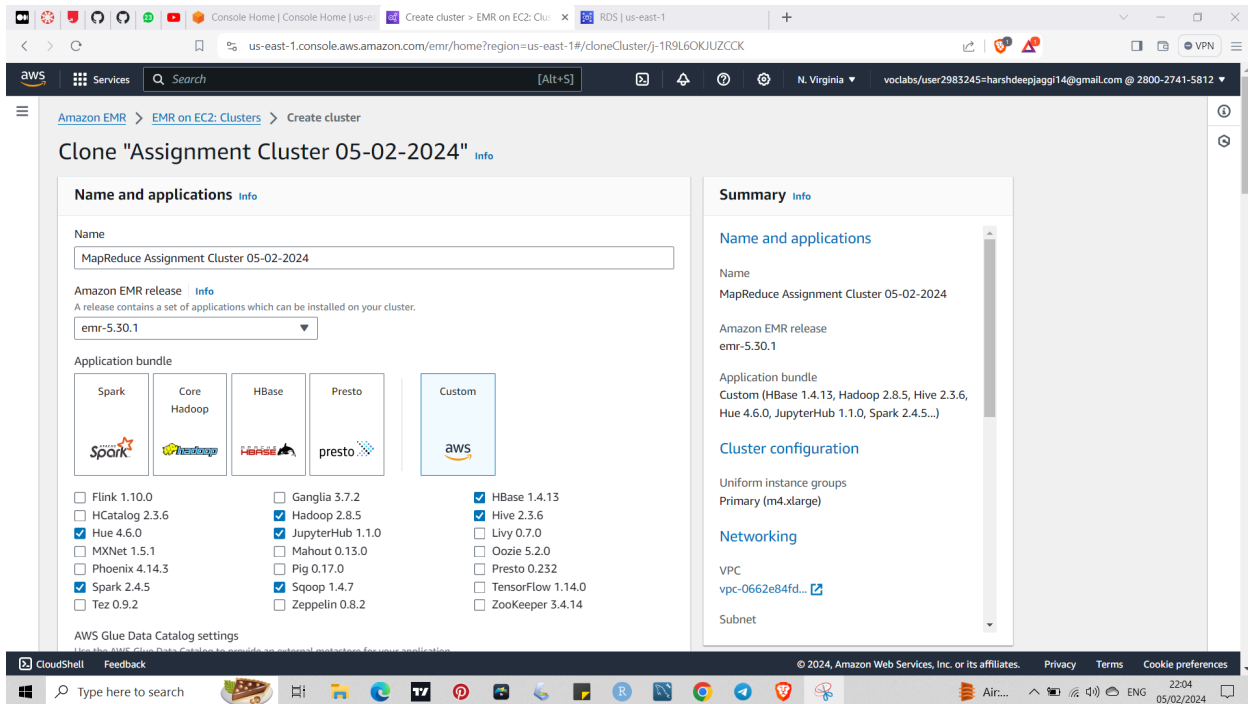
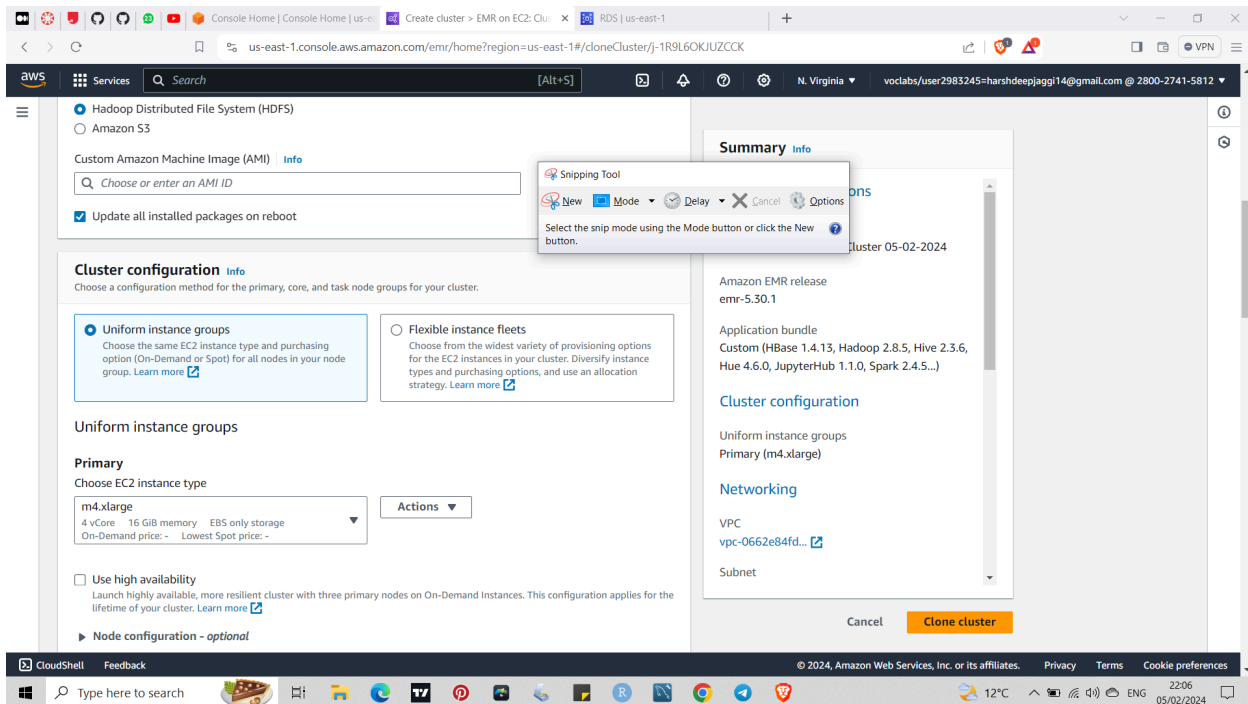


Setting up EMR Cluster

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.

EBS root volume
EBS root volume applies to the operating systems and applications that you install on the cluster.
Size (GiB)
10
10 - 100 GiB per volume General Purpose SSD (gp2)

Networking [Info](#)
Virtual private cloud (VPC) [Info](#)
vpc-0662e84fd80767c7d [Browse](#) [Create VPC](#)
Subnet [Info](#)
subnet-001df27464cf3a0ad [Browse](#) [Create subnet](#)
EC2 security groups (firewall)
Steps - optional (0) [Info](#) [Remove](#) [Edit](#) [Add](#)
Use commands and scripts to tell your cluster where to find and how to process your data. Steps run consecutively unless you enable the Concurrency option.
Cluster termination [Info](#)
☒ Manually terminate cluster
☐ Automatically terminate cluster after last step ends
☐ Automatically terminate cluster after idle time (Recommended)
☐ Use termination protection

Summary [Info](#)
Name and applications
Name
MapReduce Assignment Cluster 05-02-2024
Amazon EMR release
emr-5.30.1
Application bundle
Custom (HBase 1.4.13, Hadoop 2.8.5, Hive 2.3.6, Hue 4.6.0, JupyterHub 1.1.0, Spark 2.4.5...)
Cluster configuration
Uniform instance groups
Primary (m4.xlarge)
Networking
VPC
vpc-0662e84fd...
Subnet

[Cancel](#) [Clone cluster](#)

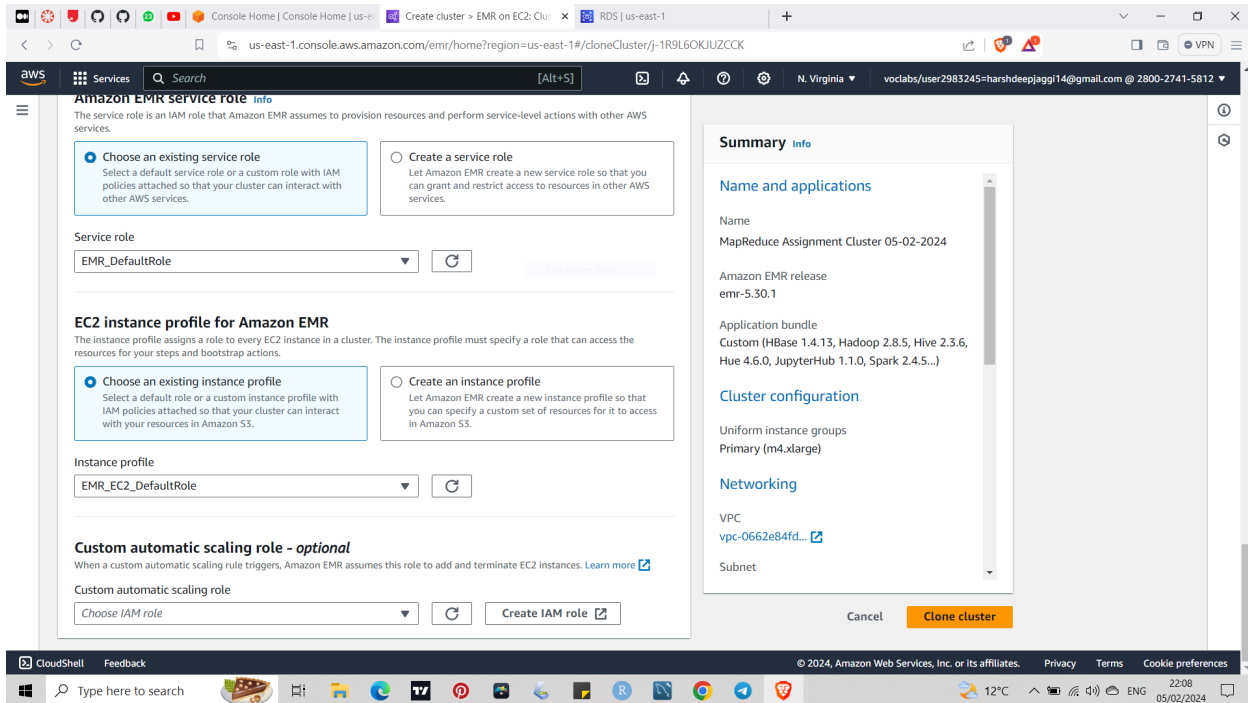
4. Selected **EC2 Key Pair** which is present in my local system & chosen Amazon EMR Service Role as **EMR_DefaultRole**.

Security configuration and EC2 key pair - optional [Info](#)
Security configuration
Select your cluster encryption, authentication, authorization, and instance metadata service settings.
[Choose a security configuration](#) [Browse](#) [Create security configuration](#)
Amazon EC2 key pair for SSH to the cluster [Info](#)
Harsh_Key_pair_demo [Browse](#) [Create key pair](#)
Identity and Access Management (IAM) roles [Info](#)
Choose or create a service role and instance profile for the EC2 instances in your cluster.
Amazon EMR service role [Info](#)
The service role is an IAM role that Amazon EMR assumes to provision resources and perform service-level actions with other AWS services.
☒ Choose an existing service role
Select a default service role or a custom role with IAM policies attached so that your cluster can interact with other AWS services.
☐ Create a service role
Let Amazon EMR create a new service role so that you can grant and restrict access to resources in other AWS services.
Service role
EMR_DefaultRole [Refresh](#)
EC2 instance profile for Amazon EMR
The instance profile assigns a role to every EC2 instance in a cluster. The instance profile must specify a role that can access the resources for your steps and bootstrap actions.

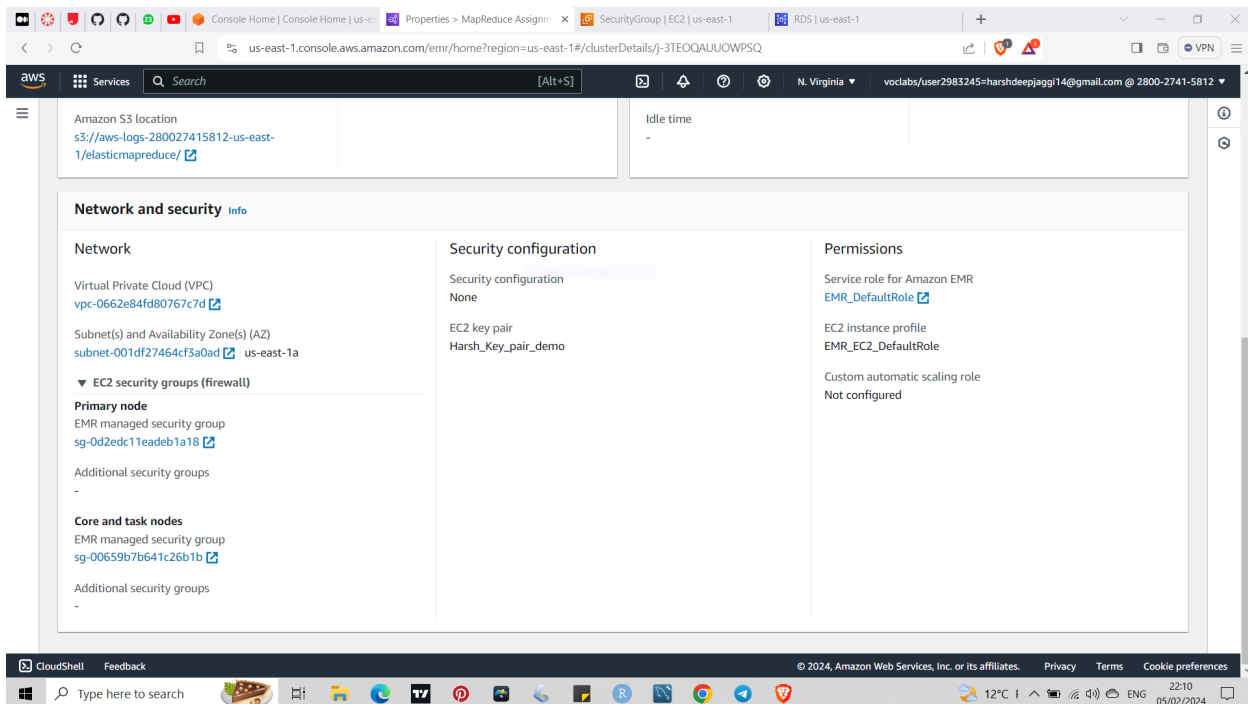
Summary [Info](#)
Name and applications
Name
MapReduce Assignment Cluster 05-02-2024
Amazon EMR release
emr-5.30.1
Application bundle
Custom (HBase 1.4.13, Hadoop 2.8.5, Hive 2.3.6, Hue 4.6.0, JupyterHub 1.1.0, Spark 2.4.5...)
Cluster configuration
Uniform instance groups
Primary (m4.xlarge)
Networking
VPC
vpc-0662e84fd...
Subnet

[Cancel](#) [Clone cluster](#)

5. Selected Instance profile as **EMR_EC2_DefaultRole** & then clicked on the **Clone Cluster** 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.

The screenshot shows the AWS Management Console interface. The left sidebar contains navigation options like EC2 Dashboard, EC2 Global View, Events, Console-to-Code, and a list of services under 'Instances', 'Images', and 'Elastic Block Store'. The main content area displays the 'sg-0d2edc11eadeb1a18 - ElasticMapReduce-master' page. The 'Details' section shows the security group name, ID, description, VPC ID, owner, and rule counts. The 'Inbound rules' tab is active, showing a table with 8 rules. The 'Edit inbound rules' button is visible in the top right of the rules section.

EC2 > Security Groups > sg-0d2edc11eadeb1a18 - ElasticMapReduce-master

sg-0d2edc11eadeb1a18 - ElasticMapReduce-master

Actions

Details

Security group name ElasticMapReduce-master	Security group ID sg-0d2edc11eadeb1a18	Description Master group for Elastic MapReduce created on 2024-01-22T19:45:49.569Z	VPC ID vpc-0662e84fd80767c7d
Owner 280027415812	Inbound rules count 8 Permission entries	Outbound rules count 1 Permission entry	

Inbound rules (8) Manage tags Edit inbound rules

Search

Name	Security group rule...	IP version	Type	Protocol	Port range

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

The screenshot shows the 'Modify Inbound Security Group Rules' page for the security group sg-0d2edc11eadeb1a18. The page lists existing rules and a new rule has been added. The new rule is for SSH, using TCP protocol, port 22, and source 0.0.0.0/0. The 'Add rule' button is visible at the bottom left. A warning message at the bottom states: 'Rules with source of 0.0.0.0 or ::/0 allow all IP addresses to access your instance. We recommend setting security group rules to allow access from known IP addresses only.'

sg-0b8c7cb351d899b0b All TCP TCP 0 - 65535 Custom Q sg-0d2edc11eadeb1a18 X Delete

sg-0f73925b96366b324 All UDP UDP 0 - 65535 Custom Q sg-0d2edc11eadeb1a18 X Delete

sg-0bc3588a0dde102bc All ICMP - IPv4 ICMP All Custom Q sg-0d2edc11eadeb1a18 X Delete

sg-05b9dd139b26d0fcc SSH TCP 22 Custom Q 0.0.0.0/0 X Delete

sg-07d325c872bf92120 Custom TCP TCP 8443 Custom Q pl-f8bd5e91 X Delete

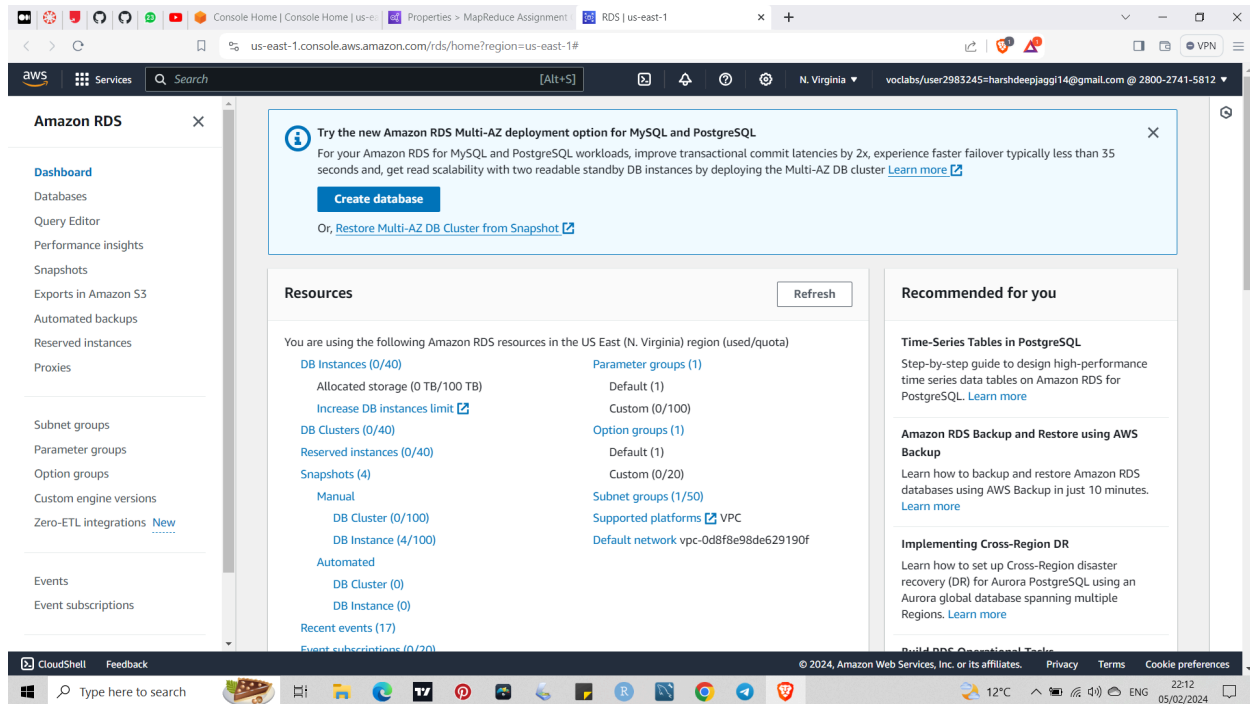
Add rule

Rules with source of 0.0.0.0 or ::/0 allow all IP addresses to access your instance. We recommend setting security group rules to allow access from known IP addresses only.

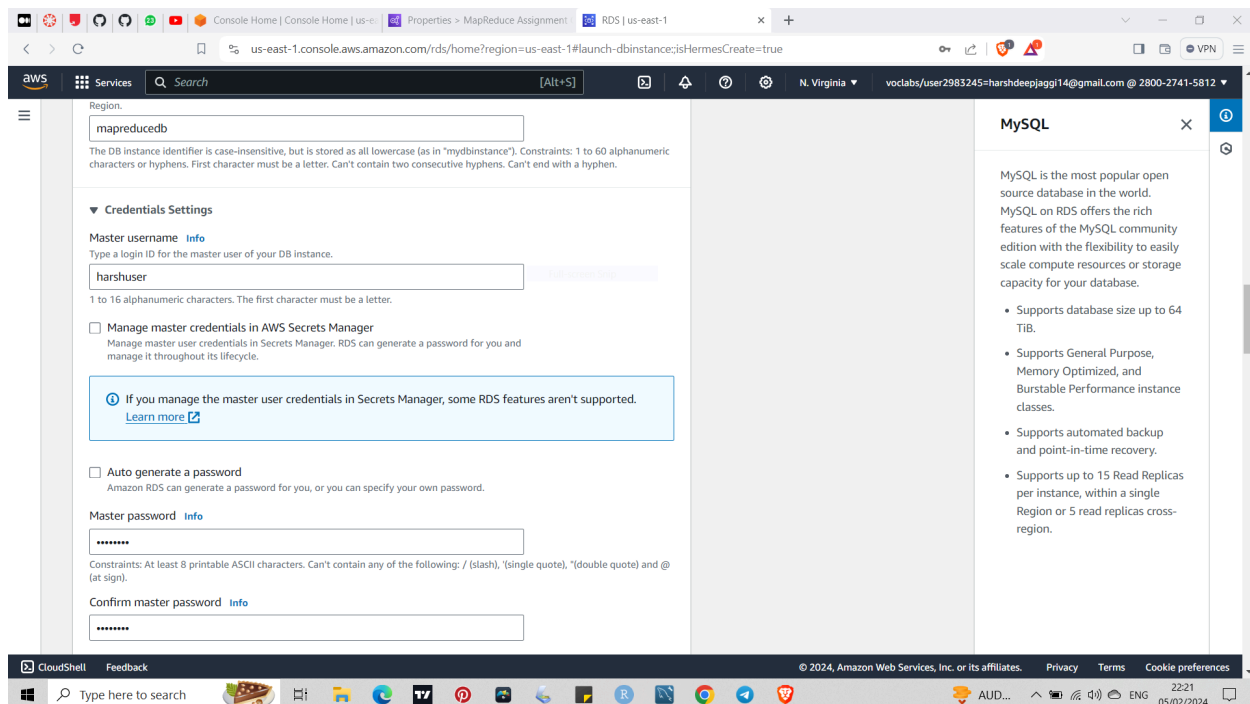
Cancel Preview changes Save rules

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

The screenshot shows the AWS Management Console for an Amazon RDS instance. The 'Instance configuration' page is active, displaying options for the DB instance class and storage. The 'DB instance class' is set to 'db.t2.micro' (1 vCPUs, 1 GiB RAM, Not EBS Optimized). The 'Storage type' is 'General Purpose SSD (gp2)' and 'Allocated storage' is '20 GiB'. A sidebar on the right provides information about MySQL on RDS, including its popularity and supported features like database size up to 64 TiB, General Purpose/Memory Optimized/Burstable Performance instance classes, automated backup and point-in-time recovery, and up to 15 Read Replicas per instance.

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

The screenshot shows the AWS Management Console for an Amazon RDS instance, specifically the 'Connectivity' page. The 'Compute resource' section has two options: 'Don't connect to an EC2 compute resource' (selected) and 'Connect to an EC2 compute resource'. The 'Virtual private cloud (VPC)' is set to 'Default VPC (vpc-0d8f8e98de629190f)'. A warning message states: 'After a database is created, you can't change its VPC.' The 'DB subnet group' is set to 'default-vpc-0d8f8e98de629190f'. The 'Public access' option is set to 'Yes'. The same 'MySQL' sidebar from the previous screenshot is visible on the right.

5. Changed Public access to Yes.

Public access [Info](#)

☒ **Yes**
RDS assigns a public IP address to the database. Amazon EC2 instances and other resources outside of the VPC can connect to your database. Resources inside the VPC can also connect to the database. Choose one or more VPC security groups that specify which resources can connect to the database.

☐ **No**
RDS doesn't assign a public IP address to the database. Only Amazon EC2 instances and other resources inside the VPC can connect to your database. Choose one or more VPC security groups that specify which resources can connect to the database.

VPC security group (firewall) [Info](#)
Choose one or more VPC security groups to allow access to your database. Make sure that the security group rules allow the appropriate incoming traffic.

☒ **Choose existing**
Choose existing VPC security groups

☐ **Create new**
Create new VPC security group

Existing VPC security groups

Choose one or more options

default X

Availability Zone [Info](#)

No preference

RDS Proxy
RDS Proxy is a fully managed, highly available database proxy that improves application scalability, resiliency, and security.

☐ **Create an RDS Proxy** [Info](#)
RDS automatically creates an IAM role and a Secrets Manager secret for the proxy. RDS Proxy has additional costs. For more information, see [Amazon RDS Proxy pricing](#).

Certificate authority - optional [Info](#)
Using a server certificate provides an extra layer of security by validating that the connection is being made to an Amazon database.

MySQL

MySQL is the most popular open source database in the world. MySQL on RDS offers the rich features of the MySQL community edition with the flexibility to easily scale compute resources or storage capacity for your database.

- Supports database size up to 64 TiB.
- Supports General Purpose, Memory Optimized, and Burstable Performance instance classes.
- Supports automated backup and point-in-time recovery.
- Supports up to 15 Read Replicas per instance, within a single Region or 5 read replicas cross-region.

Database authentication options [Info](#)

☒ **Password authentication**
Authenticates using database passwords.

☐ **Password and IAM database authentication**
Authenticates using the database password and user credentials through AWS IAM users and roles.

☐ **Password and Kerberos authentication**
Choose a directory in which you want to allow authorized users to authenticate with this DB instance using Kerberos Authentication.

Monitoring

☐ **Enable Enhanced Monitoring**
Enabling Enhanced Monitoring metrics are useful when you want to see how different processes or threads use the CPU.

Additional configuration
Database options, backup turned on, backup turned off, maintenance, CloudWatch Logs, delete protection turned off.

Estimated monthly costs

The Amazon RDS Free Tier is available to you for 12 months. Each calendar month, the free tier will allow you to use the Amazon RDS resources listed below for free:

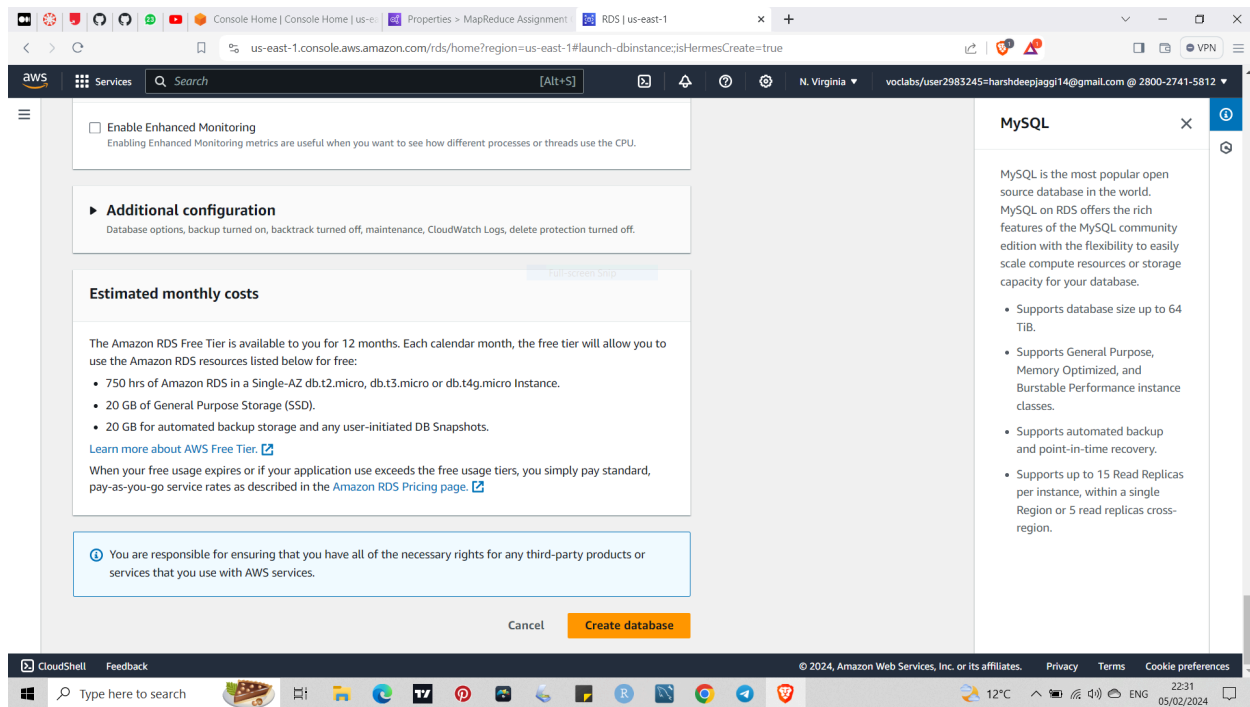
- 750 hrs of Amazon RDS in a Single-AZ db.t2.micro, db.t3.micro or db.t4g.micro Instance.

MySQL

MySQL is the most popular open source database in the world. MySQL on RDS offers the rich features of the MySQL community edition with the flexibility to easily scale compute resources or storage capacity for your database.

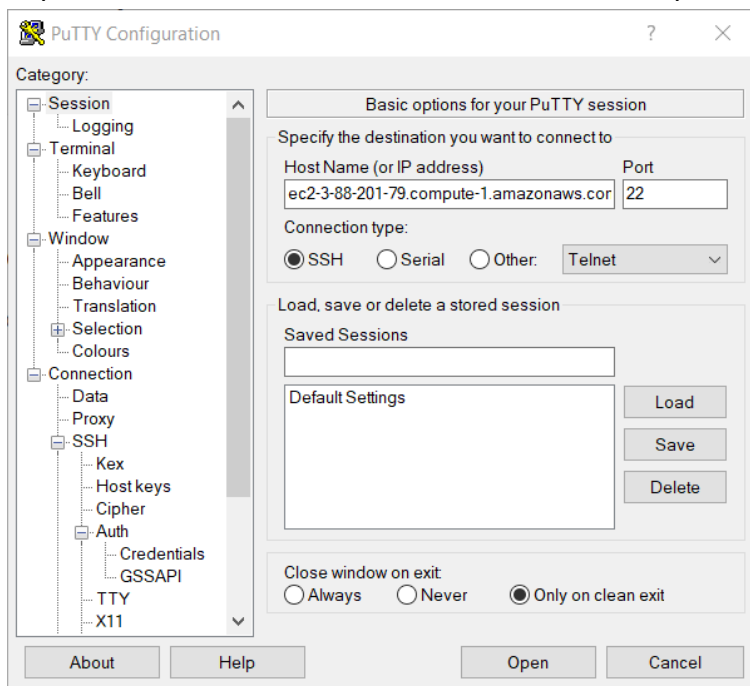
- Supports database size up to 64 TiB.
- Supports General Purpose, Memory Optimized, and Burstable Performance instance classes.
- Supports automated backup and point-in-time recovery.
- Supports up to 15 Read Replicas per instance, within a single Region or 5 read replicas cross-region.

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

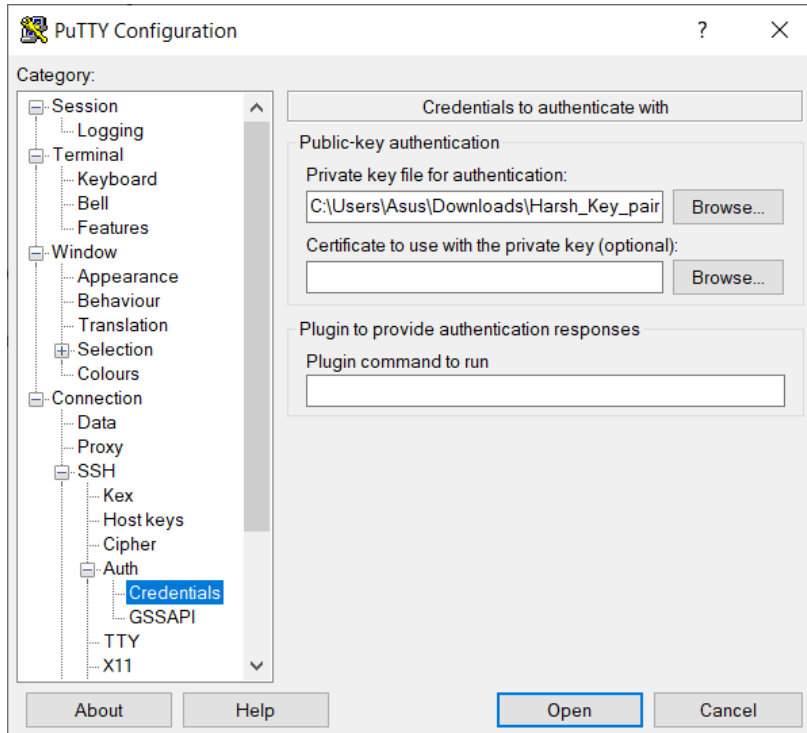


Setting-up Putty Configurations

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

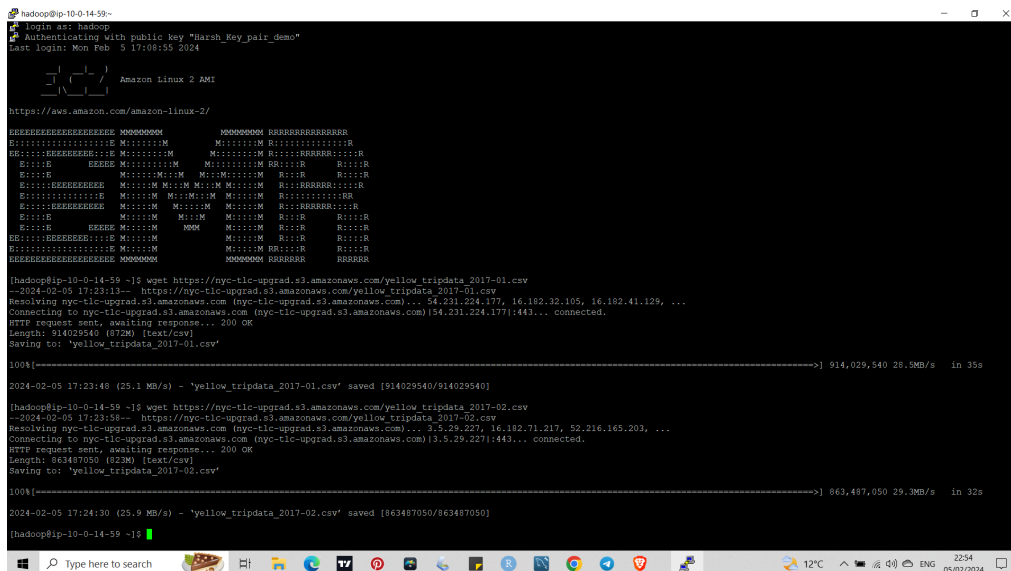


2. Provided the **Private Key File** from the local machine.



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



- 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
);


```
hadoop@ip-10-0-14-59:~$
--> (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 );
Query OK, 0 rows affected (0.04 sec)

MySQL [trip]> show tables;
+-----+
| Tables_in_trip |
+-----+
| trip_2017       |
+-----+
1 row in set (0.00 sec)

MySQL [trip]> select count(*) from trip_2017;
+-----+
| count(*) |
+-----+
|         0 |
+-----+
1 row in set (0.00 sec)

MySQL [trip]> 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;
Query OK, 9710820 rows affected, 65535 warnings (2 min 38.20 sec)
Records: 9710820 Deleted: 0 Skipped: 0 Warnings: 9710820

MySQL [trip]> 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;
Query OK, 9169775 rows affected, 65535 warnings (2 min 33.95 sec)
Records: 9169775 Deleted: 0 Skipped: 0 Warnings: 9169775

MySQL [trip]> select count(*) from trip_2017;
+-----+
| count(*) |
+-----+
| 18880595 |
+-----+
1 row in set (51.12 sec)

MySQL [trip]>
```

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