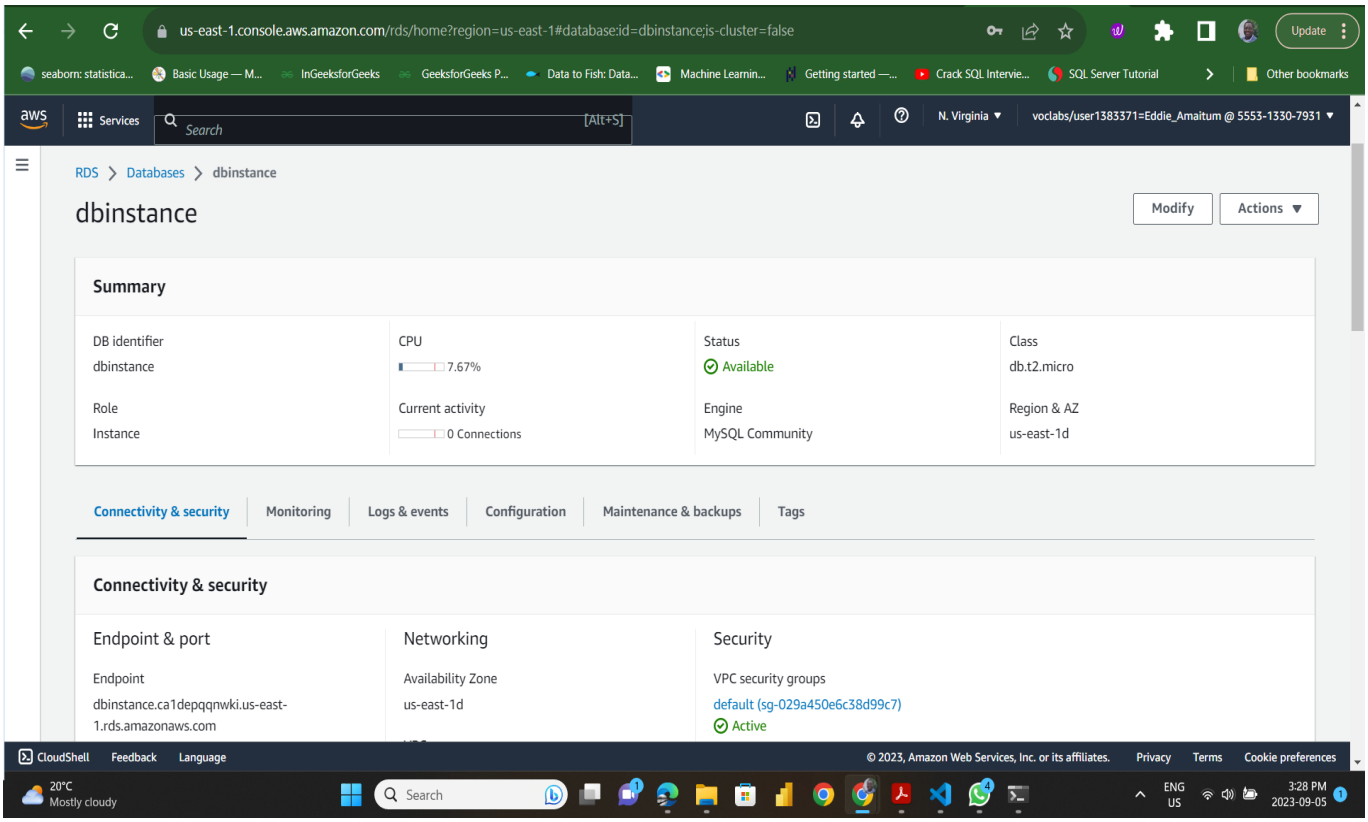
Task 1: Setting up the AWS environment and loading data

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

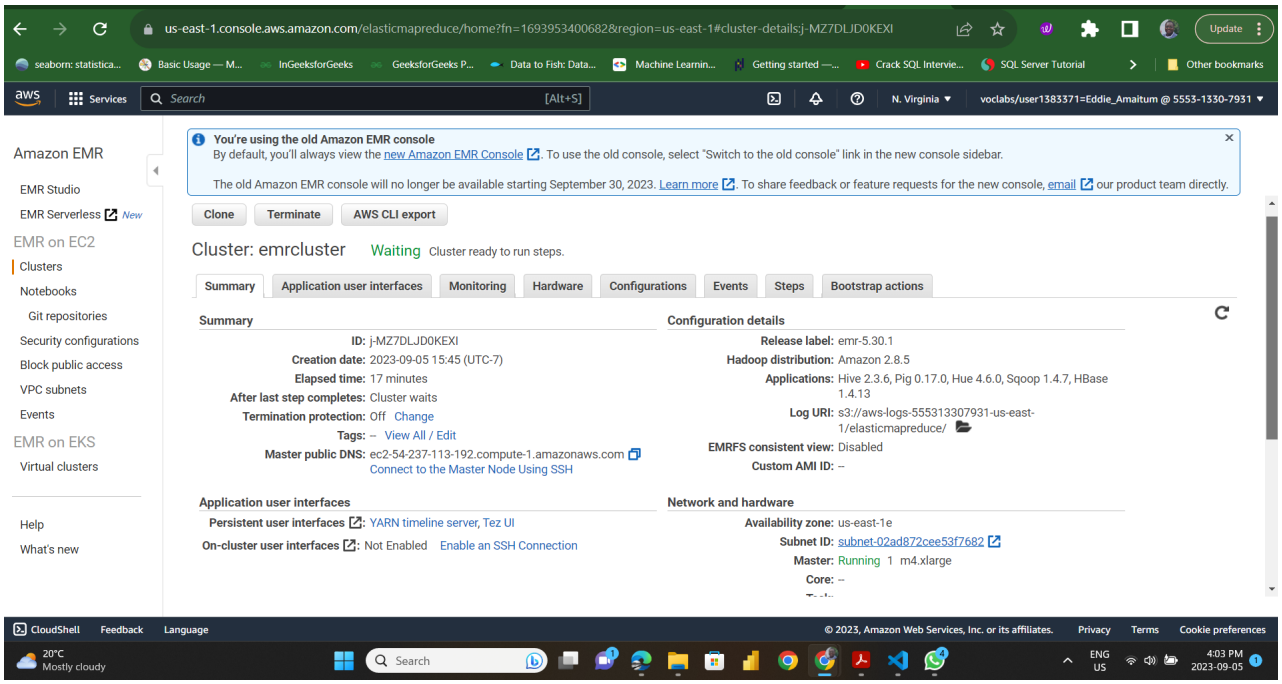
Since the dataset is huge, I uploaded the data from only two files (i.e. yellow_tripdata_2017-01.csv & yellow_tripdata_2017-02.csv) from the dataset.

Note: I created an appropriate schema for the data sets to upload them to RDS.

1. RDS instance creation in AWS

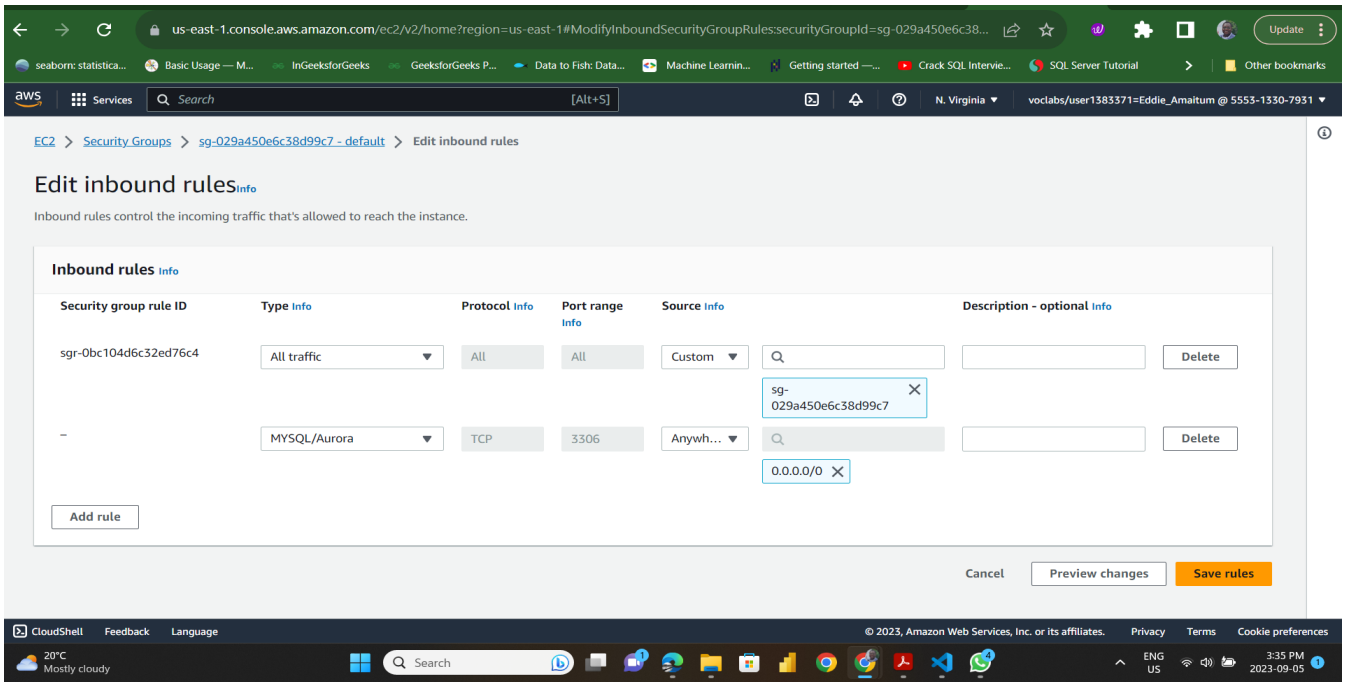


2. EMR creation,some applications selected include:
   Apache Sqoop, Apache Hbase, Hadoop

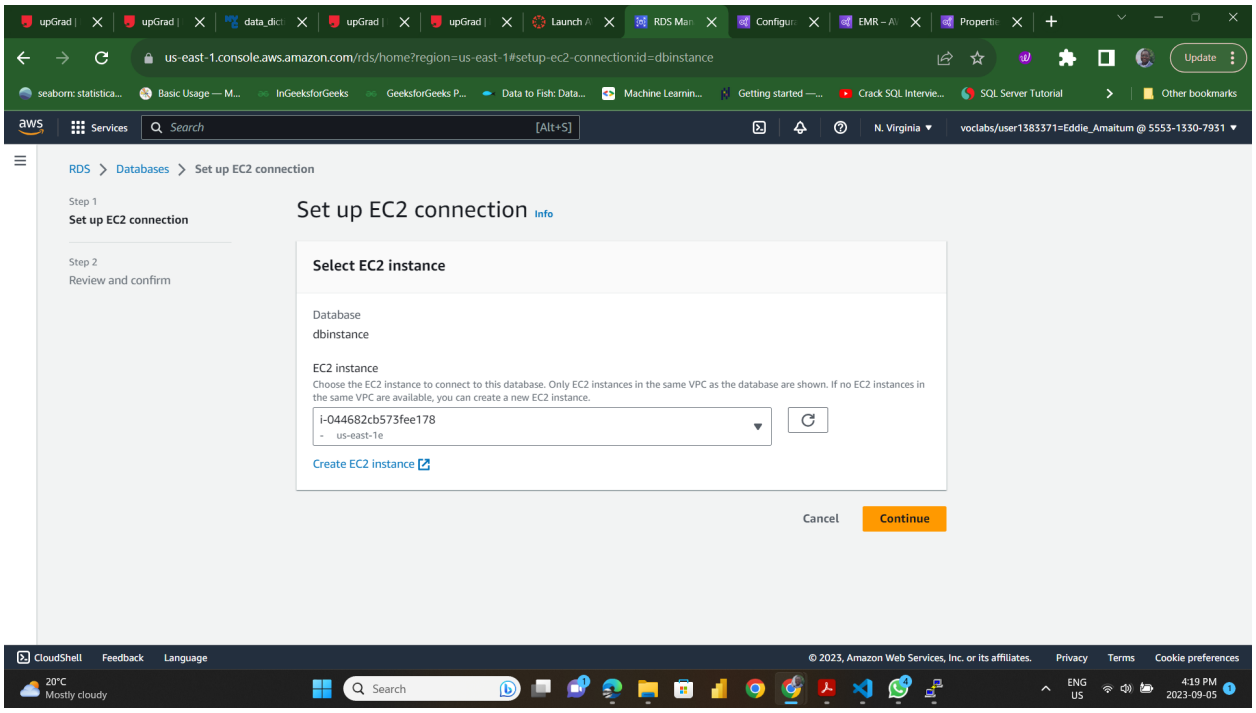


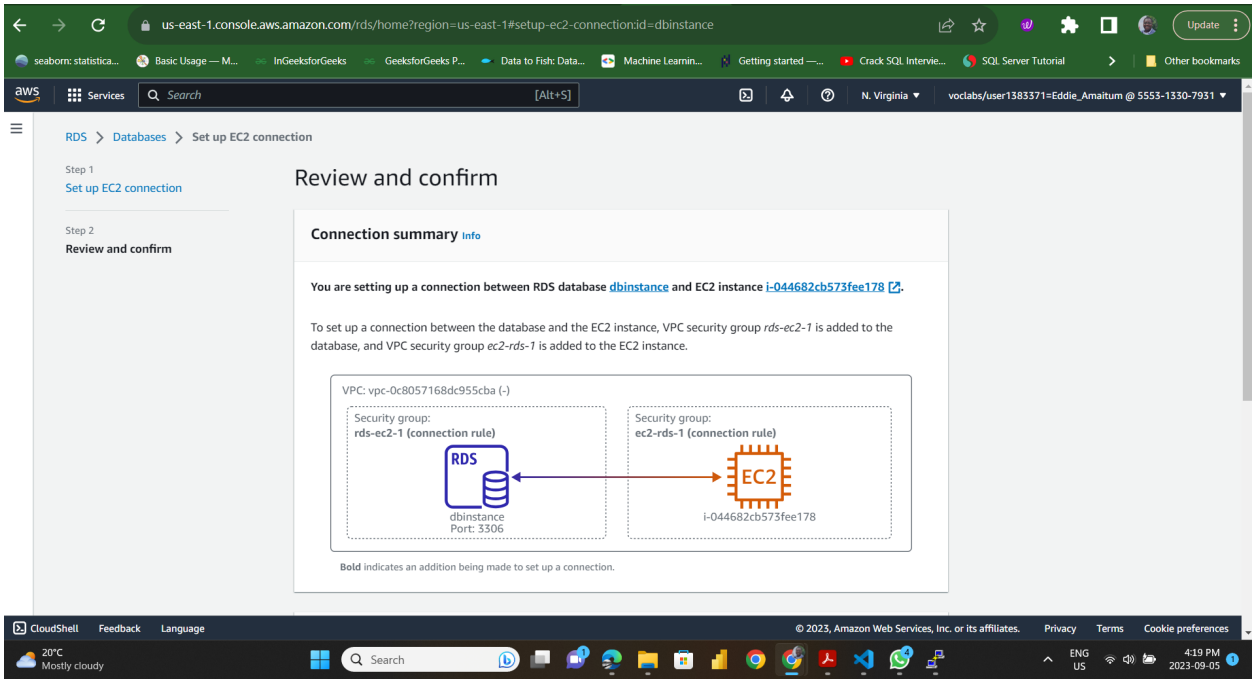
3. Connecting RDS with the EMR instance:

- We configure security group by editing inbound rules



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

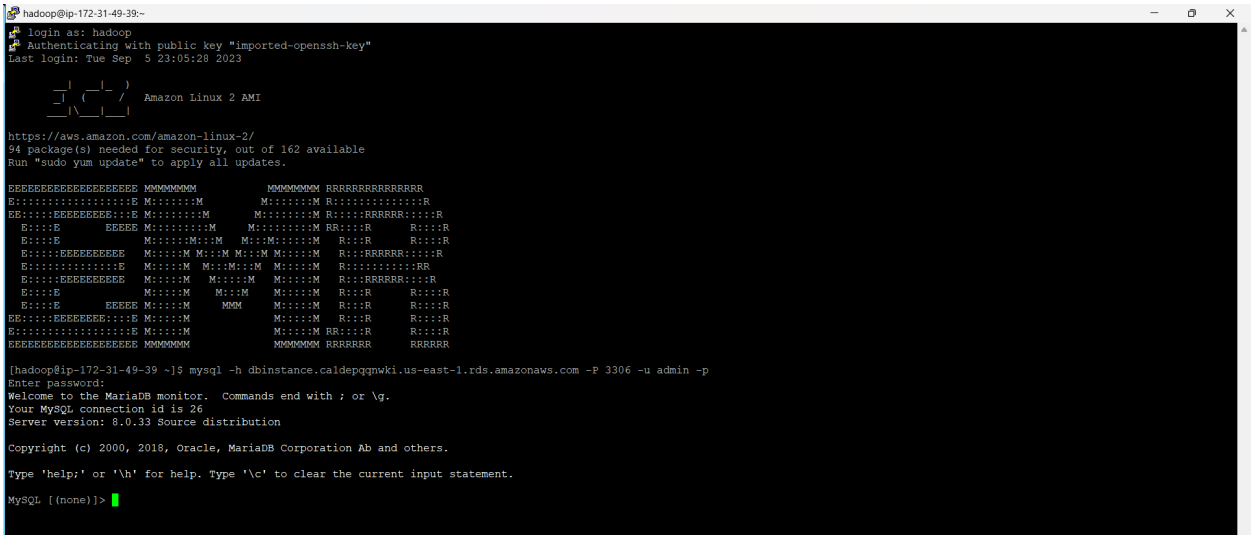




4. We then log into RDS through EMR instance using the command:

‘mysql -h dbinstance.ca1depqqnwki.us-east-1.rds.amazonaws.com -P 3306 -u admin -p’

We enter password upon request to complete login

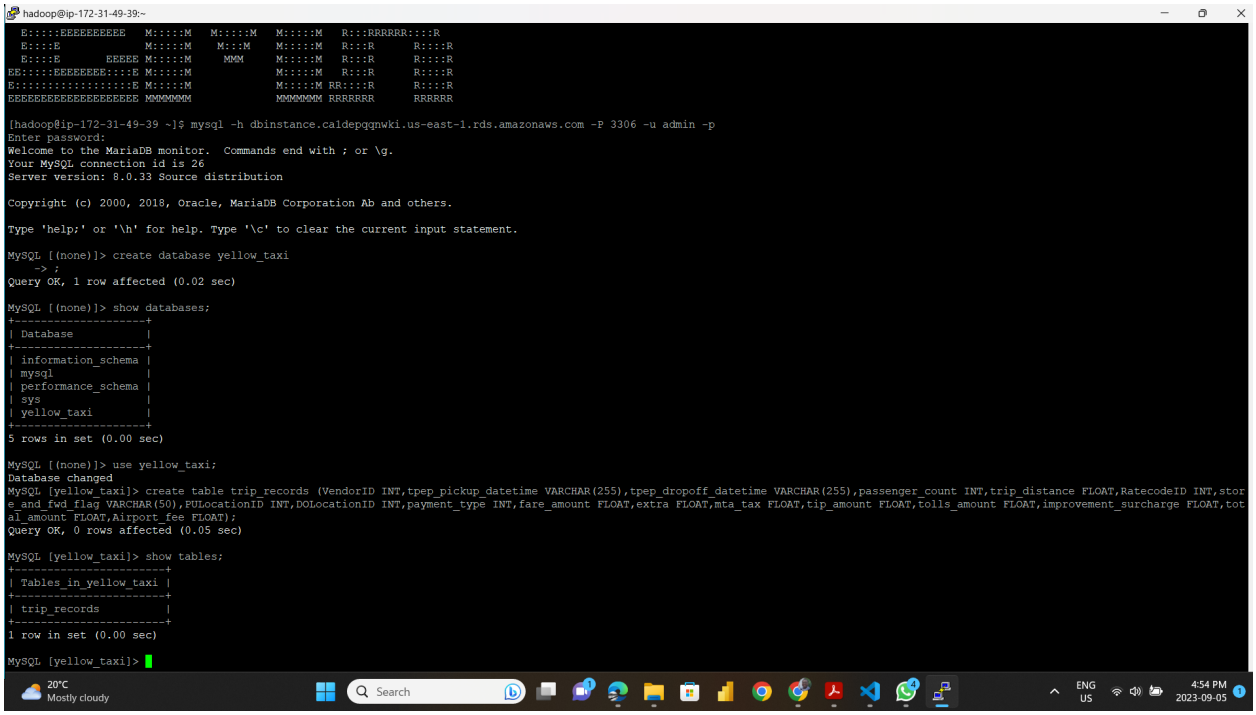


5. We then create database using code below:

> create database yellow_taxi;

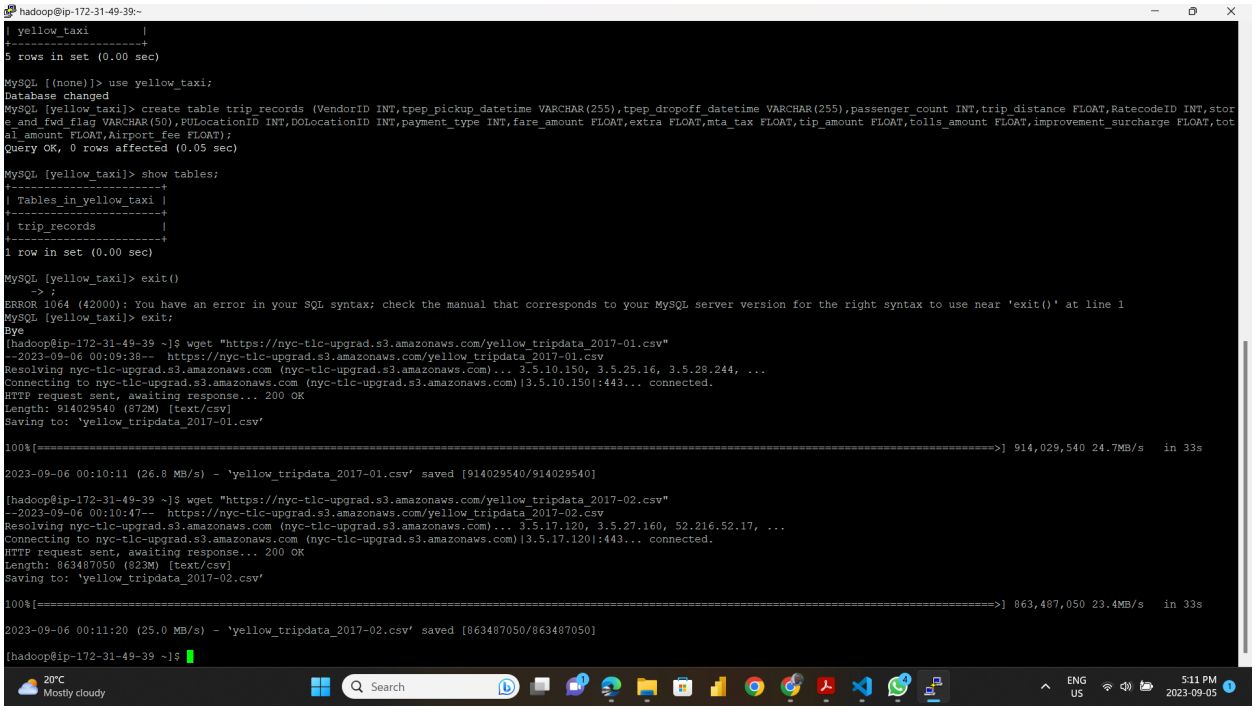
We then create table using code below;

> use yellow_taxi;
> create table trip_records (VendorID INT, tpep_pickup_datetime VARCHAR(255), tpep_dropoff_datetime VARCHAR(255),Passenger_count INT, Trip_distance FLOAT, RatecodeID INT, store_and_fwd_flag 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);



6. To download the required csv files, we use the following commands:

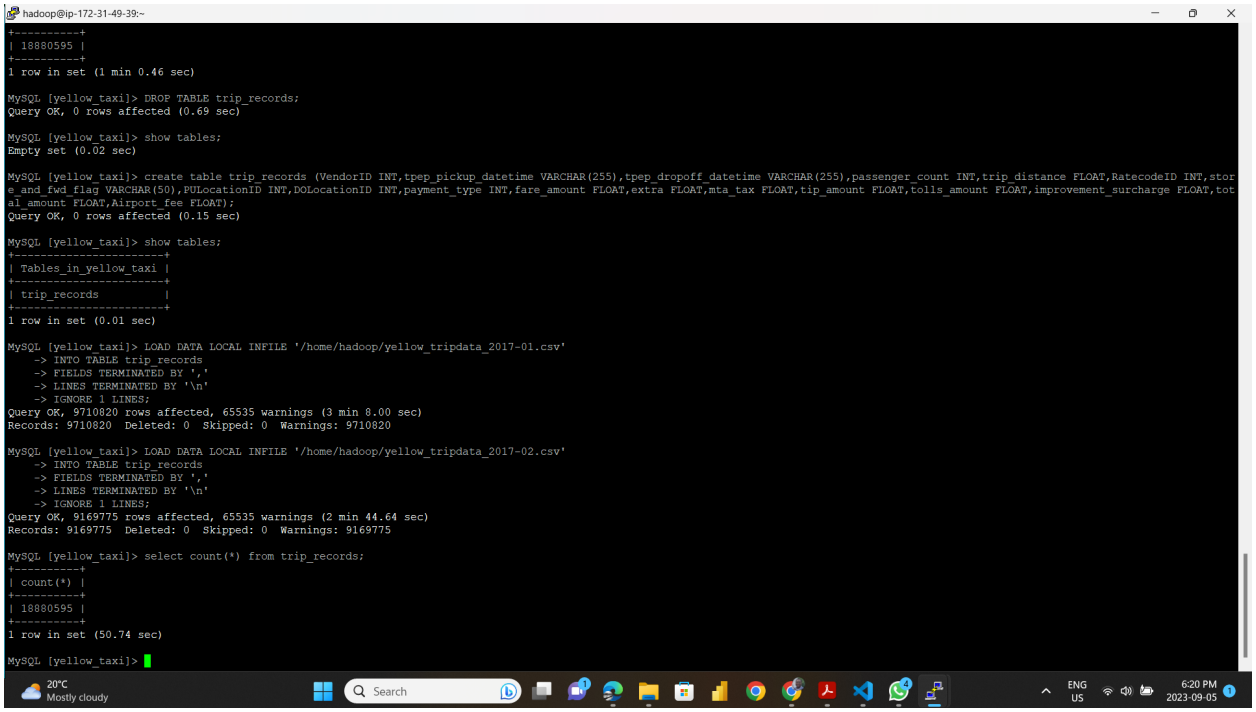
wget "https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv"
wget "https://nyc-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv"



7. To load data in MySQL table, we login and run SQL commands below:

```
>  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;
```



8. Confirming that data is loaded: to do this, we run simple SQL queries:

> select count (*) from trip_records;
> select * from trip_records limit 5;