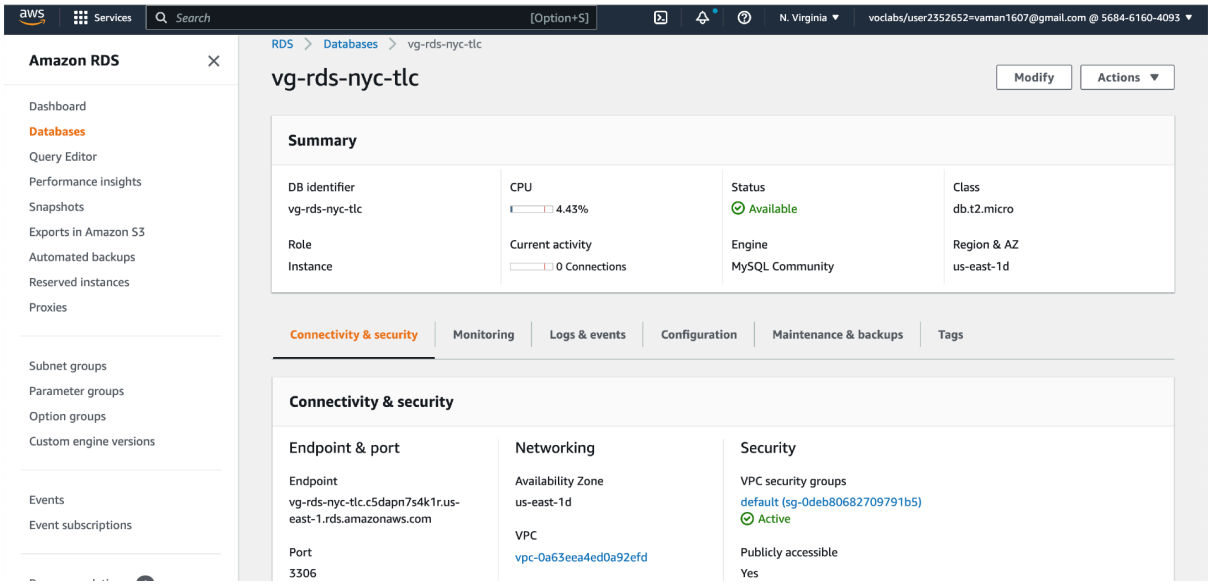
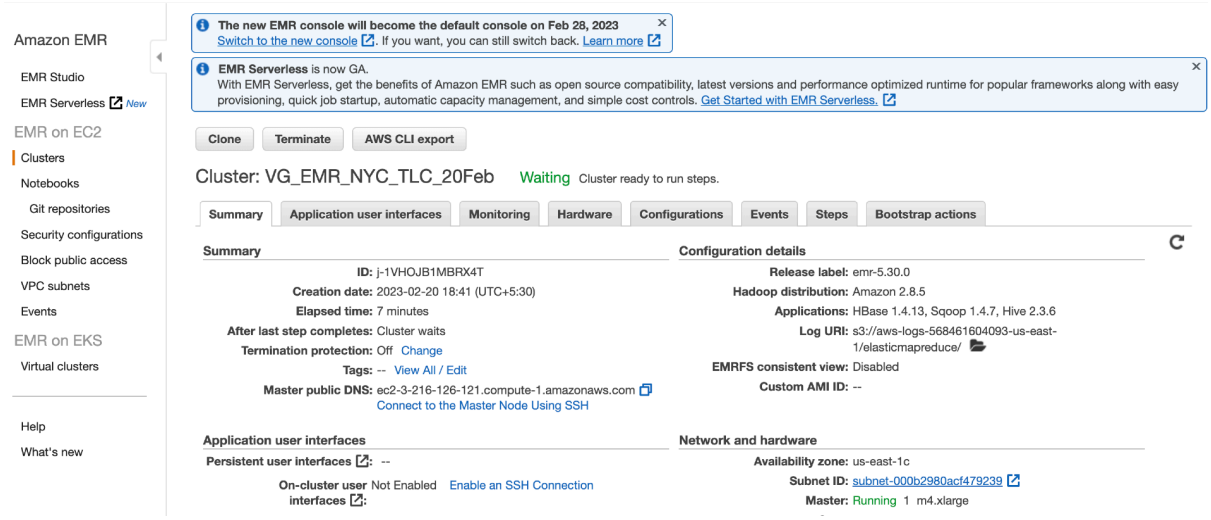
Datasets:

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

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

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

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

https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-05.csv
https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-06.csv
```

Create RDS Instance vg-rds-nyc-tlc



Create EMR Cluster::



Connect to the EMR:

Install MySQL drivers

Login as root using sudo -i

Download the datasets into root

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

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

```
Remove the header of the datasets

sed –i '1d' yellow_tripdata_2017–01.csv

sed –i '1d' yellow_tripdata_2017–02.csv

Copy the files to Hadoop FS
```

hadoop fs -put /root/yellow_tripdata_2017-01.csv /user/root/yellow_trip_data1
hadoop fs -put /root/yellow_tripdata_2017-02.csv /user/root/yellow_trip_data2

Connect to RDS instance:

mysql -h vg-rds-nyc-tlc.c5dapn7s4k1r.us-east-1.rds.amazonaws.com -u admin -p

Create database nyc_tlc

```
[root@ip-172-31-11-67 ~]# mysql -h vg-rds-nyc-tlc.c5dapn7s4k1r.us-east-1.rds.amazonaws.com -u admin -p
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 20
Server version: 8.0.28 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 nyc_tlc;
Query OK, 1 row affected (0.01 sec)

MySQL [(none)]> use nyc_tlc
Database changed
MySQL [nyc_tlc]>
```

Create a temp table without auto increment rowkey

```
Create table nyc_tlc.yellow_tripdata_temp(
VendorID int,
tpep_pickup_datetime datetime,
tpep_dropoff_datetime datetime,
passenger_count int,
trip_distance double,
RatecodeID int,
store_and_fwd_flag varchar(150),
PULocationID int,
DOLocationID int,
payment_type int,
fare_amount double,
extra double,
mta_tax double,
tip_amount double,
tolls_amount double,
improvement_surcharge double,
total_amount double,
congestion_surcharge double,
airport_fee double);
```

Create table yellow_pathdata

```
Create table nyc_tlc.yellow_tripdata(
rowkey MEDIUMINT NOT NULL AUTO_INCREMENT,
VendorID int,
tpep_pickup_datetime datetime,
tpep_dropoff_datetime datetime,
passenger_count int,
trip_distance double,
RatecodeID int,
store_and_fwd_flag varchar(150),
PULocationID int,
DOLocationID int,
payment_type int,
fare_amount double,
extra double,
mta_tax double,
tip_amount double,
tolls_amount double,
improvement_surcharge double,
total_amount double,
congestion_surcharge double,
airport_fee double,
PRIMARY KEY (rowkey)
);
```

```
Database changed
MySQL [nyc_tlc]> Create table nyc_tlc.yellow_tripdata(
    -> rowkey MEDIUMINT NOT NULL AUTO_INCREMENT,
    -> VendorID int,
    -> tpep_pickup_datetime datetime,
    -> tpep_dropoff_datetime datetime,
    -> passenger_count int,
    -> trip_distance double,
    -> RatecodeID int,
    -> store_and_fwd_flag varchar(150),
    -> PULocationID int,
    -> DOLocationID int,
    -> payment_type int,
    -> fare_amount double,
    -> extra double,
    -> mta_tax double,
    -> tip_amount double,
    -> tolls_amount double,
    -> improvement_surcharge double,
    -> total_amount double,
    -> congestion_surcharge double,
    -> airport_fee double,
    -> PRIMARY KEY (rowkey)
    -> );
Query OK, 0 rows affected (0.04 sec)

MySQL [nyc_tlc]> show tables;
+-------------------+
| Tables_in_nyc_tlc |
+-------------------+
| yellow_tripdata   |
+-------------------+
1 row in set (0.00 sec)
```

Using Scoop Export to load data from csv into temp table.

sqoop export --connect jdbc:mysql://vg-rds-nyc-tlc.c5dapn7s4k1r.us-east-1.rds.amazonaws.com:3306/nyc_tlc ****
--table yellow_tripdata_temp ****
--username admin --password admin**123 **
--export-dir /user/root/yellow_trip_data1 ****
--fields-terminated-by ',' --lines-terminated-by '\n'

sqoop export --connect jdbc:mysql://vg-rds-nyc-tlc.c5dapn7s4k1r.us-east-1.rds.amazonaws.com:3306/nyc_tlc ****
--table yellow_tripdata_temp ****
--username admin --password admin**123 **
--export-dir /user/root/yellow_trip_data2 ****
--fields-terminated-by ',' --lines-terminated-by '\n'

```
Database changed
MySQL [nyc_tlc]> show tables;
+------------------+
| Tables_in_nyc_tlc |
+------------------+
| yellow_tripdata   |
+------------------+
1 row in set (0.01 sec)

MySQL [nyc_tlc]> select count(*) from yellow_tripdata;
+----------+
| count(*) |
+----------+
| 18880595 |
+----------+


[root@ip-172-31-11-104 ~]# wc -l yellow_tripdata_2017-01.csv
9710820 yellow_tripdata_2017-01.csv
[root@ip-172-31-11-104 ~]# wc -l yellow_tripdata_2017-02.csv
9169775 yellow_tripdata_2017-02.csv
```

File	Count
yellow_tripdata_2017-01.csv	9710820
yellow_tripdata_2017-02.csv	9169775
Total	**18880595**

Inserting into yellow_tripdata which has an auto increment rowkey:

INSERT INOT yellow_tripdata
(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,
congestion_surcharge,
airport_fee double)
Select 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,
congestion_surcharge,
airport_fee double
From yellow_tripdata_temp;
```