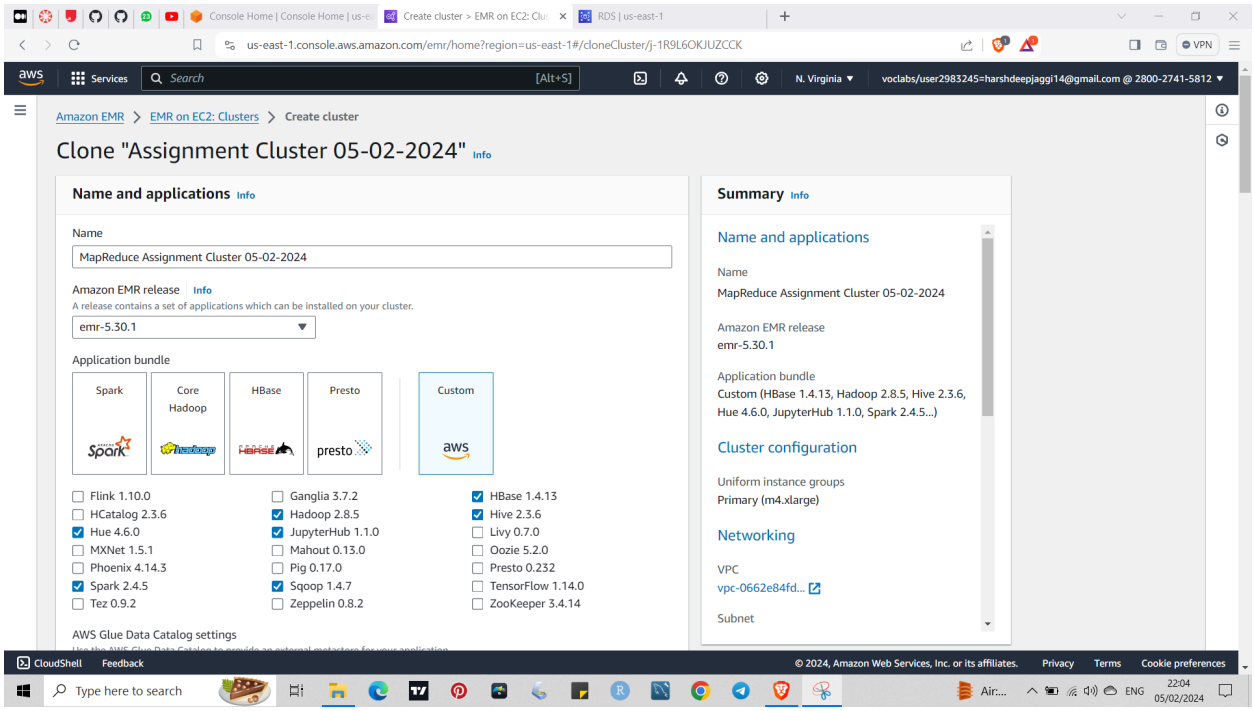
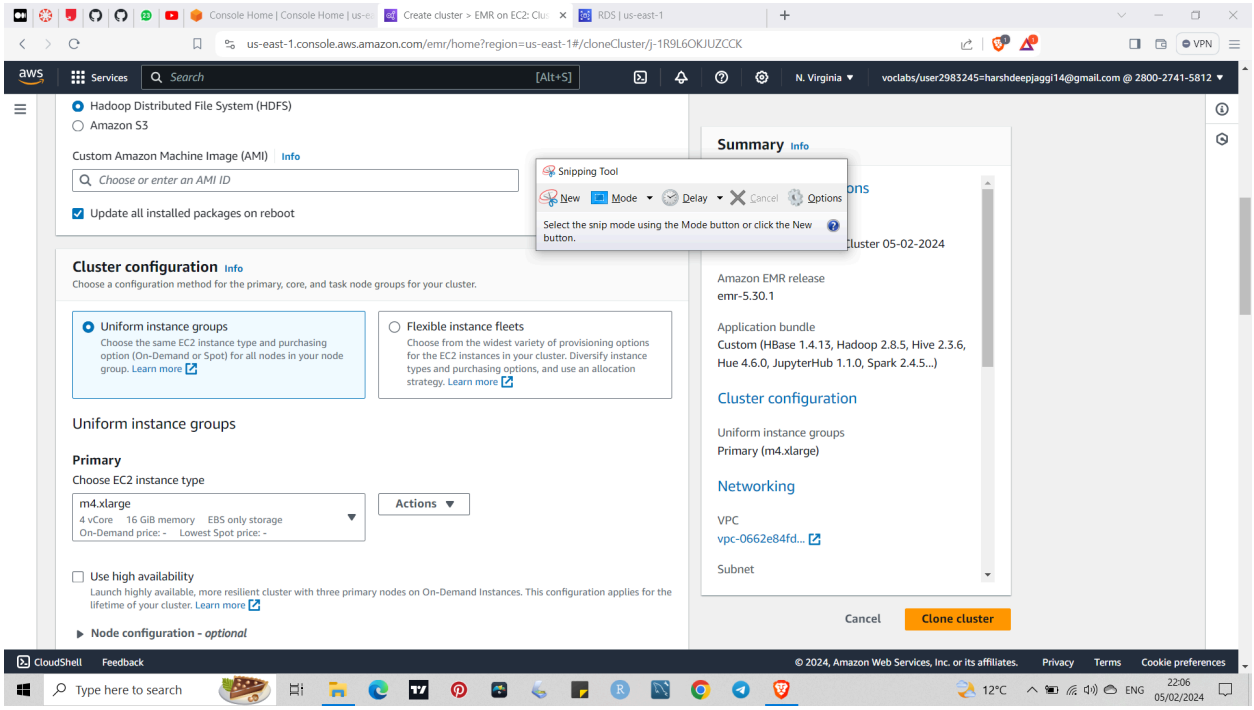
Setting up EMR Cluster

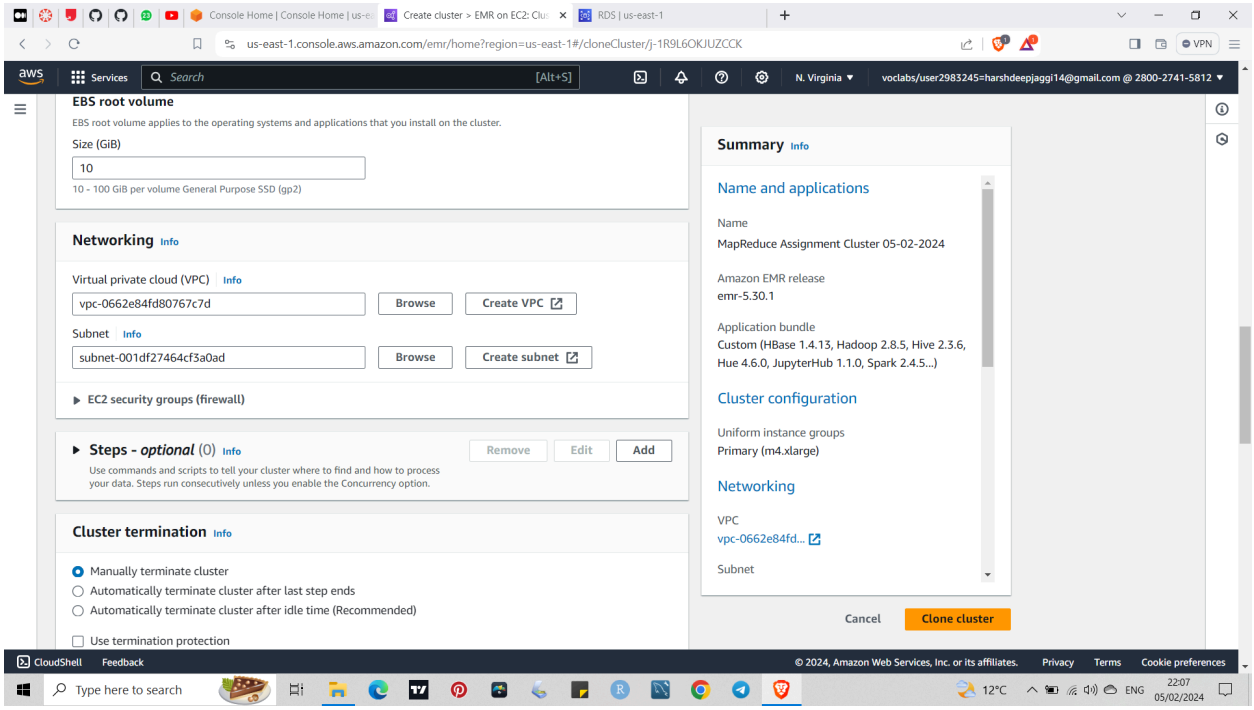
1. Creation of EMR Cluster with **Spark, Hadoop, Sqoop, HBase, Hive, JupyterHub, Hue.**



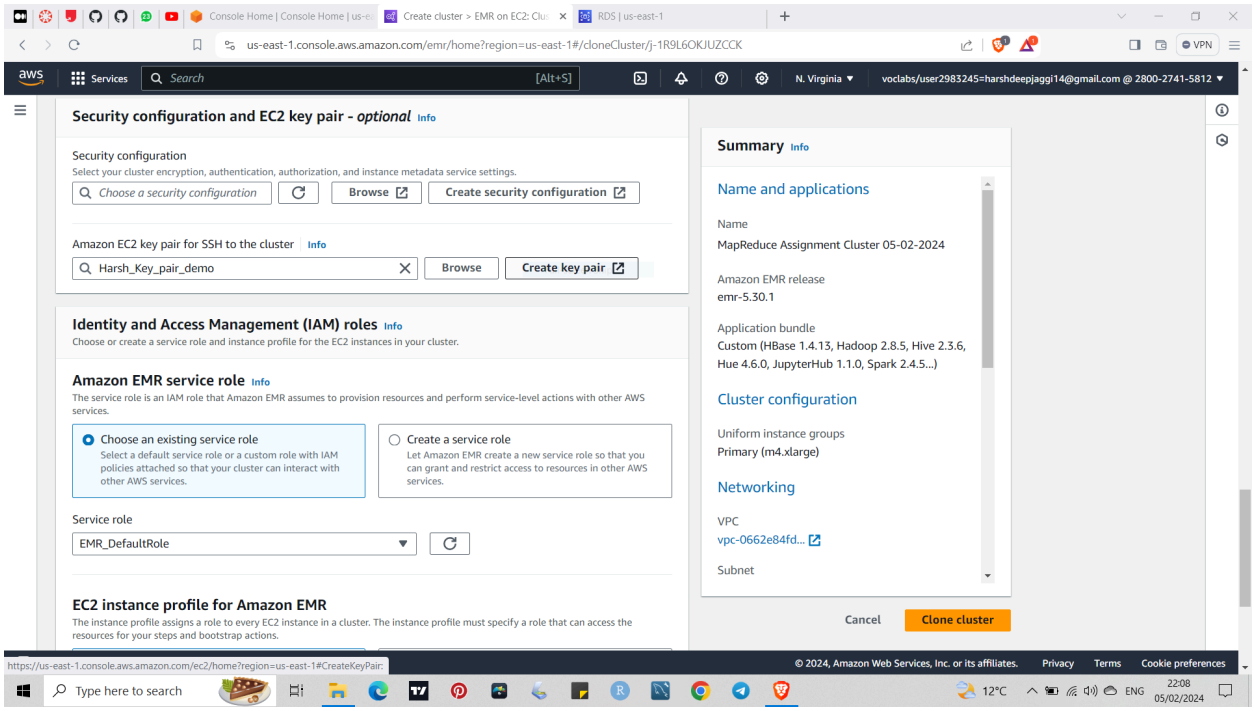
2. Selected **m4.xlarge** as EC2 instance type.



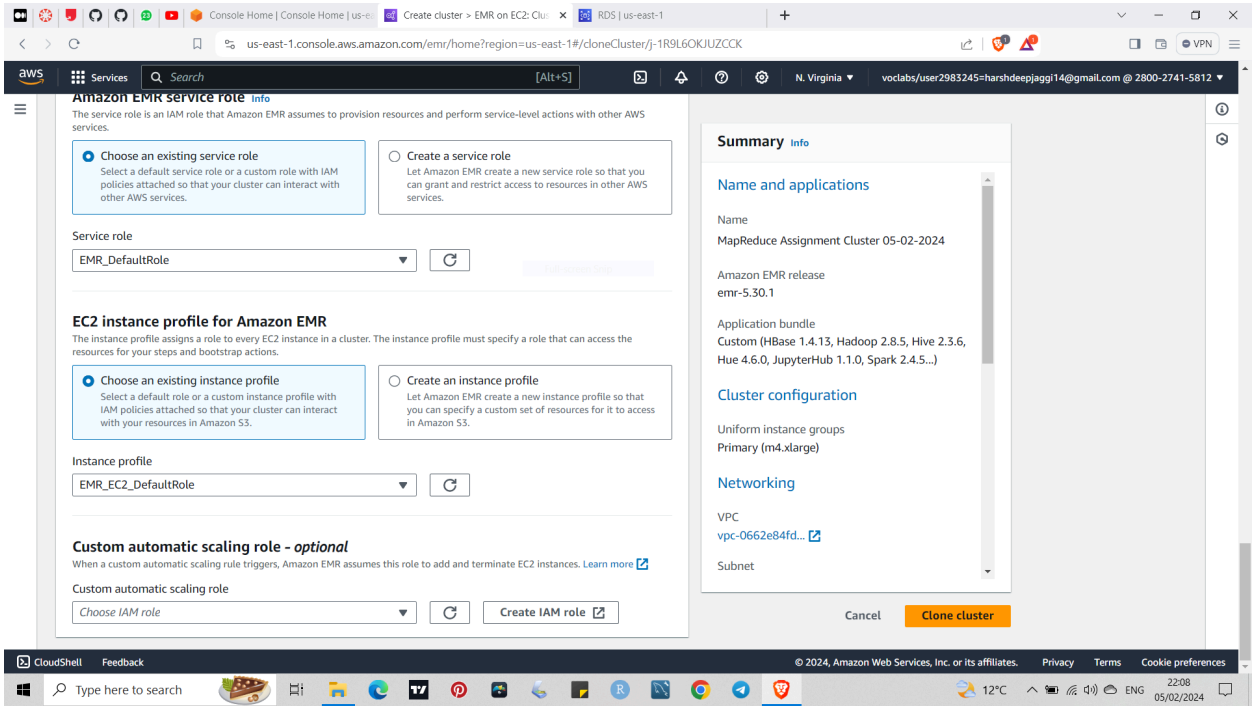
3. Selected **VPC** & **Subnet** which I had created earlier.



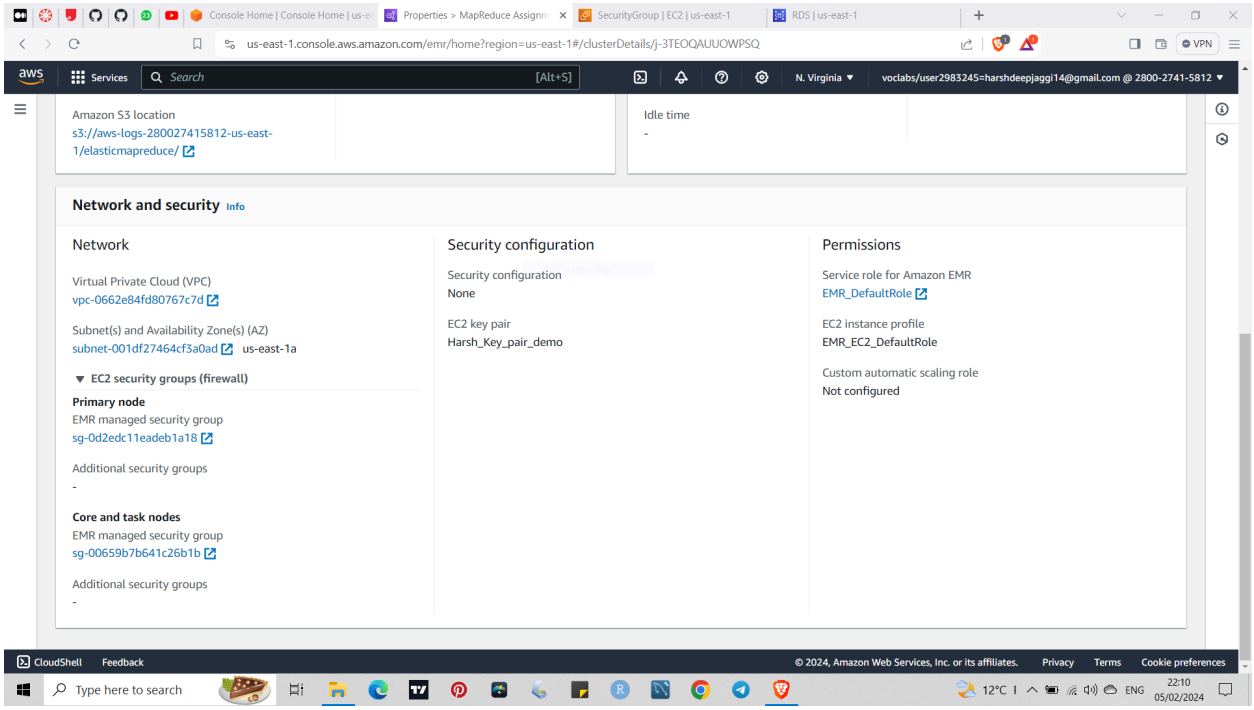
4. Selected **EC2 Key Pair** which is present in my local system & chosen Amazon EMR Service Role as **EMR_DefaultRole.**



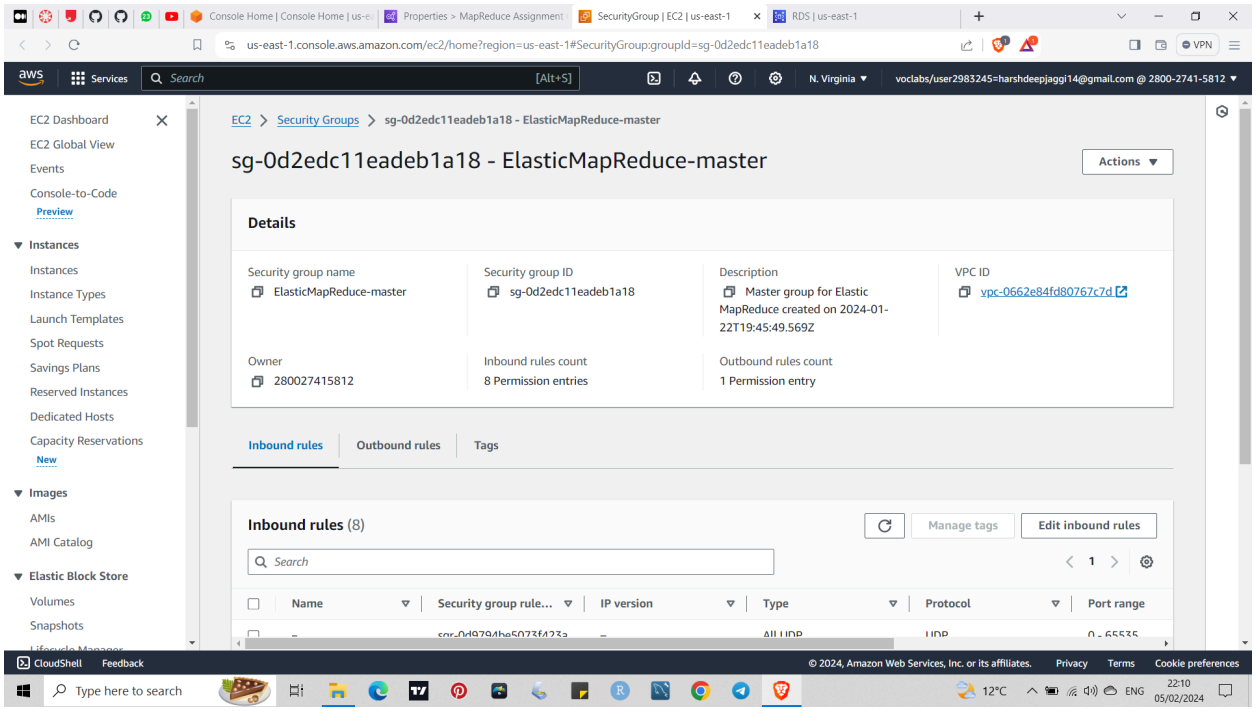
5. Selected Instance profile as **EMR_EC2_DefaultRole** & then clicked on the **Clone Custer** option at the bottom right in the Orange Box.



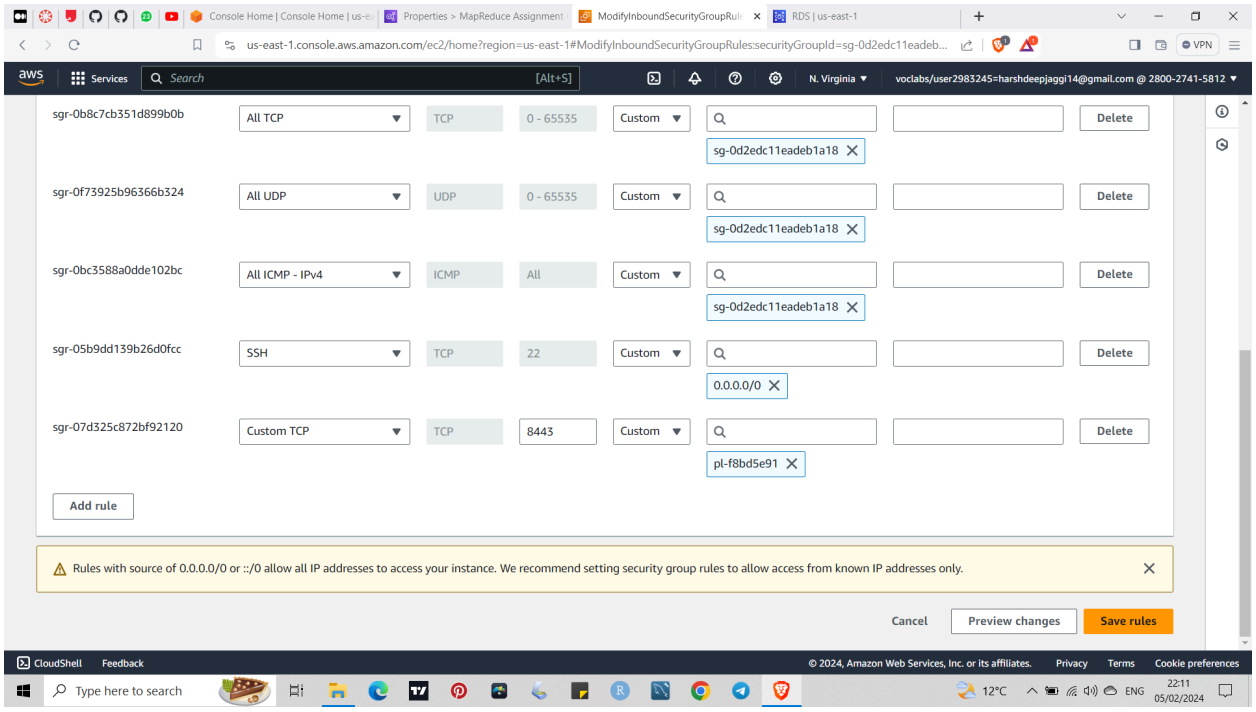
6. After clicking the clone cluster option, I changed **EC2 Security Group Settings.**



7. Click on the **Edit Inbound Rules** Option.

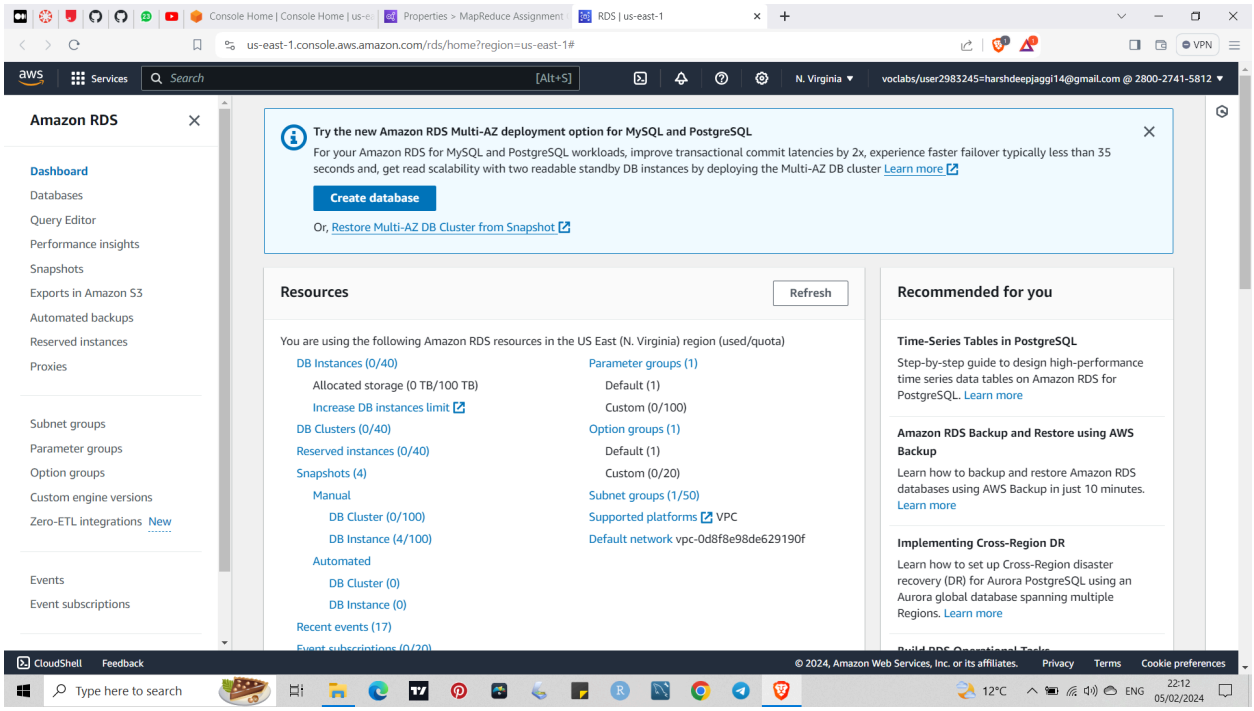


8. Added a **New Rule** with **SSH, Port: 22, AnywhereIPv-4** and then saved it.

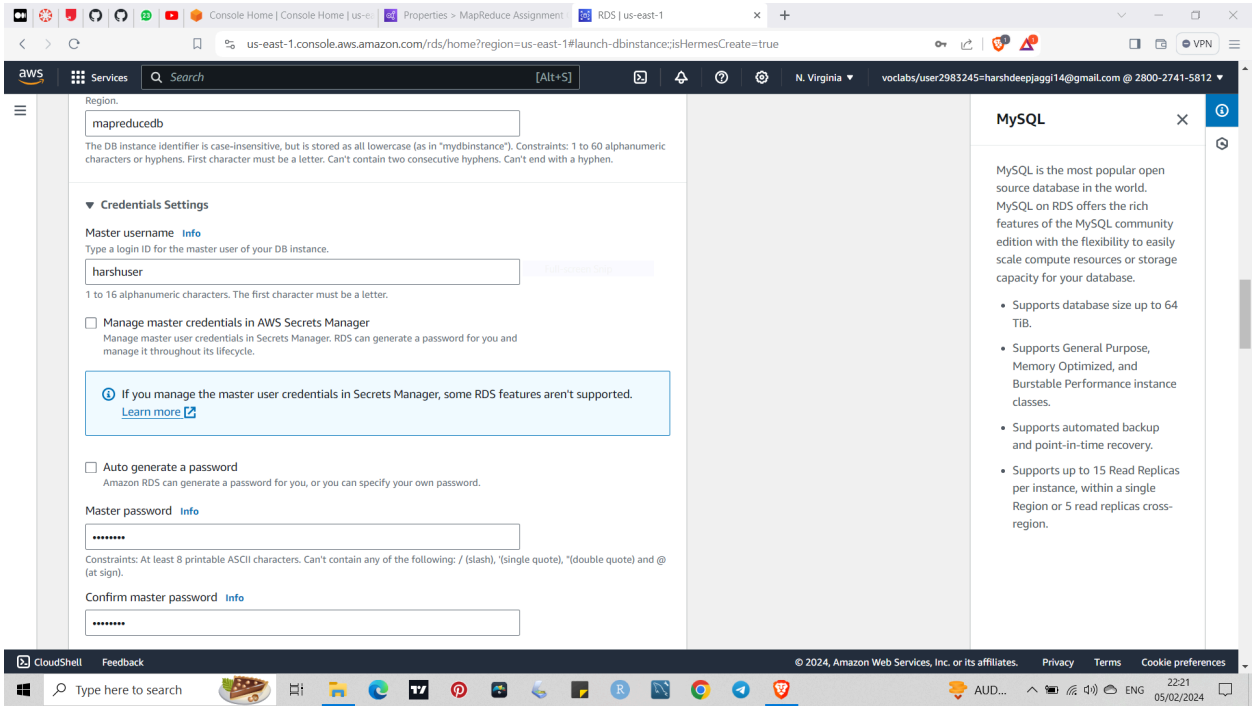


Setting up RDS Instance

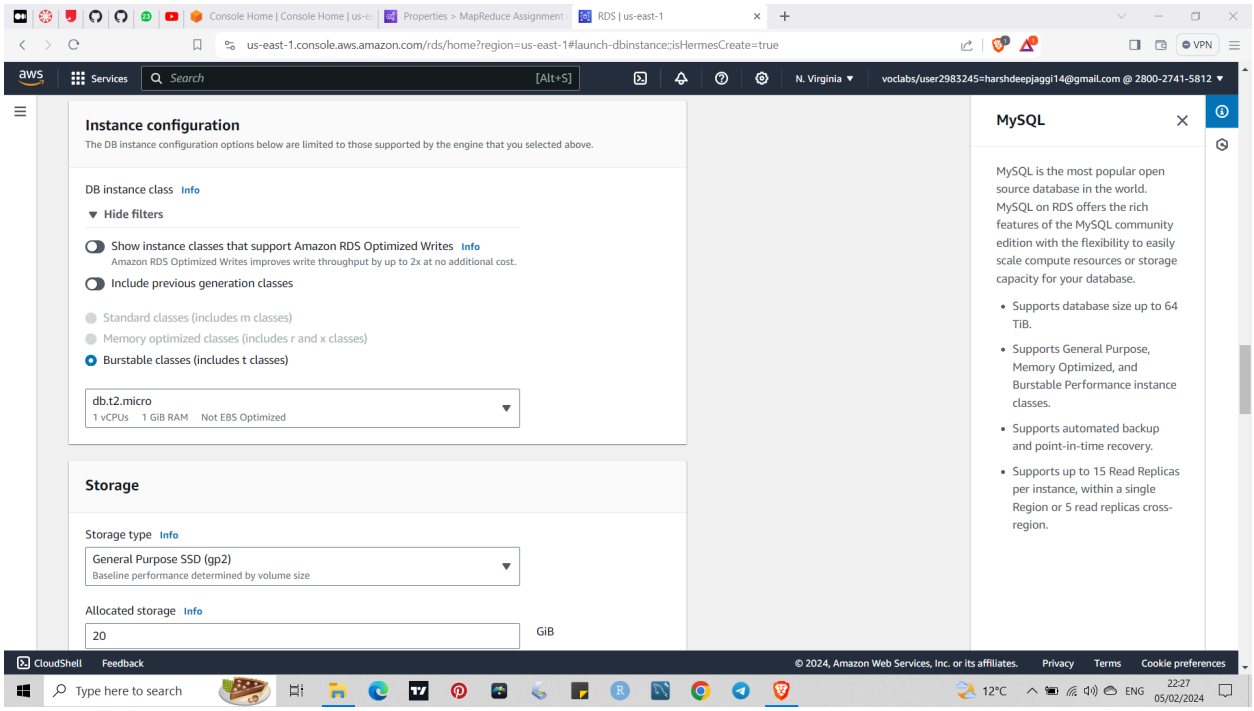
1. Click on **Create Database** Option.



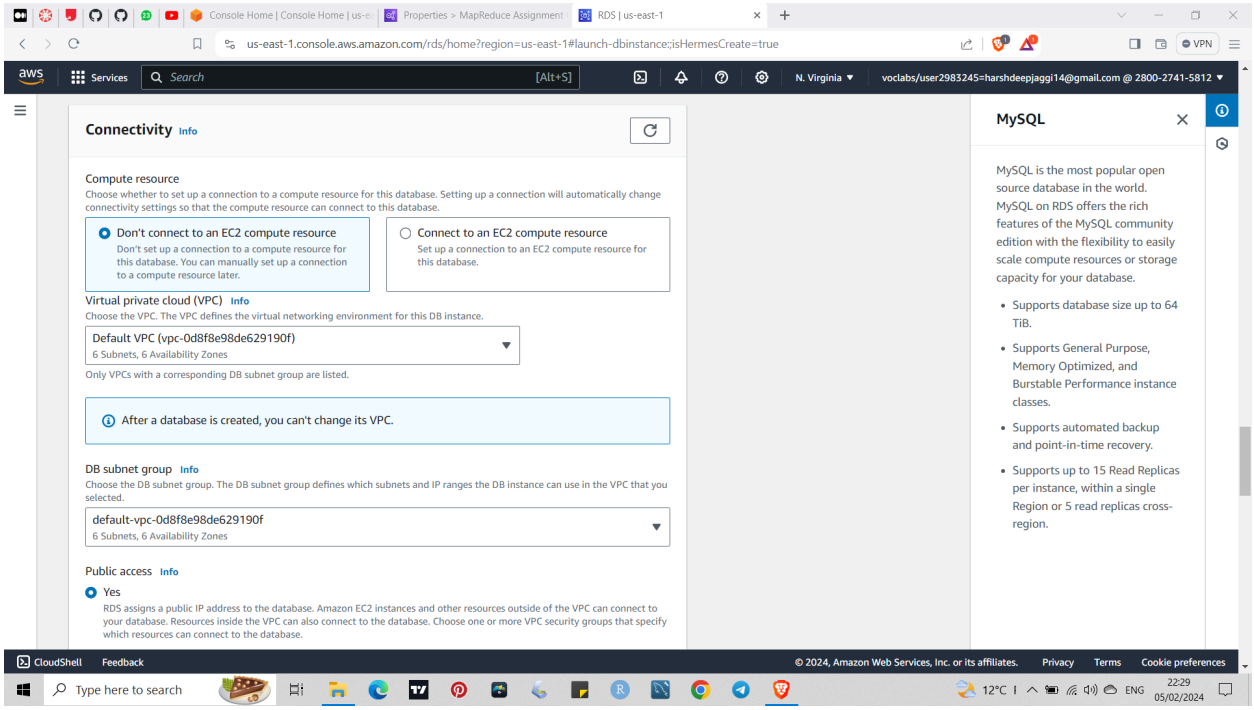
2. Providing the **DataBase** name as **mapreducedb**; **Master Username** to **harshuser**; created **password** too.



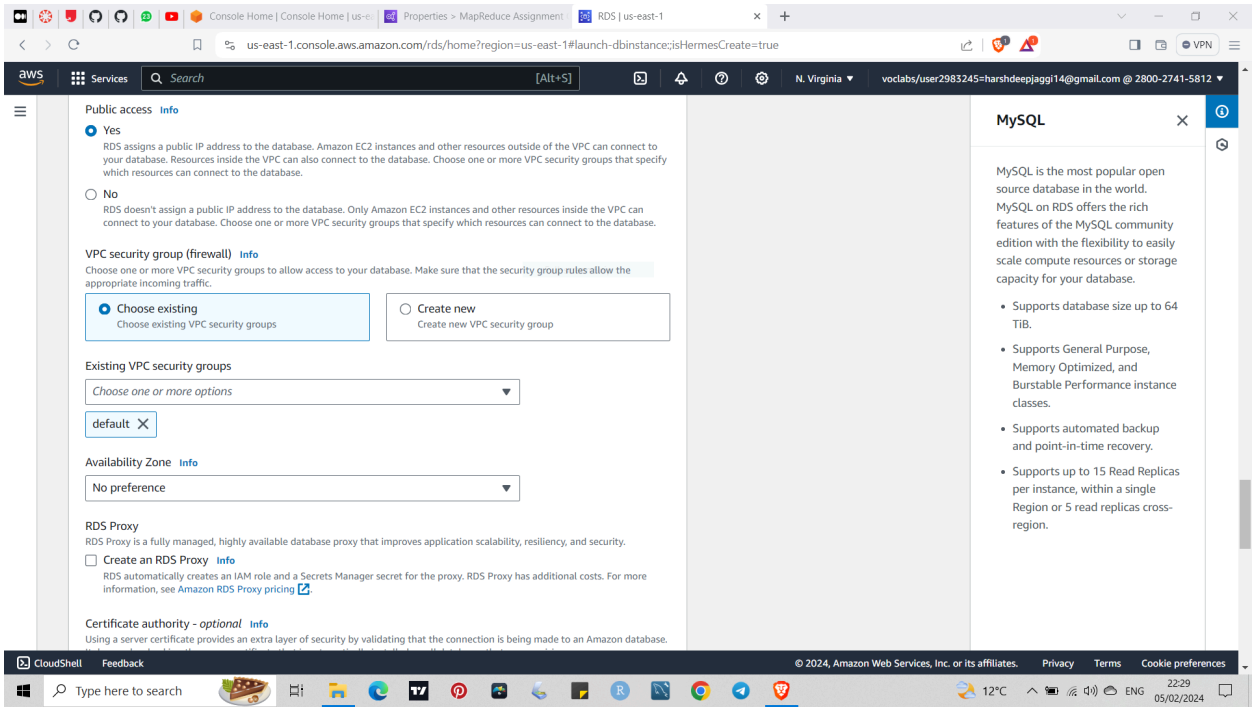
3. Selected **DB Instance class** as **db.t2.micro**

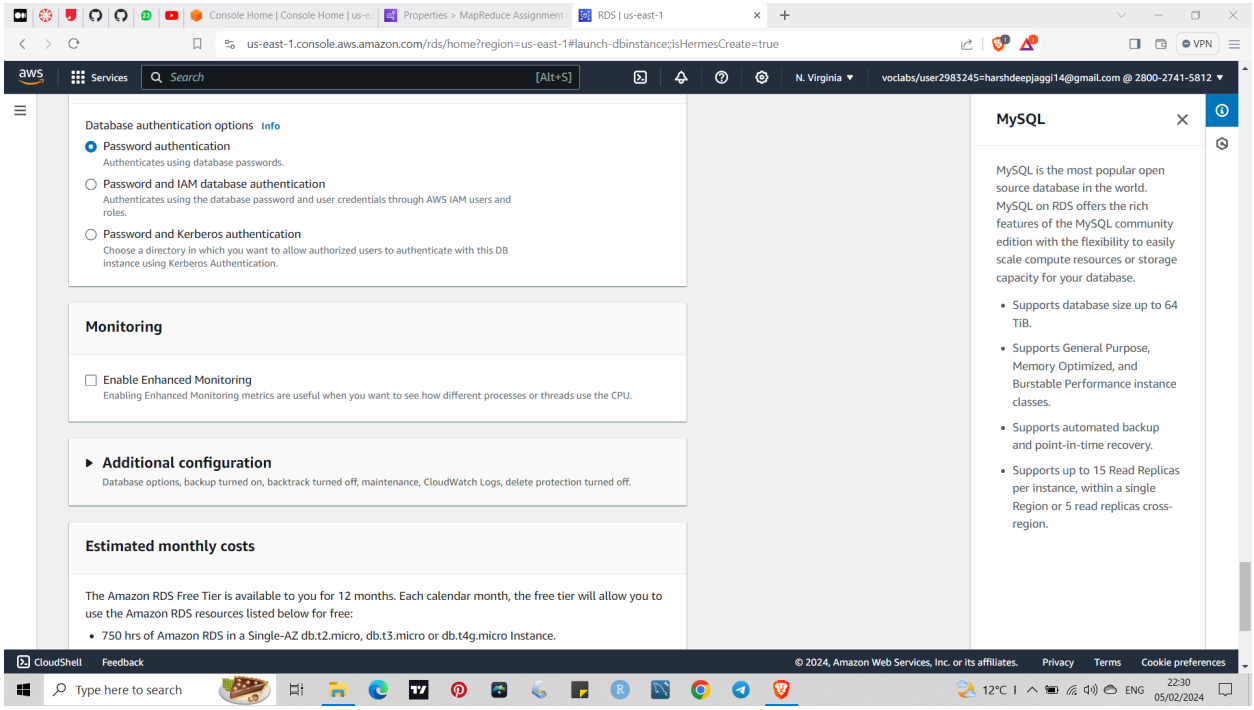


4. Set VPC and DB subnet group values as Default.

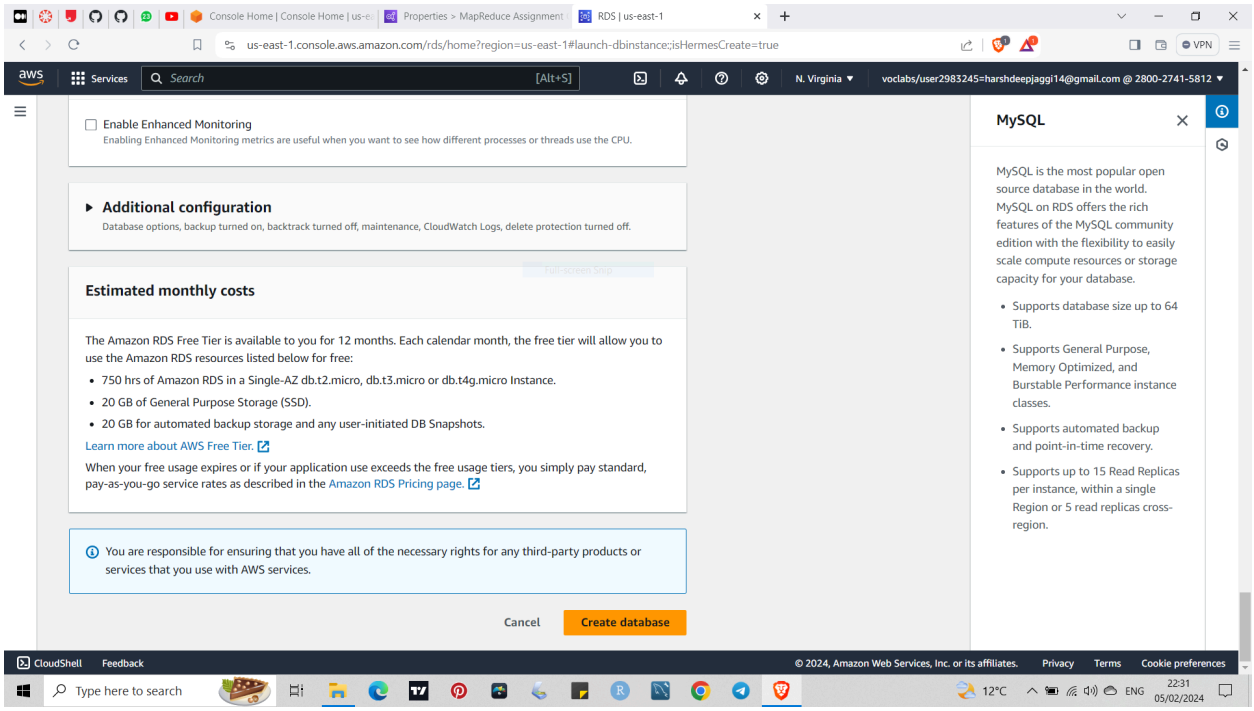


5. Changed Public access to **Yes.**



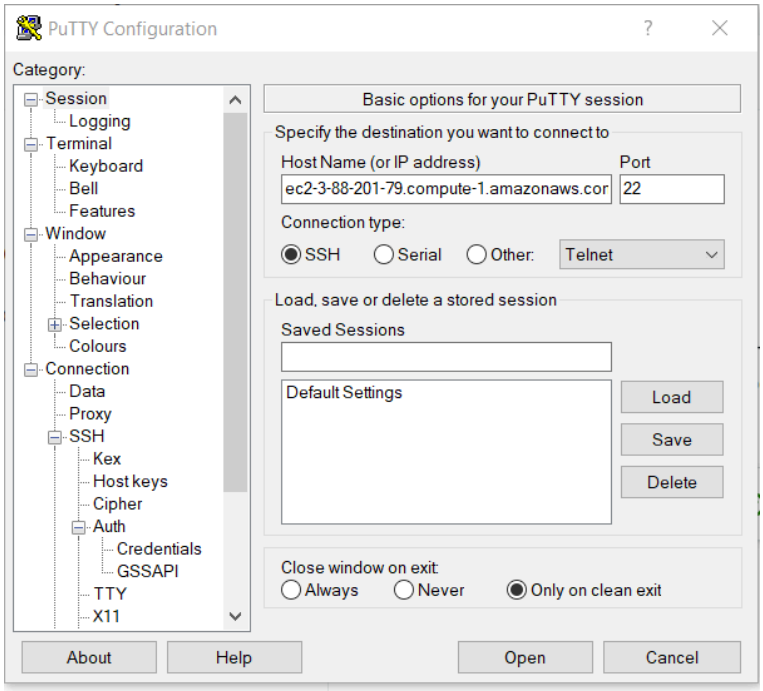


6. After doing everything Click on **Create Database.**

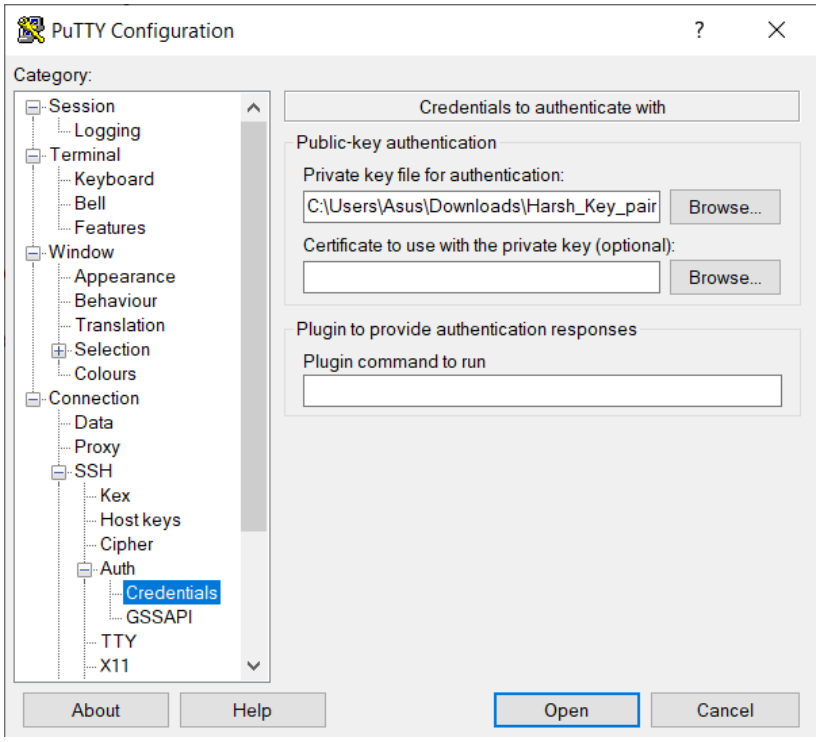


Setting-up Putty Configurations

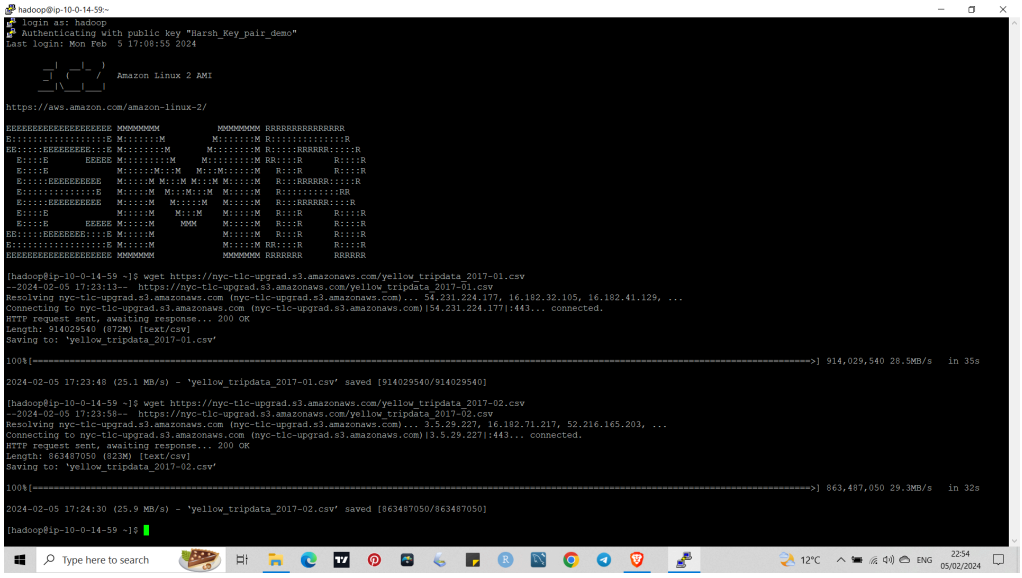
1. Copied the Public DNS from EMR Cluster and then paste it into the **Host Name.**



2. Provided the **Private Key File** from the local machine.

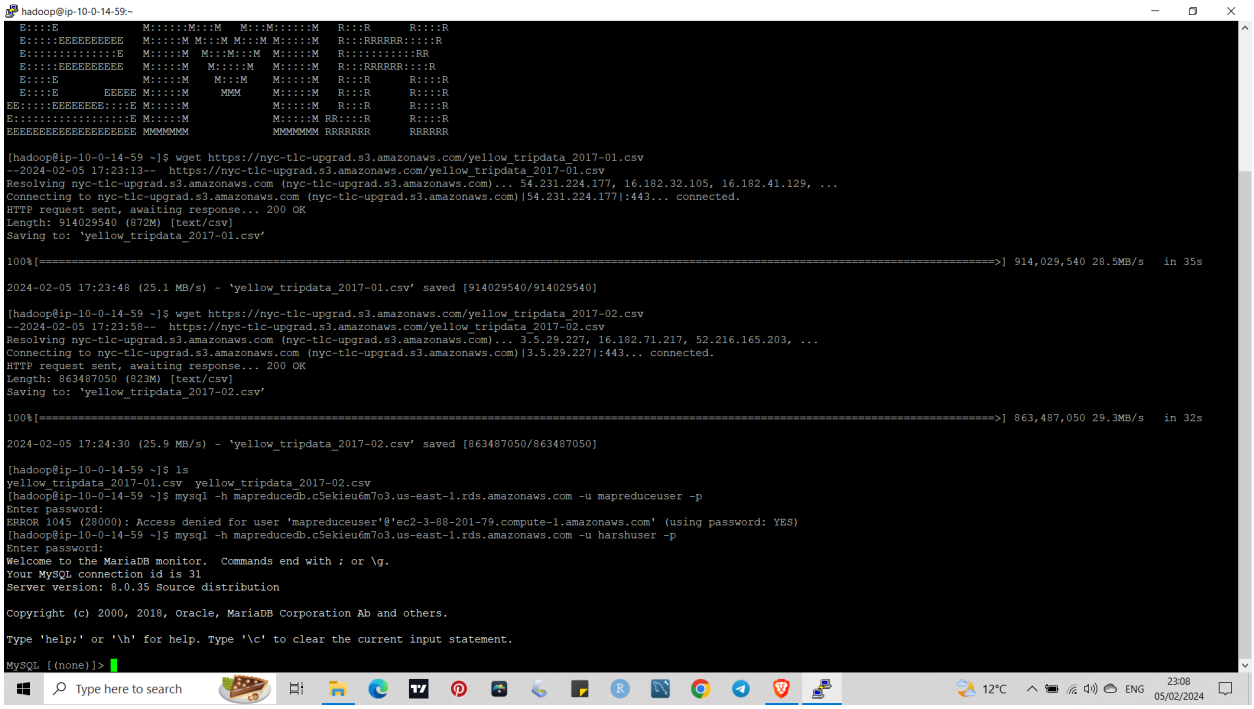


1. **Login to EMR** with putty & Used **wget command to download the csv file** into EMR.
    - wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
    - wget https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv



2. Login to MySQL:

    - Mysql -h mapreducedb.c5ekieu6m7o3.us-east-1.rds.amazonaws.com -u harshuser-p

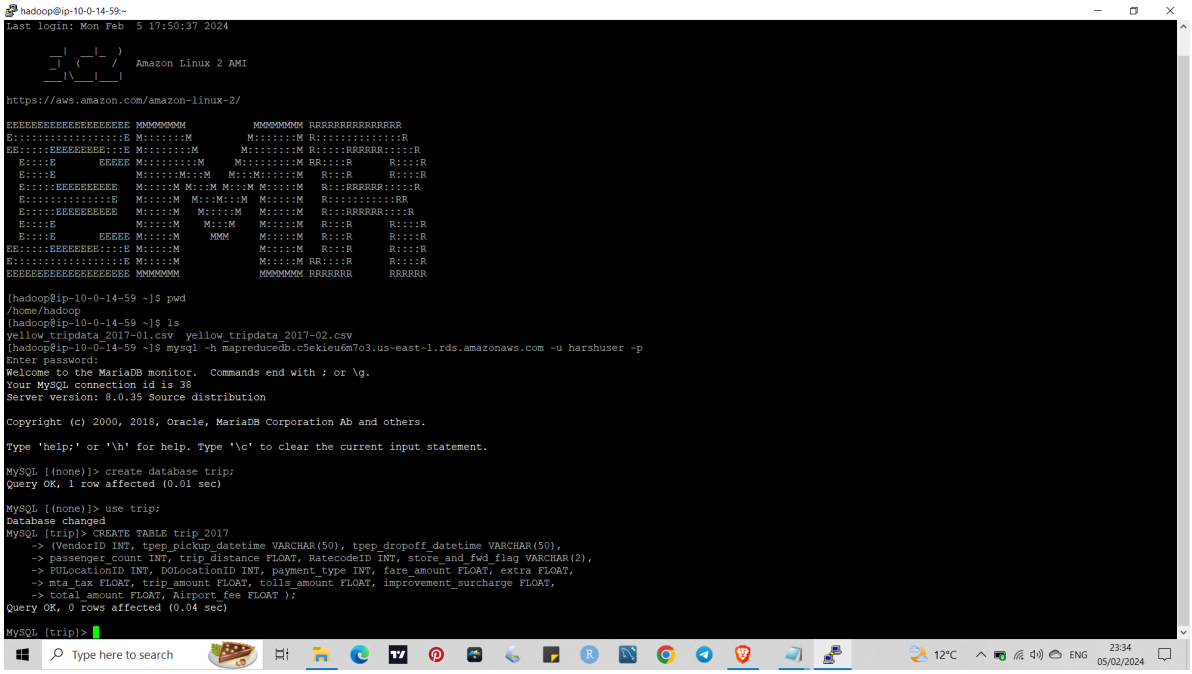


3. Performing actions: Create database, Use database and create table:

    - Create database trip;

    - Use trip;

    - CREATE TABLE trip_2017
      (
      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, trip_amount FLOAT, tolls_amount FLOAT, improvement_surcharge FLOAT, total_amount FLOAT, Airport_fee FLOAT
      );

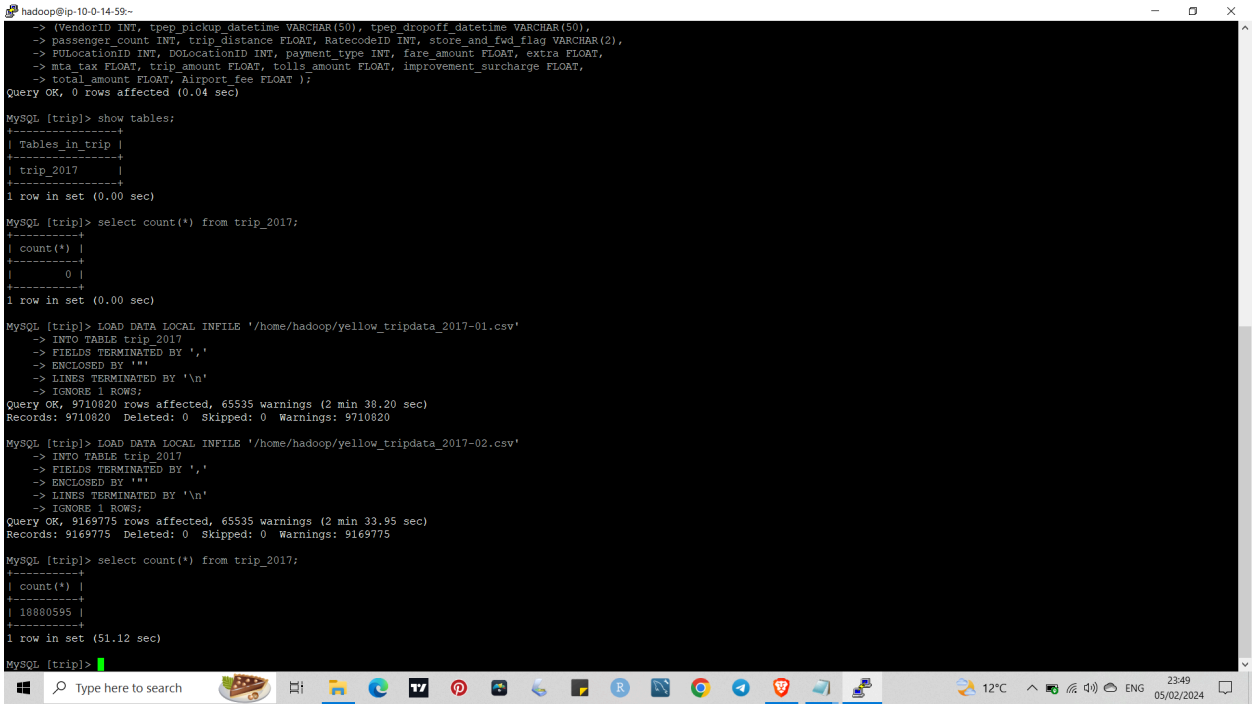


4. Loading data into MySQL Table:

- LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-01.csv'
  INTO TABLE trip_2017
  FIELDS TERMINATED BY ','
  ENCLOSED BY ' " '
  LINES TERMINATED BY '\n'
  IGNORE 1 ROWS;

- LOAD DATA LOCAL INFILE '/home/hadoop/yellow_tripdata_2017-02.csv'
  INTO TABLE trip_2017
  FIELDS TERMINATED BY ','
  ENCLOSED BY ' " '
  LINES TERMINATED BY '\n'
  IGNORE 1 ROWS;

- Select count(*) from trip_2017;



5. **Exit** the mysql so that we can proceed to the ‘Task 2’