

# Universidad Nacional Autónoma de México

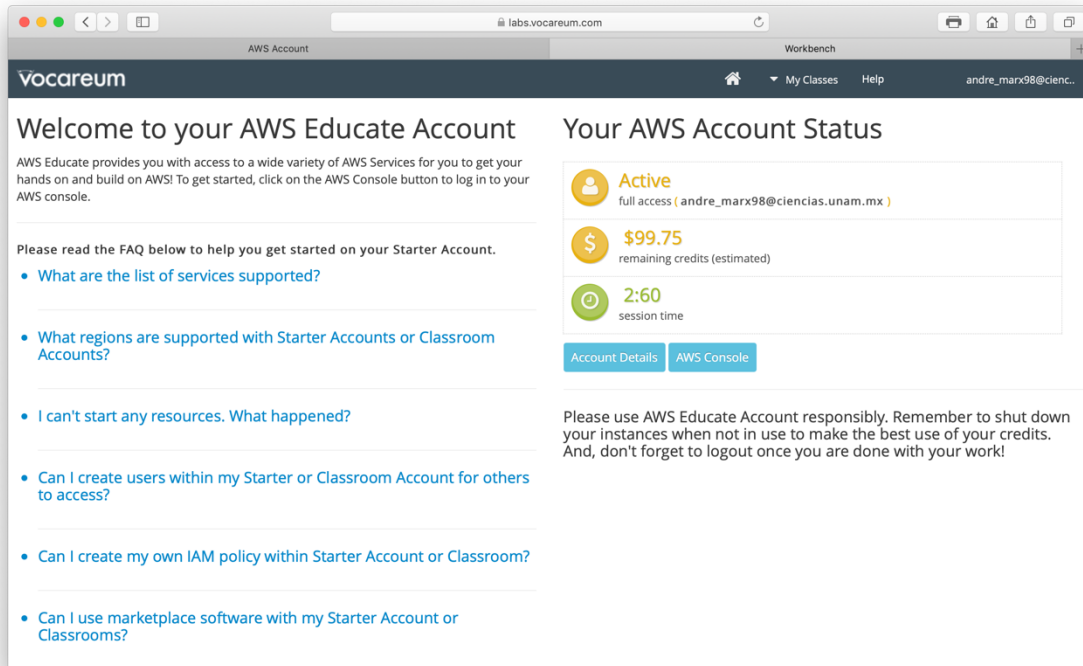
## Práctica de Cloud

Prof. Jimmy Hernández Morales

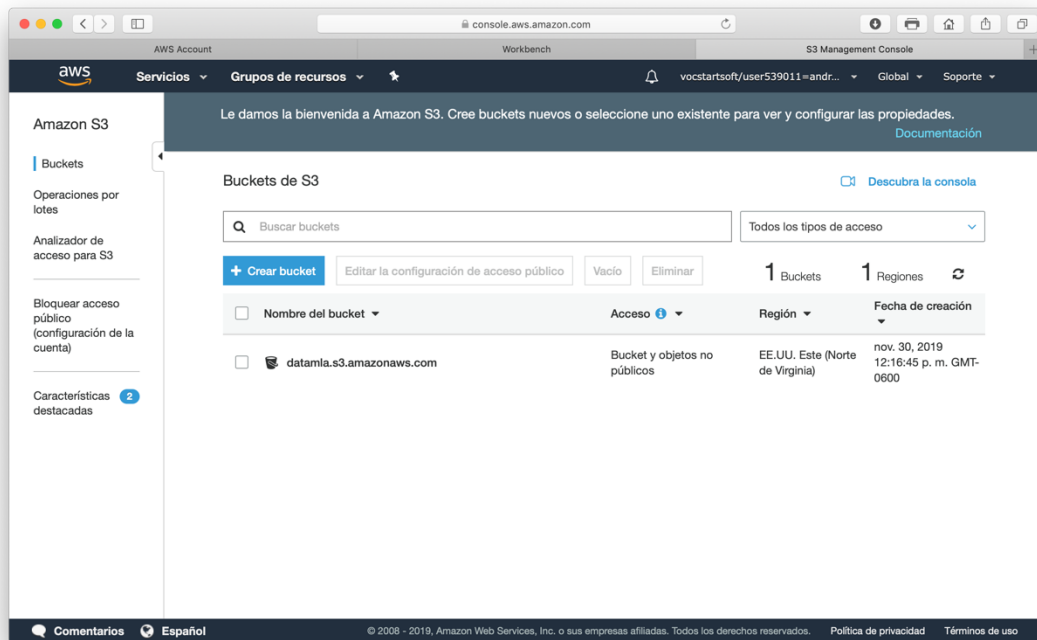
Ayudante: Miguel Hinojosa Medrano

Alumno: André Marx Puente Arévalo

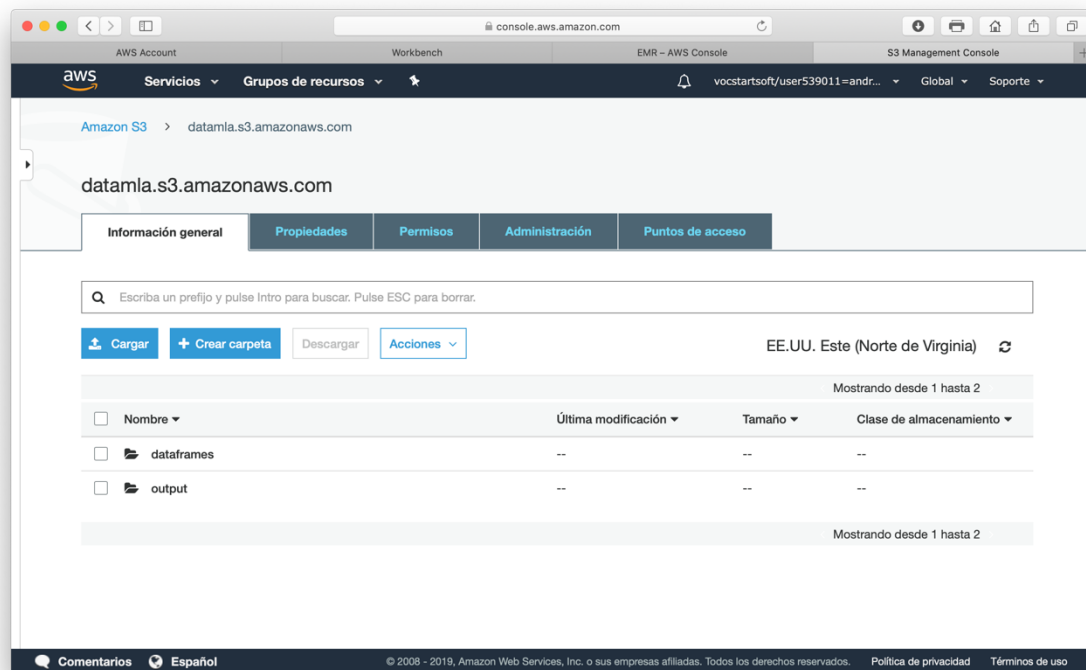
### 1. Abrimos la consola de AWS



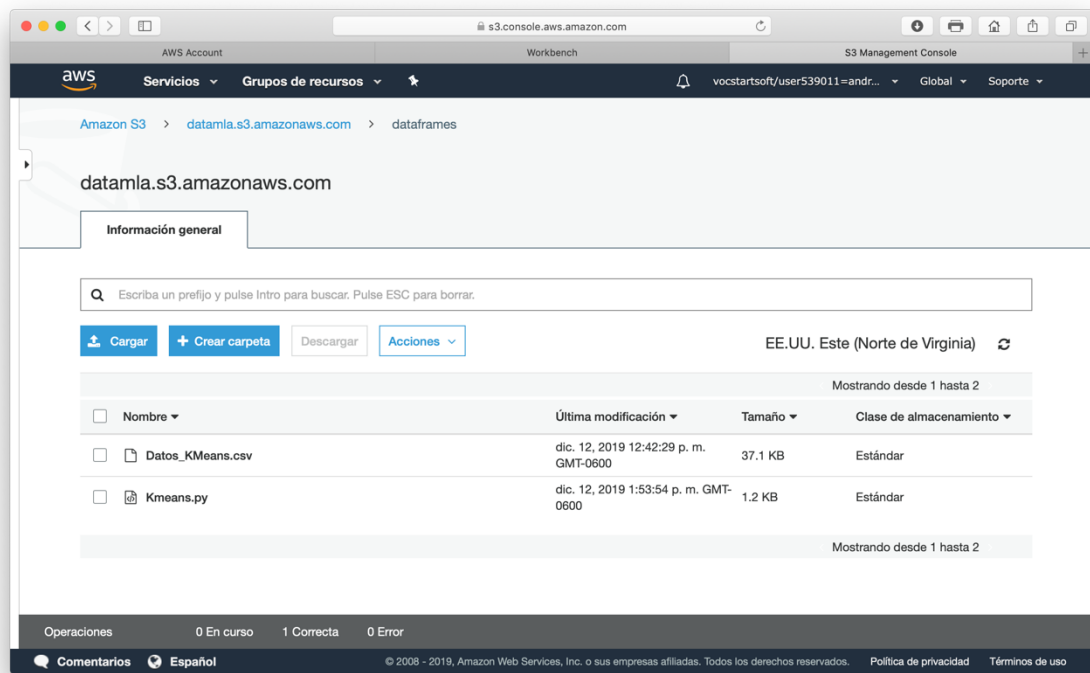
### 2. Creamos un bucket con el siguiente formato: "datamla.s3.amazonaws.com"



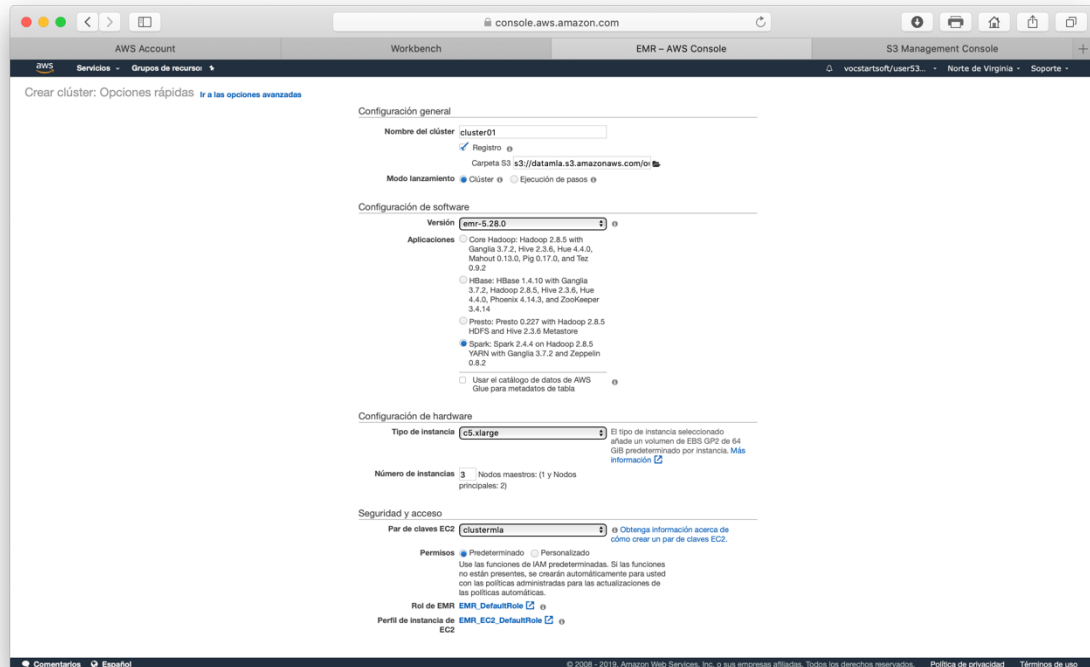
3. Creamos dos carpetas dentro del bucket “dataframes” que contendrán el archivo csv y el spcrip y “output” donde se depositarán los logs del procesamiento.



4. Almacenamos los archivos que vamos a utilizar: “Datos\_KMeans.csv” y “Kmeans.py” en S3



5. Lanzo un clúster con Elastic Map Reduce EMR, el cual contiene 1 master y 2 esclavos





## 7. Copio mi archivo “Kmeans.py” a la instancia EC2 master y finalmente lanzo el spark submit

```
Descargas — hadoop@ip-172-31-91-166:~ — ssh -i clustermla.pem hadoop@54.205.129.82 — 167x51
(base) MBP-de-Andre:Downloads AndrePuentes$ chmod go-xw clustermla.pem
(base) MBP-de-Andre:Downloads AndrePuentes$ ssh -i clustermla.pem hadoop@54.205.129.82
Last login: Thu Dec 12 19:22:34 2019

      _ _ _ _ _
     /   _   \
    / _ _ _ \
   / _ _ _ \
  / _ _ _ \
 / _ _ _ \
/ _ _ _ \

Amazon Linux AMI

https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
19 package(s) needed for security, out of 30 available
Run "sudo yum update" to apply all updates.
~bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory

EEEEEEEEEEEEEEEEEEEE MMMMMMMM MMMMMMMM RRRRRRRRRRRRRRRR
E:::EEEEEEEEEEEEEEEE E M:::M:::M M:::M:::M R:::RRRRRRRRRRRR
EE:::EEEEEEEEEEEEEEEE E M:::M:::M M:::M:::M R:::RRRRRRRRRRRR
E:::E EEEEE M:::M:::M M:::M:::M R:::RRRRRRRRRRRR
E:::E EEEEE M:::M:::M M:::M:::M R:::RRRRRRRRRRRR
E:::EEEEEEEEEEEEEEEE M:::M M:::M M:::M M:::M R:::RRRRRRRRRRRR
E:::EEEEEEEEEEEEEEEE M:::M M:::M M:::M M:::M R:::RRRRRRRRRRRR
E:::EEEEEEEEEEEEEEEE M:::M M:::M M:::M M:::M R:::RRRRRRRRRRRR
E:::E M:::M M:::M M:::M M:::M R:::RRRRRRRRRRRR
E:::E EEEEE M:::M M:::M M:::M M:::M R:::RRRRRRRRRRRR
EE:::EEEEEEEEEEEEEEEE E M:::M M:::M M:::M M:::M R:::RRRRRRRRRRRR
E:::EEEEEEEEEEEEEEEE E M:::M M:::M M:::M M:::M R:::RRRRRRRRRRRR
EEEEEEEEEEEEEEEEEEEE MMMMMMMM MMMMMMMM RRRRRRRRRRRRRRRR

[hadoop@ip-172-31-91-166 ~]$ aws s3 cp s3://datamla.s3.amazonaws.com/Tarea 5.py .
Unknown options: .
[hadoop@ip-172-31-91-166 ~]$ ls
[hadoop@ip-172-31-91-166 ~]$ aws s3 cp s3://datamla.s3.amazonaws.com/Kmeans.py .
fatal error: An error occurred (404) when calling the HeadObject operation: Key "Kmeans.py" does not exist
[hadoop@ip-172-31-91-166 ~]$ aws s3 cp s3://datamla.s3.amazonaws.com/dataframes/Kmeans.py .
download: s3://datamla.s3.amazonaws.com/dataframes/Kmeans.py to ./Kmeans.py
[hadoop@ip-172-31-91-166 ~]$ ls
Kmeans.py
[hadoop@ip-172-31-91-166 ~]$ spark-submit Kmeans.py .
19/12/12 19:58:58 INFO SparkContext: Running Spark version 2.4.4
19/12/12 19:58:58 INFO SparkContext: Submitted application: tarea5
19/12/12 19:58:58 INFO SecurityManager: Changing view acls to: hadoop
19/12/12 19:58:58 INFO SecurityManager: Changing modify acls to: hadoop
19/12/12 19:58:58 INFO SecurityManager: Changing view acls groups to:
19/12/12 19:58:58 INFO SecurityManager: Changing modify acls groups to:
19/12/12 19:58:58 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(hadoop); groups with view permissions: Set(); users with modify permissions: Set(hadoop); groups with modify permissions: Set()
19/12/12 19:58:58 INFO Utils: Successfully started service 'sparkDriver' on port 36271.
19/12/12 19:58:58 INFO SparkEnv: Registering MapOutputTracker
19/12/12 19:58:58 INFO SparkEnv: Registering BlockManagerMaster
19/12/12 19:58:58 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
19/12/12 19:58:58 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up

19/12/12 19:59:27 INFO MemoryStore: Block broadcast_51 stored as values in memory (estimated size 8.6 KB, free 1027.1 MB)
19/12/12 19:59:27 INFO MemoryStore: Block broadcast_51_piece0 stored as bytes in memory (estimated size 4.4 KB, free 1027.1 MB)
19/12/12 19:59:27 INFO BlockManagerInfo: Added broadcast_51_piece0 in memory on ip-172-31-91-166.ec2.internal:44733 (size: 4.4 KB, free: 1028.7 MB)
19/12/12 19:59:27 INFO SparkContext: Created broadcast 51 from broadcast at DAGScheduler.scala:1281
19/12/12 19:59:27 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 31 (MapPartitionsRDD[76] at collect at ClusteringEvaluator.scala:177) (first 15 tasks are for partitions Vector())
19/12/12 19:59:27 INFO YarnScheduler: Adding task set 31.0 with 1 tasks
19/12/12 19:59:27 INFO TaskSetManager: Starting task 0.0 in stage 31.0 (TID 230, ip-172-31-82-153.ec2.internal, executor 1, partition 0, NODE_LOCAL, 7778 bytes)
19/12/12 19:59:27 INFO BlockManagerInfo: Added broadcast_51_piece0 in memory on ip-172-31-82-153.ec2.internal:41369 (size: 4.4 KB, free: 2.6 GB)
19/12/12 19:59:27 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 10 to 172.31.82.153:58348
19/12/12 19:59:27 INFO TaskSetManager: Finished task 0.0 in stage 31.0 (TID 230) in 25 ms on ip-172-31-82-153.ec2.internal (executor 1) (1/1)
19/12/12 19:59:27 INFO YarnScheduler: Removed TaskSet 31.0, whose tasks have all completed, from pool
19/12/12 19:59:27 INFO DAGScheduler: ResultStage 31 (collect at ClusteringEvaluator.scala:177) finished in 0.027 s
19/12/12 19:59:27 INFO DAGScheduler: Job 20 finished: collect at ClusteringEvaluator.scala:177, took 0.249386 s
19/12/12 19:59:27 INFO TorrentBroadcast: Destroying Broadcast(48) (from destroy at ClusteringEvaluator.scala:478)
19/12/12 19:59:27 INFO BlockManagerInfo: Removed broadcast_48_piece0 on ip-172-31-91-166.ec2.internal:44733 in memory (size: 587.0 B, free: 1028.7 MB)
19/12/12 19:59:27 INFO BlockManagerInfo: Removed broadcast_48_piece0 on ip-172-31-82-153.ec2.internal:41369 in memory (size: 587.0 B, free: 2.6 GB)
('Silhouette', 0.7741374633492396)
19/12/12 19:59:27 INFO SparkContext: Invoking stop() from shutdown hook
19/12/12 19:59:27 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-91-166.ec2.internal:4040
19/12/12 19:59:27 INFO YarnClientSchedulerBackend: Interrupting monitor thread
19/12/12 19:59:27 INFO YarnClientSchedulerBackend: Shutting down all executors
19/12/12 19:59:27 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
19/12/12 19:59:27 INFO SchedulerExtensionServices: Stopping SchedulerExtensionServices
(serviceOptions=None,
services=list(),
started=false)
19/12/12 19:59:27 INFO YarnClientSchedulerBackend: Stopped
19/12/12 19:59:27 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
19/12/12 19:59:27 INFO MemoryStore: MemoryStore cleared
19/12/12 19:59:27 INFO BlockManager: BlockManager stopped
19/12/12 19:59:27 INFO BlockManagerMaster: BlockManagerMaster stopped
19/12/12 19:59:27 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
19/12/12 19:59:27 INFO SparkContext: Successfully stopped SparkContext
19/12/12 19:59:27 INFO ShutdownHookManager: Shutdown hook called
19/12/12 19:59:27 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-af1acdee-5779-4c8f-8a16-1bc579f27d32
19/12/12 19:59:27 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-af1acdee-5779-4c8f-8a16-1bc579f27d32/pyspark-4bf2c3fc-0192-43f8-baa3-5e2dca52c064
19/12/12 19:59:27 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-b3bcab79-9ed3-43e6-87d2-347e53da42eb
[hadoop@ip-172-31-91-166 ~]$
Broadcast message from root@ip-172-31-91-166
(unknown) at 20:08 ...

The system is going down for halt NOW!

Broadcast message from root@ip-172-31-91-166
(unknown) at 20:08 ...

The system is going down for halt NOW!
Connection to 54.205.129.82 closed by remote host.
Connection to 54.205.129.82 closed.
(base) MBP-de-Andre:Downloads AndrePuentes$
```

Nota: Lo último que sale en la terminal (imagen de arriba) es que detuve las instancias para que no me siguieran gastando créditos, pero si notamos, corrió todo el código, el cual terminaba cuando pintaba el Silhouette y este sale en la imagen.