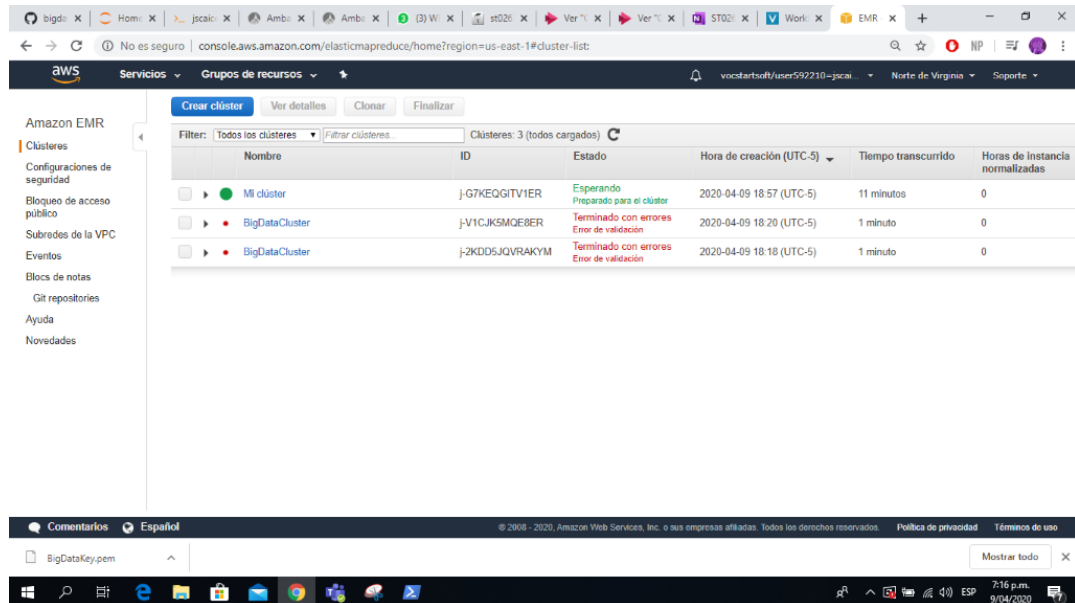


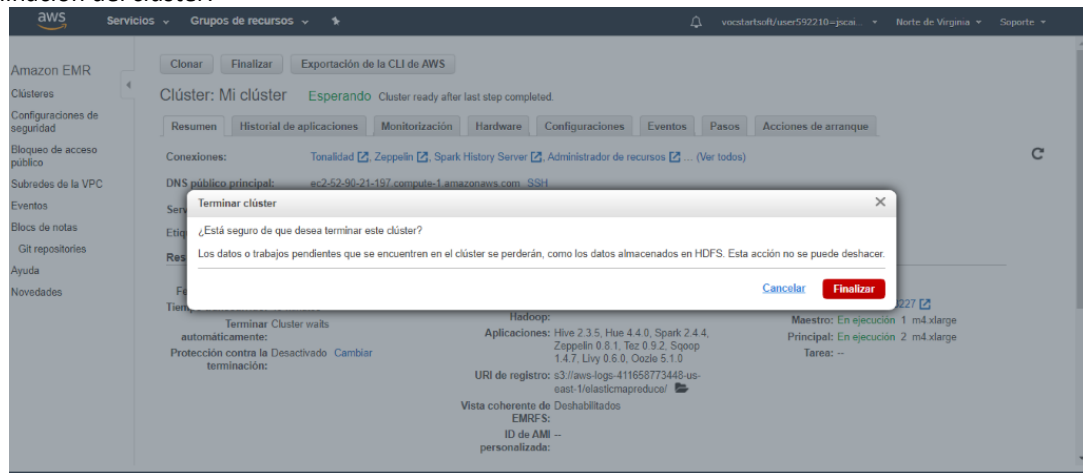
# 1. HDFS

## 1. Crear y gestionar Clusters Amazon EMR

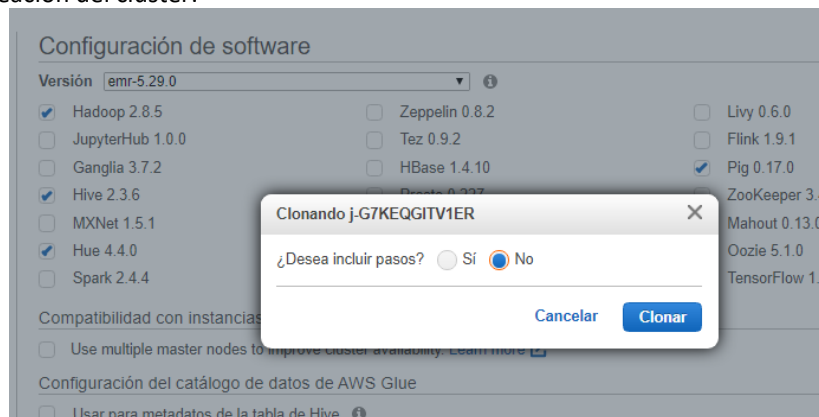
- o Creación de cluster por primera vez de forma interactiva siguiendo el video y habilitando los puertos en bloqueo de acceso público:



- o Terminación del clúster:



- o Recreación del clúster:



Editar configuración de software ⓘ

☒ Escribir la configuración ☐ Cargar JSON desde S3

`classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]`

Filter: Todos los clústeres  Clústeres: 4 (todos cargados)

	Nombre	ID	Estado	Hora de creación (UTC-5)	Tiempo transcurrido	Horas de instancia normalizadas
<input type="checkbox"/>	BigDataCluster	j-3JL0YK8FC0XB2	Comenzando	2020-04-09 19:42 (UTC-5)	0 segundos	0
<input type="checkbox"/>	Mi clúster	j-G7KEQGITV1ER	Terminado Solicitud del usuario	2020-04-09 18:57 (UTC-5)	43 minutos	24
<input type="checkbox"/>	BigDataCluster	j-V1CJIKSMQE8ER	Terminado con errores Error de validación	2020-04-09 18:20 (UTC-5)	1 minuto	0
<input type="checkbox"/>	BigDataCluster	j-2KDD5JQVRAKYM	Terminado con errores Error de validación	2020-04-09 18:18 (UTC-5)	1 minuto	0

- Por comando: Se actualizan las credenciales de aws cli

Credentials

AWS Access

Session started at: 2020-05-03T15:07:51-0700  
Session to end at: 2020-05-03T18:07:51-0700  
Remaining session time: 1h1m13s

Term: 88 days 19:07:06

AWS CLI:

Copy and paste the following into ~/.aws/credentials

```
[default]
aws_access_key_id=ASIAV7M5GPE33YPOIB4
aws_secret_access_key=Lnp75872ZyLymxQWERYQ5qXTBgb/zb2emc1k
aws_session_token=FuoGZLXVXGEEADP3HNgwXK3u4o/mCLGATZIE+LUSVjz0M0r5pNFQs/700w2a1BHG/48QABs4Fv/2tNP7KpsJqYQ5Q4Q86j62xxEReqv2T58zdfc
98S8000707ch348g3mcfAP2018a11Rv7f1r/(807149)gIB88or8e88HPzI1gCtH+KxXK4o1a7ZET+2uDuQAc4a3d8ou1P24115EURPQFH0HQ5r3tzeVY/Y0Qp
ms2XQ05B1YRqKF8BPryMRTRfju2FFALtugYHndcLl91y14gb31BTt835polFbT148H8rHq4oSh1/k3LvxUvJ/USxJfUp3RvW4h2coo1Ee0VP10f
```

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

ando Cluster ready after last step completed

Exportación de la CLI de AWS

```
aws emr create-cluster --auto-scaling-role EMR_AutoScaling_DefaultRole --
applications Name=Hadoop Name=Hive Name=Hue Name=Spark Name=Zeppelin Name=Tez
Name=Sqoop Name=Livy Name=Oozie --ebs-root-volume-size 10 --ec2-attributes
'{"KeyName":"labbigdata","InstanceProfile":"EMR_EC2_DefaultRole","SubnetId":"subn
et-093a3227","EmrManagedSlaveSecurityGroup":"sg-
068c7a98ef941d2a2","EmrManagedMasterSecurityGroup":"sg-0d23f227802a5f666"}' --
service-role EMR_DefaultRole --enable-debugging --release-label emr-5.27.0 --log-
uri 's3n://aws-logs-411658773448-us-east-1/elasticmapreduce/' --name
'BigDataCluster' --instance-groups '[{"InstanceCount":2,"EbsConfiguration":
{"EbsBlockDeviceConfigs":[{"VolumeSpecification":
{"SizeInGB":32,"VolumeType":"gp2"},"VolumesPerInstance":2}]},"InstanceGroupType":
"CORE","InstanceType":"m4.xlarge","Name":"Principal - 2"},
{"InstanceCount":1,"EbsConfiguration":{"EbsBlockDeviceConfigs":
[{"VolumeSpecification":
{"SizeInGB":32,"VolumeType":"gp2"},"VolumesPerInstance":2}]},"InstanceGroupType":
"MASTER","InstanceType":"m4.xlarge","Name":"Maestro - 1"}]' --scale-down-behavior
TERMINATE_AT_TASK_COMPLETION --region us-east-1
```

## 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/)

🏠 / Files View

ProdHDP jscaicedom

/ > user > jscaicedom > datasets

Total: 9 files or folders

+ Select All

New Folder

Upload

Search in current directory...

Name >	Size >	Last Modified >	Owner >	Group >	Permission	Erasure Coding	Enc
↶							
airlines.csv	761.8 kB	2020-04-09 15:35	jscaicedom	bigdata	-rw-r--r--		No
all-news	--	2020-04-09 15:35	jscaicedom	bigdata	drwxr-xr-x		No
gutenberg	--	2020-04-09 15:35	jscaicedom	bigdata	drwxr-xr-x		No
gutenberg-small	--	2020-04-09 15:35	jscaicedom	bigdata	drwxr-xr-x		No
onu	--	2020-04-09 15:35	jscaicedom	bigdata	drwxr-xr-x		No
otros	--	2020-04-09 15:35	jscaicedom	bigdata	drwxr-xr-x		No
papers_sample	--	2020-04-09 15:35	jscaicedom	bigdata	drwxr-xr-x		No
retail_logs	--	2020-04-09 15:35	jscaicedom	bigdata	drwxr-xr-x		No
spark	--	2020-04-09 15:35	jscaicedom	bigdata	drwxr-xr-x		No

○ Copiar datasers desde Browser AWS hacia S3/Amazon:

- Tal y como lo explica el video.

aws

Servicios

Grupos de recursos

vocstartsoft/user592210=jscaicedom

Global

Soporte

Escriba un prefijo y pulse Intro para buscar. Pulse ESC para borrar.

Cargar

+ Crear carpeta

Descargar

Acciones

EE.UU. Este (Norte de Virginia)

<input type="checkbox"/>	Nombre	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	all-news	--	--	--
<input type="checkbox"/>	gutenberg-small	--	--	--
<input type="checkbox"/>	gutenberg	--	--	--
<input type="checkbox"/>	onu	--	--	--
<input type="checkbox"/>	otros	--	--	--
<input type="checkbox"/>	papers_sample	--	--	--
<input type="checkbox"/>	retail_logs	--	--	--
<input type="checkbox"/>	spark	--	--	--

Operaciones

0 En curso

2 Correcta

0 Error

○ Gestión de archivos vía HUE en amazon EMR:

HUE

Query

Search saved documents

Jobs

admin

Upload to /user/admin/datasets/onu

Select files or drag and drop them here

Upload

New

<input type="checkbox"/>	Name	Size	Usuario	Group	Permisos	Date
<input type="checkbox"/>	.		admin	admin	drwxr-xr-x	April 09, 2020 05:20 PM
<input type="checkbox"/>	.		admin	admin	drwxr-xr-x	April 09, 2020 05:20 PM

Show 45 of 0 items

Page 1 of 1

○ Acceder al cluster via ssh:

```
hadoop@ip-172-31-90-128:~$ ssh -i /home/hadoop/.ssh/id_rsa ec2-52-90-21-197.compute-1.amazonaws.com
Warning: Permanently added 'ec2-52-90-21-197.compute-1.amazonaws.com,52.90.21.197' (ECDSA) to the list of known hosts.
Last login: Fri Apr 10 00:06:27 2020
```

