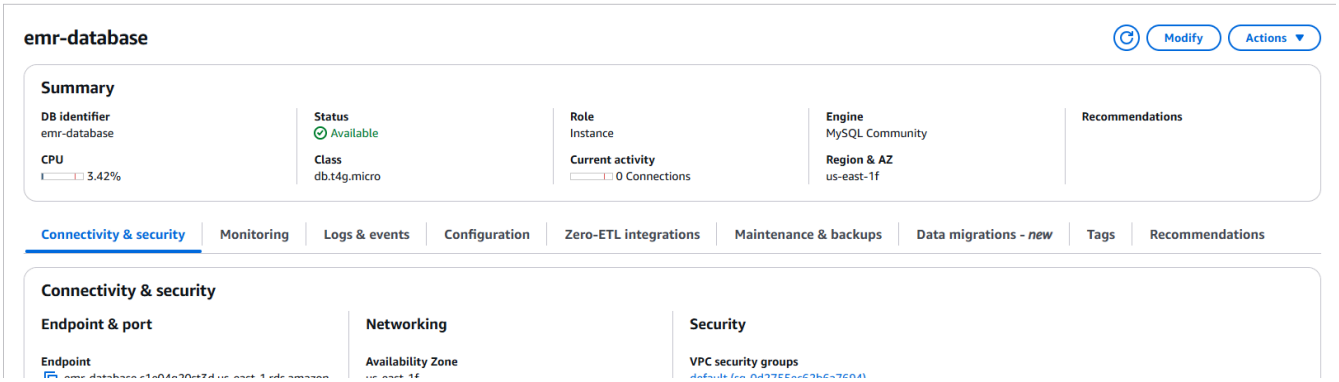
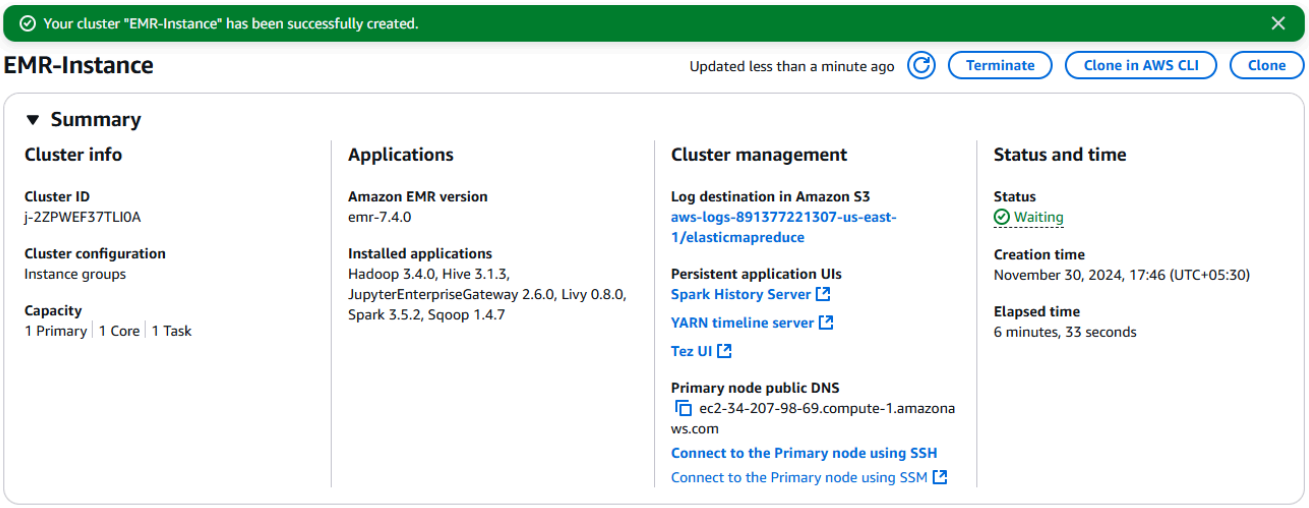
Task 1: Setting up the AWS environment and loading data

Creating an RDS instance in my AWS account and uploading the data to the RDS instance.

1. RDS instance creation in AWS

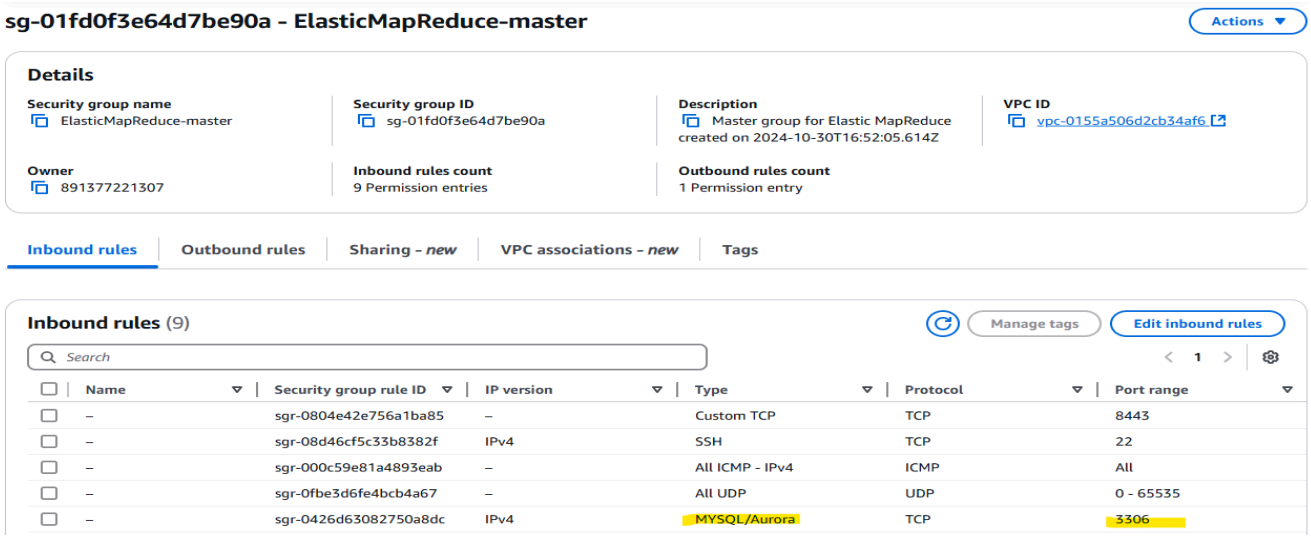


2. EMR creation,some applications selected include: Apache Sqoop, Apache Hbase, Hadoop

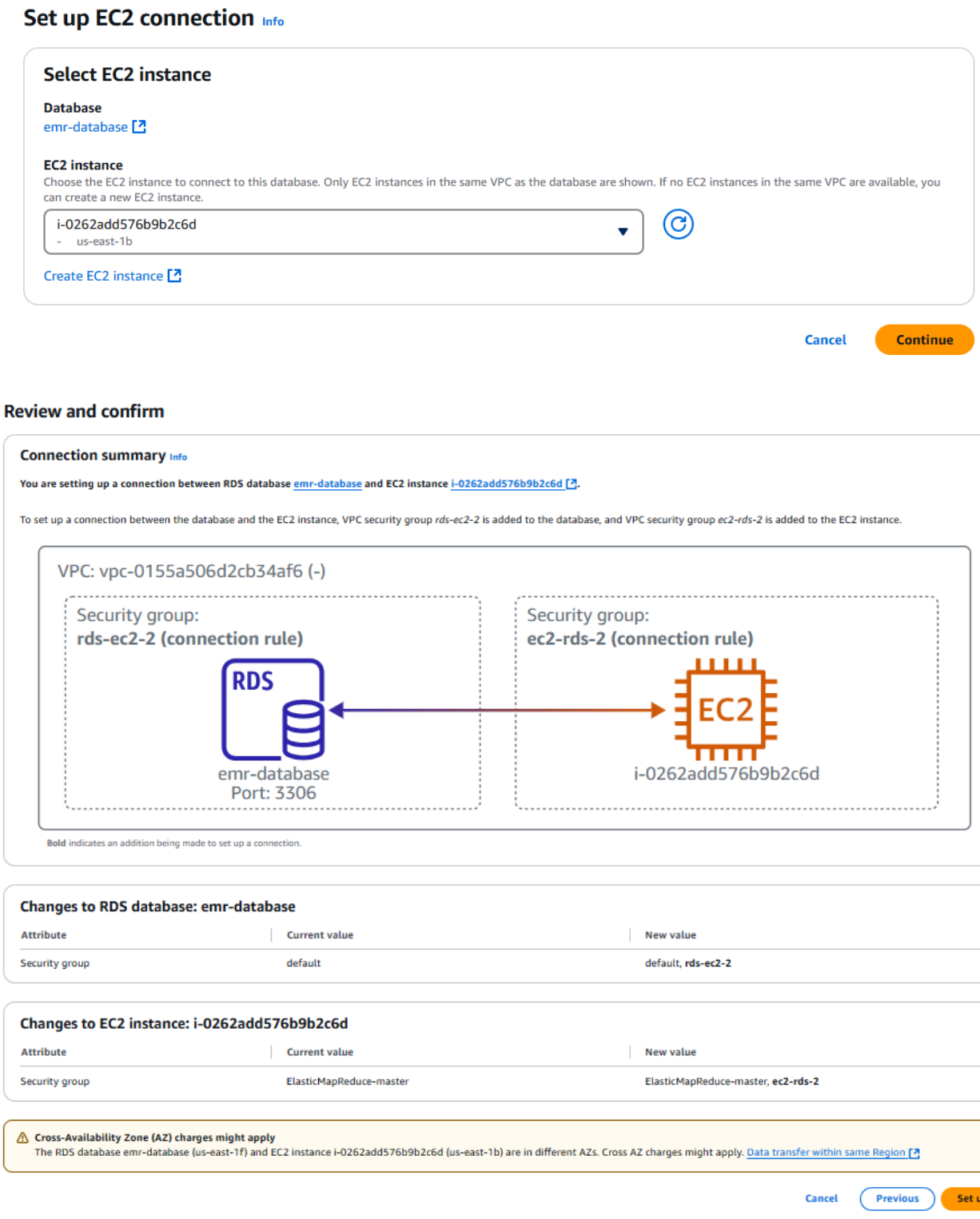


3. Connecting RDS with the EMR instance: -

We configure security group by editing inbound rules



- Then we click on ‘Action’ button on RDS menu and then ‘Set up EC2 connection’



4. We then log into RDS through EMR instance using the command

'mysql -h emr-upgrad-task.c1e04g20st3d.us-east-1.rds.amazonaws.com -P 3306 -u admin -p'

```
[hadoop@ip-172-31-32-112 ~]$ mysql -h emr-database.cle04g20st3d.us-east-1.rds.am
azonaws.com -P 3306 -u admin -p
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 28
Server version: 8.0.39 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)]>
```

5. We then create database using code below:

> create database yellow_taxi;

```
MySQL [(none)]> create database yellow_taxi;
Query OK, 1 row affected (0.007 sec)

MySQL [(none)]> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| mysql              |
| performance_schema |
| sys                |
| yellow_taxi        |
+--------------------+
5 rows in set (0.004 sec)
```

We then create table using code below;

> use yellow_taxi;
> create table trip_records (VendorID INT, tpep_pickup_datetime VARCHAR(255), tpep_dropo_datetime VARCHAR(255),Passenger_count INT, Trip_distance FLOAT, RatecodeID INT, store_and_fwd_ag VARCHAR(50),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 yellow_taxi;
Database changed
MySQL [yellow_taxi]> create table trip_records (VendorID INT, tpep_pickup_dateti
me VARCHAR(255), tpep_dropo_datetime VARCHAR(255),Passenger_count INT, Trip_dist
ance FLOAT, RatecodeID INT, store_and_fwd_ag VARCHAR(50),PULocationID INT, DOLoc
ationID INT, payment_type INT, fare_amount FLOAT, extra FLOAT,mta_tax FLOAT, tip
_amount FLOAT, tolls_amount FLOAT, improvement_surcharge FLOAT,total_amount FLOA
T, Airport_fee FLOAT );
Query OK, 0 rows affected (0.047 sec)
```

6. Downloading required csv files from internet in local using command

` 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-172-31-32-112 ~]$ wget "https://nyc-tlc-upgrad.s3.amazonaws.com/yello
--2024-11-30 12:37:48--  https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata
Resolving nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)... 5
Connecting to nyc-tlc-upgrad.s3.amazonaws.com (nyc-tlc-upgrad.s3.amazonaws.com)
HTTP request sent, awaiting response... 200 OK
Length: 910028408 (868M) [text/csv]
Saving to: 'yellow_tripdata_2017-06.csv'

yellow_tripdata_2017-06.csv                         100%[==============

2024-11-30 12:38:16 (31.4 MB/s) - 'yellow_tripdata_2017-06.csv' saved [91002840
```

7. To load data in mysql table we have to login and then run sql command:

LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv' INTO TABLE trip_records
FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;
LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv' INTO TABLE trip_records
FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 LINES;

SELECT COUNT(*) FROM taxi_records.trip_log;

```
MySQL [yellow_taxi]> SELECT COUNT(*) FROM trip_records;
+----------+
| COUNT(*) |
+----------+
| 18880595 |
+----------+
1 row in set (52.842 sec)
```