

**Laboratorio:** Instalar un clúster EMR versión 6.14.0 Hadoop/Spark

**Fecha:** 2 noviembre 2023

**Parte 1: Crear un clúster AWS EMR versión 6.14.0**

**1 Instalar AWS EMR**

1. Buscar el servicio AWS EMR: Entrar a la consola web de AWS y buscar el servicio EMR:

The screenshot shows the AWS Management Console search results for 'EMR'. The search bar at the top has 'EMR' typed into it. Below the search bar, there is a 'Services' section with a list of services: Features (3), Resources (New), Documentation (9,605), Knowledge Articles (20), Marketplace (182), Blogs (481), Events (12), and Tutorials (1). To the right of this list, the 'EMR' service is highlighted with a red circle. The 'EMR' service card includes the icon, name, star rating, and description: 'Managed Hadoop Framework'. Below the service card, there are two other service cards: 'MediaStore' and 'MediaConvert'. The entire screenshot is framed by a red border.

2. crear clúster

The screenshot shows the 'Amazon EMR' service page. On the left, there is a sidebar with navigation links: 'Amazon EMR', 'EMR Serverless', 'Clusters' (which is highlighted with a red circle), 'Notebooks and Git repos', 'Events', and 'Block public access'. The main content area is titled 'EMR on EC2: Clusters' and shows a table of clusters. The table has columns for 'Cluster ID', 'Cluster name', 'Status', and 'Create'. A red circle highlights the 'Create cluster' button. The table also includes filters for 'Filter clusters by status', 'Find clusters', and 'Filter clusters by creation date-time', along with buttons for 'Starting' and 'Bootstrapping'. The bottom of the table shows pagination with '1' and a refresh icon.

3. nombre, versión y Custom

The screenshot shows the 'Create cluster' wizard in the AWS Management Console. The 'Name and applications' step is active. The 'Name' field contains 'Cluster EMONTOYA', which is circled in red. Below it, the 'Amazon EMR release' dropdown is set to 'emr-6.14.0'. In the 'Application bundle' section, there are several options: Spark Interactive (Apache Spark logo), Core Hadoop (Apache Hadoop logo), Flink (Flink logo), HBase (Apache HBase logo), Presto (Presto logo), Trino (Trino logo), and Custom (AWS logo). The 'Custom' option is also circled in red.

4. seleccionando los paquetes adecuados para el curso y activando los catálogos Glue, Hive, Spark

Nota: seleccionar los catalogos Hive y Spark permite ver las tablas AWS Glue en EMR, y las tablas Hive se podrán ver en Glue / Athena.

Seleccione los paquetes hadoop/Spark en azul.

aws Services Search [Option+S] N. Virginia ▾ voo

Spark Interactive Core Hadoop Flink HBase Presto Trino Custom

Flink 1.17.1     Ganglia 3.7.2  
 HCatalog 3.1.3     Hadoop 3.3.3  
 Hue 4.11.0     JupyterEnterpriseGateway 2.6.0  
 Livy 0.7.1     MXNet 1.9.1  
 Phoenix 5.1.3     Pig 0.17.0  
 Spark 3.4.1     Swoop 1.4.7  
 Tez 0.10.2     Trino 422  
 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

## 5. Máquinas EC2 del Clúster

Puede dejar las máquinas por defecto m5.xlarge, en algunos momentos puede fallar la creación del clúster porque no tiene suficientes recursos, puede cambiar estas máquinas a m4.xlarge, pero por defecto dejarlas como nos sugiere la creación del clúster EMR.

aws Services Search [Option+S] N. Virginia ▾ voo

**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

m5.xlarge

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

Actions ▾

**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**

Name

Remove instance group

Actions ▾

## 6. Dejar estas opciones por defecto

**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                | <input type="checkbox"/>   |
| Task - 1 | m5.xlarge     | 1                | <input type="checkbox"/>   |

**Networking** [Info](#)

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

Subnet [Info](#)  
 [Browse](#) [Create subnet](#)

► EC2 security groups (firewall)

Dejar las siguientes opciones por defecto hasta: Software settings.

## 7. Configurar software settings

Acá va a configurar el bucket para guardar los notebooks jupyter y no se pierdan cuando se borre el clúster EMR.

Realizar una búsqueda sencilla Google: aws emr jupyterhub s3

Nos conduce al enlace: <https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-jupyterhub-s3.html>

Configurar con tu propio bucket (crear un bucket para esto)

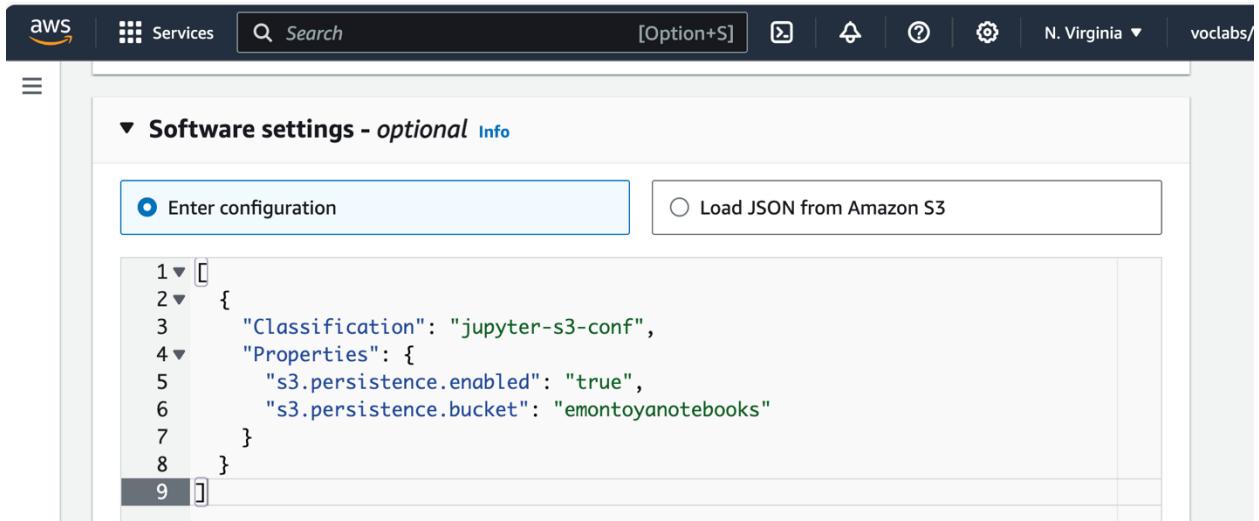
Antes:

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

Con mi bucket:

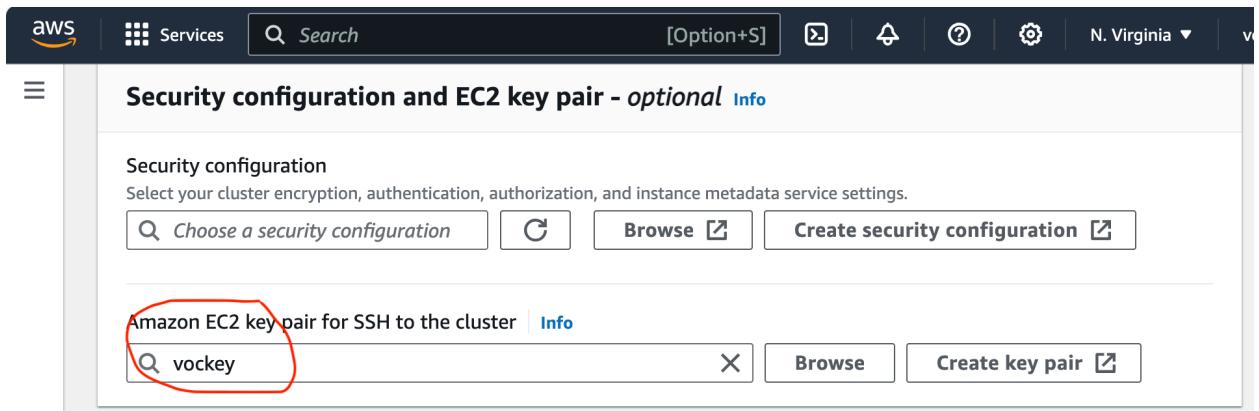
```
[  
  {  
    "Classification": "jupyter-s3-conf",  
    "Properties": {  
      "s3.persistence.enabled": "true",  
      "s3.persistence.bucket": "emontoyanotebooks"  
    }  
  }  
]
```

Y pegue esta configuración en Software Settings así:



```
1 [ ]  
2 {  
3     "Classification": "jupyter-s3-conf",  
4     "Properties": {  
5         "s3.persistence.enabled": "true",  
6         "s3.persistence.bucket": "emontoyanotebooks"  
7     }  
8 }  
9 [ ]
```

## 8. Security configuration and EC2 key pair



Choose a security configuration   Browse  Create security configuration

Amazon EC2 key pair for SSH to the cluster

## 9. IAM roles

Debe seleccionar:

Service role: EMR\_DefaultRole

Instance profile: EMR\_EC2\_DefaultRole

Custom automatic scaling role: LabRole

The screenshot shows the AWS Management Console configuration for an Amazon EMR cluster. The top navigation bar includes the AWS logo, Services menu, Search bar, and N. Virginia region selector.

### 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** (highlighted with a red box)  
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** (highlighted with a red box)  
EMR\_EC2\_DefaultRole

The screenshot shows the continuation of the AWS Management Console configuration for an Amazon EMR cluster.

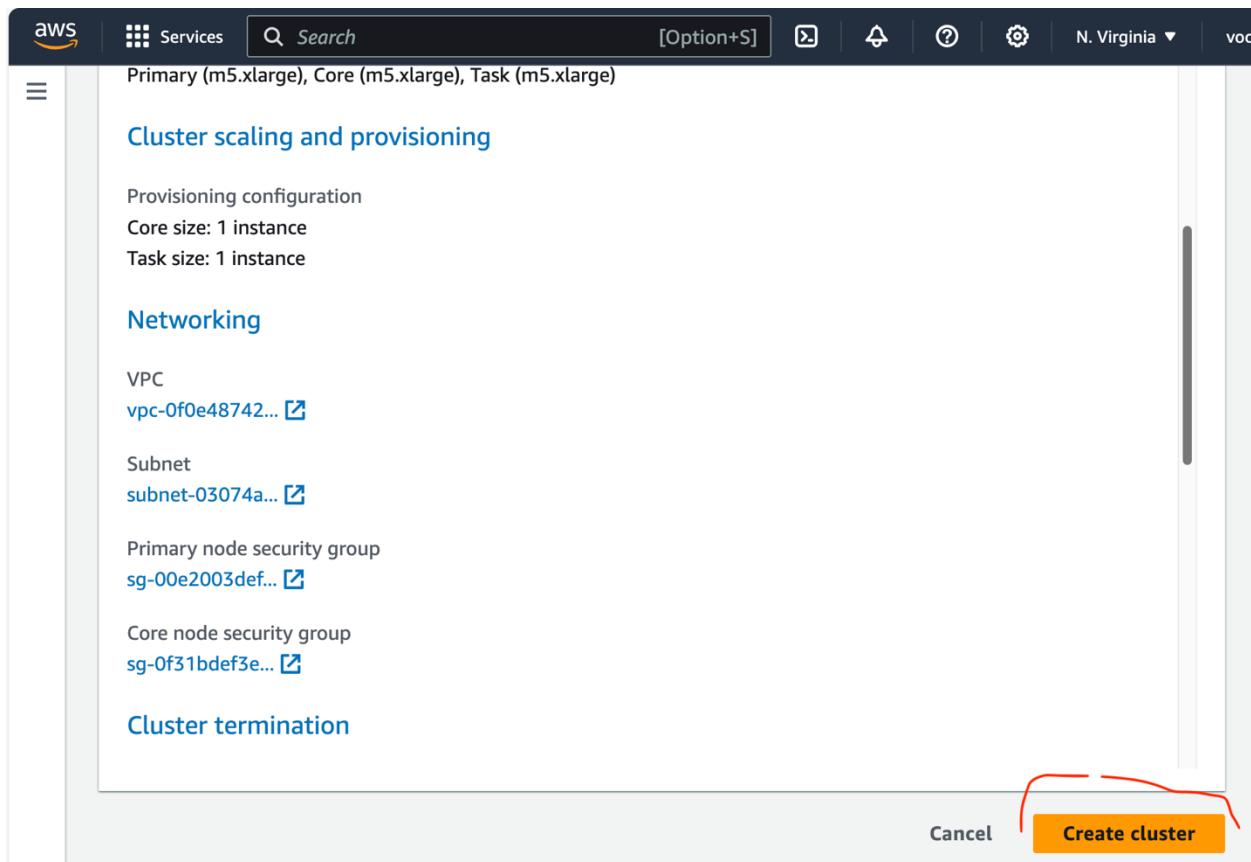
### Custom automatic scaling role - optional

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

**Custom automatic scaling role** (highlighted with a red box)  
LabRole

[Create IAM role](#)

## 10. Finalmente, a crear el clúster



Este proceso demora aproximadamente 20 minutos, tenga paciencia.

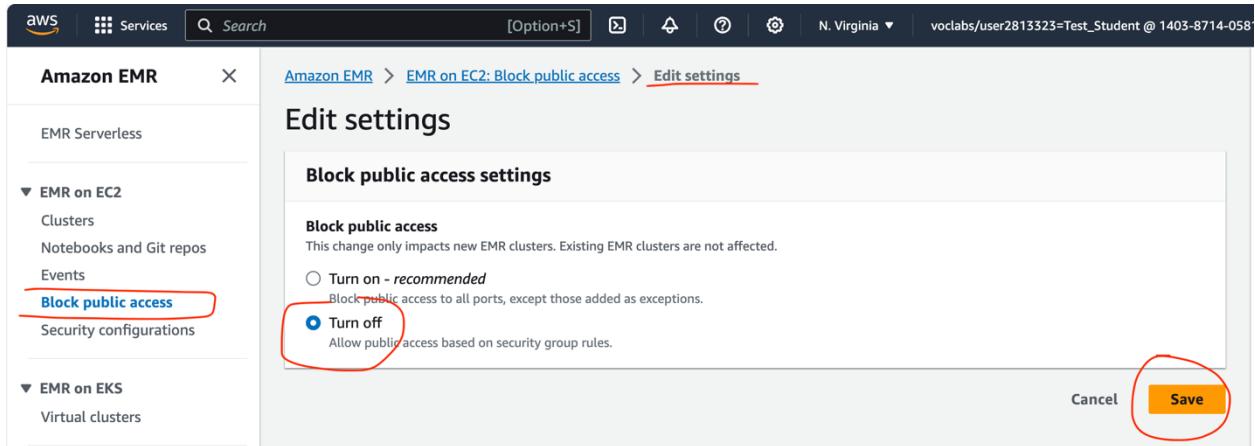
Debe salir con este mensaje de clúster exitosamente creado:

The screenshot shows the 'Amazon EMR > EMR on EC2: Clusters' page. The 'Clusters (2) Info' section displays two clusters: 'Cluster EMONTOYA' (Status: Waiting, Ready to run steps) and 'st1800-emontoya' (Status: Terminated, User request). A red circle highlights the 'Waiting' status of the first cluster.

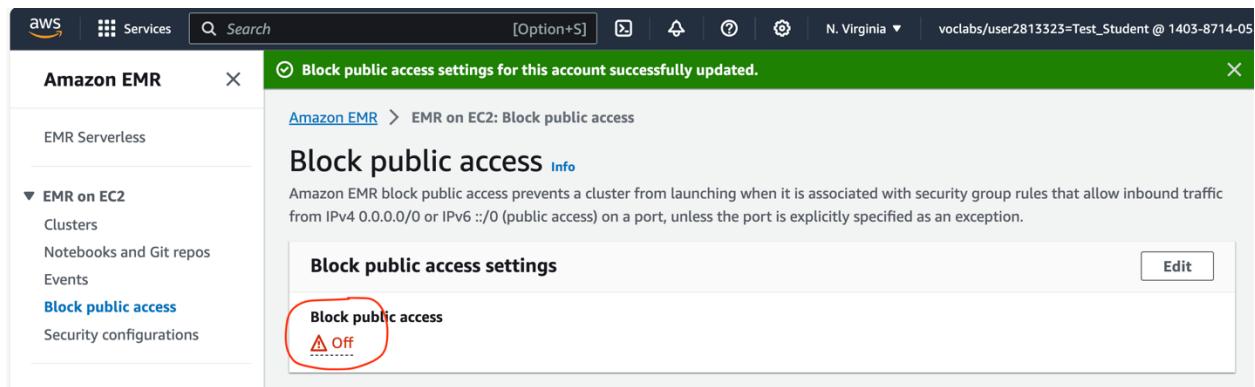
| Cluster ID      | Cluster name     | Status     |
|-----------------|------------------|------------|
| j-3DRPEB3XMBVAV | Cluster EMONTOYA | Waiting    |
| j-J5917MHJFP6H  | st1800-emontoya  | Terminated |

11. Debe abrir todos los puertos TCP para acceso al clúster así

(nota: esto solo se hace una vez, cada vez que crea, destruya o clone un clúster, ya quedan abiertos)



The screenshot shows the AWS Amazon EMR service interface. On the left, there's a sidebar with 'Amazon EMR' at the top, followed by 'EMR Serverless', 'EMR on EC2' (with 'Clusters', 'Notebooks and Git repos', 'Events', and 'Block public access' listed; 'Block public access' is highlighted with a red box), and 'EMR on EKS' (with 'Virtual clusters'). The main content area has a breadcrumb path: 'Amazon EMR > EMR on EC2: Block public access > Edit settings'. The title 'Edit settings' is above a section titled 'Block public access settings'. Under this, there's a 'Block public access' section with two options: 'Turn on - recommended' (unchecked) and 'Turn off' (checked). Below the checked option is the text 'Allow public access based on security group rules.' At the bottom right are 'Cancel' and 'Save' buttons, with 'Save' also highlighted with a red box.



The screenshot shows the AWS Amazon EMR service interface after saving changes. The main content area displays a green success message: 'Block public access settings for this account successfully updated.' Below this, it says 'Amazon EMR > EMR on EC2: Block public access'. The 'Block public access' section shows the status 'Off' with a warning icon (a triangle with an exclamation mark). There is an 'Edit' button to the right of the status. The sidebar on the left is identical to the previous screenshot.

Tambien debe abrir los puertos de las aplicaciones de hadoop/Spark en el Security Group del nodo MASTER del cluster.

(nota: esto solo se hace una vez, cada vez que crea, destruya o clone un clúster, ya quedan abiertos)

Donde ubico el nodo master?

Dar Click en el cluster que acabas de crear, mirar la IP y nombre de la máquina EC2:

**Cluster EMONTOYA**

Updated less than a minute ago

Status: Waiting

Primary node public DNS: ec2-54-237-12-70.compute-1.amazonaws.com

Luego entras al servicio EC2 de dicha máquina Master, y va a modificar el Security Group para agregar los siguientes puertos de las aplicaciones:

**Application user interfaces**

On-cluster application UIs

- HDFS Name Node
- Hue
- JupyterHub
- Livy
- Resource Manager
- Spark History Server
- Tez UI
- Zeppelin

UI URLs:

- http://ec2-54-237-12-70.compute-1.amazonaws.com:9870/
- http://ec2-54-237-12-70.compute-1.amazonaws.com:8888/
- https://ec2-54-237-12-70.compute-1.amazonaws.com:9443/
- http://ec2-54-237-12-70.compute-1.amazonaws.com:8998/
- http://ec2-54-237-12-70.compute-1.amazonaws.com:8088/
- http://ec2-54-237-12-70.compute-1.amazonaws.com:18080/
- http://ec2-54-237-12-70.compute-1.amazonaws.com:8080/tez-ui
- http://ec2-54-237-12-70.compute-1.amazonaws.com:8890/

Además, abrir los puertos TCP:

22  
14000  
9870

En AWS EC2, les debería mostrar 3 máquinas:

The screenshot shows the AWS EC2 Instances page. The left sidebar is collapsed. The main area displays a table of instances. The first instance has an ID of i-0091fa028073a0d56, is running, and is an m5.xlarge type. The second instance has an ID of i-04d11b04008320812, is running, and is an m5.xlarge type. The third instance has an ID of i-030219cc3d31044dc, is running, and is an m5.xlarge type. All three instances have 2/2 checks passed and no alarms.

This screenshot shows the same EC2 Instances page as above, but with more detailed information for each instance. The first instance's Public IPv4 DNS is ec2-54-91-221-183.compute-1.amazonaws.com. The second instance's Public IPv4 DNS is ec2-54-237-12-70.compute-1.amazonaws.com, which is highlighted with a red box. The third instance's Public IPv4 DNS is ec2-3-90-200-139.compute-1.amazonaws.com. The security group for the second instance is also highlighted with a red box, labeled 'ElasticMapReduce-master'.

Entrar a la pestaña de seguridad de la Instancia EC2 del nodo master:

The screenshot shows the Instance summary for the master node (i-04d11b04008320812). The Security tab at the bottom is circled in red. The page displays various details about the instance, including its instance ID, public and private IP addresses, instance state (Running), and VPC ID (vpc-0f0e487421d53c205).

The screenshot shows the 'Security' tab of an AWS EMR cluster's details page. It displays the IAM Role (EMR\_EC2\_DefaultRole), Owner ID (140387140581), and Launch time (Thu Nov 02 2023 07:10:12 GMT-0500). Under 'Security groups', the entry 'sg-00e2003def25b8438 (ElasticMapReduce-master)' is highlighted with a red oval.

The screenshot shows the 'Edit inbound rules' page for the security group 'sg-00e2003def25b8438 - ElasticMapReduce-master'. It includes a breadcrumb navigation: EC2 > Security Groups > sg-00e2003def25b8438 - ElasticMapReduce-master > Edit inbound rules. The main section is titled 'Inbound rules' and contains columns for Security group rule ID, Type, Protocol, Port range, and Source. A red oval highlights the 'Add rule' button at the bottom left.

Uno a uno, va adicionando los puertos, aca se adiciono el puerto 22, haga lo mismo para los demás puertos:

The screenshot shows the 'Edit inbound rules' page with a new rule being added. The 'Type' dropdown is set to 'Custom TCP', the 'Port range' field contains '22', and the 'Source' field contains 'Anywhere'. A red oval highlights the '0.0.0.0/0' entry in the source field. The 'Add rule' button at the bottom left is also circled in red.

## Parte 2: Borrar y recrear clúster

Los clúster EMR en amazon, son temporales.

Los clúster EMR no se pueden pausar

Cada que no requiera trabajar más con un clúster, DEBE BORRARLO:

Pero la próxima vez que lo requiera, puede Clonar y crear nuevamente un clúster, teniendo en cuenta la configuración de otro clúster previamente creado, esta es la opción que debe utilizar.

Amazon EMR > EMR on EC2: Clusters

| Clusters (1/2) <a href="#">Info</a> |                            |                              |  |   |                              | <a href="#">View details</a> | <a href="#">Terminate</a> | <a href="#">Clone</a> | <a href="#">Create cluster</a> |
|-------------------------------------|----------------------------|------------------------------|--|---|------------------------------|------------------------------|---------------------------|-----------------------|--------------------------------|
|                                     | <a href="#">Cluster ID</a> | <a href="#">Cluster name</a> | <a href="#">Status</a>                   | <a href="#">Creation time (UTC-05:00)</a> | <a href="#">Elapsed time</a> |                              |                           |                       |                                |
| <input checked="" type="checkbox"/> | j-3DRPEB3XMBVAV            | Cluster EMONTOYA             | <span>Terminating</span><br>User request | November 02, 2023, 07:10                  | 3 hours, 1                   |                              |                           |                       |                                |
| <input type="checkbox"/>            | j-J5917MHJFP6H             | st1800-emontoya              | <span>Terminated</span><br>User request  | September 29, 2023, 19:32                 | 1 hour, 32                   |                              |                           |                       |                                |

Amazon EMR > EMR on EC2: Clusters > Create cluster

### Clone "Cluster EMONTOYA" [Info](#)

**Name and applications** [Info](#)

Name: Cluster EMONTOYA

Amazon EMR release [Info](#)  
A release contains a set of applications which can be installed on your cluster.

emr-6.14.0

Application bundle

|   |   |   |   |  |   |  |
|---|---|---|---|--|---|--|
| Spark Interactive<br> | Core Hadoop<br> | Flink<br> | HBase<br> | Presto<br> | Trino<br> | Custom<br> |
|---|---|---|---|--|---|--|

Flink 1.17.1  
 HCatalog 3.1.3  
 Hue 4.11.0  
 Livy 0.7.1  
 Phoenix 5.1.3  
 Spark 3.4.1  
 Tez 0.10.2  
 ZooKeeper 3.5.10

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

**Summary** [Info](#)

**Name and applications**

Name: Cluster EMONTOYA

Amazon EMR release: emr-6.14.0

Application bundle: Custom (Flink 1.17.1, HCatalog 3.1.3, Hadoop 3.3.3, Hive 3.1.3, Hue 4.11.0, JupyterHub...)

Amazon Linux release: 2.0.20230906.0

**Cluster configuration**

Instance groups: Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge)

**Cluster scaling and provisioning**

[Clone cluster](#)

Cada vez que lo Clone, debe crear nuevamente el usuario hadoop / con su password de preferencia, así como realizar el arreglo del archivo hue.ini para cambiar el puerto 14000 a 9870 (esto lo entenderá más adelante)

### Parte 3: Ingresar al clúster EMR por Hue

Utilice la aplicación hue, por el puerto 8888 desde un browser a la ip o nombre del nodo master.

Fijarse en la aplicación del clúster HUE:

The screenshot shows the AWS CloudWatch Metrics interface. At the top, there's a search bar and a navigation bar with tabs: Services, Applications (which is selected and highlighted with a red box), Configurations, Monitoring, Events, and Tags (0). Below the navigation bar, there's a section titled "Application user interfaces" with a "Info" link. It explains that applications installed on the cluster publish user interfaces as websites. Two options are shown: "On-cluster application UIs" (selected) and "Persistent application UIs". Under "Live Application UIs", it says "These on-cluster application UIs are available without SSH tunneling." and lists "Application UIs" like "Spark History Server UI". A section titled "Application UIs on the primary node" indicates that these require SSH tunneling and lists applications: "HDFS Name Node", "Hue", "JupyterHub", "Livy", "Resource Manager", "Spark History Server", "Tez UI", and "Zeppelin". To the right of this table is a button "Enable an SSH connection". Three red arrows point from the text "Y darle click a la URL de HUE, en este ejemplo:" to the URLs for Hue, JupyterHub, and Zeppelin.

Y darle click a la URL de HUE, en este ejemplo:

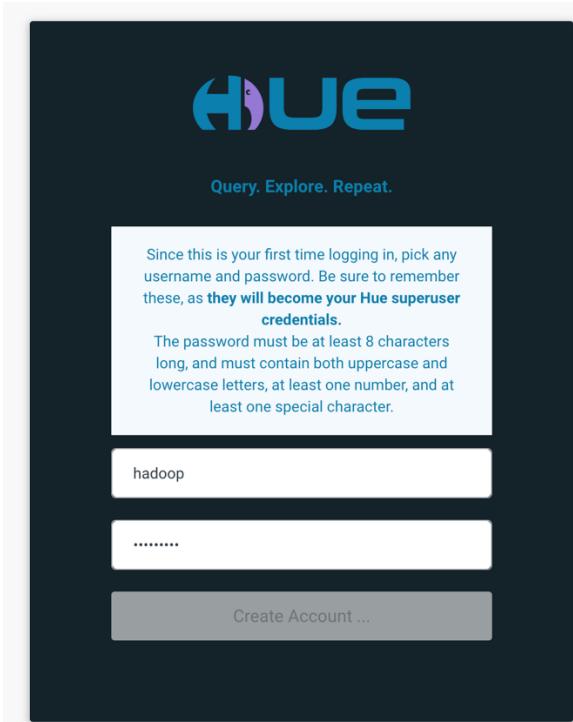
<http://ec2-54-237-12-70.compute-1.amazonaws.com:8888>

La primera vez, me pide crear un usuario y clave:

Username: hadoop

Password: <>el que quiera>>

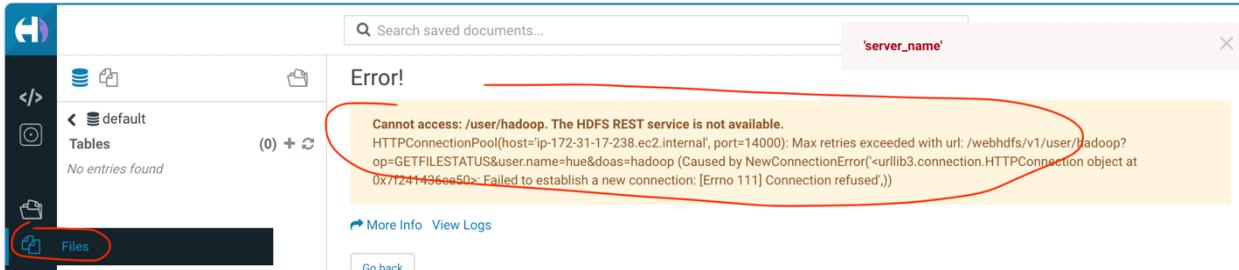
Nota: el usuario tiene que ser 'hadoop'



Deberá salir una interfaz así:

A screenshot of the Hue interface after logging in. On the left, there is a sidebar with various icons. In the center, under the "MySQL" section, it says "Add a name..." and "Add a desc...". A search bar at the top says "Search saved documents..." and has the placeholder "'server\_name'". Below the search bar, it says "No databases found ?" and "Example: SELECT \* FROM tablename, or press CTRL + space". Under the "Databases" section, it says "(0) + ⌂" and "Error loading databases.". On the right, there is a "Tables" section with the message "No tables identified." and a "Query History" section with the message "You don't have any saved queries."

Podrá acceder los servicios Hive, Spark, S3, pero el servicio de Files (archivos) HDFS falla, y le saldrá un error así:



Esto se corrige, entrando por SSH al nodo master del clúster, y realizando esto:

a. Entrar al EC2 nodo master por SSH

```

aws Services Search [Option+5]
Amazon Linux 2 AMI

https://aws.amazon.com/amazon-linux-2/
36 package(s) needed for security, out of 55 available
Run "sudo yum update" to apply all updates.

E:EEEEEEEEEEEEEEEEE M::::::M M:::::::M R:::::R RRRRRRRRRRRRRR
E:::-----:E M:::::M M:::::M R:::::R RRRRRRRRRRRRRR
E:EEEEE E:::::M M:::::M M:::::M R:::::R R:::::R
E:::E M:::::M M:::::M M:::::M R:::::R R:::::R
E:EEEEE E:::::M M:::::M M:::::M M:::::M R:::::RRRRRRR:::::R
E:-----:E M:::::M M:::::M M:::::M R:::::M R:::::RR
E:EEEEE E:::::M M:::::M M:::::M M:::::M R:::::RRRRRRR:::::R
E:::E M:::::M M:::::M M:::::M R:::::R R:::::R
E:EEEEE M:::::M MMM M:::::M R:::::R R:::::R
E:-----:E M:::::M M:::::M R:::::R R:::::R
E:-----:E M:::::M M:::::M R:::::R R:::::R
E:EEEEEEEEEEEEEEEEE M::::::M M:::::::M RRRRRRRRRRRRRR

[root@ip-172-31-17-238 ~]# 

i-04d11b04008320812
PublicIPs: 54.237.12.70 PrivateIPs: 172.31.17.238

```

b. Editar el archivo hue.ini

`nano /etc/hue/conf/hue.ini`

buscar la línea que contenga: 'webhdfs-url' y cambiar el puerto de 14000 a 9870

(en nano puede utilizar control-w para buscar la palabra)

```

aws Services Search [Option+S]
GNU nano 2.9.8 /etc/hue/conf/hue

# Configuration for HDFS NameNode
# -----
[[hdfs_clusters]]
# HA support by using HttpFs

[[[default]]]
# Enter the filesystem uri
fs_defaultfs = hdfs://ip-172-31-17-238.ec2.internal:8020

# NameNode logical name.
## logical_name=

# Use WebHdfs/HttpFs as the communication mechanism.
# Domain should be the NameNode or HttpFs host.
# Default port is 14000 for HttpFs.
webhdfs_url = http://ip-172-31-17-238.ec2.internal:14000/webhdfs/v1
# Change this if your HDFS cluster is Kerberos-secured
security_enabled = false

```

Cambiado y salvado:

(para salvar en nano: control-x -> Y para yes)

```

aws Services Search [Option+S]
GNU nano 2.9.8 /etc/hue/conf/hue

# Configuration for HDFS NameNode
# -----
[[hdfs_clusters]]
# HA support by using HttpFs

[[[default]]]
# Enter the filesystem uri
fs_defaultfs = hdfs://ip-172-31-17-238.ec2.internal:8020

# NameNode logical name.
## logical_name=

# Use WebHdfs/HttpFs as the communication mechanism.
# Domain should be the NameNode or HttpFs host.
# Default port is 14000 for HttpFs.
webhdfs_url = http://ip-172-31-17-238.ec2.internal:9870/webhdfs/v1
# Change this if your HDFS cluster is Kerberos-secured
security_enabled = false

# In secure mode (HTTPS), if SSL certificates from YARN Rest APIs
# have to be verified against certificate authority
## ssl_cert_ca_verify=True

```

c. Reiniciar el servicio hue

systemctl restart hue.service

Listo!!!!!! Ya puede entrar nuevamente por hue:

The screenshot shows the Apache Hue interface. On the left, there's a sidebar with icons for Tables, Views, and Files. The main area is titled 'File Browser' and shows the path '/user/hadoop'. It displays a list of files and directories under '/user/hadoop'. The table has columns for Name, Size, User, Group, Permissions, and Date. Two entries are listed: 'hdfs' (Size 0, User hdfs, Group hdfsadmingroup, Permissions drwxr-xr-x, Date November 02, 2023 05:23 AM) and '.' (Size 0, User hadoop, Group hdfsadmingroup, Permissions drwxrwxrwx, Date November 02, 2023 05:23 AM). There are also buttons for Upload and New.

Ya va a poder gestionar archivos sin problema por hue para HDFS

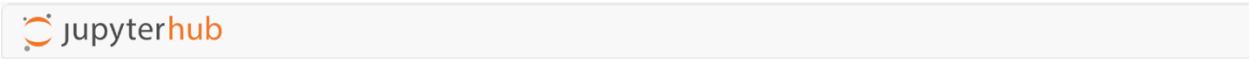
#### Parte 4: entrar a jupyter hub

Utilice la aplicación jupyterhub de:

The screenshot shows the AWS EMR Applications page. The 'Applications' tab is selected. Under 'Application user interfaces', it says 'On-cluster application UIs' are available only while the cluster is running. It lists several application UIs with their corresponding URLs:

- HDFS Name Node: <http://ec2-54-237-12-70.compute-1.amazonaws.com:9870/>
- Hue: <http://ec2-54-237-12-70.compute-1.amazonaws.com:8888/>
- JupyterHub: <https://ec2-54-237-12-70.compute-1.amazonaws.com:9443/>
- Livy: <http://ec2-54-237-12-70.compute-1.amazonaws.com:8998/>
- Resource Manager: <http://ec2-54-237-12-70.compute-1.amazonaws.com:8088/>
- Spark History Server: <http://ec2-54-237-12-70.compute-1.amazonaws.com:18080/>
- Tez UI: <http://ec2-54-237-12-70.compute-1.amazonaws.com:8080/tez-ui>
- Zeppelin: <http://ec2-54-237-12-70.compute-1.amazonaws.com:8890/>

Para este caso la URL es: <https://ec2-54-237-12-70.compute-1.amazonaws.com:9443/>



Utilice el usuario por defecto:

Username: jovyan

Password: jupyter

Tomado de: <https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-jupyterhub-user-access.html>

Y listo, ya puede realizar notebooks pyspark, verifique que las 2 variables más importantes de contexto de spark estan activas en un notebook pyspark) (primero debe crear un notebook pyspark)

A screenshot of a Jupyter Notebook interface. The title bar says "jupyterhub Untitled (autosaved)". The top menu bar includes File, Edit, View, Insert, Cell, Kernel, Widgets, Help, Logout, Control Panel, Trusted, and PySpark. The main area shows two code cells. The first cell, labeled "In [1]:", contains the command "spark" and outputs "Starting Spark application" followed by a table of application details. The second cell, labeled "In [2]:", contains the command "sc" and outputs "&lt;SparkContext master=yarn appName=livy-session-0&gt;". A third cell, labeled "In [ ]:", is at the bottom with a green border.