**Name:** Harsh Chheda

**Roll Number:** 31031521005 / 22-15405

**Class:** Msc. Computer Science

**Subject**: Cloud Computing

**Year:** 2022-23

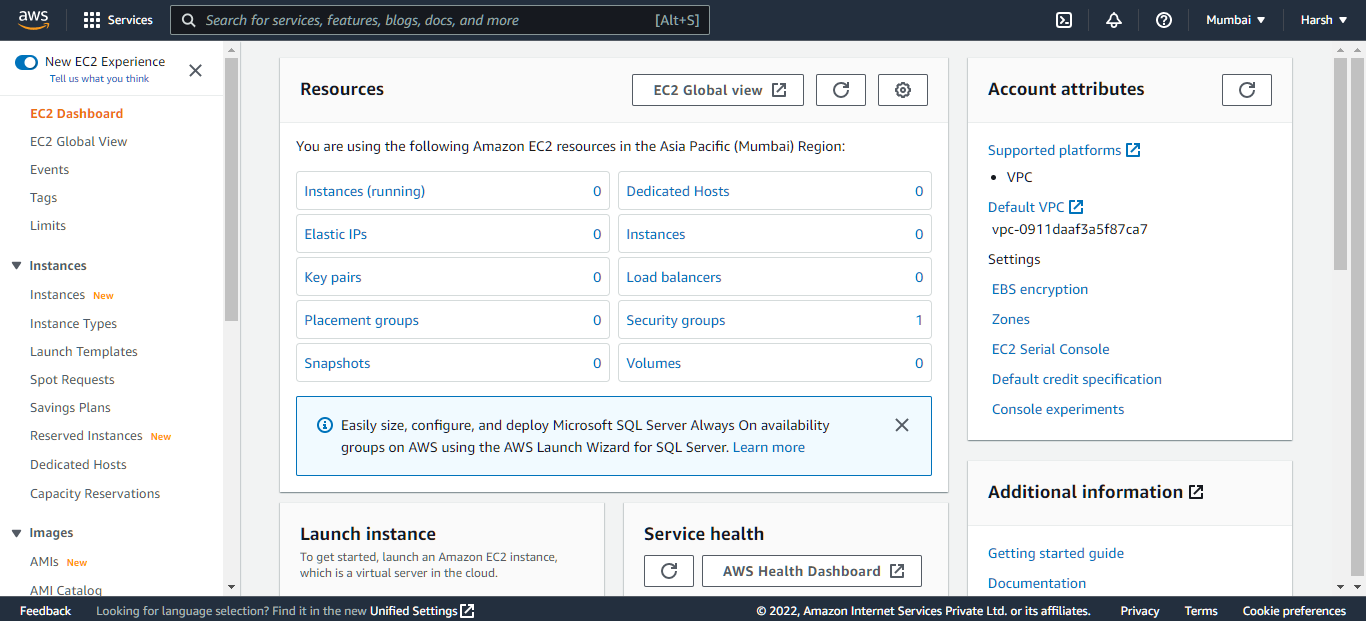
**Practical 8**

**Aim:** Demonstration of data analytics in Cloud

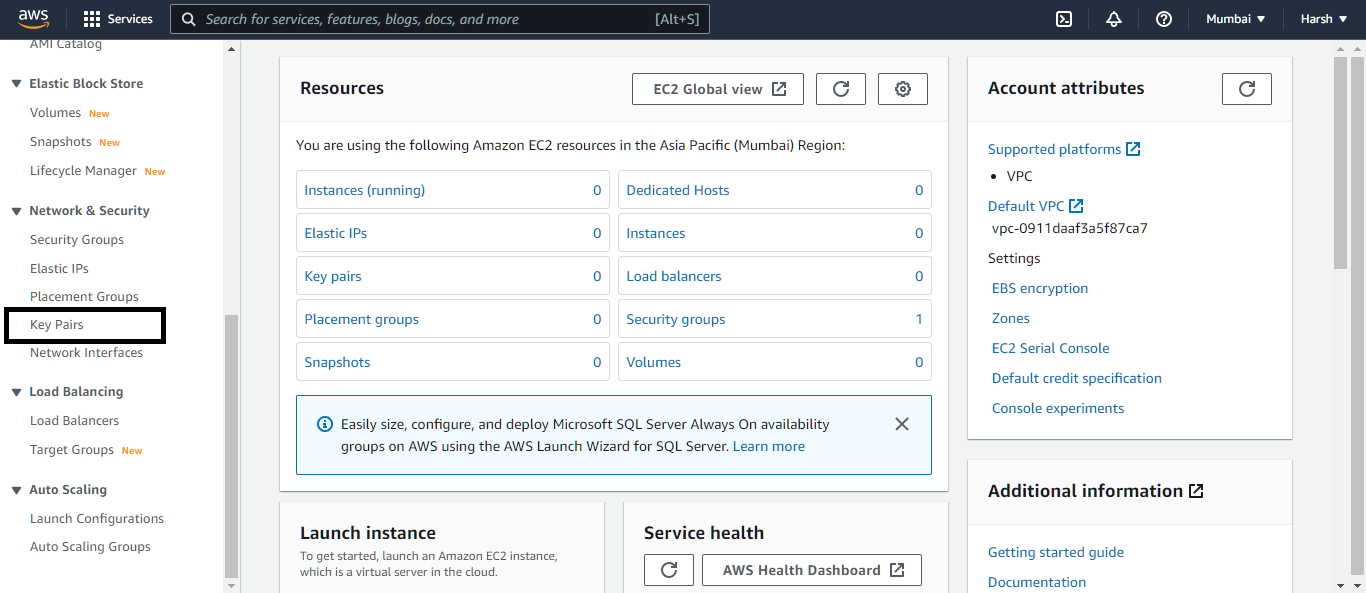
**Code:**

# Setting up EC2 Key-Pairs

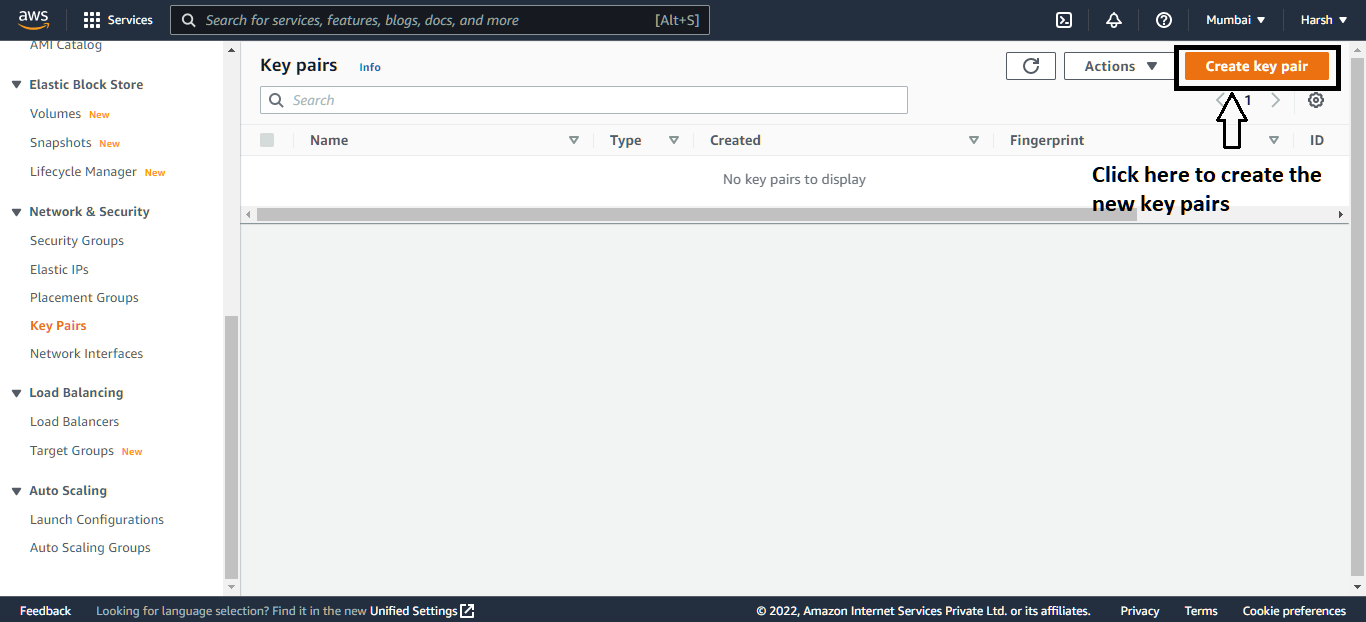
## Step 1: Go to [**Amazon EC2 Console**](https://ap-south-1.console.aws.amazon.com/ec2/home)



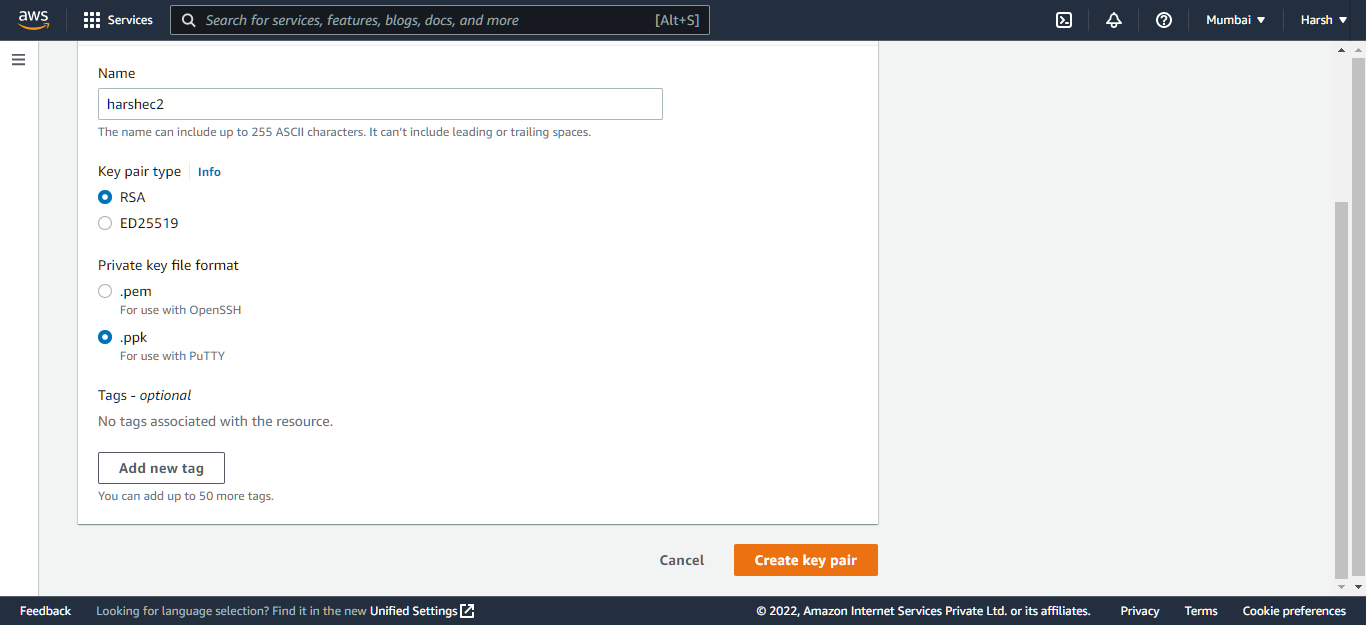
## Step 2: In the Navigation pane, click Key Pairs under Network and Security Section



## Step 3: On the Key Pairs page, click Create Key Pair

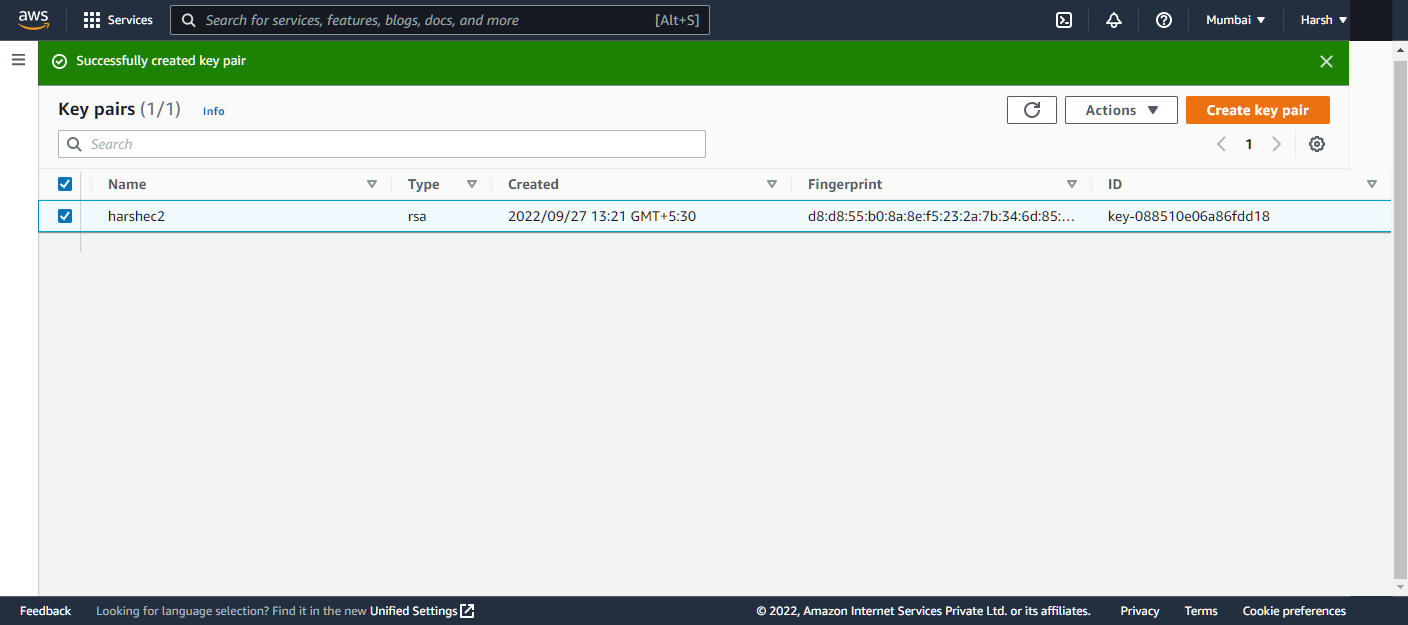


## Step 4: In the Create Key Pair dialog box, enter a name for your key pair, such as, mykeypair



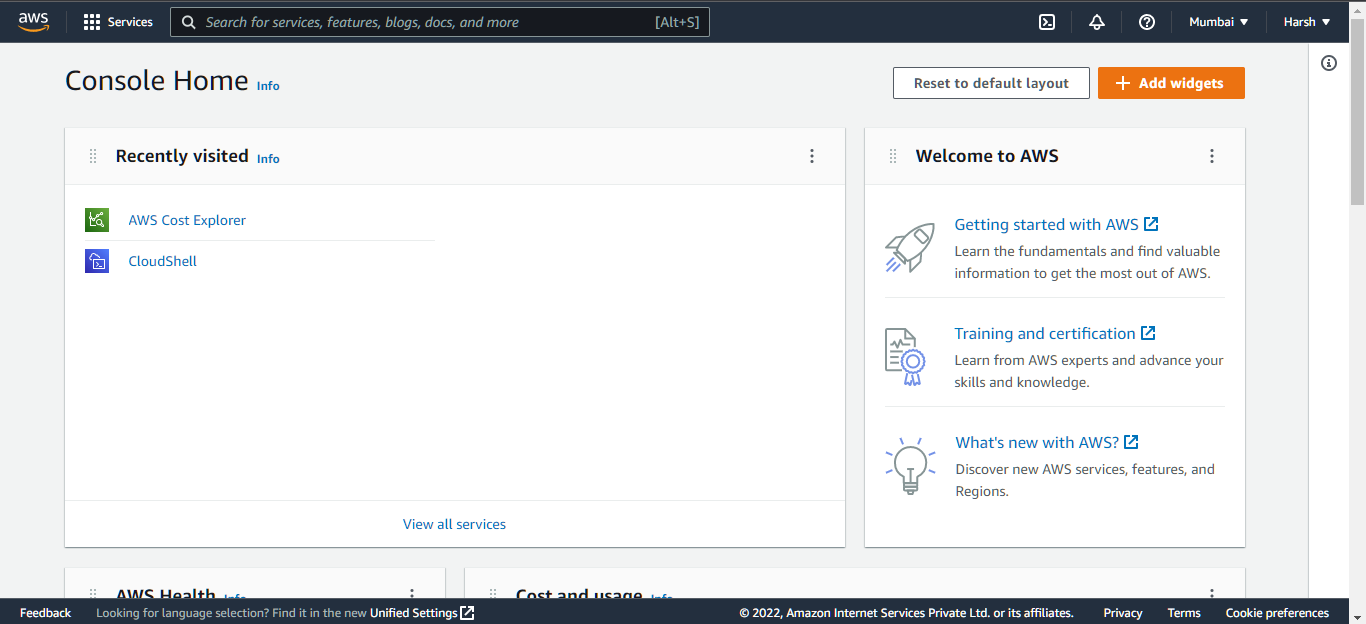
## Step 5: Click Create key Pair

## Step 6: Save the resulting PEM file in a safe location

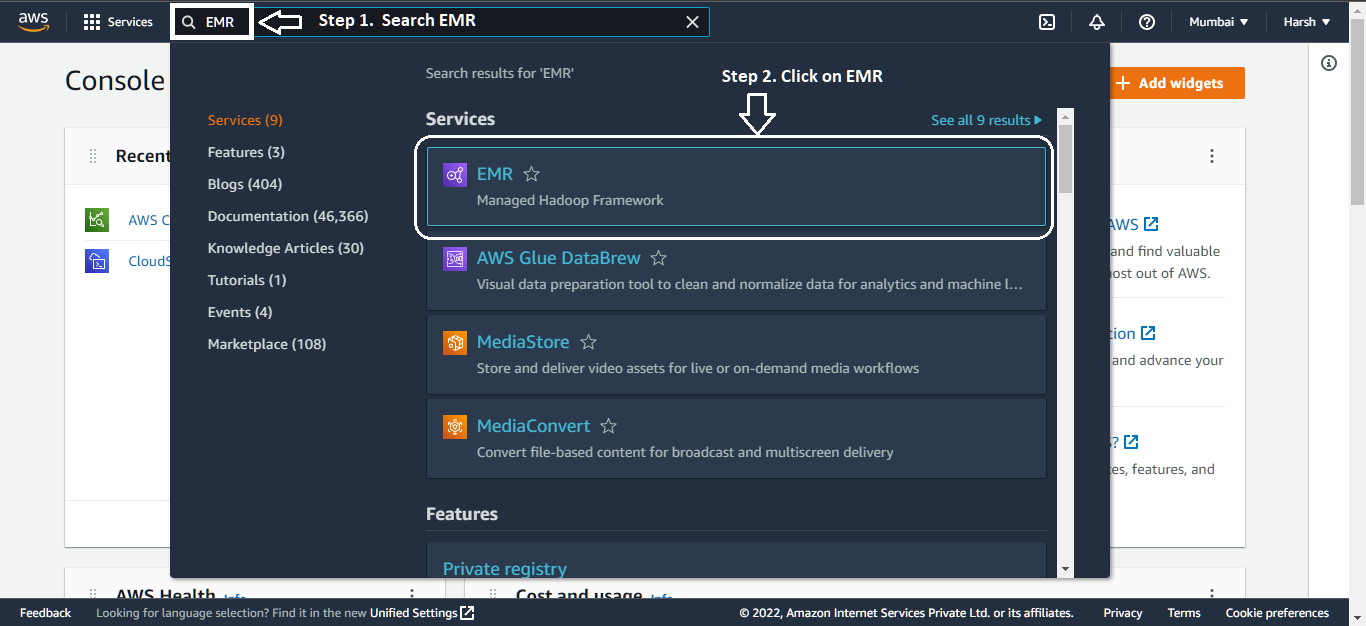


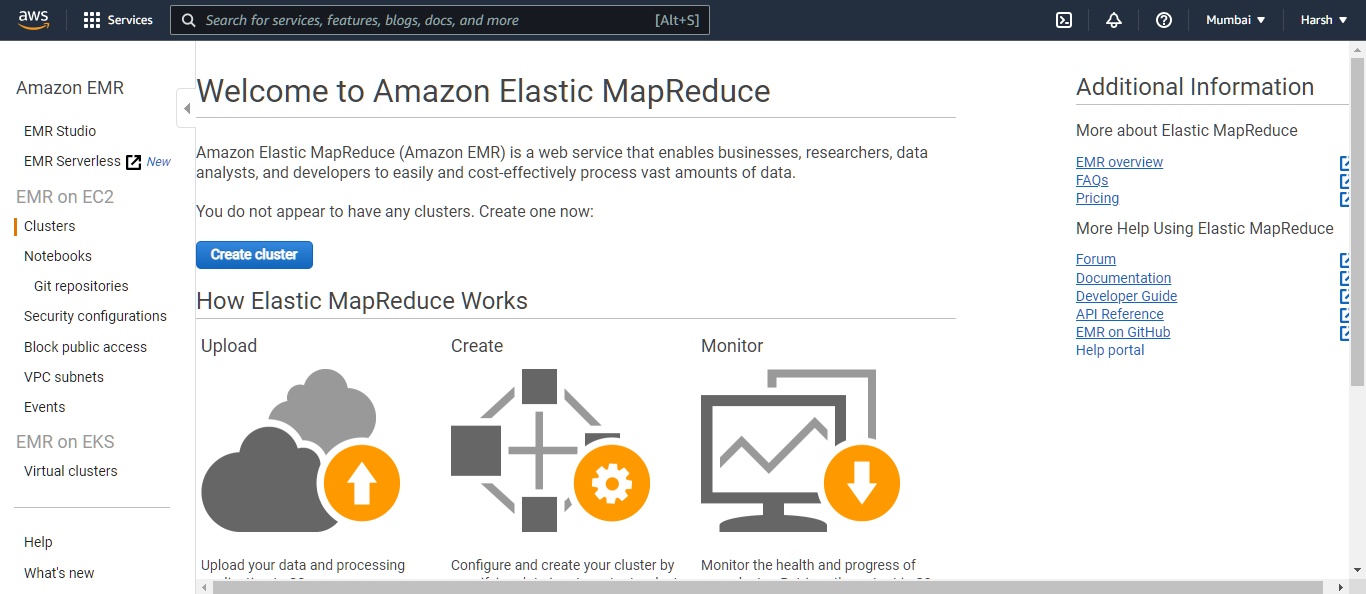
# Setting up your environment on Amazon EMR

## Step 1: Create an [AWS account](https://aws.amazon.com/?source=post_page---------------------------) and sign in to the console.

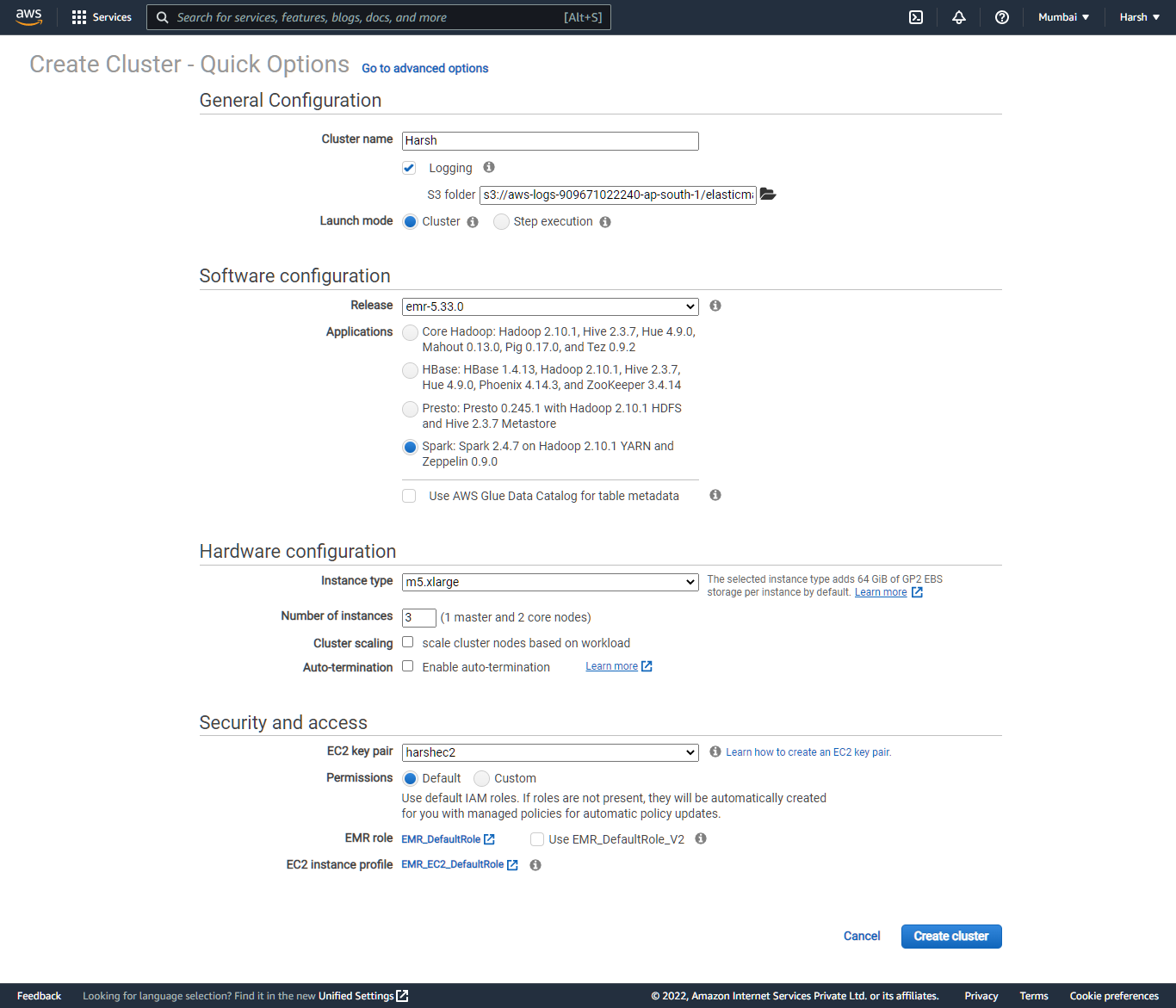


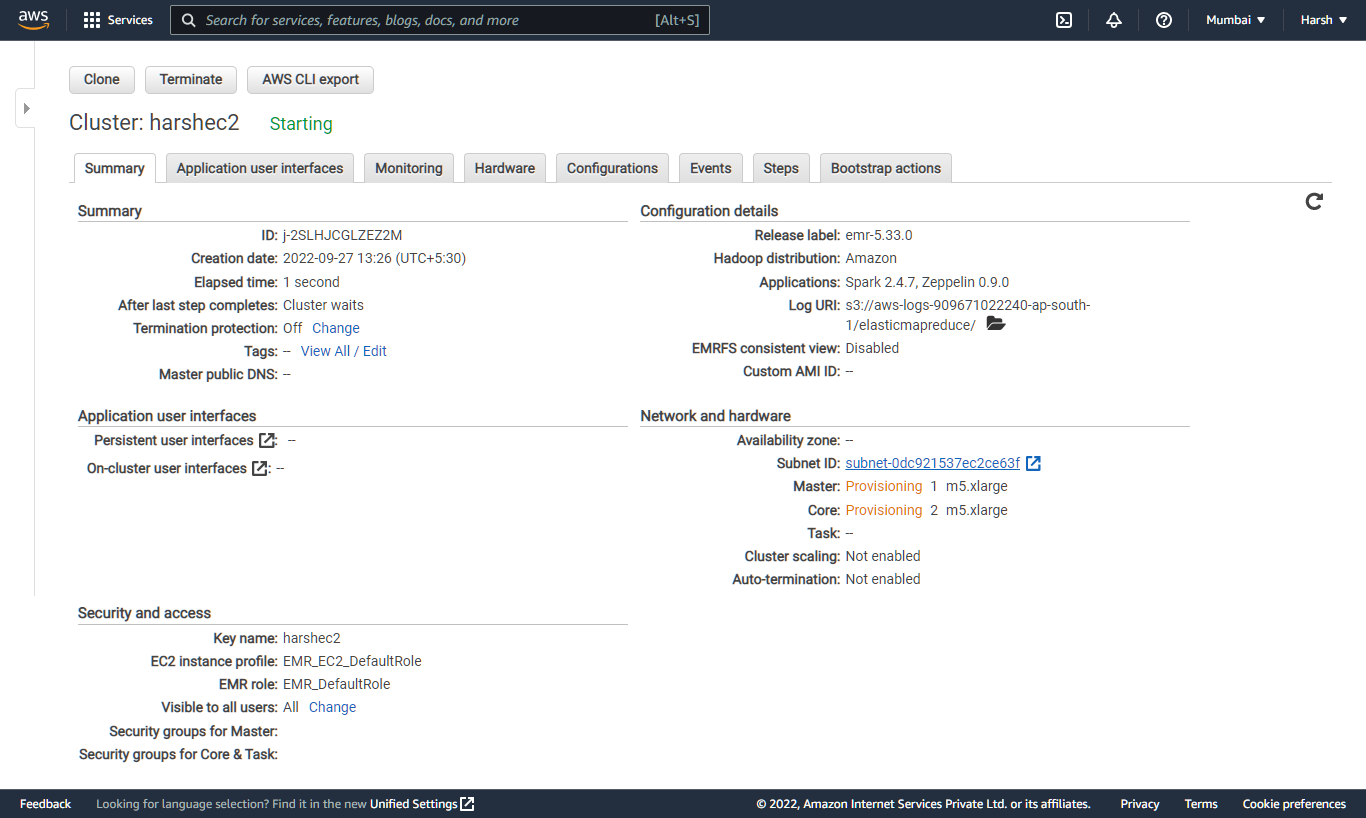
## Step 2: Search **EMR** in the Search Box





## Step 3: Creating new cluster



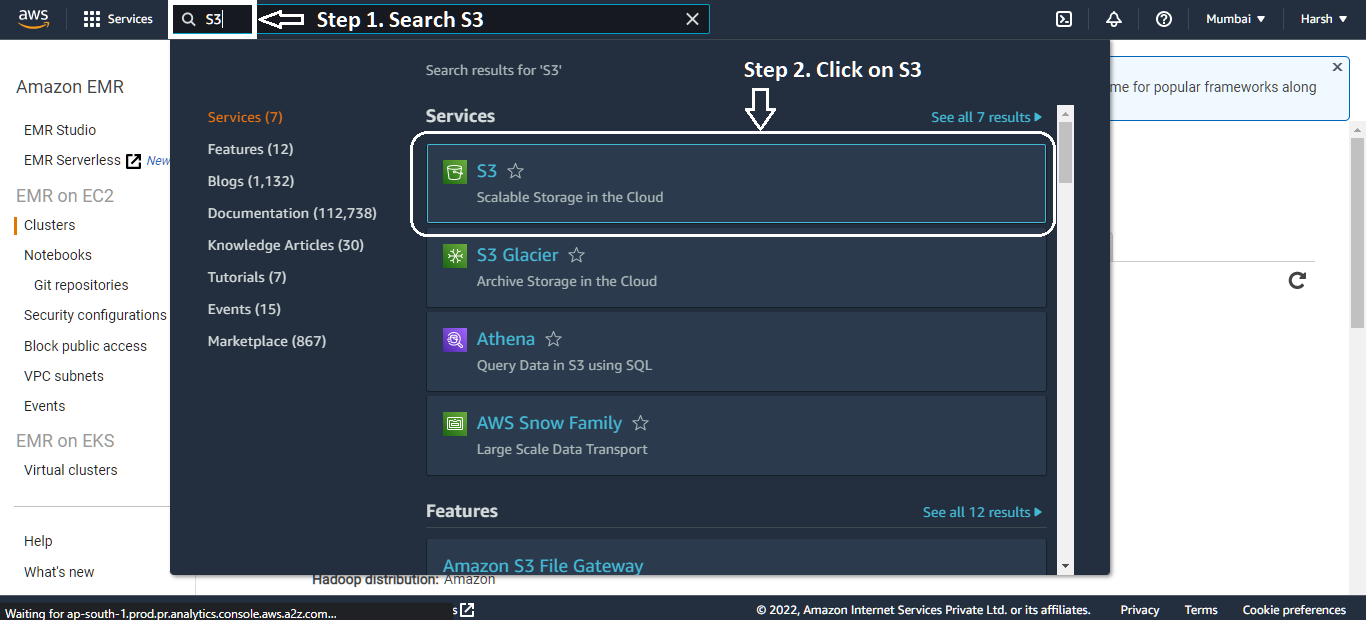


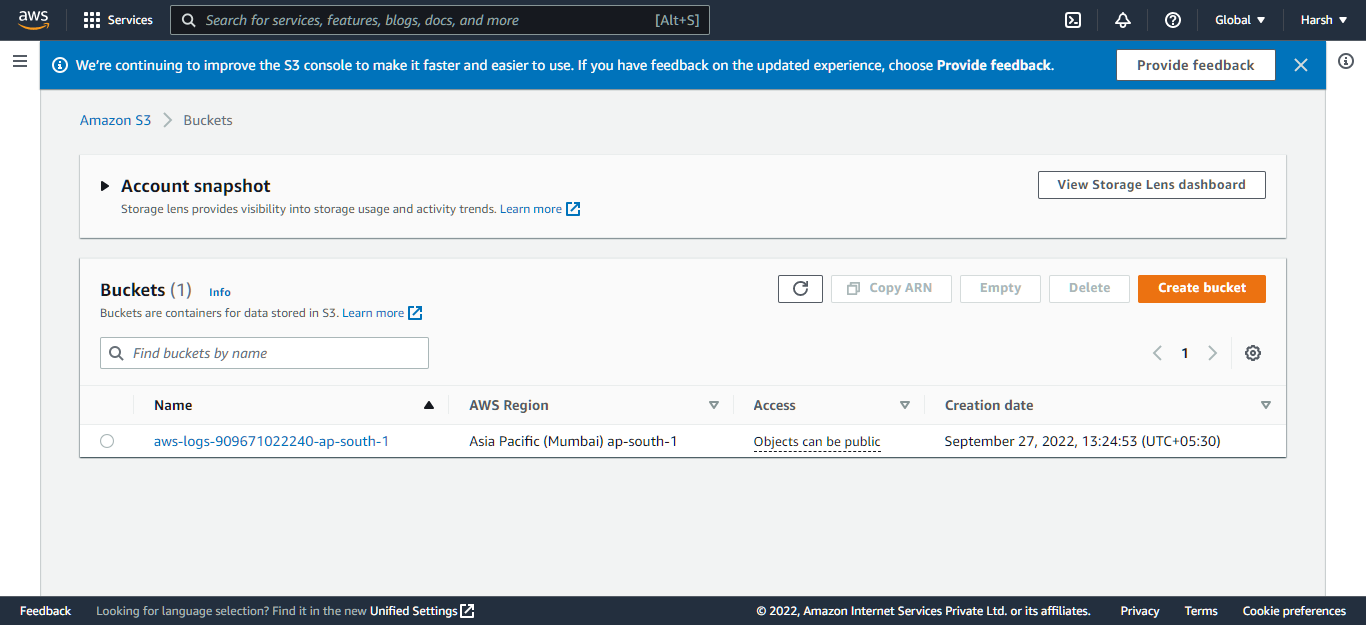
# Downloading Dataset

## Step 1: [**Click Here**](https://info.stackoverflowsolutions.com/rs/719-EMH-566/images/stack-overflow-developer-survey-2020.zip) to download dataset

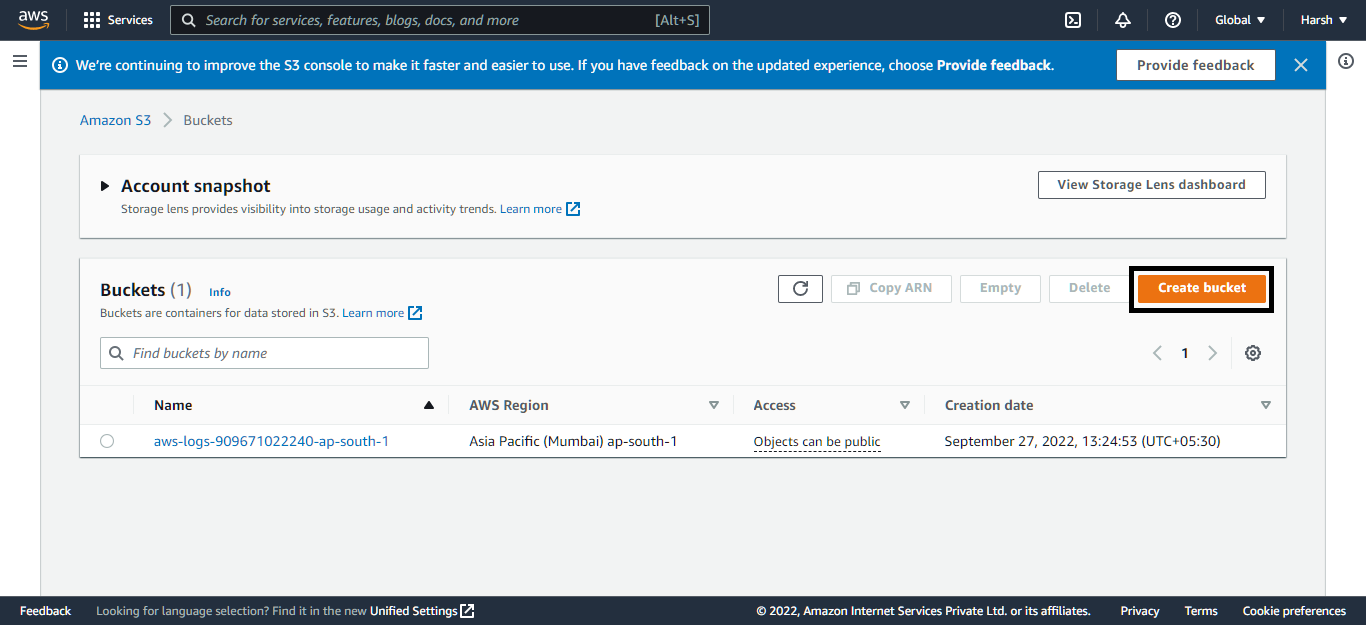
# Setting up S3 Environment

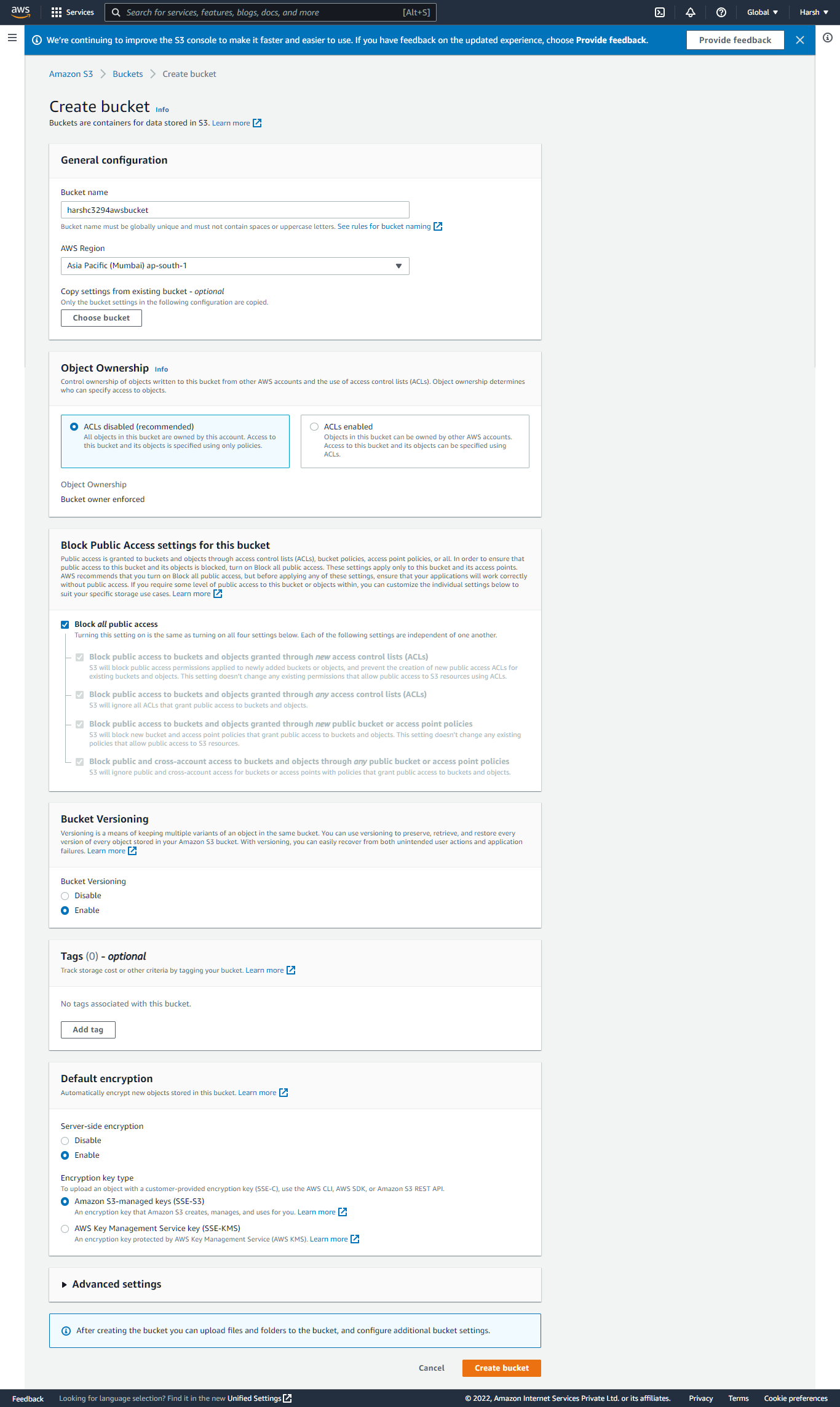
## Step 1: Search **S3** in the Search Box





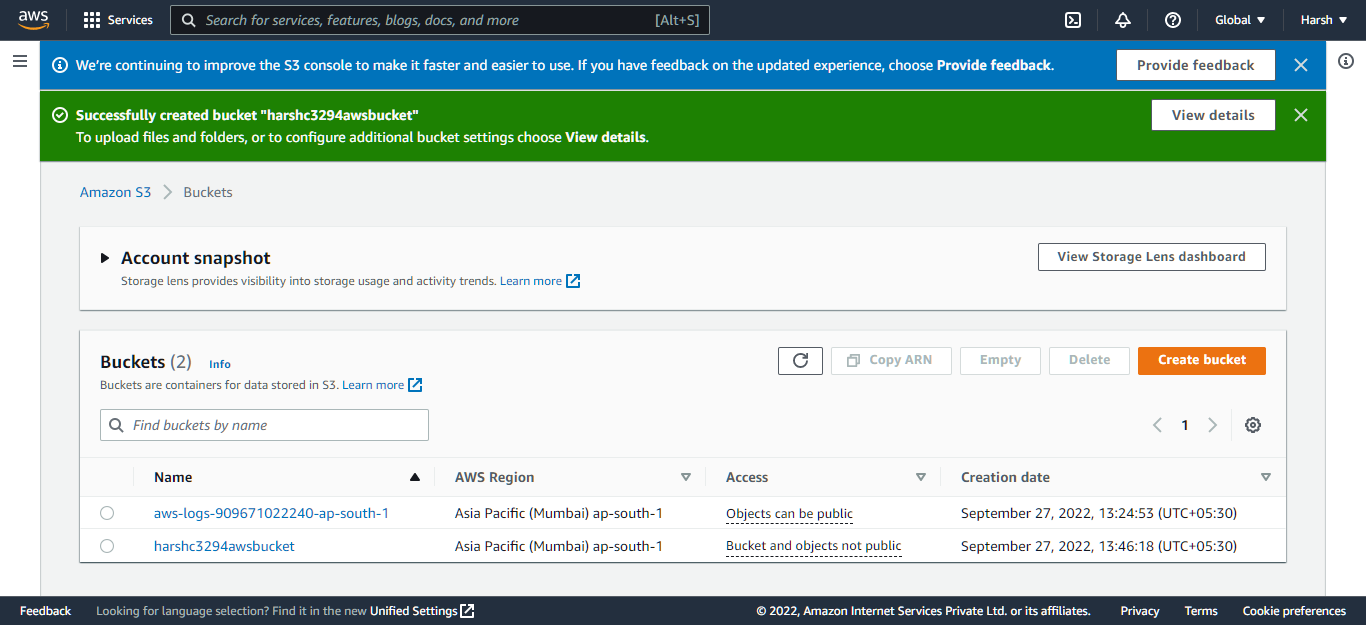
## Step 2: To create new Bucket Click on **Create Bucket**





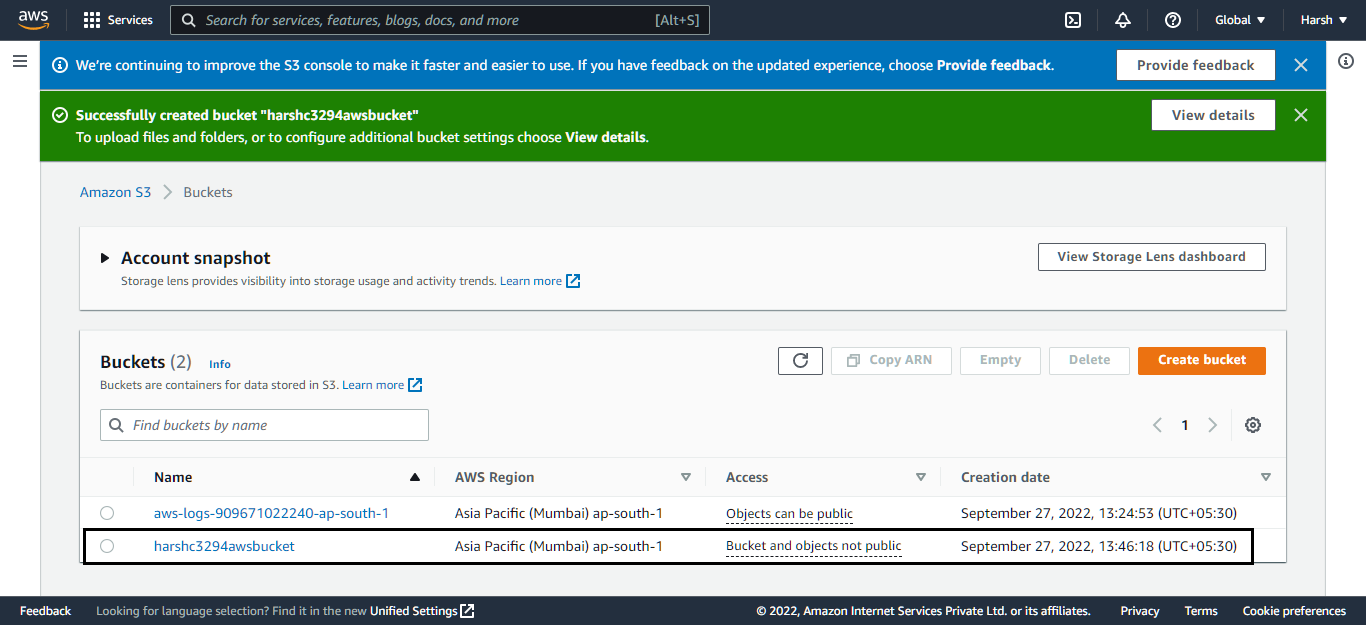
## Step 3: Click on Create Bucket

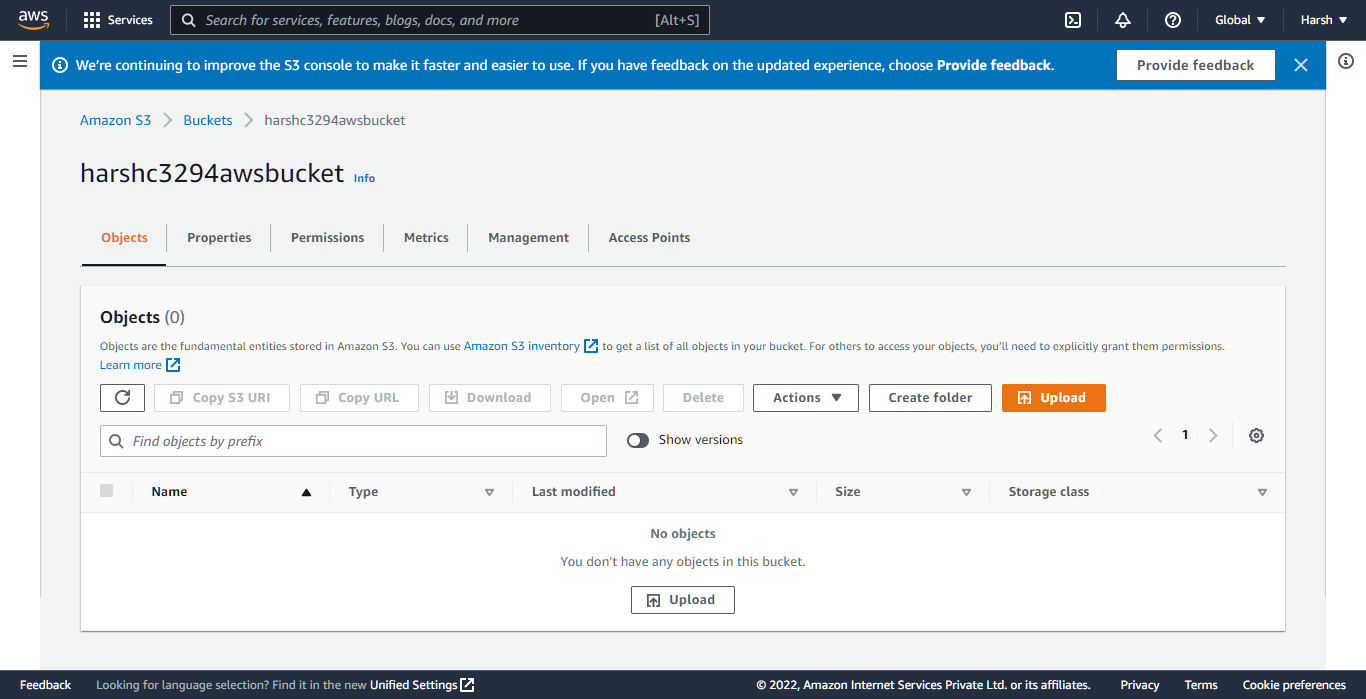
## Step 4: Once the bucket is created you will be able to see the bucket



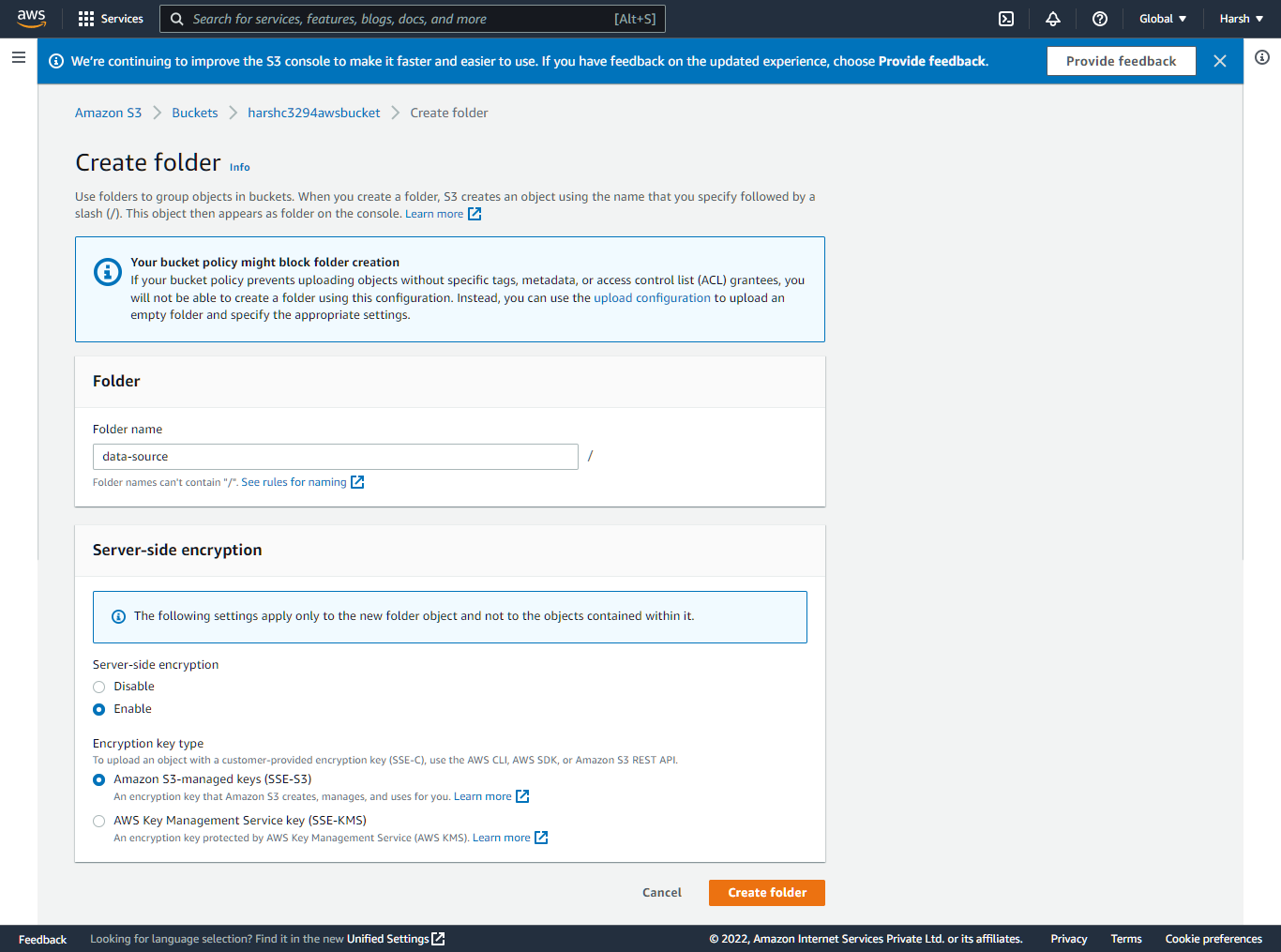
# Uploading Dataset to the S3 Bucket

## Step 1: Click and open the S3 bucket

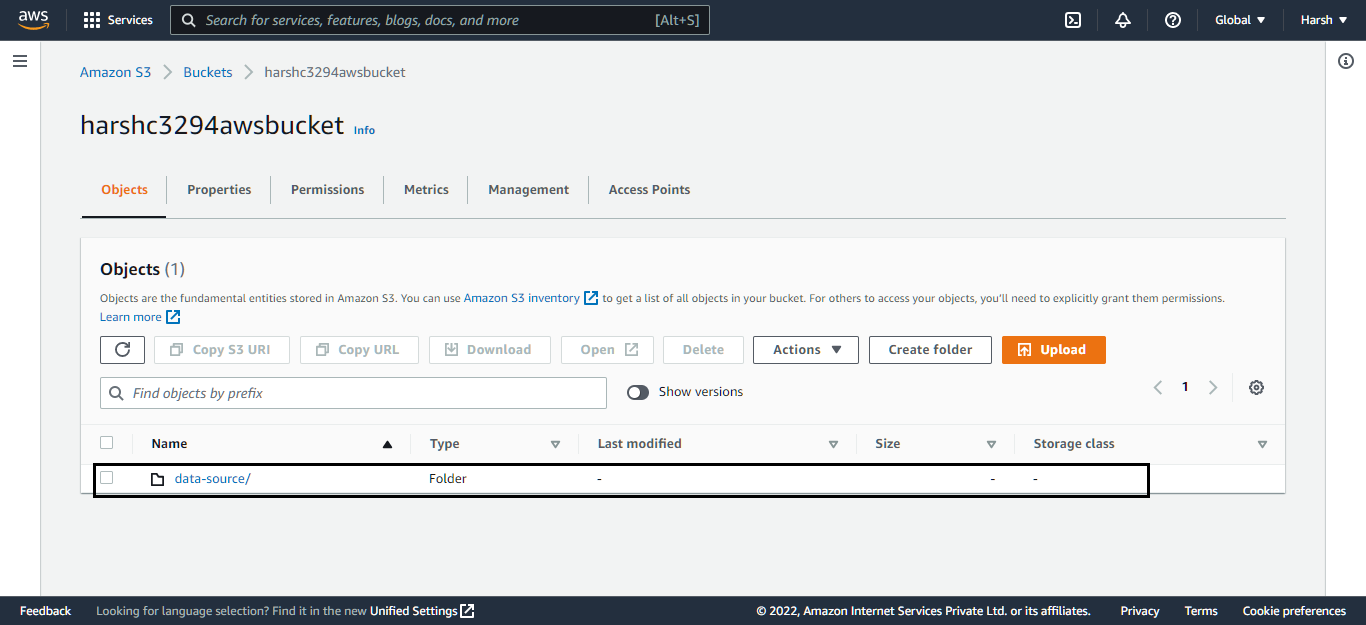


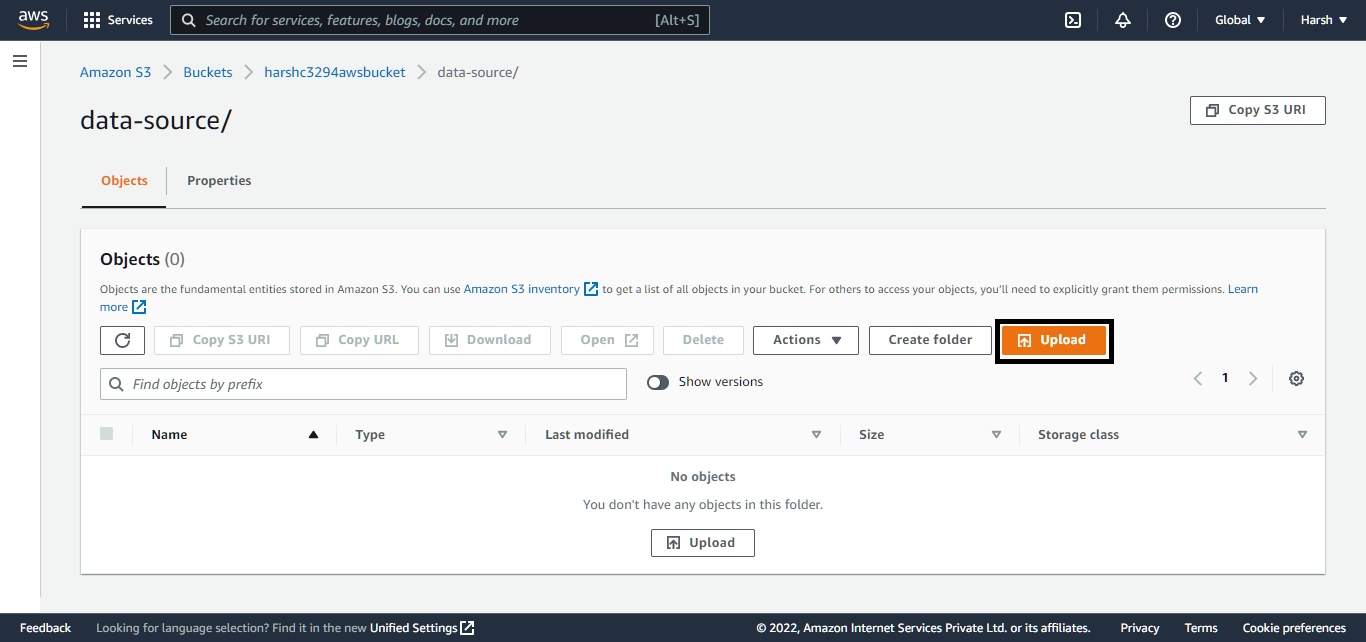


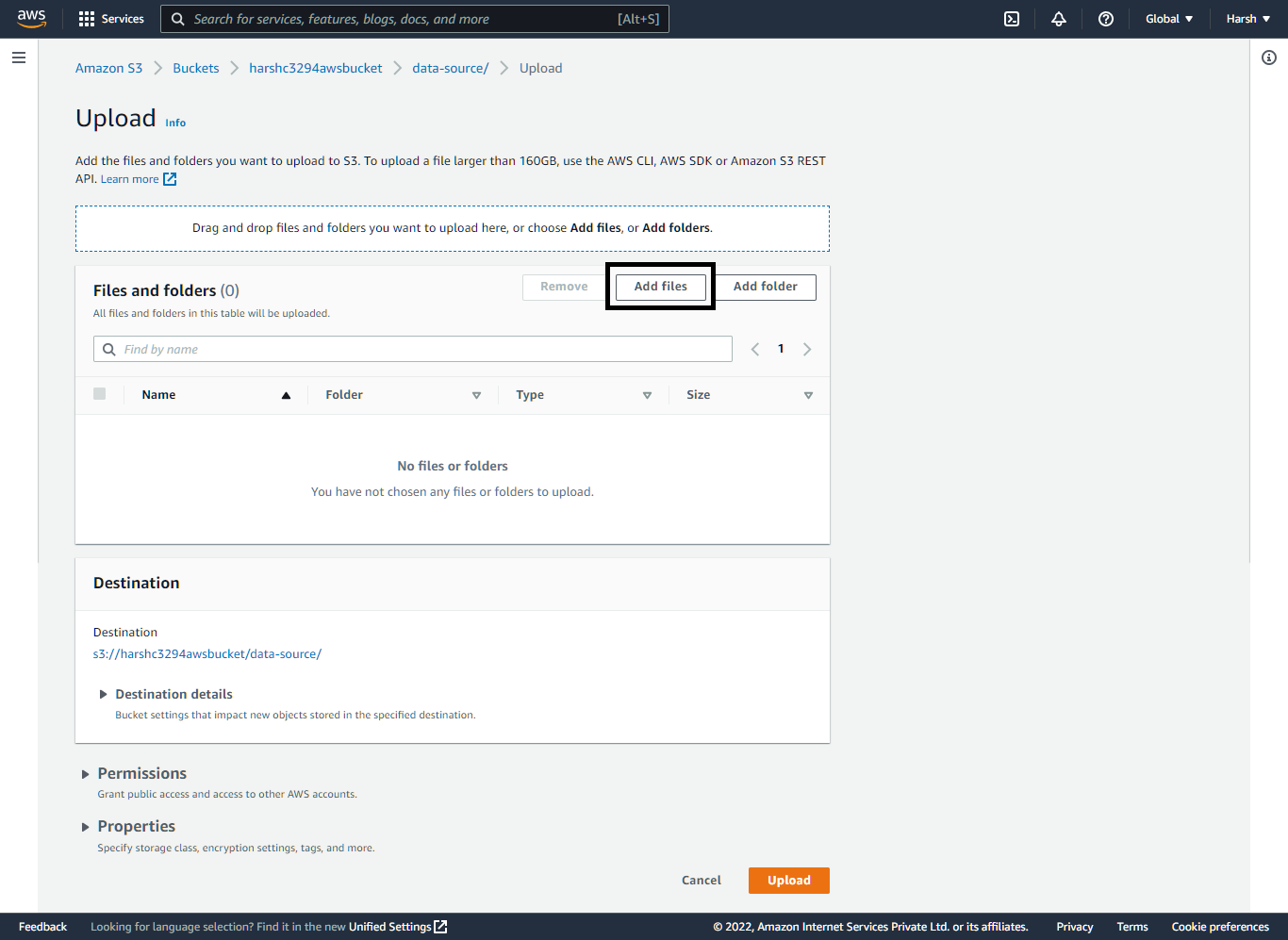
## Step 2: Click on Create Folder and create the new folder with the name data-source and click on Create folder.

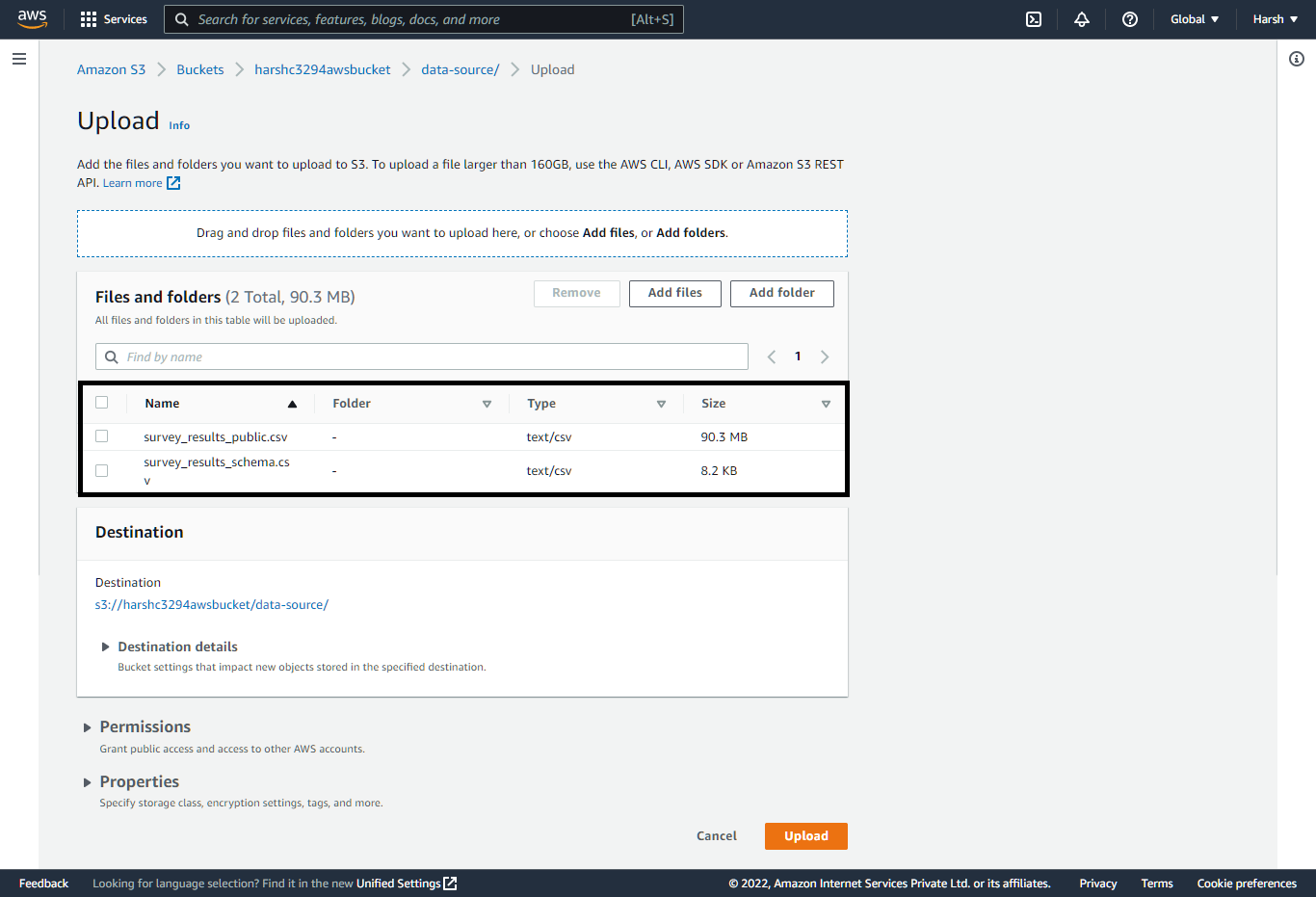


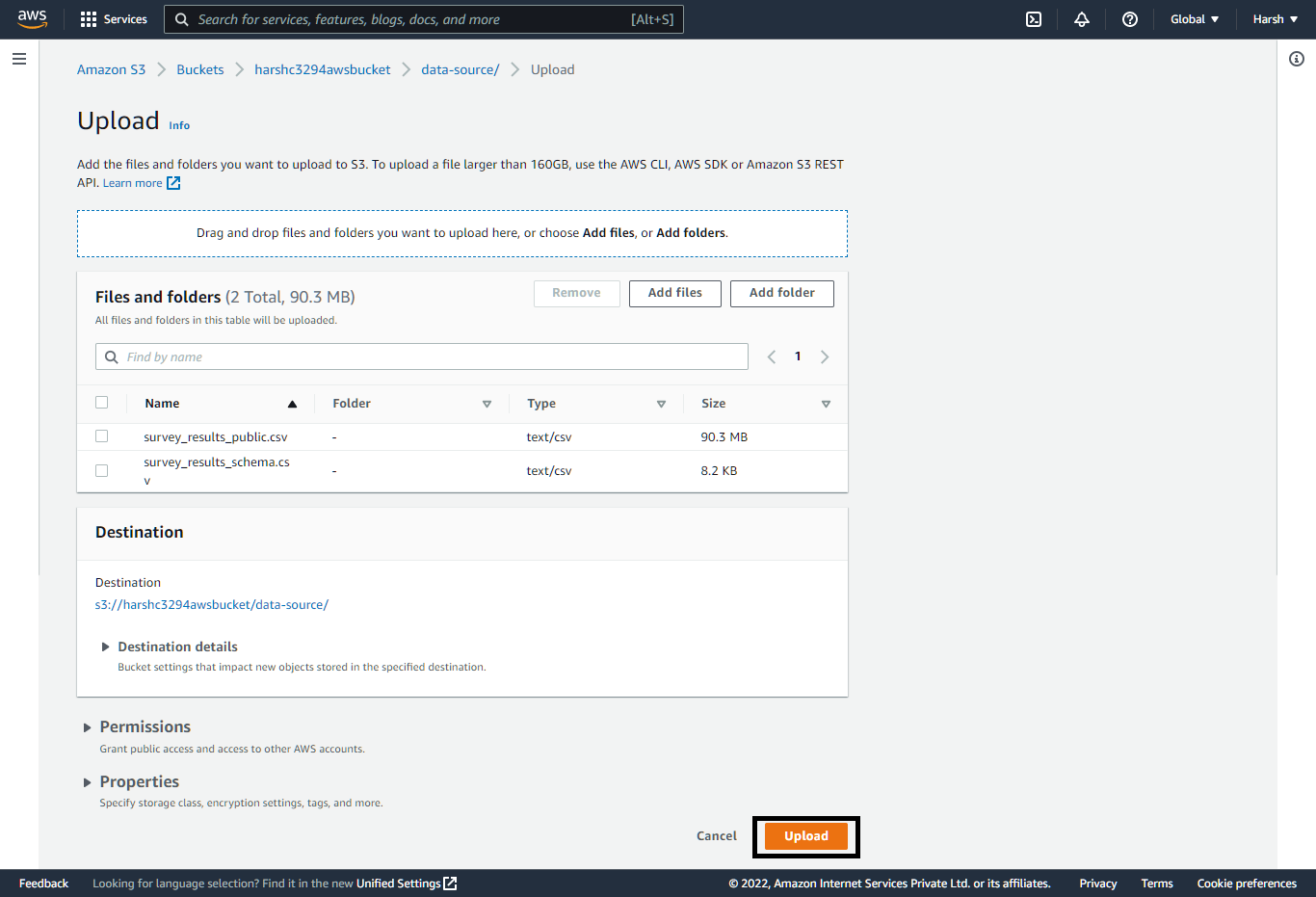
## Step 3: Now open the Folder and upload the dataset.

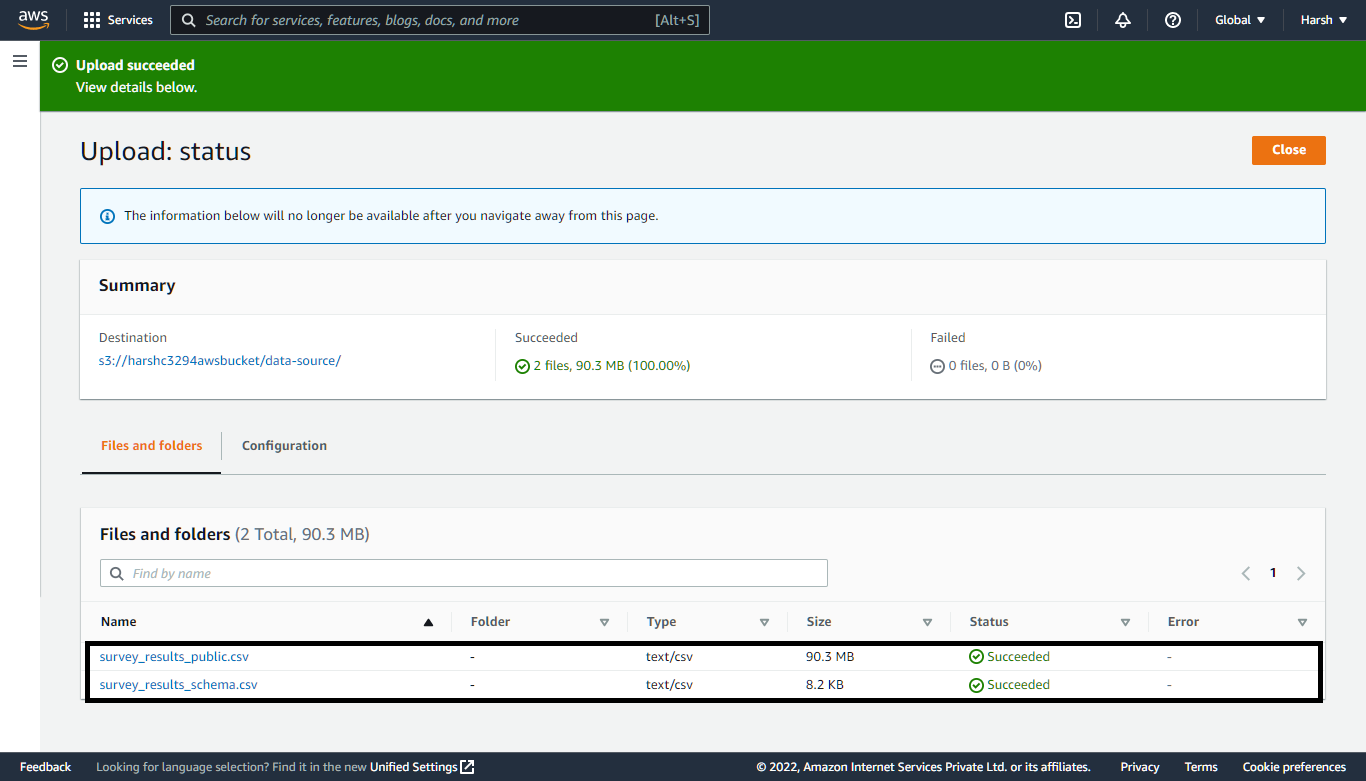












# Code

*from* pyspark.sql *import* SparkSession

*from* pyspark.sql.functions *import* col

*# FIND THE S3 URI IN THE S3 BUCKET*

*# PATH  S3\_URI/FILENAME*

S3\_DATA\_SOURCE\_PATH="s3://harshc3294awsbucket/data-source/survey\_results\_public.csv"

S3\_DATA\_OUTPUT\_PATH="s3://harshc3294awsbucket/data-output"

def main ():

    spark= SparkSession.builder.appName("HarshDemoApp").getOrCreate()

    all\_data=spark.read.csv(S3\_DATA\_SOURCE\_PATH,*header*=True)

    print("The total number of records int the source data : %s" % all\_data.count())

    selected\_data = all\_data.where((col("Country")=="United States") & (col("WorkWeekHrs")>45))

    print("The number of engineers who worked more than 45 hours a week in the US are: %s" % selected\_data.count())

    selected\_data.write.mode("overwrite").parquet(S3\_DATA\_OUTPUT\_PATH)

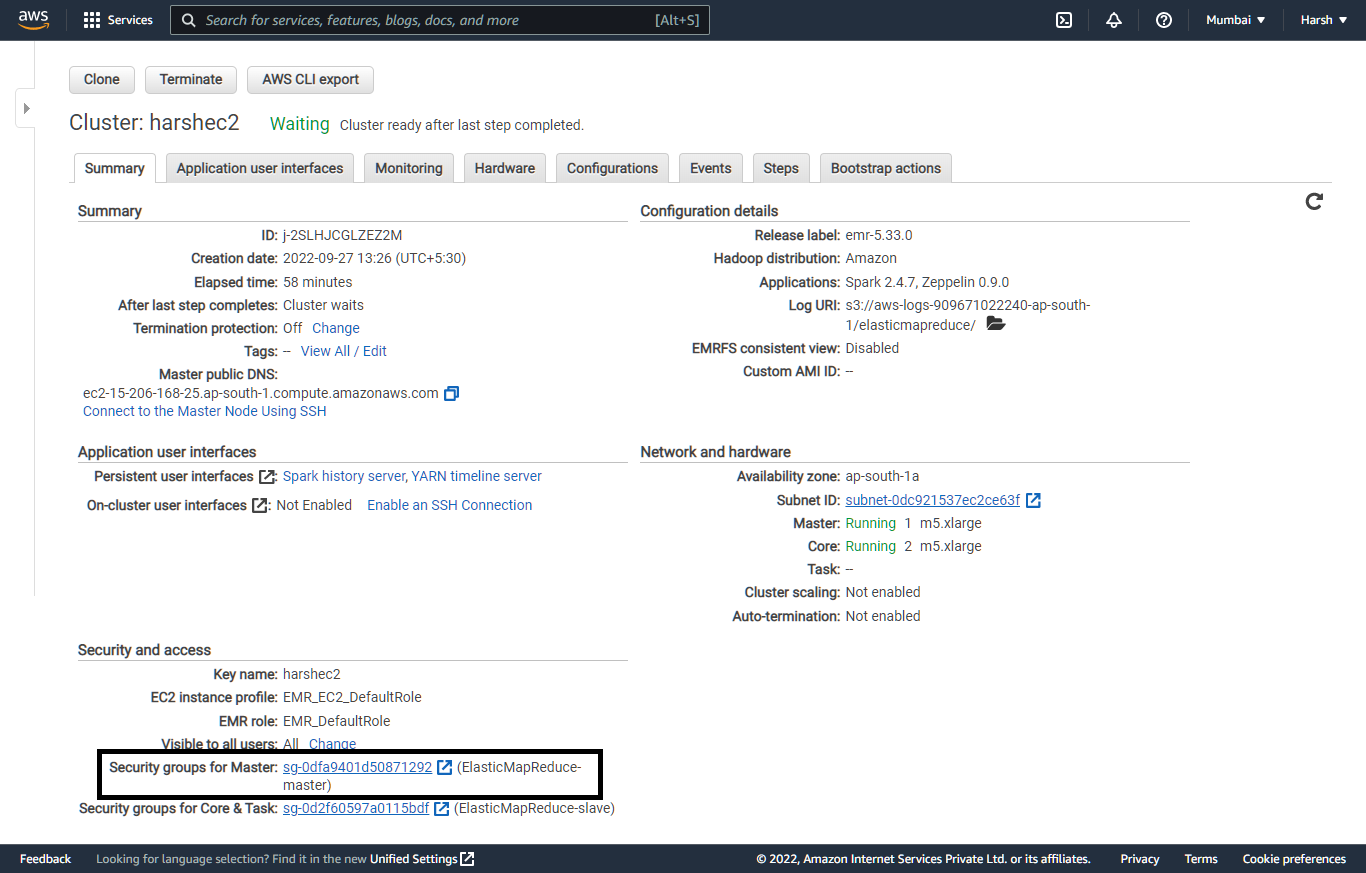
    print("Selected data was successfully saved to S3 %s"% S3\_DATA\_OUTPUT\_PATH)

*if* \_\_name\_\_==   "\_\_main\_\_":

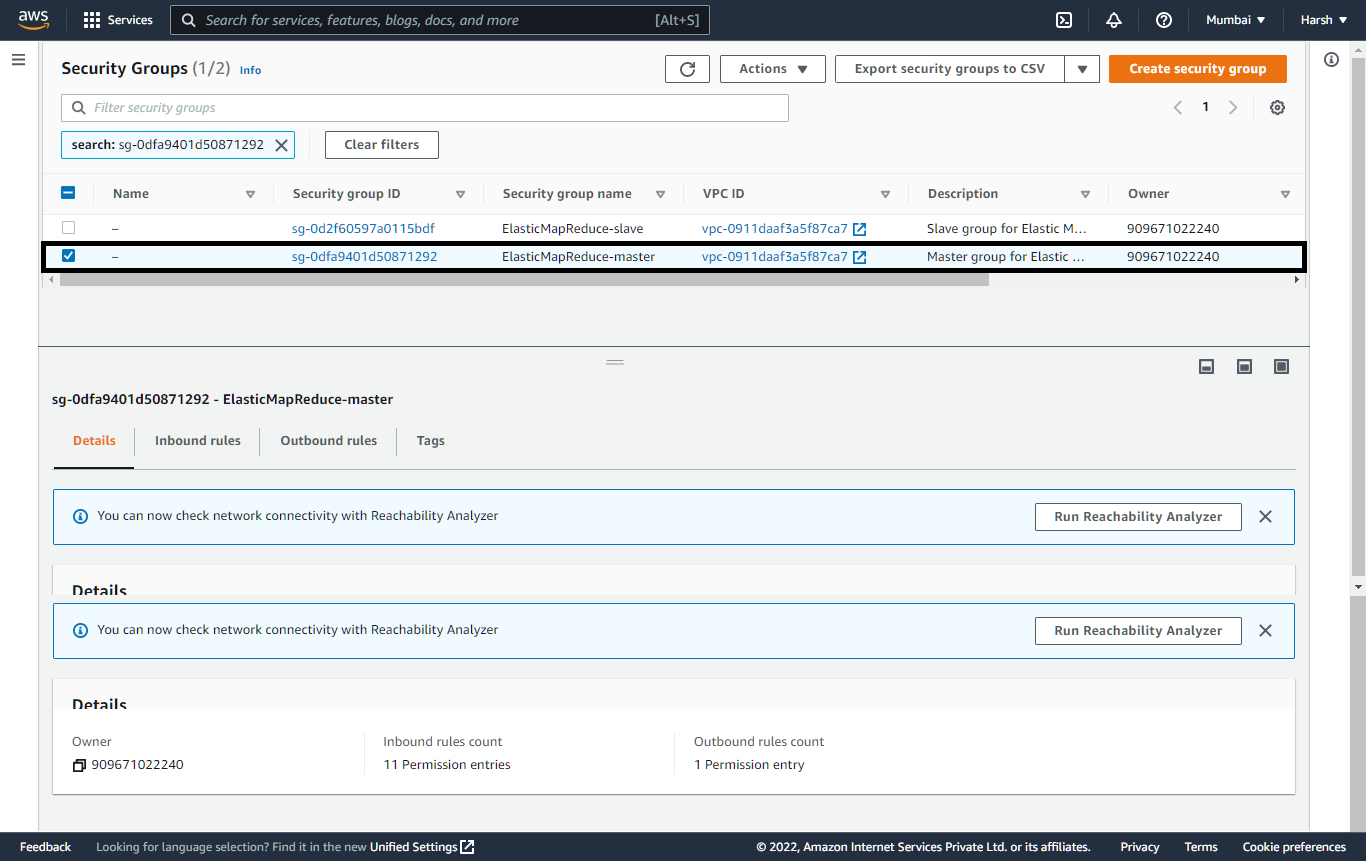
     main()

# Setting up the Security in EMR

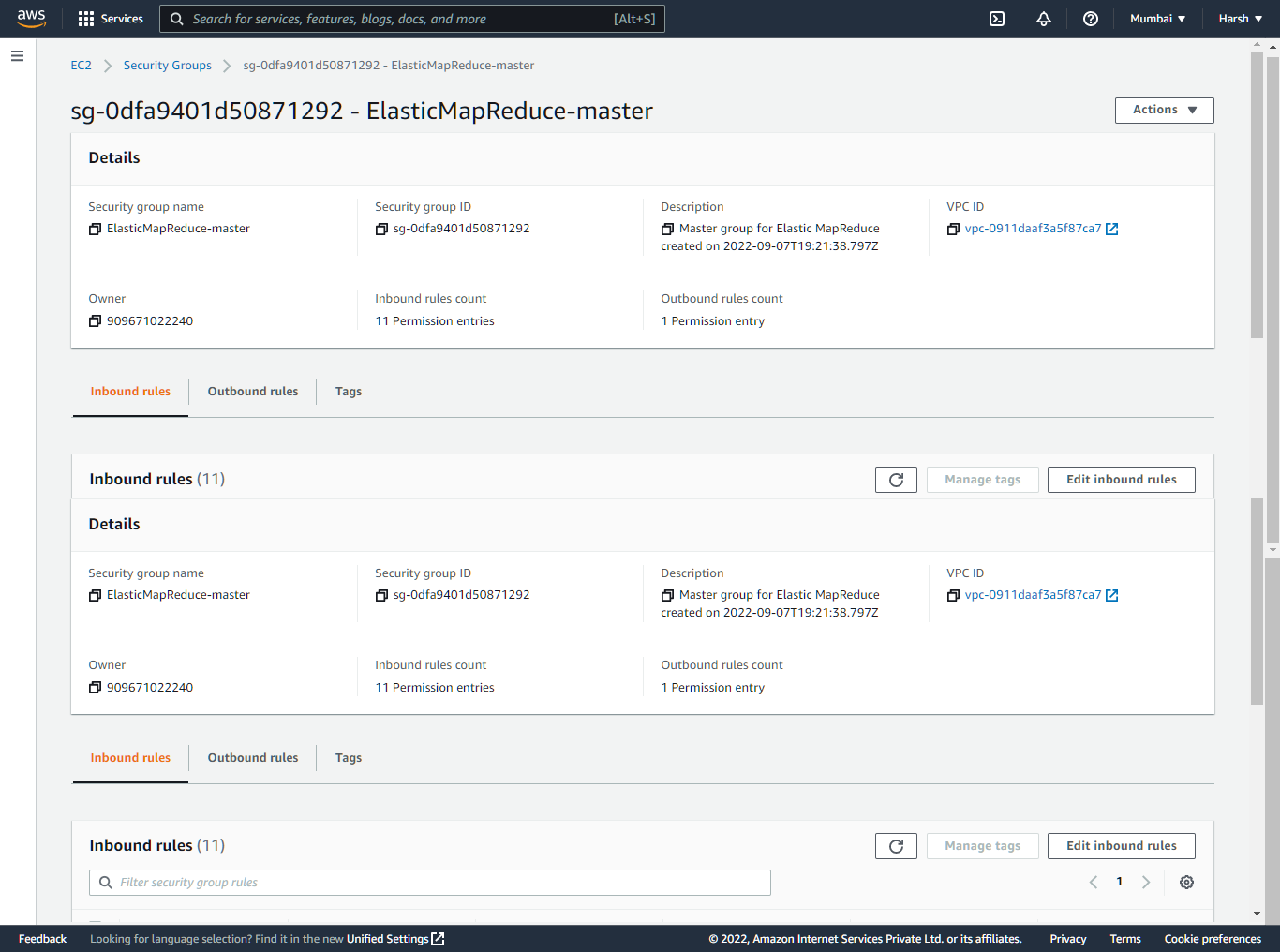
## Step 1: Open EMR and click on the instance that is created. Scroll Down to **Security groups for Master**



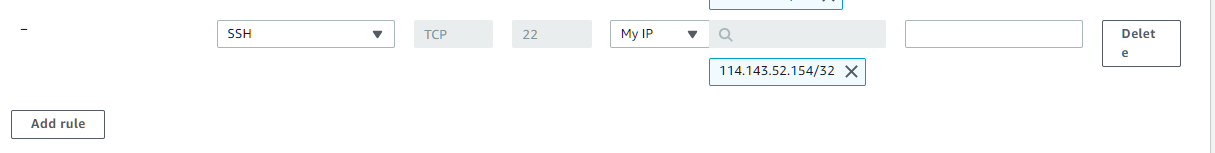
## Step 2: Select the master node security group



## Step 3. Click on Inbound Rules and click on edit inbound rules



## Step 4: Add new Rule for SSH and click on Save Rule.



# Running PYSPARK Cluster

## Step 1: Open EMR and click on the instances that is created and click on **Connect to the Master Node Using SSH**

## Step 2: Download PuTTY.exe to your computer from: <http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>

## Step 3: Start PuTTY.

## Step 4: In the Category list, click Session.

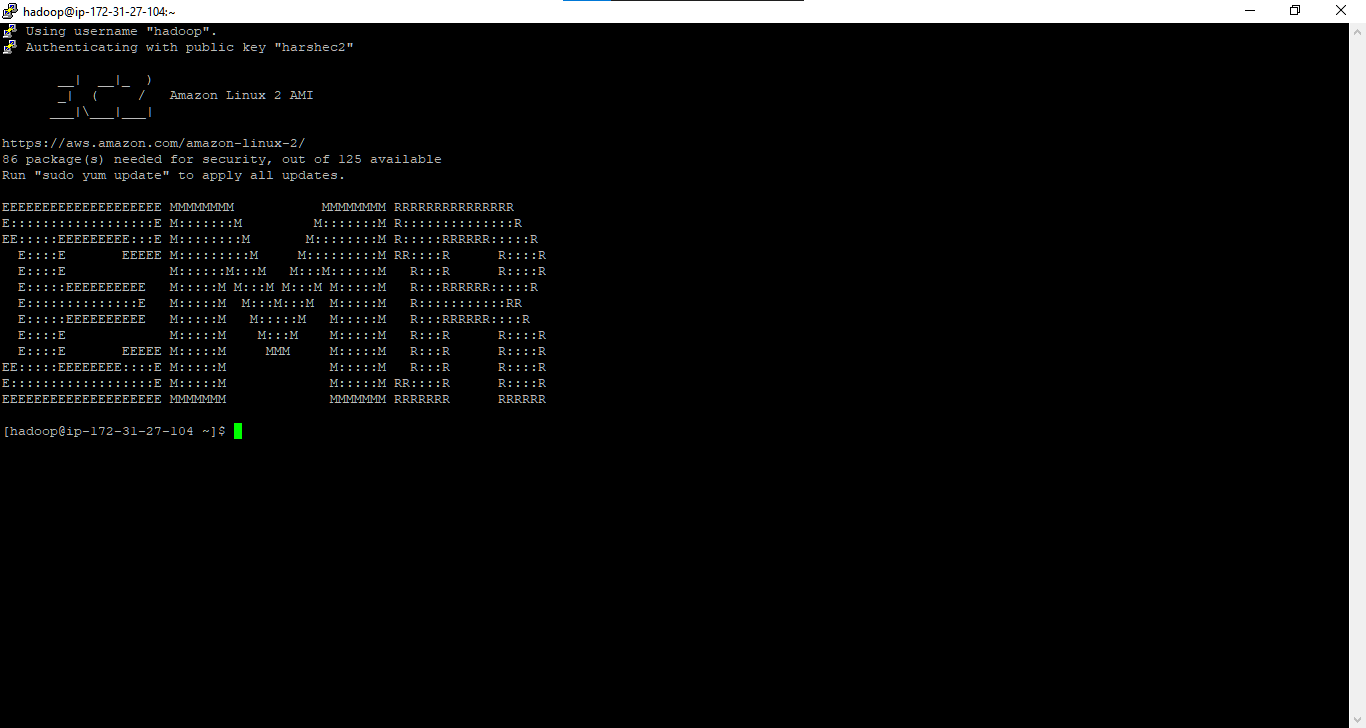
## Step 5: In the Host Name field, type [hadoop@ec2-15-206-168-25.ap-south-1.compute.amazonaws.com](mailto:hadoop@ec2-15-206-168-25.ap-south-1.compute.amazonaws.com)

## Step 6: In the Category list, expand Connection > SSH, and then click Auth.

## Step 7: For Private key file for authentication, click Browse and select the private key file (**harshec2.ppk**) used to launch the cluster.

## Step 8: Click Open.

## Step 9: Click Yes to dismiss the security alert.



## Step 10: open vi main.py

## Step 11: copy the code and press i in the terminal and paste the code

## Step 12: ESC then :wq

## Step 13: spark-submit main.py

# Output

