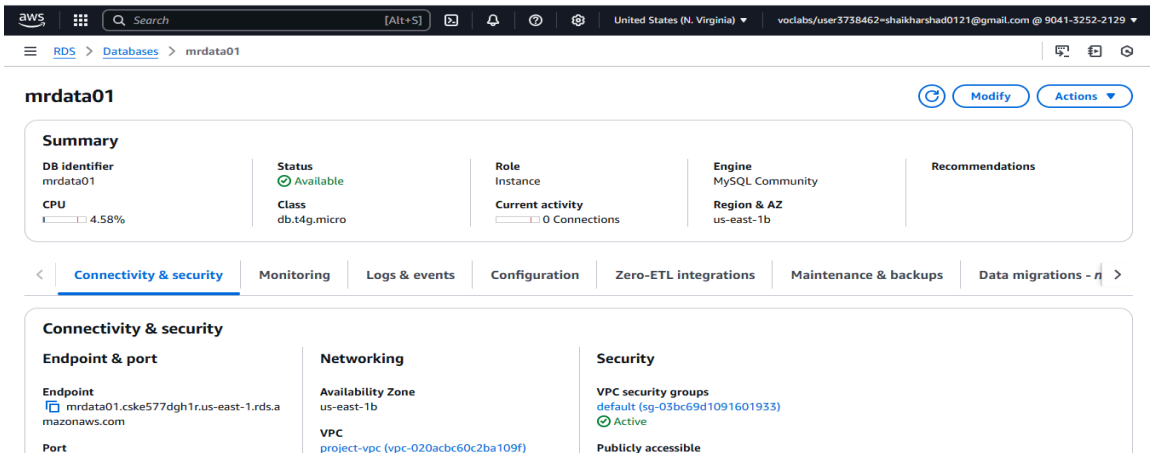
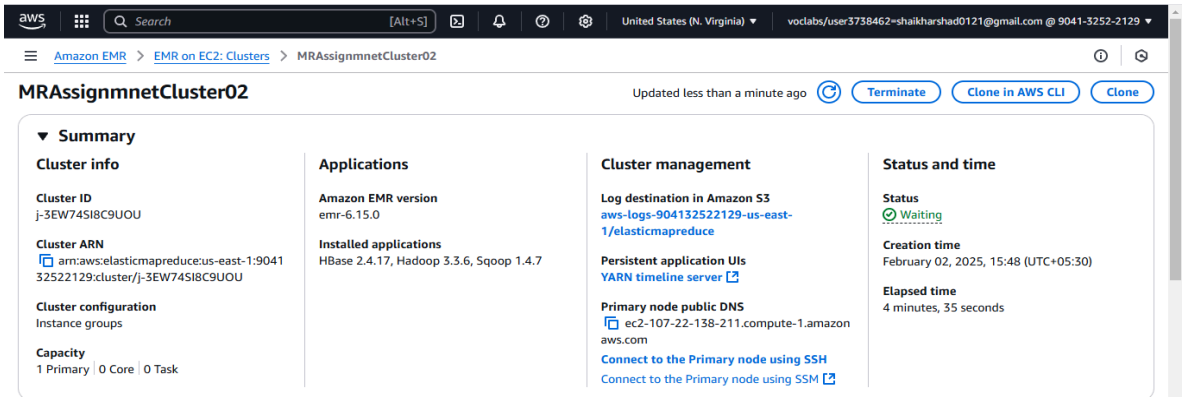
TASK 01

Create an RDS instance in your AWS account and upload the data from two files (yellow_tripdata_2017-01.csv & yellow_tripdata_2017-02.csv) from the dataset. Make sure to create an appropriate schema for the data sets before uploading them to RDS.

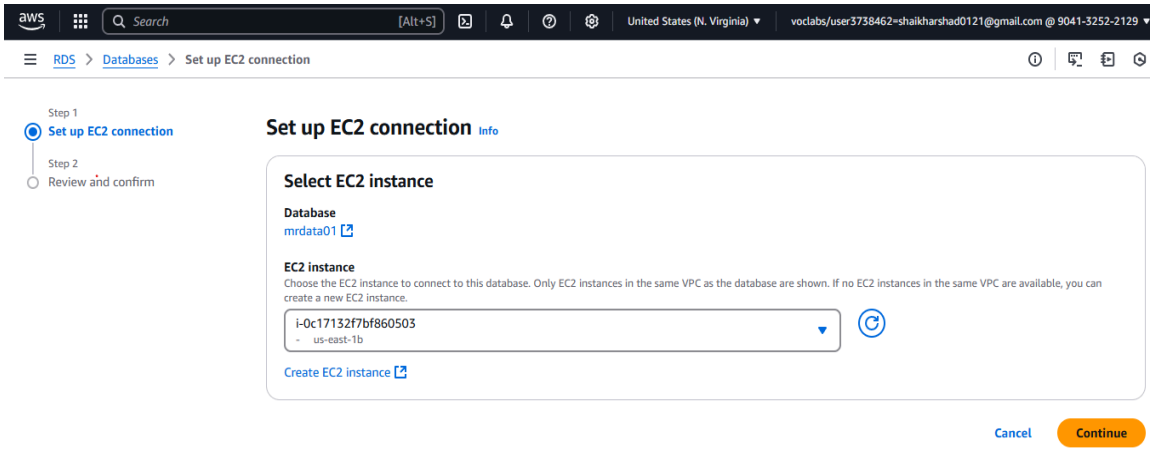
A. Create a RDS instance in AWS

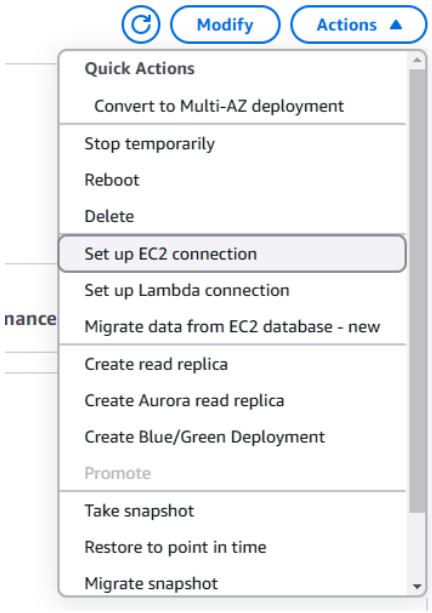


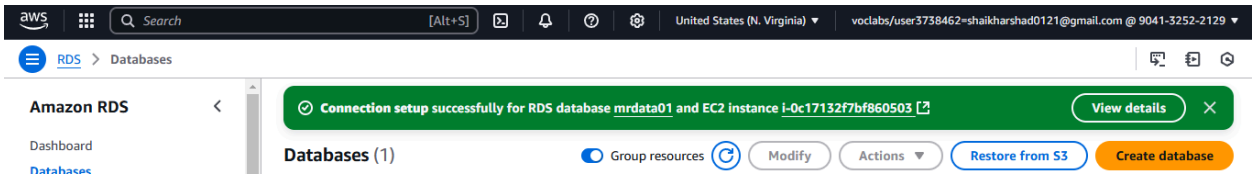
B. Create EMR instance with Hadoop, Apache Sqoop and Apache HBase



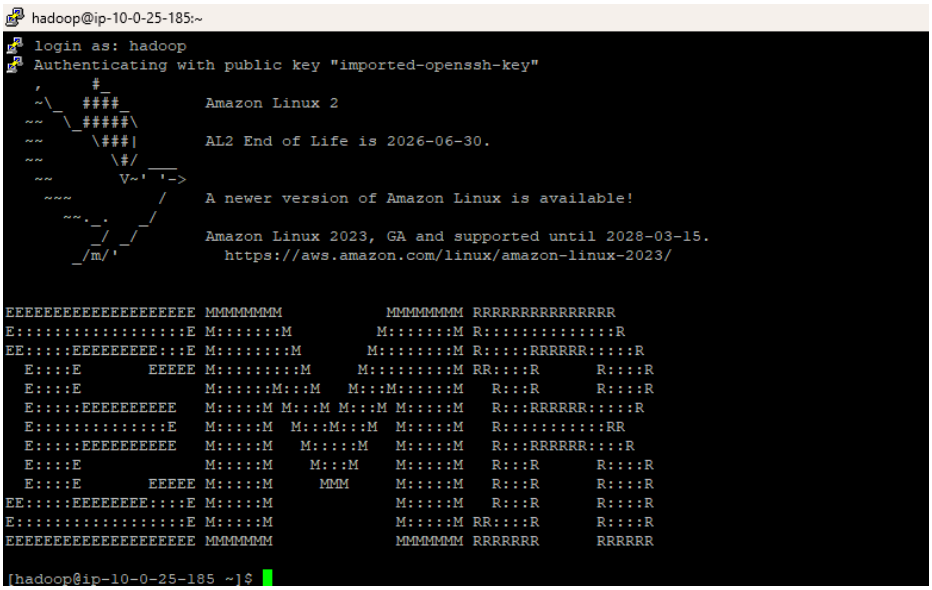
C. After creating a RDS instance, Click Action → Set up EC2 connection. Also make sure to have MySQL added in the security groups.





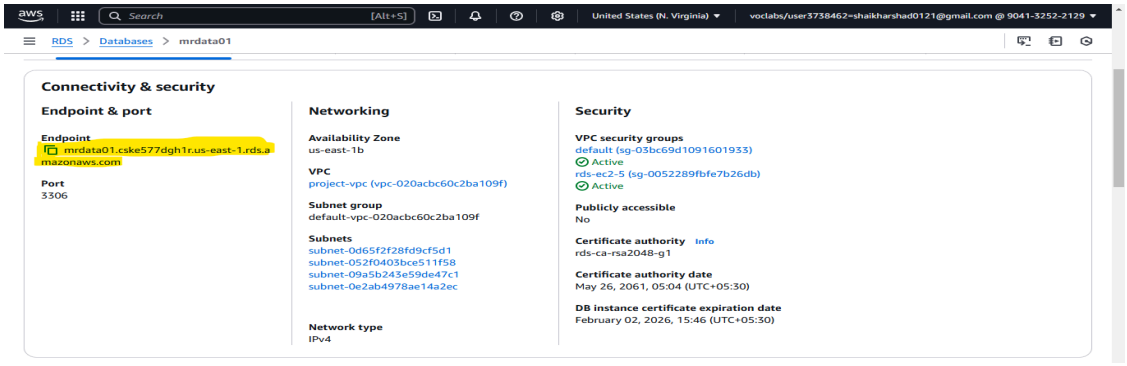


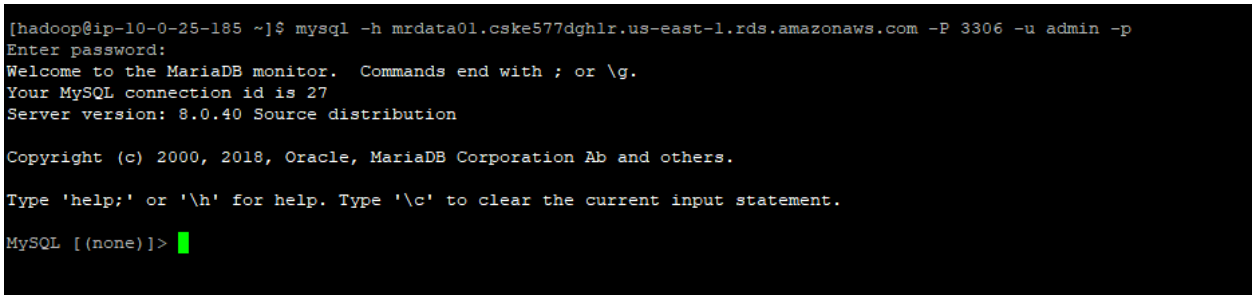
D. Login to the terminal using putty.



E. Use the following syntax to connect to MySql:

mysql -h **RDS Endpoint** -P 3306 -u admin -p





F. Create the database and the table schema

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

MySQL [(none)]> CREATE DATABASE ride_data;
Query OK, 1 row affected (0.01 sec)

MySQL [(none)]> USE ride_data;
Database changed
MySQL [ride_data]> CREATE TABLE ride_log (
    ->     VendorID INT,
    ->     pickup_datetime VARCHAR(50),
    extra FLOAT,
    mta_tax FLOAT,
    ->     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 (0.04 sec)

MySQL [ride_data]>
```

G. Get both the 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

```
MySQL [ride_data]> exit;
Bye
[hadoop@ip-10-0-25-185 ~]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
--2025-02-02 10:30:15--  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)... 16.15.177.5, 3.5.29.41, 52.216.213.129, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|16.15.177.5|: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 36.3MB/s   in 24s

2025-02-02 10:30:40 (36.2 MB/s) - 'yellow_tripdata_2017-01.csv' saved [914029540/914029540]

[hadoop@ip-10-0-25-185 ~]$ wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
--2025-02-02 10:30:43--  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)... 3.5.7.144, 52.216.53.57, 52.216.216.1, ...
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)|3.5.7.144|: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 39.5MB/s   in 22s

2025-02-02 10:31:05 (38.0 MB/s) - 'yellow_tripdata_2017-02.csv' saved [863487050/863487050]

[hadoop@ip-10-0-25-185 ~]$
```

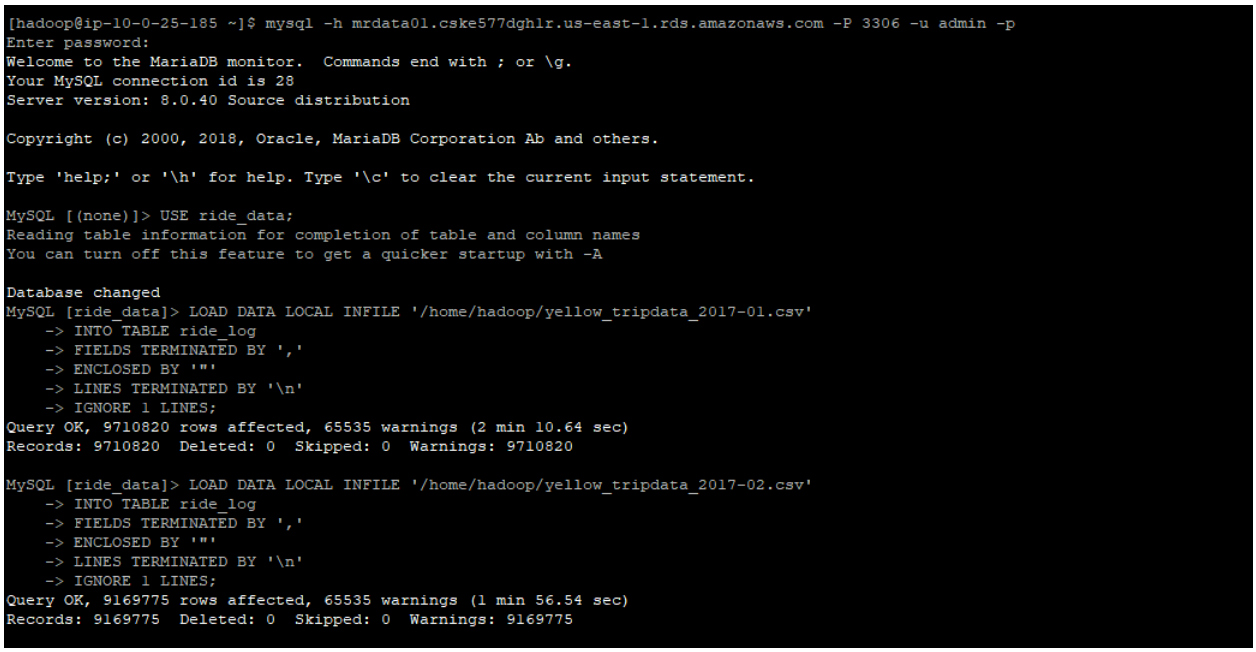
H. Login to MySQL again and write the following syntax to load the data:

LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv'
INTO TABLE ride_log
FIELDS TERMINATED BY ','
ENCLOSED BY ""

```
LINES TERMINATED BY '\n'
IGNORE 1 LINES;

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

Once this syntax run the data is loaded into our RDS table.



```
Records: 9169775  Deleted: 0  Skipped: 0  Warnings: 9169775

MySQL [ride_data]> SELECT COUNT(*) FROM ride_log;
+----------+
| COUNT(*) |
+----------+
| 18880595 |
+----------+
1 row in set (54.52 sec)

MySQL [ride_data]>
MySQL [ride_data]> SELECT * FROM ride_log LIMIT 10;
+----------+---------------------+---------------------+-----------------+---------------+------------+--------------------+--------------+--------------+--------------+
| VendorID | pickup_datetime     | 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 |
|        1 | 2017-01-01 00:20:52 | 2017-01-01 00:24:59 |               2 |           0.7 |          1 | N                  |          236 |          262 |            2 |
|          5 |   0.5 |     0.5 |          0 |           0 |                   0.3 |         6.3 |          0 |
|        1 | 2017-01-01 00:33:49 | 2017-01-01 00:42:38 |               2 |           1.6 |          1 | N                  |          236 |          238 |            1 |
|          8 |   0.5 |     0.5 |       1.85 |           0 |                   0.3 |       11.15 |          0 |
|        1 | 2017-01-01 00:48:22 | 2017-01-01 00:52:15 |               2 |           0.6 |          1 | N                  |          238 |          239 |            1 |
|          5 |   0.5 |     0.5 |       1.25 |           0 |                   0.3 |        7.55 |          0 |
|        1 | 2017-01-01 00:57:12 | 2017-01-01 01:06:28 |               2 |             1 |          1 | N                  |          239 |           48 |            1 |
|        7.5 |   0.5 |     0.5 |       1.75 |           0 |                   0.3 |       10.55 |          0 |
|        1 | 2017-01-01 00:10:25 | 2017-01-01 00:29:06 |               1 |             1 |          1 | N                  |          246 |           48 |            2 |
|         12 |   0.5 |     0.5 |          0 |           0 |                   0.3 |        13.3 |          0 |
+----------+---------------------+---------------------+-----------------+---------------+------------+--------------------+--------------+--------------+--------------+
10 rows in set (0.08 sec)

MySQL [ride_data]>
```