

Tutorial creación clúster EMR en AWS

Pablo Moreno Quintero

Escuela de Ciencias e Ingeniería, Universidad EAFIT

Pregrado en Ingeniería de Sistemas

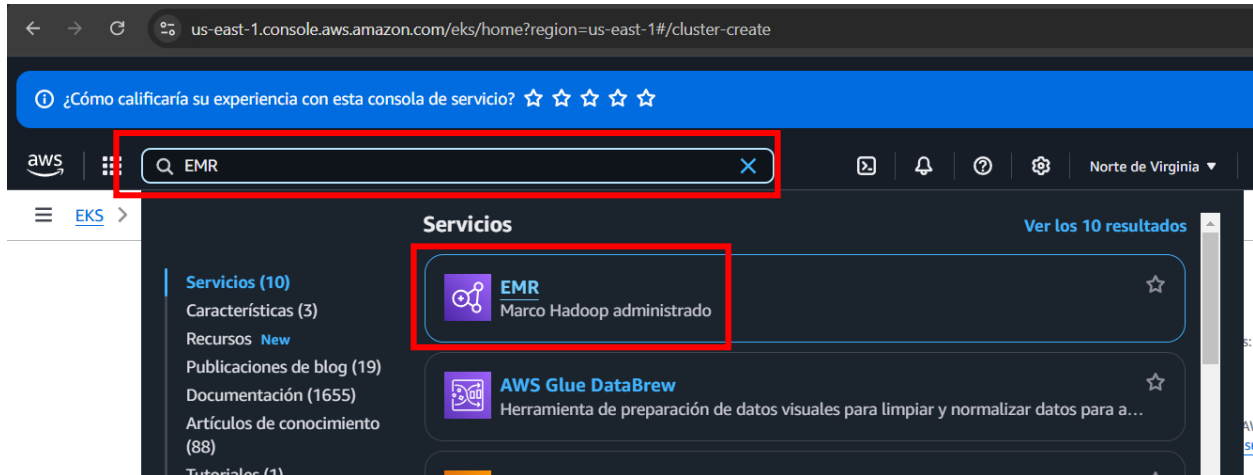
Edwin Nelson Montoya Munera

23 de noviembre de 2024

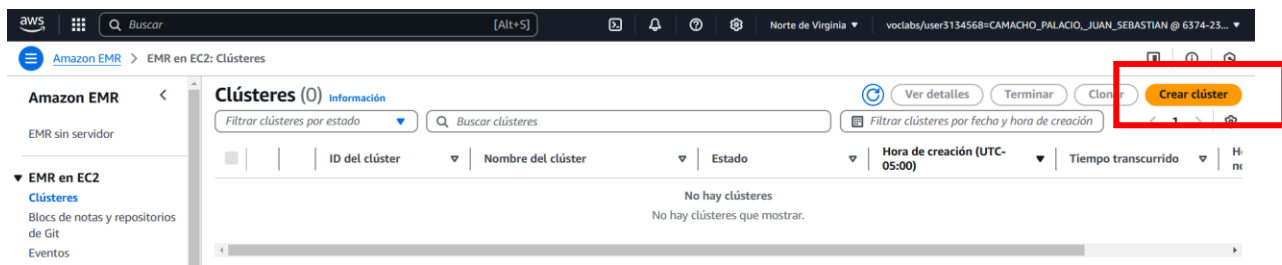
1. Creación del clúster y su configuración inicial

Dirigirse a la consola de administración de servicios de AWS

En la barra de búsqueda buscar EMR y seleccionar la primera opción



Ya en la interfaz de EMR, dar click en “Crear clúster”



Crear el clúster con la siguiente información

▼ **Nombre y aplicaciones - obligatorio** [Información](#)
Asigne un nombre a su clúster y elija las aplicaciones que desea instalar en él.

Nombre

My cluster PMORENOQ

Versión de Amazon EMR [Información](#)
Una versión contiene un conjunto de aplicaciones que se puede instalar en el clúster.

emr-7.3.0

Paquete de aplicaciones

Spark
Interactive

Core
Hadoop

Flink








HBase

Presto

Trino

Custom

Paquete de aplicaciones

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

- | | | |
|---|--|--|
| <input type="checkbox"/> AmazonCloudWatchAgent 1.300032.2 | <input type="checkbox"/> Flink 1.18.1 | <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 0.8.0 | <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 0.285 | <input checked="" type="checkbox"/> Spark 3.5.1 |
| <input checked="" type="checkbox"/> Sqoop 1.4.7 | <input type="checkbox"/> TensorFlow 2.16.1 | <input type="checkbox"/> Tez 0.10.2 |
| <input type="checkbox"/> Trino 442 | <input checked="" type="checkbox"/> Zeppelin 0.11.1 | <input checked="" type="checkbox"/> ZooKeeper 3.9.1 |

Configuración del Catálogo de datos de AWS Glue

Utilice el Catálogo de datos de AWS Glue para proporcionar un meta-almacén externo a la aplicación.

- ☒ Usar para metadatos de la tabla de Hive
- ☒ Usar para metadatos de la tabla de Spark

▼ Configuración del clúster - *obligatorio* [Información](#)

Elija un método de configuración para los grupos principales, centrales y de nodos tareas para su clúster.

- ☒ **Grupos de instancias uniformes**
Elija el mismo tipo de instancia de EC2 y la misma opción de compra (bajo demanda o de spot) para todos los nodos de su grupo de nodos. [Más información](#)

- ☐ **Flotas de instancias flexibles**
Elija entre la más amplia variedad de opciones de aprovisionamiento para las instancias de EC2 de su clúster. Diversifique los tipos de instancias y las opciones de compra, y utilice una estrategia de asignación. [Más información](#)

Grupos de instancias uniformes

Principal

Elegir tipo de instancia de EC2

m4.xlarge
4 vCore 16 GiB memoria
Únicamente EBS almacenamiento
Precio bajo demanda: -
Precio de spot más bajo: -

Acciones ▼

▼ Aprovisionamiento y escalado de clústeres - *obligatorio* [Información](#)

Elija cómo Amazon EMR debe dimensionar su clúster.

Elija una opción

☒ **Establecer el tamaño del clúster manualmente**
 Utilice esta opción si conoce los patrones de la carga de trabajo de antemano.

☐ **Utilizar escalado administrado por EMR**
 Supervise las métricas clave de la carga de trabajo de modo que EMR pueda optimizar el tamaño del clúster y la utilización de los recursos.

☐ **Utilizar el escalamiento automático personalizado**
 Para escalar mediante programación los nodos principales y los nodos de tarea, cree políticas de escalamiento automático personalizadas.

Configuración de aprovisionamiento

Establezca el tamaño del principal y tarea grupos de instancias. Amazon EMR intenta aprovisionar esta capacidad al lanzar el clúster.

Nombre	Tipo de instancia	Tamaño de instancia(s)	Utilizar la opción de compra de spot
Central	m4.xlarge	<input type="text" value="1"/>	<input type="checkbox"/>
Tarea - 1	m4.xlarge	<input type="text" value="1"/>	<input type="checkbox"/>

▼ Redes - *obligatorio* [Información](#)

Elija la configuración de red que determina la forma en que usted y otras entidades se comunican con su clúster.

Virtual Private Cloud (VPC) [Información](#)

[Examinar](#)

[Crear VPC](#)

Subred [Información](#)

[Examinar](#)

[Crear subred](#)

► **Grupos de seguridad de EC2 (firewall)**

▼ **Redes - obligatorio** [Información](#)
Elija la configuración de red que determina la forma en que usted y otras entidades se comunican con su clúster.

Virtual Private Cloud (VPC) [Información](#)
vpc-0ea2eb1224155e8dc [Examinar](#) [Crear VPC](#)

Subred [Información](#)
subnet-045d7ff634786d20b [Examinar](#) [Crear subred](#)

▼ **Grupos de seguridad de EC2 (firewall)**

Nodo principal

Grupos de seguridad administrados de EMR
EMR actualizará automáticamente el grupo seleccionado.

Crear ElasticMapReduce-Primary ▼

Grupos de seguridad adicionales - opcional
Seleccione hasta 4 grupos de seguridad adicionales.

Elegir grupos de seguridad adicionales ▼

Nodos principales y de tareas

Grupos de seguridad administrados de EMR
EMR actualizará automáticamente el grupo seleccionado.

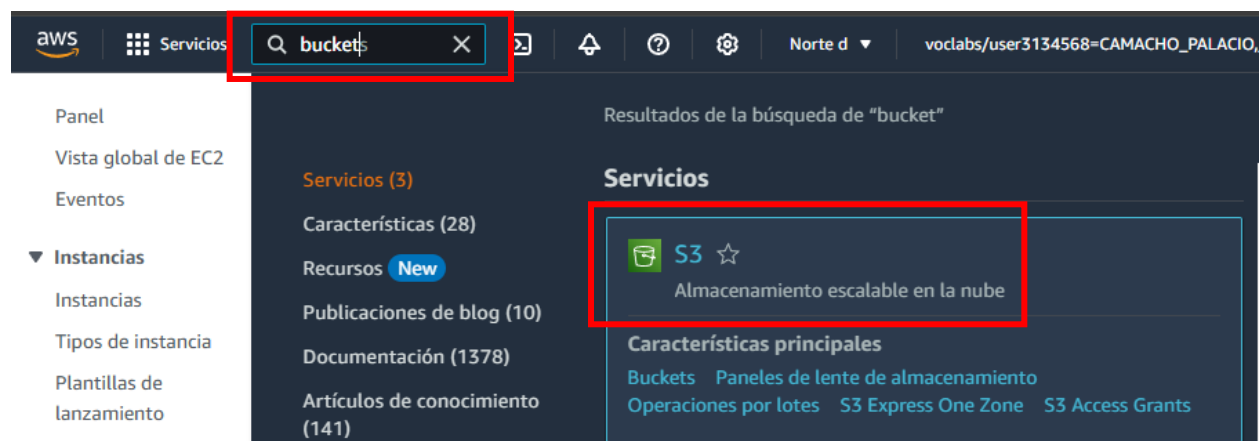
Crear ElasticMapReduce-Core ▼

Grupos de seguridad adicionales - opcional
Seleccione hasta 4 grupos de seguridad adicionales.

Elegir grupos de seguridad adicionales ▼

2. En caso de no tener ningún bucket seguir los siguientes pasos

En la barra de búsqueda buscar “Buckets” o “S3” y seleccionar la opción “S3”



Al estar en la interfaz de S3, dar
os click en la opción “Crear bucket”

Amazon S3 > Buckets > Crear bucket

Crear bucket

Información

Los buckets son contenedores de datos almacenados en S3.

Configuración general

Región de AWS

EE. UU. Este (Norte de Virginia) us-east-1

Tipo de bucket

Información

☒ **Uso general**
Recomendado para la mayoría de los casos de uso y patrones de acceso. Los buckets de uso general son del tipo de bucket de S3 original. Permiten una combinación de clases de almacenamiento que almacenan objetos de forma redundante en múltiples zonas de disponibilidad.

☐ **Directorio**
Recomendado para casos de uso de baja latencia. Estos buckets utilizan únicamente la clase de almacenamiento S3 Express One Zone, que proporciona un procesamiento más rápido de los datos dentro de una única zona de disponibilidad.

Nombre del bucket

Información

labsbucket-pmorenoq

El nombre del bucket debe ser único dentro del espacio de nombres global y seguir las reglas de nomenclatura del bucket. [Consulte las reglas para la asignación de nombres de buckets](#)

Copiar la configuración del bucket existente: *opcional*

Solo se copia la configuración del bucket en los siguientes ajustes.

Elegir el bucket

Formato: s3://bucket/prefijo

Cifrado predeterminado

Información

El cifrado del lado del servidor se aplica automáticamente a los nuevos objetos almacenados en este bucket.

Tipo de cifrado

Información

☒ Cifrado del servidor con claves administradas de Amazon S3 (SSE-S3)

☐ Cifrado del servidor con claves de AWS Key Management Service (SSE-KMS)

☐ Cifrado de doble capa del servidor con claves de AWS Key Management Service (DSSE-KMS)
Proteja sus objetos con dos capas de cifrado independientes. Para obtener más información sobre los precios, consulte [DSSE-KMS pricing](#) (Precios de DSSE-KMS) en la pestaña Storage (Almacenamiento) de la [página de precios de Amazon S3](#)

Clave de bucket

El uso de una clave de bucket de S3 para SSE-KMS reduce los costos de cifrado al reducir las llamadas a AWS KMS. Las claves de bucket de S3 no son compatibles con DSSE-KMS. [Más información](#)

☐ Desactivar

☒ Habilitar

► Configuración avanzada

Después de crear el bucket, puede cargar archivos y carpetas, y configurar ajustes adicionales en él.

Cancelar

Crear bucket

Amazon S3 > Buckets

El bucket "labbucket-pmorenoq" se creó correctamente
Para cargar archivos y carpetas, o para configurar ajustes adicionales del bucket, elija [Ver detalles](#).

Instantánea de la cuenta: *actualizada cada 24 horas* Todas las regiones de AWS
[Ver panel de Storage Lens](#)
 Storage Lens ofrece visibilidad sobre el uso del almacenamiento y las tendencias de la actividad. [Más información](#)

Buckets de uso general | Buckets de directorio

Buckets de uso general (1) Información Todas las regiones de AWS

[Copiar ARN](#) [Vaciar](#) [Eliminar](#) [Crear bucket](#)

Los buckets son contenedores de datos almacenados en S3.

Buscar buckets por nombre

Nombre	Región de AWS	Analizador de acceso de IAM	Fecha de creación
labbucket-pmorenoq	EE. UU. Este (Norte de Virginia) us-east-1	Ver analizador para us-east-1	21 Nov 2024 4:33:26 PM -05

*Después de creado el bucket debemos guardar el nombre de nuestro bucket para utilizarlo mas adelante

- Volvemos a nuestra interfaz de creación del clúster y agregamos el siguiente código en la configuración de software

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

▼ Configuración de software [Información](#)
Anule las configuraciones predeterminadas para aplicaciones específicas de su clúster.

☒ Ingresar la configuración ☐ Cargar JSON desde Simple Storage Service (Amazon S3)

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

JSON Ln 6, Col 22 ✖ : 0 ⚠ : 0

Debemos establecer previamente el modo de acceso a nuestras instancias, en mi caso es mediante la llave pem “labsKey” para hacer conexiones mediante ssh.

▼ Configuración de seguridad y par de claves de EC2 [Información](#)
Elija una configuración de seguridad o cree una nueva que pueda reutilizar con otros clústeres.

Configuración de seguridad
Seleccione la configuración del servicio de cifrado, autenticación, autorización y metadatos de instancia del clúster.

Par de claves de Amazon EC2 para el protocolo SSH al clúster [Información](#)

▼ Roles de Identity and Access Management (IAM) - *obligatorio* [Información](#)

Elija o cree un rol de servicio y un perfil de instancia para las instancias de EC2 del clúster.

Rol de servicio de Amazon EMR [Información](#)

El rol de servicio es un rol de IAM que Amazon EMR asume para aprovisionar recursos y realizar acciones de nivel de servicio con otros servicios de AWS.

- ☒ Elegir un rol de servicio existente
- Seleccione un rol de servicio predeterminado o un rol personalizado con políticas de IAM asociadas para que el clúster pueda interactuar con otros servicios de AWS.

- ☐ Crear un rol de servicio
- Deje que Amazon EMR cree un nuevo rol de servicio para que pueda conceder y restringir el acceso a los recursos de otros servicios de AWS.

Rol de servicio

EMR_DefaultRole



Perfil de instancia de EC2 para Amazon EMR

El perfil de instancia asigna un rol a cada instancia de EC2 de un clúster. El perfil de instancia debe especificar un rol que pueda acceder a los recursos de los pasos y las acciones de arranque.

- ☒ Elegir un perfil de instancia existente
- Seleccione un rol predeterminado o un perfil de instancia personalizado con políticas de IAM asociadas para que el clúster pueda interactuar con sus recursos de Amazon S3.

- ☐ Crear un perfil de instancia
- Deje que Amazon EMR cree un nuevo perfil de instancia para que pueda especificar un conjunto personalizado de recursos a los que tendrá acceso en Amazon S3.

Perfil de instancia

EMR_EC2_DefaultRole



Rol de escalamiento automático personalizado - *opcional*

Cuando se activa una regla de escalamiento automático personalizada, Amazon EMR asume esta función para agregar y finalizar instancias de EC2. [Más información](#)

Rol de escalamiento automático personalizado

EMR_AutoScaling_DefaultRole



[Crear rol de IAM](#)

Posterior a tener nuestra configuración finalizada damos click en crear clúster y nos debe llevar a la siguiente interfaz.

The screenshot shows the AWS Management Console interface for the 'My cluster PMORENOQ' page. The top navigation bar includes the AWS logo, a search bar, and the user's account information. The breadcrumb trail shows 'Amazon EMR' > 'EMR en EC2: Clústeres' > 'My cluster PMORENOQ'. A green banner at the top indicates that the cluster was created successfully.

My cluster PMORENOQ
Se ha actualizado hace menos de un minuto

Resumen

Información del clúster	Aplicaciones	Administración de clústeres	Estado y hora
ID del clúster j-JH8I77W1D70T Configuración del clúster Grupos de instancias Capacidad 1 Primary (Principal) 1 Principal 1 Tarea	Versión de Amazon EMR emr-7.3.0 Aplicaciones instaladas HCatalog 3.1.3, Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2.6.0, JupyterHub 1.5.0, Livy 0.8.0, Spark 3.5.1, Sqoop 1.4.7, Zeppelin 0.11.1, ZooKeeper 3.9.1	Destino del registro en Amazon S3 aws-logs-637423661635-us-east-1/elasticmapreduce DNS público del nodo principal -	Estado Comenzando Hora de creación 21 de noviembre de 2024 16:38 (UTC-05:00) Tiempo transcurrido 0 segundos

Propiedades | Acciones de arranque | Instancias (hardware) | Pasos | Aplicaciones | Config

Registros de clúster Información

Archivar los archivos de registro en Amazon S3
Activado

Ubicación de Amazon S3
s3://aws-logs-637423661635-us-east-1/elasticmapreduce/

Cifrado para registros
Desactivado

Terminación del clúster y reemplazo de nodos Información

Editar

Opción de terminación
Terminar automáticamente el clúster después del tiempo de inactividad

Tiempo de inactividad
1 hora

Protección contra la terminación
Desactivado

Clústeres (1) Información

Ver detalles | Terminar | Clonar | Crear clúster

Filtrar clústeres por estado | Buscar clústeres | Filtrar clústeres por fecha y hora de creación

	ID del clúster	Nombre del clúster	Estado	Hora de creación
<input type="checkbox"/>	j-JH8I77W1D70T	My cluster PMORENOQ	Comenzando Preparación de clúster	21 de noviembre 05:00 16:38

Luego de haber creado nuestro clúster debemos esperar de 10 a 20 minutos para poderlo usar

The screenshot shows the Amazon EMR console. The breadcrumb navigation indicates 'EMR en EC2: Clústeres'. The main heading is 'Clústeres (1) Información'. Below this, there are filters and a search bar. A table lists the clusters. The first cluster, 'j-JH8I77W1D70T' with name 'My cluster PMORENOQ', is highlighted with a red box around its status 'Esperando' (Waiting) and the sub-status 'Listo para ejecutar pasos' (Ready to execute steps). The creation time is '21 de noviembre de 2020 16:38'.

ID del clúster	Nombre del clúster	Estado	Hora de creación (UTC-05:00)
j-JH8I77W1D70T	My cluster PMORENOQ	Esperando Listo para ejecutar pasos	21 de noviembre de 2020 16:38

Accedemos a la barra de navegación a la izquierda y buscamos la opción de “Bloquear acceso al público”

The screenshot shows the 'Bloquear el acceso público' (Block public access) configuration page in the Amazon EMR console. The left sidebar shows the navigation menu with 'Bloquear el acceso público' highlighted with a red box. The main content area shows the configuration for blocking public access, which is currently 'Activado' (Activated). A red box highlights the 'Editar' (Edit) button. Below this, there is a section for 'Excepciones de rango de puertos' (Port range exceptions) with a list of ports, including 22.

Bloquear el acceso público

El bloqueo del acceso público de Amazon EMR impide el lanzamiento de un clúster cuando está asociado a reglas del grupo de seguridad que permiten el tráfico entrante desde IPv4 0.0.0.0/0 o IPv6 ::/0 (acceso público) en un puerto, a menos que el puerto se especifique explícitamente como una excepción.

Configuración del bloqueo del acceso público

Bloquear el acceso público
 Activado

Editar

Excepciones de rango de puertos
 Un clúster se puede lanzar con reglas del grupo de seguridad que permiten el tráfico entrante desde todas las direcciones IP públicas en estos puertos. El puerto 22 se agrega como una excepción de manera predeterminada por SSH.

22

Cuando estemos editando el acceso al público lo debemos desactivar el bloqueo

aws [Logo] [Icon] [Icon] [Icon] [Icon] [Icon] [Icon] Norte de voclabs/user3134568=CAMACHO_PALACIO_JUAN_S

Amazon EMR > EMR en EC2: Bloquear el acceso público > Editar configuración

Amazon EMR <

EMR sin servidor

▼ **EMR en EC2**

- Clústeres
- Blocs de notas y repositorios de Git
- Eventos
- [Bloquear el acceso público](#)
- Configuraciones de seguridad

▼ **EMR en EKS**

- Clústeres virtuales

Editar configuración

Configuración del bloqueo del acceso público

Bloquear el acceso público
Este cambio solo afecta a los nuevos clústeres de EMR. Los clústeres de EMR existentes no se ven afectados.

☐ Activar (recomendado)
Bloquear el acceso público a todos los puertos, excepto los que se agregan como excepciones.

☒ **Desactivar**
Permitir el acceso público en función de las reglas del grupo de seguridad.

Cancelar **Guardar**

Debemos ver lo siguiente en pantalla

aws [Logo] [Icon] [Icon] [Icon] [Icon] [Icon] [Icon] Norte de voclabs/user3134568=CAMACHO_PALACIO_JUAN_S

Amazon EMR > EMR en EC2: Bloquear el acceso público

Amazon EMR <

EMR sin servidor

▼ **EMR en EC2**

- Clústeres
- Blocs de notas y repositorios de Git
- Eventos
- [Bloquear el acceso público](#)
- Configuraciones de seguridad

▼ **EMR en EKS**

- Clústeres virtuales

✓ La configuración del bloqueo del acceso público para esta cuenta se ha actualizado correctamente.

Bloquear el acceso público [Información](#)

El bloqueo del acceso público de Amazon EMR impide el lanzamiento de un clúster cuando está asociado a reglas del grupo de seguridad que permiten el tráfico entrante desde IPv4 0.0.0.0/0 o IPv6 ::/0 (acceso público) en un puerto, a menos que el puerto se especifique explícitamente como una excepción.

Configuración del bloqueo del acceso público [Editar](#)

Bloquear el acceso público
⚠ Desactivado

4. Accediendo a nuestro clúster mediante EC2

Accedemos a nuestro clúster y buscamos la dirección del DNS público de nuestro nodo máster

The screenshot shows the Amazon EMR console for cluster **PMORENOQ**. The cluster is in the **Esperando** (Waiting) state. The **Resumen** (Summary) section displays the following information:

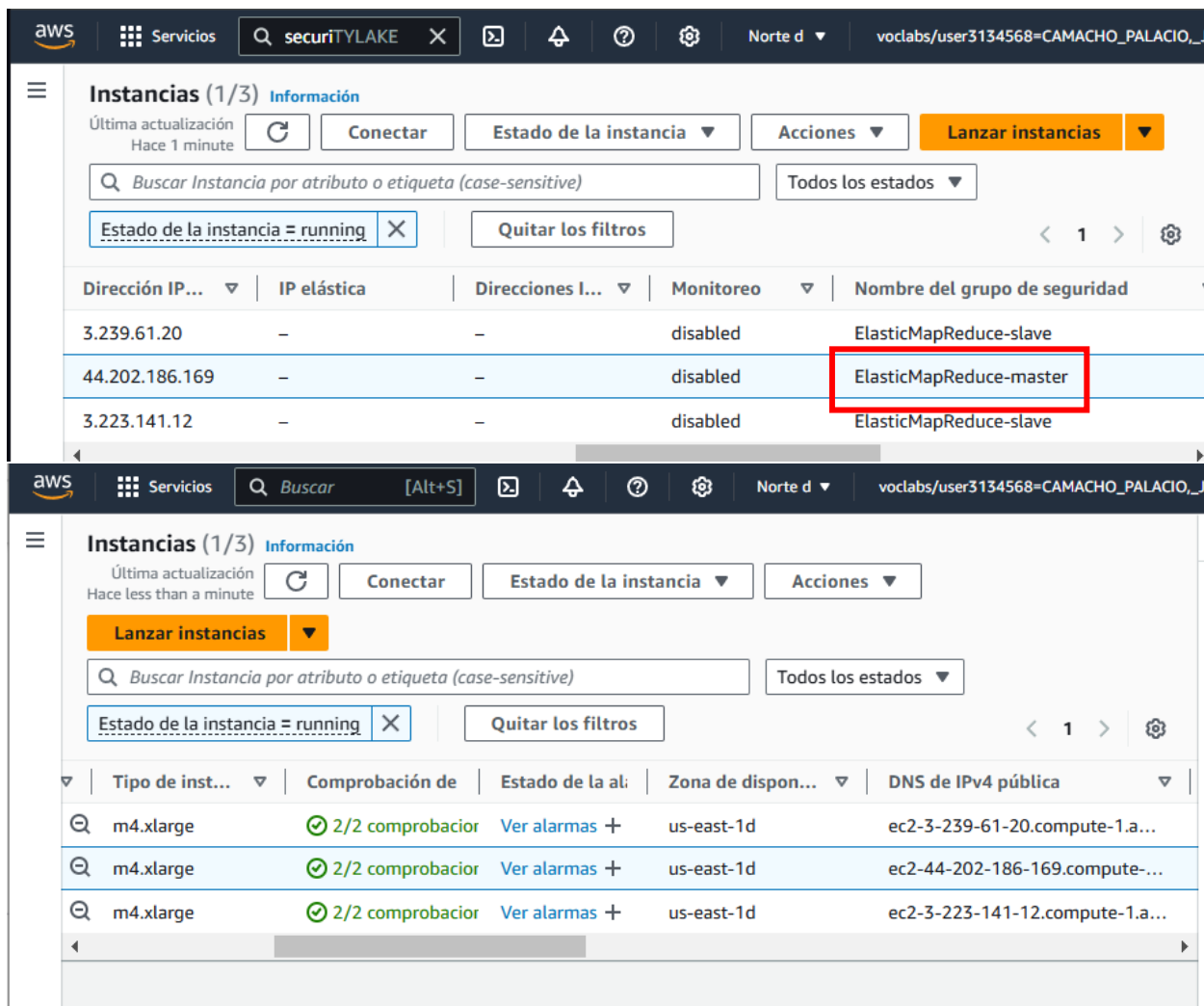
- Información del clúster:**
 - ID del clúster: j-JH8I77W1D70T
 - Configuración del clúster: Grupos de instancias
 - Capacidad: 1 Primary (Principal) | 1 Principal | 1 Tarea
- Aplicaciones:**
 - Versión de Amazon EMR: emr-7.3.0
 - Aplicaciones instaladas: HCatalog 3.1.3, Hadoop 3.3.6, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2.6.0, JupyterHub 1.5.0, Livy 0.8.0, Spark 3.5.1, Sqoop 1.4.7, Zeppelin 0.11.1, ZooKeeper 3.9.1
- Administración de clústeres:**
 - Destino del registro en Amazon S3: [aws-logs-637423661635-us-east-1/elasticmapreduce](#)
 - IU de aplicación persistente: [Servidor de historial de Spark](#)
 - Servidor de línea de tiempo de YARN: [Conectarse al nodo principal mediante SSH](#)
 - UI de Tez: [Conectarse al nodo principal mediante SSM](#)
- Estado y hora:**
 - Estado: ✓ Esperando
 - Hora de creación: 21 de noviembre de 2024 16:38 (UTC-05:00)
 - Tiempo transcurrido: 34 minutos, 6 segundos

The **DNS público del nodo principal** is highlighted with a red box and shows the address: [ec2-44-202-186-169.compute-1.amazonaws.com](#).

En la barra de búsqueda buscamos “EC2” y vamos a la opción del mismo nombre

The screenshot shows the AWS search bar with the text **ec2** entered. The search results are displayed under the **Servicios** (Services) tab, showing **EC2** as the top result. The **EC2** service is highlighted with a red box, and the description **Servidores virtuales en la nube** (Virtual servers in the cloud) is visible.

Buscamos cual es nuestro nodo master



The image displays two screenshots of the AWS Management Console, specifically the EC2 Instances page, used to identify the master node of an Elastic MapReduce cluster.

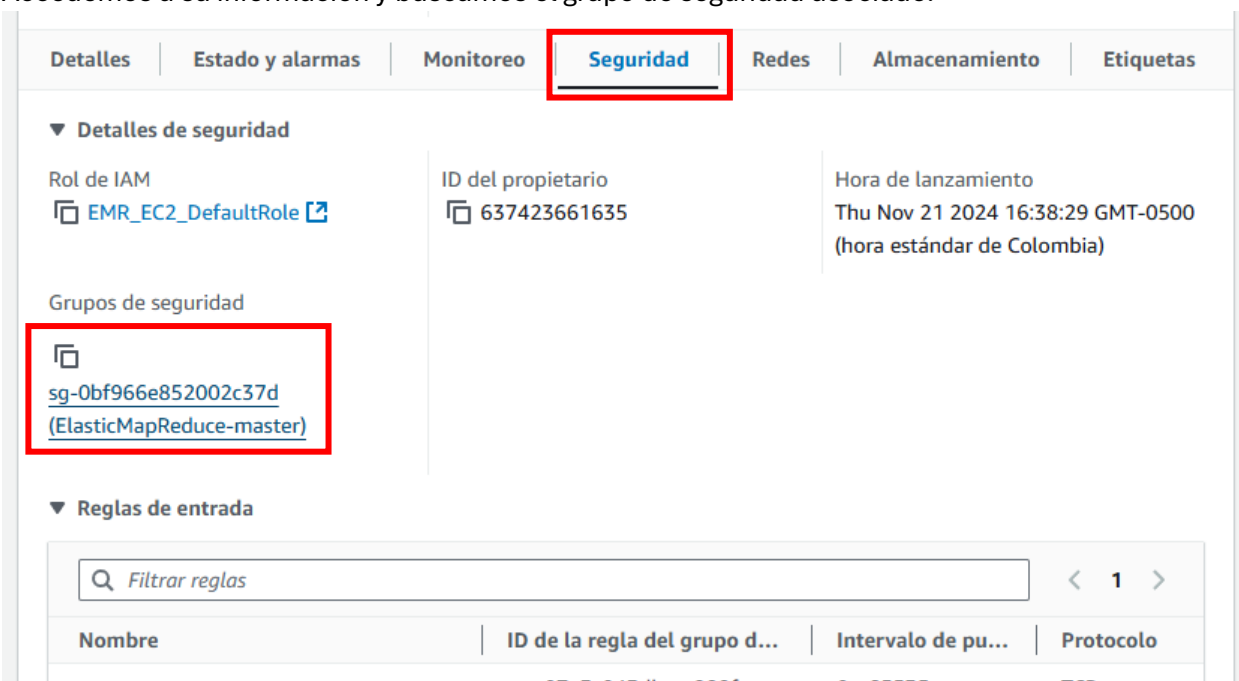
Top Screenshot: Shows a list of three EC2 instances. The instance with IP address 44.202.186.169 is highlighted, and its name, **ElasticMapReduce-master**, is circled in red. The other two instances are named ElasticMapReduce-slave.

Dirección IP...	IP elástica	Direcciones I...	Monitoreo	Nombre del grupo de seguridad
3.239.61.20	-	-	disabled	ElasticMapReduce-slave
44.202.186.169	-	-	disabled	ElasticMapReduce-master
3.223.141.12	-	-	disabled	ElasticMapReduce-slave

Bottom Screenshot: Shows a more detailed view of the same three instances. All instances are of type m4.xlarge, in the us-east-1d zone, and have a public IPv4 DNS address.

Tipo de inst...	Comprobación de	Estado de la al	Zona de dispon...	DNS de IPv4 pública
m4.xlarge	2/2 comprobaci...	Ver alarmas +	us-east-1d	ec2-3-239-61-20.compute-1.a...
m4.xlarge	2/2 comprobaci...	Ver alarmas +	us-east-1d	ec2-44-202-186-169.compute-...
m4.xlarge	2/2 comprobaci...	Ver alarmas +	us-east-1d	ec2-3-223-141-12.compute-1.a...

Accedemos a su información y buscamos el grupo de seguridad asociado.



The screenshot shows the AWS IAM console interface. The 'Seguridad' tab is selected. Under 'Detalles de seguridad', the 'Rol de IAM' is 'EMR_EC2_DefaultRole'. The 'ID del propietario' is '637423661635'. The 'Hora de lanzamiento' is 'Thu Nov 21 2024 16:38:29 GMT-0500 (hora estándar de Colombia)'. In the 'Grupos de seguridad' section, the security group 'sg-0bf966e852002c37d (ElasticMapReduce-master)' is highlighted. Below, the 'Reglas de entrada' section shows a table with one rule.

Nombre	ID de la regla del grupo d...	Intervalo de pu...	Protocolo
	sg-0bf966e852002c37d	0-65535	TCP

Accedemos al grupo de seguridad y editamos las reglas de entrada.

Debemos brindar acceso a los puertos TCP

- 8890
- 22
- 9870
- 8888
- 14000
- 9443

Debemos ver que todos los puertos están abiertos

Reglas de entrada (13)

Administrar etiquetas

Editar reglas de entrada

Buscar

1

	▼	Protocolo	▼	Intervalo de puertos	▼	Origen	▼	Descripción
do		TCP		8890		0.0.0.0/0		–
		TCP		0 - 65535		sg-05bf5280dabb233...		–
Pv4		ICMP		Todo		sg-05bf5280dabb233...		–
		TCP		22		0.0.0.0/0		–
		UDP		0 - 65535		sg-0bf966e852002c37...		–
Pv4		ICMP		Todo		sg-0bf966e852002c37...		–
do		TCP		8443		pl-f8bd5e91		–
do		TCP		9870		0.0.0.0/0		–
do		TCP		8888		0.0.0.0/0		–
		UDP		0 - 65535		sg-05bf5280dabb233...		–
		TCP		0 - 65535		sg-0bf966e852002c37...		–
do		TCP		14000		0.0.0.0/0		–
do		TCP		9443		0.0.0.0/0		–

5. Acceder a los servicios del clúster

Debemos volver a nuestro clúster EMR, dar click en nuestro clúster ya activo y dirigirnos hacia aplicaciones


Debemos encontrar la siguiente información

Damos click en el link que nos lleva al servicio HUE

IU de la aplicación en el nodo principal

Habilitar una conexión SSH

Estas requieren que el túnel de SSH esté habilitado.

Aplicación	URL de la IU 
Administrador de recursos	http://ec2-3-88-211-26.compute-1.amazonaws.com:8088/
JupyterHub	https://ec2-3-88-211-26.compute-1.amazonaws.com:9443/
Livy	http://ec2-3-88-211-26.compute-1.amazonaws.com:8998/
Nodo del nombre de HDFS	http://ec2-3-88-211-26.compute-1.amazonaws.com:9870/
Servidor de historial de Spark	http://ec2-3-88-211-26.compute-1.amazonaws.com:18080/
Tonalidad	http://ec2-3-88-211-26.compute-1.amazonaws.com:8888/
Zeppelin	http://ec2-3-88-211-26.compute-1.amazonaws.com:8890/

IU de aplicaciones en los nodos principales y de tareas

Aplicación	URL de la IU
Administrador de nodos	http://ec2-000-000-000-000.compute-1.amazonaws.com:8042/
Nodo de datos de HDFS	http://ec2-000-000-000-000.compute-1.amazonaws.com:9864/

Aplicaciones instaladas (11)

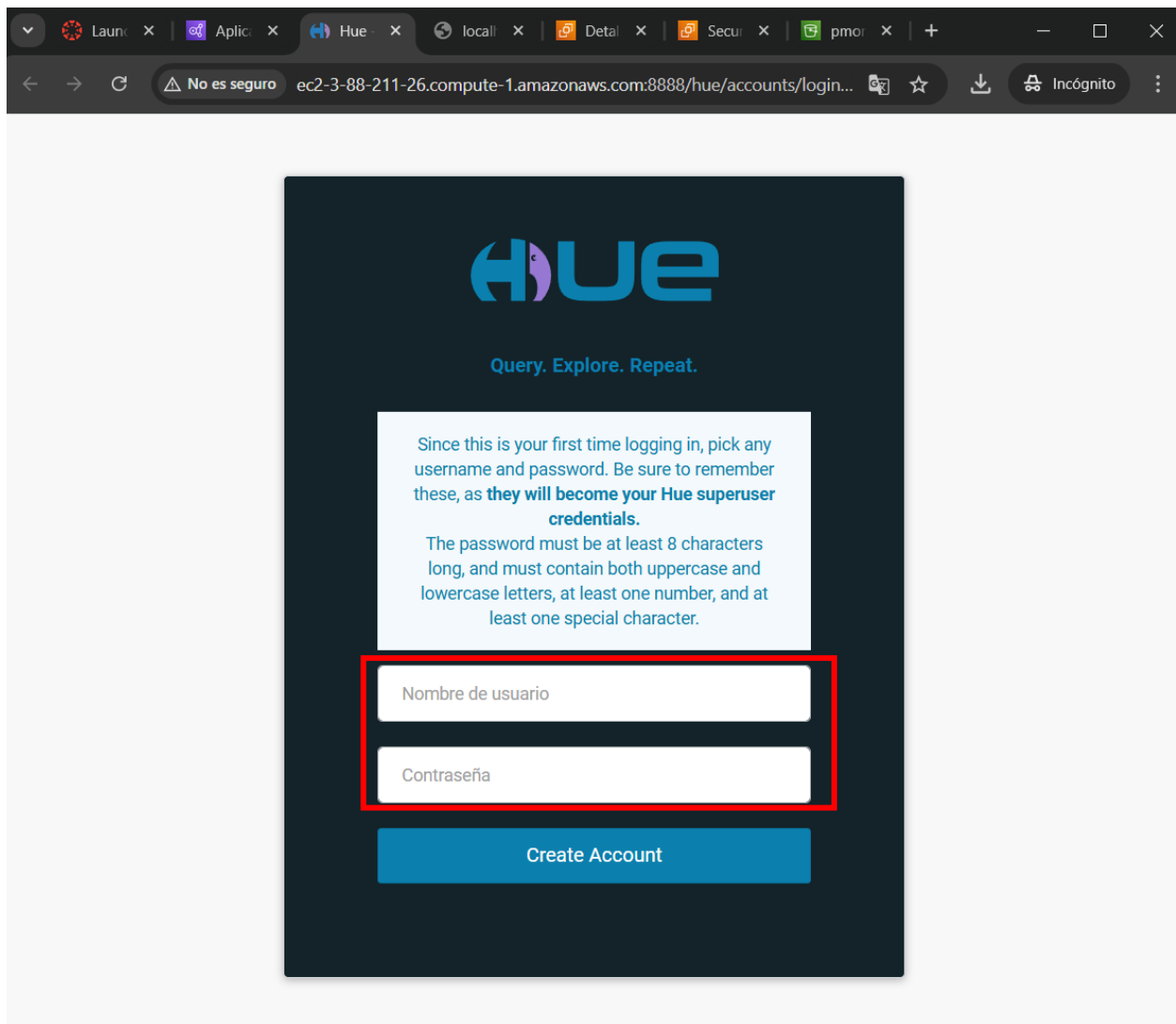
Hadoop 3.3.6	HCatalog 3.1.3	Hive 3.1.3	Hue 4.11.0
JupyterEnterpriseGateway 2.6.0	JupyterHub 1.5.0	Livy 0.8.0	Spark 3.5.1
Sqoop 1.4.7	Zeppelin 0.11.1	ZooKeeper 3.9.1	

Al acceder al link debemos ver la siguiente interfaz

Cuando iniciemos por primera vez al servicio nos pedirá crear credenciales

Usuario sugerido: hadoop

Contraseña: <usar la que desee>



Laun: x | Aplic: x | Hue: x | local: x | Detal: x | Secu: x | pmo: x | +

← → ↻ No es seguro ec2-3-88-211-26.compute-1.amazonaws.com:8888/hue/accounts/login... ☆ ↕ Incógnito

HUE

Query. Explore. Repeat.

Since this is your first time logging in, pick any username and password. Be sure to remember these, as **they will become your Hue superuser credentials.**

The password must be at least 8 characters long, and must contain both uppercase and lowercase letters, at least one number, and at least one special character.

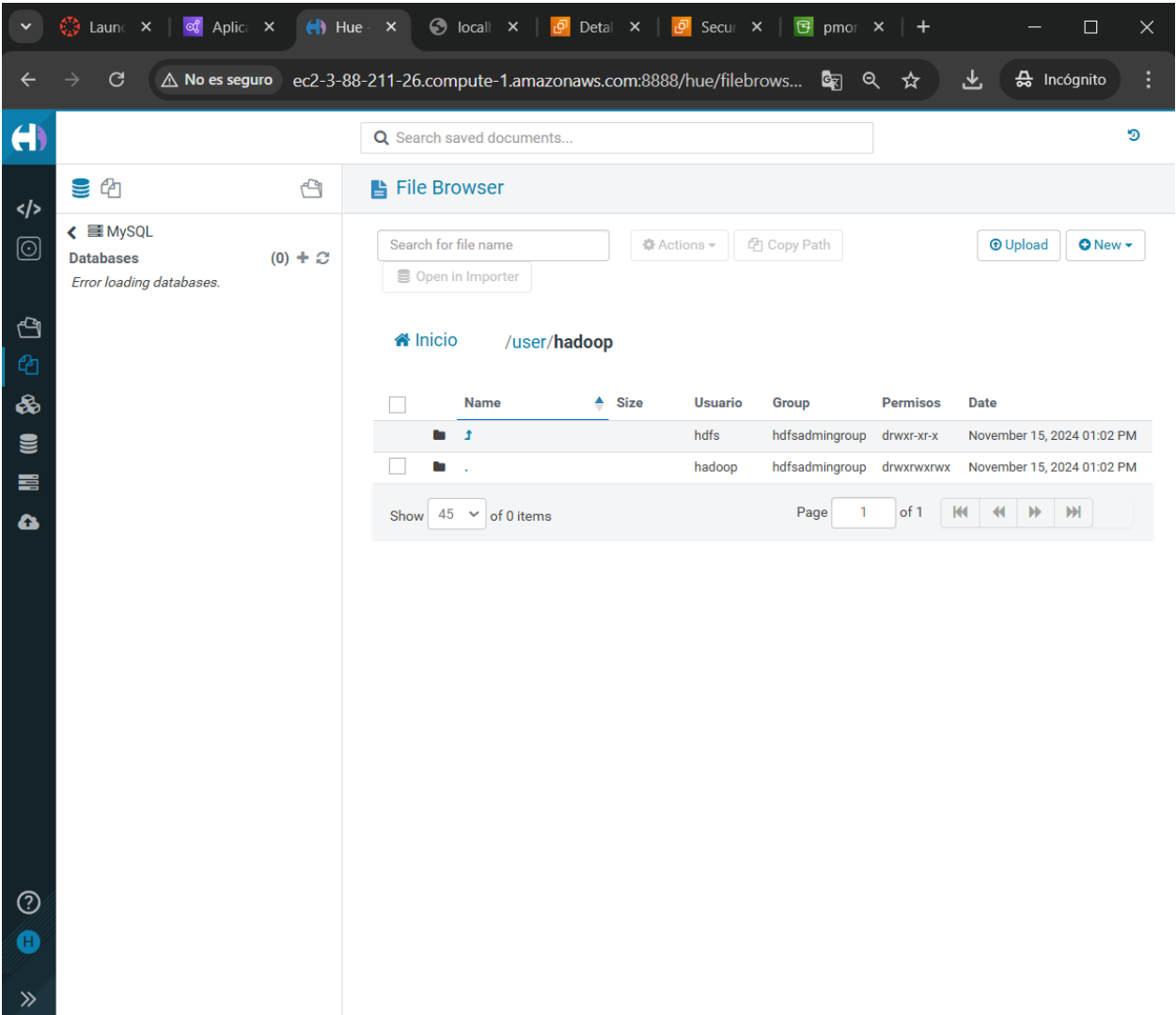
Nombre de usuario

Contraseña

Create Account

Luego de acceder debemos ver la siguiente interfaz

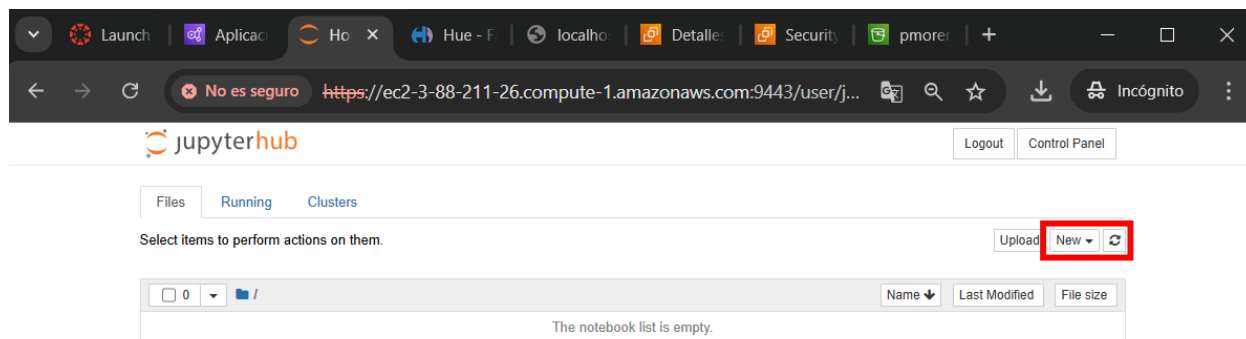
Ya nuestro servicio de HUE está listo para ser utilizado



Ahora Para acceder al servicio de jupyterHub

IU de la aplicación en el nodo principal		Habilitar una conexión SSH
Estas requieren que el túnel de SSH esté habilitado.		
Aplicación	URL de la IU	
Administrador de recursos	http://ec2-3-88-211-26.compute-1.amazonaws.com:8088/	
JupyterHub	https://ec2-3-88-211-26.compute-1.amazonaws.com:9443/	
Livy	http://ec2-3-88-211-26.compute-1.amazonaws.com:8998/	
Nodo del nombre de HDFS	http://ec2-3-88-211-26.compute-1.amazonaws.com:9870/	
Servidor de historial de Spark	http://ec2-3-88-211-26.compute-1.amazonaws.com:18080/	
Tonalidad	http://ec2-3-88-211-26.compute-1.amazonaws.com:8888/	
Zeppelin	http://ec2-3-88-211-26.compute-1.amazonaws.com:8890/	

Debemos llegar a la siguiente interfaz



Para probar el servicio damos en “nuevo” y seleccionamos un archivo PySpark

Luego introducimos la siguiente información y ejecutamos el programa

