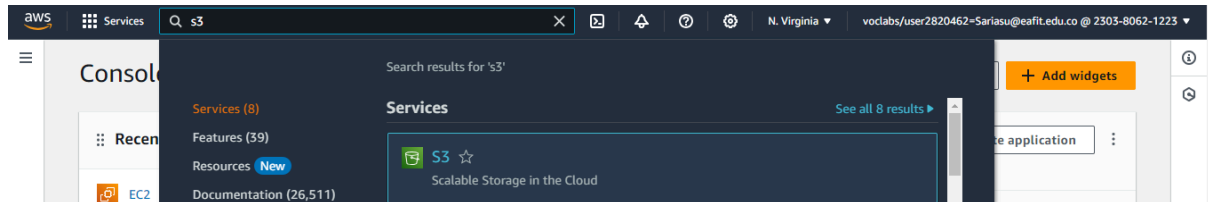


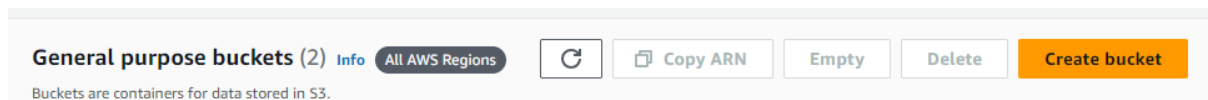
Sebastián Arias Usma.
C.C 1017932811
Evidencias laboratorios 3 big data
st0263-241 Tópicos de telematica

Reto # 5: Laboratorio EMR y Laboratorio de archivos por HDFS

Paso #1: Debemos crear un bucket S3 en amazon por lo que debemos buscar en el aws academy este espacio de la siguiente manera.



Paso #2: Debemos crear nuestro bucket de s3 para el reto #5 el cual luego conectaremos con nuestro EMR, seleccionamos el botón de Create bucket.



Paso#3: Debemos configurar nuestro bucket de la siguiente manera.

- Seleccionamos el tipo de bucket

Create bucket [Info](#)

Buckets are containers for data stored in S3.

General configuration

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

Bucket type [Info](#)

☒ **General purpose**
Recommended for most use cases and access patterns. General purpose buckets are the original S3 bucket type. They allow a mix of storage classes that redundantly store objects across multiple Availability Zones.

☐ **Directory - New**
Recommended for low-latency use cases. These buckets use only the S3 Express One Zone storage class, which provides faster processing of data within a single Availability Zone.

Bucket name [Info](#)

Bucket name must be unique within the global namespace and follow the bucket naming rules. [See rules for bucket naming](#)

Copy settings from existing bucket - *optional*
Only the bucket settings in the following configuration are copied.

Format: s3://bucket/prefix

- Seleccionamos la manera en como se va a comportar el bucket.

Object Ownership [Info](#)

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

- Permitimos que sea de acceso publico.

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.


Turning off block all public access might result in this bucket and the objects within becoming public
AWS recommends that you turn on block all public access, unless public access is required for specific and verified use cases such as static website hosting.

☒ I acknowledge that the current settings might result in this bucket and the objects within becoming public.

- Dejamos estas configuraciones por defecto.

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 - optional (0)

You can use bucket tags to track storage costs and organize buckets. [Learn more](#)

No tags associated with this bucket.

Add tag

Default encryption [Info](#)

Server-side encryption is automatically applied to new objects stored in this bucket.

Encryption type [Info](#)

☒ Server-side encryption with Amazon S3 managed keys (SSE-S3)

☐ Server-side encryption with AWS Key Management Service keys (SSE-KMS)

☐ Dual-layer server-side encryption with AWS Key Management Service keys (DSSE-KMS)

Secure your objects with two separate layers of encryption. For details on pricing, see [DSSE-KMS pricing](#) on the **Storage** tab of the [Amazon S3 pricing page](#).

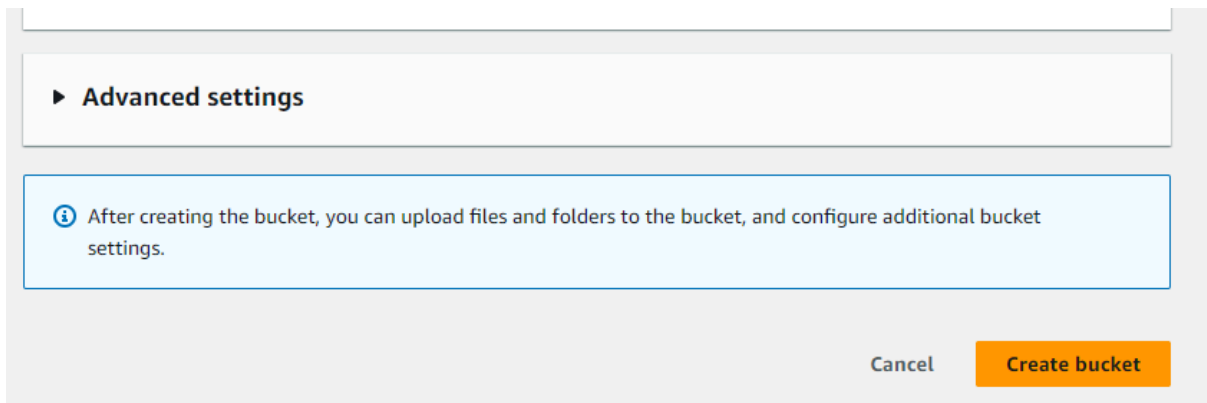
Bucket Key

Using an S3 Bucket Key for SSE-KMS reduces encryption costs by lowering calls to AWS KMS. S3 Bucket Keys aren't supported for DSSE-KMS. [Learn more](#)

☐ Disable

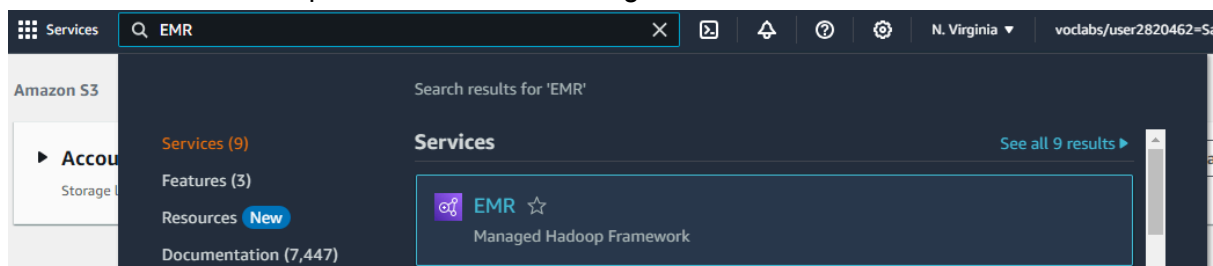
☒ Enable

- Y ahora creamos el nuestro s3 buket.

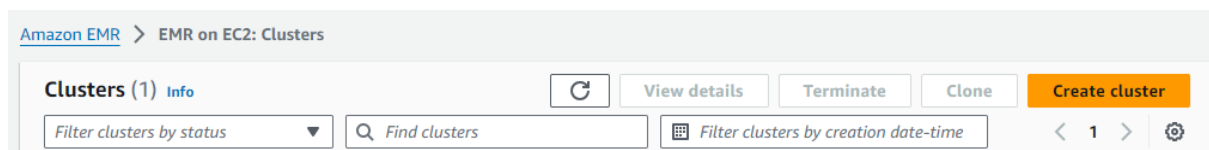


Paso #4: Ahora debemos ir al apartado de EMR en nuestro AWS y crear uno de la siguiente manera.

- Accedemos al apartado de EMR de la siguiente manera.



- Ahora vamos a darle click al botón de crear cluster.



- Crearemos el cluster con la siguiente configuración.
- Seleccionamos que version de EMR deseamos, en este caso la 6.15.0 y las aplicaciones que necesitaremos en el EMR

Name

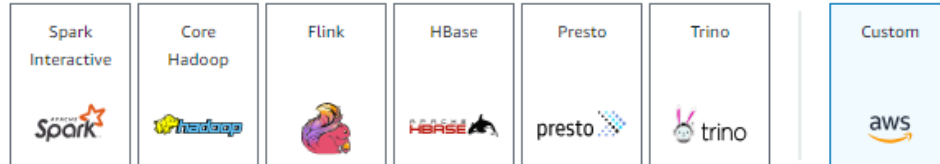
EMR-retos

Amazon EMR release [Info](#)

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

emr-6.15.0

Application bundle



- | | | |
|--|--|--|
| <input checked="" type="checkbox"/> Flink 1.17.1 | <input type="checkbox"/> Ganglia 3.7.2 | <input type="checkbox"/> HBase 2.4.17 |
| <input checked="" type="checkbox"/> HCatalog 3.1.3 | <input checked="" type="checkbox"/> Hadoop 3.3.6 | <input checked="" type="checkbox"/> Hive 3.1.3 |
| <input checked="" type="checkbox"/> Hue 4.11.0 | <input checked="" type="checkbox"/> JupyterEnterpriseGateway 2.6.0 | <input checked="" type="checkbox"/> JupyterHub 1.5.0 |
| <input checked="" type="checkbox"/> Livy 0.7.1 | <input type="checkbox"/> MXNet 1.9.1 | <input type="checkbox"/> Oozie 5.2.1 |
| <input type="checkbox"/> Phoenix 5.1.3 | <input type="checkbox"/> Pig 0.17.0 | <input type="checkbox"/> Presto 0.283 |
| <input checked="" type="checkbox"/> Spark 3.4.1 | <input checked="" type="checkbox"/> Sqoop 1.4.7 | <input type="checkbox"/> TensorFlow 2.11.0 |
| <input checked="" type="checkbox"/> Tez 0.10.2 | <input type="checkbox"/> Trino 426 | <input checked="" type="checkbox"/> Zeppelin 0.10.1 |
| <input checked="" type="checkbox"/> ZooKeeper 3.5.10 | | |

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

- Ahora configuraremos los atributos del cluster EMR de la siguiente manera

▼ 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: -

Actions ▼

- ☐ **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

Core

Choose EC2 instance type

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

Actions ▼

► Node configuration - *optional*

Task 1 of 1

Remove instance group

Name

Task - 1

Choose EC2 instance type

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

Actions ▼

► Node configuration - *optional*

Add task instance group

You can add up to 47 more task instance groups.

EBS root volume

EBS root volume applies to the operating systems and applications that you install on the cluster. [EBS root volume ratio constraints](#)

Size (GiB)

15

15 - 100 GiB per volume
General Purpose SSD (gp3)

IOPS

3000

3000 - 16000 IOPS per volume.
Choose a maximum ratio of
500:1 between IOPS and
volume size.

Throughput (MiB/s)

125

125 - 1000 MiB/s per volume.
Choose a maximum ratio of
0.25:1 between throughput
and IOPS.

- Ahora ajustaremos la escalabilidad de nuestro cluster EMR de la siguiente manera.

▼ 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	<input type="text" value="1"/>	<input type="checkbox"/>
Task - 1	m5.xlarge	<input type="text" value="1"/>	<input type="checkbox"/>


- Ahora configuraremos los volúmenes de redes o las VP, VPC que necesitamos

▼ Networking - *required* [Info](#)

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


Virtual private cloud (VPC) [Info](#)

[Browse](#)

[Create VPC](#) 

Subnet [Info](#)

[Browse](#)

[Create subnet](#) 

► **EC2 security groups (firewall)**

- Determinaremos la manera en la que dejara de trabajar nuestro cluster que sera de manera manual.

▼ Cluster termination and node replacement [Info](#)

Choose termination settings and protect your cluster from accidental shutdown.

Termination option

- ☒ Manually terminate cluster
- ☐ Automatically terminate cluster after last step ends
- ☐ Automatically terminate cluster after idle time (Recommended)

☐ Use termination protection

Protects your cluster from accidental termination. If on, you must first turn off protection to terminate the cluster. We recommend turning on termination protection for your long running clusters.

Unhealthy node replacement - *new* [Info](#)

- ☒ Turn on
Amazon EMR gracefully stops processes on unhealthy nodes to minimize data loss and job interruptions. It quickly replaces unhealthy nodes with new EC2 instances to keep your jobs running smoothly.
- ☐ Turn off
Amazon EMR adds unhealthy nodes to a denylist while keeping them in the cluster, allowing you continued access for troubleshooting.

- Luego debemos configurar el EMS en nuestro aws con este código en el apartado de configurar Software Settings para conectarlo con nuestro bucket s3.

▼ Software settings [Info](#)

Override the default configurations for specific applications on your cluster.

☒ Enter configuration

☐ Load JSON from Amazon S3

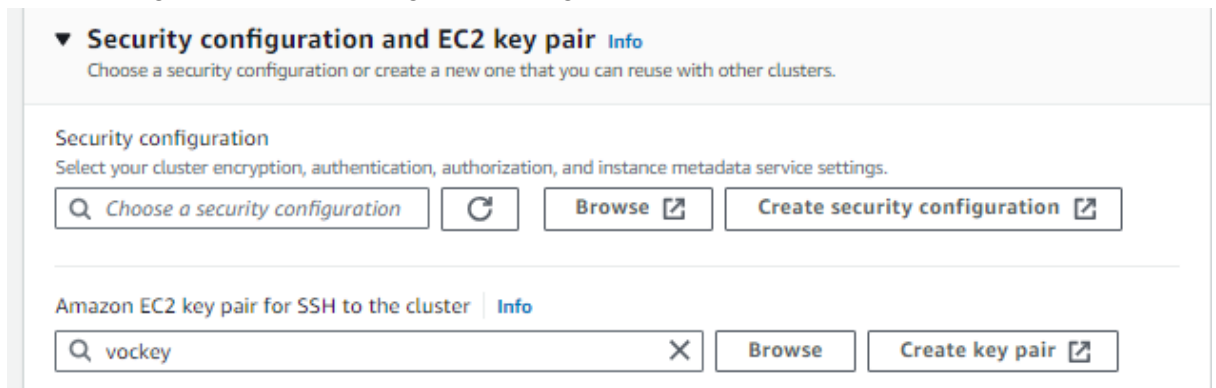
```
1 [
2   {
3     "classification": "jupyter-s3-conf",
4     "properties": {
5       "s3.persistance.bucket": "reto5-bucket",
6       "s3.persistance.enabled": "true"
7     }
8   }
9 ]
```

JSON Ln 1, Col 1




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

- Ahora configuraremos nuestro grupo de seguridad.



▼ Security configuration and EC2 key pair [Info](#)
Choose a security configuration or create a new one that you can reuse with other clusters.

Security configuration
Select your cluster encryption, authentication, authorization, and instance metadata service settings.

Amazon EC2 key pair for SSH to the cluster [Info](#)

- Y por ultimo dejamos que nuestro cluster cree las instancias necesarias para acceder a nuestro EMR.

▼ Identity and Access Management (IAM) roles - *required* [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 ▼

↺

Custom automatic scaling role - *optional*
When a custom automatic scaling rule triggers, Amazon EMR assumes this role to add and terminate EC2 instances. [Learn more](#) [↗](#)

Custom automatic scaling role

LabRole ▼

↺

Create IAM role [↗](#)

- Con estos pasos ya hemos creado nuestro EMR cluster.

Luego modificamos el security group de nuestro EMR desde el E2C para obtener las Ips públicas de nuestro HUE, debemos agregar estos security groups.

- 8443
- 9870
- 8888
- 14000
- 9443
- 8890

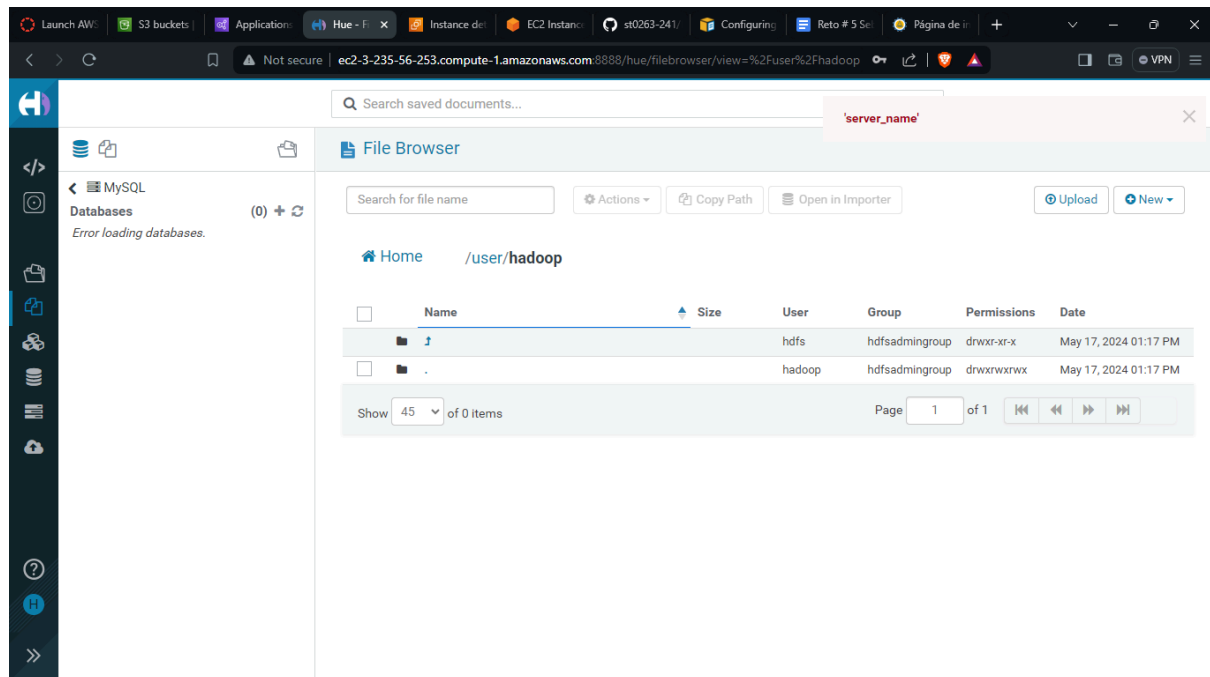
Todas estas con ipv4

Paso # 5:

- Ahora accederemos a nuestro EMR desde nuestro SSH y haremos los siguientes comandos para poder acceder a nuestro HUE y subir nuestros archivos
- Cada vez que clonemos el cluster debemos copiar estos comandos en nuestro bash de nuestro EMR

```
[root@ip-172-31-73-11 ~]# sed -i 's/.ec2.internal:14000/.ec2.internal:9870/' /etc/hue/conf/hue.ini
[root@ip-172-31-73-11 ~]# systemctl restart hue.service
```

Ahora revisamos nuestra página de hue y debería salir de esta manera:



Paso #6: Luego hacemos yum update para instalar git y luego de este realizamos el yum install git, para clonar el repositorio del reto.

- git clone <https://github.com/st0263eafit/st0263-241.git>

Y cuando tenemos clonado el repo realizamos los comandos hdfs para el hadoop

- hdfs dfs -ls /
- hdfs dfs -ls /user
- hdfs dfs -ls /user/hadoop
- hdfs dfs -ls /user/hadoop/datasets

al ejecutar el ultimo comando saldra error ya que no tenemos la carpeta datasets por lo que debemos crearla con el siguiente comando

- hdfs dfs -mkdir /user/hadoop/datasets














Cuando creamos esta carpeta crearemos la primera carpeta de datasets la cual se llamara gutenber-small.

- hdfs dfs -mkdir /user/hadoop/datasets/gutenber-small

Cuando creamos esta carpeta procederemos a insertar los primeros archivos del dataset con este comando.




- hdfs dfs -put /home/ec2-home/st0263-241/bigdata/datasets/gutenber-small/*.txt /user/hadoop/datasets/gutenber-small/

y tendremos la carpeta creada en nuestro Hue

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 ↑		hadoop	hdfsadmingroup	drwxrwxrwx	May 17, 2024 01:55 PM
<input type="checkbox"/>	 .		root	hdfsadmingroup	drwxr-xr-x	May 17, 2024 01:58 PM
<input type="checkbox"/>	 airlines.csv	761.8 KB	hadoop	hdfsadmingroup	-rw-r--r--	May 17, 2024 01:58 PM
<input type="checkbox"/>	 all-news		hadoop	hdfsadmingroup	drwxr-xr-x	May 17, 2024 01:57 PM
<input type="checkbox"/>	 covid19		hadoop	hdfsadmingroup	drwxr-xr-x	May 17, 2024 01:58 PM
<input type="checkbox"/>	 flights		hadoop	hdfsadmingroup	drwxr-xr-x	May 17, 2024 01:58 PM
<input checked="" type="checkbox"/>	 gutenberg		hadoop	hdfsadmingroup	drwxr-xr-x	May 17, 2024 01:58 PM
<input type="checkbox"/>	 gutenberg-small		root	hdfsadmingroup	drwxr-xr-x	May 17, 2024 01:58 PM
<input type="checkbox"/>	 onu		hadoop	hdfsadmingroup	drwxr-xr-x	May 17, 2024 01:58 PM
<input type="checkbox"/>	 otros		hadoop	hdfsadmingroup	drwxr-xr-x	May 17, 2024 01:58 PM
<input type="checkbox"/>	 retail_logs		hadoop	hdfsadmingroup	drwxr-xr-x	May 17, 2024 01:58 PM
<input type="checkbox"/>	 sample_data.csv	567 bytes	hadoop	hdfsadmingroup	-rw-r--r--	May 17, 2024 01:58 PM
<input type="checkbox"/>	 spark		hadoop	hdfsadmingroup	drwxr-xr-x	May 17, 2024 01:58 PM

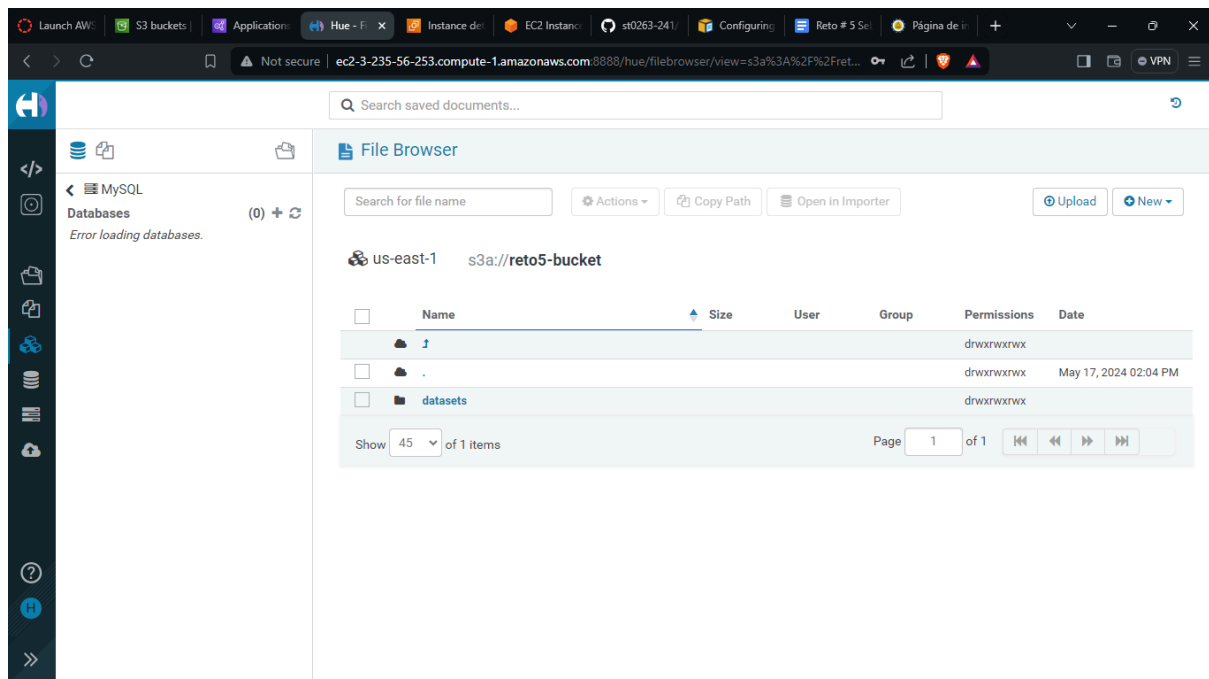
Por ultimo subiremos estas mismas carpetas a nuestro bucket de s3 en nuestro hue el cual deberia salir asi

 us-east-1 s3a://

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 .				drwxrwxrwx	
<input type="checkbox"/>	 aws-logs-230380621223-us-east-1				drwxrwxrwx	
<input type="checkbox"/>	 reto5-bucket				drwxrwxrwx	

Show 45 of 2 items Page 1 of 1

ingresamos a el “reto5-bucket”



Launch AWS S3 buckets Application Hue - F Instance do EC2 Instanc st0263-241 Configuring Reto # 5 Sel Página de ir




Not secure ec2-3-235-56-253.compute-1.amazonaws.com:8888/hue/filebrowser/view=s3a%3A%2F%2Fret... VPN

Search saved documents...

File Browser

Search for file name Actions Copy Path Open in Importer Upload New

us-east-1 s3a://reto5-bucket

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 ↑				drwxrwxrwx	
<input type="checkbox"/>	 .				drwxrwxrwx	May 17, 2024 02:04 PM
<input type="checkbox"/>	 datasets				drwxrwxrwx	

Show 45 of 1 items Page 1 of 1

y subiremos la carpeta de datasets que tambien se subio en el apartado de files para que se vea de la siguiente manera

