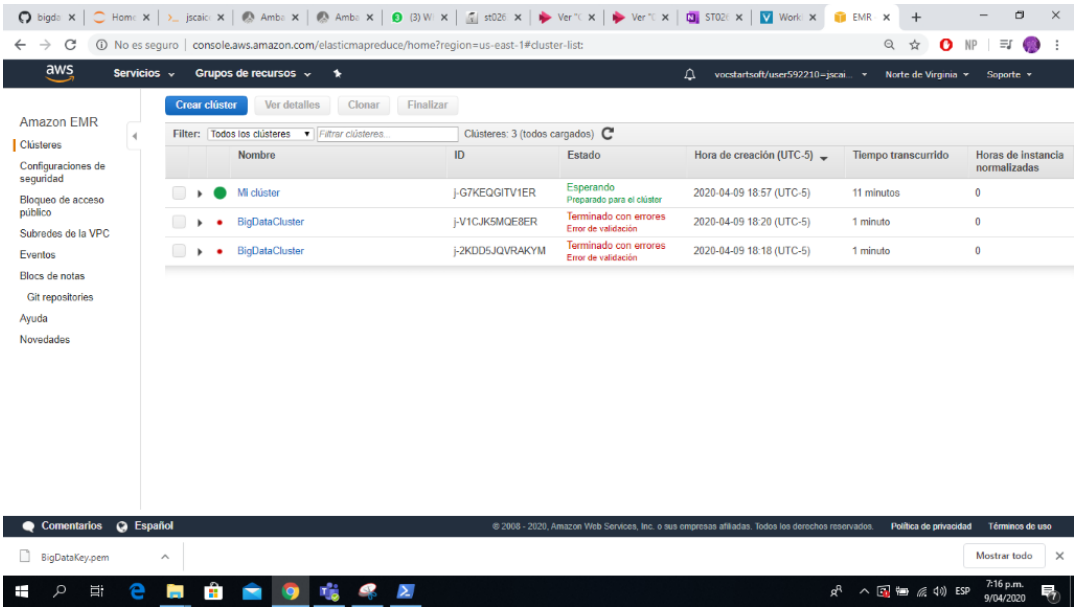
1. HDFS

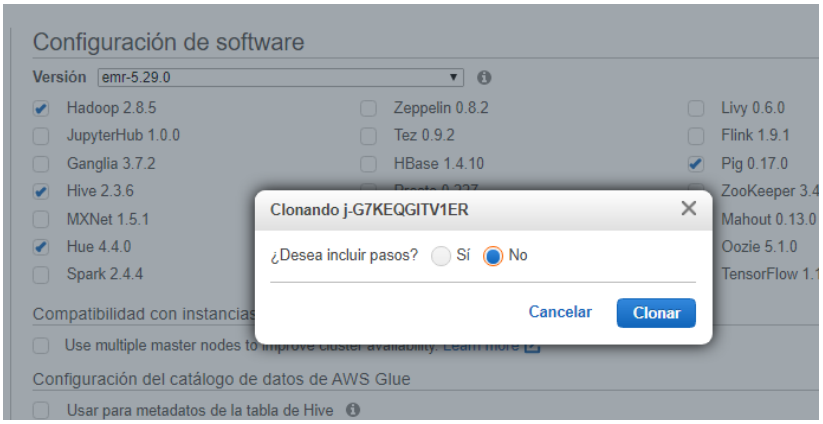
1. Crear y gestionar Clusters Amazon EMR

- Creación de cluster por primera vez de forma interactiva siguiendo el video y habilitando los puertos en bloqueo de acceso público:

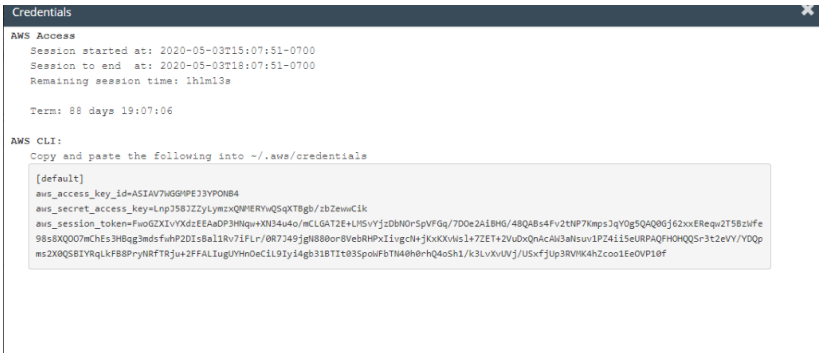


- Terminación del clúster:

- Recreación del clúster:



- Por comando: Se actualizan las credenciales de aws cli

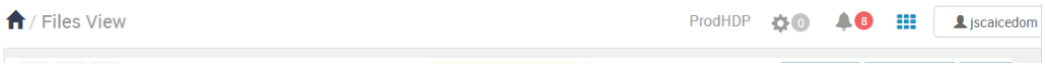


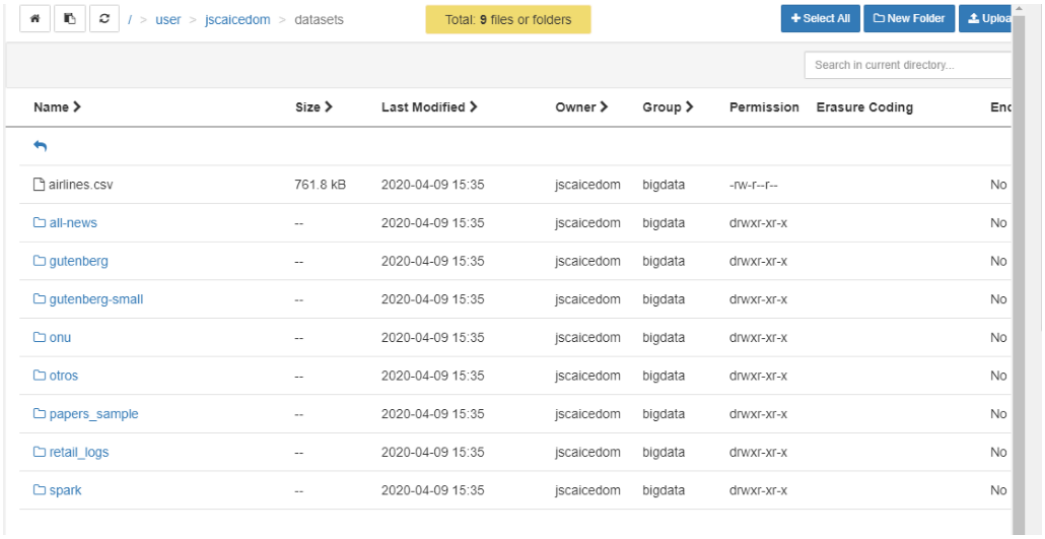
Comando para recreación de clúster por comando, y para la destrucción es necesario el ID o nombre

2. Gestión de archivos en S3 y HDFS

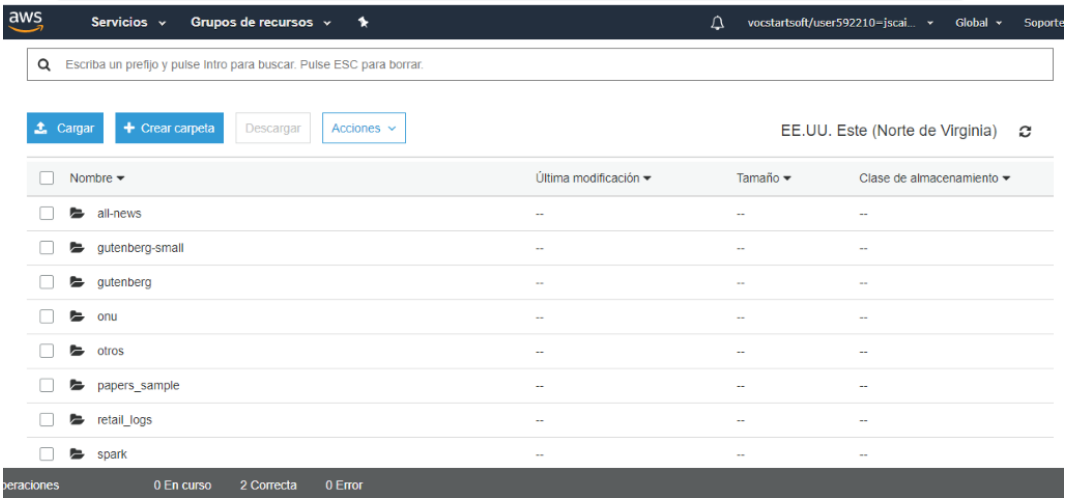
- Copiar datasets desde Shell en la 192.168.10.116 hacia HDFS/DCA:

  - Ingresé al dca por medio de jupyter y creé la carpeta de datasets en hdfs (hdfs dfs –mkdir /user/jscaicedom/datasets)
  - Descargué el github con los datos y los copié a hdfs (hdfs dfs -copyFromLocal * hdfs:///user/jscaicedom/datasets/)

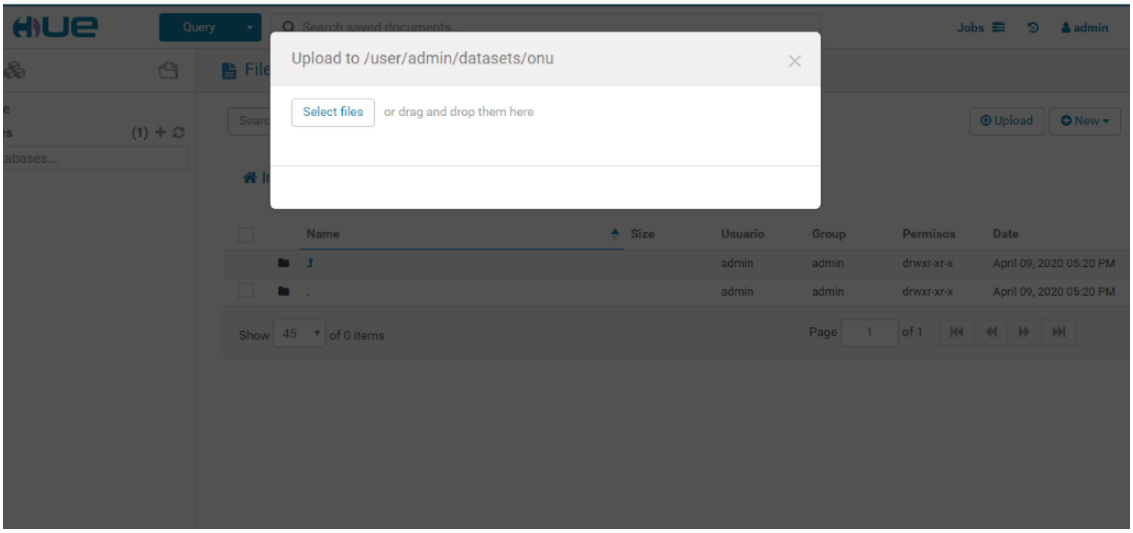




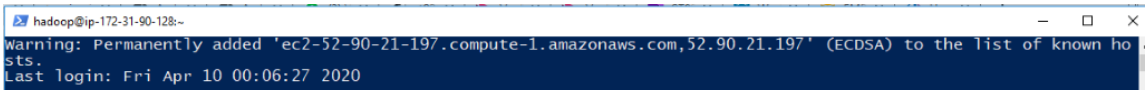
- Copiar datasers desde Browser AWS hacia S3/Amazon:
  - Tal y como lo explica el video.



- Gestión de archivos vía HUE en amazon EMR:



- Acceder al cluster via ssh:



```
       __|  __|_  )
       _|  (     /   Amazon Linux AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
33 package(s) needed for security, out of 54 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R::::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R::::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R::::RRRRRR:::R
  E::::E             M:::::M    M:::M    M:::::M   R::::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R::::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R::::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-90-128 ~]$ hdfs lfs -ls /user
Error: Could not find or load main class lfs
[hadoop@ip-172-31-90-128 ~]$ hdfs dfs -ls /user
Found 9 items
drwxrwxrwx   - hadoop   hadoop          0 2020-04-10 00:03 /user/hadoop
drwxr-xr-x   - mapred   mapred          0 2020-04-10 00:03 /user/history
drwxrwxrwx   - hdfs     hadoop          0 2020-04-10 00:03 /user/hive
drwxrwxrwx   - hue      hue             0 2020-04-10 00:03 /user/hue
drwxrwxrwx   - livy     livy            0 2020-04-10 00:03 /user/livy
```

- Descargar repositorio de github en:
  - EMR: En el nodo master del cluster EMR (ssh)

```
Complete!
[hadoop@ip-172-31-94-233 ~]$ git clone https://github.com/st0263eafit/bigdata.git
Cloning into 'bigdata'...
remote: Enumerating objects: 35, done.
remote: Counting objects: 100% (35/35), done.
remote: Compressing objects: 100% (24/24), done.
remote: Total 133 (delta 12), reused 31 (delta 9), pack-reused 98
Receiving objects: 100% (133/133), 65.79 MiB | 41.33 MiB/s, done.
Resolving deltas: 100% (12/12), done.
[hadoop@ip-172-31-94-233 ~]$ ls
bigdata
[hadoop@ip-172-31-94-233 ~]$ cd big
-bash: cd: big: No such file or directory
[hadoop@ip-172-31-94-233 ~]$ cd bigdata/
[hadoop@ip-172-31-94-233 bigdata]$ ls
00-aws  01-hdfs  02-mapreduce  datasets  README.md
[hadoop@ip-172-31-94-233 bigdata]$ c
[hadoop@ip-172-31-94-233 bigdata]$ ls -l
total 20
drwxrwxr-x  2 hadoop hadoop 4096 Apr 10 01:21 00-aws
drwxrwxr-x  4 hadoop hadoop 4096 Apr 10 01:21 01-hdfs
drwxrwxr-x  2 hadoop hadoop 4096 Apr 10 01:21 02-mapreduce
drwxrwxr-x 10 hadoop hadoop 4096 Apr 10 01:21 datasets
-rw-rw-r--  1 hadoop hadoop  986 Apr 10 01:21 README.md
```

  - DCA: En la 192.168.10.116 (Jupyterhub)

```
[jscaicedom@hdpjupyter bigdata]$ ls
00-aws  01-hdfs  02-mapreduce  datasets  README.md
[jscaicedom@hdpjupyter bigdata]$ ls -l
total 8
drwxr-xr-x  2 jscaicedom domain users 4096 Apr  9 16:23 00-aws
drwxr-xr-x  4 jscaicedom domain users   77 Apr  9 16:23 01-hdfs
drwxr-xr-x  2 jscaicedom domain users  247 Apr  9 16:23 02-mapreduce
drwxr-xr-x 10 jscaicedom domain users  159 Apr  9 16:23 datasets
-rw-r--r--  1 jscaicedom domain users  986 Apr  9 16:23 README.md
[jscaicedom@hdpjupyter bigdata]$
```