

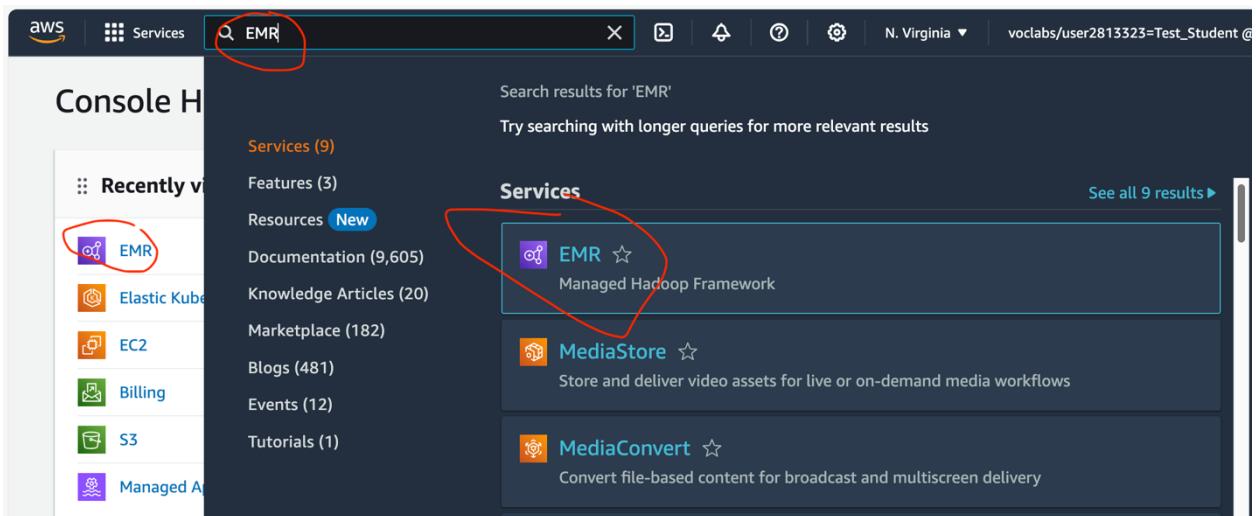
Laboratorio: Instalar un clúster EMR versión 7.3.0 Hadoop/Spark

Fecha: 16 octubre 2024

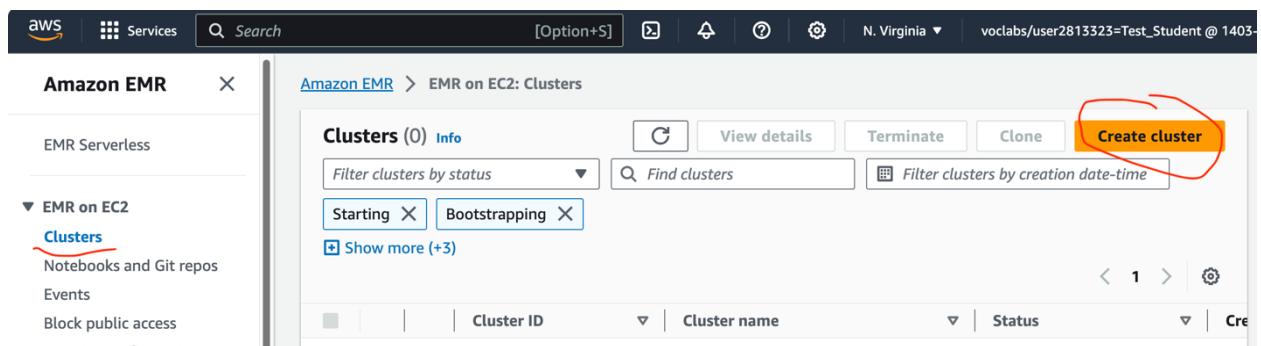
Parte 1: Crear un clúster AWS EMR versión 7.3.0

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

The screenshot shows the 'Create cluster' wizard in the AWS EMR console. The 'Name and applications - required' section is active. The 'Name' field contains 'My cluster EMONTOYA'. The 'Amazon EMR release' dropdown is set to 'emr-7.3.0'. In the 'Application bundle' section, several options are listed: Spark Interactive (Apache Spark logo), Core Hadoop (hadoop logo), Flink (Flink logo), HBase (HBase logo), Presto (presto logo), Trino (trino logo), and a circled 'Custom' option (aws logo). A red circle highlights the 'Name' field, the 'emr-7.3.0' dropdown, and the 'Custom' option.

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.

Application bundle

- Spark Interactive
- Core Hadoop
- Flink
- HBase
- Presto
- Trino
- Custom

AmazonCloudWatchAgent 1.300032.2

HCatalog 3.1.3

Hue 4.11.0

Livy 0.8.0

Pig 0.17.0

Sqoop 1.4.7

Trino 442

Flink 1.18.1

Hadoop 3.3.6

JupyterEnterpriseGateway 2.6.0

Oozie 5.2.1

Presto 0.285

TensorFlow 2.16.1

Zeppelin 0.11.1

HBase 2.4.17

Hive 3.1.3

JupyterHub 1.5.0

Phoenix 5.1.3

Spark 3.5.1

Tez 0.10.2

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

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.

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 ▾

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

Task - 1

Remove instance group

Choose EC2 instance type

m5.xlarge

4 vCore 16 GiB memory EBS only storage

On-Demand price: - Lowest Spot price: -

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](#)

vpc-0f0e487421d53c205 [Browse](#) [Create VPC](#)

Subnet [Info](#)

subnet-03074aab481e8b97e [Browse](#) [Create subnet](#)

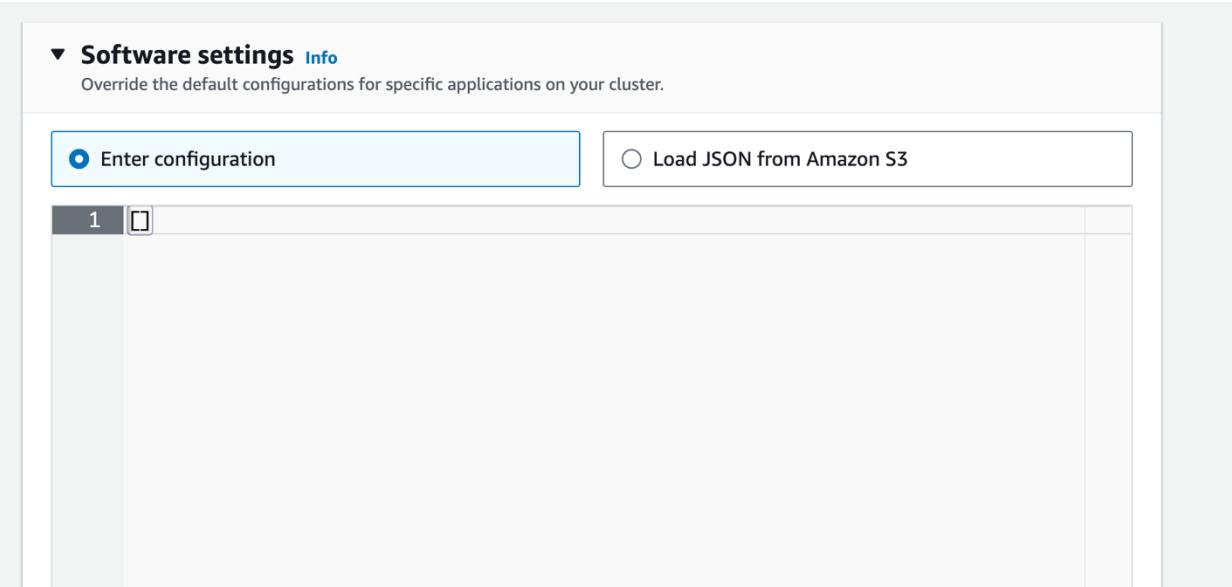
► EC2 security groups (firewall)

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

The screenshot shows the AWS EMR setup interface. Under the 'EC2 security groups (firewall)' section, there is a 'Change notice' box stating: 'We've updated the names of some security groups to use more inclusive language. For example, groups that included terms like "master" and "slave" now use the terms "primary" and "core" instead.' Below this, under 'Primary node', there is an 'EMR-managed security group' dropdown containing 'ElasticMapReduce-Primary' and its ID 'sg-0d8ee0443043c005d'. To the right, there is an 'Additional security groups - optional' section with a dropdown labeled 'Choose additional security groups'. Under 'Core and task nodes', there is another 'EMR-managed security group' dropdown containing 'ElasticMapReduce-Core' and its ID 'sg-01ad8a27e3ec82827'. To the right, there is an 'Additional security groups - optional' section with a dropdown labeled 'Choose additional security groups'.

Dejar las siguientes opciones por defecto hasta: Software settings.

7. Configurar software settings



▼ **Software settings** Info
Override the default configurations for specific applications on your cluster.

Enter configuration Load JSON from Amazon S3

```
1 [ ]
```

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

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](#)

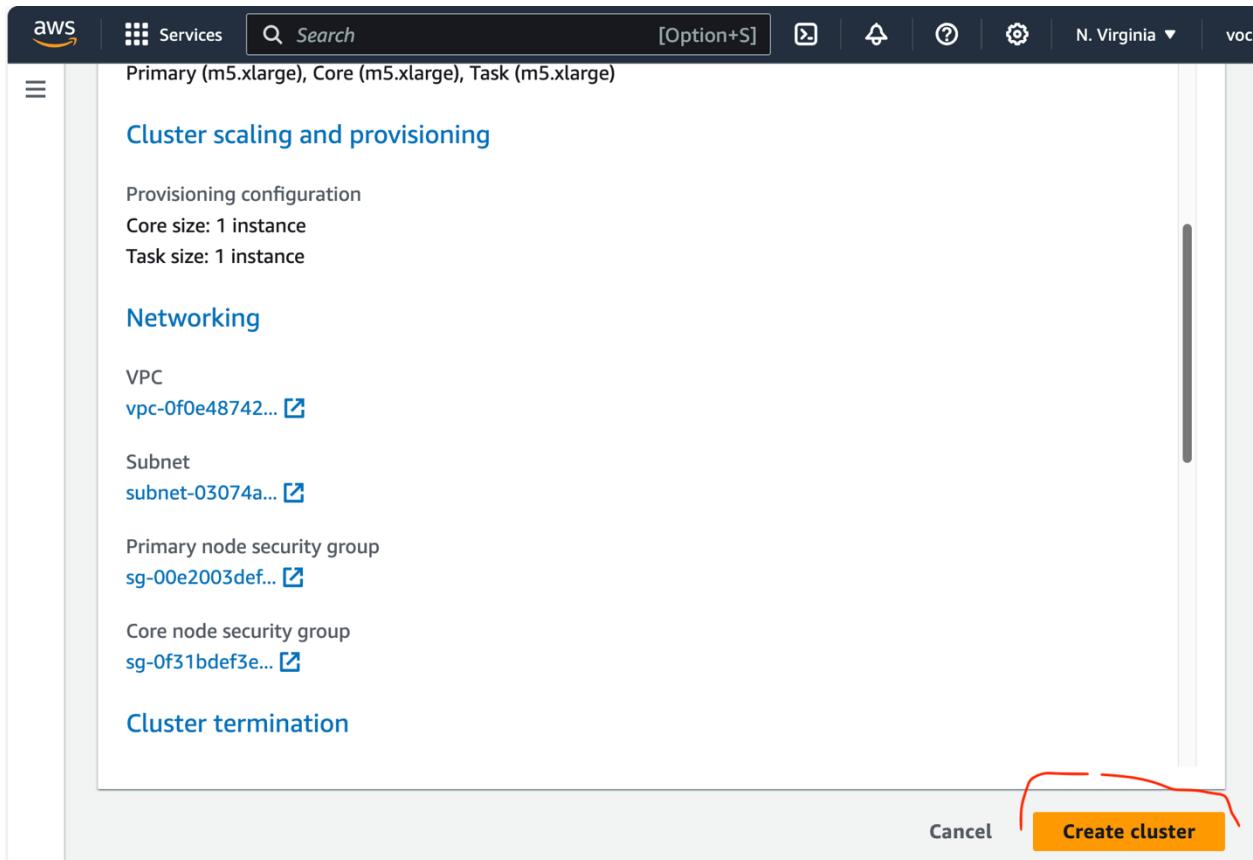
Custom automatic scaling role

EMR_AutoScaling_DefaultRole



Create IAM role

10. Finalmente, a crear el clúster



Este proceso demora aproximadamente 20 minutos, tenga paciencia.

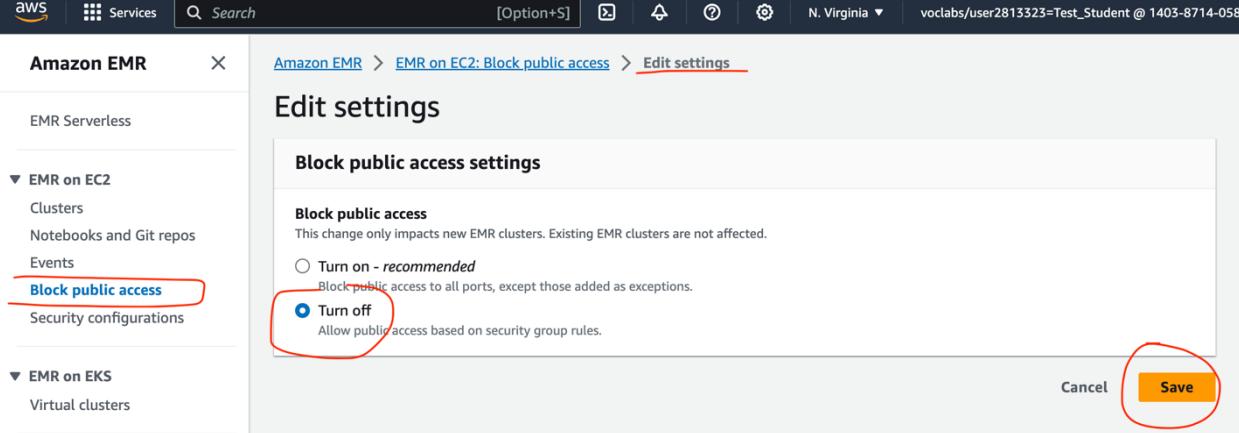
Debe salir con este mensaje de clúster exitosamente creado:

The screenshot shows the 'Clusters (2)' list:

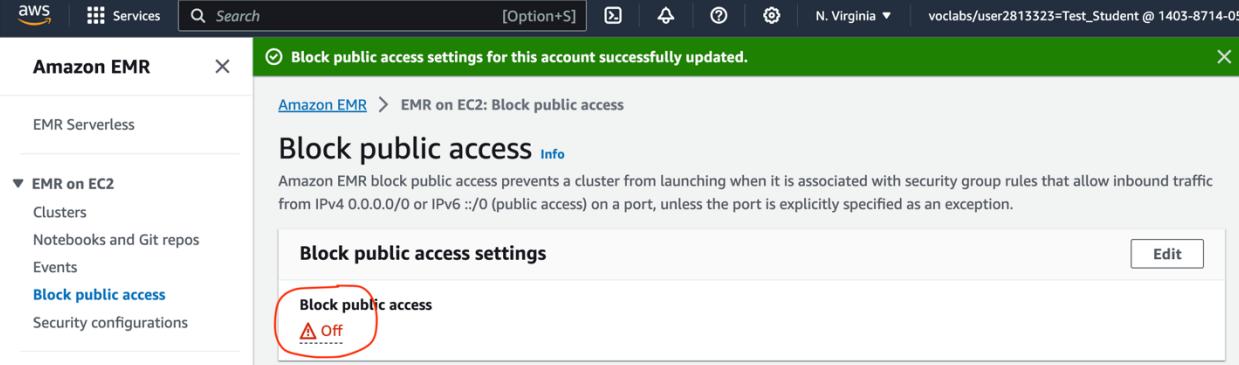
Cluster ID	Cluster name	Status
j-3DRPEB3XMBVAV	Cluster EMONTOYA	Waiting Ready to run steps
j-J5917MHJFP6H	st1800-emontoya	Terminated User request

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 'Edit settings' page for 'Block public access' under 'EMR on EC2'. The 'Turn off' radio button is selected, and the 'Save' button is highlighted with a red circle.



The screenshot shows a success message: 'Block public access settings for this account successfully updated.' The 'Block public access' status is shown as 'Off' with a warning icon, and the 'Edit' button is visible.

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:

Cluster EMONTOYA

Updated less than a minute ago

Status: Waiting

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

Connect to the Primary node using SSH

Connect to the Primary node using SSM

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 Info

Applications installed on your Amazon EMR cluster publish user interfaces (UI) as websites. You can use these to monitor cluster activity.

On-cluster application UIs
On-cluster UIs are available only while your cluster is running. Use the following links to get started. To access all the application UIs, set up SSH tunneling.

Persistent application UIs
Persistent UIs don't require SSH tunneling. They are hosted off of the cluster and are available for 30 days after an application ends.

Live Application UIs
These on-cluster application UIs are available without SSH tunneling.

[Application UIs](#)

[Spark History Server UI](#)

Application UIs on the primary node
These require SSH tunneling to be enabled.

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:

The screenshot shows the AWS EC2 Instances page with three running m5.xlarge instances listed:

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-030219cc3d3104dc	Running	m5.xlarge	2/2 checks passed	No alarms

The screenshot shows the AWS EC2 Instances page with the Public IPv4 DNS column highlighted, displaying the public DNS names for each instance.

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

The screenshot shows the AWS EC2 Instance summary page for instance i-04d11b04008320812. The Security tab is highlighted with a red circle.

Instance summary for i-04d11b04008320812

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 recommendations. [Learn more]
Auto-assigned IP address 54.237.12.70 [Public IP]	VPC ID ypc-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-0500 (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.

Inbound rules <small>Info</small>	Security group rule ID	Type <small>Info</small>	Protocol <small>Info</small>	Port range <small>Info</small>	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.

Amazon EMR > EMR on EC2: Clusters

Clusters (1/2) Info		C	View details	Terminate	Clone	Create cluster
		Filter clusters by status	<input type="text"/> Find clusters	Filter clusters by creation date-time	< 1 >	
<input checked="" type="checkbox"/>	j-3DRPEB3XMBVAV	Cluster ID	Cluster name	Status	Creation time (UTC-05:00)	Elapsed t
<input checked="" type="checkbox"/>	j-J5917MHJFP6H	j-3DRPEB3XMBVAV	Cluster EMONTOYA	Terminating User request	November 02, 2023, 07:10	3 hours, 1
<input type="checkbox"/>	j-J5917MHJFP6H	st1800-emontoya		Terminated 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-7.3.0

Application bundle

Spark Interactive	Core Hadoop	Flink	HBase	Presto	Trino	Custom

Flink 1.17.1 Ganglia 3.7.2 HBase 2.4.17
 HCatalog 3.1.3 Hadoop 3.3.3 Hive 3.1.3
 Hue 4.11.0 JupyterEnterpriseGateway 2.6.0 JupyterHub 1.5.0
 Livy 0.7.1 MXNet 1.9.1 Oozie 5.2.1
 Phoenix 5.1.3 Pig 0.17.0 Presto 0.281
 Spark 3.4.1 Sqoop 1.4.7 TensorFlow 2.11.0
 Tez 0.10.2 Trino 422 Zeppelin 0.10.1
 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, JupyterEnter...)

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 console with the 'Metrics Insights' tab selected. The main pane displays a log entry from a Lambda function named 'LambdaFunction'. The log message contains several URLs, some of which are highlighted with red arrows pointing to them. The URLs include:

- <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/>

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

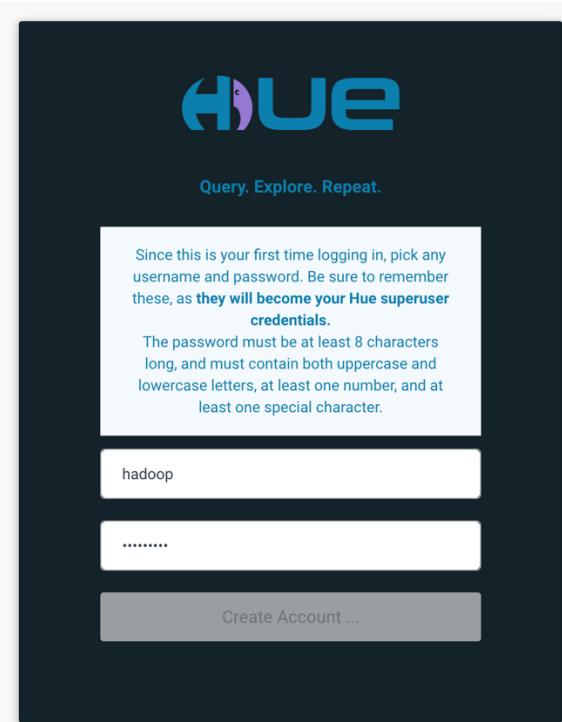
<http://ec2-54-237-12-70.compute-1.amazonaws.com:8888>

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

Username: hadoop

Password: <>el que quiera>>

Nota: el usuario tiene que ser 'hadoop'

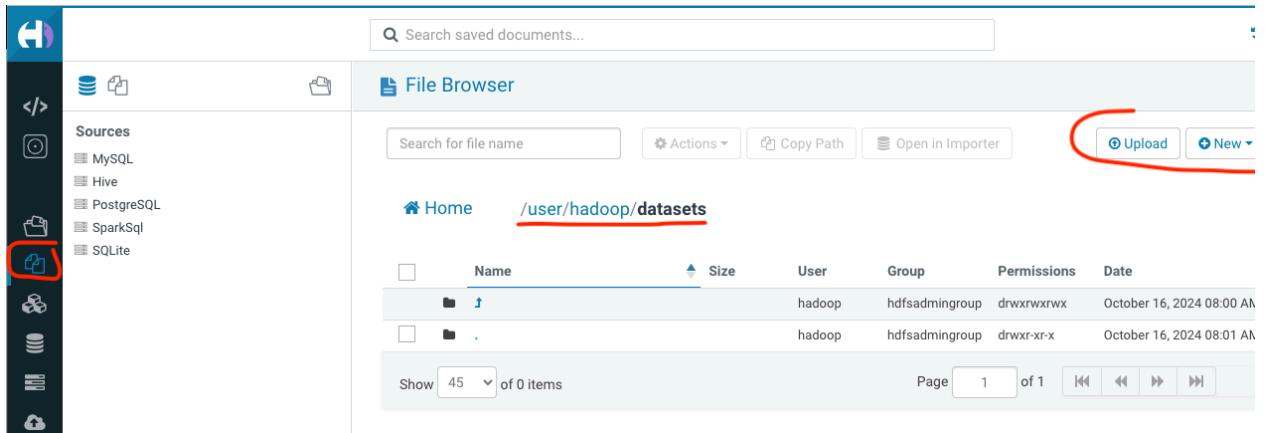


Deberá salir una interfaz así:

The image shows the Hue interface after logging in. On the left is a sidebar with various icons. The main area is titled "MySQL" and shows a search bar with "server_name". It displays a message: "Example: SELECT * FROM tablename, or press CTRL + space". Below this is a "Databases" section with "(0)" and a note "Error loading databases.". To the right is a "Tables" section with "No tables identified". At the bottom, there are tabs for "Query History" and "Saved Queries", with the latter being active.

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

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

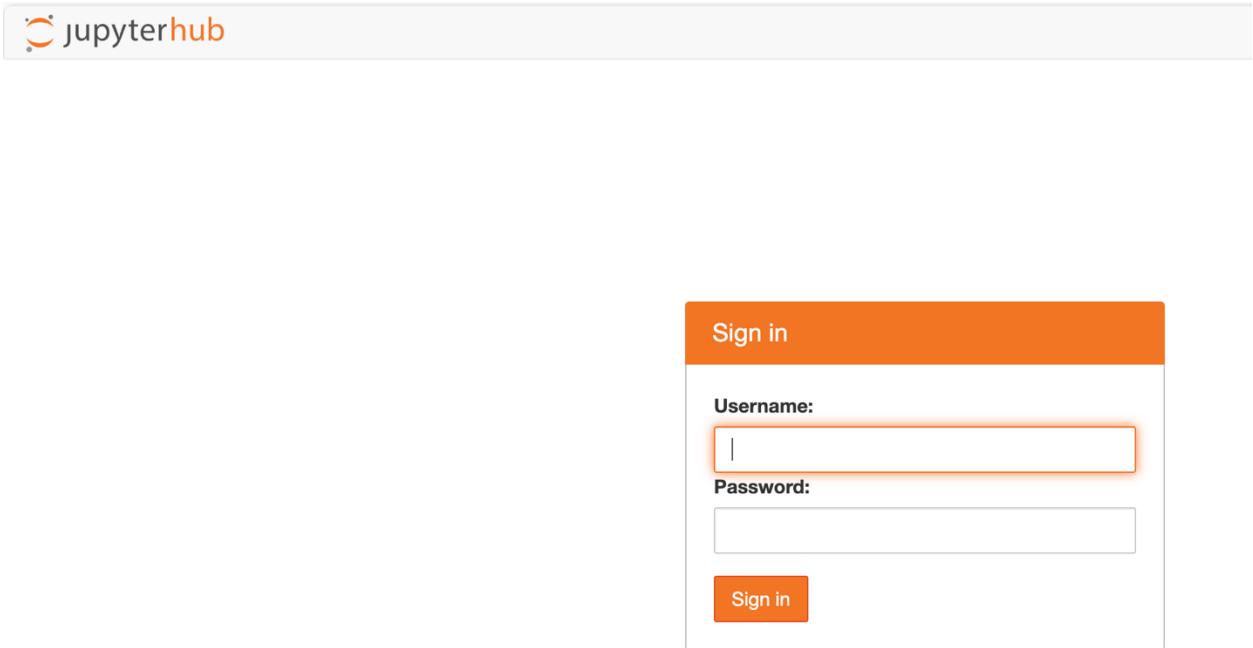


Parte 4: entrar a jupyter hub

Utilice la aplicación jupyterhub de:

The screenshot shows the AWS Management Console with the AWS logo at the top. Below it, the navigation bar includes Services, Search, and tabs for Properties, Bootstrap actions, Instances (Hardware), Steps, Applications (which is highlighted with a red box), Configurations, Monitoring, Events, and Tags (0). The Applications tab displays information about application user interfaces. It shows two sections: 'On-cluster application UIs' (selected) and 'Persistent application UIs'. Under 'On-cluster application UIs', it lists several services with their corresponding UI URLs. Three specific URLs are highlighted with red arrows pointing to them: <http://ec2-54-237-12-70.compute-1.amazonaws.com:9870/>, <https://ec2-54-237-12-70.compute-1.amazonaws.com:9443/>, and <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 includes File, Edit, View, Insert, Cell, Kernel, Widgets, Help, Logout, Control Panel, Trusted, and PySpark. The toolbar below has icons for file operations like Open, Save, and Run. 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	✓