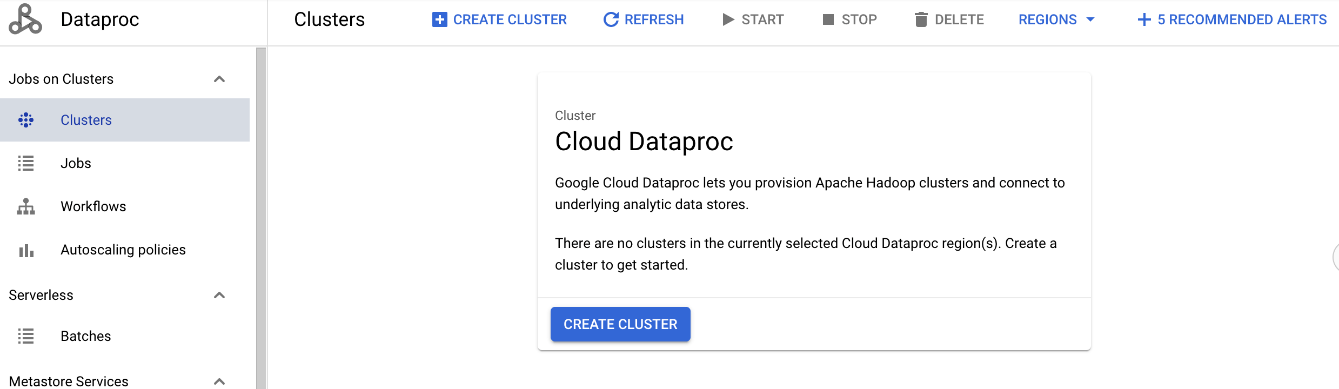
# Cloud Dataproc

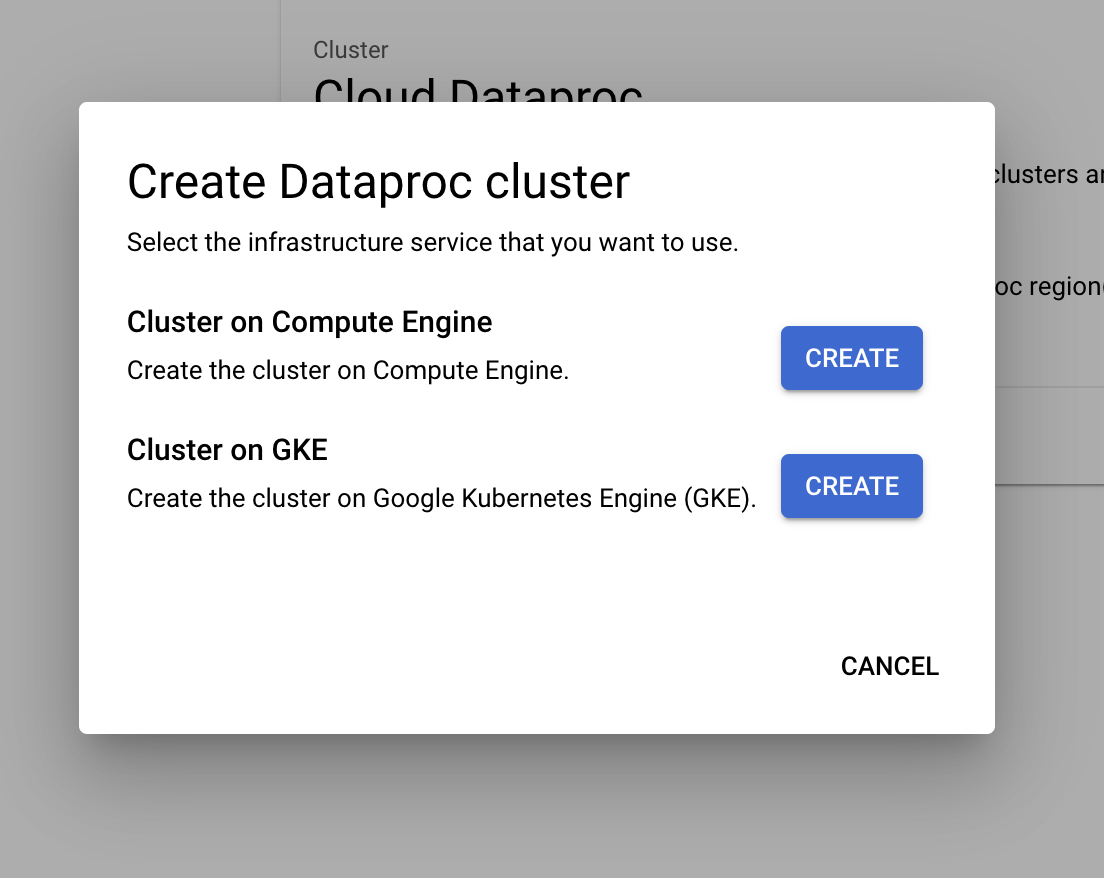
## Cluster de Cloud Dataproc

Buscaremos Dataproc en el menú de navegación, y le pondremos un pin para encontrarlo rápidamente la siguiente vez. Habilite la API en caso que no esté habilitada con anterioridad.

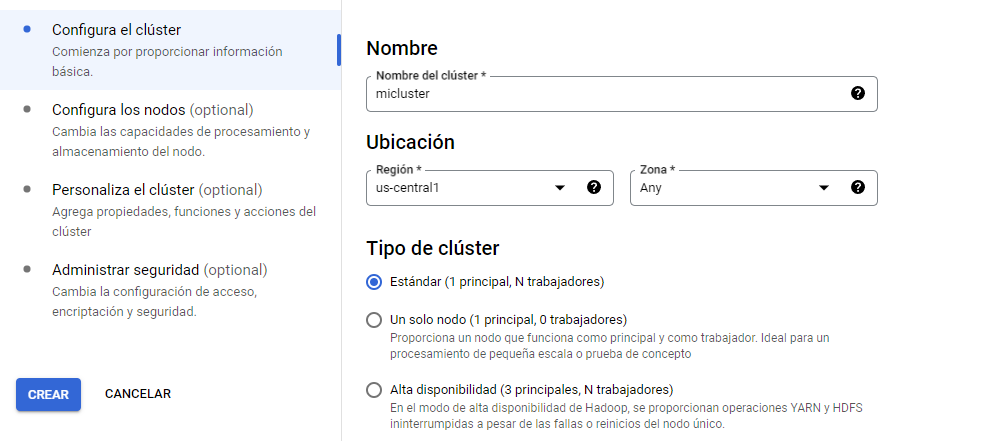
Haz click en **CREATE CLUSTER**.



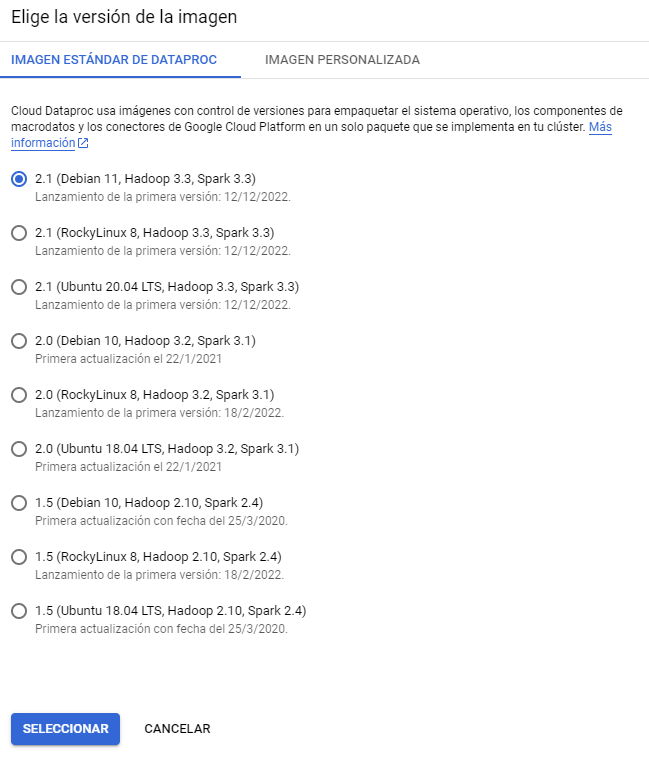
Luego da la opción entre usar Compute Engine o GKE, elegiremos Compute Engine.



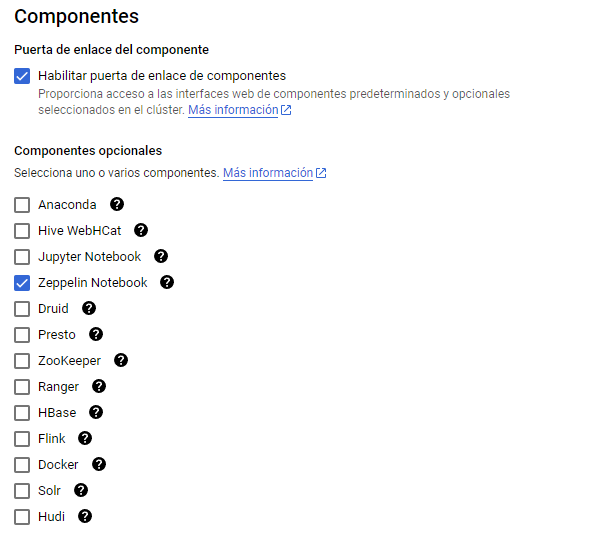
Le pondremos un nombre a nuestro cluster, por ejemplo “micluster”, seleccionamos región, el tipo de cluster dejaremos el **Standard**.



Luego pon **CAMBIAR** en la sección **Control de Versiones** y selecciona 2.1 (Debian 11, Hadoop 3.3, Spark 3.3). Aquí es donde ajustas las versiones de los software que necesitas y si no están puedes proveer una imagen hecha por ti. Pon **SELECT**.



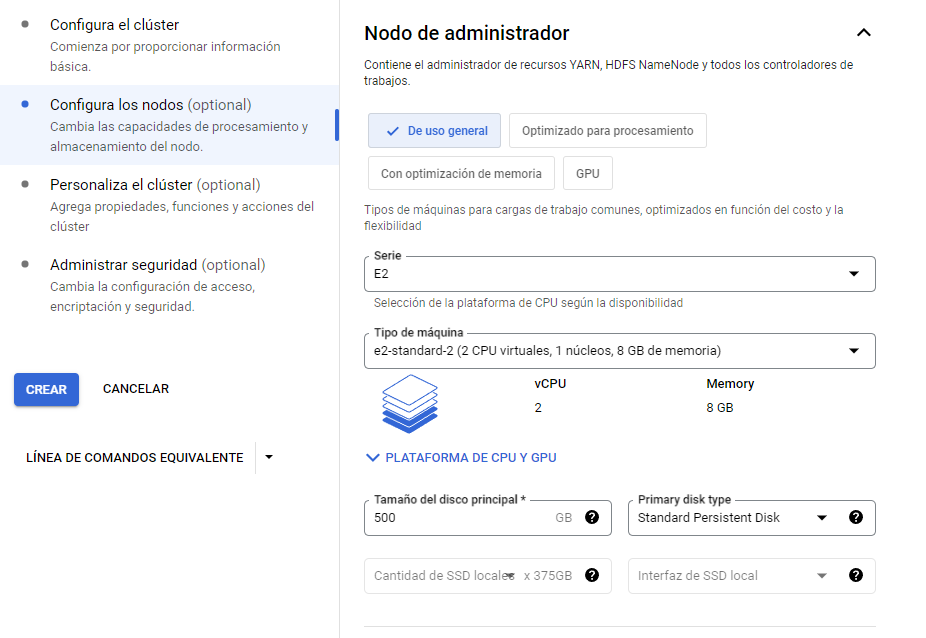
En la sección **Components**, haz click en **Enable component gateway**. Y en **Optional Components** selecciona **Jupyter Notebooks**.

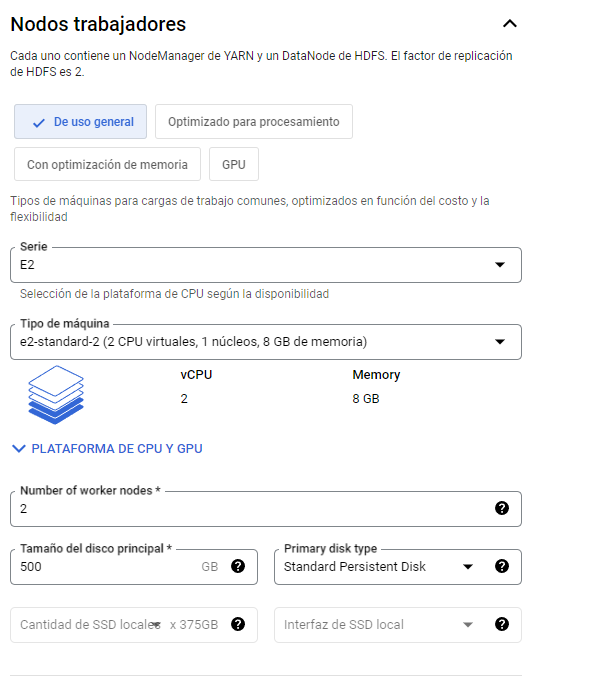


Explora las configuraciones opcionales, pero deja los valores por defecto. Haz click en Equivalent Command Line para ver el gcloud command que crea el cluster que configuraste sin necesidad de tantos clicks. Puedes copiarlo, es útil el caso de que tengas que cerrar el trabajo antes de terminar el lab y luego quieras partir sin perder minutos en configuración.

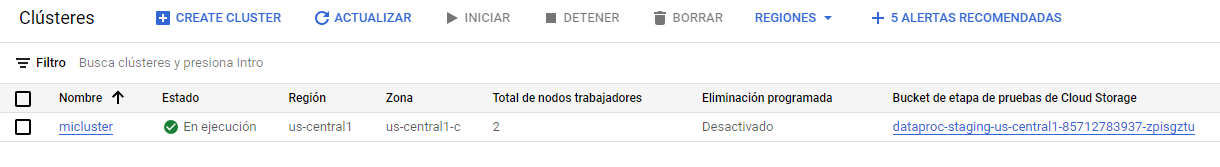


Cambiamos el master y los nodos a los más pequeños:



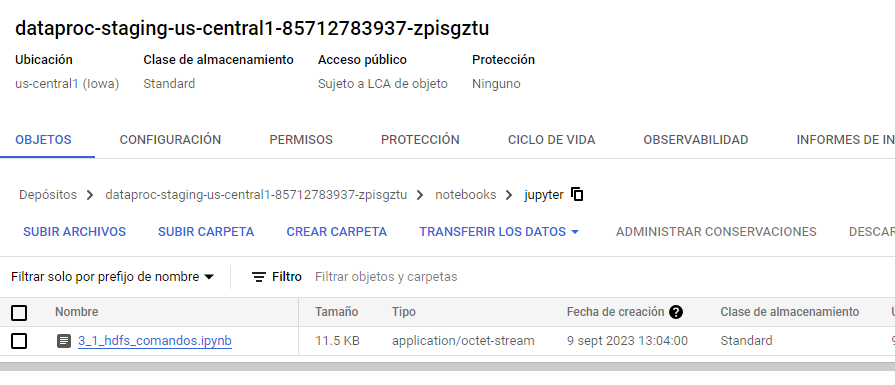


Luego presiona **CREAR** para crear el cluster. Deberías ver luego de unos minutos el cluster en estado Running, como en la siguiente figura.



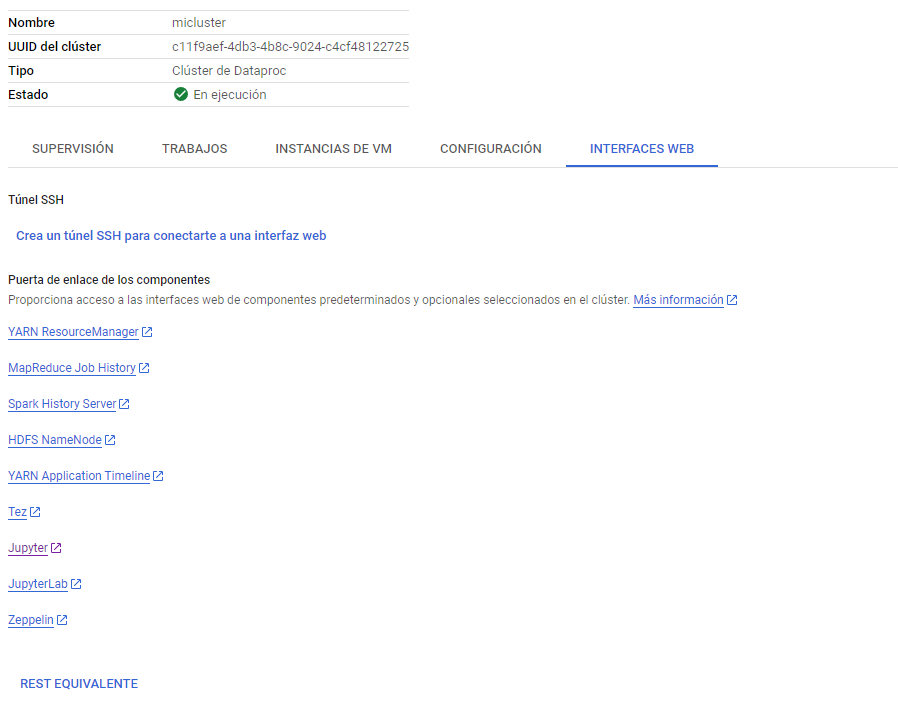
## **Usando software instalado en Dataproc (Jupyter)**

Este notebook tenemos que almacenarlo en el bucket que usa Cloud Dataproc, lo puedes encontrar entre los datos del cluster:

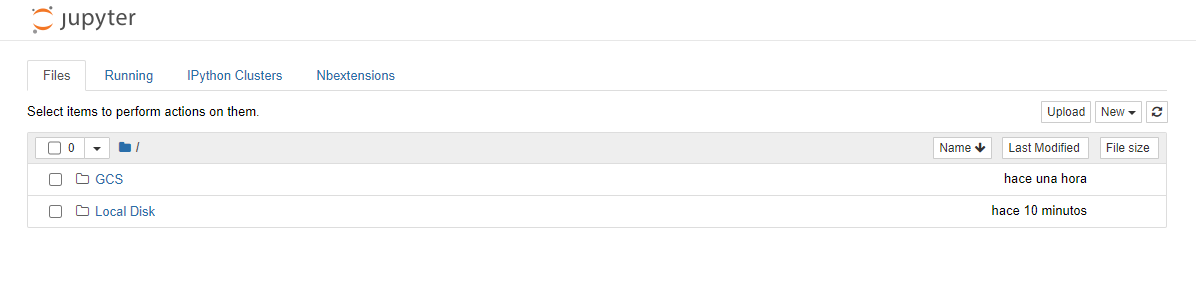


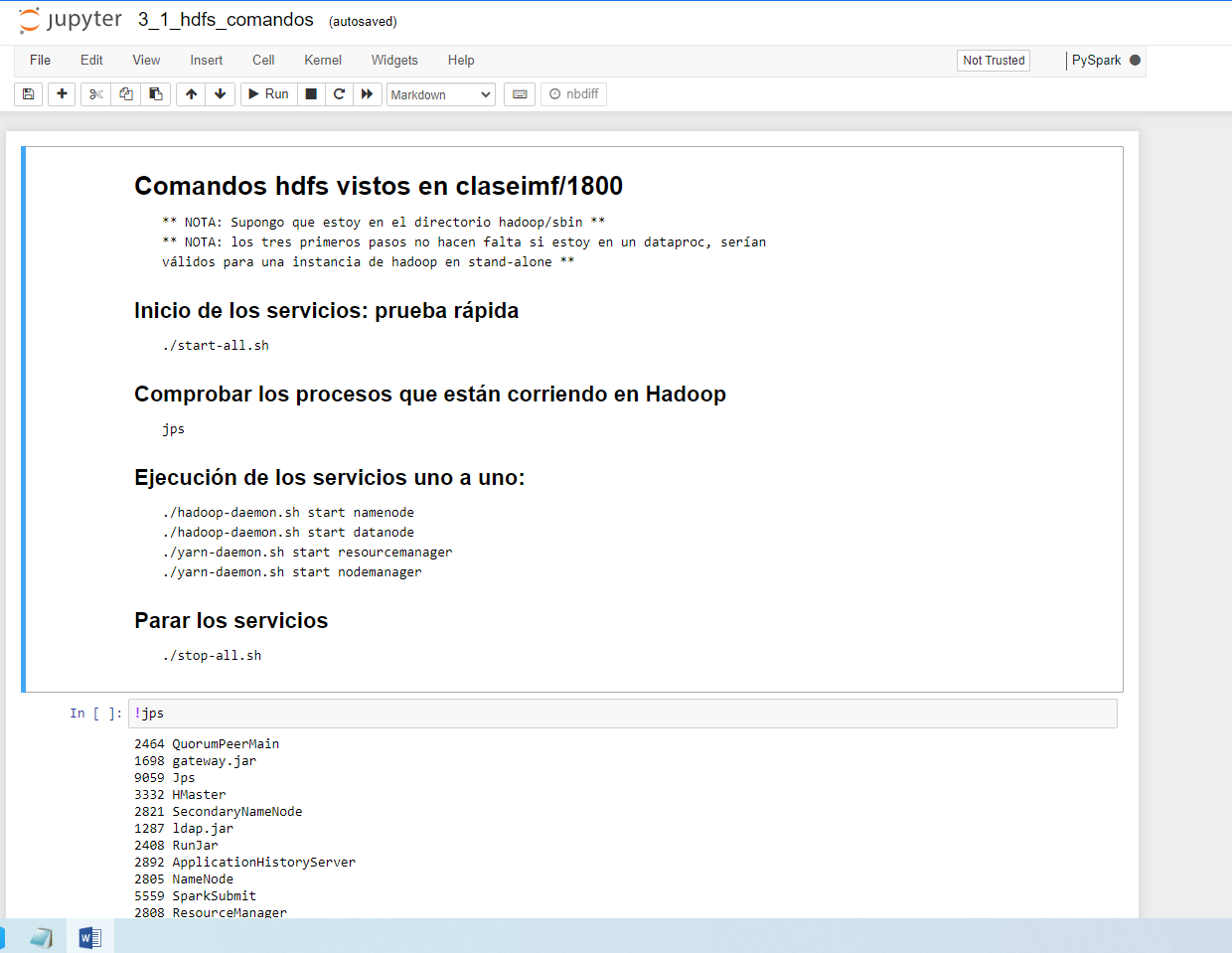
Subir el fichero 3\_1\_hdfs\_comandos.ipynb.

Sabemos que el cluster tiene instalado Hadoop, Spark, entre otros, de hecho tenemos acceso a su interfaz gráfica al entrar a los detalles del cluster, en **Web Interfaces > Component gateway**. Haz click sobre el que dice **Jupyter**.



Te aparecerá algo como, dado que se el contenido de la carpeta notebooks/jupyter que está en Cloud Storage y tiene el notebook que acabas de subir:

Haz click en el notebook dentro de GCS para abrirlo.



**Para lanzar los procesos HADOOP**

Subimos los siguientes ficheros al bucket vía consola:

* mapper.py
* reducer.py
* input.txt

Abrimos la consola SSH:

mkdir imf

Copiamos los ficheros a local

gsutil cp gs://dataproc-staging-us-central1-85712783937-zpisgztu/mapper.py mapper.py

gsutil cp gs://dataproc-staging-us-central1-85712783937-zpisgztu/reducer.py reducer.py

gsutil cp gs://dataproc-staging-us-central1-85712783937-zpisgztu/input.txt input.txt

Damos permisos de ejecución a los ficheros:

chmod 777 \*

PRUEBA DE LOS SCRIPTS EN LOCAL:

cat input.txt | python3 mapper.py | sort | python3 reducer.py

Cargamos el fichero input.txt al hdfs:

hdfs dfs -cp gs://dataproc-staging-us-central1-85712783937-zpisgztu/input.txt /user/dataproc

Lanzamos el map-reduce en hadoop:

mapred streaming -files mapper.py,reducer.py -mapper "python3 mapper.py" -reducer "python3 reducer.py" -input **hdfs:///**user/dataproc/input.txt -output **hdfs:///**user/dataproc/salida.txt

Dará una salida similar a esta:

javier\_gonzalez\_mendez@micluster-m:~/imf$ mapred streaming -files mapper.py,reducer.py -mapper "python3 mapper.py" -reducer "python3 reducer.py" -input hdfs:///user/dataproc/input.txt -output hdfs:///user/dataproc/salida.txt

WARNING: HADOOP\_JOB\_HISTORYSERVER\_OPTS has been replaced by MAPRED\_HISTORYSERVER\_OPTS. Using value of HADOOP\_JOB\_HISTORYSERVER\_OPTS.

packageJobJar: [] [/usr/lib/hadoop/hadoop-streaming-3.3.3.jar] /tmp/streamjob13448534728843822667.jar tmpDir=null

2023-09-10 11:22:21,376 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at micluster-m.c.proven-gasket-391120.internal./10.128.0.14:8032

2023-09-10 11:22:21,617 INFO client.AHSProxy: Connecting to Application History server at micluster-m.c.proven-gasket-391120.internal./10.128.0.14:10200

2023-09-10 11:22:22,375 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at micluster-m.c.proven-gasket-391120.internal./10.128.0.14:8032

2023-09-10 11:22:22,376 INFO client.AHSProxy: Connecting to Application History server at micluster-m.c.proven-gasket-391120.internal./10.128.0.14:10200

2023-09-10 11:22:22,687 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/javier\_gonzalez\_mendez/.staging/job\_1694337034639\_0021

2023-09-10 11:22:23,283 INFO mapred.FileInputFormat: Total input files to process : 1

2023-09-10 11:22:23,406 INFO mapreduce.JobSubmitter: number of splits:9

2023-09-10 11:22:23,778 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1694337034639\_0021

2023-09-10 11:22:23,778 INFO mapreduce.JobSubmitter: Executing with tokens: []

2023-09-10 11:22:24,228 INFO conf.Configuration: resource-types.xml not found

2023-09-10 11:22:24,230 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

2023-09-10 11:22:24,416 INFO impl.YarnClientImpl: Submitted application application\_1694337034639\_0021

2023-09-10 11:22:24,507 INFO mapreduce.Job: The url to track the job: http://micluster-m.c.proven-gasket-391120.internal.:8088/proxy/application\_1694337034639\_0021/

2023-09-10 11:22:24,510 INFO mapreduce.Job: Running job: job\_1694337034639\_0021

2023-09-10 11:22:37,738 INFO mapreduce.Job: Job job\_1694337034639\_0021 running in uber mode : false

2023-09-10 11:22:37,739 INFO mapreduce.Job: map 0% reduce 0%

2023-09-10 11:22:46,885 INFO mapreduce.Job: map 11% reduce 0%

2023-09-10 11:22:53,949 INFO mapreduce.Job: map 44% reduce 0%

2023-09-10 11:23:01,012 INFO mapreduce.Job: map 56% reduce 0%

2023-09-10 11:23:08,060 INFO mapreduce.Job: map 78% reduce 0%

2023-09-10 11:23:09,067 INFO mapreduce.Job: map 89% reduce 0%

2023-09-10 11:23:15,108 INFO mapreduce.Job: map 100% reduce 0%

2023-09-10 11:23:24,166 INFO mapreduce.Job: map 100% reduce 33%

2023-09-10 11:23:31,209 INFO mapreduce.Job: map 100% reduce 67%

2023-09-10 11:23:32,215 INFO mapreduce.Job: map 100% reduce 100%

2023-09-10 11:23:34,239 INFO mapreduce.Job: Job job\_1694337034639\_0021 completed successfully

2023-09-10 11:23:34,357 INFO mapreduce.Job: Counters: 55

File System Counters

FILE: Number of bytes read=293

FILE: Number of bytes written=3490297

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=1701

HDFS: Number of bytes written=172

HDFS: Number of read operations=42

HDFS: Number of large read operations=0

HDFS: Number of write operations=9

HDFS: Number of bytes read erasure-coded=0

Job Counters

Killed reduce tasks=1

Launched map tasks=9

Launched reduce tasks=3

Data-local map tasks=9

Total time spent by all maps in occupied slots (ms)=283604688

Total time spent by all reduces in occupied slots (ms)=107865732

Total time spent by all map tasks (ms)=86544

Total time spent by all reduce tasks (ms)=32916

Total vcore-milliseconds taken by all map tasks=86544

Total vcore-milliseconds taken by all reduce tasks=32916

Total megabyte-milliseconds taken by all map tasks=283604688

Total megabyte-milliseconds taken by all reduce tasks=107865732

Map-Reduce Framework

Map input records=4

Map output records=27

Map output bytes=221

Map output materialized bytes=437

Input split bytes=846

Combine input records=0

Combine output records=0

Reduce input groups=20

Reduce shuffle bytes=437

Reduce input records=27

Reduce output records=20

Spilled Records=54

Shuffled Maps =27

Failed Shuffles=0

Merged Map outputs=27

GC time elapsed (ms)=1013

CPU time spent (ms)=12220

Physical memory (bytes) snapshot=6008360960

Virtual memory (bytes) snapshot=56607043584

Total committed heap usage (bytes)=5549064192

Peak Map Physical memory (bytes)=609275904

Peak Map Virtual memory (bytes)=4721360896

Peak Reduce Physical memory (bytes)=375566336

Peak Reduce Virtual memory (bytes)=4722372608

Shuffle Errors

BAD\_ID=0

CONNECTION=0

IO\_ERROR=0

WRONG\_LENGTH=0

WRONG\_MAP=0

WRONG\_REDUCE=0

File Input Format Counters

Bytes Read=855

File Output Format Counters

Bytes Written=172

2023-09-10 11:23:34,366 INFO streaming.StreamJob: Output directory: hdfs:///user/dataproc/salida.txt

Comprobamos la salida del script:

hdfs dfs -ls /user/dataproc

hdfs dfs -cat /user/dataproc/salida.txt/