EAFIT

Special topics in telematics

Laboratory 0-1

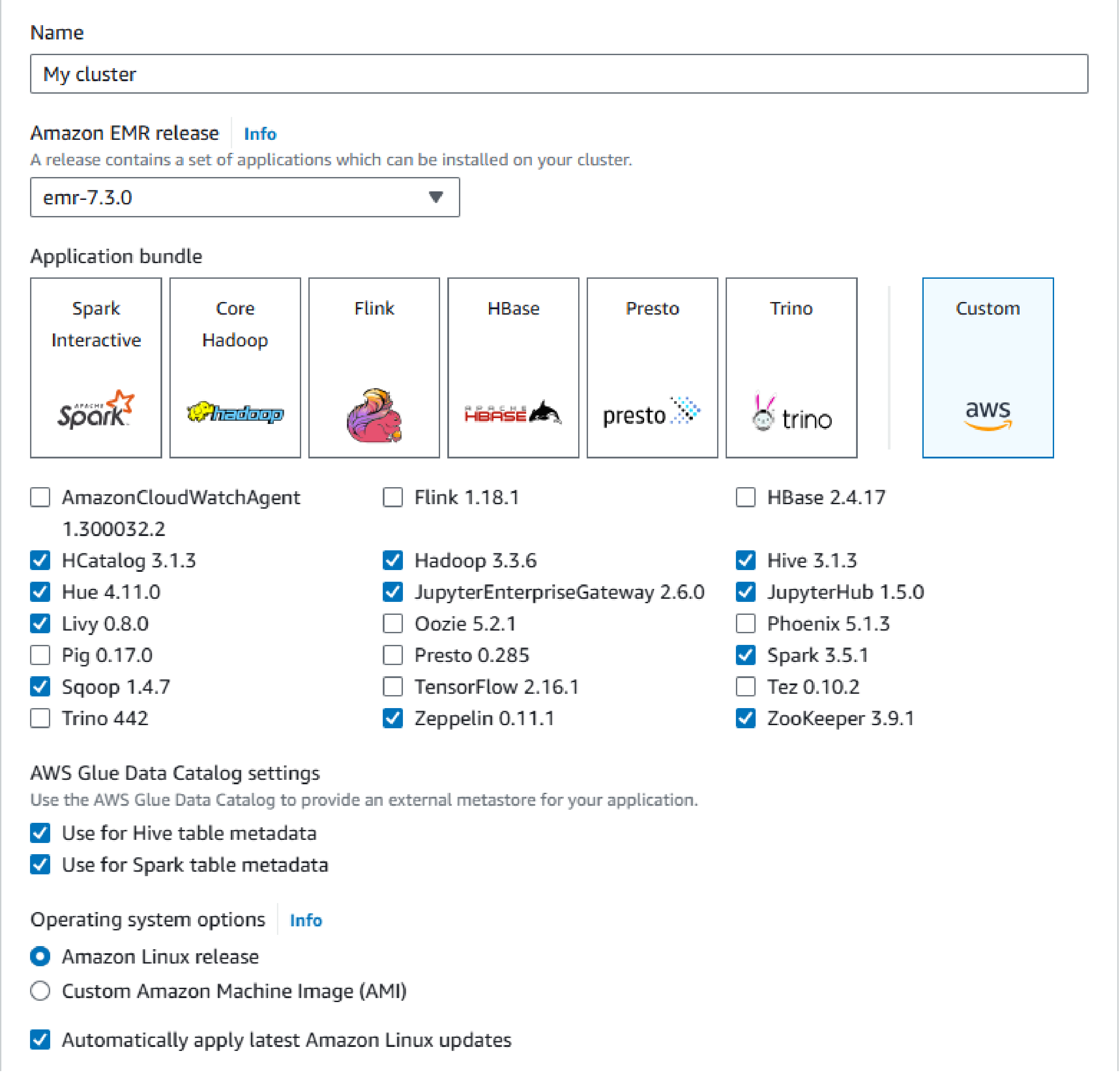
Luisa María Álvarez García – Computer Science

(lmalvarez8@eafit.edu.co)

Teacher: Edwin Nelson Montoya

Medellin, November 11, 2024

1. Go to the EMR cluster section and select the “create option”.
2. Follow the steps on the creation, be sure to select all image bundles:



Name

My cluster

Amazon EMR release Info

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

emr-7.3.0 ▼

Application bundle

☐ AmazonCloudWatchAgent 1.300032.2
☐ Flink 1.18.1
☐ HBase 2.4.17
☑ HCatalog 3.1.3
☑ Hadoop 3.3.6
☑ Hive 3.1.3
☑ Hue 4.11.0
☑ JupyterEnterpriseGateway 2.6.0
☑ JupyterHub 1.5.0
☑ Livy 0.8.0
☐ Oozie 5.2.1
☐ Phoenix 5.1.3
☐ Pig 0.17.0
☐ Presto 0.285
☑ Spark 3.5.1
☑ Sqoop 1.4.7
☐ TensorFlow 2.16.1
☐ Tez 0.10.2
☐ Trino 442
☑ Zeppelin 0.11.1
☑ ZooKeeper 3.9.1

AWS Glue Data Catalog settings

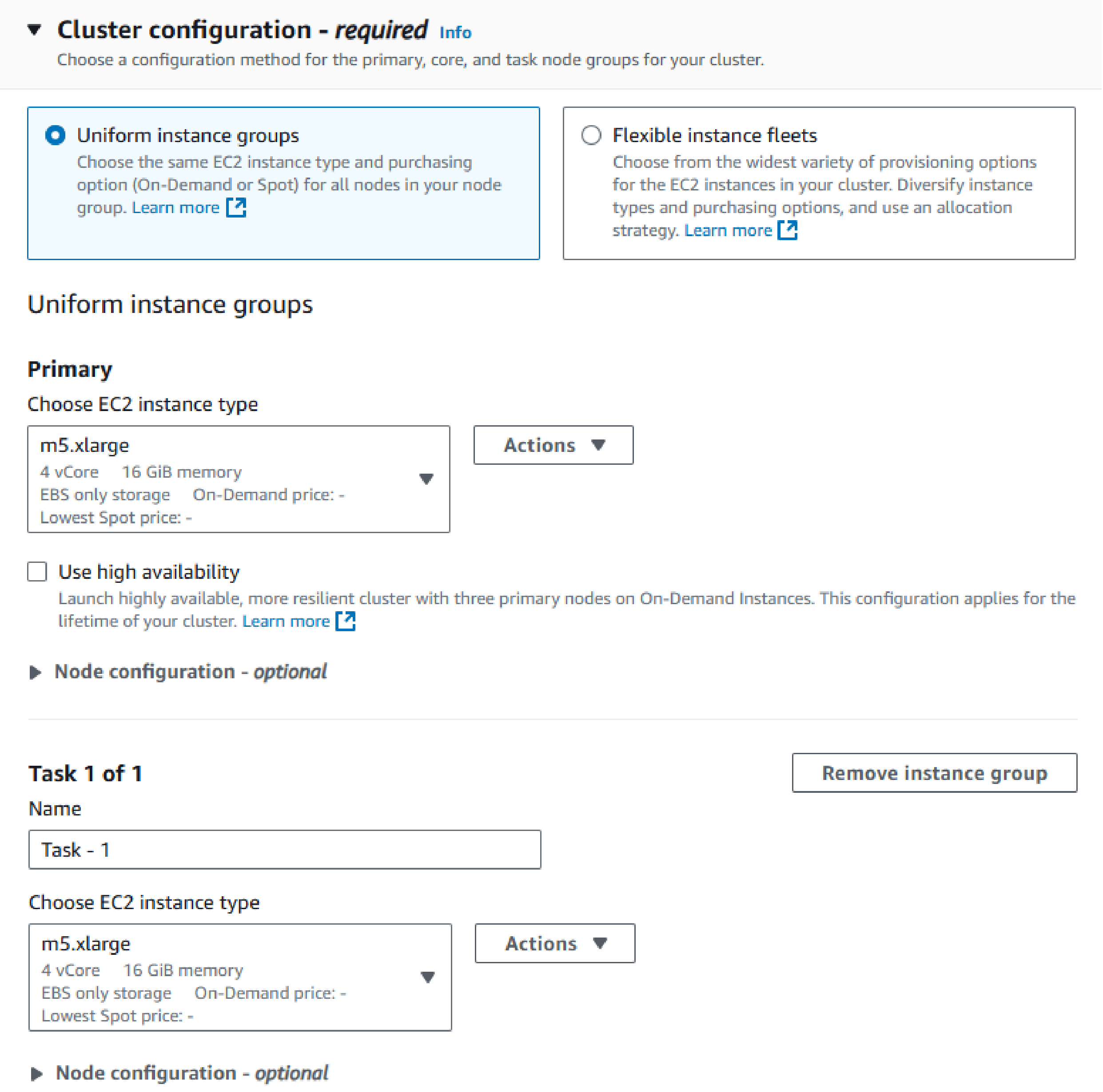
Use the AWS Glue Data Catalog to provide an external metastore for your application.

☑ Use for Hive table metadata
☑ Use for Spark table metadata

Operating system options Info

● Amazon Linux release
○ Custom Amazon Machine Image (AMI)

☑ Automatically apply latest Amazon Linux updates



▼ **Cluster configuration - *required*** Info

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

● Uniform instance groups
Choose the same EC2 instance type and purchasing option (On-Demand or Spot) for all nodes in your node group. Learn more

○ Flexible instance fleets
Choose from the widest variety of provisioning options for the EC2 instances in your cluster. Diversify instance types and purchasing options, and use an allocation strategy. Learn more

Uniform instance groups

Primary

Choose EC2 instance type

m5.xlarge
4 vCore 16 GiB memory
EBS only storage On-Demand price: -
Lowest Spot price: -

□ Use high availability
Launch highly available, more resilient cluster with three primary nodes on On-Demand Instances. This configuration applies for the lifetime of your cluster. Learn more

▶ **Node configuration - *optional***

Task 1 of 1

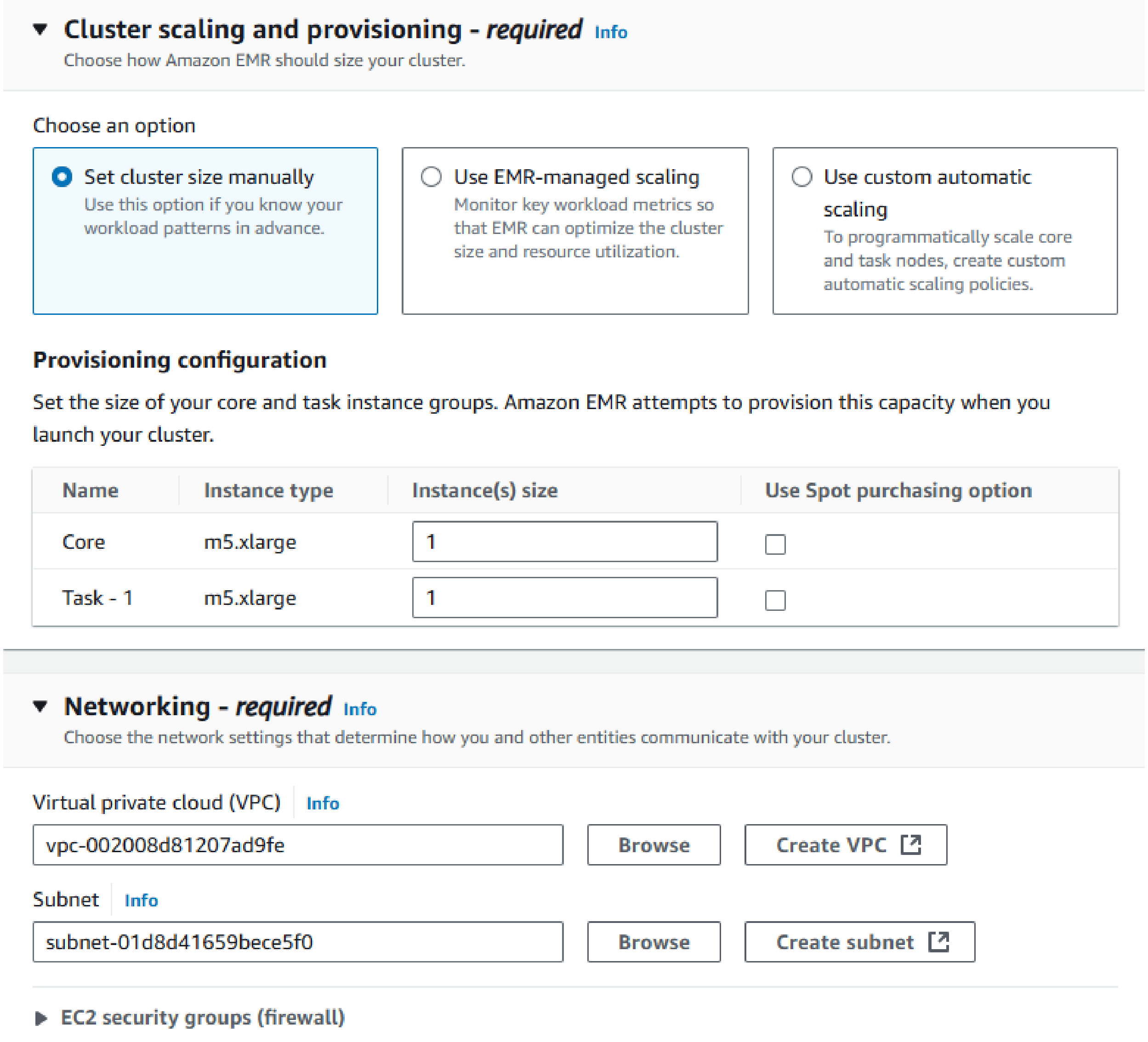
Name

Task - 1

Choose EC2 instance type

m5.xlarge
4 vCore 16 GiB memory
EBS only storage On-Demand price: -
Lowest Spot price: -

▶ **Node configuration - *optional***



▼ **Cluster scaling and provisioning - *required*** Info

Choose how Amazon EMR should size your cluster.

Choose an option

● 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.

○ Use custom automatic scaling
To programmatically scale core and task nodes, create custom automatic scaling policies.

Provisioning configuration

Set the size of your core and task instance groups. Amazon EMR attempts to provision this capacity when you launch your cluster.

Name	Instance type	Instance(s) size	Use Spot purchasing option
Core	m5.xlarge	1	□
Task - 1	m5.xlarge	1	□

▼ **Networking - *required*** Info

Choose the network settings that determine how you and other entities communicate with your cluster.

Virtual private cloud (VPC) Info

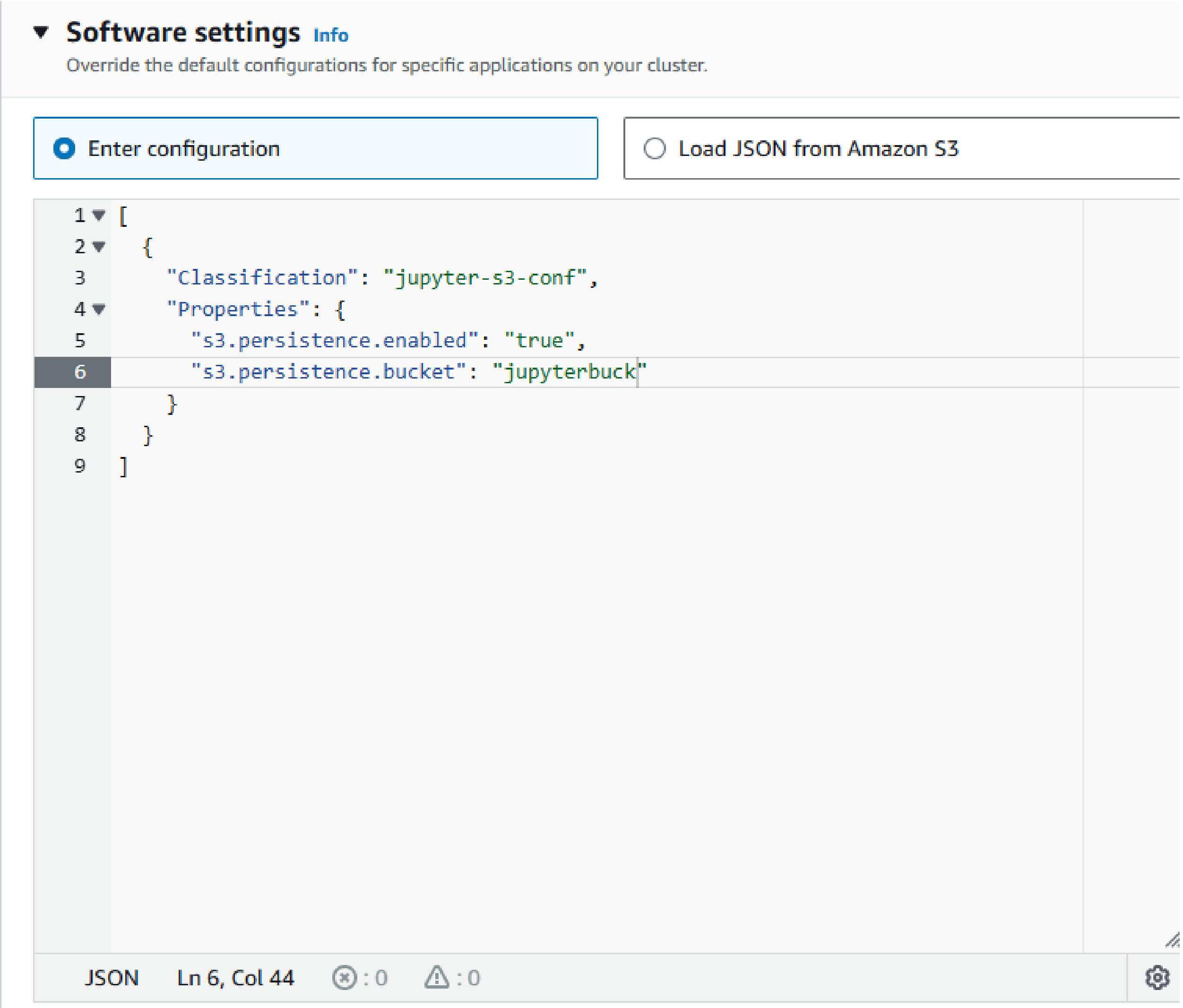
vpc-002008d81207ad9fe

Subnet Info

subnet-01d8d41659bece5f0

▶ EC2 security groups (firewall)

3. Like in this case the bucket name in s3 is called “jupyterbuck” the configuration of the software settings will be:



Override the default configurations for specific applications on your cluster.

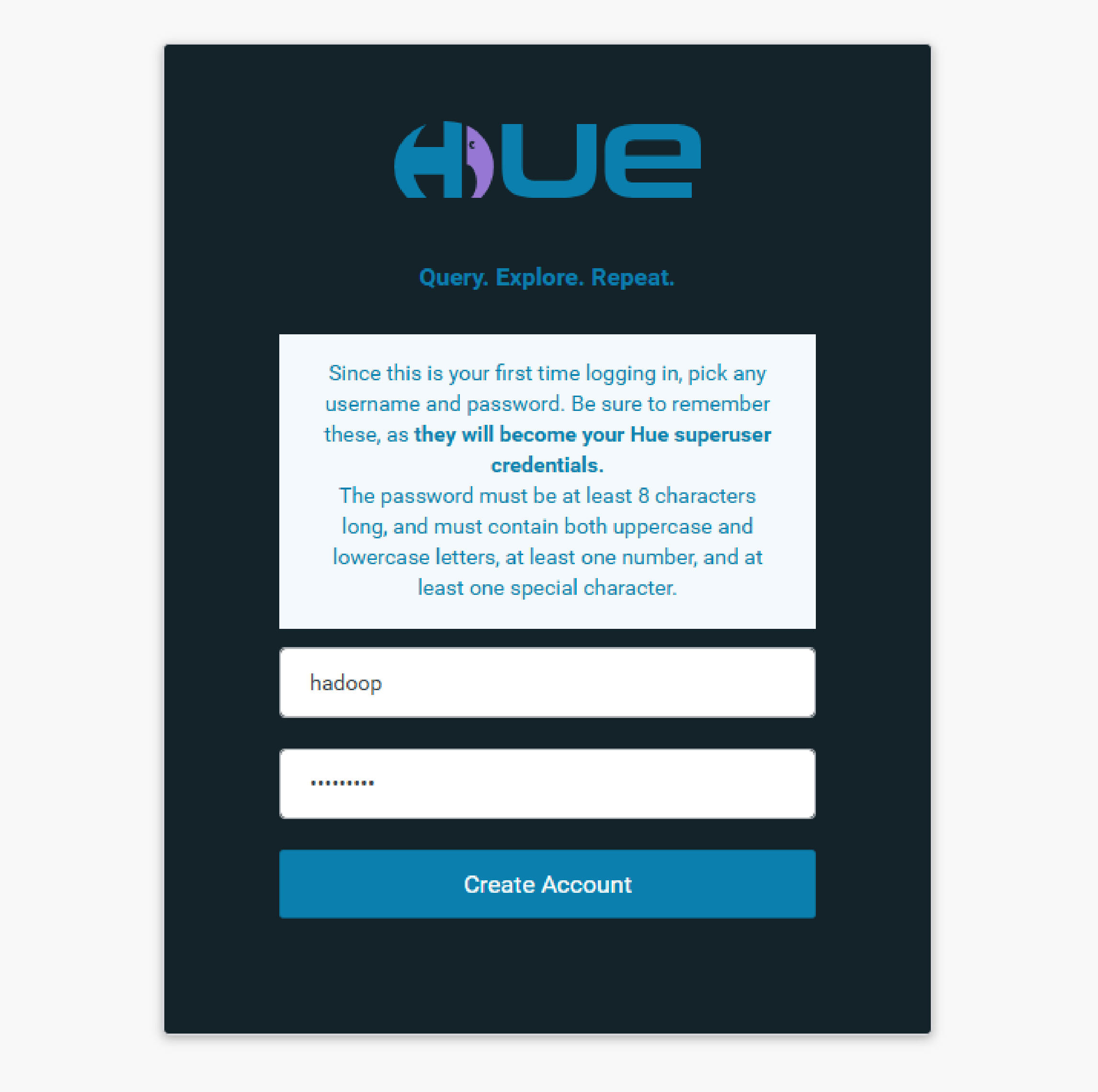
● Enter configuration   ○ Load JSON from Amazon S3

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

After that create the cluster.

4. Find the master node and in the security group add TCP:
   - 22
   - 14000
   - 8888

5. Open the hue application of the cluster and enter using the user “hadoop” and choose your own password, remember also open the port hue is running, you can see it in the emr applications:



Query. Explore. Repeat.

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.

hadoop

•••••••••

6. Now open the ec2 and search the main node of the cluster, you can also see it name into the emr cluster information, and into it install github using yum running the command:

```
sudo yum install git -y
```

7. Now clone the repository and enter to the Hadoop user and run the commands:

```
git clone https://github.com/st0263eafit/st0263-242.git
hdfs dfs -ls /
hdfs dfs -ls /user
hdfs dfs -ls /user/hadoop
hdfs dfs -ls /user/hadoop/datasets
```

If the folders does not exist, then run and upload the files of the folder *bigdata/datasets/gutenberg-small*:

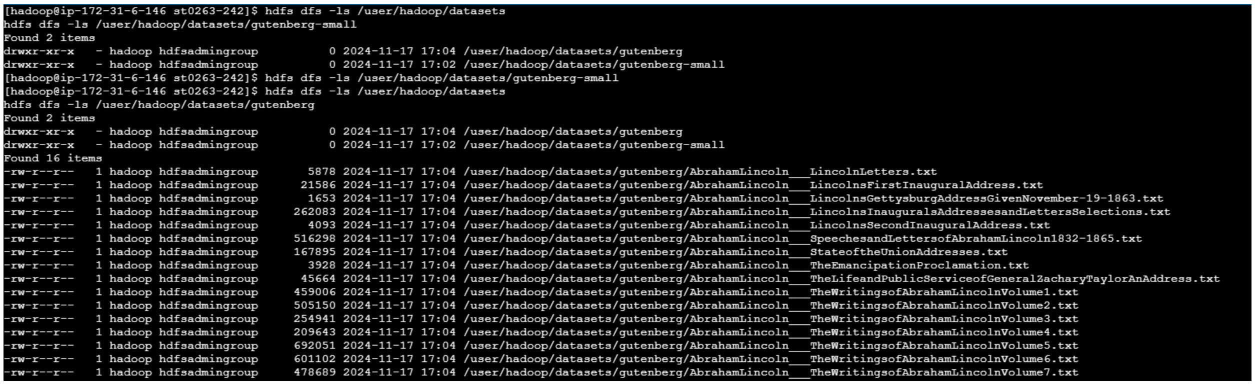
```
hdfs dfs -mkdir /user/hadoop/datasets
hdfs dfs -mkdir /user/hadoop/datasets/Gutenberg

hdfs dfs -put  st0263-242/bigdata/datasets/gutenberg-small/*.txt
/user/hadoop/datasets/gutenberg/
```

8. Now to check the existence of the files, run:

```
hdfs dfs -ls /user/hadoop/datasets
hdfs dfs -ls /user/hadoop/datasets/gutenberg-small
```

You will see something like this:

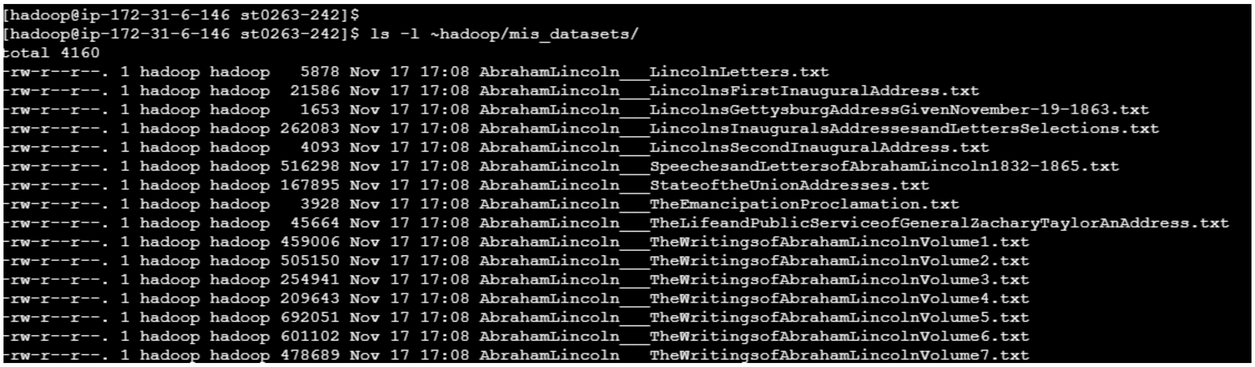


```
[hadoop@ip-172-31-6-146 st0263-242]$ hdfs dfs -ls /user/hadoop/datasets
hdfs dfs -ls /user/hadoop/datasets/gutenberg-small
Found 2 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2024-11-17 17:04 /user/hadoop/datasets/gutenberg
drwxr-xr-x   - hadoop hdfsadmingroup          0 2024-11-17 17:02 /user/hadoop/datasets/gutenberg-small
[hadoop@ip-172-31-6-146 st0263-242]$ hdfs dfs -ls /user/hadoop/datasets/gutenberg-small
[hadoop@ip-172-31-6-146 st0263-242]$ hdfs dfs -ls /user/hadoop/datasets
hdfs dfs -ls /user/hadoop/datasets/gutenberg
Found 2 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2024-11-17 17:04 /user/hadoop/datasets/gutenberg
drwxr-xr-x   - hadoop hdfsadmingroup          0 2024-11-17 17:02 /user/hadoop/datasets/gutenberg-small
Found 16 items
-rw-r--r--   1 hadoop hdfsadmingroup       5878 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___LincolnLetters.txt
-rw-r--r--   1 hadoop hdfsadmingroup      21586 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___LincolnsFirstInauguralAddress.txt
-rw-r--r--   1 hadoop hdfsadmingroup       1653 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___LincolnsGettysburgAddressGivenNovember-19-1863.txt
-rw-r--r--   1 hadoop hdfsadmingroup     262083 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___LincolnsInauguralsAddressesandLettersSelections.txt
-rw-r--r--   1 hadoop hdfsadmingroup       4093 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___LincolnsSecondInauguralAddress.txt
-rw-r--r--   1 hadoop hdfsadmingroup     516298 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___SpeechesandLettersofAbrahamLincoln1832-1865.txt
-rw-r--r--   1 hadoop hdfsadmingroup     167895 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___StateoftheUnionAddresses.txt
-rw-r--r--   1 hadoop hdfsadmingroup       3928 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___TheEmancipationProclamation.txt
-rw-r--r--   1 hadoop hdfsadmingroup      45664 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___TheLifeandPublicServiceofGeneralZacharyTaylorAnAddress.txt
-rw-r--r--   1 hadoop hdfsadmingroup     459006 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume1.txt
-rw-r--r--   1 hadoop hdfsadmingroup     505150 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume2.txt
-rw-r--r--   1 hadoop hdfsadmingroup     254941 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume3.txt
-rw-r--r--   1 hadoop hdfsadmingroup     209643 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume4.txt
-rw-r--r--   1 hadoop hdfsadmingroup     692051 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume5.txt
-rw-r--r--   1 hadoop hdfsadmingroup     601102 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume6.txt
-rw-r--r--   1 hadoop hdfsadmingroup     478689 2024-11-17 17:04 /user/hadoop/datasets/gutenberg/AbrahamLincoln___TheWritingsofAbrahamLincolnVolume7.txt
```

9. To upload HDFS files, first create the folder you will save the files, in my case, it is called "mis_databases" and copy the files using get.

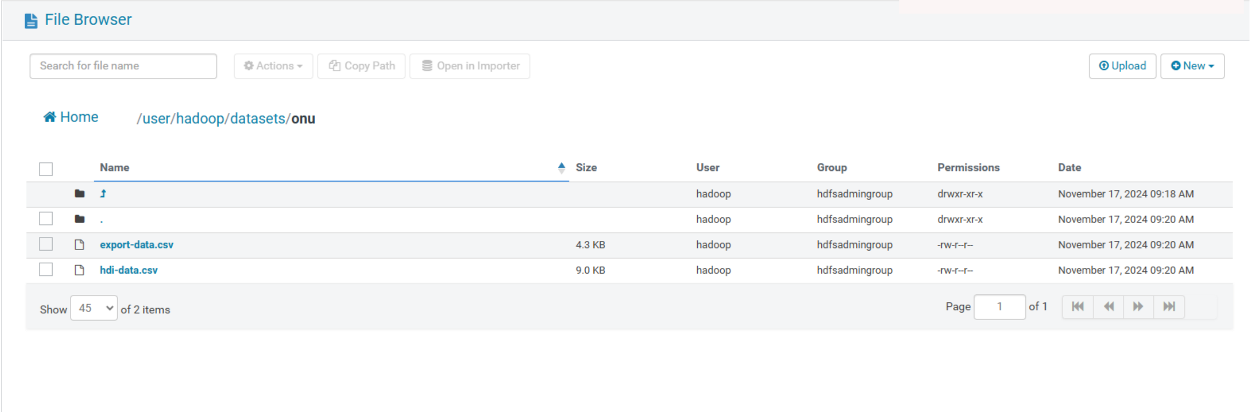
```
hdfs dfs -get /user/hadoop/datasets/gutenberg/*.txt ~hadoop/mis_datasets/
ls -l  ~hadoop/mis_datasets/
```

Your console will look like:



```
[hadoop@ip-172-31-6-146 st0263-242]$
[hadoop@ip-172-31-6-146 st0263-242]$ ls -l ~hadoop/mis_datasets/
total 4160
-rw-r--r--. 1 hadoop hadoop   5878 Nov 17 17:08 AbrahamLincoln___LincolnLetters.txt
-rw-r--r--. 1 hadoop hadoop  21586 Nov 17 17:08 AbrahamLincoln___LincolnsFirstInauguralAddress.txt
-rw-r--r--. 1 hadoop hadoop   1653 Nov 17 17:08 AbrahamLincoln___LincolnsGettysburgAddressGivenNovember-19-1863.txt
-rw-r--r--. 1 hadoop hadoop 262083 Nov 17 17:08 AbrahamLincoln___LincolnsInauguralsAddressesandLettersSelections.txt
-rw-r--r--. 1 hadoop hadoop   4093 Nov 17 17:08 AbrahamLincoln___LincolnsSecondInauguralAddress.txt
-rw-r--r--. 1 hadoop hadoop 516298 Nov 17 17:08 AbrahamLincoln___SpeechesandLettersofAbrahamLincoln1832-1865.txt
-rw-r--r--. 1 hadoop hadoop 167895 Nov 17 17:08 AbrahamLincoln___StateoftheUnionAddresses.txt
-rw-r--r--. 1 hadoop hadoop   3928 Nov 17 17:08 AbrahamLincoln___TheEmancipationProclamation.txt
-rw-r--r--. 1 hadoop hadoop  45664 Nov 17 17:08 AbrahamLincoln___TheLifeandPublicServiceofGeneralZacharyTaylorAnAddress.txt
-rw-r--r--. 1 hadoop hadoop 459006 Nov 17 17:08 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume1.txt
-rw-r--r--. 1 hadoop hadoop 505150 Nov 17 17:08 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume2.txt
-rw-r--r--. 1 hadoop hadoop 254941 Nov 17 17:08 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume3.txt
-rw-r--r--. 1 hadoop hadoop 209643 Nov 17 17:08 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume4.txt
-rw-r--r--. 1 hadoop hadoop 692051 Nov 17 17:08 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume5.txt
-rw-r--r--. 1 hadoop hadoop 601102 Nov 17 17:08 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume6.txt
-rw-r--r--. 1 hadoop hadoop 478689 Nov 17 17:08 AbrahamLincoln___TheWritingsofAbrahamLincolnVolume7.txt
```

10. Now you can create and upload files into hue, just go to files, select the new icon and you can choose between the creation of a file or a folder, for example this is how the create of the 'onu' folder should look like:



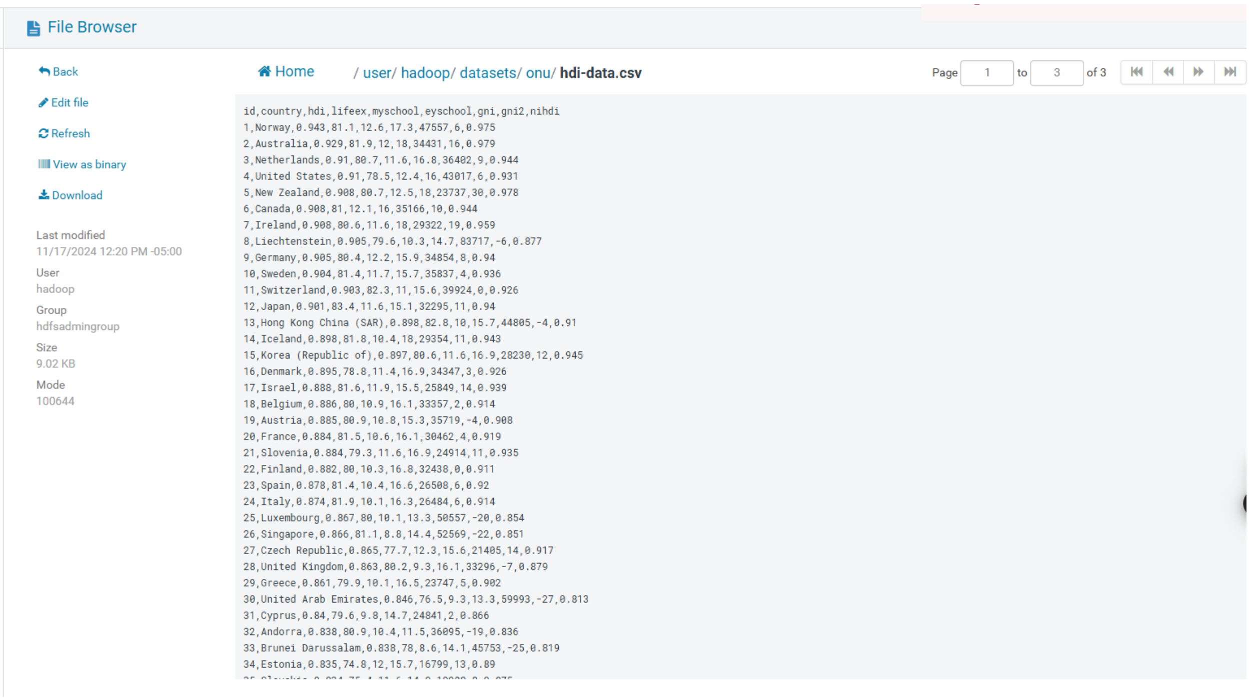
/user/hadoop/datasets/**onu**

Name	Size	User	Group	Permissions	Date
↑		hadoop	hdfsadmingroup	drwxr-xr-x	November 17, 2024 09:18 AM
.		hadoop	hdfsadmingroup	drwxr-xr-x	November 17, 2024 09:20 AM
export-data.csv	4.3 KB	hadoop	hdfsadmingroup	-rw-r--r--	November 17, 2024 09:20 AM
hdi-data.csv	9.0 KB	hadoop	hdfsadmingroup	-rw-r--r--	November 17, 2024 09:20 AM

Show 45 of 2 items

Page 1 of 1

11. Now, you can upload files, navigate to your destination path and choose the upload button, and select the files to upload, in this case the file was called ‘hdi-data.csv’.



/ user/ hadoop/ datasets/ onu/ **hdi-data.csv**

Last modified
11/17/2024 12:20 PM -05:00

User
hadoop

Group
hdfsadmingroup

Size
9.02 KB

Mode
100644

```
id,country,hdi,lifeex,myschool,eyschool,gni,gni2,nihdi
1,Norway,0.943,81.1,12.6,17.3,47557,6,0.975
2,Australia,0.929,81.9,12,18,34431,16,0.979
3,Netherlands,0.91,80.7,11.6,16.8,36402,9,0.944
4,United States,0.91,78.5,12.4,16,43017,6,0.931
5,New Zealand,0.908,80.7,12.5,18,23737,30,0.978
6,Canada,0.908,81,12.1,16,35166,10,0.944
7,Ireland,0.908,80.6,11.6,18,29322,19,0.959
8,Liechtenstein,0.905,79.6,10.3,14.7,83717,-6,0.877
9,Germany,0.905,80.4,12.2,15.9,34854,8,0.94
10,Sweden,0.904,81.4,11.7,15.7,35837,4,0.936
11,Switzerland,0.903,82.3,11,15.6,39924,0,0.926
12,Japan,0.901,83.4,11.6,15.1,32295,11,0.94
13,Hong Kong China (SAR),0.898,82.8,10,15.7,44805,-4,0.91
14,Iceland,0.898,81.8,10.4,18,29354,11,0.943
15,Korea (Republic of),0.897,80.6,11.6,16.9,28230,12,0.945
16,Denmark,0.895,78.8,11.4,16.9,34347,3,0.926
17,Israel,0.888,81.6,11.9,15.5,25849,14,0.939
18,Belgium,0.886,80,10.9,16.1,33357,2,0.914
19,Austria,0.885,80.9,10.8,15.3,35719,-4,0.908
20,France,0.884,81.5,10.6,16.1,30462,4,0.919
21,Slovenia,0.884,79.3,11.6,16.9,24914,11,0.935
22,Finland,0.882,80,10.3,16.8,32438,0,0.911
23,Spain,0.878,81.4,10.4,16.6,26508,6,0.92
24,Italy,0.874,81.9,10.1,16.3,26484,6,0.914
25,Luxembourg,0.867,80,10.1,13.3,50557,-20,0.854
26,Singapore,0.866,81.1,8.8,14.4,52569,-22,0.851
27,Czech Republic,0.865,77.7,12.3,15.6,21405,14,0.917
28,United Kingdom,0.863,80.2,9.3,16.1,33296,-7,0.879
29,Greece,0.861,79.9,10.1,16.5,23747,5,0.902
30,United Arab Emirates,0.846,76.5,9.3,13.3,59993,-27,0.813
31,Cyprus,0.84,79.6,9.8,14.7,24841,2,0.866
32,Andorra,0.838,80.9,10.4,11.5,36095,-19,0.836
33,Brunei Darussalam,0.838,78,8.6,14.1,45753,-25,0.819
34,Estonia,0.835,74.8,12,15.7,16799,13,0.89
```