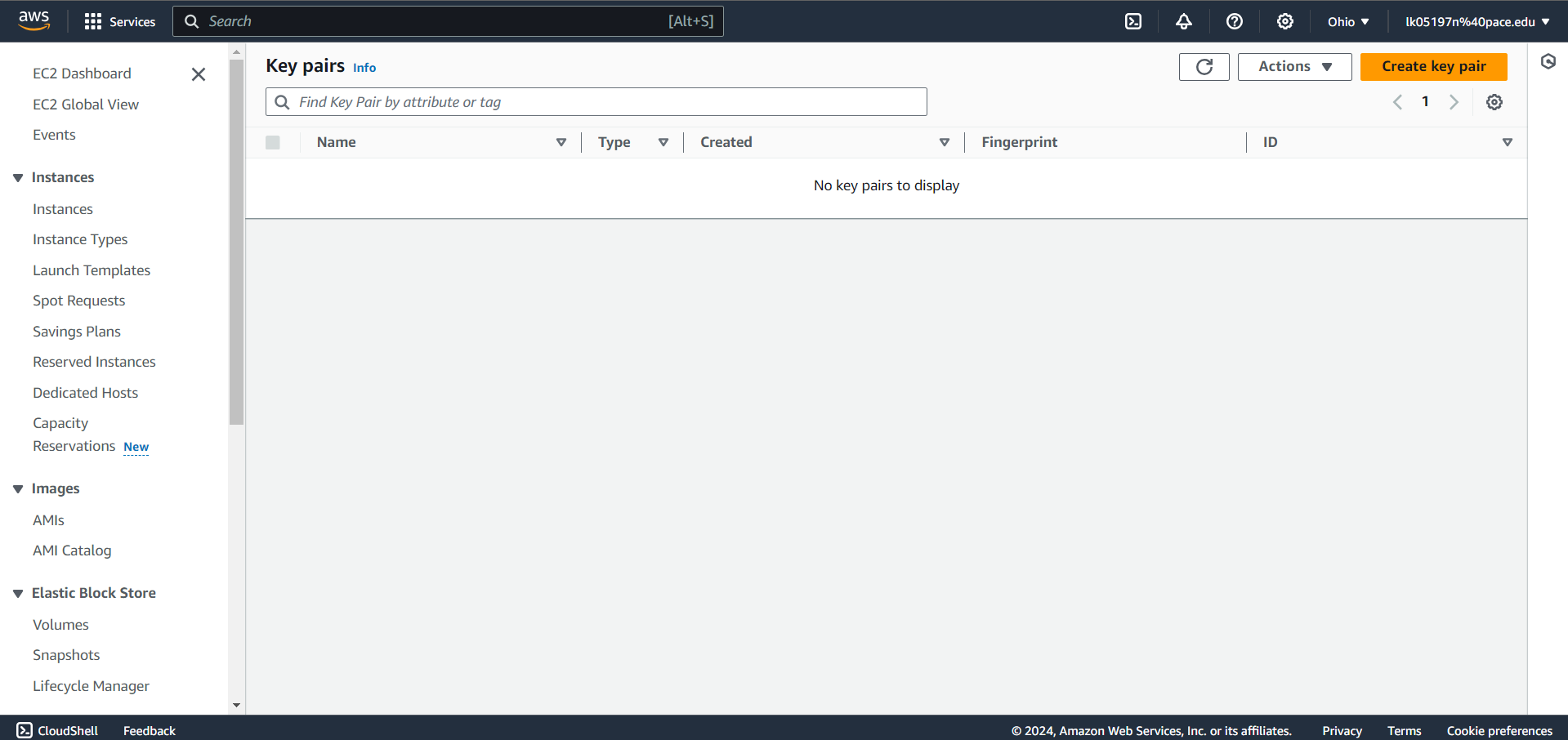
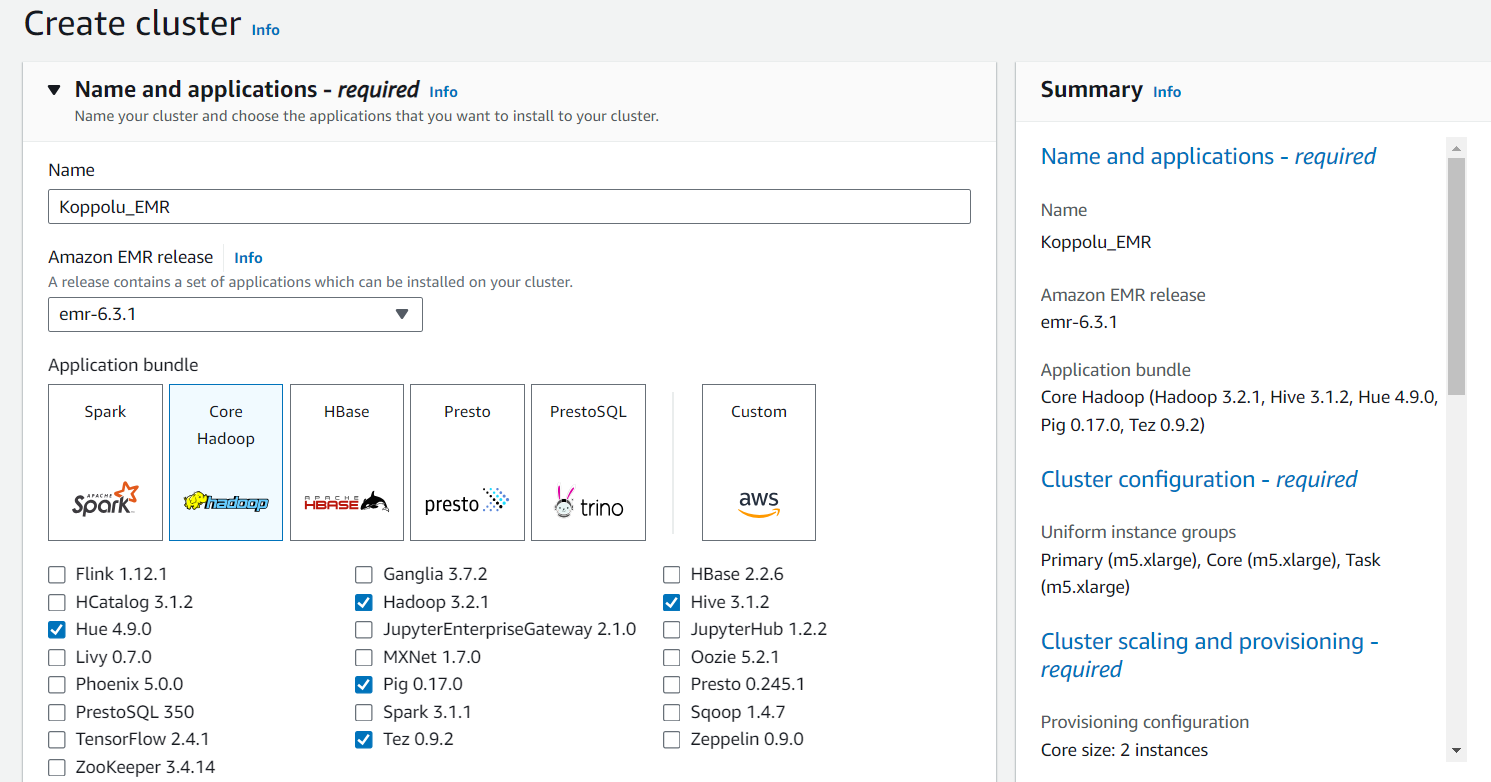
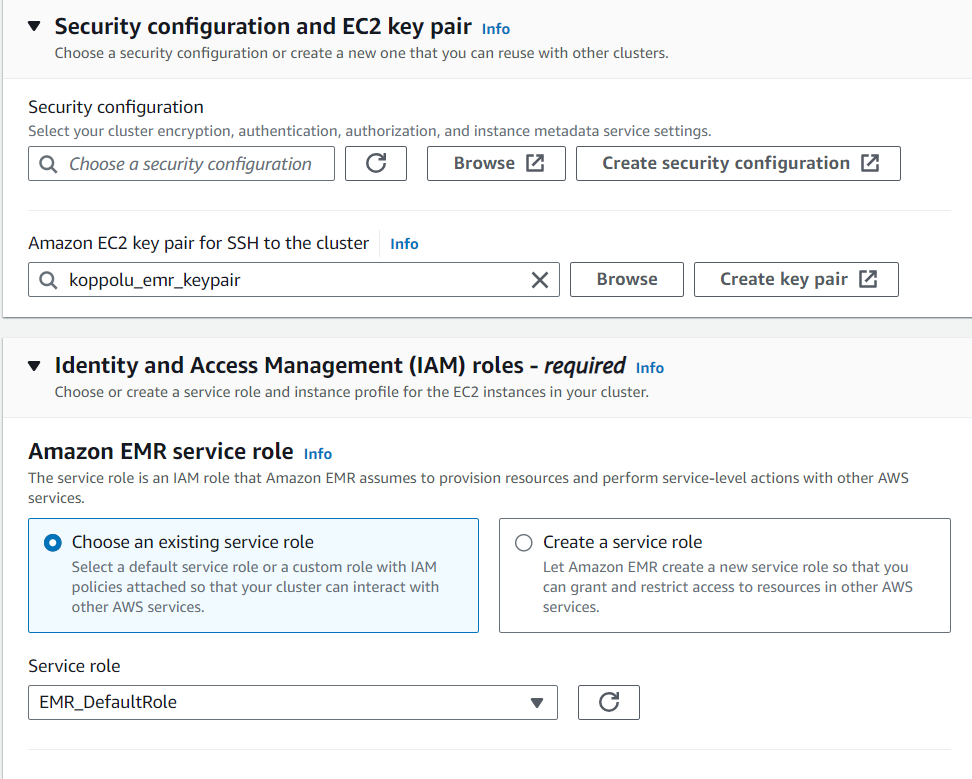
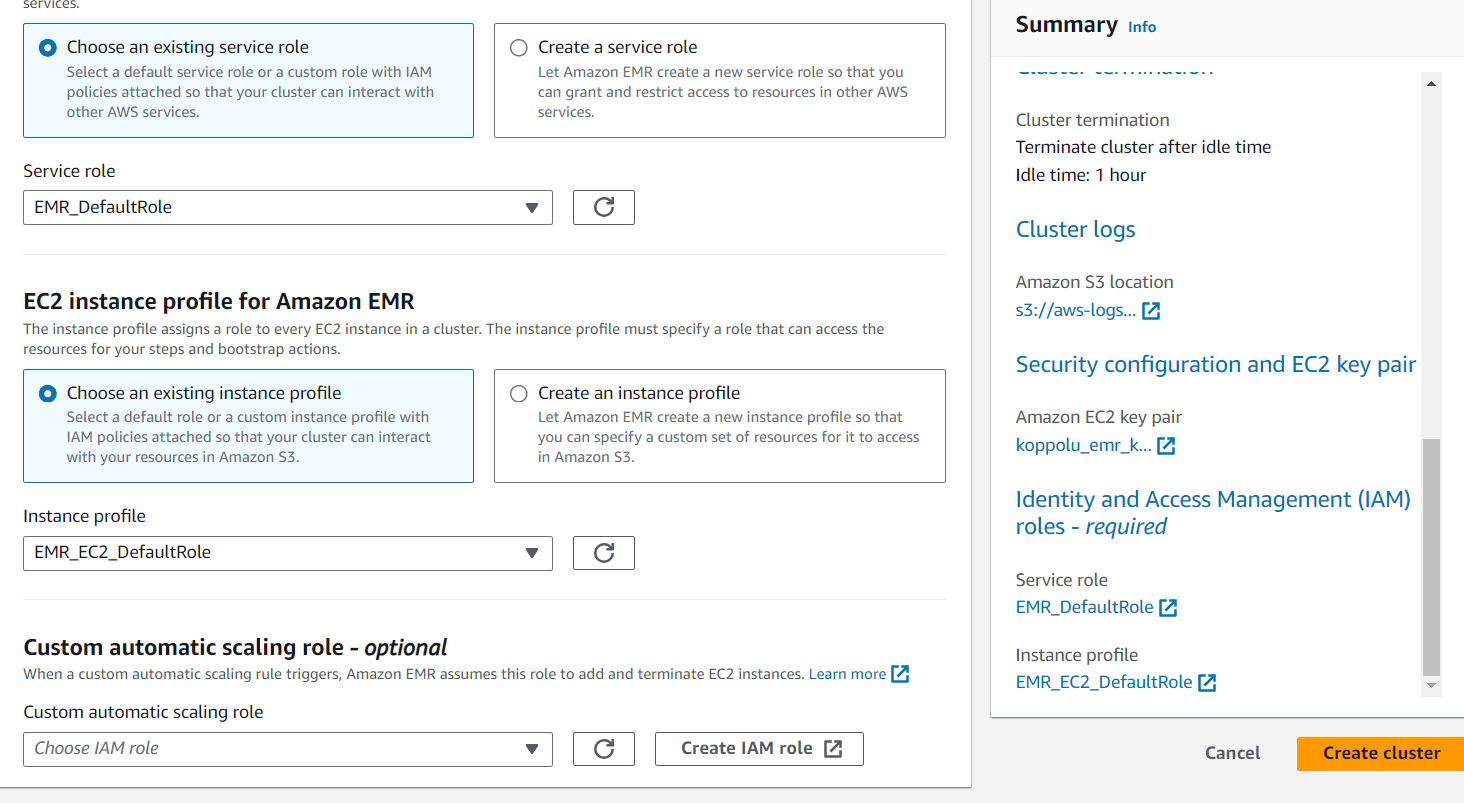
1. Creating an “EC2 Key Pair” first to connect to the EMR cluster via any SSH client (Eg: Putty)
2. Clicked on EC2
3. Clicked on Key Pairs as shown below:



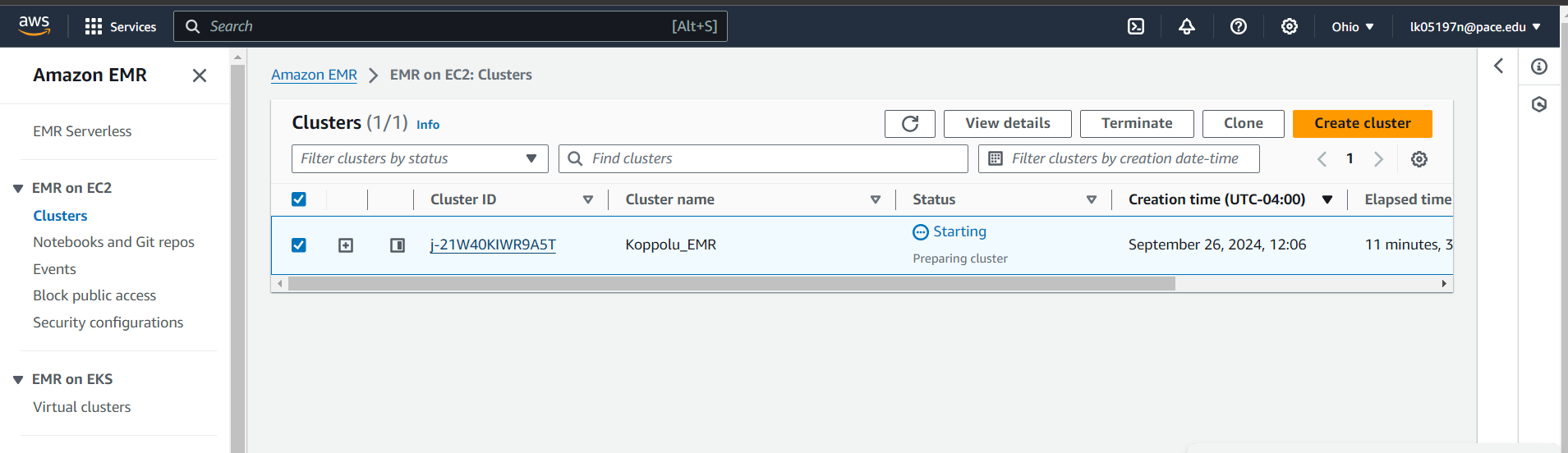
1. Clicked on “Create Key Pair” -> Provide Name and keep the Private key file format as .ppk (Use this keypair to connect via putty)
2. Click on “Create key pair”
3. Now we need to create an EMR cluster by following the below steps
4. Clicked on Create Cluster
   * Provided a Name to the cluster with my last name
   * Select any Release of EMR (Eg: emr-6.3.1)
   * Keep the instance type as m5.xlarge and you can change the “Number of Instances” as per your requirement. We will keep it as 3
   * Within “EC2 key pair” field; select the key pair which you created in Step 1 -> Click on Create cluster



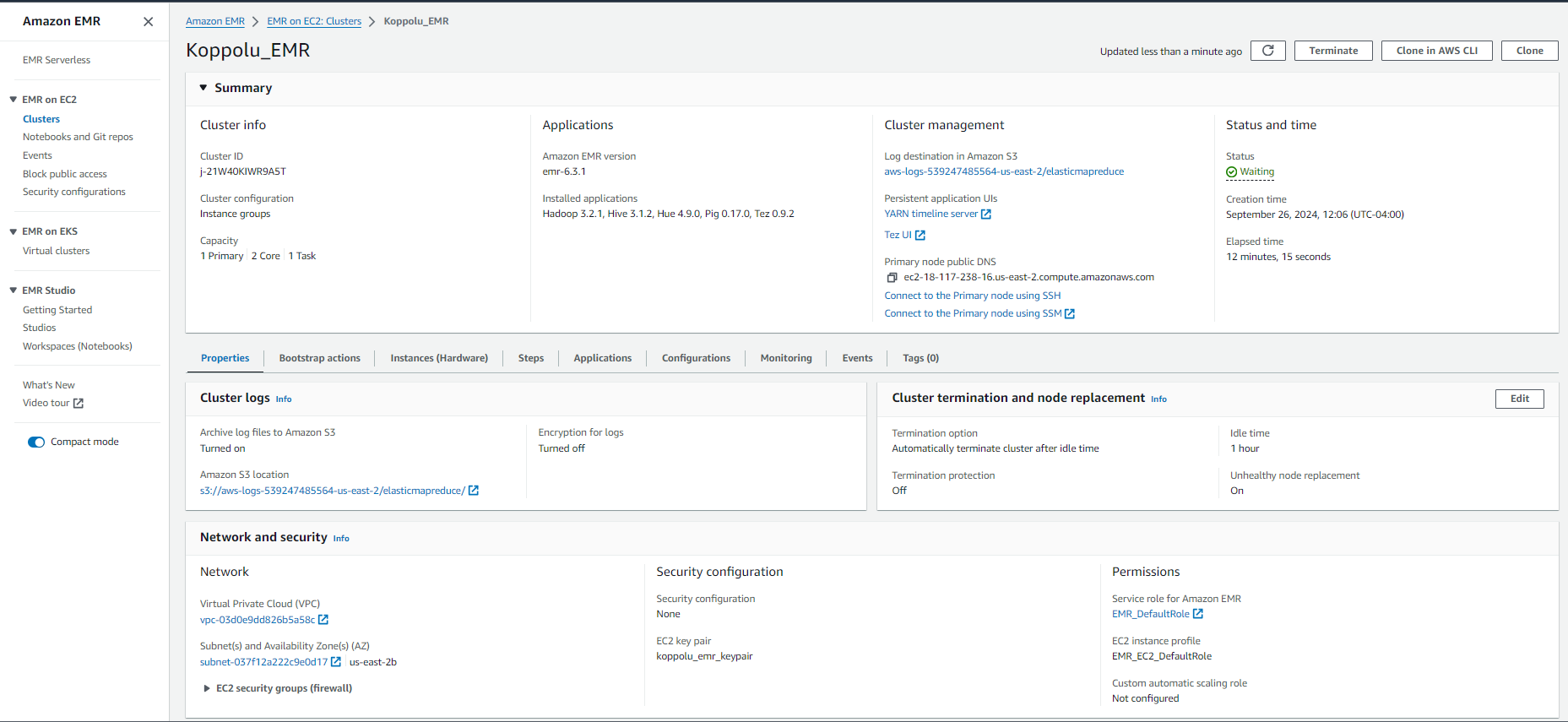




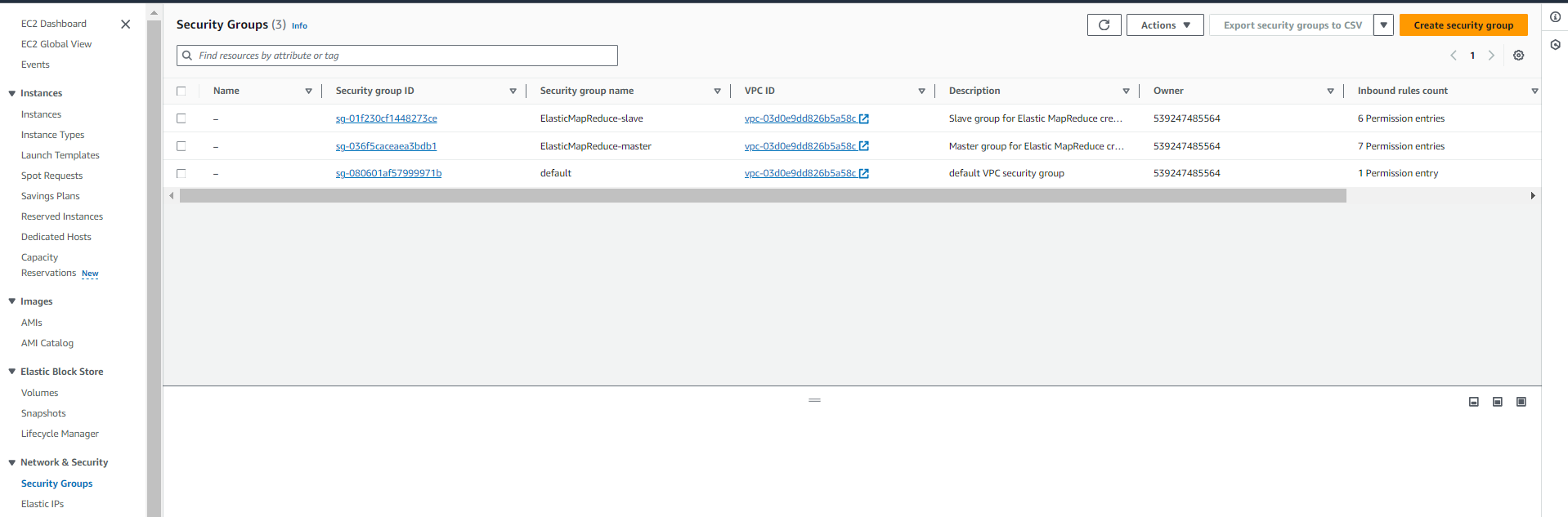
1. Now we need to add an inbound rule to the security group of Master Node which would allow all inbound connections on Port 22.
2. Go to EMR and click on “Clusters”
3. Click on the small arrow of the cluster which you have created



1. Click on “View cluster details” and scroll down and check the details within “Security and Access”



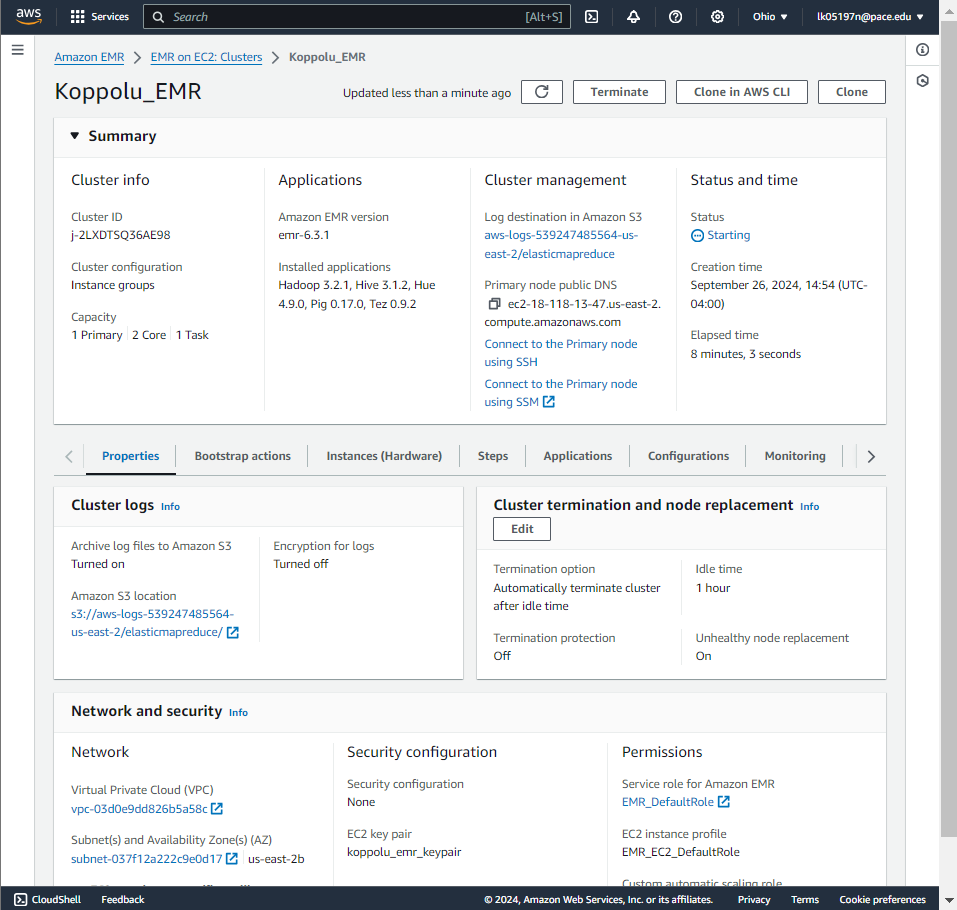
1. Click on the security groups for Master
2. Click on the “Security group id” for ElasticMapReduce-master



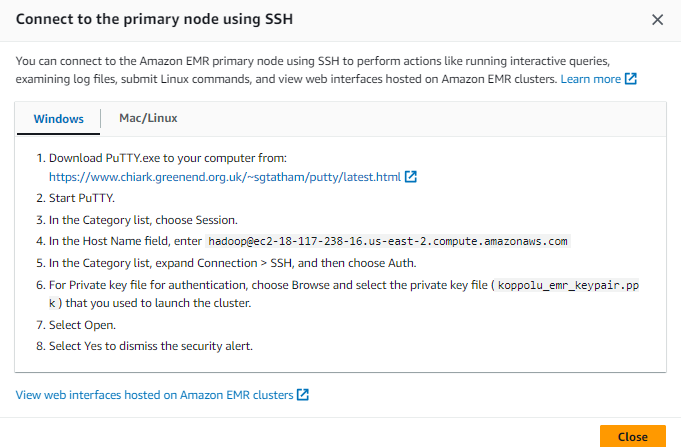
1. Scroll Down and click on “Edit inbound rules” -> Scroll down and click on “Add rule” -> Select Type as “SSH” and Source as “Anywhere-IPv4” as shown below:



1. Click on Save rules
2. Now connect to the EMR master node via putty to perform Hive table creation and ingestion of data
3. Search “EMR” -> Click on Clusters -> Click on the small arrow for the cluster which you created -> Click on “View Cluster Details”



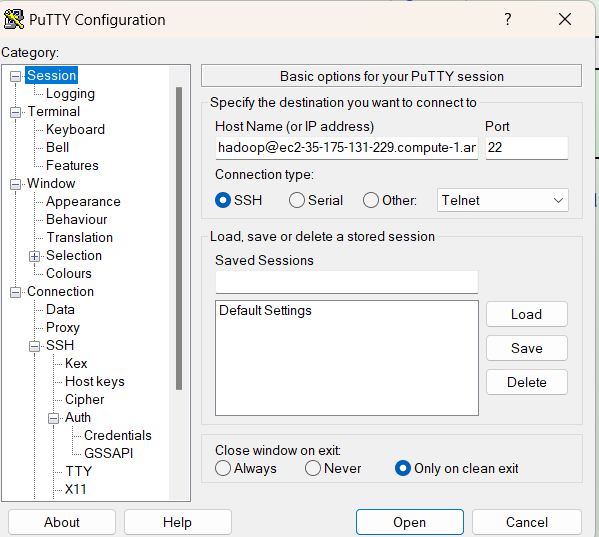
1. Click on “Connect to Master Node using SSH”

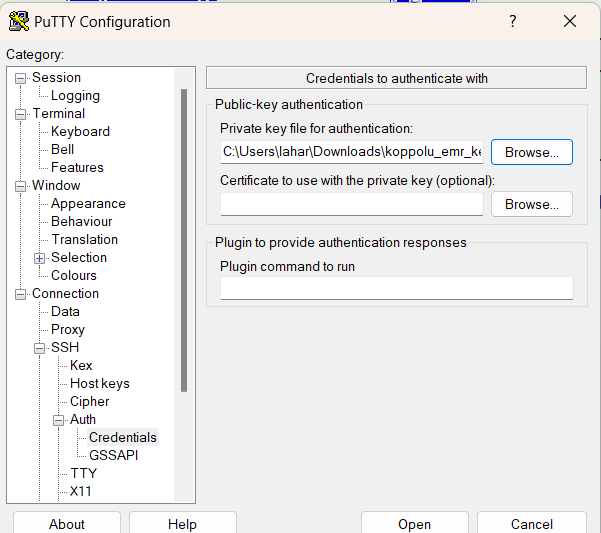


1. Click on Putty on your local machine -> Enter the Host Name [As per your HostNames]-> Make sure you select the keypair which you created in Step 1 to do ssh

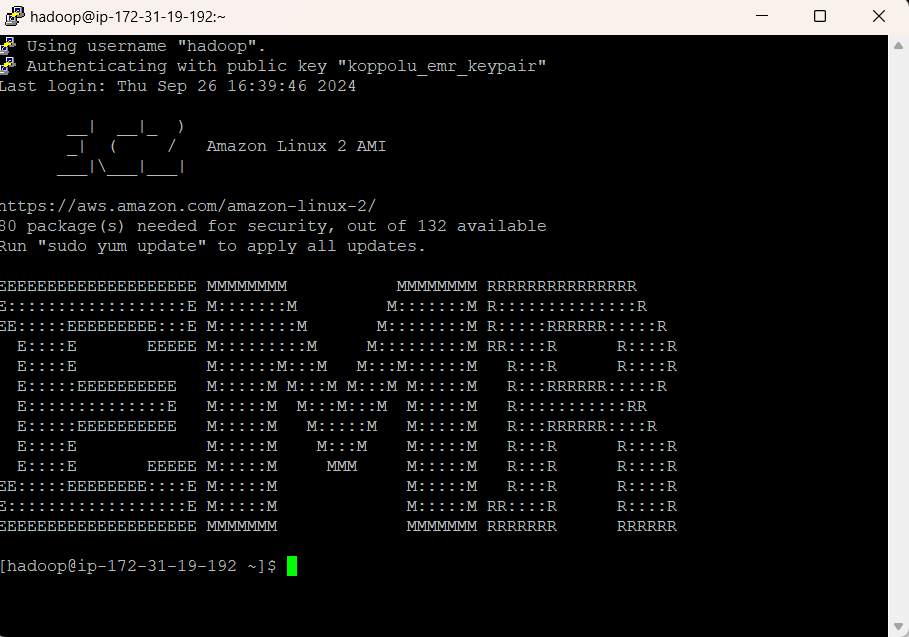
Graphical user interface, application

Description automatically generated





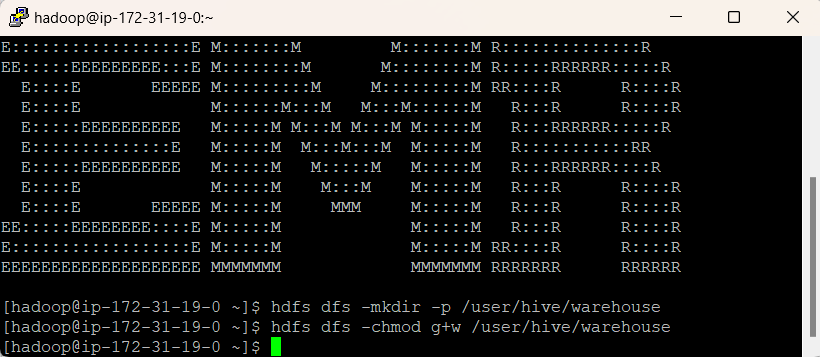
1. Click on Open -> Click on “Yes” and you are now in your Master Node



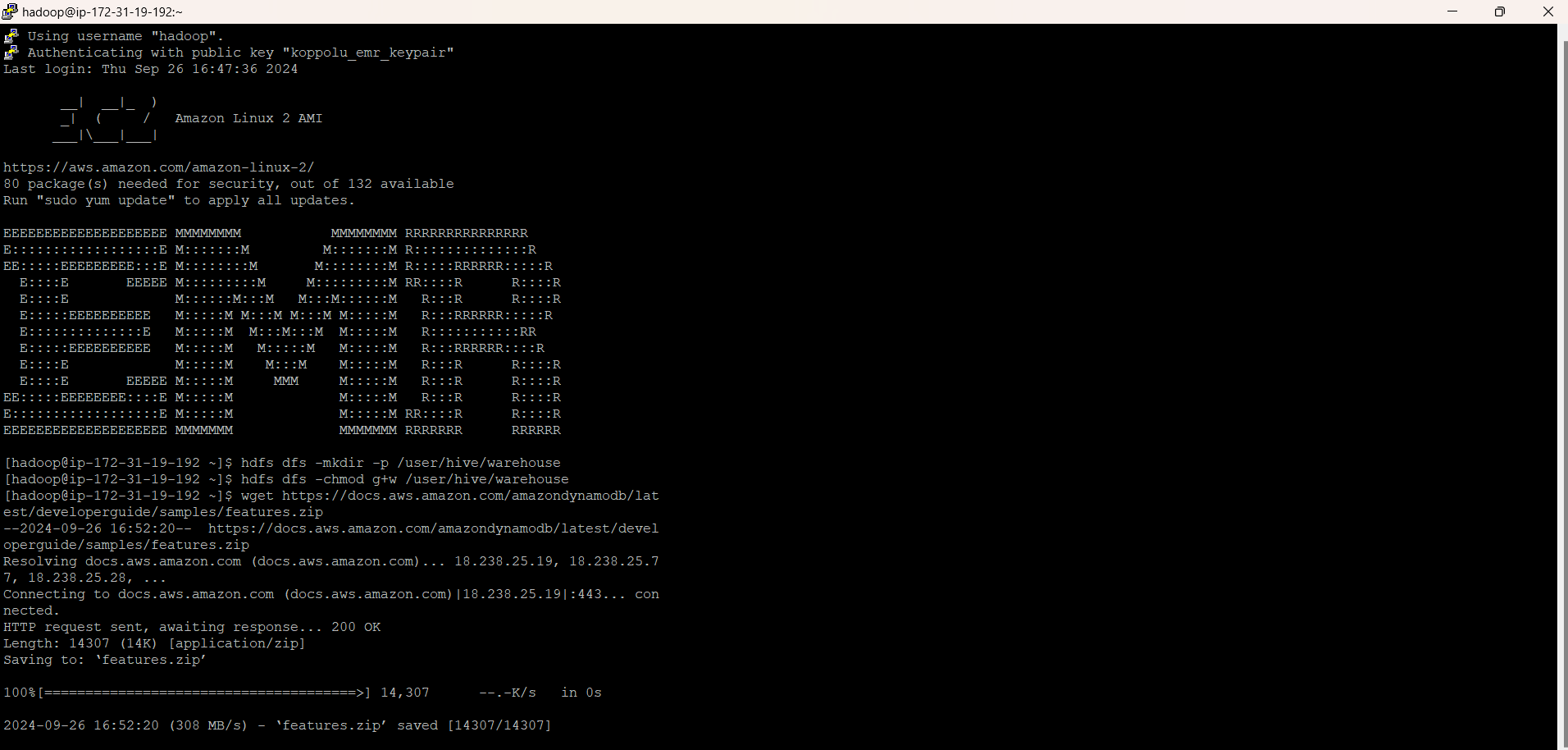
1. Create an hdfs directory and grant permission to it
2. Type the below commands on your master node:

**hdfs dfs -mkdir -p /user/hive/warehouse**

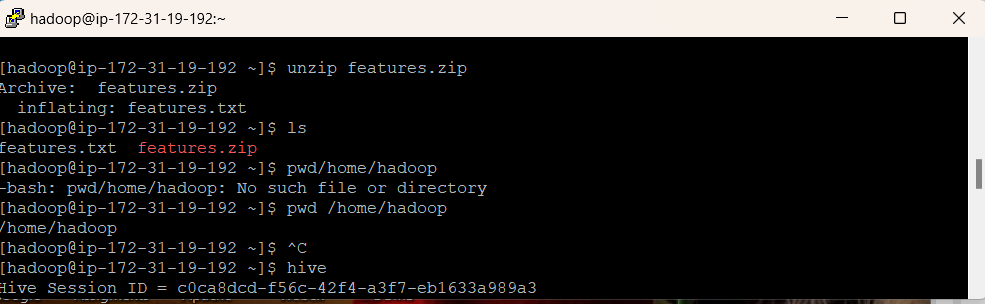
1. **hdfs dfs -chmod g+w /user/hive/warehouse**

****

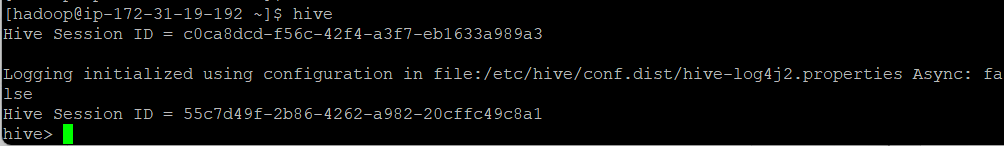
1. Extract the features.zip file and unzip it
2. Run the below command:
3. **wget https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/samples/features.zip**



1. Run command: **unzip features.zip**

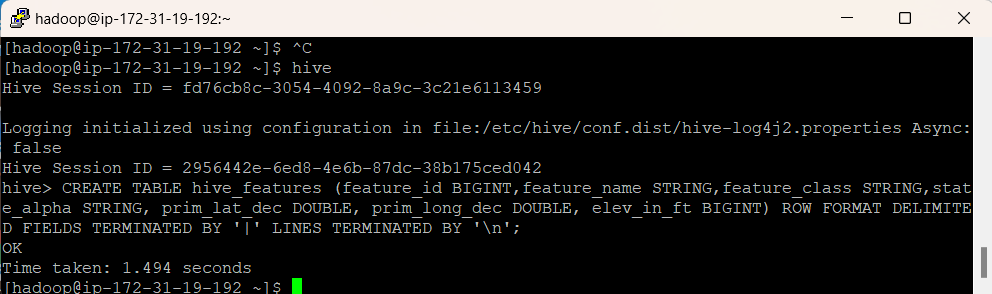


1. Create a Hive Table and ingest data into it:
2. Type “Hive” and hit enter



Create a table named “hive\_features” using the below DDL:

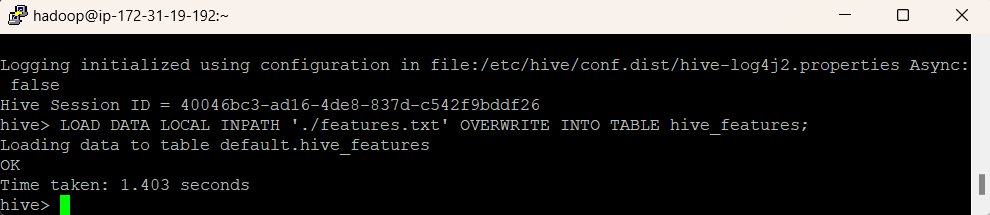
CREATE TABLE hive\_features (feature\_id BIGINT,feature\_name STRING ,feature\_class STRING,state\_alpha STRING, prim\_lat\_dec DOUBLE, prim\_long\_dec DOUBLE ,elev\_in\_ft BIGINT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' LINES TERMINATED BY '\n';



1. Load the “features.txt” data within this table

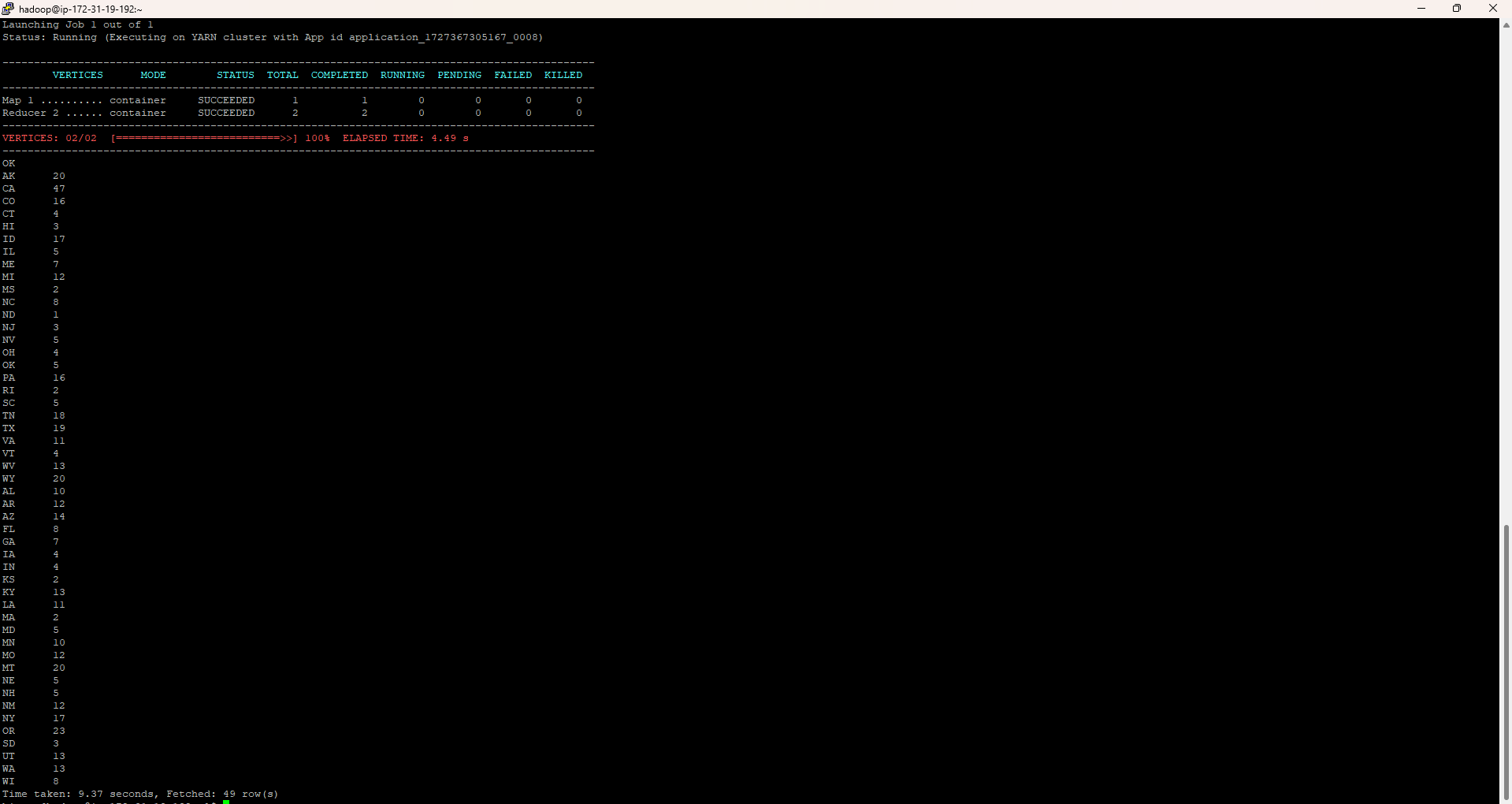
LOAD DATA LOCAL INPATH './features.txt' OVERWRITE INTO

TABLE hive\_features;

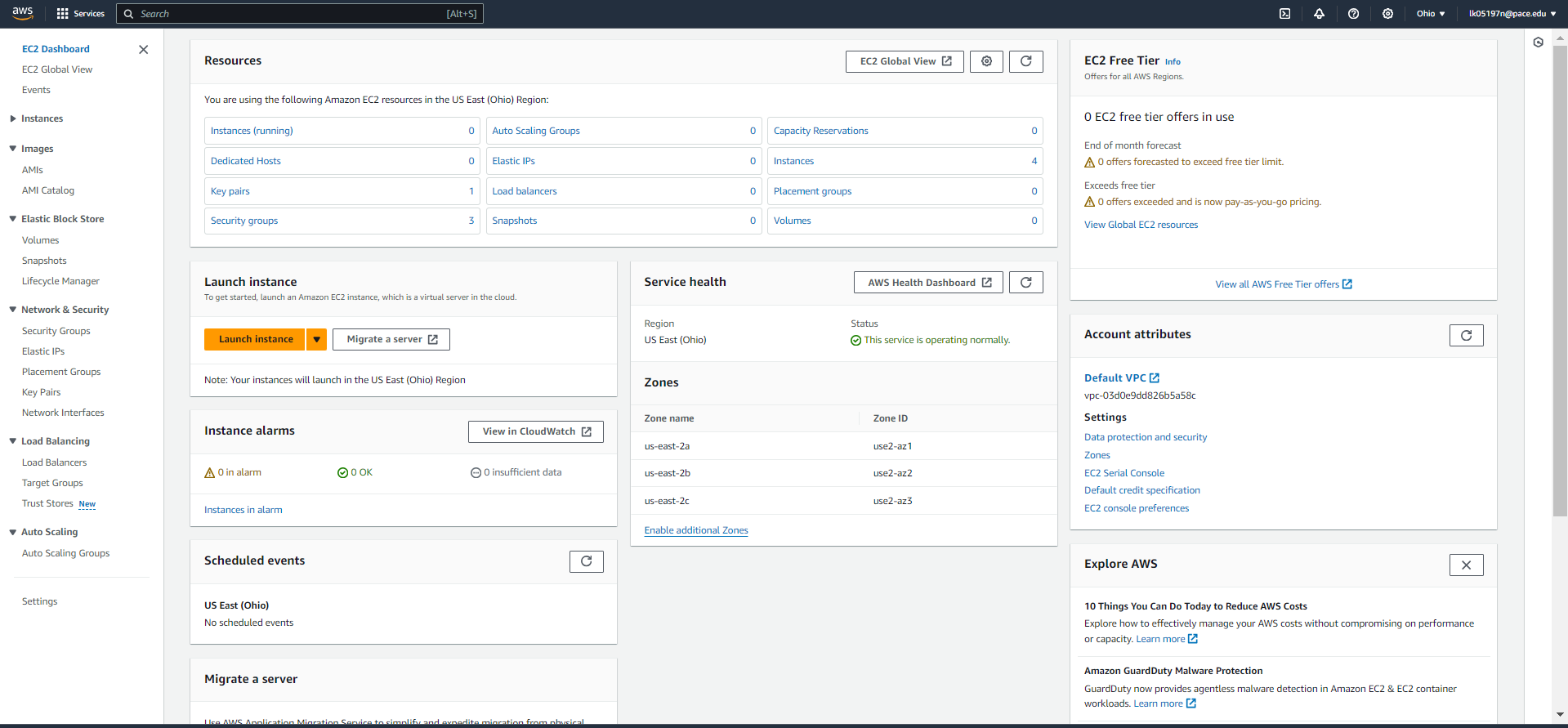


1. Run the below query to display a list of states and number of geographic features in each of these states:

SELECT state\_alpha, COUNT(\*) FROM hive\_features GROUP BY state\_alpha;



**TERMINATED**

****