1. Downloading files to EMR using wget and listing them.

2. Logging in to RDS and selecting data base.

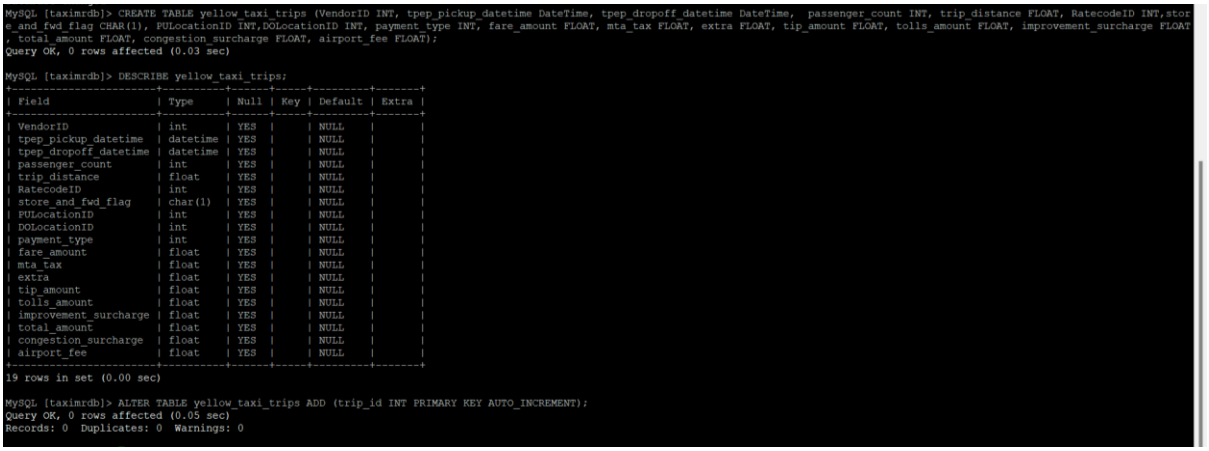
   mysql -h nyc-tlc-mr-db.cv4nmpnzjxab.us-east-1.rds.amazonaws.com -P 3306 -u admin -p

   use taximrdb;

3. Creating table and adding autoincrement column.

   CREATE TABLE yellow_taxi_trips (VendorID INT, tpep_pickup_datetime DateTime, tpep_dropoff_datetime DateTime, passenger_count INT, trip_distance FLOAT, RatecodeID INT,store_and_fwd_flag CHAR(1), PULocationID INT,DOLocationID INT, payment_type INT, fare_amount FLOAT, mta_tax FLOAT, extra FLOAT, tip_amount FLOAT, tolls_amount FLOAT, improvement_surcharge FLOAT, total_amount FLOAT, congestion_surcharge FLOAT, airport_fee FLOAT);

   ALTER TABLE yellow_taxi_trips ADD (trip_id INT PRIMARY KEY AUTO_INCREMENT);



4. Loading data from the files
   LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv' INTO TABLE yellow_taxi_trips FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;

```
MySQL [taximrdb]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv' INTO TABLE yellow_taxi_trips FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9710820 rows affected, 65535 warnings (1 min 51.40 sec)
Records: 9710820  Deleted: 0  Skipped: 0  Warnings: 29132460

MySQL [taximrdb]>
```

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

```
MySQL [taximrdb]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv' INTO TABLE yellow_taxi_trips FIELDS TERMINATED BY ',' LINES
TERMINATED BY '\n' IGNORE 1 LINES;
Query OK, 9169775 rows affected, 65535 warnings (2 min 41.39 sec)
Records: 9169775  Deleted: 0  Skipped: 0  Warnings: 27509325

MySQL [taximrdb]>
```

5. Checking count and data

```
MySQL [taximrdb]> SELECT COUNT(*) FROM yellow_taxi_trips;
+----------+
| COUNT(*) |
+----------+
| 18880595 |
+----------+
1 row in set (37.40 sec)

MySQL [taximrdb]>
```

   SELECT trip_id, VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, RatecodeID, store_and_fwd_flag, PULocationID, DOLocationID, payment_type, fare_amount, mta_tax, extra, tip_amount, tolls_amount, improvement_surcharge, total_amount, congestion_surcharge, airport_fee from yellow_taxi_trips LIMIT 5;

```
root@ip-172-31-58-150:/home/hadoop
MySQL [taximrdb]> SELECT trip_id, VendorID, tpep_pickup_datetime, tpep_dropoff_datetime,  passenger_count, trip_distance, RatecodeID, store_and_fwd_flag, PULocationID, DOLocationID, payment
_type, fare_amount, mta_tax, extra, tip_amount, tolls_amount, improvement_surcharge, total_amount, congestion_surcharge, airport_fee from yellow_taxi_trips LIMIT 5;
+---------+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+--------------+--------------+--------------+-------------+
| trip_id | VendorID | tpep_pickup_datetime | tpep_dropoff_datetime | passenger_count | trip_distance | RatecodeID | store_and_fwd_flag | PULocationID | DOLocationID | payment_type | fare_amount | mta_tax | extra | tip_amount | tolls_amount | improvement_surcharge | total_amount | congestion_surcharge | airport_fee |
+---------+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+--------------+--------------+--------------+-------------+---------+-------+------------+--------------+-----------------------+--------------+----------------------+-------------+
|       1 |        1 | 2017-01-01 00:32:05  | 2017-01-01 00:37:48   |               1 |           1.2 |          1 | N                  |          140 |          236 |            2 |         6.5 |     0.5 |   0.5 |          0 |            0 |                   0.3 |          7.8 |                    0 |           0 |
|       2 |        1 | 2017-01-01 00:43:25  | 2017-01-01 00:47:42   |               2 |           0.7 |          1 | N                  |          237 |          140 |            2 |           5 |     0.5 |   0.5 |          0 |            0 |                   0.3 |          6.3 |                    0 |           0 |
|       3 |        1 | 2017-01-01 00:49:10  | 2017-01-01 00:53:53   |               2 |           0.8 |          1 | N                  |          140 |          237 |            2 |         5.5 |     0.5 |   0.5 |          0 |            0 |                   0.3 |          6.8 |                    0 |           0 |
|       4 |        1 | 2017-01-01 00:36:42  | 2017-01-01 00:41:09   |               1 |           1.1 |          1 | N                  |           41 |           42 |            2 |           6 |     0.5 |   0.5 |          0 |            0 |                   0.3 |          7.3 |                    0 |           0 |
|       5 |        1 | 2017-01-01 00:07:41  | 2017-01-01 00:18:16   |               1 |             3 |          1 | N                  |           48 |          263 |            2 |          11 |     0.5 |   0.5 |          0 |            0 |                   0.3 |         12.3 |                    0 |           0 |
+---------+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+--------------+--------------+--------------+-------------+---------+-------+------------+--------------+-----------------------+--------------+----------------------+-------------+
5 rows in set (0.01 sec)

MySQL [taximrdb]> SELECT trip_id, VendorID, tpep_pickup_datetime, tpep_dropoff_datetime,  passenger_count, trip_distance, RatecodeID, store_and_fwd_flag, PULocationID, DOLocationID, payment
_type from yellow_taxi_trips LIMIT 5;
+---------+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+--------------+--------------+--------------+
| trip_id | VendorID | tpep_pickup_datetime | tpep_dropoff_datetime | passenger_count | trip_distance | RatecodeID | store_and_fwd_flag | PULocationID | DOLocationID | payment_type |
+---------+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+--------------+--------------+--------------+
|       1 |        1 | 2017-01-01 00:32:05  | 2017-01-01 00:37:48   |               1 |           1.2 |          1 | N                  |          140 |          236 |            2 |
|       2 |        1 | 2017-01-01 00:43:25  | 2017-01-01 00:47:42   |               2 |           0.7 |          1 | N                  |          237 |          140 |            2 |
|       3 |        1 | 2017-01-01 00:49:10  | 2017-01-01 00:53:53   |               2 |           0.8 |          1 | N                  |          140 |          237 |            2 |
|       4 |        1 | 2017-01-01 00:36:42  | 2017-01-01 00:41:09   |               1 |           1.1 |          1 | N                  |           41 |           42 |            2 |
|       5 |        1 | 2017-01-01 00:07:41  | 2017-01-01 00:18:16   |               1 |             3 |          1 | N                  |           48 |          263 |            2 |
+---------+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+--------------+--------------+--------------+
5 rows in set (0.00 sec)

MySQL [taximrdb]> SELECT trip_id, fare_amount, mta_tax, extra, tip_amount, tolls_amount, improvement_surcharge, total_amount, congestion_surcharge, airport_fee from yellow_taxi_trips LIMIT
5;
+---------+-------------+---------+-------+------------+--------------+-----------------------+--------------+----------------------+-------------+
| trip_id | fare_amount | mta_tax | extra | tip_amount | tolls_amount | improvement_surcharge | total_amount | congestion_surcharge | airport_fee |
+---------+-------------+---------+-------+------------+--------------+-----------------------+--------------+----------------------+-------------+
|       1 |         6.5 |     0.5 |   0.5 |          0 |            0 |                   0.3 |          7.8 |                    0 |           0 |
|       2 |           5 |     0.5 |   0.5 |          0 |            0 |                   0.3 |          6.3 |                    0 |           0 |
|       3 |         5.5 |     0.5 |   0.5 |          0 |            0 |                   0.3 |          6.8 |                    0 |           0 |
|       4 |           6 |     0.5 |   0.5 |          0 |            0 |                   0.3 |          7.3 |                    0 |           0 |
|       5 |          11 |     0.5 |   0.5 |          0 |            0 |                   0.3 |         12.3 |                    0 |           0 |
+---------+-------------+---------+-------+------------+--------------+-----------------------+--------------+----------------------+-------------+
5 rows in set (0.00 sec)
```

6. Row count in file – including headers

```
MySQL [taximrdb]> exit;
Bye
[root@ip-172-31-58-150 hadoop]# wc -l yellow_tripdata_2017-01.csv
9710821 yellow_tripdata_2017-01.csv
[root@ip-172-31-58-150 hadoop]# wc -l yellow_tripdata_2017-02.csv
9169776 yellow_tripdata_2017-02.csv
[root@ip-172-31-58-150 hadoop]#
```