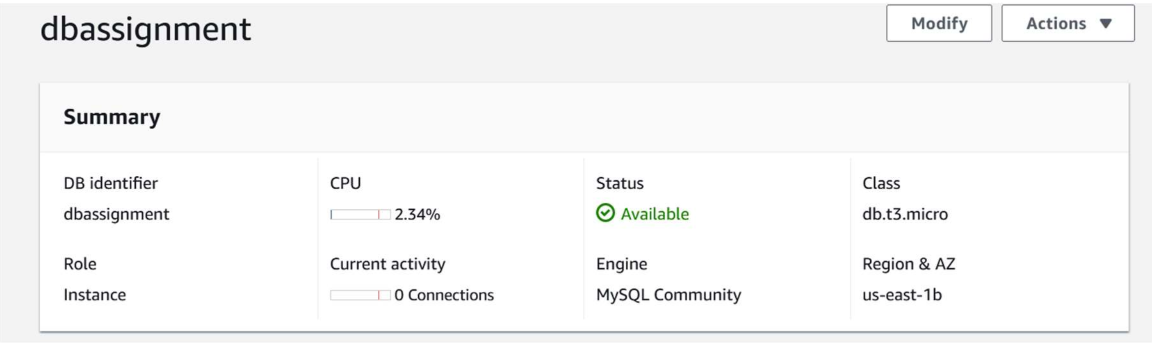
Task 1

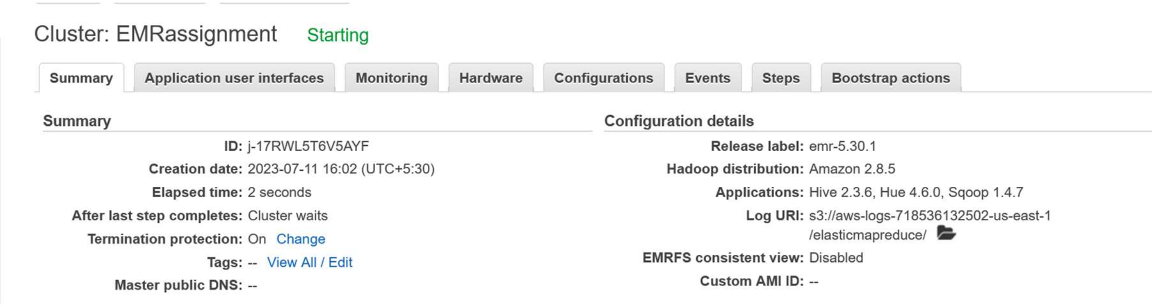


dbassignment

Summary

DB identifier	CPU	Status	Class
dbassignment	2.34%	✓ Available	db.t3.micro
Role	Current activity	Engine	Region & AZ
Instance	0 Connections	MySQL Community	us-east-1b

RDS INSTANCE



Cluster: EMRassignment Starting

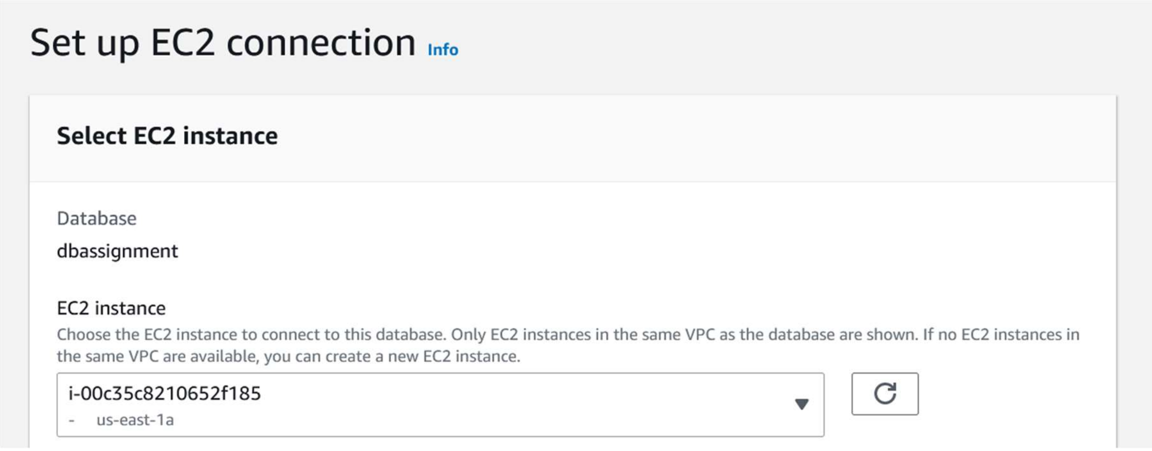
Summary

ID: j-17RWL5T6V5AYF
Creation date: 2023-07-11 16:02 (UTC+5:30)
Elapsed time: 2 seconds
After last step completes: Cluster waits
Termination protection: On Change
Tags: -- View All / Edit
Master public DNS: --

Configuration details

Release label: emr-5.30.1
Hadoop distribution: Amazon 2.8.5
Applications: Hive 2.3.6, Hue 4.6.0, Sqoop 1.4.7
Log URI: s3://aws-logs-718536132502-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

EMR INSTANCE



Set up EC2 connection

Select EC2 instance

Database
dbassignment

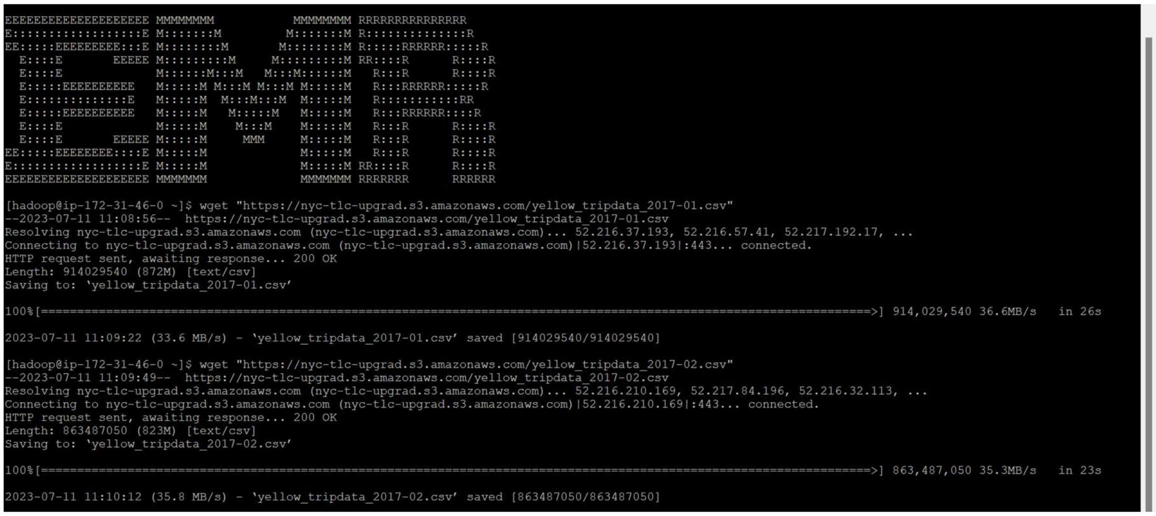
EC2 instance
Choose the EC2 instance to connect to this database. Only EC2 instances in the same VPC as the database are shown. If no EC2 instances in the same VPC are available, you can create a new EC2 instance.

i-00c35c8210652f185
- us-east-1a

CONNECTING RDS DATABASE TO EC2 INSTANACE

CODE FOR DOWNLOADING DATASET:

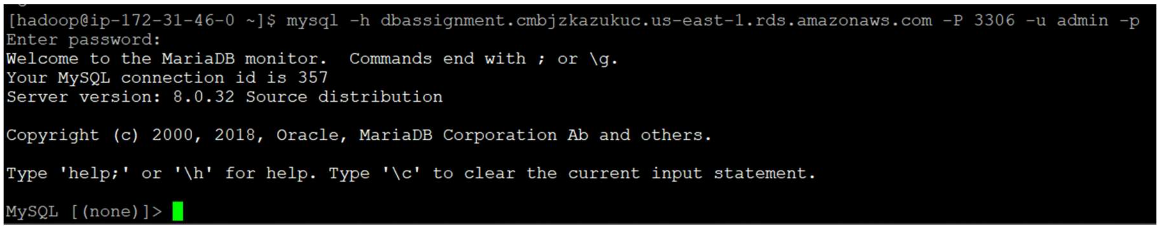
```
wget "https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv"

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

wget "https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-03.csv"

wget "https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-04.csv"

wget "https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-05.csv"

wget "https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-06.csv"
```



```
EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-46-0 ~]$ wget "https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv"
--2023-07-11 11:08:56--  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)... 52.216.37.193, 52.216.57.41, 52.217.192.17, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|52.216.37.193|: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 36.6MB/s   in 26s

2023-07-11 11:09:22 (33.6 MB/s) - ‘yellow_tripdata_2017-01.csv’ saved [914029540/914029540]

[hadoop@ip-172-31-46-0 ~]$ wget "https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv"
--2023-07-11 11:09:49--  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)... 52.216.210.169, 52.217.84.196, 52.216.32.113, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|52.216.210.169|: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 35.3MB/s   in 23s

2023-07-11 11:10:12 (35.8 MB/s) - ‘yellow_tripdata_2017-02.csv’ saved [863487050/863487050]
```

CODE FOR CONNECTING EMR TO RDS DATABASE:

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

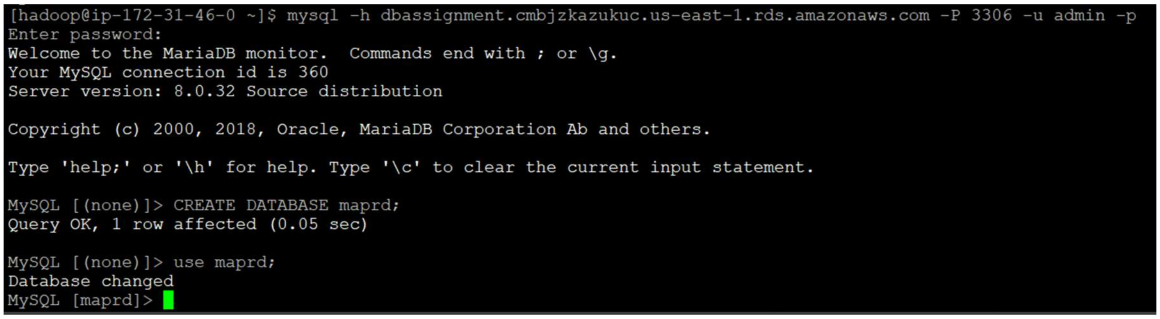


```
[hadoop@ip-172-31-46-0 ~]$ mysql -h dbassignment.cmbjzkazukuc.us-east-1.rds.amazonaws.com -P 3306 -u admin -p
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 357
Server version: 8.0.32 Source distribution

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MySQL [(none)]>
```

CODE FOR CREATING SCHEMA FOR RDS TABLE:

Create database maprd;

Use maprd;



```
[hadoop@ip-172-31-46-0 ~]$ mysql -h dbassignment.cmbjzkazukuc.us-east-1.rds.amazonaws.com -P 3306 -u admin -p
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 360
Server version: 8.0.32 Source distribution

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MySQL [(none)]> CREATE DATABASE maprd;
Query OK, 1 row affected (0.05 sec)

MySQL [(none)]> use maprd;
Database changed
MySQL [maprd]> 
```

CREATE A TABLE IN THE DATABASE

CREATE TABLE trip_log

(

VendorID INT,

tpep_pickup_datetime VARCHAR(50),

tpep_dropoff_datetime VARCHAR(50),

Passenger_count INT,

Trip_distance FLOAT,

RatecodeID INT,

store_and_fwd_flag VARCHAR(2),

PULocationID INT,

DOLocationID INT,

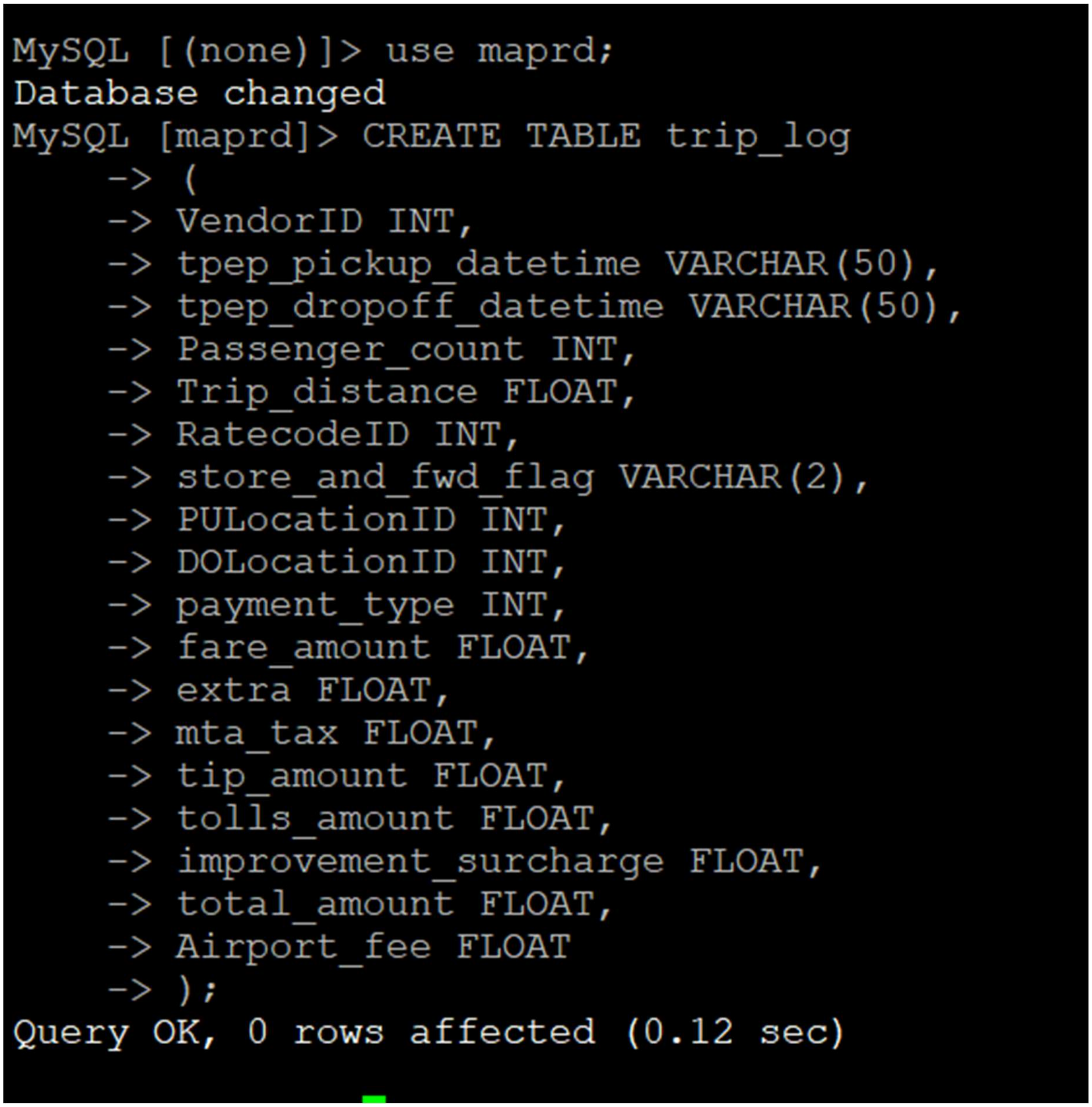
payment_type INT,

fare_amount FLOAT,

extra FLOAT,

mta_tax FLOAT,

```
tip_amount FLOAT,

tolls_amount FLOAT,

improvement_surcharge FLOAT,

total_amount FLOAT,

Airport_fee FLOAT

);
```



```
MySQL [(none)]> use maprd;
Database changed
MySQL [maprd]> CREATE TABLE trip_log
    -> (
    -> VendorID INT,
    -> tpep_pickup_datetime VARCHAR(50),
    -> tpep_dropoff_datetime VARCHAR(50),
    -> Passenger_count INT,
    -> Trip_distance FLOAT,
    -> RatecodeID INT,
    -> store_and_fwd_flag VARCHAR(2),
    -> 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
    -> );
Query OK, 0 rows affected (0.12 sec)
```

UPLOADING DATASETS TO THE DATABASE

Here we 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).

LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv'

INTO TABLE trip_log

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

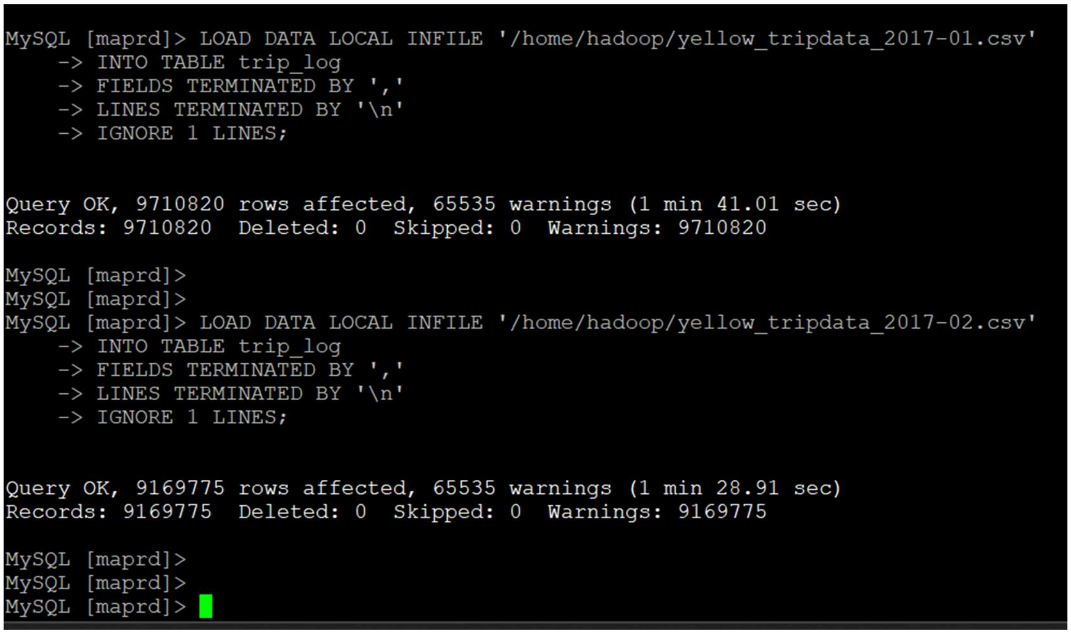
LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv'

INTO TABLE trip_log

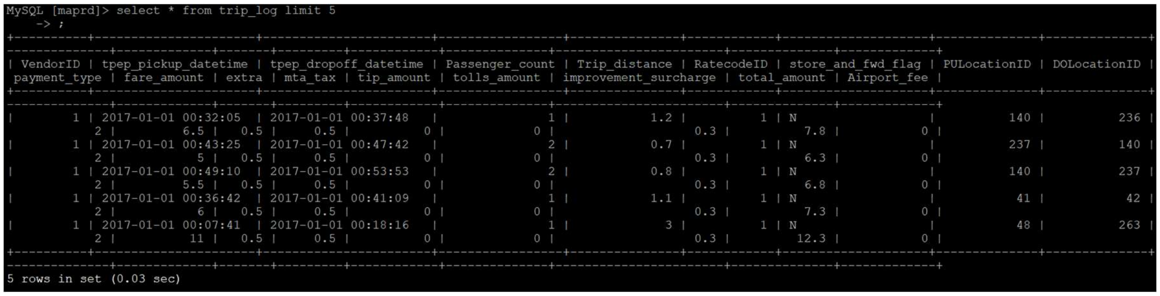
FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

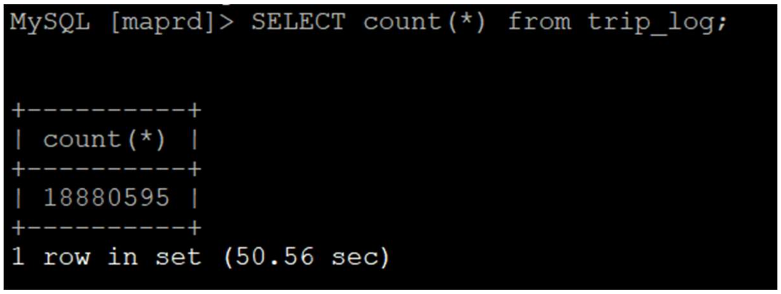
IGNORE 1 LINES;



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


Query OK, 9710820 rows affected, 65535 warnings (1 min 41.01 sec)
Records: 9710820  Deleted: 0  Skipped: 0  Warnings: 9710820

MySQL [maprd]>
MySQL [maprd]>
MySQL [maprd]> LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv'
    -> INTO TABLE trip_log
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES;


Query OK, 9169775 rows affected, 65535 warnings (1 min 28.91 sec)
Records: 9169775  Deleted: 0  Skipped: 0  Warnings: 9169775

MySQL [maprd]>
MySQL [maprd]>
MySQL [maprd]>
```



```
MySQL [maprd]> select * from trip_log limit 5
    -> ;
+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+--------------+--------------+
--------------+-------------+-------+---------+------------+--------------+-----------------------+--------------+-------------+
| VendorID | tpep_pickup_datetime | tpep_dropoff_datetime | Passenger_count | Trip_distance | RatecodeID | store_and_fwd_flag | PULocationID | DOLocationID |
 payment_type | fare_amount | extra | mta_tax | tip_amount | tolls_amount | improvement_surcharge | total_amount | Airport_fee |
+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+--------------+--------------+
--------------+-------------+-------+---------+------------+--------------+-----------------------+--------------+-------------+
|        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 |
|        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 |
|        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 |
|        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 |
|        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 |
+----------+----------------------+-----------------------+-----------------+---------------+------------+--------------------+--------------+--------------+
--------------+-------------+-------+---------+------------+--------------+-----------------------+--------------+-------------+
5 rows in set (0.03 sec)
```



```
MySQL [maprd]> SELECT count(*) from trip_log;


+----------+
| count(*) |
+----------+
| 18880595 |
+----------+
1 row in set (50.56 sec)
```

CODE FOR CREATING HBASE TABLE:

```
Hbase shell

create ‘tlc_data’,'trip_info'
```

CODE FOR IMPORTING INTO HBASE TABLE FROM RDS TABLE USING SQOOP:

```
sqoop import \

--connect "jdbc:mysql:// dbassignment.cmbjzkazukuc.us-east1.rds.amazonaws.com/maprd" \

--username admin -P \

--table trip_log \

--hbase-table YELLOW_TLC_DATA\

--columns VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,Passenger_count,Trip_distance,PULocationId,DOLocationID,RateCodeID,Store_and_fwd_flag,Payment_type,Fare_amount,Extra,MTA_tax,Improvement_surchange,Tip_amount,Tolls_amount,Total_amount,Congstion_Surcharge,Airport_fee,id \

--column-family trip_info \

--hbase-row-key id
```