Name: Harsh Chheda

Roll Number: 31031521005 / 22-15405

Class: Msc. Computer Science

Subject: Cloud Computing

Year: 2022-23

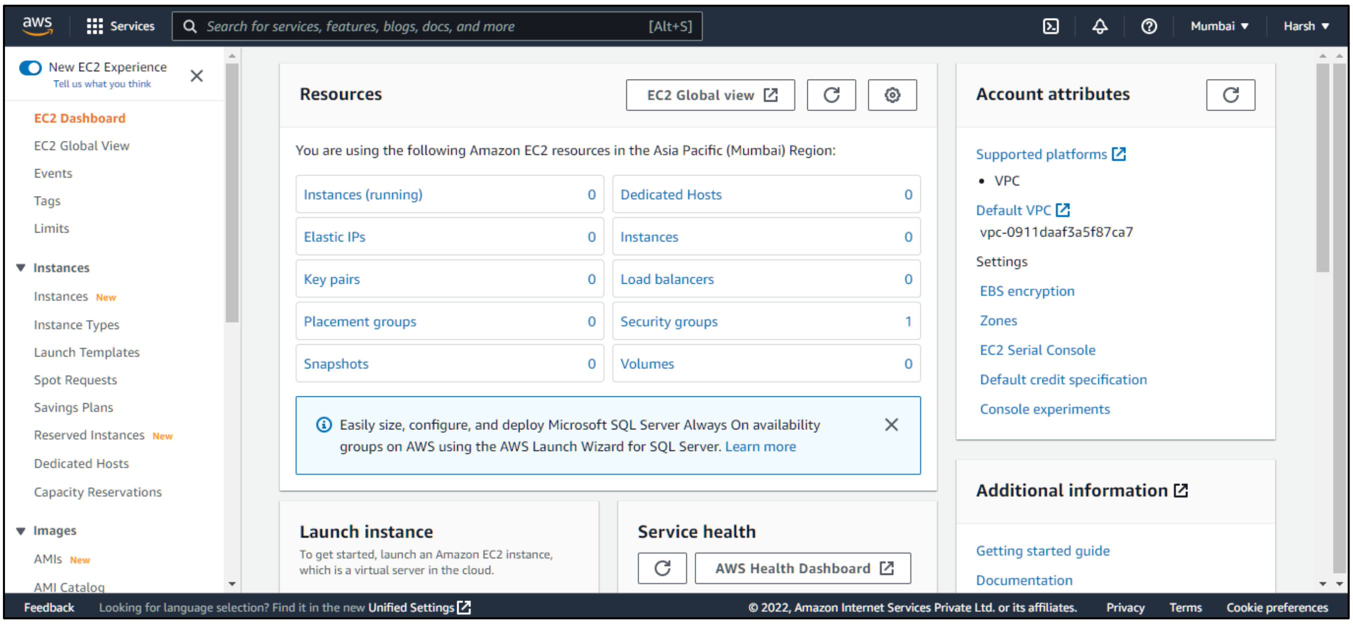
Practical 8

Aim: Demonstration of data analytics in Cloud

Code:

Setting up EC2 Key-Pairs

Step 1: Go to Amazon EC2 Console



Resources

You are using the following Amazon EC2 resources in the Asia Pacific (Mumbai) Region:

Instances (running)	0	Dedicated Hosts	0
Elastic IPs	0	Instances	0
Key pairs	0	Load balancers	0
Placement groups	0	Security groups	1
Snapshots	0	Volumes	0

Easily size, configure, and deploy Microsoft SQL Server Always On availability groups on AWS using the AWS Launch Wizard for SQL Server. Learn more

Account attributes

- VPC

vpc-0911daaf3a5f87ca7

Launch instance

To get started, launch an Amazon EC2 instance, which is a virtual server in the cloud.

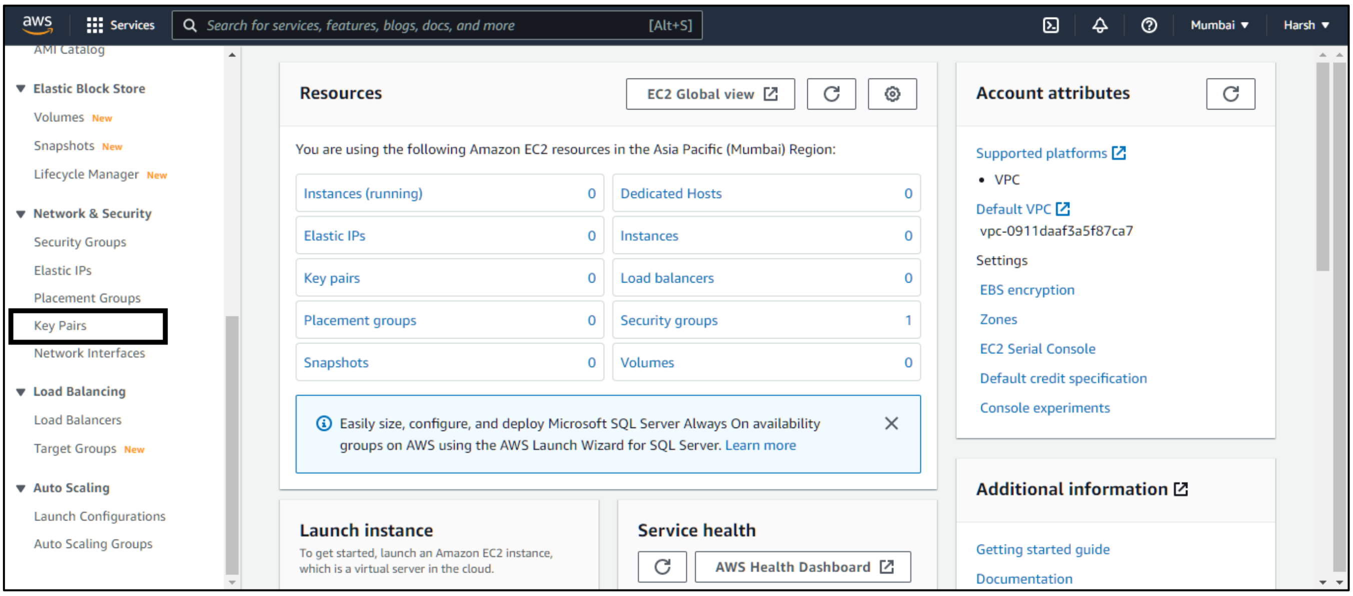
Service health

Additional information

Looking for language selection? Find it in the new Unified Settings

© 2022, Amazon Internet Services Private Ltd. or its affiliates.

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



Resources

You are using the following Amazon EC2 resources in the Asia Pacific (Mumbai) Region:

Instances (running)	0	Dedicated Hosts	0
Elastic IPs	0	Instances	0
Key pairs	0	Load balancers	0
Placement groups	0	Security groups	1
Snapshots	0	Volumes	0

Easily size, configure, and deploy Microsoft SQL Server Always On availability groups on AWS using the AWS Launch Wizard for SQL Server. Learn more

Account attributes

- VPC

vpc-0911daaf3a5f87ca7

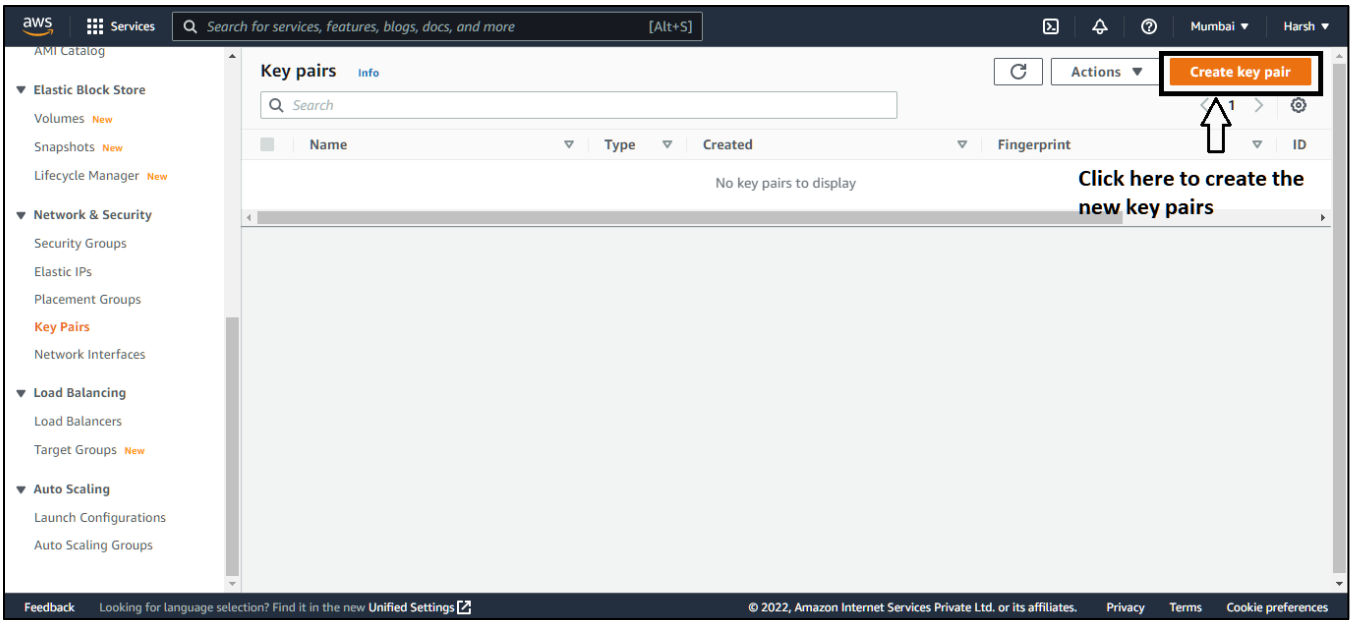
Launch instance

To get started, launch an Amazon EC2 instance, which is a virtual server in the cloud.

Service health

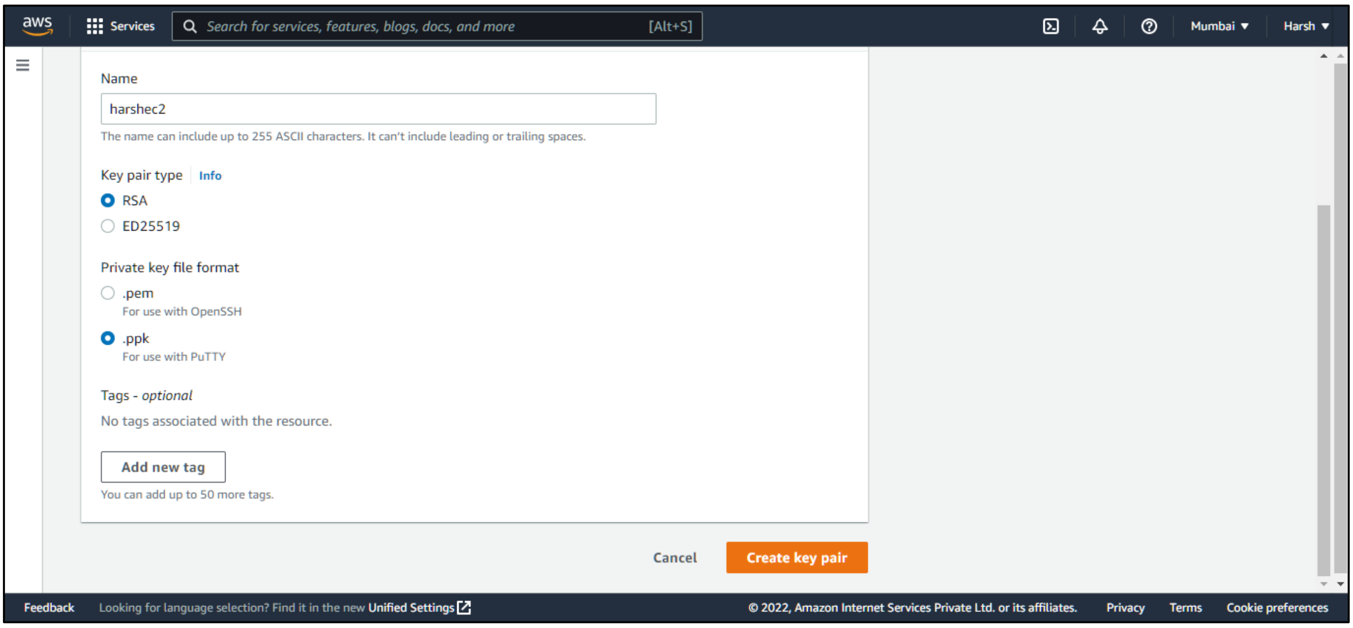
Additional information

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



Click here to create the new key pairs

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


aws
Services
Search for services, features, blogs, docs, and more [Alt+S]
Mumbai
Harsh
Add new tag
Cancel
Create key pair
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

Name

harshec2

The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.

Key pair type Info

● RSA
○ ED25519

Private key file format

○ .pem
For use with OpenSSH
● .ppk
For use with PuTTY

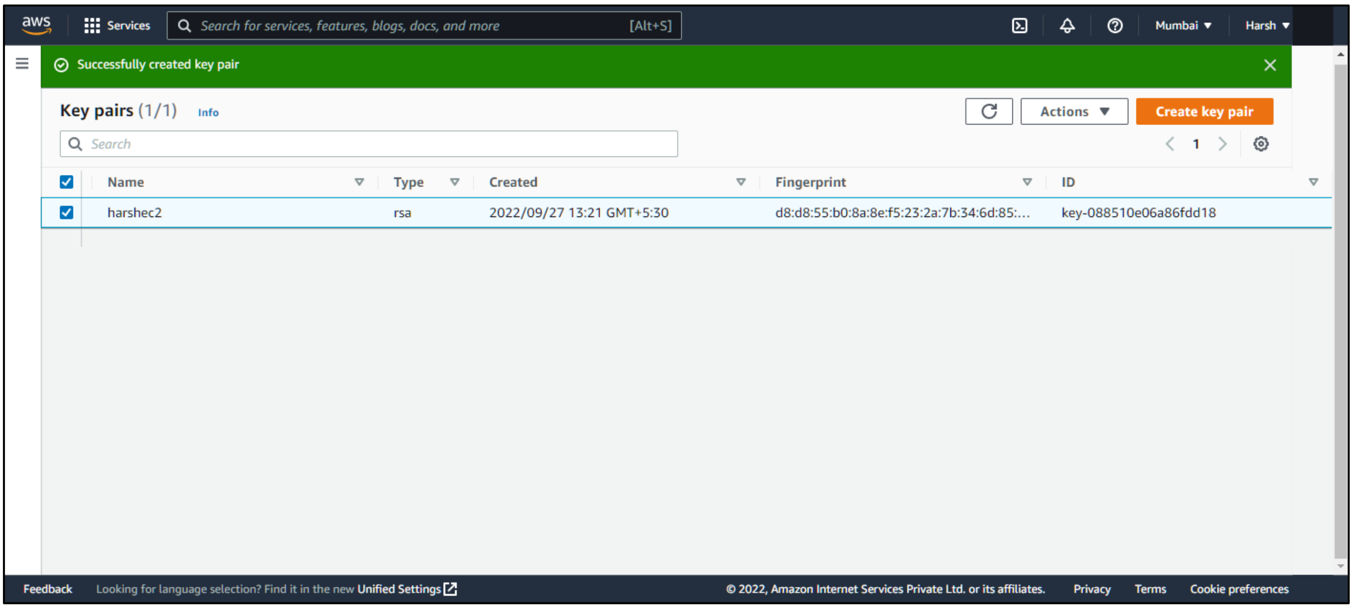
Tags - *optional*

No tags associated with the resource.

You can add up to 50 more tags.

Step 5: Click Create key Pair

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


aws
Services
Search for services, features, blogs, docs, and more [Alt+S]
Mumbai ▼
Harsh ▼

Successfully created key pair

Key pairs (1/1) Info

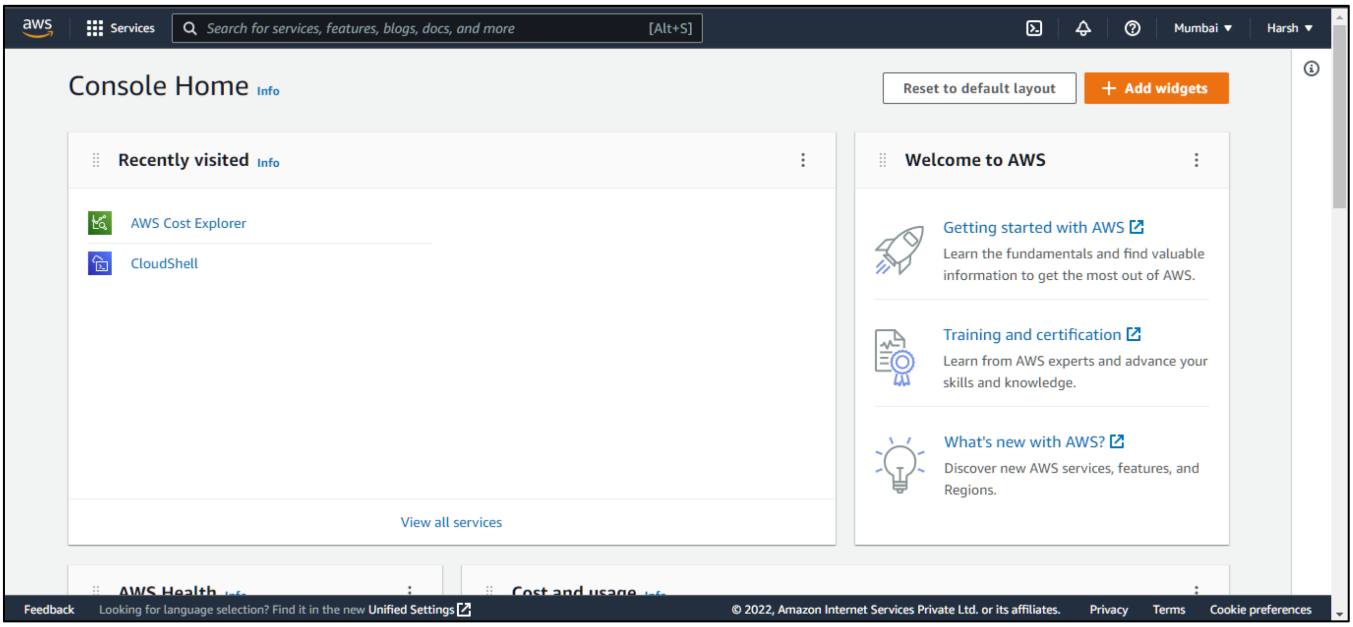
Actions ▼
Create key pair
Search

☑	Name	Type	Created	Fingerprint	ID
☑	harshec2	rsa	2022/09/27 13:21 GMT+5:30	d8:d8:55:b0:8a:8e:f5:23:2a:7b:34:6d:85:...	key-088510e06a86fdd18

Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

Setting up your environment on Amazon EMR

Step 1: Create an AWS account and sign in to the console.



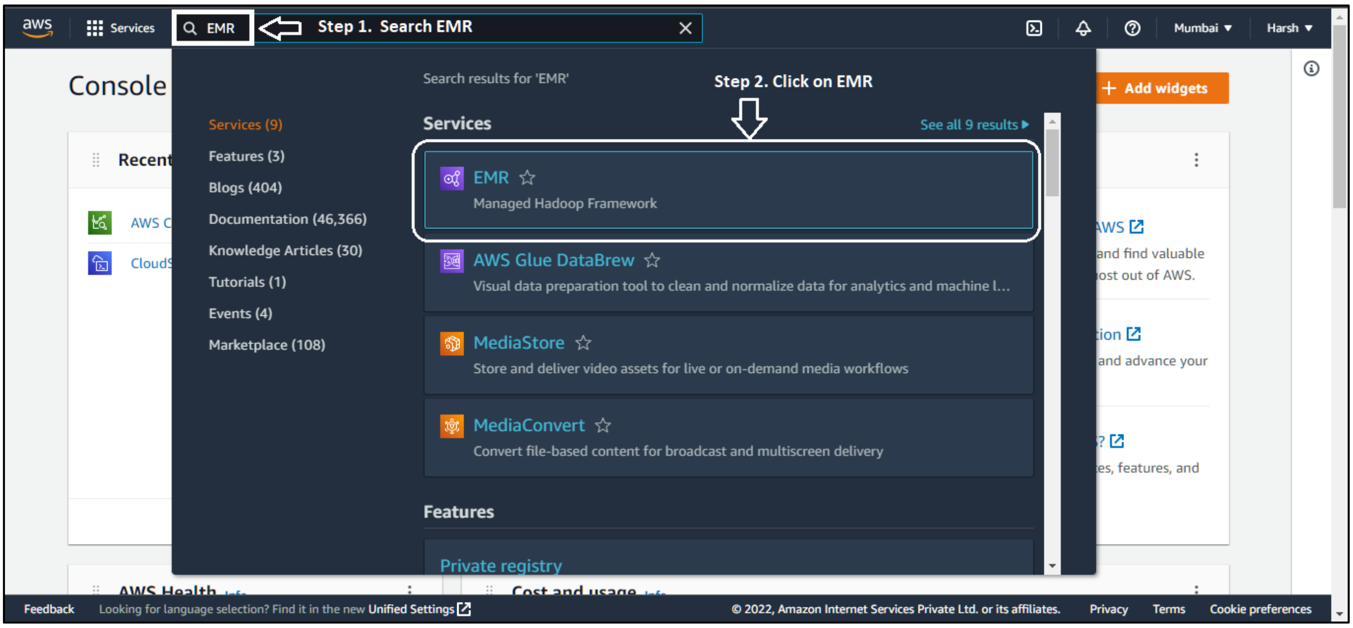
Getting started with AWS
Learn the fundamentals and find valuable information to get the most out of AWS.

Training and certification
Learn from AWS experts and advance your skills and knowledge.

What's new with AWS?
Discover new AWS services, features, and Regions.

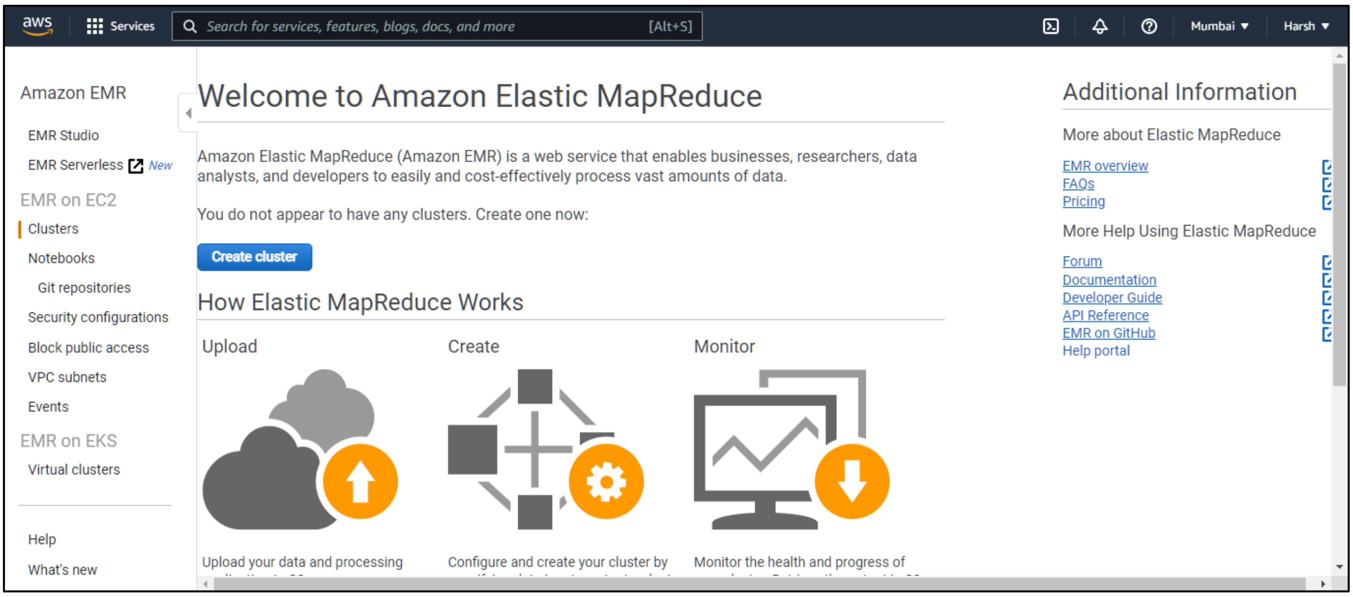
AWS Health
Cost and usage
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

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



Step 1. Search EMR

Step 2. Click on EMR

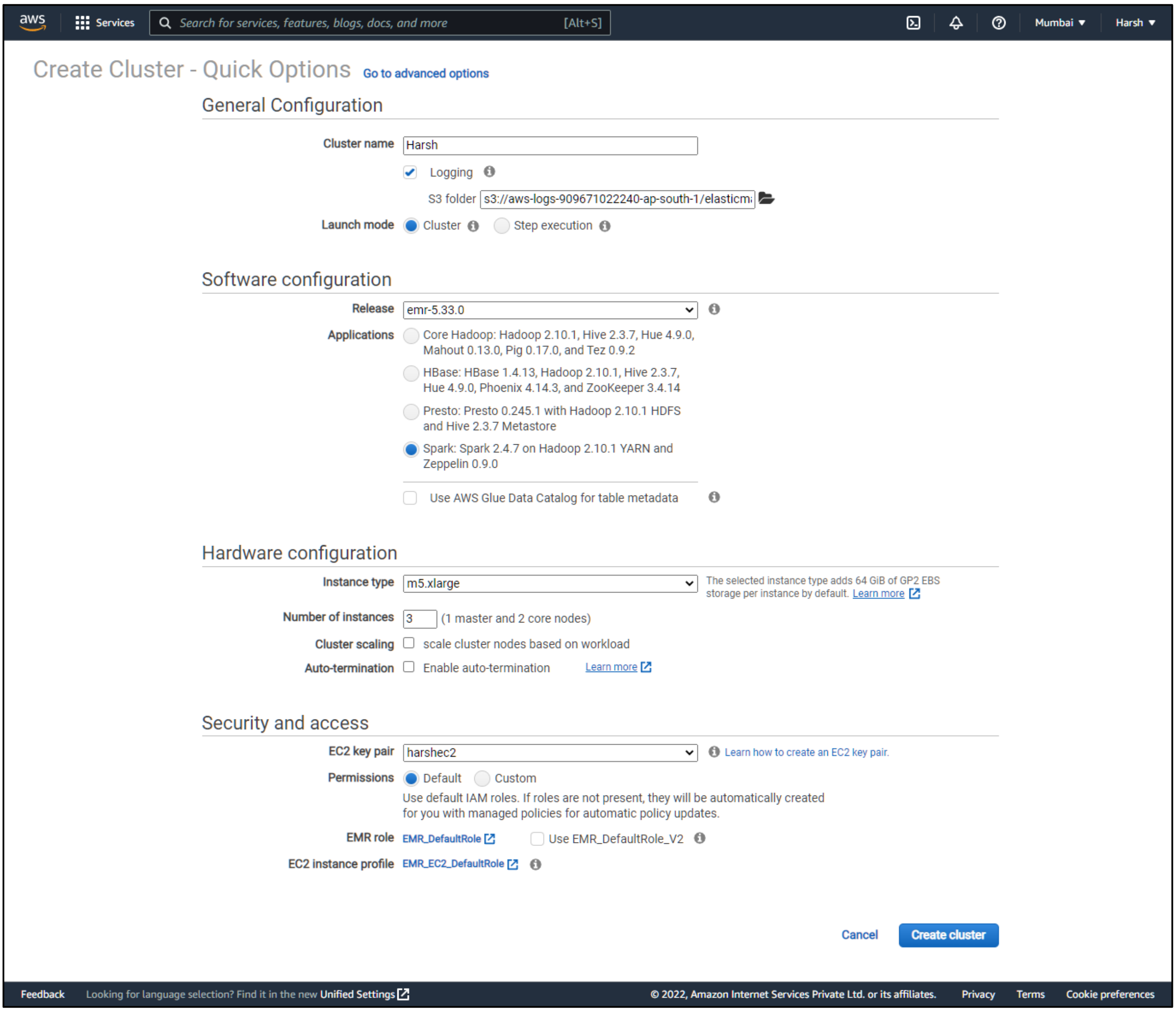


Welcome to Amazon Elastic MapReduce

Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.

You do not appear to have any clusters. Create one now:

Step 3: Creating new cluster


aws
Services
Search for services, features, blogs, docs, and more [Alt+S]
Mumbai ▾
Harsh ▾

Create Cluster - Quick Options Go to advanced options

General Configuration

Cluster name: Harsh

☑ Logging

S3 folder: s3://aws-logs-909671022240-ap-south-1/elasticm

Launch mode: ● Cluster ○ Step execution

Software configuration

Release: emr-5.33.0

Applications:
- ○ Core Hadoop: Hadoop 2.10.1, Hive 2.3.7, Hue 4.9.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
- ○ HBase: HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.7, Hue 4.9.0, Phoenix 4.14.3, and ZooKeeper 3.4.14
- ○ Presto: Presto 0.245.1 with Hadoop 2.10.1 HDFS and Hive 2.3.7 Metastore
- ● Spark: Spark 2.4.7 on Hadoop 2.10.1 YARN and Zeppelin 0.9.0

□ Use AWS Glue Data Catalog for table metadata

Hardware configuration

Instance type: m5.xlarge

The selected instance type adds 64 GiB of GP2 EBS storage per instance by default. Learn more

Number of instances: 3 (1 master and 2 core nodes)

Cluster scaling: □ scale cluster nodes based on workload

Auto-termination: □ Enable auto-termination Learn more

Security and access

EC2 key pair: harshec2

Learn how to create an EC2 key pair.

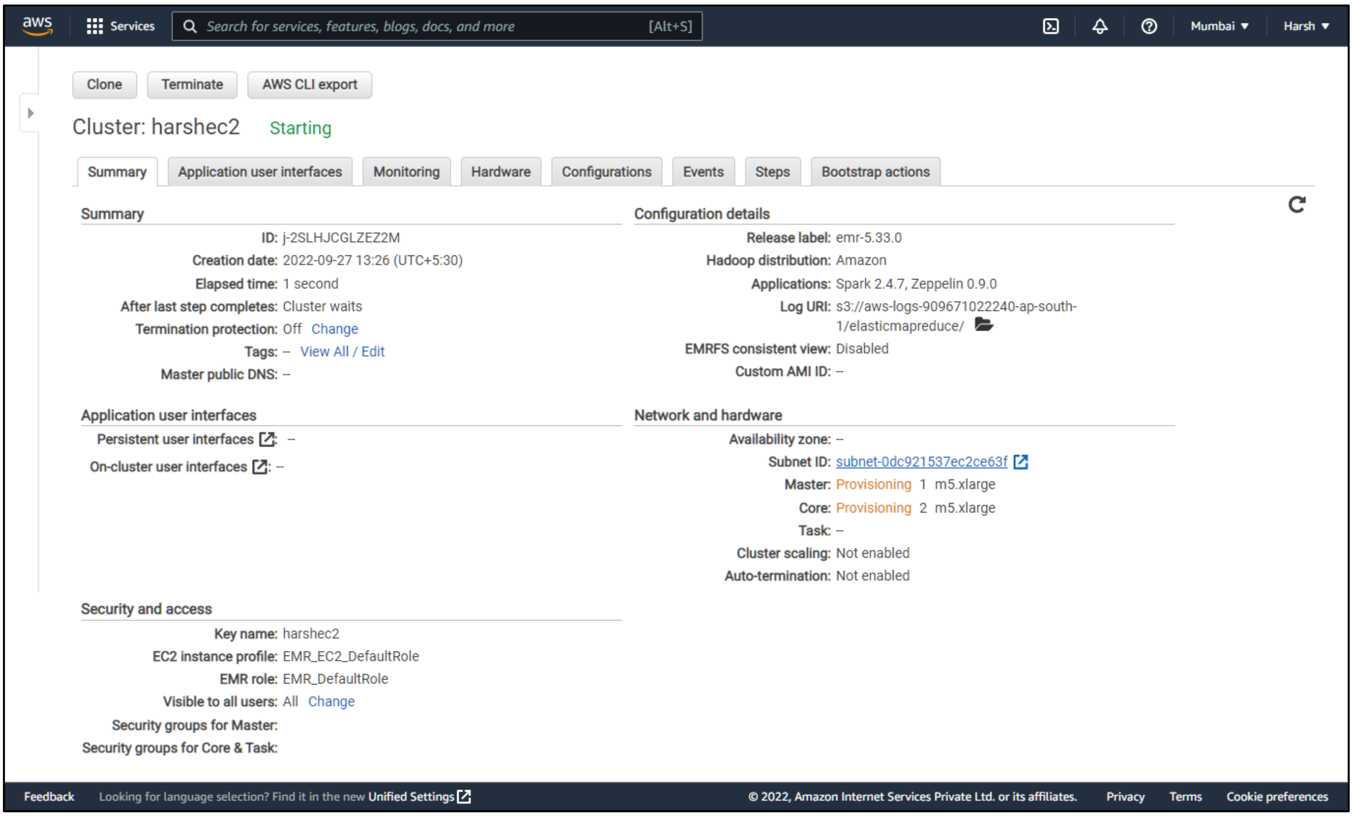
Permissions: ● Default ○ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role: EMR_DefaultRole □ Use EMR_DefaultRole_V2

EC2 instance profile: EMR_EC2_DefaultRole

Cancel
Create cluster
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences


aws
Services
Search for services, features, blogs, docs, and more [Alt+S]
Mumbai ▼
Harsh ▼
Clone
Terminate
AWS CLI export
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

Cluster: harshec2 Starting

Summary

ID: j-2SLHJCGLZEZ2M
Creation date: 2022-09-27 13:26 (UTC+5:30)
Elapsed time: 1 second
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: --

Configuration details

Release label: emr-5.33.0
Hadoop distribution: Amazon
Applications: Spark 2.4.7, Zeppelin 0.9.0
Log URI: s3://aws-logs-909671022240-ap-south-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

Persistent user interfaces: --
On-cluster user interfaces: --

Network and hardware

Availability zone: --
Subnet ID: subnet-0dc921537ec2ce63f
Master: Provisioning 1 m5.xlarge
Core: Provisioning 2 m5.xlarge
Task: --
Cluster scaling: Not enabled
Auto-termination: Not enabled

Security and access

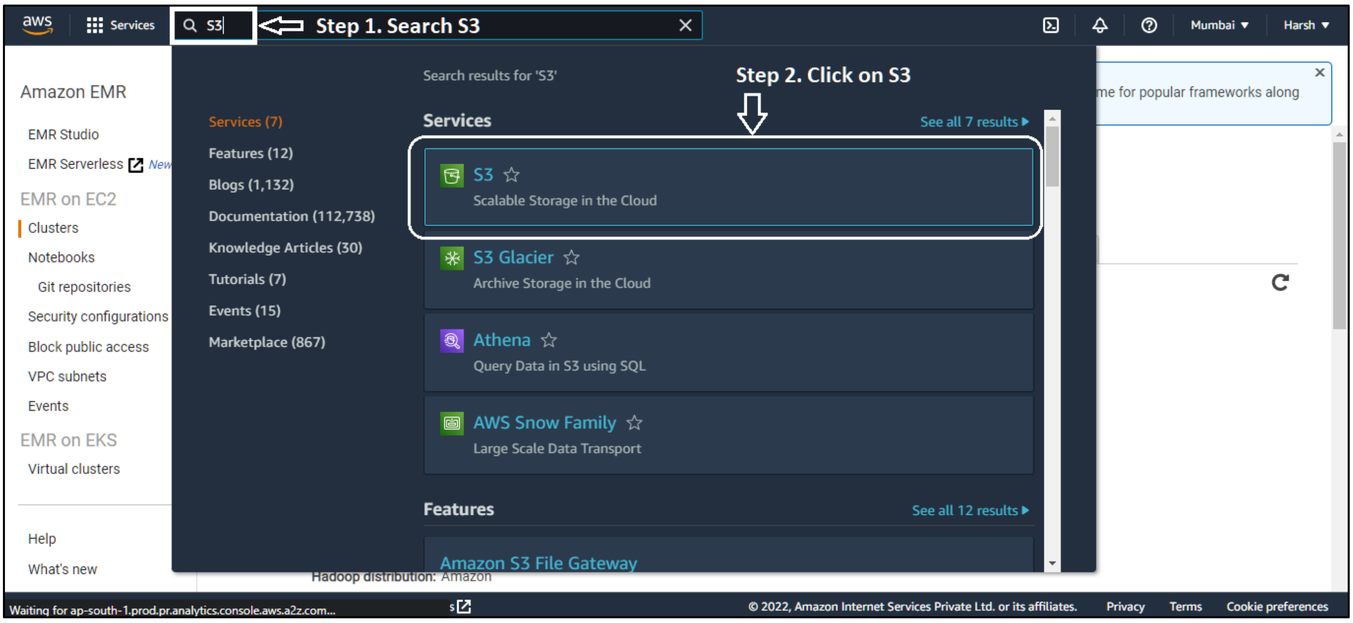
Key name: harshec2
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master:
Security groups for Core & Task:

Downloading Dataset

Step 1: Click Here to download dataset

Setting up S3 Environment

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



S3

Step 1. Search S3

Search results for 'S3'

Step 2. Click on S3

Services

S3 ☆
Scalable Storage in the Cloud

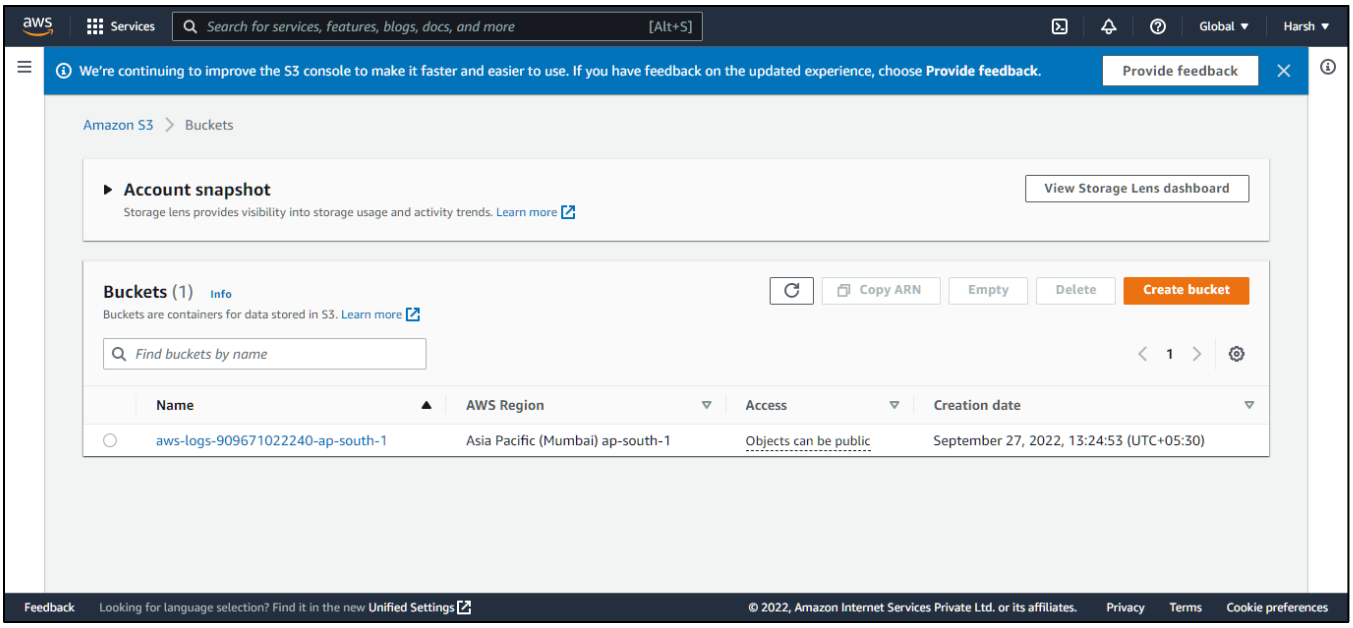
S3 Glacier ☆
Archive Storage in the Cloud

Athena ☆
Query Data in S3 using SQL

AWS Snow Family ☆
Large Scale Data Transport

Features

Amazon S3 File Gateway



We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

Provide feedback

Amazon S3 > Buckets

Account snapshot

Storage lens provides visibility into storage usage and activity trends. Learn more

View Storage Lens dashboard

Buckets (1) Info

Buckets are containers for data stored in S3. Learn more

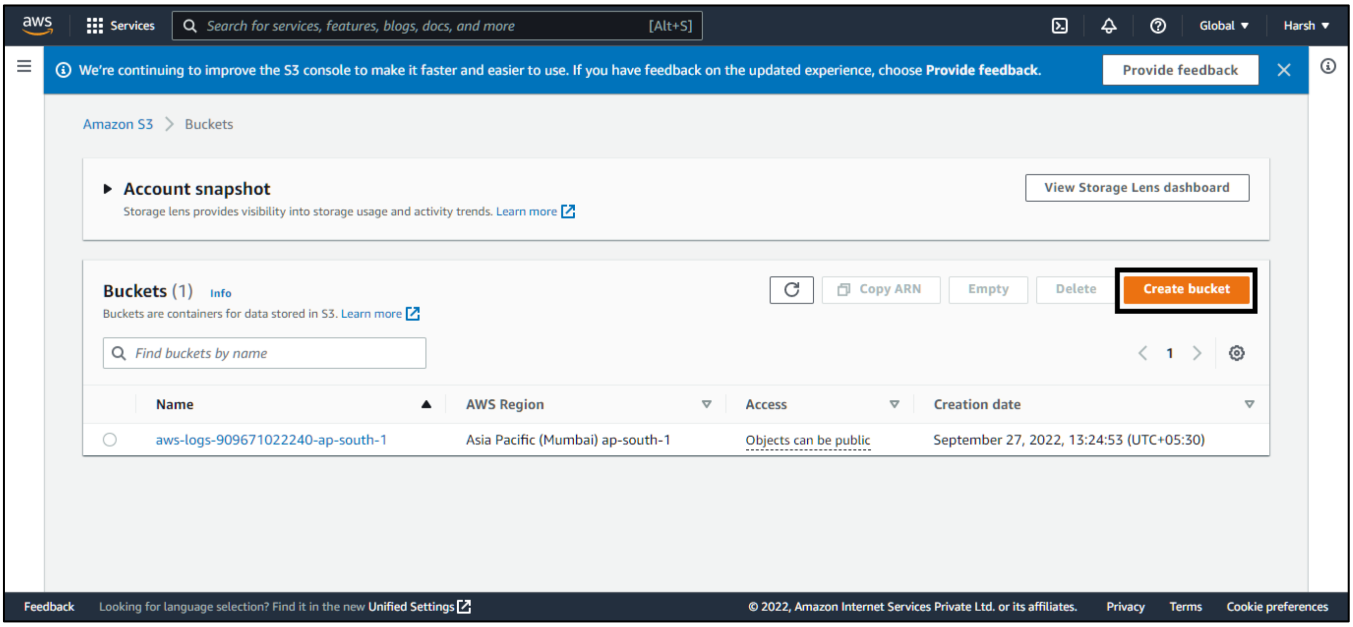
Copy ARN | Empty | Delete | Create bucket

Find buckets by name

	Name	AWS Region	Access	Creation date
○	aws-logs-909671022240-ap-south-1	Asia Pacific (Mumbai) ap-south-1	Objects can be public	September 27, 2022, 13:24:53 (UTC+05:30)

Feedback Looking for language selection? Find it in the new **Unified Settings** © 2022, Amazon Internet Services Private Ltd. or its affiliates. Privacy Terms Cookie preferences

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



aws Services Search for services, features, blogs, docs, and more [Alt+S] Global Harsh

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

Provide feedback

Amazon S3 > Buckets

Account snapshot

Storage lens provides visibility into storage usage and activity trends. Learn more

View Storage Lens dashboard

Buckets (1) Info

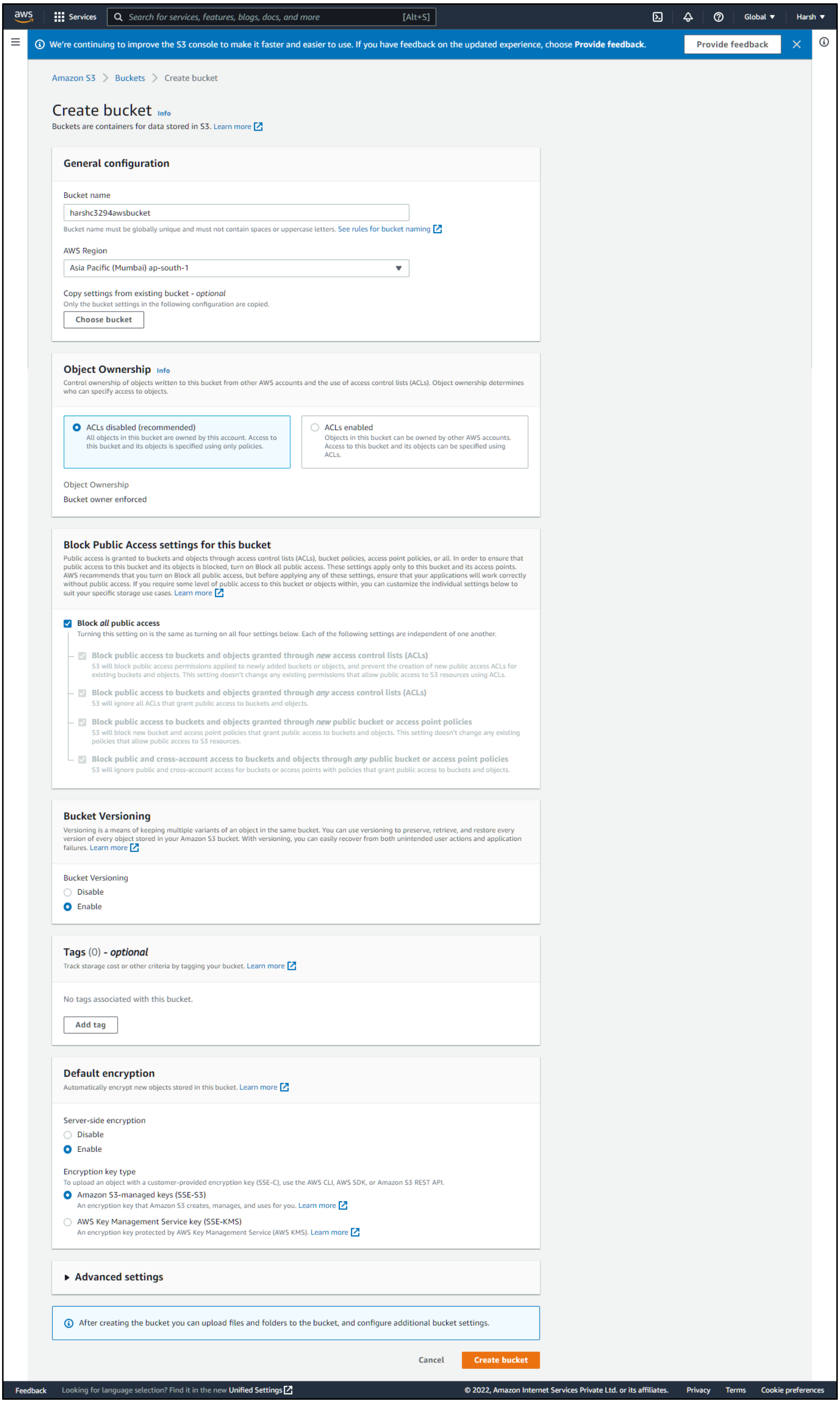
Buckets are containers for data stored in S3. Learn more

Copy ARN | Empty | Delete | Create bucket

Find buckets by name

	Name	AWS Region	Access	Creation date
○	aws-logs-909671022240-ap-south-1	Asia Pacific (Mumbai) ap-south-1	Objects can be public	September 27, 2022, 13:24:53 (UTC+05:30)

Feedback Looking for language selection? Find it in the new **Unified Settings** © 2022, Amazon Internet Services Private Ltd. or its affiliates. Privacy Terms Cookie preferences



We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

Create bucket

Buckets are containers for data stored in S3. Learn more

General configuration

Bucket name

harshc3294awsbucket

Bucket name must be globally unique and must not contain spaces or uppercase letters. See rules for bucket naming

AWS Region

Asia Pacific (Mumbai) ap-south-1

Copy settings from existing bucket - *optional*

Only the bucket settings in the following configuration are copied.

Object Ownership

Control ownership of objects written to this bucket from other AWS accounts and the use of access control lists (ACLs). Object ownership determines who can specify access to objects.

● ACLs disabled (recommended)
All objects in this bucket are owned by this account. Access to this bucket and its objects is specified using only policies.

○ ACLs enabled
Objects in this bucket can be owned by other AWS accounts. Access to this bucket and its objects can be specified using ACLs.

Object Ownership

Bucket owner enforced

Block Public Access settings for this bucket

Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure that public access to this bucket and its objects is blocked, turn on Block all public access. These settings apply only to this bucket and its access points. AWS recommends that you turn on Block all public access, but before applying any of these settings, ensure that your applications will work correctly without public access. If you require some level of public access to this bucket or objects within, you can customize the individual settings below to suit your specific storage use cases. Learn more

☑ **Block *all* public access**
Turning this setting on is the same as turning on all four settings below. Each of the following settings are independent of one another.

- ☑ **Block public access to buckets and objects granted through *new* access control lists (ACLs)**
  S3 will block public access permissions applied to newly added buckets or objects, and prevent the creation of new public access ACLs for existing buckets and objects. This setting doesn't change any existing permissions that allow public access to S3 resources using ACLs.
- ☑ **Block public access to buckets and objects granted through *any* access control lists (ACLs)**
  S3 will ignore all ACLs that grant public access to buckets and objects.
- ☑ **Block public access to buckets and objects granted through *new* public bucket or access point policies**
  S3 will block new bucket and access point policies that grant public access to buckets and objects. This setting doesn't change any existing policies that allow public access to S3 resources.
- ☑ **Block public and cross-account access to buckets and objects through *any* public bucket or access point policies**
  S3 will ignore public and cross-account access for buckets or access points with policies that grant public access to buckets and objects.

Bucket Versioning

Versioning is a means of keeping multiple variants of an object in the same bucket. You can use versioning to preserve, retrieve, and restore every version of every object stored in your Amazon S3 bucket. With versioning, you can easily recover from both unintended user actions and application failures. Learn more

Bucket Versioning

○ Disable
● Enable

Tags (0) - *optional*

Track storage cost or other criteria by tagging your bucket. Learn more

No tags associated with this bucket.

Default encryption

Automatically encrypt new objects stored in this bucket. Learn more

Server-side encryption

○ Disable
● Enable

Encryption key type

To upload an object with a customer-provided encryption key (SSE-C), use the AWS CLI, AWS SDK, or Amazon S3 REST API.

● Amazon S3-managed keys (SSE-S3)
An encryption key that Amazon S3 creates, manages, and uses for you. Learn more

○ AWS Key Management Service key (SSE-KMS)
An encryption key protected by AWS Key Management Service (AWS KMS). Learn more

▸ **Advanced settings**

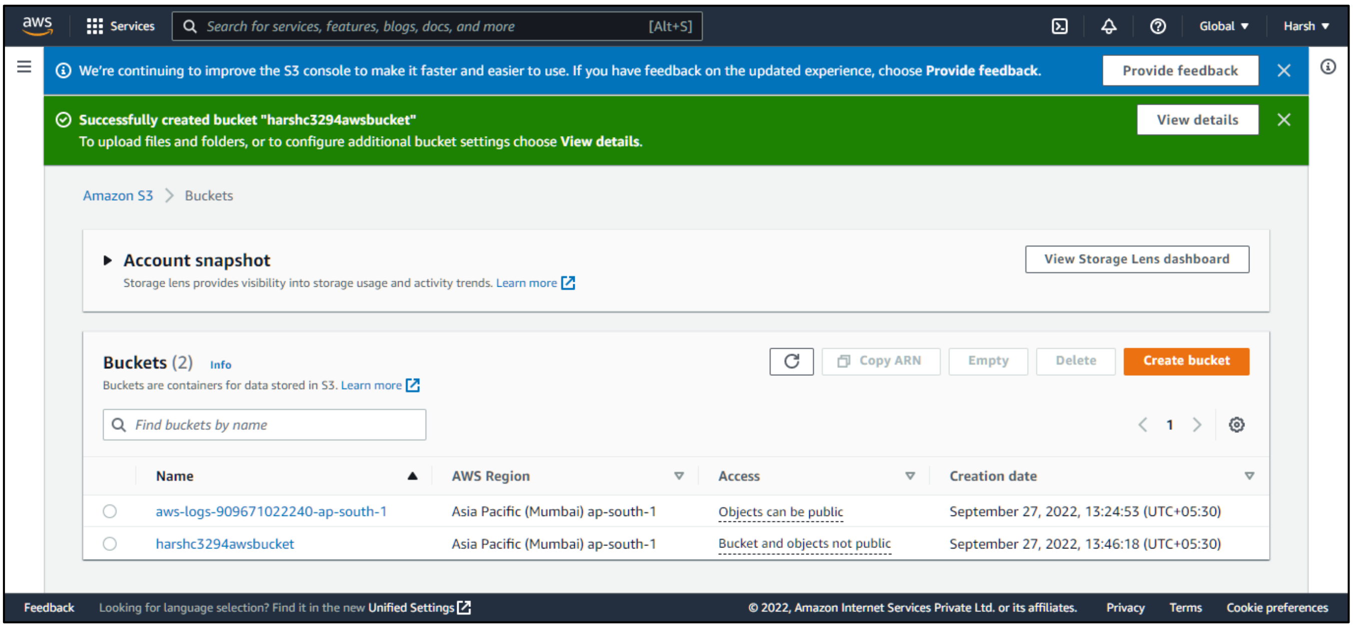
After creating the bucket you can upload files and folders to the bucket, and configure additional bucket settings.

Looking for language selection? Find it in the new **Unified Settings**

© 2022, Amazon Internet Services Private Ltd. or its affiliates.

Step 3: Click on Create Bucket

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



We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

Successfully created bucket "harshc3294awsbucket"
To upload files and folders, or to configure additional bucket settings choose **View details**.

Amazon S3 > Buckets

Account snapshot
Storage lens provides visibility into storage usage and activity trends. Learn more

Buckets (2) Info
Buckets are containers for data stored in S3. Learn more

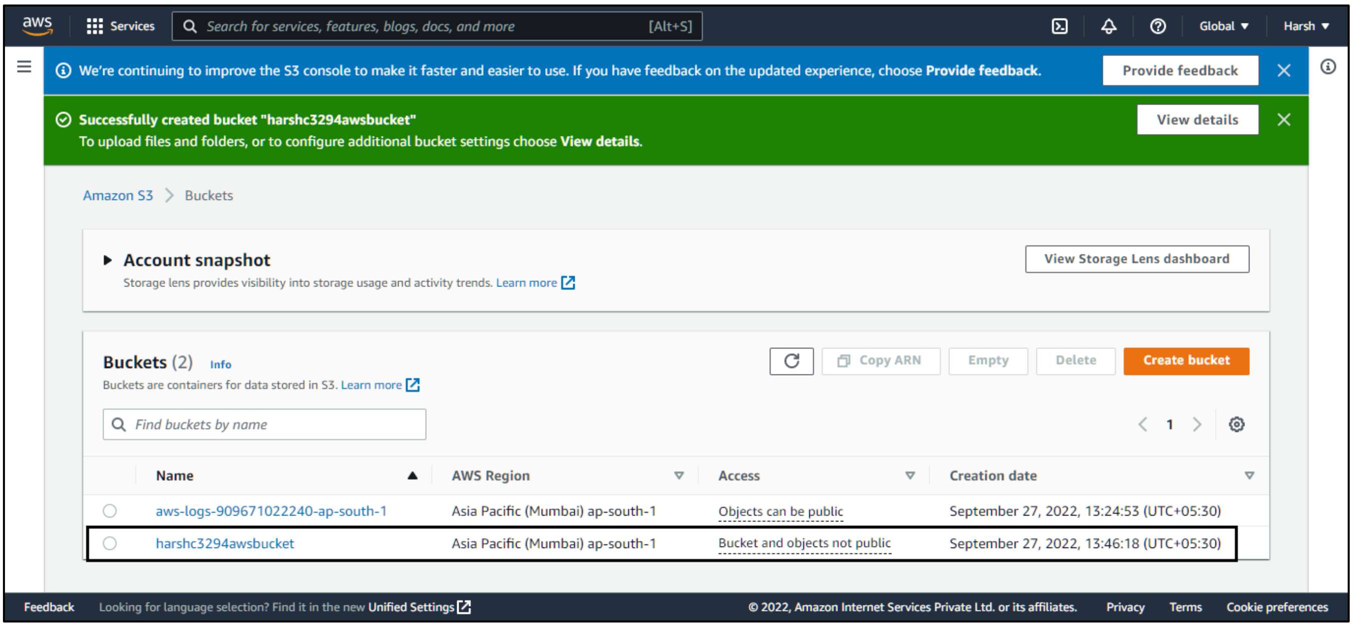
	Name	AWS Region	Access	Creation date
○	aws-logs-909671022240-ap-south-1	Asia Pacific (Mumbai) ap-south-1	Objects can be public	September 27, 2022, 13:24:53 (UTC+05:30)
○	harshc3294awsbucket	Asia Pacific (Mumbai) ap-south-1	Bucket and objects not public	September 27, 2022, 13:46:18 (UTC+05:30)

Looking for language selection? Find it in the new **Unified Settings**

© 2022, Amazon Internet Services Private Ltd. or its affiliates.

Uploading Dataset to the S3 Bucket

Step 1: Click and open the S3 bucket



We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

Successfully created bucket "harshc3294awsbucket"
To upload files and folders, or to configure additional bucket settings choose **View details**.

Amazon S3 > Buckets

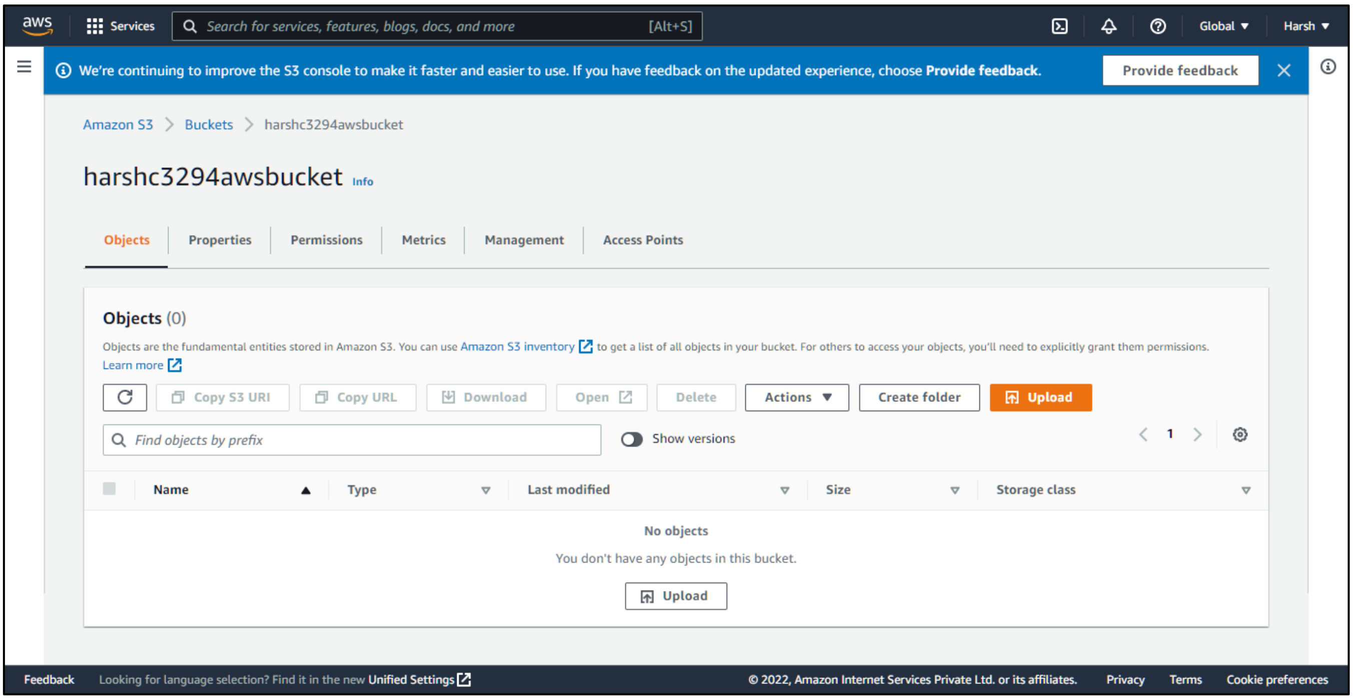
Account snapshot
Storage lens provides visibility into storage usage and activity trends. Learn more

Buckets (2) Info
Buckets are containers for data stored in S3. Learn more

	Name	AWS Region	Access	Creation date
○	aws-logs-909671022240-ap-south-1	Asia Pacific (Mumbai) ap-south-1	Objects can be public	September 27, 2022, 13:24:53 (UTC+05:30)
○	harshc3294awsbucket	Asia Pacific (Mumbai) ap-south-1	Bucket and objects not public	September 27, 2022, 13:46:18 (UTC+05:30)

Looking for language selection? Find it in the new **Unified Settings**

© 2022, Amazon Internet Services Private Ltd. or its affiliates.



aws
Services
Search for services, features, blogs, docs, and more [Alt+S]
Global ▼
Harsh ▼

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback.**

Provide feedback
Amazon S3 > Buckets > harshc3294awsbucket

harshc3294awsbucket Info

Objects
Properties
Permissions
Metrics
Management
Access Points

Objects (0)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

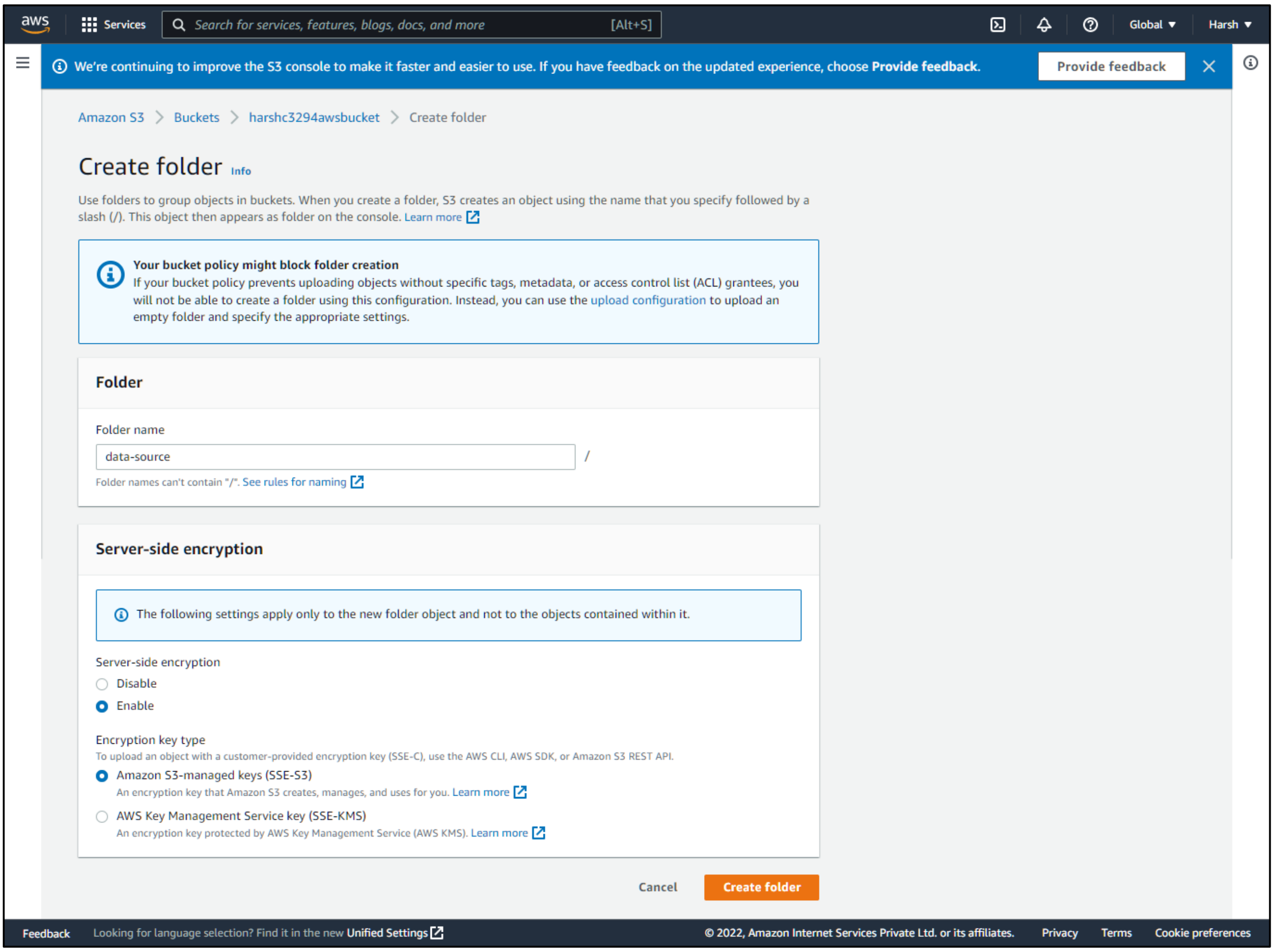
Name ▲	Type	Last modified	Size	Storage class

No objects

You don't have any objects in this bucket.

Upload
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

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



aws
Services
Search for services, features, blogs, docs, and more [Alt+S]
Global ▼
Harsh ▼

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback.**

Provide feedback
Amazon S3 > Buckets > harshc3294awsbucket > Create folder

Create folder Info

Use folders to group objects in buckets. When you create a folder, S3 creates an object using the name that you specify followed by a slash (/). This object then appears as folder on the console. Learn more

Your bucket policy might block folder creation

If your bucket policy prevents uploading objects without specific tags, metadata, or access control list (ACL) grantees, you will not be able to create a folder using this configuration. Instead, you can use the upload configuration to upload an empty folder and specify the appropriate settings.

Folder

Folder name

data-source /

Folder names can't contain "/". See rules for naming

Server-side encryption

The following settings apply only to the new folder object and not to the objects contained within it.

Server-side encryption

○ Disable

● Enable

Encryption key type

To upload an object with a customer-provided encryption key (SSE-C), use the AWS CLI, AWS SDK, or Amazon S3 REST API.

● Amazon S3-managed keys (SSE-S3)

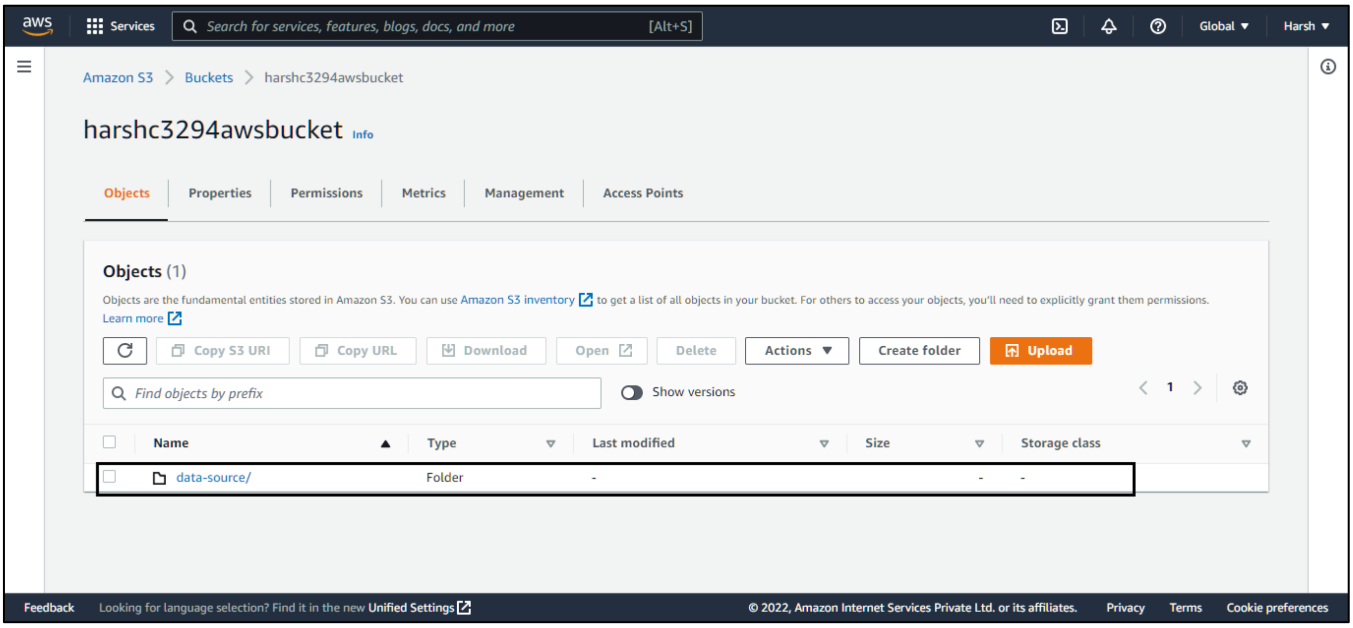
An encryption key that Amazon S3 creates, manages, and uses for you. Learn more

○ AWS Key Management Service key (SSE-KMS)

An encryption key protected by AWS Key Management Service (AWS KMS). Learn more

Cancel
Create folder
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

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

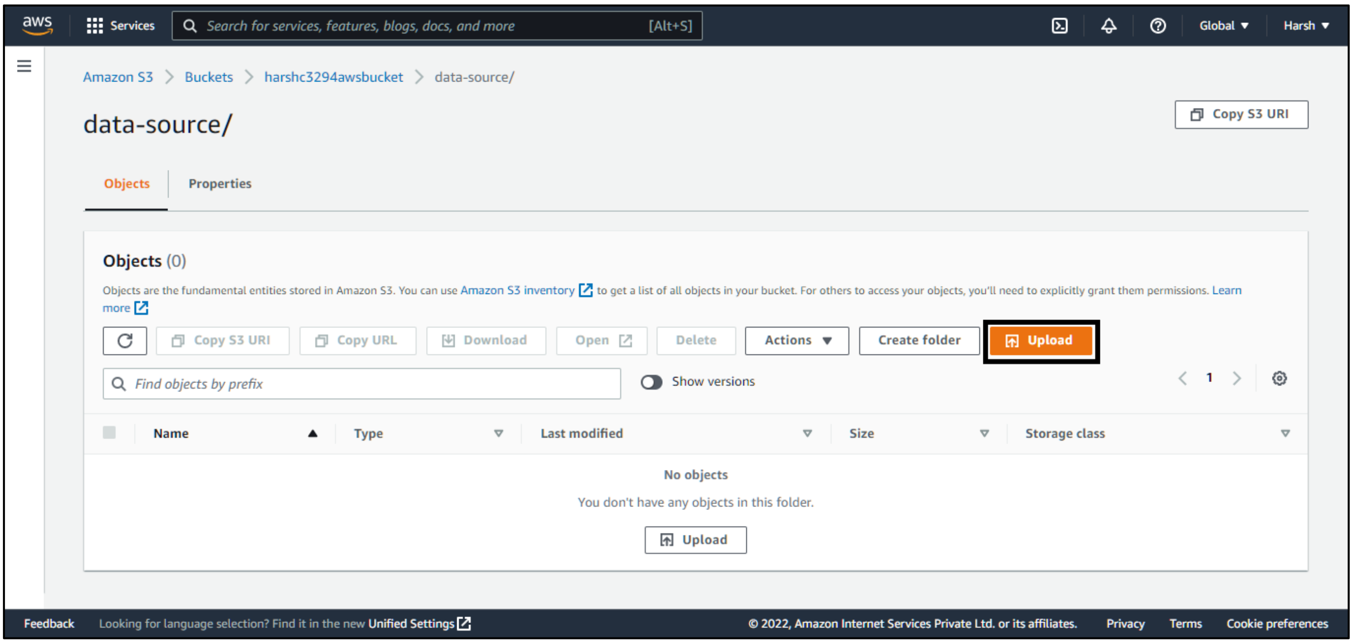


Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	data-source/	Folder	-	-	-

Looking for language selection? Find it in the new Unified Settings

© 2022, Amazon Internet Services Private Ltd. or its affiliates.



Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

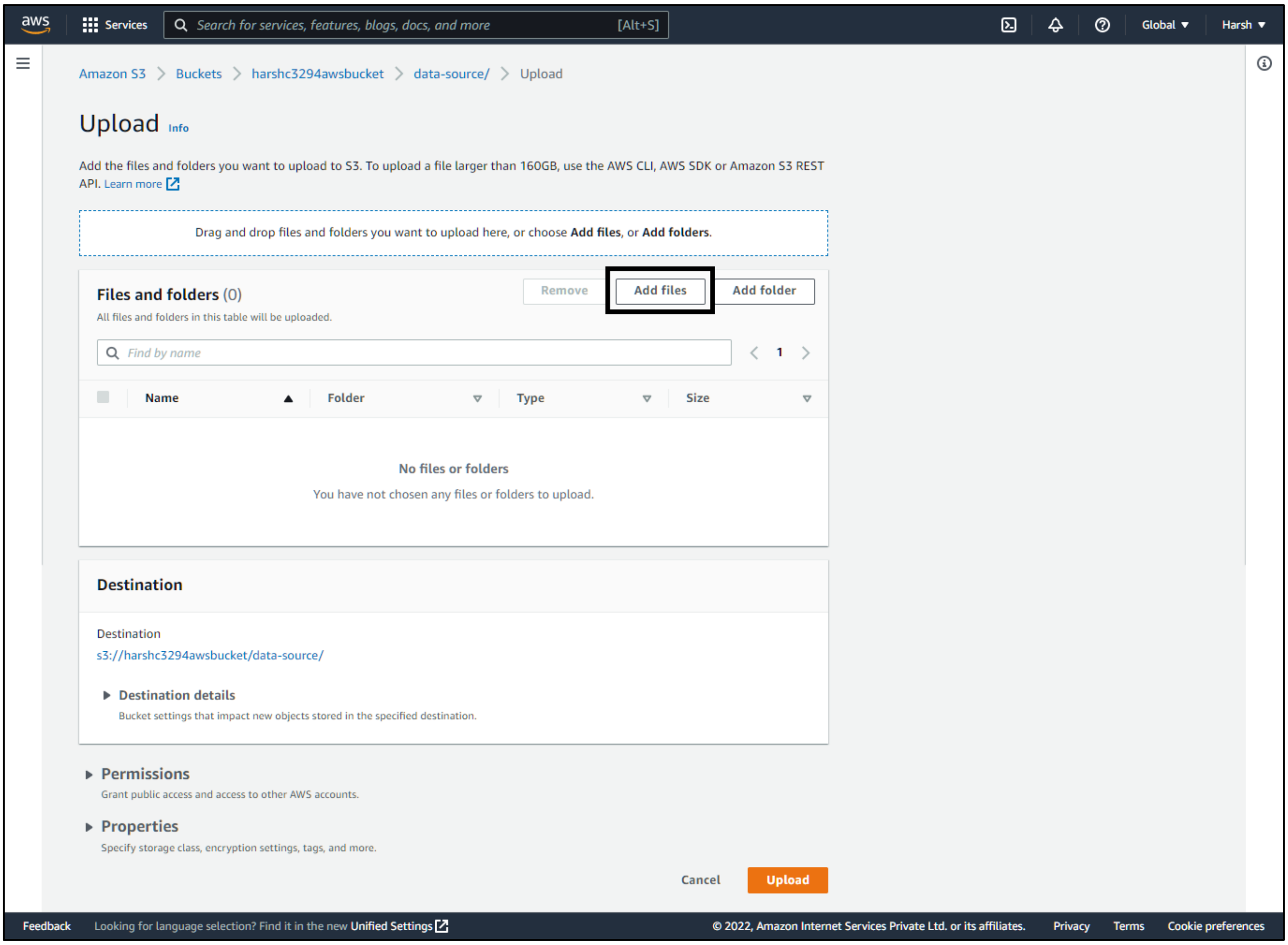
□	Name	Type	Last modified	Size	Storage class

No objects

You don't have any objects in this folder.

Looking for language selection? Find it in the new Unified Settings

© 2022, Amazon Internet Services Private Ltd. or its affiliates.


aws
Services
Search for services, features, blogs, docs, and more [Alt+S]
Global
Harsh

Amazon S3 > Buckets > harshc3294awsbucket > data-source/ > Upload

Upload Info

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. Learn more

Drag and drop files and folders you want to upload here, or choose **Add files**, or **Add folders**.

Files and folders (0)

All files and folders in this table will be uploaded.

Name	Folder	Type	Size

No files or folders

You have not chosen any files or folders to upload.

Destination

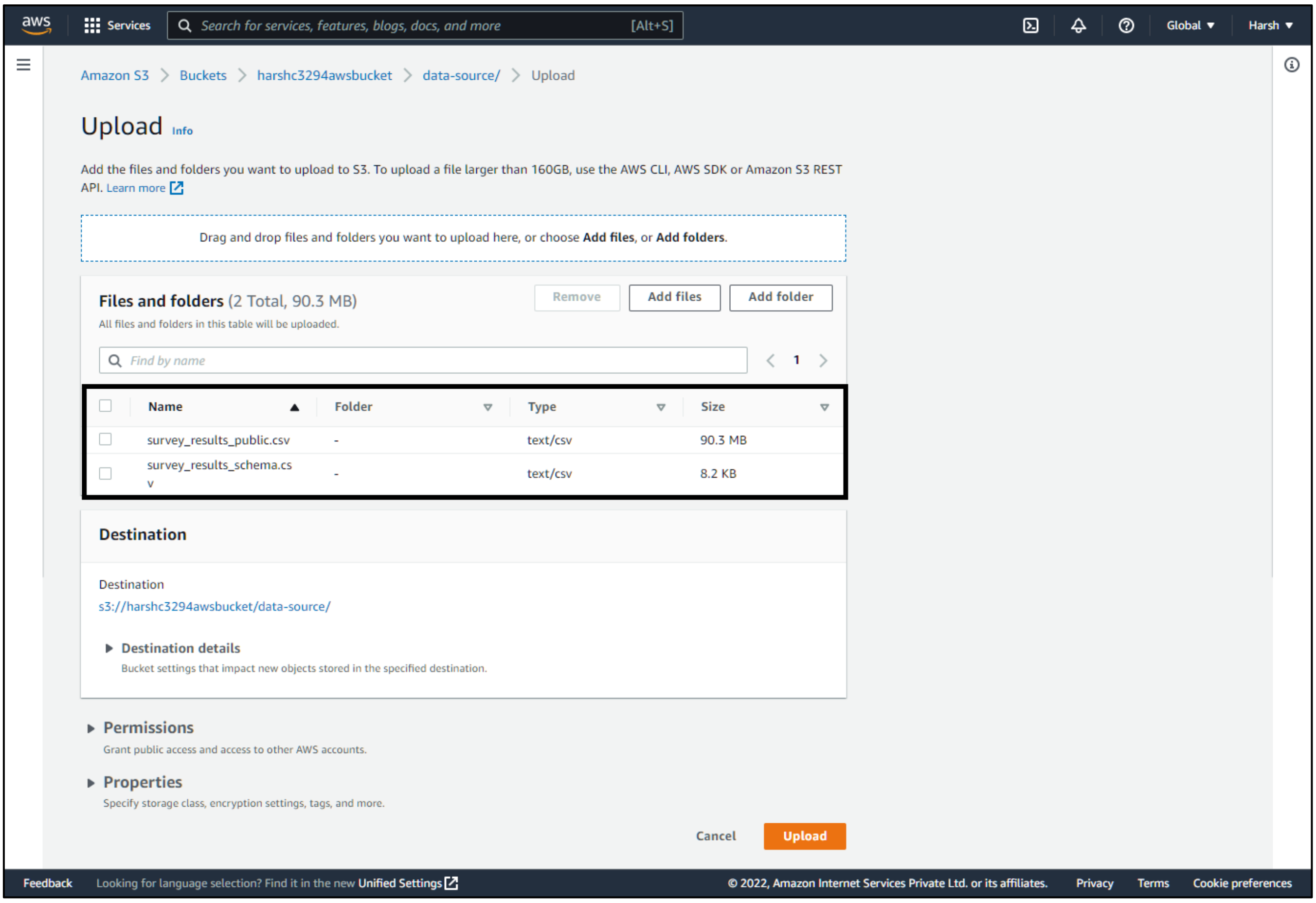
Destination
s3://harshc3294awsbucket/data-source/

▸ Destination details
Bucket settings that impact new objects stored in the specified destination.

▸ Permissions
Grant public access and access to other AWS accounts.

▸ Properties
Specify storage class, encryption settings, tags, and more.

Cancel
Upload
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences


aws
Services
Search for services, features, blogs, docs, and more [Alt+S]
Global
Harsh

Amazon S3 > Buckets > harshc3294awsbucket > data-source/ > Upload

Upload Info

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. Learn more

Drag and drop files and folders you want to upload here, or choose **Add files**, or **Add folders**.

Files and folders (2 Total, 90.3 MB)

All files and folders in this table will be uploaded.

□	Name	Folder	Type	Size
□	survey_results_public.csv	-	text/csv	90.3 MB
□	survey_results_schema.csv	-	text/csv	8.2 KB

Destination

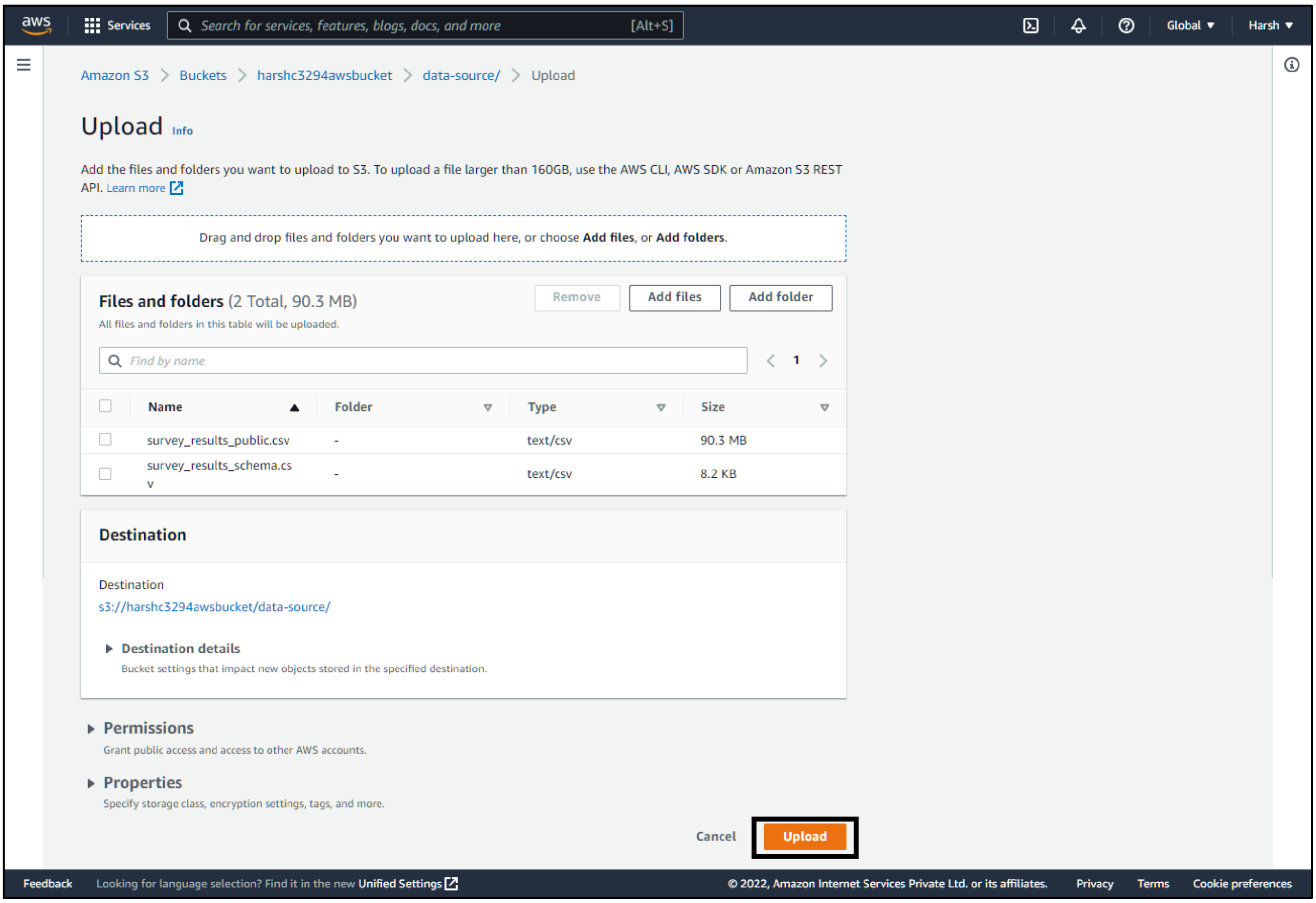
Destination
s3://harshc3294awsbucket/data-source/

Destination details
Bucket settings that impact new objects stored in the specified destination.

Permissions
Grant public access and access to other AWS accounts.

Properties
Specify storage class, encryption settings, tags, and more.

Cancel
Upload
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences


Amazon S3 > Buckets > harshc3294awsbucket > data-source/ > Upload

Upload Info

Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. Learn more

Drag and drop files and folders you want to upload here, or choose **Add files**, or **Add folders**.

Files and folders (2 Total, 90.3 MB)

All files and folders in this table will be uploaded.

□	Name	Folder	Type	Size
□	survey_results_public.csv	-	text/csv	90.3 MB
□	survey_results_schema.csv	-	text/csv	8.2 KB

Destination

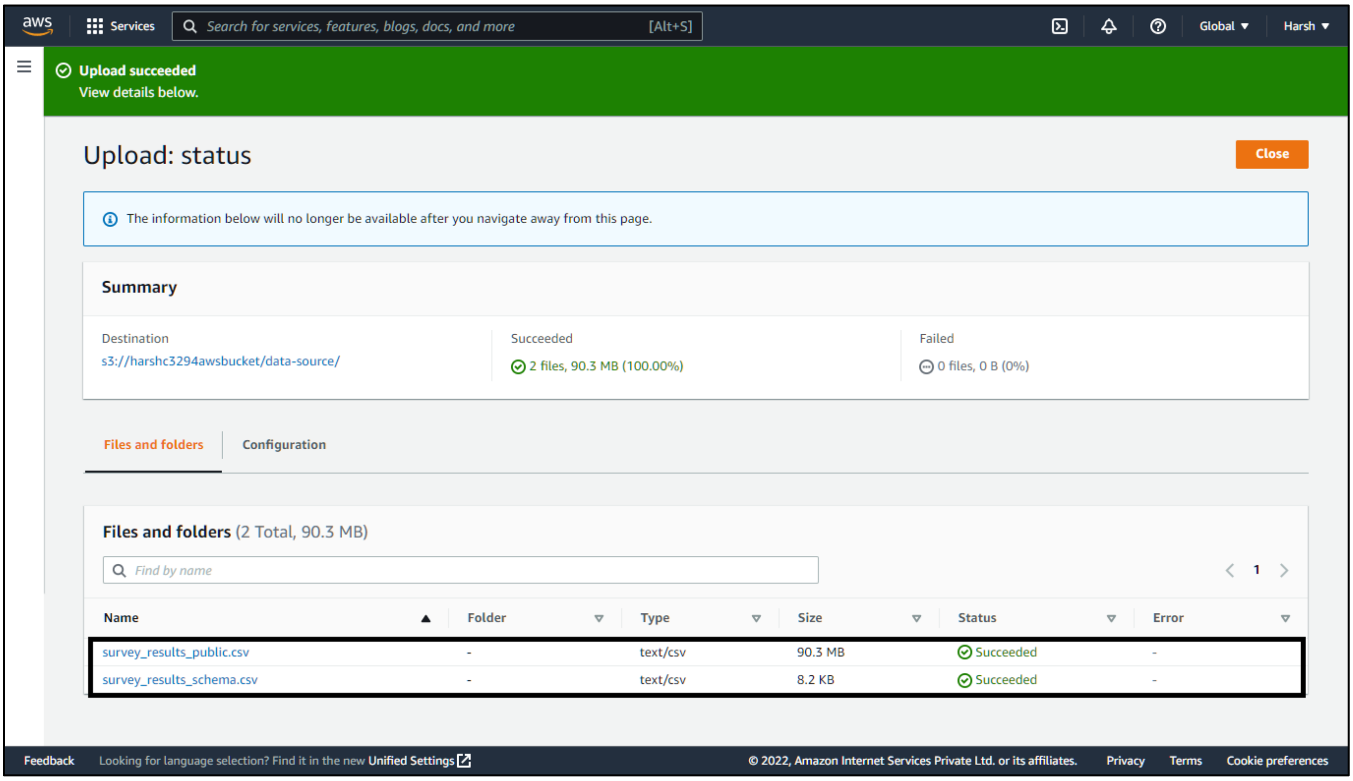
Destination
s3://harshc3294awsbucket/data-source/

Destination details
Bucket settings that impact new objects stored in the specified destination.

Permissions
Grant public access and access to other AWS accounts.

Properties
Specify storage class, encryption settings, tags, and more.

Cancel Upload


aws
Services
Search for services, features, blogs, docs, and more [Alt+S]
Global ▾
Harsh ▾

Upload succeeded
View details below.

Upload: status

The information below will no longer be available after you navigate away from this page.

Summary

Destination	Succeeded	Failed
s3://harshc3294awsbucket/data-source/	2 files, 90.3 MB (100.00%)	0 files, 0 B (0%)

Files and folders (2 Total, 90.3 MB)

Name	Folder	Type	Size	Status	Error
survey_results_public.csv	-	text/csv	90.3 MB	Succeeded	-
survey_results_schema.csv	-	text/csv	8.2 KB	Succeeded	-

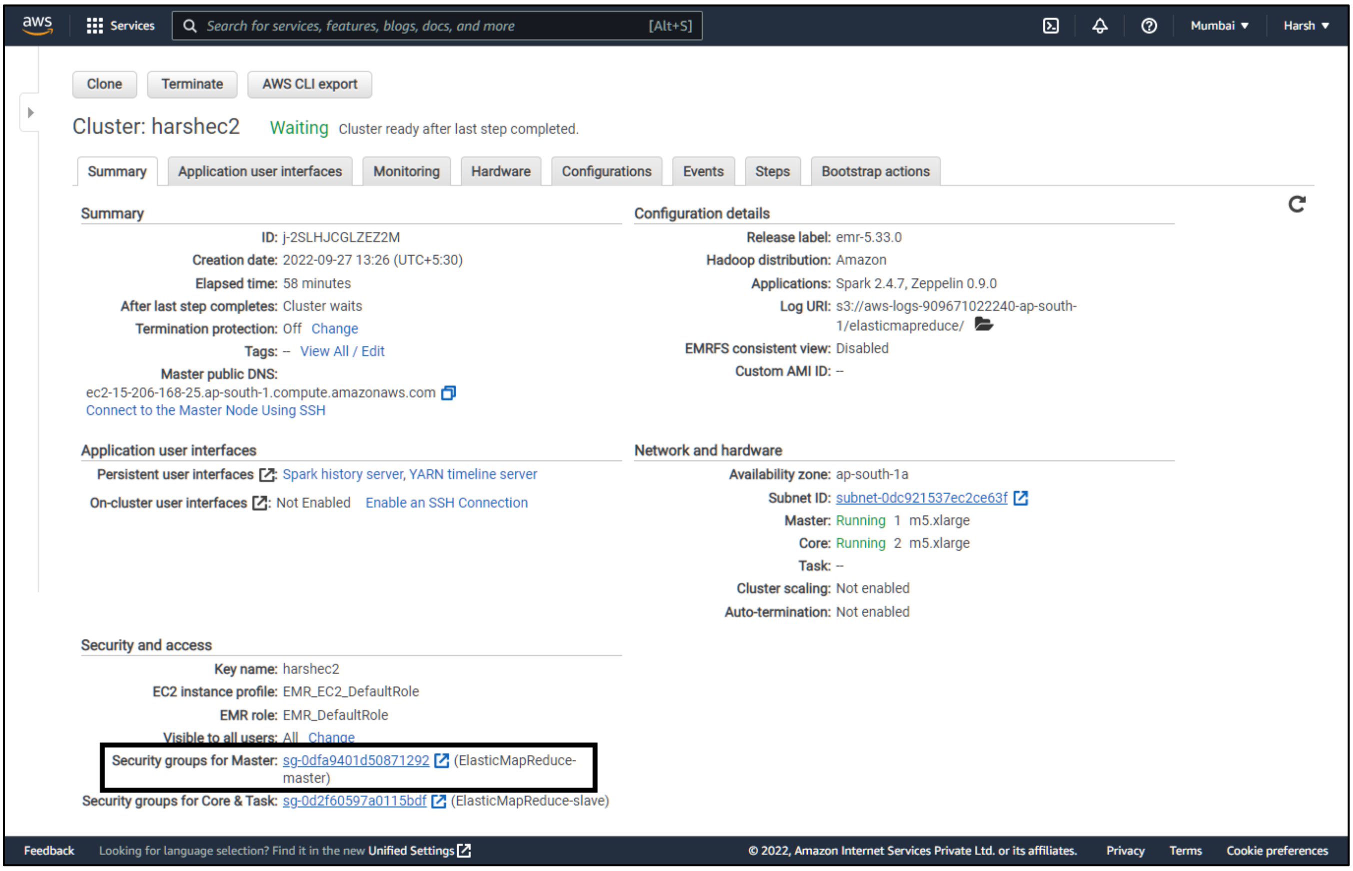
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

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**


aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Mumbai ▾
Harsh ▾
Clone
Terminate
AWS CLI export
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

Cluster: harshec2 Waiting Cluster ready after last step completed.

Summary

ID: j-2SLHJCGLZEZ2M
Creation date: 2022-09-27 13:26 (UTC+5:30)
Elapsed time: 58 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS:
ec2-15-206-168-25.ap-south-1.compute.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.33.0
Hadoop distribution: Amazon
Applications: Spark 2.4.7, Zeppelin 0.9.0
Log URI: s3://aws-logs-909671022240-ap-south-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

Persistent user interfaces: Spark history server, YARN timeline server
On-cluster user interfaces: Not Enabled Enable an SSH Connection

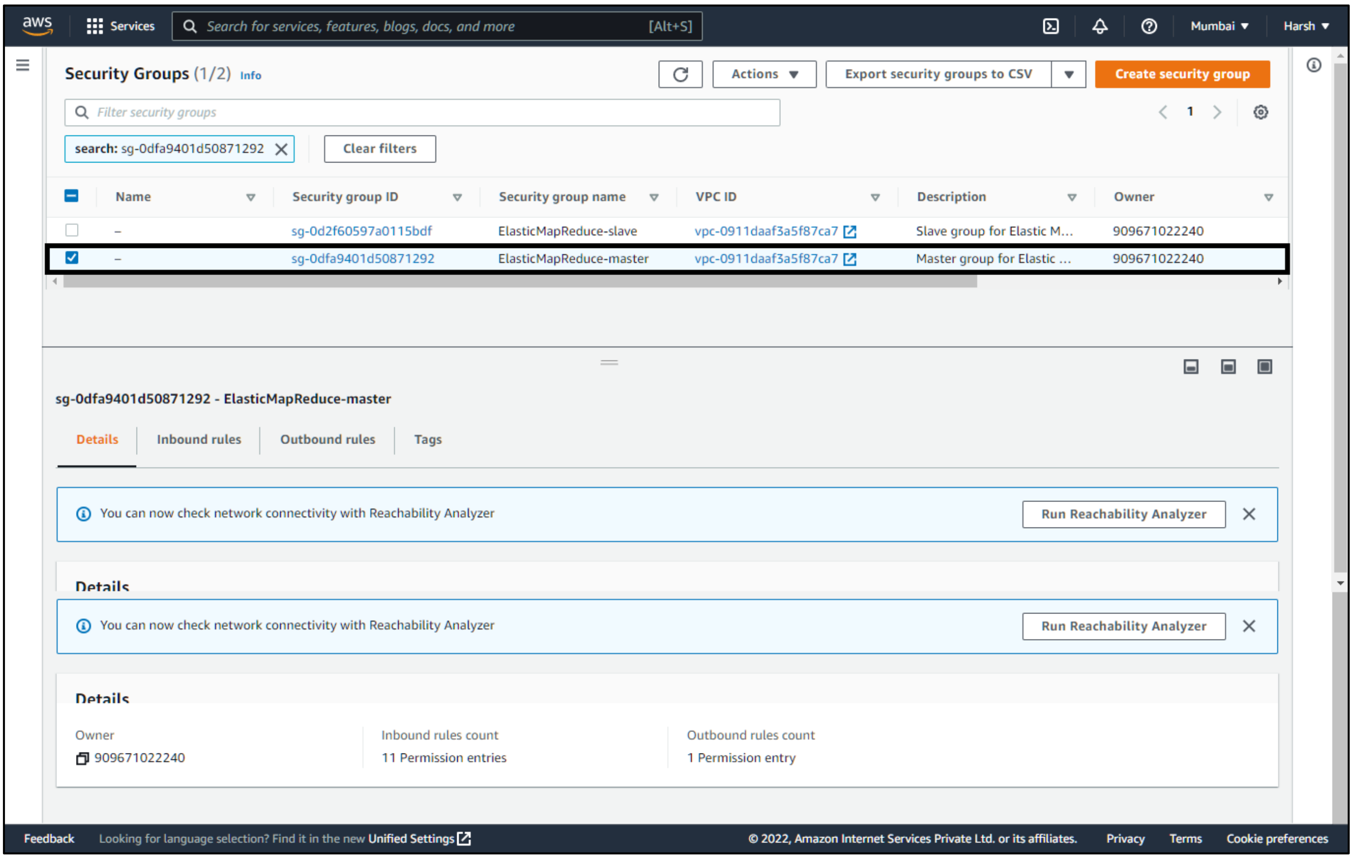
Network and hardware

Availability zone: ap-south-1a
Subnet ID: subnet-0dc921537ec2ce63f
Master: Running 1 m5.xlarge
Core: Running 2 m5.xlarge
Task: --
Cluster scaling: Not enabled
Auto-termination: Not enabled

Security and access

Key name: harshec2
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master: sg-0dfa9401d50871292 (ElasticMapReduce-master)
Security groups for Core & Task: sg-0d2f60597a0115bdf (ElasticMapReduce-slave)

Step 2: Select the master node security group



Security Groups (1/2) Info

search: sg-0dfa9401d50871292

☑	Name	Security group ID	Security group name	VPC ID	Description	Owner
□	–	sg-0d2f60597a0115bdf	ElasticMapReduce-slave	vpc-0911daaf3a5f87ca7	Slave group for Elastic M...	909671022240
☑	–	sg-0dfa9401d50871292	ElasticMapReduce-master	vpc-0911daaf3a5f87ca7	Master group for Elastic ...	909671022240

sg-0dfa9401d50871292 - ElasticMapReduce-master

You can now check network connectivity with Reachability Analyzer

Details

You can now check network connectivity with Reachability Analyzer

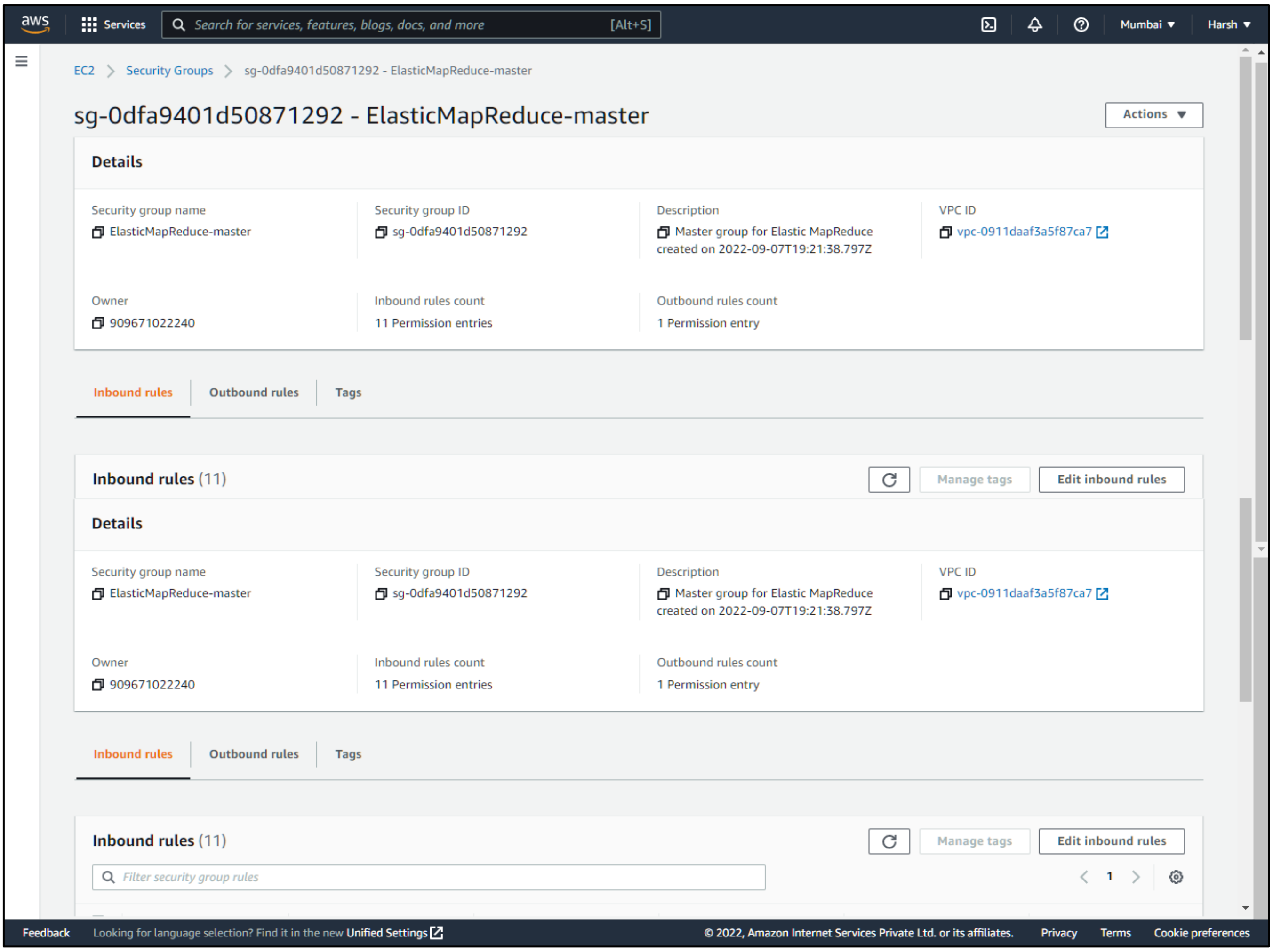
Details

Owner	Inbound rules count	Outbound rules count
909671022240	11 Permission entries	1 Permission entry

Looking for language selection? Find it in the new Unified Settings

© 2022, Amazon Internet Services Private Ltd. or its affiliates. Privacy Terms Cookie preferences

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


aws
Services
Search for services, features, blogs, docs, and more [Alt+S]
Mumbai
Harsh
EC2 > Security Groups > sg-0dfa9401d50871292 - ElasticMapReduce-master
Actions

sg-0dfa9401d50871292 - ElasticMapReduce-master

Details

Security group name ElasticMapReduce-master	Security group ID sg-0dfa9401d50871292	Description Master group for Elastic MapReduce created on 2022-09-07T19:21:38.797Z	VPC ID vpc-0911daaf3a5f87ca7
Owner 909671022240	Inbound rules count 11 Permission entries	Outbound rules count 1 Permission entry	

Inbound rules (11)

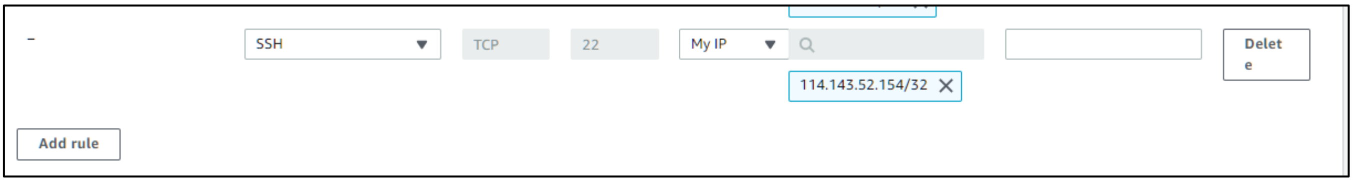
Details

Security group name ElasticMapReduce-master	Security group ID sg-0dfa9401d50871292	Description Master group for Elastic MapReduce created on 2022-09-07T19:21:38.797Z	VPC ID vpc-0911daaf3a5f87ca7
Owner 909671022240	Inbound rules count 11 Permission entries	Outbound rules count 1 Permission entry	

Inbound rules (11)

Inbound rules
Outbound rules
Tags
Manage tags
Edit inbound rules
Filter security group rules
1
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Internet Services Private Ltd. or its affiliates.
Privacy
Terms
Cookie preferences

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



SSH	TCP	22	My IP	114.143.52.154/32

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.

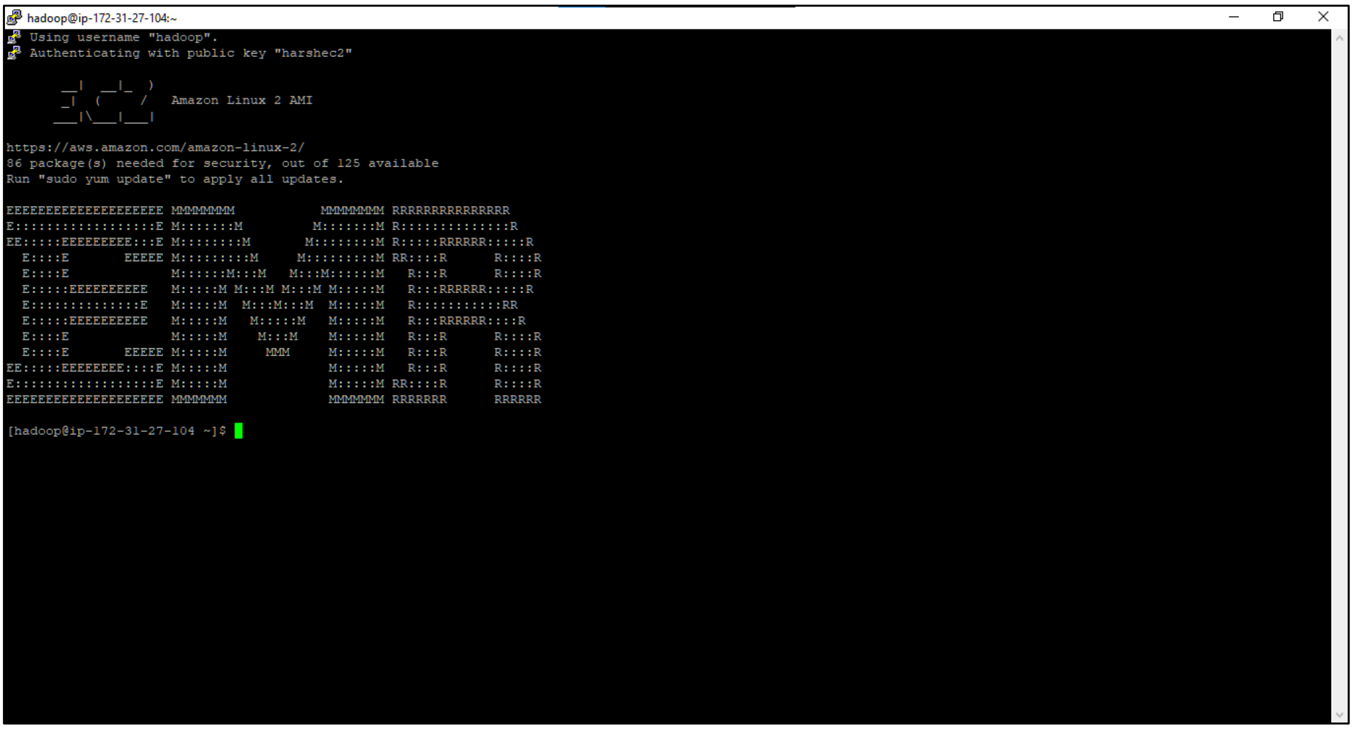
Step 5: In the Host Name field, type hadoop@ec2-15-206-168-25.ap-south-1.compute.amazonaws.com

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.



hadoop@ip-172-31-27-104:~

```
Using username "hadoop".
Authenticating with public key "harshec2"

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
86 package(s) needed for security, out of 125 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-27-104 ~]$
```

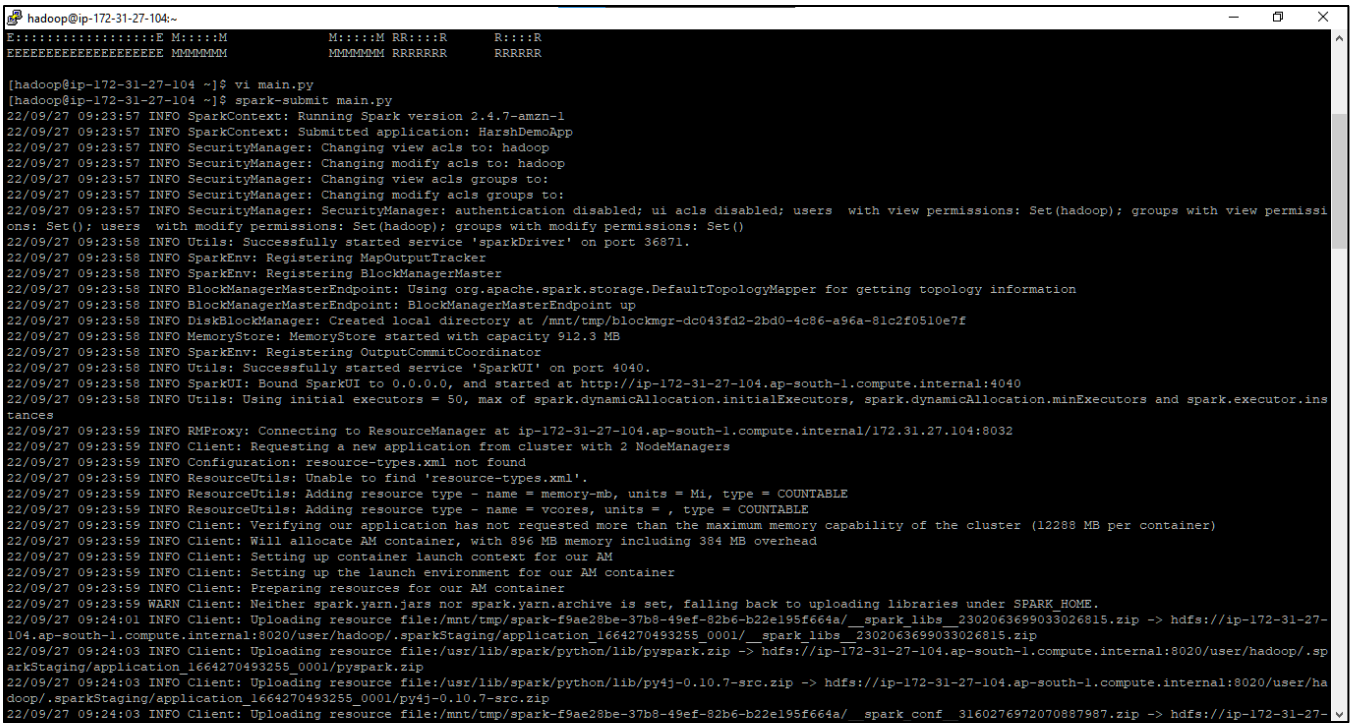
Step 10: open vi main.py

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

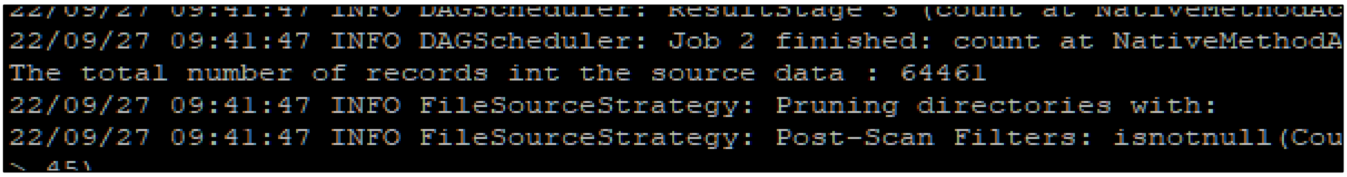
Step 12: ESC then :wq

Step 13: spark-submit main.py

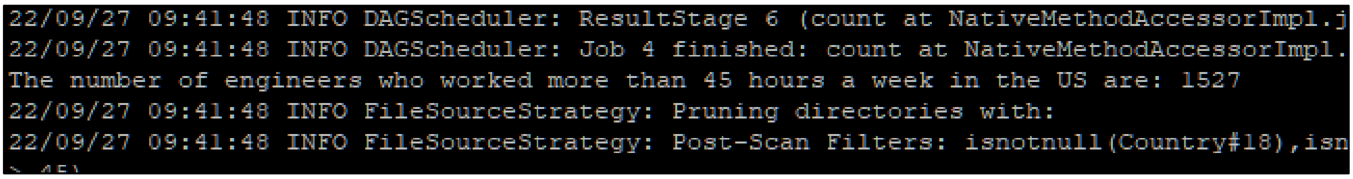
Output



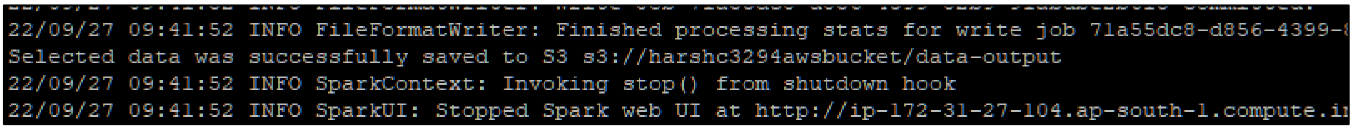
```
E::::::::::::::::::::E M:::::M                M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEEE MMMMMMM                MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-27-104 ~]$ vi main.py
[hadoop@ip-172-31-27-104 ~]$ spark-submit main.py
22/09/27 09:23:57 INFO SparkContext: Running Spark version 2.4.7-amzn-1
22/09/27 09:23:57 INFO SparkContext: Submitted application: HarshDemoApp
22/09/27 09:23:57 INFO SecurityManager: Changing view acls to: hadoop
22/09/27 09:23:57 INFO SecurityManager: Changing modify acls to: hadoop
22/09/27 09:23:57 INFO SecurityManager: Changing view acls groups to:
22/09/27 09:23:57 INFO SecurityManager: Changing modify acls groups to:
22/09/27 09:23:57 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(hadoop); groups with view permissi
ons: Set(); users  with modify permissions: Set(hadoop); groups with modify permissions: Set()
22/09/27 09:23:58 INFO Utils: Successfully started service 'sparkDriver' on port 36871.
22/09/27 09:23:58 INFO SparkEnv: Registering MapOutputTracker
22/09/27 09:23:58 INFO SparkEnv: Registering BlockManagerMaster
22/09/27 09:23:58 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
22/09/27 09:23:58 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
22/09/27 09:23:58 INFO DiskBlockManager: Created local directory at /mnt/tmp/blockmgr-dc043fd2-2bd0-4c86-a96a-81c2f0510e7f
22/09/27 09:23:58 INFO MemoryStore: MemoryStore started with capacity 912.3 MB
22/09/27 09:23:58 INFO SparkEnv: Registering OutputCommitCoordinator
22/09/27 09:23:58 INFO Utils: Successfully started service 'SparkUI' on port 4040.
22/09/27 09:23:58 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://ip-172-31-27-104.ap-south-1.compute.internal:4040
22/09/27 09:23:58 INFO Utils: Using initial executors = 50, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spark.executor.ins
tances
22/09/27 09:23:59 INFO RMProxy: Connecting to ResourceManager at ip-172-31-27-104.ap-south-1.compute.internal/172.31.27.104:8032
22/09/27 09:23:59 INFO Client: Requesting a new application from cluster with 2 NodeManagers
22/09/27 09:23:59 INFO Configuration: resource-types.xml not found
22/09/27 09:23:59 INFO ResourceUtils: Unable to find 'resource-types.xml'.
22/09/27 09:23:59 INFO ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
22/09/27 09:23:59 INFO ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
22/09/27 09:23:59 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (12288 MB per container)
22/09/27 09:23:59 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
22/09/27 09:23:59 INFO Client: Setting up container launch context for our AM
22/09/27 09:23:59 INFO Client: Setting up the launch environment for our AM container
22/09/27 09:23:59 INFO Client: Preparing resources for our AM container
22/09/27 09:23:59 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
22/09/27 09:24:01 INFO Client: Uploading resource file:/mnt/tmp/spark-f9ae28be-37b8-49ef-82b6-b22e195f664a/__spark_libs__2302063699033026815.zip -> hdfs://ip-172-31-27-
104.ap-south-1.compute.internal:8020/user/hadoop/.sparkStaging/application_1664270493255_0001/__spark_libs__2302063699033026815.zip
22/09/27 09:24:03 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/pyspark.zip -> hdfs://ip-172-31-27-104.ap-south-1.compute.internal:8020/user/hadoop/.sp
arkStaging/application_1664270493255_0001/pyspark.zip
22/09/27 09:24:03 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/py4j-0.10.7-src.zip -> hdfs://ip-172-31-27-104.ap-south-1.compute.internal:8020/user/ha
doop/.sparkStaging/application_1664270493255_0001/py4j-0.10.7-src.zip
22/09/27 09:24:03 INFO Client: Uploading resource file:/mnt/tmp/spark-f9ae28be-37b8-49ef-82b6-b22e195f664a/__spark_conf__3160276972070887987.zip -> hdfs://ip-172-31-27-
```



```
22/09/27 09:41:47 INFO DAGScheduler: ResultStage 5 (count at NativeMethodAc
22/09/27 09:41:47 INFO DAGScheduler: Job 2 finished: count at NativeMethodA
The total number of records int the source data : 64461
22/09/27 09:41:47 INFO FileSourceStrategy: Pruning directories with:
22/09/27 09:41:47 INFO FileSourceStrategy: Post-Scan Filters: isnotnull(Cou
> 45)
```



```
22/09/27 09:41:48 INFO DAGScheduler: ResultStage 6 (count at NativeMethodAccessorImpl.j
22/09/27 09:41:48 INFO DAGScheduler: Job 4 finished: count at NativeMethodAccessorImpl.
The number of engineers who worked more than 45 hours a week in the US are: 1527
22/09/27 09:41:48 INFO FileSourceStrategy: Pruning directories with:
22/09/27 09:41:48 INFO FileSourceStrategy: Post-Scan Filters: isnotnull(Country#18),isn
> 45)
```



```
22/09/27 09:41:52 INFO FileFormatWriter: Finished processing stats for write job 71a55dc8-d856-4399-
Selected data was successfully saved to S3 s3://harshc3294awsbucket/data-output
22/09/27 09:41:52 INFO SparkContext: Invoking stop() from shutdown hook
22/09/27 09:41:52 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-27-104.ap-south-1.compute.in
```