

## Creación de cluster

The screenshot shows the AWS EMR console with the following details:

- Cluster ID:** j-2DBDPY7BW9FA
- Cluster ARN:** arn:aws:elasticmapreduce:us-east-1:923015016075:cluster/j-2DBDPY7BW9FA
- Cluster configuration:** Instance groups (1 Primary, 1 Core, 1 Task)
- Operating system:** Amazon Linux release 2023.9.20251014.0
- Applications:** Amazon EMR version emr-7.11.0, installed applications include Flink 1.20.0, HCatalog 3.1.3, Hadoop 3.4.1, Hive 3.1.3, Hue 4.11.0, JupyterHub 1.5.0, Livy 0.8.0, Spark 3.5.6, Tez 0.10.2, Zeppelin 0.11.1.
- Cluster management:** Log destination in Amazon S3 emrsamuel, Persistent application UEs (Spark History Server, YARN Timeline Server, Tez UI).
- Status and time:** Status Waiting, Creation time November 09, 2025, 22:03 (UTC-05:00), Elapsed time 1 hour, 9 minutes.
- Cluster logs:** Archive log files to Amazon S3 Turned on, Amazon S3 location s3://emrsamuel/. Encryption for logs Turned off.
- Cluster termination and node replacement:** Termination option Automatically terminate cluster after idle time (Idle time 1 hour), Termination protection Off, Unhealthy node replacement On.

## Acceso ssh

```
LAST LOGIN: Mon Nov 17 02:00:08 2025

The terminal session shows the user has logged in via SSH to the Amazon Linux 2023 cluster. The session includes a welcome banner, a detailed ASCII art logo, and a standard Linux login message. The user then runs an 'hdfs dfs -ls /user' command, which lists the contents of the '/user' directory, showing various HDFS files and directories owned by users like hadoop, mapred, hue, livy, oozie, root, spark, and zeppelin, all belonging to the 'hdfsadmingroup' group.
```

## Security groups del cluster

ID de la regla del grupo de seguridad	Tipo	Protocolo	Intervalo de puertos	Origen	Descripción: opcional
sgr-0d56431f19ca7120b	TCP personalizado	TCP	9870	Personas...	0.0.0.0/0
sgr-0efdb63a66b23966e	TCP personalizado	TCP	8890	Personas...	0.0.0.0/0
sgr-052288fefc6cb2e17	TCP personalizado	TCP	8443	Personas...	pl-f8bd5e91
sgr-011a04e586d9e8960	TCP personalizado	TCP	8888	Personas...	0.0.0.0/0
sgr-01d7a6ebb354c7b95	TCP personalizado	TCP	9443	Personas...	0.0.0.0/0
sgr-03f32252e1aa8df9c	Todos los UDP	UDP	0 - 65535	Personas...	0.0.0.0/0
sgr-0cbc9ee8a7bde42dc	Todos los ICMP IPv4	ICMP	Todo	Personas...	sg-094a6e3d4ffc780d8
sgr-0201718b901fc665e	SSH	TCP	22	Personas...	sg-094a6e3d4ffc780d8
sgr-0897276182a3fc0b9	TCP personalizado	TCP	14000	Personas...	0.0.0.0/0

## Actividades del video

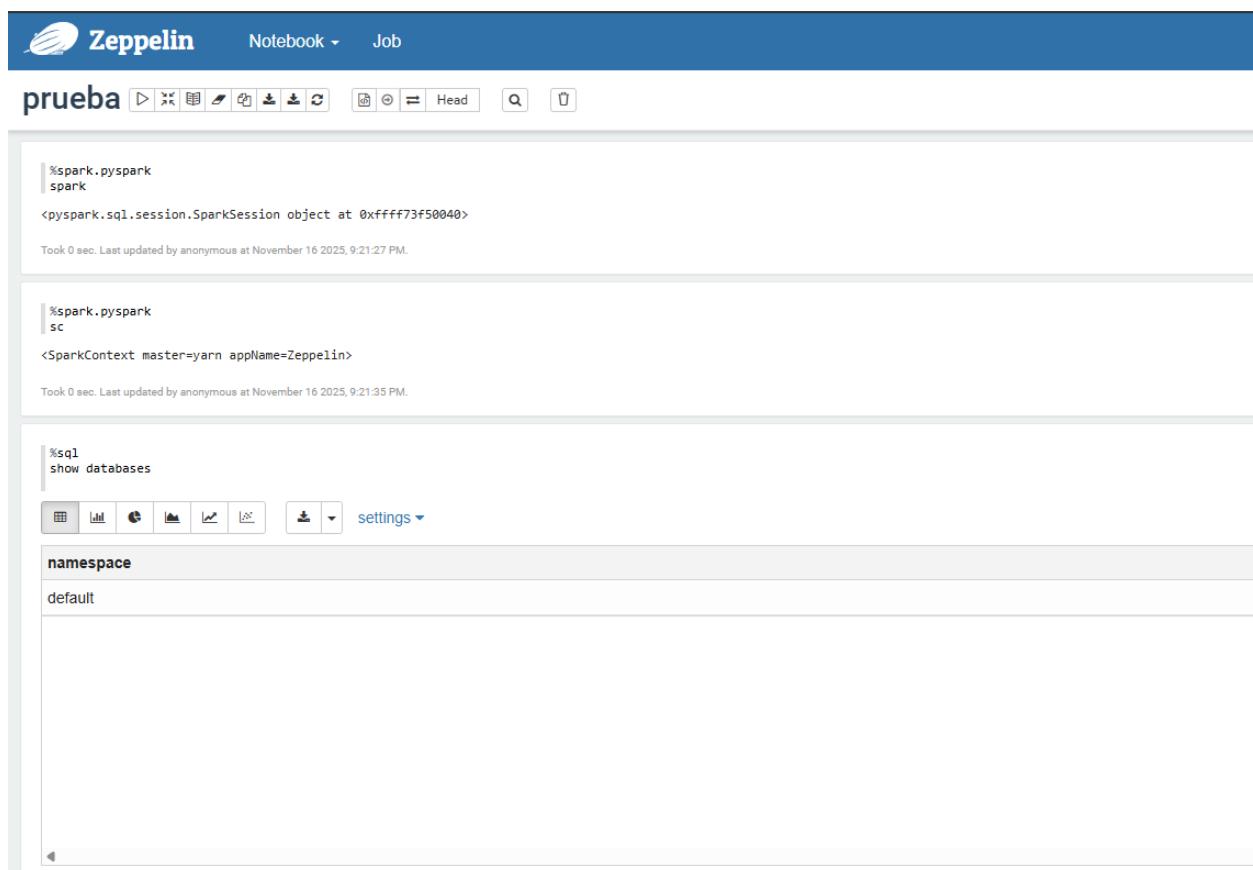
### Acceso a JupyterHub

```
In [1]: spark
Starting Spark application
ID      YARN Application ID  Kind  State  Spark UI  Driver log  User  Current session?
0  application_1763344109923_0001  pyspark  idle          None    ✓

SparkSession available as 'spark'.
<pyspark.sql.session.SparkSession object at 0xfffff92ac3730>

In [2]: sc
<SparkContext master=yarn appName=livy-session-0>
```

## Acceso a Athena



The screenshot shows the Zeppelin Notebook interface with a blue header bar. The header includes the Zeppelin logo, the word "Zeppelin", a "Notebook" dropdown, and a "Job" button. Below the header, the notebook title is "prueba". The interface has three main sections:

- Section 1:** Contains the command "%spark.pyspark" and its output, which shows the creation of a SparkSession object at memory address 0xfffff73f50040. It also indicates a duration of 0 seconds and was last updated at November 16 2025, 9:21:27 PM.
- Section 2:** Contains the command "%spark.pyspark" and its output, which shows the creation of a SparkContext master=yarn appName=Zeppelin. It also indicates a duration of 0 seconds and was last updated at November 16 2025, 9:21:35 PM.
- Section 3:** Contains the command "%sql show databases" and its output, which lists the database "default". The output table has two columns: "namespace" and "default".