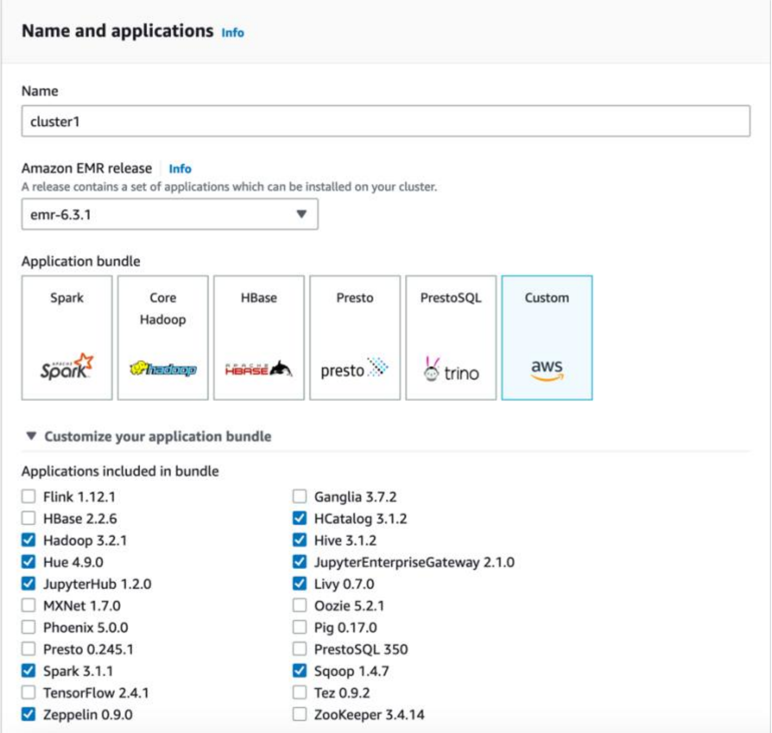
Reto 5-1

1. Creación y configuración del clúster en AWS.



Name and applications Info

Name

cluster1

Amazon EMR release | Info

A release contains a set of applications which can be installed on your cluster.

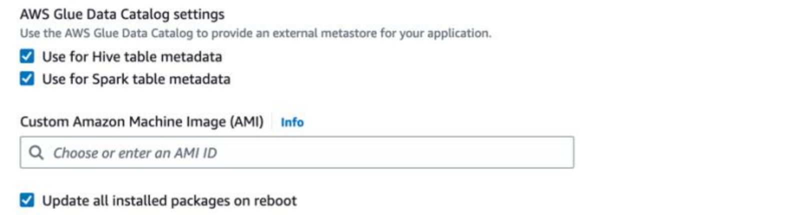
emr-6.3.1 ▼

Application bundle

▼ Customize your application bundle

Applications included in bundle

☐ Flink 1.12.1
☐ HBase 2.2.6
☑ Hadoop 3.2.1
☑ Hue 4.9.0
☑ JupyterHub 1.2.0
☐ MXNet 1.7.0
☐ Phoenix 5.0.0
☐ Presto 0.245.1
☑ Spark 3.1.1
☐ TensorFlow 2.4.1
☑ Zeppelin 0.9.0
☐ Ganglia 3.7.2
☑ HCatalog 3.1.2
☑ Hive 3.1.2
☑ JupyterEnterpriseGateway 2.1.0
☑ Livy 0.7.0
☐ Oozie 5.2.1
☐ Pig 0.17.0
☐ PrestoSQL 350
☑ Sqoop 1.4.7
☐ Tez 0.9.2
☐ ZooKeeper 3.4.14



AWS Glue Data Catalog settings

Use the AWS Glue Data Catalog to provide an external metastore for your application.

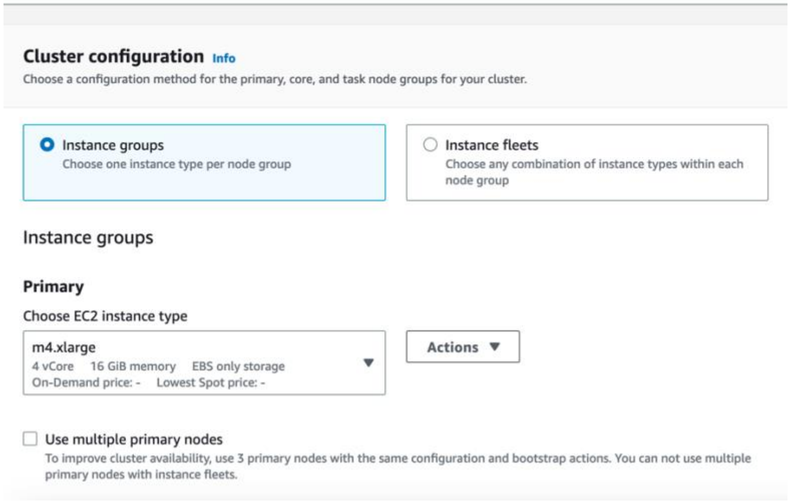
☑ Use for Hive table metadata

☑ Use for Spark table metadata

Custom Amazon Machine Image (AMI) | Info

Choose or enter an AMI ID

☑ Update all installed packages on reboot



Cluster configuration Info

Choose a configuration method for the primary, core, and task node groups for your cluster.

● Instance groups
Choose one instance type per node group

○ Instance fleets
Choose any combination of instance types within each node group

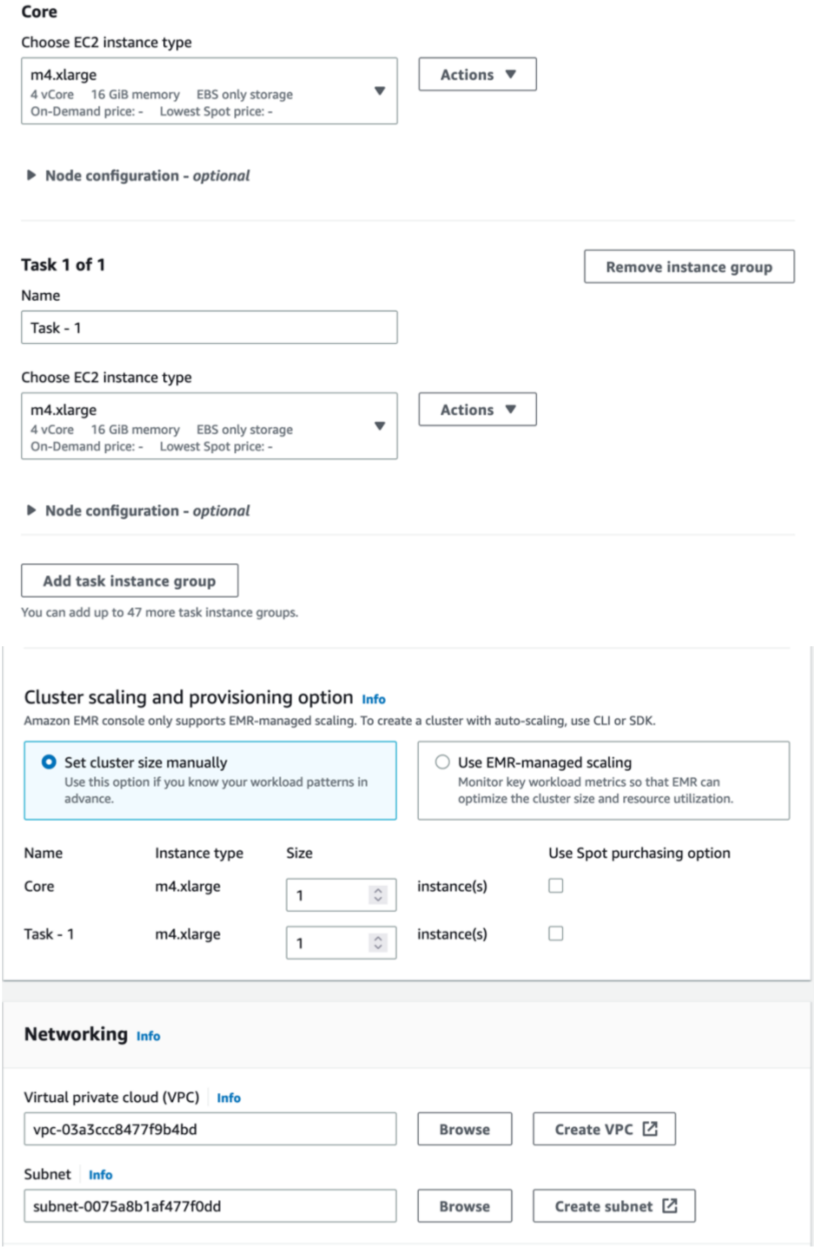
Instance groups

Primary

Choose EC2 instance type

m4.xlarge
4 vCore 16 GiB memory EBS only storage
On-Demand price: - Lowest Spot price: - ▼

☐ Use multiple primary nodes
To improve cluster availability, use 3 primary nodes with the same configuration and bootstrap actions. You can not use multiple primary nodes with instance fleets.



Core

Choose EC2 instance type

m4.xlarge
4 vCore 16 GiB memory EBS only storage
On-Demand price: - Lowest Spot price: -

▶ **Node configuration - *optional***

Task 1 of 1

Name

Task - 1

Choose EC2 instance type

m4.xlarge
4 vCore 16 GiB memory EBS only storage
On-Demand price: - Lowest Spot price: -

▶ **Node configuration - *optional***

You can add up to 47 more task instance groups.

Cluster scaling and provisioning option

Amazon EMR console only supports EMR-managed scaling. To create a cluster with auto-scaling, use CLI or SDK.

● Set cluster size manually
Use this option if you know your workload patterns in advance.

○ Use EMR-managed scaling
Monitor key workload metrics so that EMR can optimize the cluster size and resource utilization.

Name	Instance type	Size		Use Spot purchasing option
Core	m4.xlarge	1	instance(s)	□
Task - 1	m4.xlarge	1	instance(s)	□

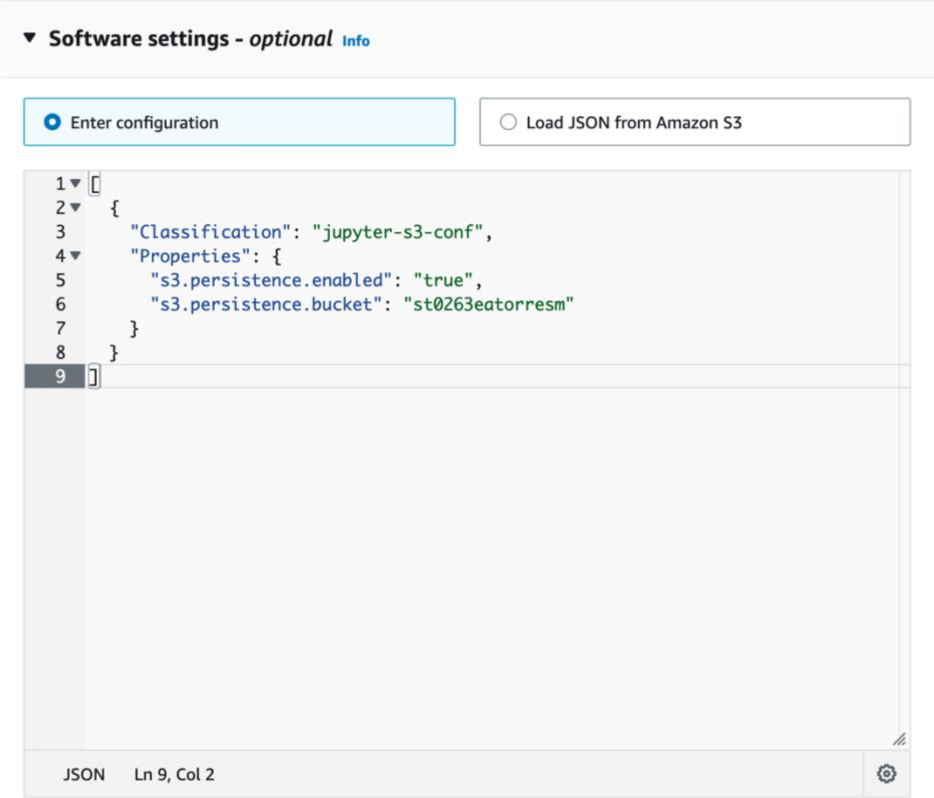
Networking

Virtual private cloud (VPC)

vpc-03a3ccc8477f9b4bd

Subnet

subnet-0075a8b1af477f0dd

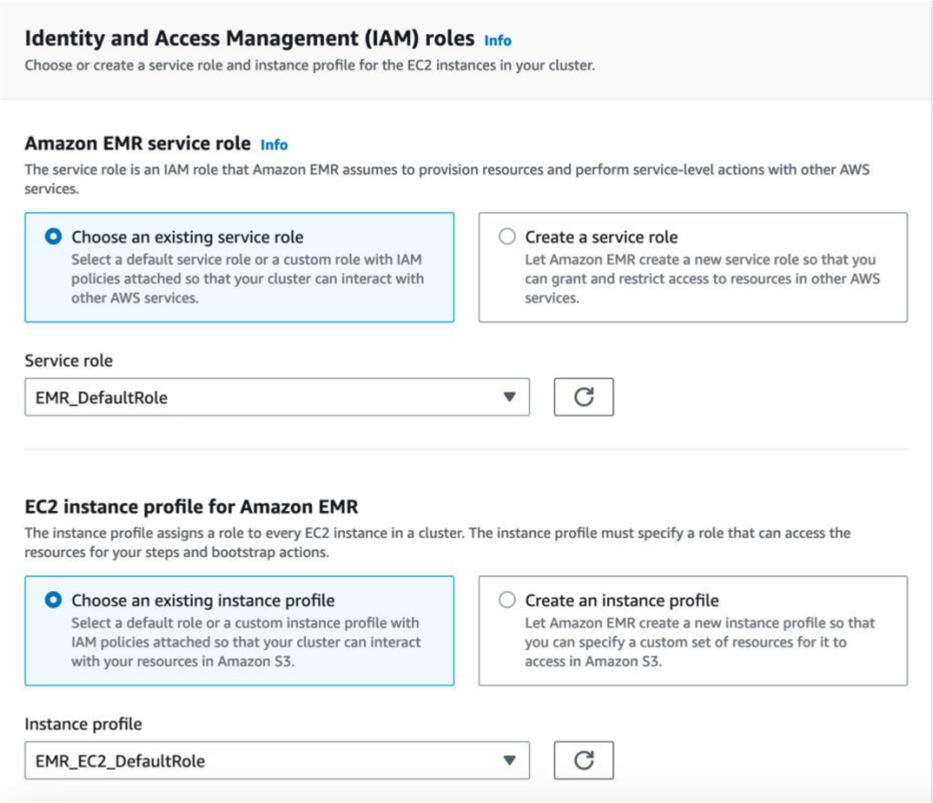


▼ **Software settings - *optional*** Info

● Enter configuration ○ Load JSON from Amazon S3

```
[
  {
    "Classification": "jupyter-s3-conf",
    "Properties": {
      "s3.persistence.enabled": "true",
      "s3.persistence.bucket": "st0263eatorresm"
    }
  }
]
```

JSON Ln 9, Col 2



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 ▼

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.

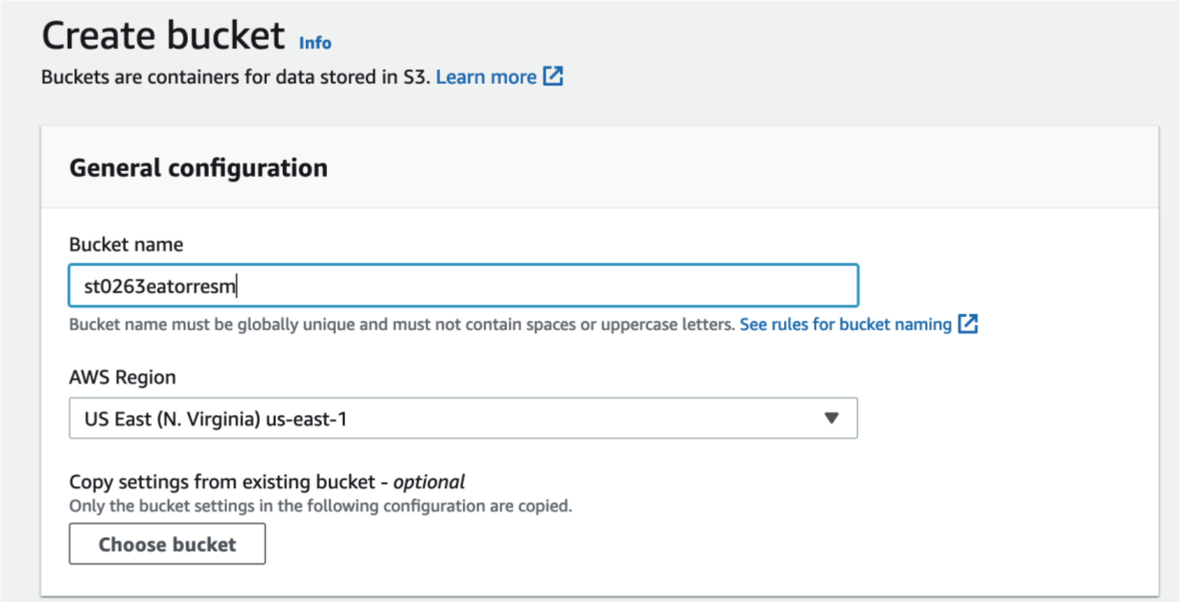
● Choose an existing instance profile
Select a default role or a custom instance profile with IAM policies attached so that your cluster can interact with your resources in Amazon S3.

○ Create an instance profile
Let Amazon EMR create a new instance profile so that you can specify a custom set of resources for it to access in Amazon S3.

Instance profile

EMR_EC2_DefaultRole ▼

2. Creación del bucket



Create bucket

Buckets are containers for data stored in S3. Learn more

General configuration

Bucket name

st0263eatorresm

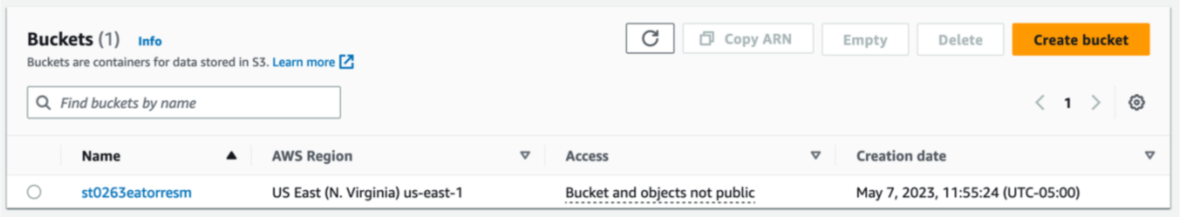
Bucket name must be globally unique and must not contain spaces or uppercase letters. See rules for bucket naming

AWS Region

US East (N. Virginia) us-east-1

Copy settings from existing bucket - *optional*

Only the bucket settings in the following configuration are copied.

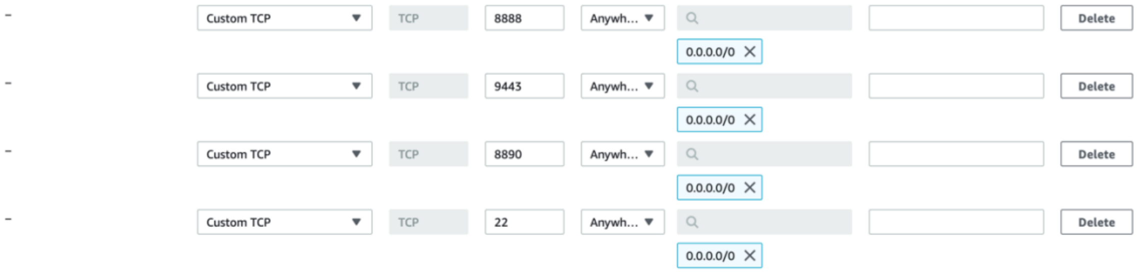


Buckets (1)

Buckets are containers for data stored in S3. Learn more

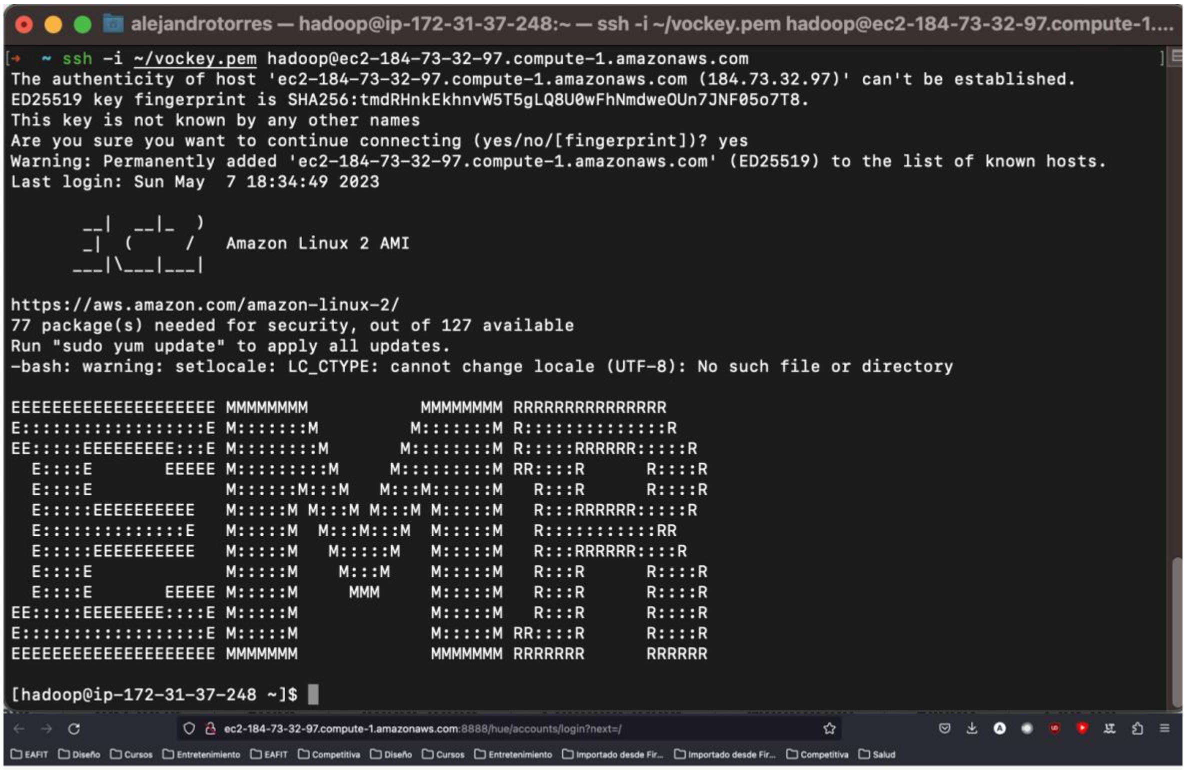
	Name	AWS Region	Access	Creation date
○	st0263eatorresm	US East (N. Virginia) us-east-1	Bucket and objects not public	May 7, 2023, 11:55:24 (UTC-05:00)

3. Configuración de seguridad.



-	Custom TCP	TCP	8888	Anywh...	0.0.0.0/0		Delete
-	Custom TCP	TCP	9443	Anywh...	0.0.0.0/0		Delete
-	Custom TCP	TCP	8890	Anywh...	0.0.0.0/0		Delete
-	Custom TCP	TCP	22	Anywh...	0.0.0.0/0		Delete

4. Conexión con el clúster, Hue y Zeppelin.


alejandrotorres — hadoop@ip-172-31-37-248:~ — ssh -i ~/vockey.pem hadoop@ec2-184-73-32-97.compute-1....

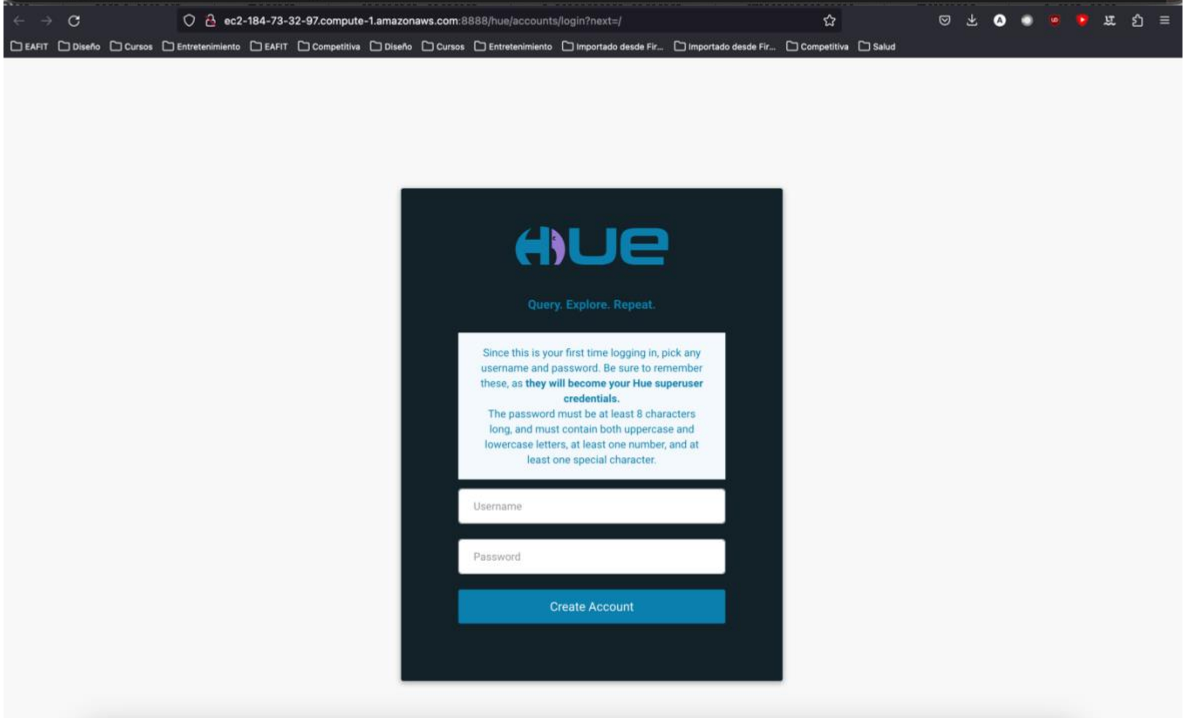
```
➜  ~ ssh -i ~/vockey.pem hadoop@ec2-184-73-32-97.compute-1.amazonaws.com
The authenticity of host 'ec2-184-73-32-97.compute-1.amazonaws.com (184.73.32.97)' can't be established.
ED25519 key fingerprint is SHA256:tmdRHnkEkhnvW5T5gLQ8U0wFhNmdweOUn7JNF05o7T8.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-184-73-32-97.compute-1.amazonaws.com' (ED25519) to the list of known hosts.
Last login: Sun May  7 18:34:49 2023

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

https://aws.amazon.com/amazon-linux-2/
77 package(s) needed for security, out of 127 available
Run "sudo yum update" to apply all updates.
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory

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-172-31-37-248 ~]$ 
```

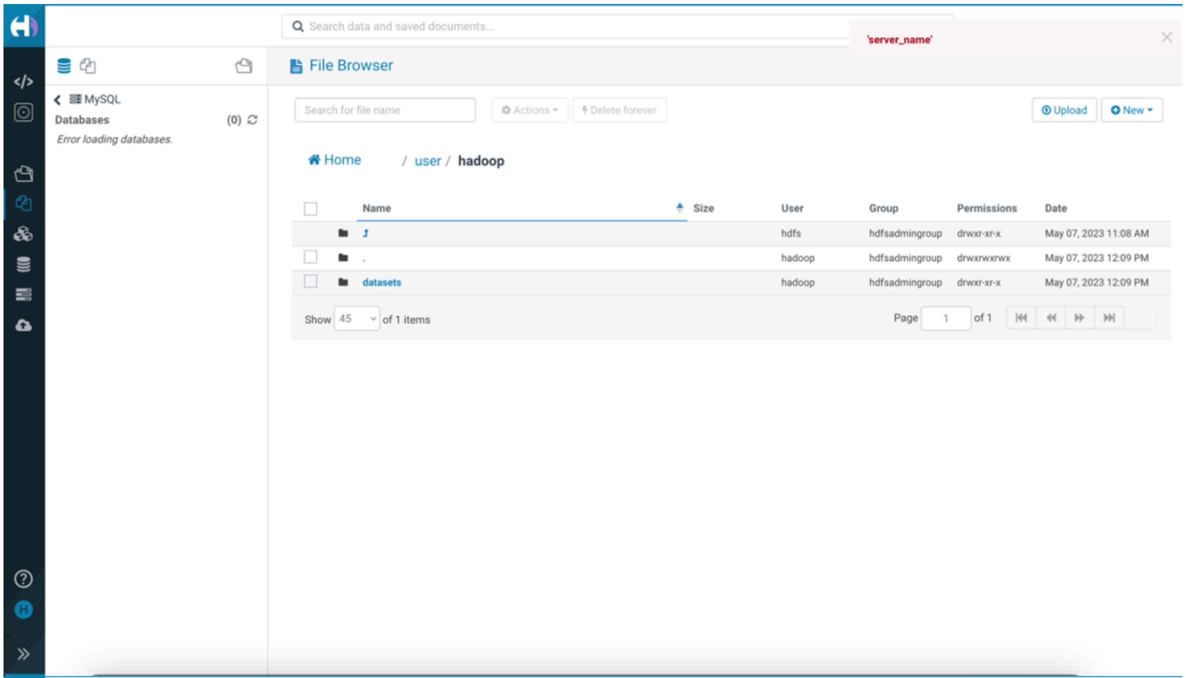


Since this is your first time logging in, pick any username and password. Be sure to remember these, as **they will become your Hue superuser credentials.**
The password must be at least 8 characters long, and must contain both uppercase and lowercase letters, at least one number, and at least one special character.

Username

Password

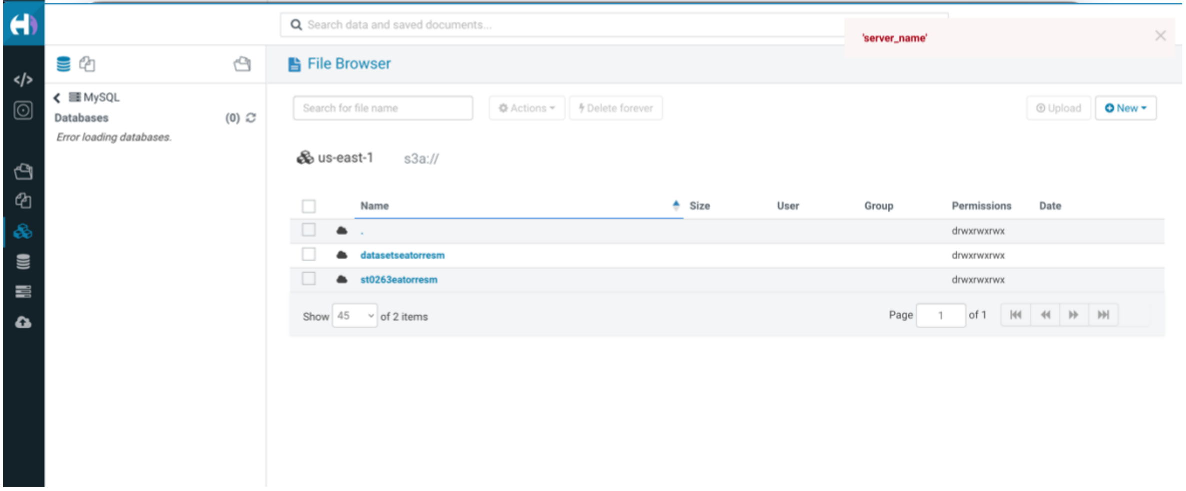
Create Account



Home / user / hadoop

Name	Size	User	Group	Permissions	Date
↥		hdfs	hdfsadmingroup	drwxr-xr-x	May 07, 2023 11:08 AM
.		hadoop	hdfsadmingroup	drwxrwxrwx	May 07, 2023 12:09 PM
datasets		hadoop	hdfsadmingroup	drwxr-xr-x	May 07, 2023 12:09 PM

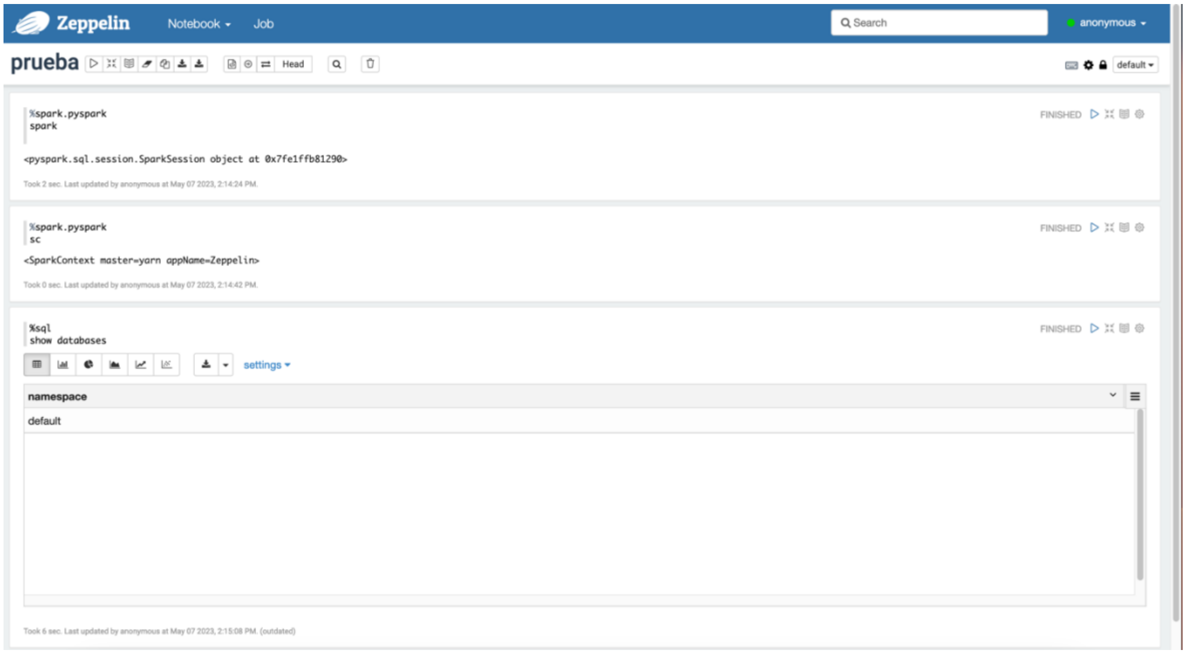
Show 45 of 1 items Page 1 of 1



us-east-1 s3a://

Name	Size	User	Group	Permissions	Date
.				drwxrwxrwx	
datasetseatorresm				drwxrwxrwx	
st0263eatorresm				drwxrwxrwx	

Show 45 of 2 items Page 1 of 1



```
%spark.pyspark
spark
```

```
<pyspark.sql.session.SparkSession object at 0x7fe1ffb81290>
```

Took 2 sec. Last updated by anonymous at May 07 2023, 2:14:24 PM.

```
%spark.pyspark
sc
```

```
<SparkContext master=yarn appName=Zeppelin>
```

Took 0 sec. Last updated by anonymous at May 07 2023, 2:14:42 PM.

```
%sql
show databases
```

namespace
default

Took 6 sec. Last updated by anonymous at May 07 2023, 2:15:08 PM. (outdated)