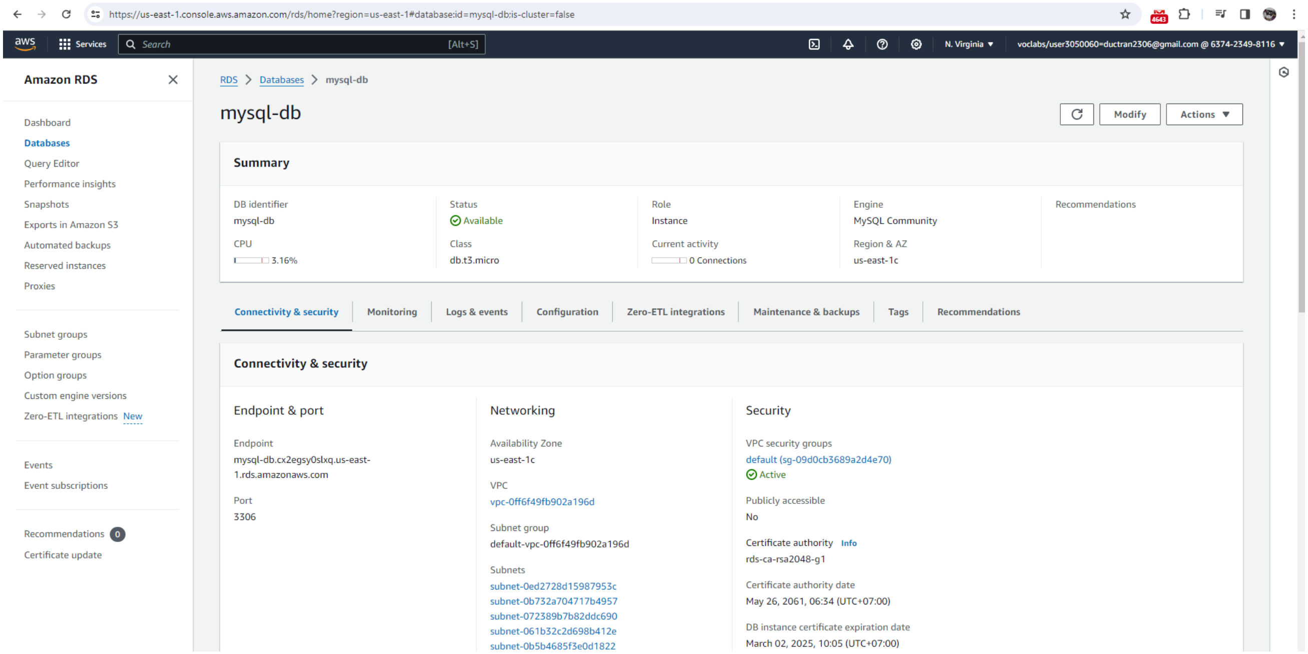
MapReduce Case Study Task 1

1) Create RDS Instance



https://us-east-1.console.aws.amazon.com/rds/home?region=us-east-1#database:id=mysql-db;is-cluster=false

voclabs/user3050060=ductran2306@gmail.com @ 6374-2349-8116

mysql-db

Summary

DB identifier: mysql-db
CPU: 3.16%
Status: Available
Class: db.t3.micro
Role: Instance
Current activity: 0 Connections
Engine: MySQL Community
Region & AZ: us-east-1c
Recommendations

Connectivity & security

Endpoint & port
Endpoint: mysql-db.cx2egsy0slxq.us-east-1.rds.amazonaws.com
Port: 3306

Networking
Availability Zone: us-east-1c
VPC: vpc-0ff6f49fb902a196d
Subnet group: default-vpc-0ff6f49fb902a196d
Subnets: subnet-0ed2728d15987953c, subnet-0b732a704717b4957, subnet-072389b7b82ddc690, subnet-061b32c2d698b412e, subnet-0b5b4685f3e0d1822

Security
VPC security groups: default (sg-09d0cb3689a2d4e70) Active
Publicly accessible: No
Certificate authority Info: rds-ca-rsa2048-g1
Certificate authority date: May 26, 2061, 06:34 (UTC+07:00)
DB instance certificate expiration date: March 02, 2025, 10:05 (UTC+07:00)

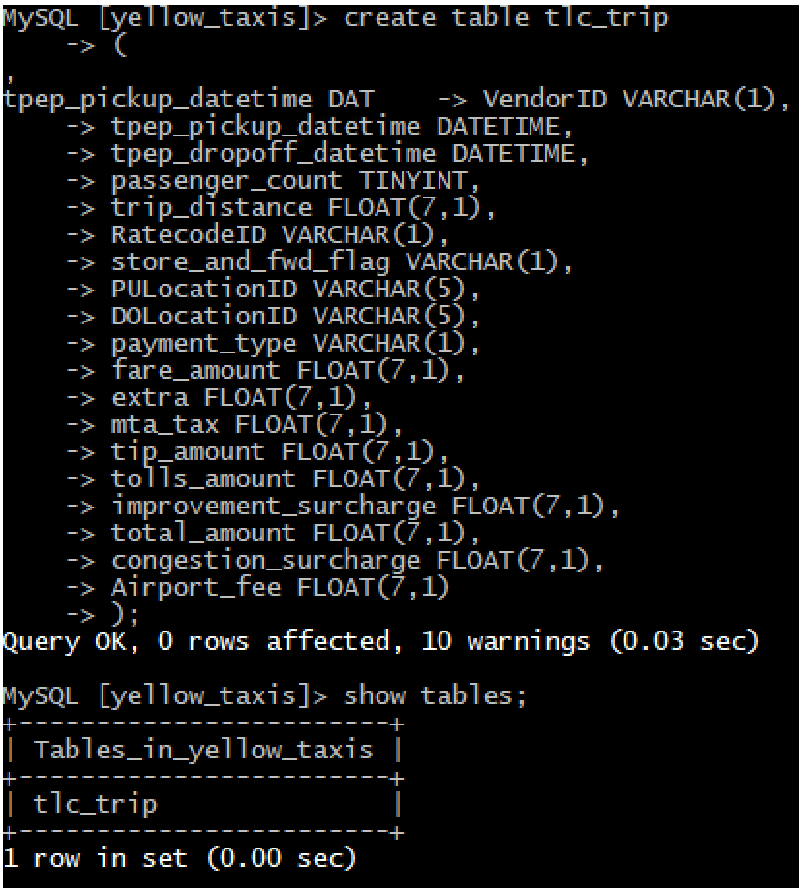
Create a database name "yellow_taxis":

```
create database yellow_taxis;
use yellow_taxis;
```

Create a table name "tlc_trip":

```
create table tlc_trip
(
VendorID VARCHAR(1),
tpep_pickup_datetime DATETIME,
tpep_dropoff_datetime DATETIME,
passenger_count TINYINT,
trip_distance FLOAT(7,1),
RatecodeID VARCHAR(1),
store_and_fwd_flag VARCHAR(1),
PULocationID VARCHAR(5),
DOLocationID VARCHAR(5),
payment_type VARCHAR(1),
fare_amount FLOAT(7,1),
extra FLOAT(7,1),
mta_tax FLOAT(7,1),
tip_amount FLOAT(7,1),
tolls_amount FLOAT(7,1),
improvement_surcharge FLOAT(7,1),
total_amount FLOAT(7,1),
congestion_surcharge FLOAT(7,1),
Airport_fee FLOAT(7,1)
);
```

```
show tables;
```



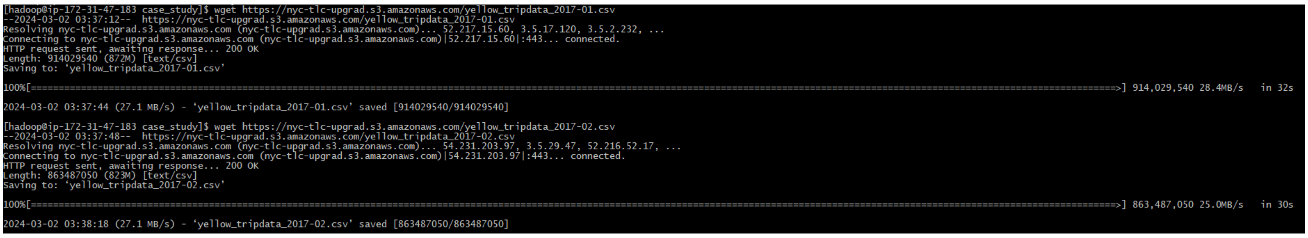
```
MySQL [yellow_taxis]> create table tlc_trip
    -> (
,
tpep_pickup_datetime DAT    -> VendorID VARCHAR(1),
    -> tpep_pickup_datetime DATETIME,
    -> tpep_dropoff_datetime DATETIME,
    -> passenger_count TINYINT,
    -> trip_distance FLOAT(7,1),
    -> RatecodeID VARCHAR(1),
    -> store_and_fwd_flag VARCHAR(1),
    -> PULocationID VARCHAR(5),
    -> DOLocationID VARCHAR(5),
    -> payment_type VARCHAR(1),
    -> fare_amount FLOAT(7,1),
    -> extra FLOAT(7,1),
    -> mta_tax FLOAT(7,1),
    -> tip_amount FLOAT(7,1),
    -> tolls_amount FLOAT(7,1),
    -> improvement_surcharge FLOAT(7,1),
    -> total_amount FLOAT(7,1),
    -> congestion_surcharge FLOAT(7,1),
    -> Airport_fee FLOAT(7,1)
    -> );
Query OK, 0 rows affected, 10 warnings (0.03 sec)

MySQL [yellow_taxis]> show tables;
+------------------------+
| Tables_in_yellow_taxis |
+------------------------+
| tlc_trip               |
+------------------------+
1 row in set (0.00 sec)
```

2) Load data into the table

First, download the data set in EMR instance:

```
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
```



```
[hadoop@ip-172-31-47-183 case_study]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
--2024-03-02 03:37:12--  https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
Resolving nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)... 52.217.15.60, 3.5.17.120, 3.5.2.232, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|52.217.15.60|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 914029540 (872M) [text/csv]
Saving to: ‘yellow_tripdata_2017-01.csv’

100%[===================================================================================================================================================================>] 914,029,540 28.4MB/s   in 32s

2024-03-02 03:37:44 (27.1 MB/s) - ‘yellow_tripdata_2017-01.csv’ saved [914029540/914029540]

[hadoop@ip-172-31-47-183 case_study]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
--2024-03-02 03:37:48--  https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
Resolving nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)... 54.231.203.97, 3.5.29.47, 52.216.52.17, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|54.231.203.97|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 863487050 (823M) [text/csv]
Saving to: ‘yellow_tripdata_2017-02.csv’

100%[===================================================================================================================================================================>] 863,487,050 25.0MB/s   in 30s

2024-03-02 03:38:18 (27.1 MB/s) - ‘yellow_tripdata_2017-02.csv’ saved [863487050/863487050]
```

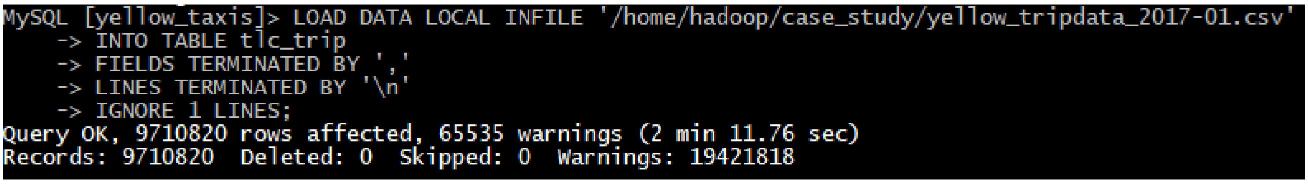
Then, load the data set to the table

a) Connect to mysql database from EMR instane:

```
mysql -h mysql-db.cx2egsy0slxq.us-east-1.rds.amazonaws.com -P 3306 -u admin -p
```

b) In mysql console, load the data in to the table tlc_trip created in the previous step.

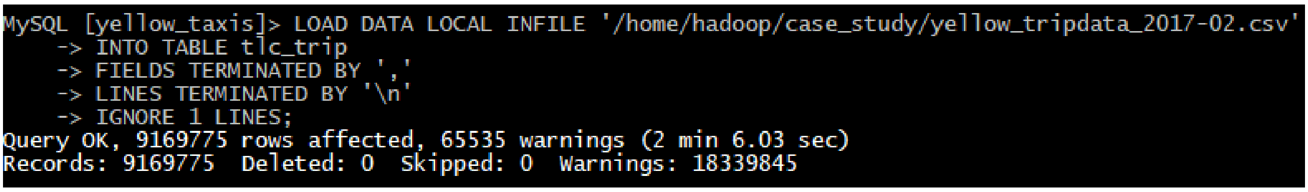
```
LOAD DATA LOCAL INFILE '/home/hadoop/case_study/yellow_tripdata_2017-01.csv'
INTO TABLE tlc_trip
FIELDS TERMINATED BY ','  -- each value delimited by character ','
LINES TERMINATED BY '\n'  -- each record end with '\n'
IGNORE 1 LINES;           -- ignore the first header line
```



```
MySQL [yellow_taxis]> LOAD DATA LOCAL INFILE '/home/hadoop/case_study/yellow_tripdata_2017-01.csv'
    -> INTO TABLE tlc_trip
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 9710820 rows affected, 65535 warnings (2 min 11.76 sec)
Records: 9710820  Deleted: 0  Skipped: 0  Warnings: 19421818
```

Same for the second csv file:

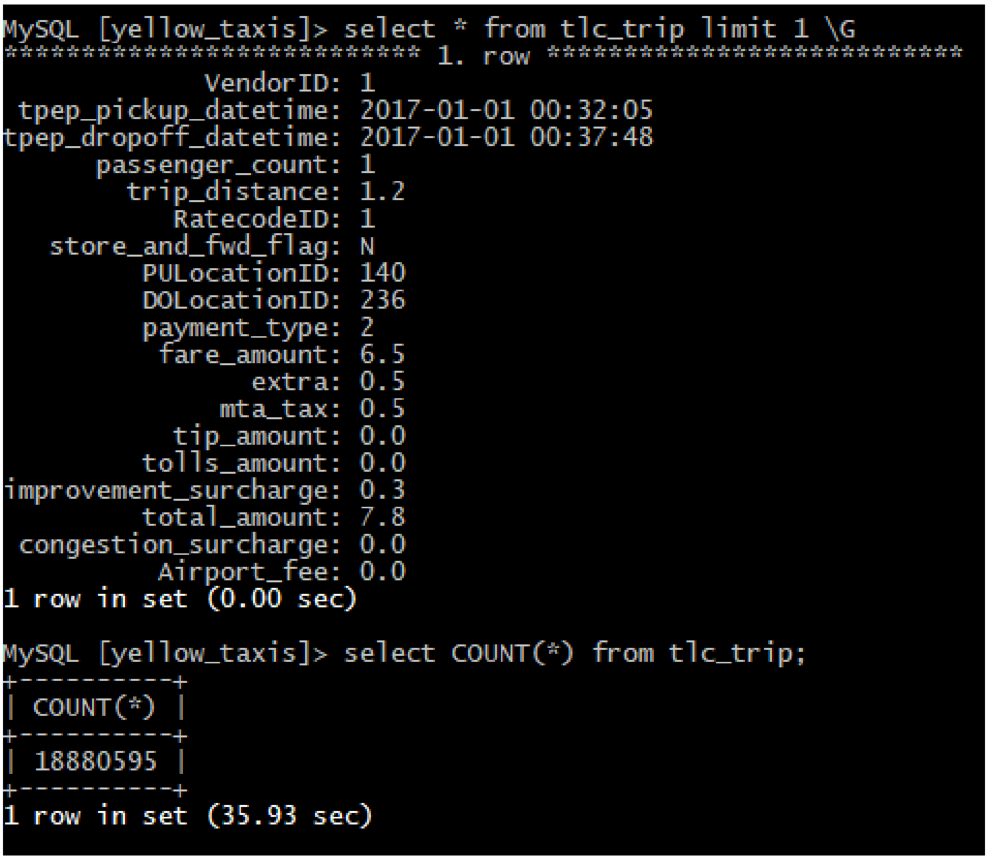
```
LOAD DATA LOCAL INFILE '/home/hadoop/case_study/yellow_tripdata_2017-02.csv'
INTO TABLE tlc_trip
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
IGNORE 1 LINES;
```



```
MySQL [yellow_taxis]> LOAD DATA LOCAL INFILE '/home/hadoop/case_study/yellow_tripdata_2017-02.csv'
    -> INTO TABLE tlc_trip
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;
Query OK, 9169775 rows affected, 65535 warnings (2 min 6.03 sec)
Records: 9169775  Deleted: 0  Skipped: 0  Warnings: 18339845
```

Make sure the data load successfully

```
select * from tlc_trip limit 1 \G
select COUNT(*) from tlc_trip;
```

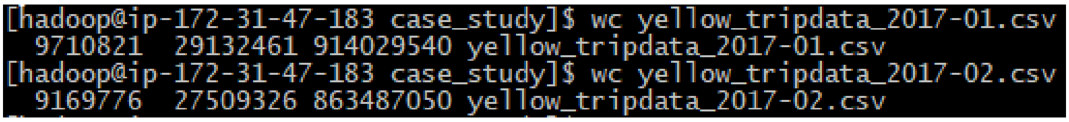


```
MySQL [yellow_taxis]> select * from tlc_trip limit 1 \G
*************************** 1. row ***************************
             VendorID: 1
 tpep_pickup_datetime: 2017-01-01 00:32:05
tpep_dropoff_datetime: 2017-01-01 00:37:48
      passenger_count: 1
        trip_distance: 1.2
           RatecodeID: 1
   store_and_fwd_flag: N
         PULocationID: 140
         DOLocationID: 236
         payment_type: 2
          fare_amount: 6.5
                extra: 0.5
              mta_tax: 0.5
           tip_amount: 0.0
         tolls_amount: 0.0
improvement_surcharge: 0.3
         total_amount: 7.8
 congestion_surcharge: 0.0
          Airport_fee: 0.0
1 row in set (0.00 sec)

MySQL [yellow_taxis]> select COUNT(*) from tlc_trip;
+----------+
| COUNT(*) |
+----------+
| 18880595 |
+----------+
1 row in set (35.93 sec)
```

Validate the count with the original data set in EMR instance:

```
wc yellow_tripdata_2017-01.csv
wc yellow_tripdata_2017-02.csv
```



```
[hadoop@ip-172-31-47-183 case_study]$ wc yellow_tripdata_2017-01.csv
  9710821  29132461 914029540 yellow_tripdata_2017-01.csv
[hadoop@ip-172-31-47-183 case_study]$ wc yellow_tripdata_2017-02.csv
  9169776  27509326 863487050 yellow_tripdata_2017-02.csv
```