

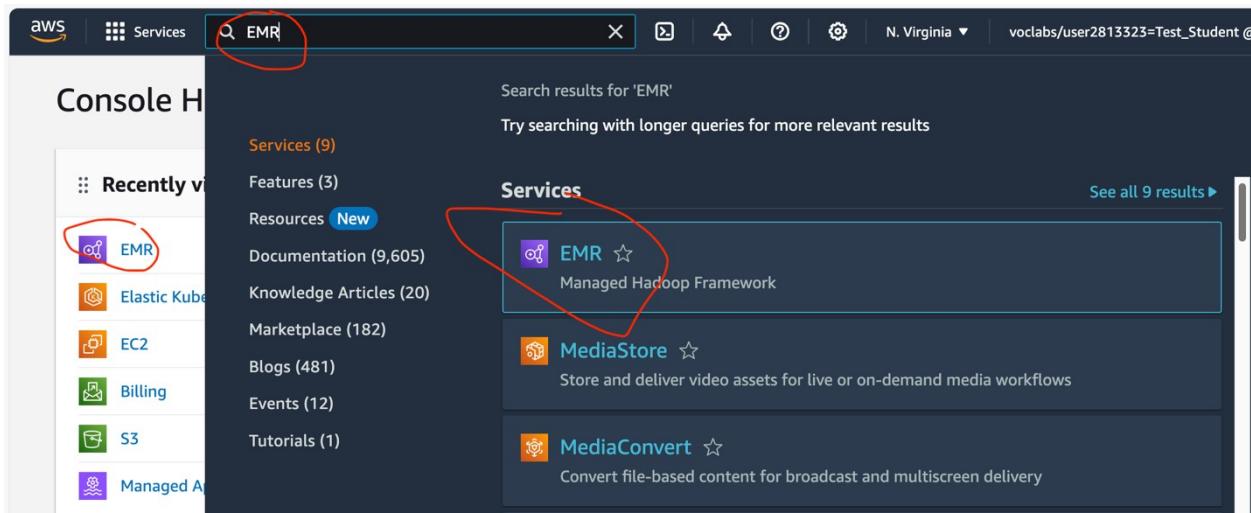
Laboratorio: Instalar un clúster EMR versión 7.9.0 Hadoop/Spark

Fecha: 22 mayo 2025

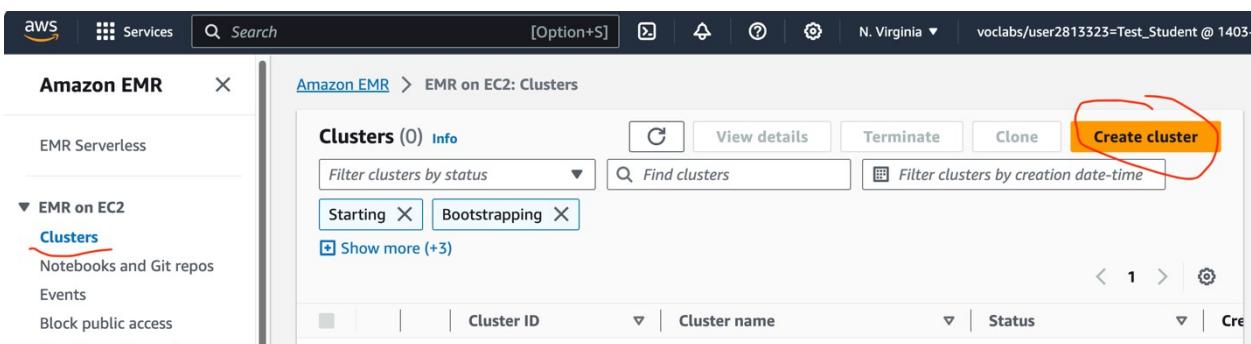
Parte 1: Crear un clúster AWS EMR versión 7.9.0

1 Instalar AWS EMR

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



2. crear clúster



3. nombre, versión y Custom

Name
myeks

Amazon EMR release | Info
A release contains a set of applications which can be installed on your cluster.
emr-7.9.0

Application bundle

Spark Interactive	Core Hadoop	Flink	HBase	Presto	Trino	Custom

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.

Amazon CloudWatch Agent 1.300032.2

AmazonCloudWatchAgent 1.300032.2

HCatalog 3.1.3

Hue 4.11.0

Livy 0.8.0

Pig 0.17.0

TensorFlow 2.16.1

Zeppelin 0.11.1

Flink 1.20.0

Hadoop 3.4.1

JupyterEnterpriseGateway 2.6.0

Oozie 5.2.1

Presto 0.287

Tez 0.10.2

ZooKeeper 3.9.3

HBase 2.6.2

Hive 3.1.3

JupyterHub 1.5.0

Phoenix 5.2.1

Spark 3.5.5

Trino 467

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

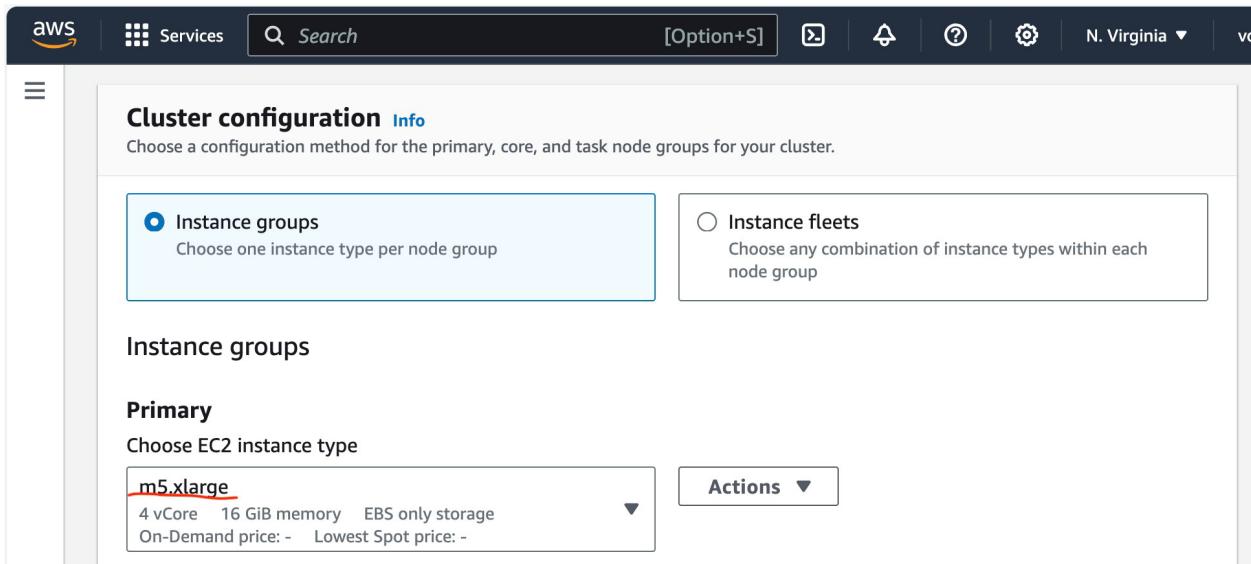
Amazon Linux release

Custom Amazon Machine Image (AMI)

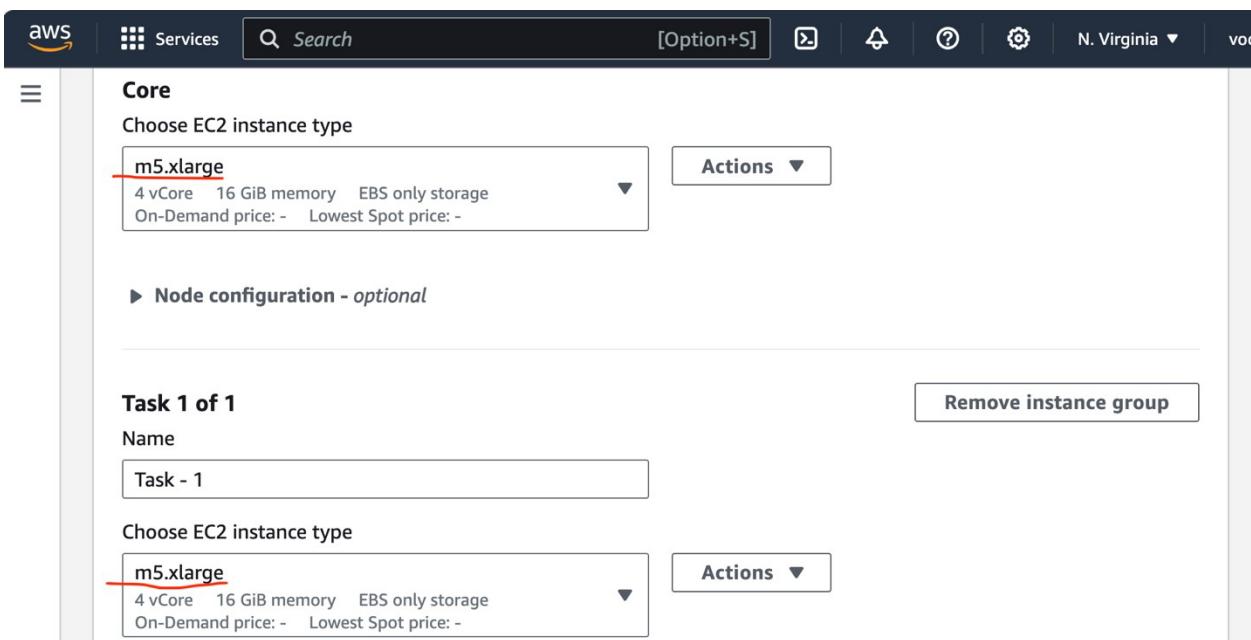
Automatically apply latest Amazon Linux updates

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.



The screenshot shows the 'Cluster configuration' page in the AWS Management Console. The 'Instance groups' option is selected, indicated by a blue outline. Below it, a note says 'Choose one instance type per node group'. To the right, the 'Instance fleets' option is shown with a note 'Choose any combination of instance types within each node group'. Under 'Instance groups', there's a section for 'Primary' where 'm5.xlarge' is selected from a dropdown menu. The dropdown details '4 vCore 16 GiB memory EBS only storage' and shows 'On-Demand price: - Lowest Spot price: -'. An 'Actions' button is also visible.



The screenshot shows the 'Core' configuration section of the cluster setup. It asks to 'Choose EC2 instance type' and lists 'm5.xlarge' as the selected option. The dropdown details '4 vCore 16 GiB memory EBS only storage' and shows 'On-Demand price: - Lowest Spot price: -'. An 'Actions' button is present. Below this, there's a section titled 'Node configuration - optional'. At the bottom, under 'Task 1 of 1', there's a 'Name' field containing 'Task - 1' and a 'Remove instance group' button.

6. Dejar estas opciones por defecto

The screenshot shows the 'Provisioning configuration' section of the AWS EMR console. It includes a table for instance groups and a 'Networking' section with VPC and subnet details.

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
vpc-0f0e487421d53c205

Subnet Info
subnet-03074aab481e8b97e

EC2 security groups (firewall)

Tener en cuenta el EC2 security groups (firewall), para más adelante adicionar los diferentes puertos para las aplicaciones:

Primary node

EMR-managed security group
EMR will automatically update the selected group.
ElasticMapReduce-Primary
sg-0d8ee0443043c005d

Additional security groups - *optional*
Select up to 4 additional security groups.
Choose additional security groups

Core and task nodes

EMR-managed security group
EMR will automatically update the selected group.
ElasticMapReduce-Core
sg-01ad8a27e3ec82827

Additional security groups - *optional*
Select up to 4 additional security groups.
Choose additional security groups

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": "su-bucket"  
   }  
 }  
]
```

Y pegue esta configuración en Software Settings así:

The screenshot shows the AWS Software Settings interface. At the top, there are tabs for 'Services' and 'Search'. Below the tabs, there is a 'Software settings - optional' section. Underneath this, there are two options: 'Enter configuration' (selected) and 'Load JSON from Amazon S3'. The 'Enter configuration' field contains the following JSON code:

```
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

The screenshot shows the AWS Security configuration and EC2 key pair interface. At the top, there are tabs for 'Services' and 'Search'. Below the tabs, there is a 'Security configuration and EC2 key pair - optional' section. Underneath this, there is a 'Security configuration' section with a note: 'Select your cluster encryption, authentication, authorization, and instance metadata service settings.' There are three buttons: 'Choose a security configuration', 'Browse', and 'Create security configuration'. Below this, there is an 'Amazon EC2 key pair for SSH to the cluster' section. A search bar here contains the value 'vockey', which is circled in red. There are also 'Browse' and 'Create key pair' buttons.

9. IAM roles

Debe seleccionar:

Service role: EMR_DefaultRole

Instance profile: EMR_EC2_DefaultRole

Custom automatic scaling role: EMR_AutoScaling_DefaultRole

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

EMR_AutoScaling_DefaultRole



Create IAM role

10. Si desea instalar librerías adicionales en python para ejecutarse en los notebooks de jupyterhub, debe adicionar este script.sh en la sección de:

Bootstrap Actions(0)

▼ **Bootstrap actions (0)** Info

Use bootstrap actions to install software or customize your instance configuration.

Name	Amazon S3 location	Arguments
No bootstrap actions You don't have any bootstrap actions to display.		

Add

Add bootstrap action X

Name

Script location
For best performance, store custom bootstrap actions in the same AWS Region as your cluster.

Arguments - optional
Provide arguments for your bootstrap action scripts. These arguments send references to the scripts on to Amazon EMR.

Cancel **Add bootstrap action**

El contenido del archivo: `install-my-jupyter-libraries.sh`

```
#!/bin/bash
sudo python3 -m pip install boto3 nltk scipy scikit-learn pandas
```

Asegúrese antes de crear o clonar el cluster EMR, tener copiado el archivo: `install-my-jupyter-libraries.sh` en el bucket S3 requerido (en el github de la materia se encuentra un ejemplo de este archivo)

11. Finalmente, a crear el clúster

The screenshot shows the 'Cluster scaling and provisioning' step of the EKS cluster creation wizard. It details the provisioning configuration: Core size: 1 instance and Task size: 1 instance. Under 'Networking', it lists the VPC (vpc-0f0e48742...), Subnet (subnet-03074a...), Primary node security group (sg-00e2003def...), and Core node security group (sg-0f31bdef3e...). At the bottom right, there is a red curved arrow pointing to the orange 'Create cluster' button.

Este proceso demora aproximadamente 20 minutos, tenga paciencia.

Debe salir con este mensaje de clúster exitosamente creado:

The screenshot shows the 'EMR on EC2: Clusters' page in the Amazon EMR console. It displays two clusters: 'Cluster EMONTOYA' (Status: Waiting) and 'st1800-emontoya' (Status: Terminated). A red circle highlights the 'Waiting' status of the first cluster.

Cluster ID	Cluster name	Status
j-3DRPEB3XMBVAV	Cluster EMONTOYA	Waiting Ready to run steps
j-J5917MHJFP6H	st1800-emontoya	Terminated User request

12. 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 Management Console with the search bar at the top. The left sidebar under 'Amazon EMR' has 'EMR Serverless' and 'EMR on EC2' expanded. 'Block public access' is selected and highlighted with a red box. The main content area shows the 'Edit settings' for 'EMR on EC2: Block public access'. The 'Block public access' section contains two options: 'Turn on - recommended' (unchecked) and 'Turn off' (checked). Below the checked option, it says 'Allow public access based on security group rules.' At the bottom right are 'Cancel' and 'Save' buttons, both of which are circled in red.

The screenshot shows the AWS Management Console with the search bar at the top. The left sidebar under 'Amazon EMR' has 'EMR Serverless' and 'EMR on EC2' expanded. 'Block public access' is selected and highlighted with a red box. A green success message banner at the top says 'Block public access settings for this account successfully updated.' The main content area shows the 'Block public access' section with the status 'Off' highlighted with a red circle. There is also an 'Edit' button.

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

(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 clúster que acabas de crear, mirar la IP y nombre de la máquina EC2:

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	UI URL
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/

Además, abrir los puertos TCP:

22
14000
9870

En AWS EC2, les debería mostrar 3 máquinas:

Screenshot of the AWS EC2 Instances page showing three running m5.xlarge instances. The search bar filters for 'Instance state = running'. The table includes columns for Name, Instance ID, Instance state, Instance type, Status check, and Alarm status.

Name	Instance ID	Instance state	Instance type	Status check	Alarm status
	i-0091fa028073a0d56	Running	m5.xlarge	2/2 checks passed	No alarms
	i-04d11b04008320812	Running	m5.xlarge	2/2 checks passed	No alarms
	i-030219cc3d31044dc	Running	m5.xlarge	2/2 checks passed	No alarms

Screenshot of the AWS EC2 Instances page showing three instances with their Public IPv4 DNS and Public IPv4 addresses. The table includes columns for Public IPv4 DNS, Public IPv4 address, Elastic IP, IPv6 IPs, Monitoring, Security group name, and Key name.

Public IPv4 DNS	Public IPv4 address	Elastic IP	IPv6 IPs	Monitoring	Security group name	Key name
ec2-54-91-221-183.compute-1.amazonaws.com	54.91.221.183	-	-	disabled	ElasticMapReduce-slave	vockey
ec2-54-237-12-70.compute-1.amazonaws.com	54.237.12.70	-	-	disabled	ElasticMapReduce-master	vockey
ec2-3-90-200-139.compute-1.amazonaws.com	3.90.200.139	-	-	disabled	ElasticMapReduce-slave	vockey

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

Screenshot of the AWS EC2 Instance summary page for instance i-04d11b04008320812. The 'Security' tab is highlighted with a red circle. The page displays various instance details such as Instance ID, Public IPv4 address, Instance state, and IAM Role.

Instance summary for i-04d11b04008320812

Updated less than a minute ago

Instance ID i-04d11b04008320812	Public IPv4 address 54.237.12.70 [open address]	Private IPv4 addresses 172.31.17.238
IPv6 address -	Instance state Running	Public IPv4 DNS ec2-54-237-12-70.compute-1.amazonaws.com [open address]
Hostname type IP name: ip-172-31-17-238.ec2.internal	Private IP DNS name (IPv4 only) ip-172-31-17-238.ec2.internal	Elastic IP addresses -
Answer private resource DNS name -	Instance type m5.xlarge	AWS Compute Optimizer finding Opt-in to AWS Compute Optimizer for recommendation s. [Learn more]
Auto-assigned IP address 54.237.12.70 [Public IP]	VPC ID vpc-0f0e487421d53c205	Auto Scaling Group name -
IAM Role EMR_EC2_DefaultRole	Subnet ID subnet-03074aab481e8b97e	
IMDSv2 Optional		

Details **Security** Networking Storage Status checks Monitoring Tags

Details Security Networking Storage Status checks Monitoring Tags

▼ Security details

IAM Role EMR_EC2_DefaultRole Owner ID 140387140581 Launch time Thu Nov 02 2023 07:10:12 GMT-05:00

Security groups sg-00e2003def25b8438 (ElasticMapReduce-master)

aws Services Search [Option+S]

[EC2](#) > [Security Groups](#) > [sg-00e2003def25b8438 - ElasticMapReduce-master](#) > [Edit inbound rules](#)

Edit inbound rules Info

Inbound rules control the incoming traffic that's allowed to reach the instance.

Security group rule ID	Type <small>Info</small>	Protocol <small>Info</small>	Port range	Source <small>Info</small>
------------------------	--------------------------	------------------------------	------------	----------------------------

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

-

Custom TCP ▾

TCP

22

Anyw... ▾

0.0.0.0/0 X

Add rule

Delete

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.

The screenshot shows the 'Clusters (1/2)' page in the Amazon EMR console. It lists two clusters: 'Cluster EMONTOYA' (status: Terminating) and 'st1800-emontoya' (status: Terminated). The 'Cluster EMONTOYA' row has a checked checkbox. The 'Clone' button in the top right corner is highlighted with a red circle.

The screenshot shows the 'Clone "Cluster EMONTOYA"' configuration page. In the 'Name and applications' section, 'Cluster EMONTOYA' is selected. The 'Summary' section on the right shows the same cluster details. The 'Cluster scaling and provisioning' section is partially visible at the bottom right. The 'Clone cluster' button in the bottom right corner is highlighted with a red circle.

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 Management Console interface for an EMR cluster. The top navigation bar includes 'Services', 'Search', and tabs for 'Properties', 'Bootstrap actions', 'Instances (Hardware)', 'Steps', 'Applications' (which is highlighted with a blue border), 'Configurations', 'Monitoring', 'Events', and 'Tags (0)'. The main content area is titled 'Application user interfaces' with a 'Info' link. It explains that applications installed on the cluster publish user interfaces as websites for monitoring cluster activity. Two options are shown: 'On-cluster application UIs' (selected) and 'Persistent application UIs'. The 'On-cluster application UIs' section lists several services with their respective URLs. Red arrows from the surrounding text point to the URLs for 'Hue', 'JupyterHub', and 'Zeppelin'.

Application	UI URL
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/

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

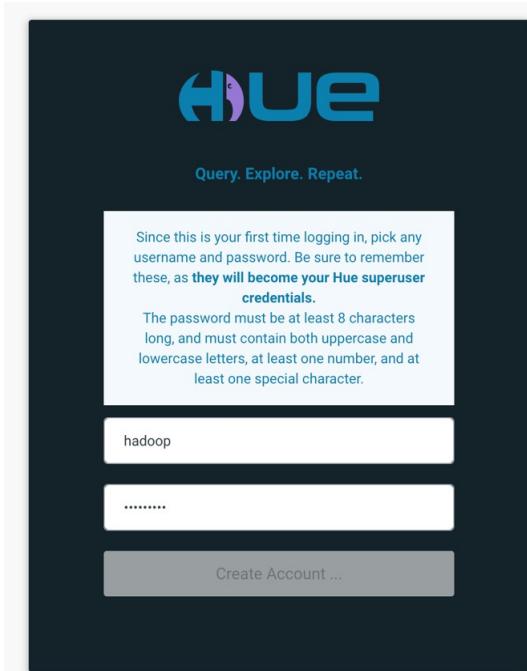
<http://ec2-54-237-12-70.compute-1.amazonaws.com:8888>

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 web interface. On the left is a sidebar with various icons for different services. The main area shows a MySQL connection. The "Databases" section says "(0)" and "Error loading databases.". The "Tables" section says "No tables found" and "No tables identified.". At the bottom, there are tabs for "Query History" and "Saved Queries", with the latter being active. A message states "You don't have any saved queries.".

Podrá acceder los servicios Hive, Spark, S3, y HDFS.

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

The screenshot shows the Hue File Browser interface. On the left, there's a sidebar with various icons and a list of 'Sources' including MySQL, Hive, PostgreSQL, SparkSQL, and SQLite. A red circle highlights the 'File' icon in the sidebar. The main area is titled 'File Browser' and shows the path '/user/hadoop/datasets'. It contains two items: a folder named 't' owned by hadoop with permissions drwxrwxrwx, and a folder named '.' owned by hadoop with permissions drwxr-xr-x. Below the table, it says 'Show 45 of 0 items'. At the top right, there are 'Upload' and 'New' buttons, which are also circled in red.

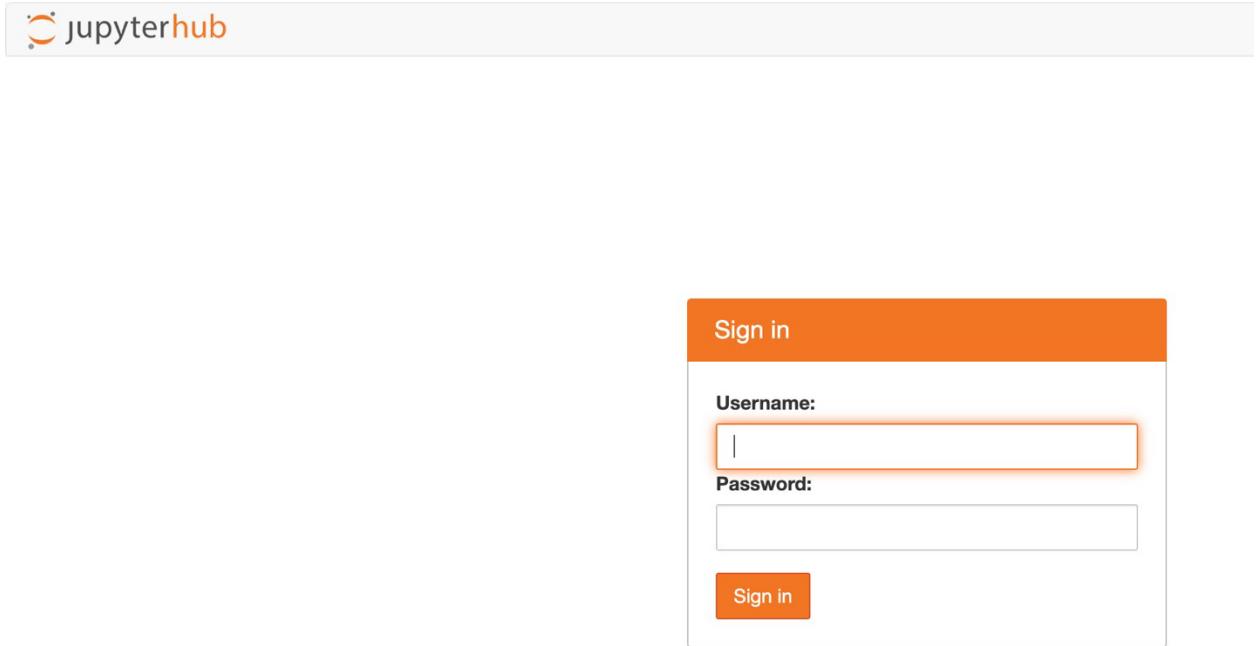
Parte 4: entrar a jupyter hub

Utilice la aplicación jupyterhub de:

The screenshot shows the AWS EMR cluster configuration page under the 'Applications' tab. It lists 'On-cluster application UIs' and 'Persistent application UIs'. The 'On-cluster application UIs' section is highlighted with a red box and three red arrows pointing to the UI URLs for JupyterHub, Livy, and Zeppelin. The UI URLs are listed as follows:

- JupyterHub: <http://ec2-54-237-12-70.compute-1.amazonaws.com:9443/>
- Livy: <http://ec2-54-237-12-70.compute-1.amazonaws.com:8888/>
- 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 esta activas en un notebook así: (primero debe crear un notebook pyspark)

A screenshot of a Jupyter Notebook interface. The title bar says "jupyterhub Untitled (autosaved)". The toolbar includes File, Edit, View, Insert, Cell, Kernel, Widgets, Help, Logout, Control Panel, Trusted, and PySpark. The main area shows two code cells. Cell [1] contains the command "spark" and outputs "Starting Spark application" followed by a table of application details. Cell [2] contains the command "sc" and outputs "<SparkContext master=yarn appName=livy-session-0>".

ID	YARN Application ID	Kind	State	Spark UI	Driver log	User	Current session?
0	application_1698927808120_0001	pyspark	idle	Link	Link	None	✓