Data Ingestion Tasks:

Task 1. Create an RDS instance in your AWS account and upload the data to the RDS instance.

Since the dataset is huge, you need to upload the data from only two files (*i.e.* `yellow_tripdata_2017-01.csv & yellow_tripdata_2017-02.csv`) from the dataset.

Note: You will need to create an appropriate schema for the data sets to upload them to RDS (you can find the data dictionary in the previous segments. The steps to work with RDS in given in the Additional Resource or the documentation can be referred here).

```
Create database demo;

create table yellow_tripdata
(
VendorID              int,
tpep_pickup_datetime  datetime,
tpep_dropoff_datetime datetime,
passenger_count       int,
trip_distance         float,
RatecodeID            int,
store_and_fwd_flag    VARCHAR(255),
PULocationID          int,
DOLocationID          int,
payment_type          int,
fare_amount           float,
extra                 float,
mta_tax               float,
tip_amount            float,
tolls_amount          float,
improvement_surcharge float,
total_amount          float,
Airport_fee           float,
PRIMARY KEY (tpep_pickup_datetime));
```

```
MySQL [demo]> desc yellow_tripdata;
+-----------------------+--------------+------+-----+---------+-------+
| Field                 | Type         | Null | Key | Default | Extra |
+-----------------------+--------------+------+-----+---------+-------+
| VendorID              | int          | YES  |     | NULL    |       |
| tpep_pickup_datetime  | datetime     | NO   | PRI | NULL    |       |
| tpep_dropoff_datetime | datetime     | YES  |     | NULL    |       |
| passenger_count       | int          | YES  |     | NULL    |       |
| trip_distance         | float        | YES  |     | NULL    |       |
| RatecodeID            | int          | YES  |     | NULL    |       |
| store_and_fwd_flag    | varchar(255) | YES  |     | NULL    |       |
| PULocationID          | int          | YES  |     | NULL    |       |
| DOLocationID          | int          | YES  |     | NULL    |       |
| payment_type          | int          | YES  |     | NULL    |       |
| fare_amount           | float        | YES  |     | NULL    |       |
| extra                 | float        | YES  |     | NULL    |       |
| mta_tax               | float        | YES  |     | NULL    |       |
| tip_amount            | float        | YES  |     | NULL    |       |
| tolls_amount          | float        | YES  |     | NULL    |       |
| improvement_surcharge | float        | YES  |     | NULL    |       |
| total_amount          | float        | YES  |     | NULL    |       |
| Airport_fee           | float        | YES  |     | NULL    |       |
+-----------------------+--------------+------+-----+---------+-------+
18 rows in set (0.00 sec)
```

LOAD DATA local INFILE '/home/hadoop/Yellow_tripdata/yellow_tripdata_2017-01.csv'
INTO TABLE yellow_tripdata
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES;

```
MySQL [demo]> LOAD DATA  local INFILE '/home/hadoop/Yellow_tripdata/yellow_tripdata_2017-01.csv'
    -> INTO TABLE yellow_tripdata
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;


Query OK, 2383471 rows affected, 65535 warnings (1 min 54.72 sec)
Records: 9710820  Deleted: 0  Skipped: 7327349  Warnings: 17038169
```

LOAD DATA local INFILE '/home/hadoop/Yellow_tripdata/yellow_tripdata_2017-02.csv'
INTO TABLE yellow_tripdata
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES;

```
MySQL [demo]> LOAD DATA  local INFILE '/home/hadoop/Yellow_tripdata/yellow_tripdata_2017-02.csv'
    -> INTO TABLE yellow_tripdata
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 2167423 rows affected, 65535 warnings (1 min 48.90 sec)
Records: 9169775  Deleted: 0  Skipped: 7002352  Warnings: 16172127
```

```
MySQL [demo]> select count(*) from yellow_tripdata;
+----------+
| count(*) |
+----------+
|  4550894 |
+----------+
1 row in set (0.73 sec)
```

Task 2. Use Sqoop command to ingest the data from RDS into the HBase Table.

create 'Sqoop_to_Hbase_Test','Trip_details'

```
hbase(main):004:0> create 'Sqoop_to_Hbase_Test','Trip_details'
0 row(s) in 1.2530 seconds

=> Hbase::Table - Sqoop_to_Hbase_Test
hbase(main):005:0> list
TABLE
Sqoop_to_Hbase_Test
1 row(s) in 0.0070 seconds

=> ["Sqoop_to_Hbase_Test"]
```

```
hbase(main):005:0> describe 'Sqoop_to_Hbase_Test'
Table Sqoop_to_Hbase_Test is ENABLED
Sqoop_to_Hbase_Test
COLUMN FAMILIES DESCRIPTION
{NAME => 'Trip_details', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL
=> 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.0360 seconds
```

/usr/bin/sqoop import --connect jdbc:mysql://database-1-instance-1.cgzzrjwdbd6o.us-east-1.rds.amazonaws.com:3306/demo --table yellow_tripdata --hbase-table 'Sqoop_to_Hbase_Test' --column-family Trip_details --username root --password XXXXXX --hbase-create-table --hbase-row-key tpep_pickup_datetime -m 4

```
[hadoop@ip-172-31-51-126 Yellow_tripdata]$ /usr/bin/sqoop import --connect jdbc:mysql://database-1-instance-1.cgzzrjwdbd6o.us-east-1.rds.amazonaws.
com:3306/demo --table yellow_tripdata --hbase-table 'Sqoop_to_Hbase_Test' --column-family Trip_details --username root --password welcome1 --hbase-
create-table --hbase-row-key tpep_pickup_datetime -m 4
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
23/03/14 16:45:20 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
23/03/14 16:45:20 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
23/03/14 16:45:20 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
23/03/14 16:45:20 INFO tool.CodeGenTool: Beginning code generation
23/03/14 16:45:21 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `yellow_tripdata` AS t LIMIT 1
23/03/14 16:45:21 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `yellow_tripdata` AS t LIMIT 1
23/03/14 16:45:21 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-hadoop/compile/48db3d93948c99e8d2dd429df7cfab34/yellow_tripdata.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
23/03/14 16:45:24 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoop/compile/48db3d93948c99e8d2dd429df7cfab34/yellow_tripdata.jar
23/03/14 16:45:24 WARN manager.MySQLManager: It looks like you are importing from mysql.
23/03/14 16:45:24 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
23/03/14 16:45:24 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
23/03/14 16:45:24 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
23/03/14 16:45:24 INFO mapreduce.ImportJobBase: Beginning import of yellow_tripdata
```

```
23/03/14 16:45:31 INFO mapreduce.Job: Running job: job_1678807349766_0003
23/03/14 16:45:39 INFO mapreduce.Job: Job job_1678807349766_0003 running in uber mode : false
23/03/14 16:45:39 INFO mapreduce.Job:  map 0% reduce 0%
23/03/14 16:50:26 INFO mapreduce.Job:  map 25% reduce 0%
23/03/14 16:50:28 INFO mapreduce.Job:  map 50% reduce 0%
23/03/14 16:56:13 INFO mapreduce.Job:  map 75% reduce 0%
23/03/14 16:56:23 INFO mapreduce.Job:  map 100% reduce 0%
23/03/14 16:56:24 INFO mapreduce.Job: Job job_1678807349766_0003 completed successfully
23/03/14 16:56:24 INFO mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=1040439
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=717
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=5
                Other local map tasks=5
                Total time spent by all maps in occupied slots (ms)=61447488
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=1280156
                Total vcore-milliseconds taken by all map tasks=1280156
                Total megabyte-milliseconds taken by all map tasks=1966319616
```

```
        Job Counters
                Killed map tasks=1
                Launched map tasks=5
                Other local map tasks=5
                Total time spent by all maps in occupied slots (ms)=61447488
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=1280156
                Total vcore-milliseconds taken by all map tasks=1280156
                Total megabyte-milliseconds taken by all map tasks=1966319616
        Map-Reduce Framework
                Map input records=4550894
                Map output records=4550894
                Input split bytes=717
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=9780
                CPU time spent (ms)=396180
                Physical memory (bytes) snapshot=2818007040
                Virtual memory (bytes) snapshot=13336571904
                Total committed heap usage (bytes)=2558001152
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
23/03/14 16:56:24 INFO mapreduce.ImportJobBase: Transferred 0 bytes in 657.316 seconds (0 bytes/sec)
23/03/14 16:56:24 INFO mapreduce.ImportJobBase: Retrieved 4550894 records.
```

Task 3. Bulk import data from next two files in the dataset on your EMR cluster to your HBase Table using the relevant codes.

Note: For the above task 3, you just need to import data from the subsequent 2 csv files (*i.e.* `yellow_tripdata_2017-03.csv & yellow_tripdata_2017-04.csv`**)** on your EMR cluster.

Comment 1: First line of the data is Vendor ID and it can not use as RowKey.. hence, for this experiment, deleted first column and but I have choose tpep_pickup_datetime

cut -d"," -f2- yellow_tripdata_2017-03.csv > yellow_tripdata_2017-03_updated.csv
cut -d"," -f2- yellow_tripdata_2017-04.csv > yellow_tripdata_2017-04_updated.csv

```
hbase(main):001:0> create 'Hbase_Bulk_loading',{NAME => 'cf'}
0 row(s) in 1.6130 seconds

=> Hbase::Table - Hbase_Bulk_loading
```

```
[hadoop@ip-172-31-52-222 Yellow_tripdata]$ cut -d"," -f2- yellow_tripdata_2017-03.csv > yellow_tripdata_2017-03_updated.csv
[hadoop@ip-172-31-52-222 Yellow_tripdata]$ cut -d"," -f2- yellow_tripdata_2017-04.csv > yellow_tripdata_2017-04_updated.csv
```

```
[hadoop@ip-172-31-52-222 Yellow_tripdata]$ hdfs dfs -ls /user/hadoop/
[hadoop@ip-172-31-52-222 Yellow_tripdata]$ hdfs dfs -put yellow_tripdata_2017-03_updated.csv /user/hadoop/
[hadoop@ip-172-31-52-222 Yellow_tripdata]$ hdfs dfs -put yellow_tripdata_2017-04_updated.csv /user/hadoop/
```

Comment 2 : Happy Base is not industry standard. Hence utilized the inbuilt options to load the data to Hbase.

Bulk loading:
hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator=',' -Dimporttsv.columns='HBASE_ROW_KEY,cf:tpep_dropoff_datetime,cf:passenger_count,cf:trip_distance,cf:RatecodeID,cf:store_and_fwd_flag,cf:PULocationID,cf:DOLocationID,cf:payment_type,cf:fare_amount,cf:extra,cf:mta_tax,cf:tip_amount,cf:tolls_amount,cf:improvement_surcharge,cf:total_amount,cf:congestion_surcharge,cf:airport_fee' Hbase_Bulk_loading /user/hadoop/yellow_tripdata_2017-03_updated.csv

```
[hadoop@ip-172-31-52-222 Yellow_tripdata]$ hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator=',' -Dimporttsv.columns='HBASE_ROW_KEY,cf:tpep_dropoff_datetime,cf:passenger_count,cf:trip_distance,cf:RatecodeID,cf:store_and_fwd_flag,cf:PULocationID,cf:DOLocationID,cf:payment_type,cf:fare_amount,cf:extra,cf:mta_tax,cf:tip_amount,cf:tolls_amount,cf:improvement_surcharge,cf:total_amount,cf:congestion_surcharge,cf:airport_fee' Hbase_Bulk_loading /user/hadoop/yellow_tripdata_2017-03_updated.csv
```

```
Current count: 2387000, row: 2017-03-31 23:46:57
2387778 row(s) in 88.8830 seconds

=> 2387778
```

hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator=',' -Dimporttsv.columns='HBASE_ROW_KEY,cf:tpep_dropoff_datetime,cf:passenger_count,cf:trip_distance,cf:RatecodeID,cf:store_and_fwd_flag,cf:PULocationID,cf:DOLocationID,cf:payment_type,cf:fare_amount,cf:extra,cf:mta_tax,cf:tip_amount,cf:tolls_amount,cf:improvement_surcharge,cf:total_amount,cf:congestion_surcharge,cf:airport_fee' Hbase_Bulk_loading /user/hadoop/yellow_tripdata_2017-04_updated.csv

```
[hadoop@ip-172-31-52-222 ~]$ hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator=',' -Dimporttsv.columns='HBASE_ROW_KEY,cf:tpep_dropoff_datetime,cf:passenger_count,cf:trip_distance,cf:RatecodeID,cf:store_and_fwd_flag,cf:PULocationID,cf:DOLocationID,cf:payment_type,cf:fare_amount,cf:extra,cf:mta_tax,cf:tip_amount,cf:tolls_amount,cf:improvement_surcharge,cf:total_amount,cf:congestion_surcharge,cf:airport_fee' Hbase_Bulk_loading /user/hadoop/yellow_tripdata_2017-04_updated.csv
```

```
Current count: 4730000, row: 2017-04-30 23:53:35
4730332 row(s) in 179.6380 seconds

=> 4730332
```

MapReduce Tasks:

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

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

Ans : VeriFone Inc. has more trips and for 2017-01 month it is got revenue: 82859254.

```
[hadoop@ip-172-31-63-245 Yellow_tripdata]$ cat yellow_tripdata_2017-01.csv | python3 mapphase_4a.py | python3 reducephase_4a.py
VeriFone Inc. has more trips
count: 10625798
revenue: 82859254.39585336
```

Will attach the mapper (mapphase_4a.py) and reducer (reducephase_4a.py) code for the same.

b. Which pickup location generates the most revenue?

```
hadoop jar /lib/hadoop-mapreduce/hadoop-streaming-2.10.1-amzn-4.jar \
-file mapper_4b.py -mapper 'python3 mapper_4b.py' \
-file reducer_4b.py -reducer 'python3 reducer_4b.py' \
-input /user/hadoop/yellow_tripdata_2017-0?.csv \
-output /user/hadoop/output_2
```

Ans : Location 132 with the most revenue is : 77196748.87979728

Will attach all the required mapper_4b.py and reducer_4b.py.

```
[hadoop@ip-172-31-63-245 Yellow_tripdata]$ hdfs dfs -cat /user/hadoop/output_2/*
Location 230 with the most revenue is : 31638128.67054441
Location 132 with the most revenue is : 77196748.87979728
Location 142 with the most revenue is : 22624103.050367367
Location 161 with the most revenue is : 32910770.160582744
Location 162 with the most revenue is : 29439821.990483917
Location 163 with the most revenue is : 22988111.080351263
Location 236 with the most revenue is : 26541250.06052153
Location 138 with the most revenue is : 64480263.96003238
Location 79 with the most revenue is : 25537135.580411006
Location 239 with the most revenue is : 20043140.81030514
Location 186 with the most revenue is : 29804472.580543276
Location 100 with the most revenue is : 16463658.800200181
Location 50 with the most revenue is : 8199828.849957439
Location 24 with the most revenue is : 2459204.5900002224
Location 43 with the most revenue is : 10537984.630019233
```

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

Ans : Payment method 1(i.e. Credit card) is used most for payment methods and its count is: 39754212.
Will attach mapper_4c.py and reducer_4c.py files.

```
[hadoop@ip-172-31-63-245 Yellow_tripdata]$ cat yellow_tripdata_2017-0?.csv | python3 mapper_4c.py | python3 reducer_4c.py
payment method: 5        Num_of transactions : 3
payment method: payment_type     Num_of transactions : 5
payment method: 4        Num_of transactions : 88794
payment method: 3        Num_of transactions : 306912
payment method: 2        Num_of transactions : 18832370
payment method: 1        Num_of transactions : 39754212
```

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

hadoop jar /lib/hadoop-mapreduce/hadoop-streaming-2.10.1-amzn-4.jar \
-file mapper_4d.py -mapper 'python3 mapper_4d.py' \
-file reducer_4d.py -reducer 'python3 reducer_4d.py' \
-input /user/hadoop/yellow_tripdata_2017-0?.csv \
-output /user/hadoop/output_4

Will attach mapper_4d.py and reducer_4d.py.

[hadoop@ip-172-31-63-245 Yellow_tripdata]$ hdfs dfs -cat /user/hadoop/output_4/*
for location id 104 the average amount is: 0.23 mins 0.0 secs
for location id 113 the average amount is: 0.13 mins 0.0 secs
for location id 122 the average amount is: 0.22 mins 0.0 secs
for location id 131 the average amount is: 0.11 mins 0.0 secs
for location id 140 the average amount is: 0.1 mins 0.0 secs
for location id 17 the average amount is: 0.83 mins 0.0 secs
for location id 203 the average amount is: 0.13 mins 0.0 secs
for location id 212 the average amount is: 0.1 mins 0.0 secs
for location id 221 the average amount is: 0.07 mins 0.0 secs
for location id 230 the average amount is: 0.07 mins 0.0 secs
for location id 26 the average amount is: 0.46 mins 0.0 secs
for location id 35 the average amount is: 0.56 mins 0.0 secs
for location id 44 the average amount is: 0.27 mins 0.0 secs
for location id 53 the average amount is: 0.32 mins 0.0 secs
for location id 62 the average amount is: 0.25 mins 0.0 secs
for location id 71 the average amount is: 0.23 mins 0.0 secs
for location id 80 the average amount is: 0.19 mins 0.0 secs
for location id 105 the average amount is: 0.19 mins 0.0 secs
for location id 114 the average amount is: 0.14 mins 0.0 secs
for location id 123 the average amount is: 0.13 mins 0.0 secs
for location id 132 the average amount is: 0.33 mins 0.0 secs
for location id 141 the average amount is: 0.09 mins 0.0 secs
for location id 150 the average amount is: 0.12 mins 0.0 secs
for location id 18 the average amount is: 0.8 mins 0.0 secs
for location id 204 the average amount is: 0.02 mins 0.0 secs
for location id 213 the average amount is: 0.08 mins 0.0 secs
for location id 222 the average amount is: 0.13 mins 0.0 secs
for location id 231 the average amount is: 0.07 mins 0.0 secs
for location id 240 the average amount is: 0.06 mins 0.0 secs
for location id 27 the average amount is: 0.33 mins 0.0 secs
for location id 36 the average amount is: 0.47 mins 0.0 secs

for location id 45 the average amount is: 0.41 mins 0.0 secs
for location id 54 the average amount is: 0.32 mins 0.0 secs
for location id 63 the average amount is: 0.3 mins 0.0 secs
for location id 72 the average amount is: 0.27 mins 0.0 secs
for location id 81 the average amount is: 0.16 mins 0.0 secs
for location id 90 the average amount is: 0.16 mins 0.0 secs
for location id 106 the average amount is: 0.13 mins 0.0 secs
for location id 115 the average amount is: 0.12 mins 0.0 secs
for location id 124 the average amount is: 0.23 mins 0.0 secs
for location id 133 the average amount is: 0.12 mins 0.0 secs
for location id 142 the average amount is: 0.1 mins 0.0 secs
for location id 151 the average amount is: 0.09 mins 0.0 secs
for location id 160 the average amount is: 0.13 mins 0.0 secs
for location id 19 the average amount is: 0.66 mins 0.0 secs
for location id 205 the average amount is: 0.12 mins 0.0 secs
for location id 214 the average amount is: 0.05 mins 0.0 secs
for location id 223 the average amount is: 0.07 mins 0.0 secs
for location id 232 the average amount is: 0.07 mins 0.0 secs
for location id 241 the average amount is: 0.06 mins 0.0 secs
for location id 250 the average amount is: 0.07 mins 0.0 secs
for location id 28 the average amount is: 1.01 mins 1.0 secs
for location id 37 the average amount is: 0.44 mins 0.0 secs
for location id 46 the average amount is: 0.38 mins 0.0 secs
for location id 55 the average amount is: 0.45 mins 0.0 secs
for location id 64 the average amount is: 0.21 mins 0.0 secs
for location id 73 the average amount is: 0.29 mins 0.0 secs
for location id 82 the average amount is: 0.18 mins 0.0 secs
for location id 91 the average amount is: 0.23 mins 0.0 secs
for location id 107 the average amount is: 0.13 mins 0.0 secs
for location id 116 the average amount is: 0.13 mins 0.0 secs
for location id 125 the average amount is: 0.13 mins 0.0 secs
for location id 134 the average amount is: 0.12 mins 0.0 secs
for location id 143 the average amount is: 0.09 mins 0.0 secs
for location id 152 the average amount is: 0.09 mins 0.0 secs
for location id 161 the average amount is: 0.1 mins 0.0 secs
for location id 170 the average amount is: 0.09 mins 0.0 secs
for location id 206 the average amount is: 0.07 mins 0.0 secs
for location id 215 the average amount is: 0.23 mins 0.0 secs
for location id 224 the average amount is: 0.06 mins 0.0 secs
for location id 233 the average amount is: 0.07 mins 0.0 secs
for location id 242 the average amount is: 0.06 mins 0.0 secs
for location id 251 the average amount is: 0.05 mins 0.0 secs
for location id 260 the average amount is: 0.06 mins 0.0 secs
for location id 29 the average amount is: 0.76 mins 0.0 secs
for location id 38 the average amount is: 1.34 mins 1.0 secs
for location id 47 the average amount is: 0.35 mins 0.0 secs
for location id 56 the average amount is: 0.36 mins 0.0 secs
for location id 65 the average amount is: 0.26 mins 0.0 secs
for location id 74 the average amount is: 0.17 mins 0.0 secs
for location id 83 the average amount is: 0.19 mins 0.0 secs

for location id 92 the average amount is: 0.19 mins 0.0 secs
for location id 108 the average amount is: 0.13 mins 0.0 secs
for location id 117 the average amount is: 0.17 mins 0.0 secs
for location id 126 the average amount is: 0.15 mins 0.0 secs
for location id 135 the average amount is: 0.13 mins 0.0 secs
for location id 144 the average amount is: 0.12 mins 0.0 secs
for location id 153 the average amount is: 0.09 mins 0.0 secs
for location id 162 the average amount is: 0.09 mins 0.0 secs
for location id 171 the average amount is: 0.08 mins 0.0 secs
for location id 180 the average amount is: 0.17 mins 0.0 secs
for location id 207 the average amount is: 0.04 mins 0.0 secs
for location id 216 the average amount is: 0.13 mins 0.0 secs
for location id 225 the average amount is: 0.07 mins 0.0 secs
for location id 234 the average amount is: 0.06 mins 0.0 secs
for location id 243 the average amount is: 0.07 mins 0.0 secs
for location id 252 the average amount is: 0.08 mins 0.0 secs
for location id 261 the average amount is: 0.08 mins 0.0 secs
for location id 39 the average amount is: 0.47 mins 0.0 secs
for location id 48 the average amount is: 0.31 mins 0.0 secs
for location id 57 the average amount is: 0.25 mins 0.0 secs
for location id 66 the average amount is: 0.27 mins 0.0 secs
for location id 75 the average amount is: 0.17 mins 0.0 secs
for location id 84 the average amount is: 0.14 mins 0.0 secs
for location id 93 the average amount is: 0.37 mins 0.0 secs
for location id 1 the average amount is: 8.19 mins 8.0 secs
for location id 109 the average amount is: 0.56 mins 0.0 secs
for location id 118 the average amount is: 0.1 mins 0.0 secs
for location id 127 the average amount is: 0.12 mins 0.0 secs
for location id 136 the average amount is: 0.08 mins 0.0 secs
for location id 145 the average amount is: 0.09 mins 0.0 secs
for location id 154 the average amount is: 0.16 mins 0.0 secs
for location id 163 the average amount is: 0.1 mins 0.0 secs
for location id 172 the average amount is: 0.09 mins 0.0 secs
for location id 181 the average amount is: 0.09 mins 0.0 secs
for location id 190 the average amount is: 0.1 mins 0.0 secs
for location id 208 the average amount is: 0.13 mins 0.0 secs
for location id 217 the average amount is: 0.08 mins 0.0 secs
for location id 226 the average amount is: 0.07 mins 0.0 secs
for location id 235 the average amount is: 0.07 mins 0.0 secs
for location id 244 the average amount is: 0.07 mins 0.0 secs
for location id 253 the average amount is: 0.09 mins 0.0 secs
for location id 262 the average amount is: 0.05 mins 0.0 secs
for location id 49 the average amount is: 0.29 mins 0.0 secs
for location id 58 the average amount is: 0.11 mins 0.0 secs
for location id 67 the average amount is: 0.29 mins 0.0 secs
for location id 76 the average amount is: 0.29 mins 0.0 secs
for location id 85 the average amount is: 0.23 mins 0.0 secs
for location id 94 the average amount is: 0.21 mins 0.0 secs
for location id 119 the average amount is: 0.14 mins 0.0 secs
for location id 128 the average amount is: 0.12 mins 0.0 secs

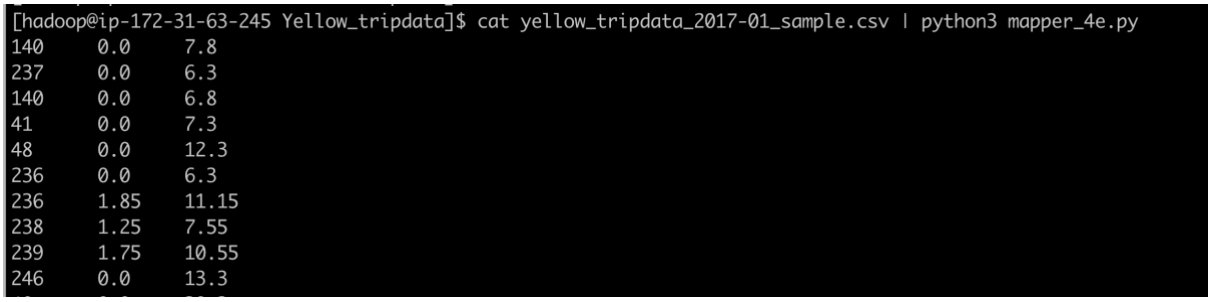
for location id 137 the average amount is: 0.1 mins 0.0 secs
for location id 146 the average amount is: 0.1 mins 0.0 secs
for location id 155 the average amount is: 0.17 mins 0.0 secs
for location id 164 the average amount is: 0.09 mins 0.0 secs
for location id 173 the average amount is: 0.08 mins 0.0 secs
for location id 182 the average amount is: 0.08 mins 0.0 secs
for location id 191 the average amount is: 0.09 mins 0.0 secs
for location id 2 the average amount is: 19.08 mins 19.0 secs
for location id 209 the average amount is: 0.09 mins 0.0 secs
for location id 218 the average amount is: 0.09 mins 0.0 secs
for location id 227 the average amount is: 0.06 mins 0.0 secs
for location id 236 the average amount is: 0.05 mins 0.0 secs
for location id 245 the average amount is: 0.04 mins 0.0 secs
for location id 254 the average amount is: 0.07 mins 0.0 secs
for location id 263 the average amount is: 0.05 mins 0.0 secs
for location id 59 the average amount is: 0.29 mins 0.0 secs
for location id 68 the average amount is: 0.23 mins 0.0 secs
for location id 77 the average amount is: 0.26 mins 0.0 secs
for location id 86 the average amount is: 0.17 mins 0.0 secs
for location id 95 the average amount is: 0.19 mins 0.0 secs
for location id 129 the average amount is: 0.11 mins 0.0 secs
for location id 138 the average amount is: 0.27 mins 0.0 secs
for location id 147 the average amount is: 0.09 mins 0.0 secs
for location id 156 the average amount is: 0.12 mins 0.0 secs
for location id 165 the average amount is: 0.11 mins 0.0 secs
for location id 174 the average amount is: 0.08 mins 0.0 secs
for location id 183 the average amount is: 0.07 mins 0.0 secs
for location id 192 the average amount is: 0.1 mins 0.0 secs
for location id 219 the average amount is: 0.21 mins 0.0 secs
for location id 228 the average amount is: 0.07 mins 0.0 secs
for location id 237 the average amount is: 0.05 mins 0.0 secs
for location id 246 the average amount is: 0.06 mins 0.0 secs
for location id 255 the average amount is: 0.06 mins 0.0 secs
for location id 264 the average amount is: 0.06 mins 0.0 secs
for location id 3 the average amount is: 7.22 mins 7.0 secs
for location id 69 the average amount is: 0.24 mins 0.0 secs
for location id 78 the average amount is: 0.17 mins 0.0 secs
for location id 87 the average amount is: 0.23 mins 0.0 secs
for location id 96 the average amount is: 0.16 mins 0.0 secs
for location id 10 the average amount is: 4.98 mins 4.0 secs
for location id 139 the average amount is: 0.43 mins 0.0 secs
for location id 148 the average amount is: 0.11 mins 0.0 secs
for location id 157 the average amount is: 0.13 mins 0.0 secs
for location id 166 the average amount is: 0.09 mins 0.0 secs
for location id 175 the average amount is: 0.07 mins 0.0 secs
for location id 184 the average amount is: 0.09 mins 0.0 secs
for location id 193 the average amount is: 0.07 mins 0.0 secs
for location id 229 the average amount is: 0.06 mins 0.0 secs
for location id 238 the average amount is: 0.05 mins 0.0 secs
for location id 247 the average amount is: 0.07 mins 0.0 secs

for location id 256 the average amount is: 0.06 mins 0.0 secs
for location id 265 the average amount is: 0.04 mins 0.0 secs
for location id 4 the average amount is: 3.75 mins 3.0 secs
for location id 79 the average amount is: 0.19 mins 0.0 secs
for location id 88 the average amount is: 0.25 mins 0.0 secs
for location id 97 the average amount is: 0.16 mins 0.0 secs
for location id 11 the average amount is: 1.56 mins 1.0 secs
for location id 149 the average amount is: 0.11 mins 0.0 secs
for location id 158 the average amount is: 0.11 mins 0.0 secs
for location id 167 the average amount is: 0.09 mins 0.0 secs
for location id 176 the average amount is: 0.17 mins 0.0 secs
for location id 185 the average amount is: 0.1 mins 0.0 secs
for location id 194 the average amount is: 0.13 mins 0.0 secs
for location id 20 the average amount is: 0.71 mins 0.0 secs
for location id 239 the average amount is: 0.05 mins 0.0 secs
for location id 248 the average amount is: 0.06 mins 0.0 secs
for location id 257 the average amount is: 0.06 mins 0.0 secs
for location id 5 the average amount is: 2.29 mins 2.0 secs
for location id 89 the average amount is: 0.18 mins 0.0 secs
for location id 98 the average amount is: 0.13 mins 0.0 secs
for location id 12 the average amount is: 2.03 mins 2.0 secs
for location id 159 the average amount is: 0.09 mins 0.0 secs
for location id 168 the average amount is: 0.08 mins 0.0 secs
for location id 177 the average amount is: 0.11 mins 0.0 secs
for location id 186 the average amount is: 0.09 mins 0.0 secs
for location id 195 the average amount is: 0.11 mins 0.0 secs
for location id 21 the average amount is: 0.98 mins 0.0 secs
for location id 249 the average amount is: 0.06 mins 0.0 secs
for location id 258 the average amount is: 0.08 mins 0.0 secs
for location id 30 the average amount is: 0.84 mins 0.0 secs
for location id 6 the average amount is: 1.25 mins 1.0 secs
for location id 99 the average amount is: 0.13 mins 0.0 secs
for location id 100 the average amount is: 0.15 mins 0.0 secs
for location id 13 the average amount is: 1.5 mins 1.0 secs
for location id 169 the average amount is: 0.09 mins 0.0 secs
for location id 178 the average amount is: 0.04 mins 0.0 secs
for location id 187 the average amount is: 0.21 mins 0.0 secs
for location id 196 the average amount is: 0.09 mins 0.0 secs
for location id 22 the average amount is: 1.0 mins 0.0 secs
for location id 259 the average amount is: 0.06 mins 0.0 secs
for location id 31 the average amount is: 0.8 mins 0.0 secs
for location id 40 the average amount is: 0.41 mins 0.0 secs
for location id 7 the average amount is: 1.89 mins 1.0 secs
for location id 101 the average amount is: 0.18 mins 0.0 secs
for location id 110 the average amount is: 0.03 mins 0.0 secs
for location id 14 the average amount is: 1.2 mins 1.0 secs
for location id 179 the average amount is: 0.08 mins 0.0 secs
for location id 188 the average amount is: 0.09 mins 0.0 secs
for location id 197 the average amount is: 0.08 mins 0.0 secs
for location id 200 the average amount is: 0.06 mins 0.0 secs

for location id 23 the average amount is: 0.56 mins 0.0 secs
for location id 32 the average amount is: 0.63 mins 0.0 secs
for location id 41 the average amount is: 0.32 mins 0.0 secs
for location id 50 the average amount is: 0.3 mins 0.0 secs
for location id 8 the average amount is: 2.19 mins 2.0 secs
for location id 102 the average amount is: 0.22 mins 0.0 secs
for location id 111 the average amount is: 0.1 mins 0.0 secs
for location id 120 the average amount is: 0.12 mins 0.0 secs
for location id 15 the average amount is: 0.97 mins 0.0 secs
for location id 189 the average amount is: 0.08 mins 0.0 secs
for location id 198 the average amount is: 0.07 mins 0.0 secs
for location id 201 the average amount is: 0.05 mins 0.0 secs
for location id 210 the average amount is: 0.09 mins 0.0 secs
for location id 24 the average amount is: 0.56 mins 0.0 secs
for location id 33 the average amount is: 0.55 mins 0.0 secs
for location id 42 the average amount is: 0.31 mins 0.0 secs
for location id 51 the average amount is: 0.32 mins 0.0 secs
for location id 60 the average amount is: 0.21 mins 0.0 secs
for location id 9 the average amount is: 9.39 mins 9.0 secs
for location id 112 the average amount is: 0.13 mins 0.0 secs
for location id 121 the average amount is: 0.12 mins 0.0 secs
for location id 130 the average amount is: 0.27 mins 0.0 secs
for location id 16 the average amount is: 1.37 mins 1.0 secs
for location id 199 the average amount is: 0.09 mins 0.0 secs
for location id 202 the average amount is: 0.08 mins 0.0 secs
for location id 211 the average amount is: 0.08 mins 0.0 secs
for location id 220 the average amount is: 0.05 mins 0.0 secs
for location id 25 the average amount is: 0.62 mins 0.0 secs
for location id 34 the average amount is: 0.47 mins 0.0 secs
for location id 43 the average amount is: 0.36 mins 0.0 secs
for location id 52 the average amount is: 0.33 mins 0.0 secs
for location id 61 the average amount is: 0.25 mins 0.0 secs
for location id 70 the average amount is: 0.36 mins 0.0 secs

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

Due to some issue, the reducer is not working.. hence, just submitting the mapper job. Attaching the code of Mapper job mapper_4e.py



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

Answer: Not able to implement the reducer code.. Just attaching the mapper job for this job.

Attachign the Mapper code : mapper_4f.py

```
[hadoop@ip-172-31-63-245 Yellow_tripdata]$ cat yellow_tripdata_2017-01_sample.csv | python mapper_4f.py
01      7.8     01      00
01      6.3     01      00
01      6.8     01      00
01      7.3     01      00
01      12.3    01      00
01      6.3     01      00
01      11.15   01      00
01      7.55    01      00
01      10.55   01      01
```