataset Files are already downloaded through 'wget' to folder /root/tripdata/.
- Python files are transferred to /home/Hadoop/tasks/ folder. And execution permission has been added.
- Happybase and MRJob packages are installed.
- Shell variables DNS_EMR and DNS_RDS are already set to desired URLs.

## Upload of Dataset files to HDFS

Following hadoop commands are used to upload dataset files (.csv) to Hadoop File Systems (HDFS):

- Goto folder where CSV files are saved on local file System.
- **hadoop fs -mkdir**: to make a directory on HDFS.
- **hadoop fs -put**: to upload/put CSV files in HDFS.
- **hadoop fs -ls**: to verify the datasets uploaded on HDFS.

```
cd /root/tripdata/
ls -ltr
wc -l yellow_tripdata_2017-*.csv
hadoop fs -mkdir /user/root/tripdata/
hadoop fs -put yellow_tripdata_2017-*.csv /user/root/tripdata/
hadoop fs -ls /user/root/tripdata/
```

*Screenshot of switching as 'root' user*

```
[root@ip-172-31-27-72 tripdata]# ls -ltr
total 5425468
-rw-r--r-- 1 root root 914029540 Nov 25  2022 yellow_tripdata_2017-01.csv
-rw-r--r-- 1 root root 863487050 Nov 25  2022 yellow_tripdata_2017-02.csv
-rw-r--r-- 1 root root 969809025 Nov 25  2022 yellow_tripdata_2017-03.csv
-rw-r--r-- 1 root root 946349441 Nov 25  2022 yellow_tripdata_2017-04.csv
-rw-r--r-- 1 root root 951965526 Nov 25  2022 yellow_tripdata_2017-05.csv
-rw-r--r-- 1 root root 910028408 Nov 25  2022 yellow_tripdata_2017-06.csv
[root@ip-172-31-27-72 tripdata]# wc -l yellow_tripdata_2017-*.csv
   9710821 yellow_tripdata_2017-01.csv
   9169776 yellow_tripdata_2017-02.csv
  10295442 yellow_tripdata_2017-03.csv
  10047136 yellow_tripdata_2017-04.csv
  10102128 yellow_tripdata_2017-05.csv
   9656994 yellow_tripdata_2017-06.csv
  58982297 total
```

```
[root@ip-172-31-27-72 ~]# hadoop fs -mkdir /user/root/tripdata/
[root@ip-172-31-27-72 tripdata]# hadoop fs -put yellow_tripdata_2017-*.csv /user
/root/tripdata/
[root@ip-172-31-27-72 tripdata]# hadoop fs -ls /user/root/tripdata/
Found 6 items
-rw-r--r--   1 root hdfsadmingroup  914029540 2023-11-28 11:34 /user/root/tripda
ta/yellow_tripdata_2017-01.csv
-rw-r--r--   1 root hdfsadmingroup  863487050 2023-11-28 11:34 /user/root/tripda
ta/yellow_tripdata_2017-02.csv
-rw-r--r--   1 root hdfsadmingroup  969809025 2023-11-28 11:34 /user/root/tripda
ta/yellow_tripdata_2017-03.csv
-rw-r--r--   1 root hdfsadmingroup  946349441 2023-11-28 11:34 /user/root/tripda
ta/yellow_tripdata_2017-04.csv
-rw-r--r--   1 root hdfsadmingroup  951965526 2023-11-28 11:34 /user/root/tripda
ta/yellow_tripdata_2017-05.csv
-rw-r--r--   1 root hdfsadmingroup  910028408 2023-11-28 11:34 /user/root/tripda
ta/yellow_tripdata_2017-06.csv
[root@ip-172-31-27-72 tripdata]#
```

## Execution of MapReduce Tasks

While performing MapReduce Tasks, python will be used and following parameters/options will be used for each MR task:

- **Python File Name:** provide .py file name as input. If file is not in same directory, then provide fully qualified file path.
- **-r hadoop:** this option indicate that script will be executing using Hadoop platform.
- **Input files:** file names on Hadoop file system. To perform all tasks all .CSV files will be used from HDFS using “hdfs:///user/root/tripdata/*.csv”
- **--output-dir directoty_name:** the path of directory where output will be saved on HDFS. Make sure directory shouldn’t be already existing.

## Question 1. Which vendors have the most trips, and what is the total revenue generated by that vendor? [mrtask_a.py]

**Command used:**

```
python /home/hadoop/tasks/mrtask_a.py -r hadoop
hdfs:///user/root/tripdata/*.csv --output-dir
/user/root/tripdata/out_mrtask_a/
```

Output of script 1:

```
[root@ip-172-31-27-72 tripdata]# hadoop fs -cat /user/root/tripdata/out_mrtask_a
/*
"2"     525037658.1365462
[root@ip-172-31-27-72 tripdata]#
```

## Screenshot of execution of script 1:

```
[root@ip-172-31-27-72 tripdata]# python /home/hadoop/tasks/mrtask_a.py -r hadoop
 hdfs:///user/root/tripdata/*.csv --output-dir /user/root/tripdata/out_mrtask_a/
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /bin/hadoop
Using Hadoop version 3.3.3
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.root.20231128.113503.861858
uploading working dir files to hdfs:///user/root/tmp/mrjob/mrtask_a.root.2023112
8.113503.861858/files/wd...
Copying other local files to hdfs:///user/root/tmp/mrjob/mrtask_a.root.20231128.
113503.861858/files/
Running step 1 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-2.1.jar] /tmp/s
treamjob9095180633402192010.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_
1701169939900_0001
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 6
  Adding a new node: /default-rack/172.31.29.89:9866
  Adding a new node: /default-rack/172.31.21.145:9866
  number of splits:43
  Submitting tokens for job: job_1701169939900_0001
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1701169939900_0001
  The url to track the job: http://ip-172-31-27-72.ec2.internal:20888/proxy/appl
ication_1701169939900_0001/
  Running job: job_1701169939900_0001
  Job job_1701169939900_0001 running in uber mode : false
   map 0% reduce 0%
   map 1% reduce 0%
   map 2% reduce 0%
   map 4% reduce 0%
```

```
   map 100% reduce 88%
   map 100% reduce 94%
   map 100% reduce 97%
   map 100% reduce 100%
  Job job_1701169939900_0001 completed successfully
  Output directory: hdfs:///user/root/tmp/mrjob/mrtask_a.root.20231128.113503.86
1858/step-output/0000
Counters: 57
        File Input Format Counters
                Bytes Read=5558093822
        File Output Format Counters
                Bytes Written=80
        File System Counters
                FILE: Number of bytes read=250627226
                FILE: Number of bytes written=517136853
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5558099799
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=80
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=184
                HDFS: Number of write operations=22
        Job Counters
                Data-local map tasks=38
                Killed map tasks=1
                Killed reduce tasks=2
                Launched map tasks=43
                Launched reduce tasks=13
                Rack-local map tasks=5
                Total megabyte-milliseconds taken by all map tasks=3260161536
                Total megabyte-milliseconds taken by all reduce tasks=3635401728
                Total time spent by all map tasks (ms)=2122501
                Total time spent by all maps in occupied slots (ms)=101880048
                Total time spent by all reduce tasks (ms)=1183399
                Total time spent by all reduces in occupied slots (ms)=113606304
                Total vcore-milliseconds taken by all map tasks=2122501
                Total vcore-milliseconds taken by all reduce tasks=1183399
        Map-Reduce Framework
                CPU time spent (ms)=2155850
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=9193
                Input split bytes=5977
                Map input records=58982297
                Map output bytes=830756102
```

```
                Map input records=58982297
                Map output bytes=830756102
                Map output materialized bytes=250629677
                Map output records=58982291
                Merged Map outputs=473
                Peak Map Physical memory (bytes)=659681280
                Peak Map Virtual memory (bytes)=3415584768
                Peak Reduce Physical memory (bytes)=6765871104
                Peak Reduce Virtual memory (bytes)=10452545536
                Physical memory (bytes) snapshot=28536864768
                Reduce input groups=2
                Reduce input records=58982291
                Reduce output records=2
                Reduce shuffle bytes=250629677
                Shuffled Maps =473
                Spilled Records=117964582
                Total committed heap usage (bytes)=26642743296
                Virtual memory (bytes) snapshot=181508653056
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Running step 2 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-2.1.jar] /tmp/s
treamjob3596698174580255324.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_
1701169939900_0002
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 11
  number of splits:37
  Submitting tokens for job: job_1701169939900_0002
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1701169939900_0002
```

```
  Submitted application application_1701169939900_0002
  The url to track the job: http://ip-172-31-27-72.ec2.internal:20888/proxy/appl
ication_1701169939900_0002/
  Running job: job_1701169939900_0002
  Job job_1701169939900_0002 running in uber mode : false
   map 0% reduce 0%
   map 5% reduce 0%
   map 8% reduce 0%
```

```
  map 100% reduce 91%
  map 100% reduce 100%
 Job job_1701169939900_0002 completed successfully
 Output directory: hdfs:///user/root/tripdata/out_mrtask_a/
Counters: 56
        File Input Format Counters
                Bytes Read=574
        File Output Format Counters
                Bytes Written=22
        File System Counters
                FILE: Number of bytes read=299
                FILE: Number of bytes written=14118936
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=7123
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=22
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=166
                HDFS: Number of write operations=22
        Job Counters
                Data-local map tasks=18
                Launched map tasks=37
                Launched reduce tasks=11
                Other local map tasks=9
                Rack-local map tasks=10
                Total megabyte-milliseconds taken by all map tasks=437735424
                Total megabyte-milliseconds taken by all reduce tasks=278286336
                Total time spent by all map tasks (ms)=284984
                Total time spent by all maps in occupied slots (ms)=13679232
                Total time spent by all reduce tasks (ms)=90588
                Total time spent by all reduces in occupied slots (ms)=8696448
                Total vcore-milliseconds taken by all map tasks=284984
                Total vcore-milliseconds taken by all reduce tasks=90588
        Map-Reduce Framework
                CPU time spent (ms)=65610
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=7380
                Input split bytes=6549
                Map input records=2
                Map output bytes=80
                Map output materialized bytes=6596
                Map output records=2
                Merged Map outputs=407
                Peak Map Physical memory (bytes)=564854784
```

```
                Peak Map Physical memory (bytes)=564854784
                Peak Map Virtual memory (bytes)=3101667328
                Peak Reduce Physical memory (bytes)=316575744
                Peak Reduce Virtual memory (bytes)=4423585792
                Physical memory (bytes) snapshot=21876469760
                Reduce input groups=1
                Reduce input records=2
                Reduce output records=1
                Reduce shuffle bytes=6596
                Shuffled Maps =407
                Spilled Records=4
                Total committed heap usage (bytes)=20449853440
                Virtual memory (bytes) snapshot=162664157184
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/root/tripdata/out_mrtask_a/
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/mrtask_a.root.20231128.
113503.861858...
Removing temp directory /tmp/mrtask_a.root.20231128.113503.861858...
[root@ip-172-31-27-72 tripdata]#
```

## Question 2. Which pickup location generates the most revenue? [mrtask_b.py]

**Command used:**

```
python /home/hadoop/tasks/mrtask_b.py -r hadoop
hdfs:///user/root/tripdata/*.csv --output-dir
/user/root/tripdata/out_mrtask_b/
```

Output of script 2:

```
[root@ip-172-31-27-72 tripdata]# hadoop fs -cat /user/root/tripdata/out_mrtask_b
/*
"132"   77196812.23979996
[root@ip-172-31-27-72 tripdata]#
```

## Screenshot of execution of script 2:

```
  map 100% reduce 98%
  map 100% reduce 99%
  map 100% reduce 100%
 Job job_1701169939900_0003 completed successfully
 Output directory: hdfs:///user/root/tmp/mrjob/mrtask_b.root.20231128.115313.22
6920/step-output/0000
Counters: 57
        File Input Format Counters
                Bytes Read=5558093822
        File Output Format Counters
                Bytes Written=8369
        File System Counters
                FILE: Number of bytes read=234858289
                FILE: Number of bytes written=486487192
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5558099799
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=8369
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=184
                HDFS: Number of write operations=22
        Job Counters
                Data-local map tasks=31
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=43
                Launched reduce tasks=12
                Rack-local map tasks=12
                Total megabyte-milliseconds taken by all map tasks=3028428288
                Total megabyte-milliseconds taken by all reduce tasks=2361787392
                Total time spent by all map tasks (ms)=1971633
                Total time spent by all maps in occupied slots (ms)=94638384
                Total time spent by all reduce tasks (ms)=768811
                Total time spent by all reduces in occupied slots (ms)=73805856
                Total vcore-milliseconds taken by all map tasks=1971633
                Total vcore-milliseconds taken by all reduce tasks=768811
        Map-Reduce Framework
                CPU time spent (ms)=1893110
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=9396
                Input split bytes=5977
                Map input records=58982297
                Map output bytes=642695664
                Map output materialized bytes=235748982
```

```
                Map output materialized bytes=235748982
                Map output records=58982291
                Merged Map outputs=473
                Peak Map Physical memory (bytes)=635793408
                Peak Map Virtual memory (bytes)=3347116032
                Peak Reduce Physical memory (bytes)=488685568
                Peak Reduce Virtual memory (bytes)=4737605632
                Physical memory (bytes) snapshot=27402612736
                Reduce input groups=264
                Reduce input records=58982291
                Reduce output records=264
                Reduce shuffle bytes=235748982
                Shuffled Maps =473
                Spilled Records=117964582
                Total committed heap usage (bytes)=25094520832
                Virtual memory (bytes) snapshot=181335953408
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Running step 2 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-2.1.jar] /tmp/s
treamjob4664139953376333306.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_
1701169939900_0004
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 11
  number of splits:30
  Submitting tokens for job: job_1701169939900_0004
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1701169939900_0004
  The url to track the job: http://ip-172-31-27-72.ec2.internal:20888/proxy/appl
ication_1701169939900_0004/
```

```
ication_1701169939900_0004/
  Running job: job_1701169939900_0004
  Job job_1701169939900_0004 running in uber mode : false
   map 0% reduce 0%
   map 7% reduce 0%
   map 23% reduce 0%
   map 40% reduce 0%
   map 50% reduce 0%
   map 57% reduce 0%
   map 63% reduce 0%
   map 67% reduce 0%
   map 70% reduce 0%
   map 73% reduce 0%
   map 77% reduce 0%
   map 87% reduce 2%
   map 93% reduce 2%
   map 93% reduce 5%
   map 97% reduce 5%
   map 100% reduce 6%
   map 100% reduce 18%
   map 100% reduce 36%
   map 100% reduce 45%
   map 100% reduce 55%
   map 100% reduce 64%
   map 100% reduce 82%
   map 100% reduce 91%
   map 100% reduce 100%
  Job job_1701169939900_0004 completed successfully
  Output directory: hdfs:///user/root/tripdata/out_mrtask_b/
Counters: 56
        File Input Format Counters
                Bytes Read=13581
        File Output Format Counters
                Bytes Written=24
        File System Counters
                FILE: Number of bytes read=5356
                FILE: Number of bytes written=12070747
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=18891
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=24
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=145
                HDFS: Number of write operations=22
        Job Counters
                Data-local map tasks=22
```

```
                Killed reduce tasks=1
                Launched map tasks=30
                Launched reduce tasks=11
                Rack-local map tasks=8
                Total megabyte-milliseconds taken by all map tasks=348281856
                Total megabyte-milliseconds taken by all reduce tasks=282107904
                Total time spent by all map tasks (ms)=226746
                Total time spent by all maps in occupied slots (ms)=10883808
                Total time spent by all reduce tasks (ms)=91832
                Total time spent by all reduces in occupied slots (ms)=8815872
                Total vcore-milliseconds taken by all map tasks=226746
                Total vcore-milliseconds taken by all reduce tasks=91832
        Map-Reduce Framework
                CPU time spent (ms)=57790
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=5712
                Input split bytes=5310
                Map input records=264
                Map output bytes=8369
                Map output materialized bytes=11837
                Map output records=264
                Merged Map outputs=330
                Peak Map Physical memory (bytes)=562794496
                Peak Map Virtual memory (bytes)=3097354240
                Peak Reduce Physical memory (bytes)=301613056
                Peak Reduce Virtual memory (bytes)=4437692416
                Physical memory (bytes) snapshot=18094403584
                Reduce input groups=1
                Reduce input records=264
                Reduce output records=1
                Reduce shuffle bytes=11837
                Shuffled Maps =330
                Spilled Records=528
                Total committed heap usage (bytes)=16774070272
                Virtual memory (bytes) snapshot=141048786944
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/root/tripdata/out_mrtask_b/
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/mrtask_b.root.20231128.
115313.226920...
Removing temp directory /tmp/mrtask_b.root.20231128.115313.226920...
```

```
[root@ip-172-31-27-72 tripdata]# python /home/hadoop/tasks/mrtask_b.py -r hadoop
 hdfs:///user/root/tripdata/*.csv --output-dir /user/root/tripdata/out_mrtask_b/
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /bin/hadoop
Using Hadoop version 3.3.3
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.root.20231128.115313.226920
uploading working dir files to hdfs:///user/root/tmp/mrjob/mrtask_b.root.2023112
8.115313.226920/files/wd...
Copying other local files to hdfs:///user/root/tmp/mrjob/mrtask_b.root.20231128.
115313.226920/files/
Running step 1 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-2.1.jar] /tmp/s
treamjob4907924884761735041.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_
1701169939900_0003
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 6
  Adding a new node: /default-rack/172.31.29.89:9866
  Adding a new node: /default-rack/172.31.21.145:9866
  number of splits:43
  Submitting tokens for job: job_1701169939900_0003
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1701169939900_0003
  The url to track the job: http://ip-172-31-27-72.ec2.internal:20888/proxy/appl
ication_1701169939900_0003/
  Running job: job_1701169939900_0003
  Job job_1701169939900_0003 running in uber mode : false
   map 0% reduce 0%
   map 1% reduce 0%
   map 4% reduce 0%
```

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

**Command used:**

```
python /home/hadoop/tasks/mrtask_c.py -r hadoop
hdfs:///user/root/tripdata/*.csv --output-dir
/user/root/tripdata/out_mrtask_c/
```

Output of script 3:

```
[root@ip-172-31-27-72 tripdata]# hadoop fs -cat /user/root/tripdata/out_mrtask_c
/*
"1"     39754212
"2"     18832370
"3"     306912
"4"     88794
"5"     3
[root@ip-172-31-27-72 tripdata]#
```

## Screenshot of execution of script 3:

```
[root@ip-172-31-27-72 tripdata]# python /home/hadoop/tasks/mrtask_c.py -r hadoop
 hdfs:///user/root/tripdata/*.csv --output-dir /user/root/tripdata/out_mrtask_c/
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /bin/hadoop
Using Hadoop version 3.3.3
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.root.20231128.120420.450706
uploading working dir files to hdfs:///user/root/tmp/mrjob/mrtask_c.root.2023112
8.120420.450706/files/wd...
Copying other local files to hdfs:///user/root/tmp/mrjob/mrtask_c.root.20231128.
120420.450706/files/
Running step 1 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-2.1.jar] /tmp/s
treamjob8350865417553021109.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_
1701169939900_0005
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 6
  Adding a new node: /default-rack/172.31.29.89:9866
  Adding a new node: /default-rack/172.31.21.145:9866
  number of splits:43
  Submitting tokens for job: job_1701169939900_0005
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1701169939900_0005
  The url to track the job: http://ip-172-31-27-72.ec2.internal:20888/proxy/appl
ication_1701169939900_0005/
  Running job: job_1701169939900_0005
  Job job_1701169939900_0005 running in uber mode : false
   map 0% reduce 0%
   map 2% reduce 0%
   map 4% reduce 0%
   map 5% reduce 0%
```

```
  map 100% reduce 97%
  map 100% reduce 100%
 Job job_1701169939900_0005 completed successfully
 Output directory: hdfs:///user/root/tmp/mrjob/mrtask_c.root.20231128.120420.450706/step-output/0000
Counters: 57
        File Input Format Counters
                Bytes Read=5558093822
        File Output Format Counters
                Bytes Written=93
        File System Counters
                FILE: Number of bytes read=22221206
                FILE: Number of bytes written=60329840
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5558099799
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=93
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=184
                HDFS: Number of write operations=22
        Job Counters
                Data-local map tasks=31
                Killed map tasks=1
                Killed reduce tasks=3
                Launched map tasks=43
                Launched reduce tasks=13
                Rack-local map tasks=12
                Total megabyte-milliseconds taken by all map tasks=2758479360
                Total megabyte-milliseconds taken by all reduce tasks=2324852736
                Total time spent by all map tasks (ms)=1795885
                Total time spent by all maps in occupied slots (ms)=86202480
                Total time spent by all reduce tasks (ms)=756788
                Total time spent by all reduces in occupied slots (ms)=72651648
                Total vcore-milliseconds taken by all map tasks=1795885
                Total vcore-milliseconds taken by all reduce tasks=756788
        Map-Reduce Framework
                CPU time spent (ms)=1679040
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=9249
                Input split bytes=5977
                Map input records=58982297
                Map output bytes=353893746
                Map output materialized bytes=22228704
                Map output records=58982291
```

```
                Map output records=58982291
                Merged Map outputs=473
                Peak Map Physical memory (bytes)=647135232
                Peak Map Virtual memory (bytes)=3383635968
                Peak Reduce Physical memory (bytes)=854114304
                Peak Reduce Virtual memory (bytes)=4680687616
                Physical memory (bytes) snapshot=27312644096
                Reduce input groups=5
                Reduce input records=58982291
                Reduce output records=5
                Reduce shuffle bytes=22228704
                Shuffled Maps =473
                Spilled Records=117964582
                Total committed heap usage (bytes)=25097142272
                Virtual memory (bytes) snapshot=181310857216
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Running step 2 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-2.1.jar] /tmp/s
treamjob3590649815195869015.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_
1701169939900_0006
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 11
  number of splits:38
  Submitting tokens for job: job_1701169939900_0006
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1701169939900_0006
  The url to track the job: http://ip-172-31-27-72.ec2.internal:20888/proxy/appl
ication_1701169939900_0006/
  Running job: job_1701169939900_0006
```

```
  Running job: job_1701169939900_0006
  Job job_1701169939900_0006 running in uber mode : false
   map 0% reduce 0%
   map 5% reduce 0%
```

```
   map 100% reduce 91%
   map 100% reduce 100%
  Job job_1701169939900_0006 completed successfully
  Output directory: hdfs:///user/root/tripdata/out_mrtask_c/
Counters: 57
        File Input Format Counters
                Bytes Read=341
        File Output Format Counters
                Bytes Written=53
        File System Counters
                FILE: Number of bytes read=307
                FILE: Number of bytes written=14413191
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=7067
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=53
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=169
                HDFS: Number of write operations=22
        Job Counters
                Data-local map tasks=19
                Killed reduce tasks=1
                Launched map tasks=38
                Launched reduce tasks=11
                Other local map tasks=6
                Rack-local map tasks=13
                Total megabyte-milliseconds taken by all map tasks=449528832
                Total megabyte-milliseconds taken by all reduce tasks=300011520
                Total time spent by all map tasks (ms)=292662
                Total time spent by all maps in occupied slots (ms)=14047776
                Total time spent by all reduce tasks (ms)=97660
                Total time spent by all reduces in occupied slots (ms)=9375360
                Total vcore-milliseconds taken by all map tasks=292662
                Total vcore-milliseconds taken by all reduce tasks=97660
        Map-Reduce Framework
                CPU time spent (ms)=65040
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=7535
                Input split bytes=6726
                Map input records=5
                Map output bytes=93
                Map output materialized bytes=6791
                Map output records=5
                Merged Map outputs=418
```

```
                Shuffled Maps =418
                Spilled Records=10
                Total committed heap usage (bytes)=20659044352
                Virtual memory (bytes) snapshot=165596667904
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/root/tripdata/out_mrtask_c/
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/mrtask_c.root.20231128.
120420.450706...
Removing temp directory /tmp/mrtask_c.root.20231128.120420.450706...
[root@ip-172-31-27-72 tripdata]#
```

## Question 4. What is the average trip time for different pickup locations? [mrtask_d.py]

**Command used:**

```
python /home/hadoop/tasks/mrtask_d.py -r hadoop
hdfs:///user/root/tripdata/*.csv --output-dir
/user/root/tripdata/out_mrtask_d/
```

## Output of script 4:

```
[root@ip-172-31-27-72 tripdata]# hadoop fs -cat /user/root/tripdata/out_mrtask_d
/*
"9"     84.49011173184354
"109"   60.776153846153846
"139"   59.65663461538462
"38"    50.988758169934655
"10"    49.76854575312459
"215"   48.546164383561575
"219"   45.452668701254666
"132"   43.77028865677146
"187"   38.9908695652174
"2"     38.16673913043478
"138"   37.32248906352792
"130"   35.166776359973134
"93"    33.988361131254464
"176"   30.77214285714286
"180"   30.03792710706149
"222"   29.712380952380936
"124"   28.92710823909534
"216"   28.41854150092807
"28"    28.17357612267257
"203"   27.16497991967872
"122"   27.16340000000001
"155"   27.028484848484837
"208"   26.576567901234576
"70"    25.207638392857064
"205"   25.105694444444435
"30"    25.10047619047619
"194"   24.988401169875782
"31"    24.915381679389302
"55"    24.754764705882344
"154"   24.453961038961026
"12"    24.358714156742952
"104"   23.62
"253"   22.277801418439726
"102"   22.112343749999994
"16"    21.921933333333314
"22"    21.920131752305675
"29"    21.90638941398865
"261"   21.84983993270276
"76"    21.723183773816853
"3"     21.662666666666667
"88"    21.651551719058492
"212"   21.123416230366495
"91"    20.863868698710462
"195"   20.830272185993728
"73"    20.808068965517243
"218"   20.67360406091371
```

```
"157"   20.50436126629425
"21"    20.498062500000017
"258"   20.483842173350588
"87"    20.44242292349143
"32"    20.141291866028716
"56"    20.112816831683134
"160"   20.065231092436992
"105"   19.978039215686277
"77"    19.923120567375882
"252"   19.827400000000008
"85"    19.751288088642653
"72"    19.67538345864663
"13"    19.50199386458798
"35"    19.435640686922053
"156"   19.40417475728155
"67"    19.37815584415583
"94"    19.335820895522385
"117"   19.317441860465117
"190"   19.285978056426284
"63"    19.201828309305366
"177"   19.160654022236766
"209"   19.135253621598896
"254"   19.044217877094965
"185"   18.767665094339616
"126"   18.56175965665236
"101"   18.519558823529415
"150"   18.49263736263736
"45"    18.465255226500695
"165"   18.458102108768028
"192"   18.249200652528547
"39"    18.23428713858421
"135"   18.103647859922184
"33"    18.09514579196476
"210"   17.9481971830986
"95"    17.861194558756317
"66"    17.815410656119045
"196"   17.65802781289507
"46"    17.60447368421053
"217"   17.581780012771365
"92"    17.523125222975413
"8"     17.490223463687144
"199"   17.451904761904764
"52"    17.332517341995178
"232"   17.193648916192842
"11"    17.169146341463424
"53"    17.147376623376626
"54"    17.03241042345278
"65"    16.96084216995242
```

```
"230"   16.93072835148817
"235"   16.9205707395498 52
"36"    16.873506835937558
"14"    16.866985547143834
"231"   16.84408361850162
"244"   16.833110001584874
"59"    16.816956521739133
"186"   16.81188624340465
"144"   16.80378325825834
"211"   16.764863380866014
"250"   16.7627138643068
"69"    16.737178217821807
"149"   16.72867796610171
"184"   16.715000000000003
"148"   16.66852167741667
"158"   16.619858904059114
"133"   16.605759265332885
"47"    16.590380952380947
"226"   16.553299039197185
"191"   16.42862559241705
"213"   16.387427385892103
"243"   16.34682820776346
"255"   16.321391988947063
"119"   16.32091586794464
"71"    16.31415123456789
"51"    16.277214285714283
"40"    16.212707273477154
"202"   16.211691743119285
"188"   16.13376315114189
"37"    16.128151186335057
"256"   16.111510198998104
"247"   16.078678936957083
"89"    15.994474166361298
"172"   15.978333333333328
"114"   15.921301255463838
"125"   15.899443291165806
"83"    15.877737104825297
"34"    15.864772727272724
"163"   15.862749532466587
"181"   15.849556176575732
"134"   15.84853457172344
"228"   15.833866762177594
"169"   15.798557858376505
"96"    15.732589928057557
"248"   15.719237472766881
"246"   15.718103380361832
"97"    15.708177323719424
"197"   15.662647058823545
```

```
"116"   15.637838088900216
"43"    15.613800756480765
"161"   15.588042766814533
"127"   15.56611605053813
"123"   15.514745958429556
"260"   15.507387337876347
"25"    15.483474098401302
"100"   15.451721719342709
"164"   15.42455613905816
"257"   15.4200867678959
"61"    15.336179751922948
"68"    15.325826679040274
"62"    15.302478902953615
"233"   15.269121152726001
"241"   15.209324546952226
"79"    15.19047779287196
"189"   15.166341057016545
"259"   15.158783382789322
"80"    15.157749331094921
"162"   15.10368900322217
"146"   15.048250149592477
"234"   15.0029415800564
"113"   14.99985117807177
"4"     14.99466014250498
"167"   14.993400431344323
"249"   14.940340757152194
"223"   14.933763179027652
"225"   14.926515780730877
"128"   14.898110831234247
"50"    14.86682801948296
"166"   14.844408927594904
"264"   14.775025363925117
"48"    14.690363703303364
"170"   14.685757417280383
"173"   14.653154648956383
"82"    14.641144189796023
"227"   14.551109677419351
"15"    14.541926605504585
"179"   14.502852709970938
"86"    14.450923076923077
"112"   14.446122047244085
"57"    14.42856589147286
"221"   14.373776824034326
"193"   14.370107348750974
"18"    14.326460431654668
"129"   14.258695788275531
"242"   14.231503340757241
"108"   14.208398058252413
```

```
"159"   14.204903417533455
"90"    14.17738139525186
"107"   14.170104175145891
"152"   14.159461091601324
"121"   14.144716049382707
"106"   14.142370135436215
"131"   14.111790633608813
"20"    14.108927680797999
"115"   14.100331950207469
"240"   14.076011235955052
"17"    14.029979363527708
"49"    13.988324643584429
"182"   13.937709923664128
"140"   13.928652733926045
"142"   13.862746518091084
"120"   13.808186968838521
"137"   13.785312000193205
"224"   13.713110282076572
"206"   13.607142857142854
"153"   13.562914653784219
"24"    13.555460474029992
"64"    13.288802816901407
"229"   13.281867582939789
"7"     13.264168184183283
"198"   13.249644917887226
"151"   13.214629242110432
"147"   13.15429389312978
"239"   13.126087703509167
"143"   13.114752633855705
"174"   13.067607913669065
"81"    13.033366834170854
"42"    12.95989324091511
"78"    12.958535127055324
"41"    12.956672551849087
"75"    12.912285947157272
"238"   12.90728719063254
"171"   12.873103448275865
"168"   12.85559367906407
"262"   12.843554106013553
"200"   12.83173708920189
"175"   12.795576923076926
"145"   12.773812445945318
"23"    12.765217391304342
"236"   12.747241449978176
"99"    12.728
"74"    12.628689617946044
"60"    12.572266666666664
"19"    12.450671641791043
```

```
"237"   12.365681640984109
"141"   12.262030708803833
"118"   12.259714285714285
"98"    12.255823293172696
"263"   12.159215394102256
"183"   12.080416666666672
"220"   12.070856031128388
"26"    11.983158301158312
"251"   11.874774193548383
"84"    11.774000000000003
"44"    11.740714285714287
"111"   11.634569536423838
"5"     11.444615384615386
"136"   11.392957359009628
"214"   9.997272727272728
"201"   9.882253521126758
"265"   9.502463293040948
"27"    8.77909090909091
"245"   8.677000000000001
"207"   8.466118649416204
"1"     8.186395126612483
"6"     7.502000000000002
"178"   6.8092463442069775
"58"    6.4692307692307685
"204"   3.5514285714285703
"110"   3.18
[root@ip-172-31-27-72 tripdata]#
```

## Screenshot of execution of script 4:

```
[root@ip-172-31-27-72 tripdata]# python /home/hadoop/tasks/mrtask_d.py -r hadoop
 hdfs:///user/root/tripdata/*.csv --output-dir /user/root/tripdata/out_mrtask_d/
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /bin/hadoop
Using Hadoop version 3.3.3
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.root.20231128.121654.036121
uploading working dir files to hdfs:///user/root/tmp/mrjob/mrtask_d.root.2023112
8.121654.036121/files/wd...
Copying other local files to hdfs:///user/root/tmp/mrjob/mrtask_d.root.20231128.
121654.036121/files/
Running step 1 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-2.1.jar] /tmp/s
treamjob3554828512971156106.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_
1701169939900_0007
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 6
  Adding a new node: /default-rack/172.31.29.89:9866
  Adding a new node: /default-rack/172.31.21.145:9866
  number of splits:43
  Submitting tokens for job: job_1701169939900_0007
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1701169939900_0007
  The url to track the job: http://ip-172-31-27-72.ec2.internal:20888/proxy/appl
ication_1701169939900_0007/
  Running job: job_1701169939900_0007
  Job job_1701169939900_0007 running in uber mode : false
   map 0% reduce 0%
   map 1% reduce 0%
   map 2% reduce 0%
```

```
   map 100% reduce 99%
   map 100% reduce 100%
  Job job_1701169939900_0007 completed successfully
  Output directory: hdfs:///user/root/tmp/mrjob/mrtask_d.root.20231128.121654.03
6121/step-output/0000
Counters: 56
        File Input Format Counters
                Bytes Read=5558093822
        File Output Format Counters
                Bytes Written=8489
        File System Counters
                FILE: Number of bytes read=293505303
                FILE: Number of bytes written=604331356
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5558099799
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=8489
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=184
                HDFS: Number of write operations=22
        Job Counters
                Data-local map tasks=31
                Killed map tasks=1
                Launched map tasks=43
                Launched reduce tasks=12
                Rack-local map tasks=12
                Total megabyte-milliseconds taken by all map tasks=10294150656
                Total megabyte-milliseconds taken by all reduce tasks=5784950784
                Total time spent by all map tasks (ms)=6701921
                Total time spent by all maps in occupied slots (ms)=321692208
                Total time spent by all reduce tasks (ms)=1883122
                Total time spent by all reduces in occupied slots (ms)=180779712
                Total vcore-milliseconds taken by all map tasks=6701921
                Total vcore-milliseconds taken by all reduce tasks=1883122
        Map-Reduce Framework
                CPU time spent (ms)=6447190
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=11459
                Input split bytes=5977
                Map input records=58982297
                Map output bytes=955309824
                Map output materialized bytes=294946124
                Map output records=58982291
                Merged Map outputs=473
```

```
                Merged Map outputs=473
                Peak Map Physical memory (bytes)=667680768
                Peak Map Virtual memory (bytes)=3383492608
                Peak Reduce Physical memory (bytes)=892526592
                Peak Reduce Virtual memory (bytes)=5095993344
                Physical memory (bytes) snapshot=28520153088
                Reduce input groups=264
                Reduce input records=58982291
                Reduce output records=264
                Reduce shuffle bytes=294946124
                Shuffled Maps =473
                Spilled Records=117964582
                Total committed heap usage (bytes)=25894584320
                Virtual memory (bytes) snapshot=181414903808
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Running step 2 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-2.1.jar] /tmp/s
treamjob6258520715673116471.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_
1701169939900_0008
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 11
  number of splits:30
  Submitting tokens for job: job_1701169939900_0008
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1701169939900_0008
  The url to track the job: http://ip-172-31-27-72.ec2.internal:20888/proxy/appl
ication_1701169939900_0008/
  Running job: job_1701169939900_0008
  Job job_1701169939900_0008 running in uber mode : false
```

```
  Job job_1701169939900_0008 running in uber mode : false
   map 0% reduce 0%
   map 7% reduce 0%
   map 27% reduce 0%
```

```
   map 100% reduce 82%
   map 100% reduce 91%
   map 100% reduce 100%
  Job job_1701169939900_0008 completed successfully
  Output directory: hdfs:///user/root/tripdata/out_mrtask_d/
Counters: 56
        File Input Format Counters
                Bytes Read=13763
        File Output Format Counters
                Bytes Written=6377
        File System Counters
                FILE: Number of bytes read=5989
                FILE: Number of bytes written=12071893
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=19073
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=6377
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=145
                HDFS: Number of write operations=22
        Job Counters
                Data-local map tasks=22
                Killed map tasks=1
                Launched map tasks=30
                Launched reduce tasks=11
                Rack-local map tasks=8
                Total megabyte-milliseconds taken by all map tasks=343414272
                Total megabyte-milliseconds taken by all reduce tasks=252503040
                Total time spent by all map tasks (ms)=223577
                Total time spent by all maps in occupied slots (ms)=10731696
                Total time spent by all reduce tasks (ms)=82195
                Total time spent by all reduces in occupied slots (ms)=7890720
                Total vcore-milliseconds taken by all map tasks=223577
                Total vcore-milliseconds taken by all reduce tasks=82195
        Map-Reduce Framework
                CPU time spent (ms)=54350
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=5554
                Input split bytes=5310
                Map input records=264
                Map output bytes=8489
                Map output materialized bytes=12371
                Map output records=264
                Merged Map outputs=330
```

```
                Merged Map outputs=330
                Peak Map Physical memory (bytes)=555757568
                Peak Map Virtual memory (bytes)=3091533824
                Peak Reduce Physical memory (bytes)=323858432
                Peak Reduce Virtual memory (bytes)=4446572544
                Physical memory (bytes) snapshot=17985179648
                Reduce input groups=1
                Reduce input records=264
                Reduce output records=264
                Reduce shuffle bytes=12371
                Shuffled Maps =330
                Spilled Records=528
                Total committed heap usage (bytes)=17041981440
                Virtual memory (bytes) snapshot=140959744000
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/root/tripdata/out_mrtask_d/
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/mrtask_d.root.20231128.
121654.036121...
Removing temp directory /tmp/mrtask_d.root.20231128.121654.036121...
[root@ip-172-31-27-72 tripdata]#
```

## Question 5. Calculate the average tips to revenue ratio of the drivers for different pickup locations in sorted format. [mrtask_e.py]

**Command used:**

```
python /home/hadoop/tasks/mrtask_e.py -r hadoop
hdfs:///user/root/tripdata/*.csv --output-dir
/user/root/tripdata/out_mrtask_e/
```

## Output of script 5:

```
[root@ip-172-31-27-72 tripdata]# hadoop fs -cat /user/root/tripdata/out_mrtask_e
/*
"30"    0.2561075875029364
"104"   0.2000665778961385
"187"   0.17978913134704155
"109"   0.17861970356364323
"5"     0.173567645642754
"172"   0.17223686242471647
"117"   0.16546115321544105
"176"   0.15889955267208697
"201"   0.1516472849404957
"58"    0.14391826665236515
"199"   0.14024694882592248
"122"   0.13240255677287158
"138"   0.13191299589402555
"52"    0.12903601946205426
"175"   0.127723077600828
"210"   0.12696291028237985
"191"   0.12487168062809852
"87"    0.12471673086150699
"125"   0.12389238738129221
"16"    0.12347342980892857
"84"    0.12333321182517408
"13"    0.12263264632334413
"178"   0.12194781013939375
"194"   0.12129343435403825
"33"    0.12071128564964351
"162"   0.1206663239677913
"40"    0.12031726919562945
"54"    0.12031650579198322
"234"   0.12013283079340571
"249"   0.11990213646579875
"107"   0.11974753027212386
"246"   0.11961317882513706
"1"     0.1193146282170377
"23"    0.11893642866783627
"231"   0.11887862152231361
"113"   0.11854772693332671
"170"   0.11808315282085192
"79"    0.1180040681963363
"252"   0.11791966686185075
"114"   0.11737171385693342
"118"   0.1172896356117838
"66"    0.11716659766188252
"255"   0.11685699515460911
"88"    0.11647722217055936
"158"   0.11639038989030531
"184"   0.11633468693321355
```

```
"15"    0.11600468842514824
"233"   0.11587691604667817
"224"   0.11526880032060012
"148"   0.11466547302826928
"90"    0.11445563465173544
"68"    0.11443748108063767
"209"   0.11438950713872204
"143"   0.11408622408856395
"211"   0.11399355397938883
"258"   0.11397415106036057
"263"   0.11388129538365796
"161"   0.11359049234213116
"64"    0.11356366975963504
"262"   0.11338221978066348
"257"   0.11311365670211546
"181"   0.11303400885954305
"144"   0.11283013521682496
"166"   0.1127867528416522
"137"   0.11277472565908724
"76"    0.11266074016140219
"264"   0.11265242617699583
"229"   0.11256386704258736
"142"   0.11242578567603717
"265"   0.11224718082317688
"163"   0.11222164046258055
"164"   0.11217684253727223
"96"    0.11215306381581779
"189"   0.11197515428642564
"106"   0.11182299192052929
"239"   0.11135965083693283
"140"   0.111184442118051
"236"   0.11108195526238318
"123"   0.11055898094072218
"238"   0.11044770762987073
"80"    0.11027187700902219
"93"    0.11009789332101834
"237"   0.1099044245936146
"186"   0.10977067019583905
"112"   0.10939659242523707
"86"    0.10935811216524567
"25"    0.10931197115088469
"50"    0.10916652254471852
"8"     0.10862824529869558
"256"   0.10853673778730165
"215"   0.10829221671091824
"70"    0.10788380927812946
"141"   0.10768193607093385
"261"   0.10766619243087137
```

```
"10"    0.10761668826559867
"230"   0.10760970767267052
"134"   0.10749874267012811
"151"   0.10736778555022615
"4"     0.10684873200923953
"34"    0.10658808178106388
"48"    0.10627040728532096
"63"    0.10608349633094753
"24"    0.10590891911111136
"36"    0.10580489050473851
"2"     0.10551996408687955
"65"    0.10543895761077816
"180"   0.10523696750375865
"43"    0.10492270664827301
"216"   0.10479778546735245
"195"   0.10463665842928088
"130"   0.10459971957672846
"150"   0.10406209924668974
"132"   0.10404480517984674
"101"   0.10391200650224484
"200"   0.10311919266153378
"221"   0.10296610645201089
"67"    0.10270469503747023
"28"    0.10225113334394163
"128"   0.10195419121098397
"100"   0.10176112040801057
"111"   0.10175458286569396
"190"   0.10147578909371589
"157"   0.10078849415880564
"35"    0.10022554243576758
"219"   0.10013130970242896
"71"    0.09976058424195912
"49"    0.09943300939917539
"45"    0.09934274516885801
"213"   0.09930916492713546
"37"    0.09921178378981346
"56"    0.09883778991554207
"97"    0.09860378966747259
"115"   0.09805645544907098
"218"   0.09787185841442576
"228"   0.09725404075366942
"202"   0.09722093931447215
"77"    0.09702969688085904
"223"   0.09657319511941946
"39"    0.09651707726856389
"243"   0.09649751817938544
"160"   0.09647705819251878
"57"    0.09639556671648793
```

```
"155"   0.09633375755470464
"145"   0.09530064643701927
"55"    0.09470551332740063
"222"   0.09461551127943343
"253"   0.09444941923267368
"133"   0.09428655857404224
"244"   0.094131612382101
"102"   0.09286325213174144
"139"   0.09244706142211928
"61"    0.09216038290045477
"31"    0.09122118342582683
"135"   0.0911299209238633
"154"   0.09108097210043571
"121"   0.09104834906722979
"197"   0.09057550505117994
"14"    0.09044594848522369
"9"     0.09026115108356447
"116"   0.09022031542003194
"75"    0.09013747412649323
"41"    0.0901371077650071
"89"    0.08958574112048633
"51"    0.08952152479566286
"203"   0.08922376129095648
"72"    0.08865379692027077
"146"   0.0885365814566892
"127"   0.08838814791010617
"12"    0.08816912666412209
"245"   0.08812313347116932
"27"    0.0880988704482764
"198"   0.08806313330584384
"156"   0.08792537666968149
"98"    0.08788658752859323
"259"   0.08717567117240305
"196"   0.08616922680963858
"108"   0.08561073480894194
"226"   0.0855443843084591
"179"   0.08477276212071948
"124"   0.08450445139029453
"232"   0.08364178523168092
"206"   0.08357280929682975
"131"   0.08338783482815317
"95"    0.0833569398816512
"152"   0.08319834814824395
"17"    0.08317536901457594
"120"   0.08303252748719832
"205"   0.08302284089547898
"251"   0.08299641950457216
"21"    0.08297404683405173
```

```
"105"   0.08289275441720313
"225"   0.08262074885723772
"46"    0.08234332088746261
"149"   0.08233859572151536
"81"    0.08090785981257764
"227"   0.08060571925597004
"171"   0.08052993314534178
"19"    0.0794349494311421
"7"     0.07831986411646918
"91"    0.07770689984709575
"214"   0.07745375602878155
"53"    0.07740291031997883
"208"   0.07718716885014376
"62"    0.07661527162660337
"177"   0.07641696352884633
"73"    0.07567600767754314
"220"   0.07548055451382524
"20"    0.07539779360644307
"185"   0.075309699831948
"212"   0.074599916120346
"188"   0.07438531115225959
"217"   0.0743390392253003
"207"   0.07416425281279072
"247"   0.07294934365908463
"126"   0.07223106849013453
"74"    0.07135080378578117
"242"   0.07134362315468414
"192"   0.0712438378066297
"240"   0.0710554459113095
"42"    0.07036241083605818
"260"   0.07000191673138784
"3"     0.06985350809560509
"165"   0.06797416237704348
"119"   0.06778234295102134
"11"    0.0661192351337737
"85"    0.0660285451126606
"26"    0.06577859806429268
"29"    0.06573652675700885
"92"    0.06480851959104446
"82"    0.064743588258187
"169"   0.06451074143663961
"44"    0.06402542001826439
"129"   0.06331017333927771
"22"    0.063033812614293
"174"   0.0627814335084246
"204"   0.06247952716655987
"241"   0.06142723563688878
"167"   0.06122252383032673
```

```
"32"    0.061097873277617114
"153"   0.06056373253577961
"99"    0.06008865539320309
"173"   0.060081458043702456
"183"   0.05853409985743484
"38"    0.05791427287264446
"83"    0.05736768211328874
"250"   0.05704400870673732
"248"   0.054061056074233
"18"    0.05241971650085587
"235"   0.05148549299609815
"168"   0.049620759878184505
"182"   0.04956118360347716
"78"    0.04945682121018886
"6"     0.04831818885848528
"69"    0.04770731790200153
"159"   0.04663800562314009
"193"   0.04602694182350401
"254"   0.04539094268441616
"147"   0.044604149528371
"136"   0.04090437862017819
"60"    0.04087201889401148
"47"    0.03560717179556608
"94"    0.03261240109944167
"59"    0.011954905847373635
"110"   0.0
[root@ip-172-31-27-72 tripdata]#
```

Screenshot of execution of script 5:

```
[root@ip-172-31-27-72 tripdata]# python /home/hadoop/tasks/mrtask_e.py -r hadoop
 hdfs:///user/root/tripdata/*.csv --output-dir /user/root/tripdata/out_mrtask_e/
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /bin/hadoop
Using Hadoop version 3.3.3
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.root.20231128.123945.072782
uploading working dir files to hdfs:///user/root/tmp/mrjob/mrtask_e.root.2023112
8.123945.072782/files/wd...
Copying other local files to hdfs:///user/root/tmp/mrjob/mrtask_e.root.20231128.
123945.072782/files/
Running step 1 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-2.1.jar] /tmp/s
treamjob8779132237336186774.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_
1701169939900_0009
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 6
  Adding a new node: /default-rack/172.31.29.89:9866
  Adding a new node: /default-rack/172.31.21.145:9866
  number of splits:43
  Submitting tokens for job: job_1701169939900_0009
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1701169939900_0009
  The url to track the job: http://ip-172-31-27-72.ec2.internal:20888/proxy/appl
ication_1701169939900_0009/
  Running job: job_1701169939900_0009
  Job job_1701169939900_0009 running in uber mode : false
   map 0% reduce 0%
   map 1% reduce 0%
   map 2% reduce 0%
   map 3% reduce 0%
```

```
  map 100% reduce 99%
  map 100% reduce 100%
 Job job_1701169939900_0009 completed successfully
 Output directory: hdfs:///user/root/tmp/mrjob/mrtask_e.root.20231128.123945.07
2782/step-output/0000
Counters: 57
        File Input Format Counters
                Bytes Read=5558093822
        File Output Format Counters
                Bytes Written=8814
        File System Counters
                FILE: Number of bytes read=354660814
                FILE: Number of bytes written=726228304
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5558099799
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=8814
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=184
                HDFS: Number of write operations=22
        Job Counters
                Data-local map tasks=32
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=43
                Launched reduce tasks=12
                Rack-local map tasks=11
                Total megabyte-milliseconds taken by all map tasks=3492334080
                Total megabyte-milliseconds taken by all reduce tasks=3002173440
                Total time spent by all map tasks (ms)=2273655
                Total time spent by all maps in occupied slots (ms)=109135440
                Total time spent by all reduce tasks (ms)=977270
                Total time spent by all reduces in occupied slots (ms)=93817920
                Total vcore-milliseconds taken by all map tasks=2273655
                Total vcore-milliseconds taken by all reduce tasks=977270
        Map-Reduce Framework
                CPU time spent (ms)=2398640
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=10632
                Input split bytes=5977
                Map input records=58982297
                Map output bytes=1082802766
                Map output materialized bytes=355687606
                Map output records=58982291
```

```
                Reduce input records=58982291
                Reduce output records=264
                Reduce shuffle bytes=355687606
                Shuffled Maps =473
                Spilled Records=117964582
                Total committed heap usage (bytes)=26347569152
                Virtual memory (bytes) snapshot=181559123968
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Running step 2 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-2.1.jar] /tmp/s
treamjob4929080415390647714.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_
1701169939900_0010
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 11
  number of splits:30
  Submitting tokens for job: job_1701169939900_0010
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1701169939900_0010
  The url to track the job: http://ip-172-31-27-72.ec2.internal:20888/proxy/appl
ication_1701169939900_0010/
  Running job: job_1701169939900_0010
  Job job_1701169939900_0010 running in uber mode : false
   map 0% reduce 0%
   map 7% reduce 0%
   map 30% reduce 0%
   map 40% reduce 0%
   map 67% reduce 0%
   map 67% reduce 2%
   map 73% reduce 2%
```

```
  map 100% reduce 91%
  map 100% reduce 100%
 Job job_1701169939900_0010 completed successfully
 Output directory: hdfs:///user/root/tripdata/out_mrtask_e/
Counters: 56
        File Input Format Counters
                Bytes Read=14261
        File Output Format Counters
                Bytes Written=6702
        File System Counters
                FILE: Number of bytes read=6032
                FILE: Number of bytes written=12072096
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=19571
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=6702
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=145
                HDFS: Number of write operations=22
        Job Counters
                Data-local map tasks=29
                Killed map tasks=1
                Launched map tasks=30
                Launched reduce tasks=11
                Rack-local map tasks=1
                Total megabyte-milliseconds taken by all map tasks=394547712
                Total megabyte-milliseconds taken by all reduce tasks=263033856
                Total time spent by all map tasks (ms)=256867
                Total time spent by all maps in occupied slots (ms)=12329616
                Total time spent by all reduce tasks (ms)=85623
                Total time spent by all reduces in occupied slots (ms)=8219808
                Total vcore-milliseconds taken by all map tasks=256867
                Total vcore-milliseconds taken by all reduce tasks=85623
        Map-Reduce Framework
                CPU time spent (ms)=55130
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=6822
                Input split bytes=5310
                Map input records=264
                Map output bytes=8814
                Map output materialized bytes=12549
                Map output records=264
                Merged Map outputs=330
                Peak Map Physical memory (bytes)=566157312
```

```
                Peak Map Physical memory (bytes)=566157312
                Peak Map Virtual memory (bytes)=3094192128
                Peak Reduce Physical memory (bytes)=340963328
                Peak Reduce Virtual memory (bytes)=4438728704
                Physical memory (bytes) snapshot=18166382592
                Reduce input groups=1
                Reduce input records=264
                Reduce output records=264
                Reduce shuffle bytes=12549
                Shuffled Maps =330
                Spilled Records=528
                Total committed heap usage (bytes)=17110138880
                Virtual memory (bytes) snapshot=141040582656
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/root/tripdata/out_mrtask_e/
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/mrtask_e.root.20231128.
123945.072782...
Removing temp directory /tmp/mrtask_e.root.20231128.123945.072782...
[root@ip-172-31-27-72 tripdata]#
```

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). [mrtask_f.py]

**Command used:**

```
python /home/hadoop/tasks/mrtask_f.py -r hadoop
hdfs:///user/root/tripdata/*.csv --output-dir
/user/root/tripdata/out_mrtask_f/
```

## Output of script 6:

```
[root@ip-172-31-27-72 tripdata]# hadoop fs -cat /user/root/tripdata/out_mrtask_f
/*
"01-Weekday-00" 17.200476132402844
"01-Weekday-01" 16.52274755885113
"01-Weekday-02" 16.16899252840724
"01-Weekday-03" 17.390409325713083
"01-Weekday-04" 19.81653275049372
"01-Weekday-05" 19.4436686252908
"01-Weekday-06" 14.915323806861682
"01-Weekday-07" 14.074264290473451
"01-Weekday-08" 14.409976521306454
"01-Weekday-09" 16.51671940328511
"01-Weekday-10" 15.132553022958056
"01-Weekday-11" 15.021462847959326
"01-Weekday-12" 15.053894328309841
"01-Weekday-13" 15.459587483198423
"01-Weekday-14" 15.668720572347757
"01-Weekday-15" 15.504986952778848
"01-Weekday-16" 16.941667344353913
"01-Weekday-17" 16.150073893063933
"01-Weekday-18" 15.463753409563774
"01-Weekday-19" 15.310948274051695
"01-Weekday-20" 15.434833196243318
"01-Weekday-21" 15.876239291962918
"01-Weekday-22" 16.116412701710537
"01-Weekday-23" 16.62915975751689
"01-Weekend-00" 15.492677914167835
"01-Weekend-01" 15.221754978674015
"01-Weekend-02" 19.676045388000507
"01-Weekend-03" 15.796174786765324
"01-Weekend-04" 16.9038321528958
"01-Weekend-05" 19.611099207268918
"01-Weekend-06" 19.365804005037397
"01-Weekend-07" 16.357491028245622
"01-Weekend-08" 14.26352577961805
"01-Weekend-09" 12.926096329396852
"01-Weekend-10" 13.052689692790315
"01-Weekend-11" 13.115258676736508
"01-Weekend-12" 13.589219273538811
"01-Weekend-13" 14.233717029228808
"01-Weekend-14" 14.86409574609936
"01-Weekend-15" 14.825027445549306
"01-Weekend-16" 15.28485232525307
"01-Weekend-17" 15.196178783743054
"01-Weekend-18" 14.523416795691354
"01-Weekend-19" 14.487575905195618
"01-Weekend-20" 15.300936980745705
"01-Weekend-21" 15.57767630335397
```

```
"01-Weekend-22" 15.761635314830924
"01-Weekend-23" 16.27716797074358
"02-Weekday-00" 16.624534852829573
"02-Weekday-01" 15.958691886305655
"02-Weekday-02" 15.517073988828718
"02-Weekday-03" 16.41745764348376
"02-Weekday-04" 19.747369349616502
"02-Weekday-05" 20.010079254499757
"02-Weekday-06" 15.056520042501958
"02-Weekday-07" 14.182037283449263
"02-Weekday-08" 14.36447896150126
"02-Weekday-09" 14.87336597880182
"02-Weekday-10" 15.377098231057548
"02-Weekday-11" 15.324369666207064
"02-Weekday-12" 15.37345136398676
"02-Weekday-13" 15.973985513973316
"02-Weekday-14" 16.124635719746674
"02-Weekday-15" 15.889883596977725
"02-Weekday-16" 18.478468871258872
"02-Weekday-17" 16.771236135408905
"02-Weekday-18" 15.541284613015696
"02-Weekday-19" 15.342079566199766
"02-Weekday-20" 15.455142471823336
"02-Weekday-21" 15.964644194797494
"02-Weekday-22" 17.228521316788783
"02-Weekday-23" 16.848290343595508
"02-Weekend-00" 15.717681829936232
"02-Weekend-01" 15.21391623774398
"02-Weekend-02" 15.219543293502197
"02-Weekend-03" 15.576802395222874
"02-Weekend-04" 16.925485039946157
"02-Weekend-05" 20.392564383557026
"02-Weekend-06" 18.961630042403595
"02-Weekend-07" 16.325909500452052
"02-Weekend-08" 14.037687088920901
"02-Weekend-09" 12.979614418717606
"02-Weekend-10" 13.15201125369243
"02-Weekend-11" 13.133083887380716
"02-Weekend-12" 13.77235457760831
"02-Weekend-13" 14.308972649330412
"02-Weekend-14" 15.044893032177265
"02-Weekend-15" 15.378423608160334
"02-Weekend-16" 15.298620216704524
"02-Weekend-17" 15.07999803121459
"02-Weekend-18" 14.529941792998763
"02-Weekend-19" 14.53254192476164
"02-Weekend-20" 15.427541050452229
"02-Weekend-21" 15.87560118580828
```

```
"02-Weekend-22" 15.878073723805395
"02-Weekend-23" 16.19622921105247
"03-Weekday-00" 17.706681781643397
"03-Weekday-01" 16.85254365720629
"03-Weekday-02" 16.29235268722667
"03-Weekday-03" 16.928700720694778
"03-Weekday-04" 20.194628690000524
"03-Weekday-05" 20.237903274086065
"03-Weekday-06" 15.758817970878994
"03-Weekday-07" 14.535676520917598
"03-Weekday-08" 14.946093599593581
"03-Weekday-09" 15.457095345228746
"03-Weekday-10" 15.930720655553303
"03-Weekday-11" 16.255371273141918
"03-Weekday-12" 16.469608540766203
"03-Weekday-13" 16.99947224397787
"03-Weekday-14" 17.224570655906135
"03-Weekday-15" 17.200087652649156
"03-Weekday-16" 18.69631188576578
"03-Weekday-17" 17.46524914055405
"03-Weekday-18" 16.6087969286072
"03-Weekday-19" 16.008619663460646
"03-Weekday-20" 15.865122145199464
"03-Weekday-21" 16.34653437981351
"03-Weekday-22" 16.84017175676268
"03-Weekday-23" 17.338514205585827
"03-Weekend-00" 16.041376225879947
"03-Weekend-01" 15.591188453014844
"03-Weekend-02" 15.379495890597605
"03-Weekend-03" 15.819462340137685
"03-Weekend-04" 17.30104406164219
"03-Weekend-05" 20.42549115801075
"03-Weekend-06" 19.884900187054683
"03-Weekend-07" 16.927217978163593
"03-Weekend-08" 14.600109987753768
"03-Weekend-09" 13.590348187835085
"03-Weekend-10" 13.509171636698433
"03-Weekend-11" 13.657631392241237
"03-Weekend-12" 14.079435136445278
"03-Weekend-13" 14.683893418234709
"03-Weekend-14" 15.424649768082656
"03-Weekend-15" 15.751678191585011
"03-Weekend-16" 15.701753990371294
"03-Weekend-17" 15.406489686081883
"03-Weekend-18" 14.795767258678966
"03-Weekend-19" 14.778656228553157
"03-Weekend-20" 15.411752798782919
"03-Weekend-21" 15.823254785309633
```

```
"03-Weekend-22" 16.548258367850696
"03-Weekend-23" 17.150321650491097
"04-Weekday-00" 18.51600196965465
"04-Weekday-01" 17.793384146651995
"04-Weekday-02" 17.199010607801274
"04-Weekday-03" 17.494598600349473
"04-Weekday-04" 20.578910667437057
"04-Weekday-05" 20.064185981232256
"04-Weekday-06" 15.382159119494295
"04-Weekday-07" 14.601019836212725
"04-Weekday-08" 14.864007056320014
"04-Weekday-09" 15.415372389452688
"04-Weekday-10" 15.88396334815962
"04-Weekday-11" 16.834124885428658
"04-Weekday-12" 16.317079540287406
"04-Weekday-13" 16.792758980696814
"04-Weekday-14" 17.066245237245802
"04-Weekday-15" 16.801182616997433
"04-Weekday-16" 18.267202421962605
"04-Weekday-17" 17.463488647217254
"04-Weekday-18" 16.28539322434687
"04-Weekday-19" 15.916875126784577
"04-Weekday-20" 15.949128335669775
"04-Weekday-21" 16.491005249441166
"04-Weekday-22" 17.00451764904585
"04-Weekday-23" 17.649619974192703
"04-Weekend-00" 16.21626103253723
"04-Weekend-01" 15.713164564227124
"04-Weekend-02" 15.41493933329952
"04-Weekend-03" 15.885438294891317
"04-Weekend-04" 17.3417382220788
"04-Weekend-05" 20.781433096337523
"04-Weekend-06" 19.674270648376968
"04-Weekend-07" 16.677376477506613
"04-Weekend-08" 14.592686115871663
"04-Weekend-09" 13.674913015697646
"04-Weekend-10" 13.888303803837
"04-Weekend-11" 14.233907112388O2
"04-Weekend-12" 14.774475374725903
"04-Weekend-13" 15.583307987749327
"04-Weekend-14" 16.29979955832673
"04-Weekend-15" 16.397938279776845
"04-Weekend-16" 16.381832489029204
"04-Weekend-17" 16.30056479583137
"04-Weekend-18" 15.39567464438236
"04-Weekend-19" 18.212213041571506
"04-Weekend-20" 16.071106033858634
"04-Weekend-21" 16.144422535030635
```

```
"04-Weekend-22" 16.46665106293407
"04-Weekend-23" 17.191823268126623
"05-Weekday-00" 18.23457004158131
"05-Weekday-01" 17.855758755200874
"05-Weekday-02" 16.872221490924478
"05-Weekday-03" 17.06451605847531
"05-Weekday-04" 20.6099448208902843
"05-Weekday-05" 20.054898770849224
"05-Weekday-06" 15.504758111315507
"05-Weekday-07" 14.779404281200922
"05-Weekday-08" 15.208048778670776
"05-Weekday-09" 15.77332287617997
"05-Weekday-10" 18.86453435571475
"05-Weekday-11" 17.20439315698898
"05-Weekday-12" 17.25955732137385
"05-Weekday-13" 17.692038700601795
"05-Weekday-14" 17.999688698611003
"05-Weekday-15" 17.763227898681198
"05-Weekday-16" 19.262635560982766
"05-Weekday-17" 18.15817859352869
"05-Weekday-18" 16.806468033528233
"05-Weekday-19" 16.186095304091058
"05-Weekday-20" 16.111827045040577
"05-Weekday-21" 16.599366790151866
"05-Weekday-22" 17.078797095009694
"05-Weekday-23" 17.680399216476616
"05-Weekend-00" 16.411521827806965
"05-Weekend-01" 15.733080021242193
"05-Weekend-02" 15.51587990881733
"05-Weekend-03" 15.987735946302118
"05-Weekend-04" 17.606750244365898
"05-Weekend-05" 21.05305800576708
"05-Weekend-06" 19.601823831129384
"05-Weekend-07" 17.06195805334479
"05-Weekend-08" 15.111026402852424
"05-Weekend-09" 14.238260294987349
"05-Weekend-10" 14.4314102345011
"05-Weekend-11" 14.473402834244306
"05-Weekend-12" 14.86433840494933
"05-Weekend-13" 15.440218918665458
"05-Weekend-14" 16.218078955207442
"05-Weekend-15" 16.257689238054954
"05-Weekend-16" 16.471052344449454
"05-Weekend-17" 16.367193368872098
"05-Weekend-18" 15.482629345535774
"05-Weekend-19" 15.252281992546063
"05-Weekend-20" 15.708283127304298
"05-Weekend-21" 15.85686128900639
```

```
"05-Weekend-22" 16.09228248037357
"05-Weekend-23" 16.92615476505912
"06-Weekday-00" 17.696089227975797
"06-Weekday-01" 17.248904450610453
"06-Weekday-02" 16.354964083387365
"06-Weekday-03" 17.028716589186207
"06-Weekday-04" 20.55753475701337
"06-Weekday-05" 19.723296020228936
"06-Weekday-06" 15.390534884904103
"06-Weekday-07" 14.70810156519479
"06-Weekday-08" 15.231173377934557
"06-Weekday-09" 15.828859202076053
"06-Weekday-10" 16.29227353990938
"06-Weekday-11" 16.682923194417818
"06-Weekday-12" 16.76927451736787
"06-Weekday-13" 17.270562528026467
"06-Weekday-14" 17.6700473401106
"06-Weekday-15" 17.52532340455445
"06-Weekday-16" 19.883915198376176
"06-Weekday-17" 18.12294319522848
"06-Weekday-18" 16.741519730359567
"06-Weekday-19" 16.00375144066547
"06-Weekday-20" 15.603608471540799
"06-Weekday-21" 16.15730435758703
"06-Weekday-22" 18.320080201372107
"06-Weekday-23" 17.304836110819803
"06-Weekend-00" 16.08440072978086
"06-Weekend-01" 15.443699956281144
"06-Weekend-02" 15.17917911471905
"06-Weekend-03" 15.731856679821455
"06-Weekend-04" 17.16856446265509
"06-Weekend-05" 20.59795795665398
"06-Weekend-06" 19.44068963122127
"06-Weekend-07" 16.739302822242074
"06-Weekend-08" 15.218634769011539
"06-Weekend-09" 14.292997910467736
"06-Weekend-10" 14.433005206515292
"06-Weekend-11" 14.862840743657562
"06-Weekend-12" 15.118000624157983
"06-Weekend-13" 16.076253307094465
"06-Weekend-14" 16.82208014966499
"06-Weekend-15" 16.90002034272814
"06-Weekend-16" 16.835363745558297
"06-Weekend-17" 16.66267436626512
"06-Weekend-18" 16.01036880684912
"06-Weekend-19" 15.486876939790466
"06-Weekend-20" 16.096447335630817
"06-Weekend-21" 16.105793675860447
```

```
"06-Weekend-22" 16.311017802236503
"06-Weekend-23" 16.913981303144364
[root@ip-172-31-27-72 tripdata]#
```

## Screenshot of execution of script 6:

```
[root@ip-172-31-27-72 tripdata]# python /home/hadoop/tasks/mrtask_f.py -r hadoop
 hdfs:///user/root/tripdata/*.csv --output-dir /user/root/tripdata/out_mrtask_f/
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in $PATH...
Found hadoop binary: /bin/hadoop
Using Hadoop version 3.3.3
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.root.20231128.125224.604580
uploading working dir files to hdfs:///user/root/tmp/mrjob/mrtask_f.root.2023112
8.125224.604580/files/wd...
Copying other local files to hdfs:///user/root/tmp/mrjob/mrtask_f.root.20231128.
125224.604580/files/
Running step 1 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-2.1.jar] /tmp/s
treamjob3663670730390827794.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_
1701169939900_0011
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 6
  Adding a new node: /default-rack/172.31.29.89:9866
  Adding a new node: /default-rack/172.31.21.145:9866
  number of splits:43
  Submitting tokens for job: job_1701169939900_0011
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1701169939900_0011
  The url to track the job: http://ip-172-31-27-72.ec2.internal:20888/proxy/appl
ication_1701169939900_0011/
  Running job: job_1701169939900_0011
  Job job_1701169939900_0011 running in uber mode : false
   map 0% reduce 0%
   map 1% reduce 0%
   map 2% reduce 0%
```

```
   map 100% reduce 99%
   map 100% reduce 100%
  Job job_1701169939900_0011 completed successfully
  Output directory: hdfs:///user/root/tmp/mrjob/mrtask_f.root.20231128.125224.60
4580/step-output/0000
Counters: 57
        File Input Format Counters
                Bytes Read=5558093822
        File Output Format Counters
                Bytes Written=12297
        File System Counters
                FILE: Number of bytes read=307165256
                FILE: Number of bytes written=630373472
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5558099799
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=12297
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=184
                HDFS: Number of write operations=22
        Job Counters
                Data-local map tasks=32
                Killed map tasks=1
                Killed reduce tasks=1
                Launched map tasks=43
                Launched reduce tasks=12
                Rack-local map tasks=11
                Total megabyte-milliseconds taken by all map tasks=7656187392
                Total megabyte-milliseconds taken by all reduce tasks=4750654464
                Total time spent by all map tasks (ms)=4984497
                Total time spent by all maps in occupied slots (ms)=239255856
                Total time spent by all reduce tasks (ms)=1546437
                Total time spent by all reduces in occupied slots (ms)=148457952
                Total vcore-milliseconds taken by all map tasks=4984497
                Total vcore-milliseconds taken by all reduce tasks=1546437
        Map-Reduce Framework
                CPU time spent (ms)=4932880
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=11831
                Input split bytes=5977
                Map input records=58982297
                Map output bytes=1538543594
                Map output materialized bytes=307328257
                Map output records=58982291
```

```
                Map output records=58982291
                Merged Map outputs=473
                Peak Map Physical memory (bytes)=664129536
                Peak Map Virtual memory (bytes)=3345104896
                Peak Reduce Physical memory (bytes)=620584960
                Peak Reduce Virtual memory (bytes)=4764602368
                Physical memory (bytes) snapshot=29084254208
                Reduce input groups=288
                Reduce input records=58982291
                Reduce output records=288
                Reduce shuffle bytes=307328257
                Shuffled Maps =473
                Spilled Records=117964582
                Total committed heap usage (bytes)=26981957632
                Virtual memory (bytes) snapshot=181251846144
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Running step 2 of 2...
  packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3-amzn-2.1.jar] /tmp/s
treamjob2921628932452517065.jar tmpDir=null
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Connecting to ResourceManager at ip-172-31-27-72.ec2.internal/172.31.27.72:803
2
  Connecting to Application History server at ip-172-31-27-72.ec2.internal/172.3
1.27.72:10200
  Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_
1701169939900_0012
  Loaded native gpl library
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7c
f53ff5f739d6b1532457f2c6cd495e8]
  Total input files to process : 11
  number of splits:31
  Submitting tokens for job: job_1701169939900_0012
  Executing with tokens: []
  resource-types.xml not found
  Unable to find 'resource-types.xml'.
  Submitted application application_1701169939900_0012
  The url to track the job: http://ip-172-31-27-72.ec2.internal:20888/proxy/appl
ication_1701169939900_0012/
  Running job: job_1701169939900_0012
```

```
  Running job: job_1701169939900_0012
  Job job_1701169939900_0012 running in uber mode : false
   map 0% reduce 0%
   map 6% reduce 0%
   map 23% reduce 0%
```

```
   map 100% reduce 73%
   map 100% reduce 91%
   map 100% reduce 100%
  Job job_1701169939900_0012 completed successfully
  Output directory: hdfs:///user/root/tripdata/out_mrtask_f/
Counters: 56
        File Input Format Counters
                Bytes Read=19915
        File Output Format Counters
                Bytes Written=9993
        File System Counters
                FILE: Number of bytes read=6931
                FILE: Number of bytes written=12368470
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=25402
                HDFS: Number of bytes read erasure-coded=0
                HDFS: Number of bytes written=9993
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=148
                HDFS: Number of write operations=22
        Job Counters
                Data-local map tasks=26
                Killed reduce tasks=1
                Launched map tasks=31
                Launched reduce tasks=11
                Rack-local map tasks=5
                Total megabyte-milliseconds taken by all map tasks=398911488
                Total megabyte-milliseconds taken by all reduce tasks=246223872
                Total time spent by all map tasks (ms)=259708
                Total time spent by all maps in occupied slots (ms)=12465984
                Total time spent by all reduce tasks (ms)=80151
                Total time spent by all reduces in occupied slots (ms)=7694496
                Total vcore-milliseconds taken by all map tasks=259708
                Total vcore-milliseconds taken by all reduce tasks=80151
        Map-Reduce Framework
                CPU time spent (ms)=58090
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=6286
                Input split bytes=5487
                Map input records=288
                Map output bytes=12297
                Map output materialized bytes=13952
                Map output records=288
                Merged Map outputs=341
```

```
                Merged Map outputs=341
                Peak Map Physical memory (bytes)=564744192
                Peak Map Virtual memory (bytes)=3097784320
                Peak Reduce Physical memory (bytes)=315330560
                Peak Reduce Virtual memory (bytes)=4440154112
                Physical memory (bytes) snapshot=18903928832
                Reduce input groups=1
                Reduce input records=288
                Reduce output records=288
                Reduce shuffle bytes=13952
                Shuffled Maps =341
                Spilled Records=576
                Total committed heap usage (bytes)=17955815424
                Virtual memory (bytes) snapshot=144109342720
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/root/tripdata/out_mrtask_f/
Removing HDFS temp directory hdfs:///user/root/tmp/mrjob/mrtask_f.root.20231128.
125224.604580...
Removing temp directory /tmp/mrtask_f.root.20231128.125224.604580...
[root@ip-172-31-27-72 tripdata]#
```