
Grandes Volúmenes de Datos



Actividad MapReduce

12/10/2021

Marcos Eladio Somoza

ÍNDICE

1. Enunciado	4
2. Parte 1.....	5
3. Parte 2.....	5
4. Parte 3.....	6
5. Parte 4.....	8
6. Parte 5.....	9

1. Enunciado

Los objetivos de esta actividad son los siguientes:

- Configurar y arrancar un clúster Hadoop en la nube.
- Aprender los comandos para trabajar con el sistema de archivos distribuido HDFS.
- Entender cómo funciona el algoritmo MapReduce para la ejecución de procesos en paralelo.

Para poder realizar de forma correcta esta actividad es necesario contar con los siguientes requisitos:

- Tener una cuenta en Google Cloud Platform.
- Un equipo con al menos 8Gb de RAM.

Los enunciados de cada paso son:

- **Paso 1.** Descargar el texto del libro "La Celestina" de la página del proyecto Gutenberg (<http://www.gutenberg.org/ebooks/1619> (Enlaces a un sitio externo.)) y cargarlo en HDFS.
- **Paso 2.** Visualizar los videos del Campus virtual que explican como ejecutar el contador de palabras con MapReduce. Replicar el ejemplo y hacerlo funcionar.
- **Paso 3.** Crear un clúster Dataproc con tres nodos en Google Cloud Platform y ejecutar el contador de palabras. Ver cuánto tiempo tarda. Eliminar el clúster.
- **Paso 4.** Crear un nuevo clúster Dataproc con el doble de nodos y ejecutar el contador de palabras. Ver cuánto tiempo tarda.
- **Paso 5.** Buscar una colección pública de tweets y, a partir del ejemplo del contador de palabras, implementar un programa MapReduce en Java que recibe un fichero de tweets como entrada, y escribe otro fichero con el conteo del número de hashtags (palabras que empiezan por una #) que contienen. Ejecutar el contador de hashtags en el clúster.

2. Parte 1

Para descargar el libro deseado en el clúster, usaremos el comando `wget`. Para poder trabajar con todos los clústers con el libro, se deberá copiar el archivo del libro a HDFS.

```
somozadev@cluster-somoza-m:~$ wget https://www.gutenberg.org/cache/epub/1619/pg1619.txt
--2021-10-14 01:30:52-- https://www.gutenberg.org/cache/epub/1619/pg1619.txt
Resolving www.gutenberg.org (www.gutenberg.org)... 152.19.134.47, 2610:28:3090:3000:0:bad:cafe:47
Connecting to www.gutenberg.org (www.gutenberg.org)|152.19.134.47|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 687503 (671K) [text/plain]
Saving to: 'pg1619.txt'

pg1619.txt                                100% [=====]

2021-10-14 01:30:53 (1.44 MB/s) - 'pg1619.txt' saved [687503/687503]

somozadev@cluster-somoza-m:~$
```

Crearemos una carpeta `datos` en HDFS mediante el comando `hdfs dfs -mkdir /datos`. Ahora, copiaremos el libro a dicha carpeta:

```
somozadev@cluster-somoza-m:~$ hdfs dfs -put pg1619.txt /datos
somozadev@cluster-somoza-m:~$ hdfs dfs -ls /datos
Found 1 items
-rw-r--r-- 2 somozadev hadoop 687503 2021-10-14 01:34 /datos/pg1619.txt
somozadev@cluster-somoza-m:~$
```

3. Parte 2

Para aplicar el contador de palabras, se deberá llamar a `wordcount` disponible en los ejemplos de `hadoop`. Para ello, se usará el siguiente comando:

Pasando como primer parámetro el libro (el fichero que se desea contar) y como segundo el directorio de salida.

```
somozadev@cluster-somoza-m:~$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount /datos/pg1619.txt /datos/out
2021-10-14 01:39:28,435 INFO client.RMProxy: Connecting to ResourceManager at cluster-somoza-m/10.132.0.25:8032
2021-10-14 01:39:28,704 INFO client.AHSProxy: Connecting to Application History server at cluster-somoza-m/10.132.0.25:10200
2021-10-14 01:39:29,013 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/somozadev/.staging/job_1634174697116_0001
2021-10-14 01:39:29,501 INFO Input.FileInputFormat: Total input files to process : 1
2021-10-14 01:39:29,662 INFO mapreduce.JobSubmitter: number of splits:1
2021-10-14 01:39:29,856 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1634174697116_0001
2021-10-14 01:39:29,858 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-14 01:39:30,144 INFO conf.Configuration: resource-types.xml not found
2021-10-14 01:39:30,144 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-14 01:39:30,653 INFO impl.YarnClientImpl: Submitted application application_1634174697116_0001
2021-10-14 01:39:30,813 INFO mapreduce.Job: The url to track the job: http://cluster-somoza-m:8088/proxy/application_1634174697116_0001/
2021-10-14 01:39:30,814 INFO mapreduce.Job: Running job: job_1634174697116_0001
```

```
2021-10-14 01:40:13,404 INFO mapreduce.Job: Job job_1634174697116_0001 completed successfully
2021-10-14 01:40:13,516 INFO mapreduce.Job: Counters: 55
  File System Counters
    FILE: Number of bytes read=463119
    FILE: Number of bytes written=2394625
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=687608
    HDFS: Number of bytes written=338966
    HDFS: Number of read operations=28
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=15
    HDFS: Number of bytes read erasure-coded=0
  Job Counters
    Killed reduce tasks=1
    Launched map tasks=1
    Launched reduce tasks=5
    Data-local map tasks=1
    Total time spent by all maps in occupied slots (ms)=22330368
    Total time spent by all reduces in occupied slots (ms)=134055936
    Total time spent by all map tasks (ms)=7269
    Total time spent by all reduce tasks (ms)=43638
    Total vcore-milliseconds taken by all map tasks=7269
    Total vcore-milliseconds taken by all reduce tasks=43638
    Total megabyte-milliseconds taken by all map tasks=22330368
    Total megabyte-milliseconds taken by all reduce tasks=134055936

  Map-Reduce Framework
    Map input records=18453
    Map output records=109859
    Map output bytes=1100473
    Map output materialized bytes=463119
    Input split bytes=105
    Combine input records=109859
    Combine output records=31284
    Reduce input groups=31284
    Reduce shuffle bytes=463119
    Reduce input records=31284
    Reduce output records=31284
    Spilled Records=62568
    Shuffled Maps =5
    Failed Shuffles=0
    Merged Map outputs=5
    GC time elapsed (ms)=970
    CPU time spent (ms)=8130
    Physical memory (bytes) snapshot=1738179560
    Virtual memory (bytes) snapshot=26336030720
    Total committed heap usage (bytes)=1390411776
    Peak Map Physical memory (bytes)=533757952
    Peak Map Virtual memory (bytes)=438894592
    Peak Reduce Physical memory (bytes)=258879488
    Peak Reduce Virtual memory (bytes)=4392288256
  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=687503
  File Output Format Counters
    Bytes Written=338966
```

Cuando termine, se guardará el resultado en la carpeta indicada /datos/out como se puede ver:

```
somozadev@cluster-somoza-m:~$ hdfs dfs -ls /datos/out
Found 6 items
-rw-r--r--  2 somozadev hadoop      0 2021-10-14 01:40 /datos/out/_SUCCESS
-rw-r--r--  2 somozadev hadoop  67541 2021-10-14 01:40 /datos/out/part-r-00000
-rw-r--r--  2 somozadev hadoop  67117 2021-10-14 01:40 /datos/out/part-r-00001
-rw-r--r--  2 somozadev hadoop  67609 2021-10-14 01:40 /datos/out/part-r-00002
-rw-r--r--  2 somozadev hadoop  67952 2021-10-14 01:40 /datos/out/part-r-00003
-rw-r--r--  2 somozadev hadoop  68747 2021-10-14 01:40 /datos/out/part-r-00004
```

Para comprobar que funcionó correctamente, se pueden visualizar los ficheros con -cat:

```
somozadev@cluster-somoza-m:~$ hdfs dfs -cat /datos/out/part-r-00000
```

```
you 25
yrada. 1
yre: 1
yre? 1
yrle 1
yua): 1
yuamos 1
yuas, 1
zapatas. 1
zapatero" 1
zapato 1
zarazas: 1
zargatona 1
zip 1
zozobra_ 1
zozobras 1
zumo 3
zumos 2
zurrido?; 1
zurujano 1
```

4. Parte 3

Para crear el clúster, hay que asignar la región adecuada, en nuestro caso (al ser europeos) europe-west1 y la zona predeterminada vale, europe-west1-b. En cuanto a la imagen del clúster, se utilizará Ubuntu (más concretamente 2.0-ubuntu18), dado que se ha estado utilizando Ubuntu durante la asignatura.

Clúster

Cloud Dataproc

Google Cloud Dataproc te permite aprovisionar clústeres de Apache Hadoop y conectarte a almacenes de datos de análisis subyacentes.

No hay clústeres en las regiones de Cloud Dataproc seleccionadas actualmente. Crea un clúster para comenzar.

CREAR CLÚSTER

Nombre

Nombre del clúster *

cluster-somoza

Ubicación

Región *

europe-west1

Zona *

europe-west1-b

Tipo de clúster

☒ Estándar (1 principal, N trabajadores)

Control de versiones

Usa una imagen personalizada para cargar paquetes preinstalados. [Más información](#)

Tipo de imagen y versión

2.0-ubuntu18

En cuanto a la arquitectura del clúster, se usará un main node con 2CPUs, 7.5Gb de ram y 500Gb de disco. También, tres worker nodes con 2CPUs, 7.5Gb de ram y 200Gb de disco cada uno.

Nodo principal

Contiene el administrador de recursos YARN, HDFS NameNode y todos los controladores de trabajos.

Familia de máquinas

USO GENERAL OPTIMIZADA PARA PROCESAMIENTO MEMORIA OPTIMIZADA

Tipos de máquinas para cargas de trabajo comunes, optimizados en función del costo y la flexibilidad

Serie N1

Con la tecnología de la plataforma de CPU Intel Skylake o uno de sus predecesores

Tipo de máquina n1-standard-2 (2 CPU virtuales, 7.5 GB de memoria)

vCPU 2 Memory 7.5 GB

PLATAFORMA DE CPU Y GPU

Tamaño del disco principal (...) 500 GB Primary disk type Standard Persistent Disk

Cantidad de SSD locales * 0 x 375GB

Nodos trabajadores

Cada uno contiene un NodeManager de YARN y un DataNode de HDFS. El factor de replicación de HDFS es 2.

Familia de máquinas

USO GENERAL OPTIMIZADA PARA PROCESAMIENTO MEMORIA OPTIMIZADA

Tipos de máquinas para cargas de trabajo comunes, optimizados en función del costo y la flexibilidad

Serie N1

Con la tecnología de la plataforma de CPU Intel Skylake o uno de sus predecesores

Tipo de máquina n1-standard-2 (2 CPU virtuales, 7.5 GB de memoria)

vCPU 2 Memory 7.5 GB

PLATAFORMA DE CPU Y GPU

Number of worker nodes 3

Tamaño del disco principal (...) 100 GB Primary disk type Standard Persistent Disk

Cantidad de SSD locales * 0 x 375GB

```



somozadev@cluster-somoza-m:~$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordc
2021-10-14 01:50:16,387 INFO client.RMProxy: Connecting to ResourceManager at cluster-somoza-m/10.132
2021-10-14 01:50:16,649 INFO client.AHSProxy: Connecting to Application History server at cluster-som
2021-10-14 01:50:16,897 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/h
2021-10-14 01:50:17,202 INFO input.FileInputFormat: Total input files to process : 1
2021-10-14 01:50:17,315 INFO mapreduce.JobSubmitter: number of splits:1
2021-10-14 01:50:17,507 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1634174697116_000
2021-10-14 01:50:17,509 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-14 01:50:17,797 INFO conf.Configuration: resource-types.xml not found
2021-10-14 01:50:17,798 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-14 01:50:17,915 INFO impl.YarnClientImpl: Submitted application application_1634174697116_000
2021-10-14 01:50:17,998 INFO mapreduce.Job: The url to track the job: http://cluster-somoza-m:8088/pr
2021-10-14 01:50:18,002 INFO mapreduce.Job: Running job: job_1634174697116_0002
2021-10-14 01:50:27,187 INFO mapreduce.Job: Job job_1634174697116_0002 running in uber mode : false
2021-10-14 01:50:27,188 INFO mapreduce.Job: map 0% reduce 0%
2021-10-14 01:50:34,273 INFO mapreduce.Job: map 100% reduce 0%
2021-10-14 01:50:40,314 INFO mapreduce.Job: map 100% reduce 20%
2021-10-14 01:50:43,339 INFO mapreduce.Job: map 100% reduce 60%
2021-10-14 01:50:44,348 INFO mapreduce.Job: map 100% reduce 100%
2021-10-14 01:50:46,370 INFO mapreduce.Job: Job job_1634174697116_0002 completed successfully
  
```

Al usar wordcount con La Celestina con esta configuración de clústers, vemos que ha empezado en el tiempo 01:50:18, y ha terminado en 01:50:46. Esto nos indica que ha tardado **28 segundos** en realizar el MapReduce con un master node y tres workers.

Ahora borraremos el clúster.

Filtro

Busca clústeres y presiona Intro

<input checked="" type="checkbox"/>	Nombre 	Estado	Región	Zona	Total de nodos tr
<input checked="" type="checkbox"/>	cluster-somoza	 Eliminación en curso	europe-west1	europe-west1-b	3

5. Parte 4

Se ha creado un clúster igual al anterior pero con el doble de nodos trabajadores (es decir, con un master node pero seis workers).

Nodos trabajadores

Cada uno contiene un NodeManager de YARN y un DataNode de HDFS. El factor de replicación de HDFS es 2.

Familia de máquinas


USO GENERAL OPTIMIZADA PARA PROCESAMIENTO MEMORIA OPTIMIZADA

Tipos de máquinas para cargas de trabajo comunes, optimizados en función del costo y la flexibilidad

Serie: N1

Con la tecnología de la plataforma de CPU Intel Skylake o uno de sus predecesores

Tipo de máquina: n1-standard-2 (2 CPU virtuales, 7.5 GB de memoria)

 vCPU: 2 Memory: 7.5 GB

✓ PLATAFORMA DE CPU Y GPU

Number of worker nodes: 6

Tamaño del disco principal (...): 500 GB Primary disk type: Standard Persistent Disk

Cantidad de SSD locales *: 0 x 375GB

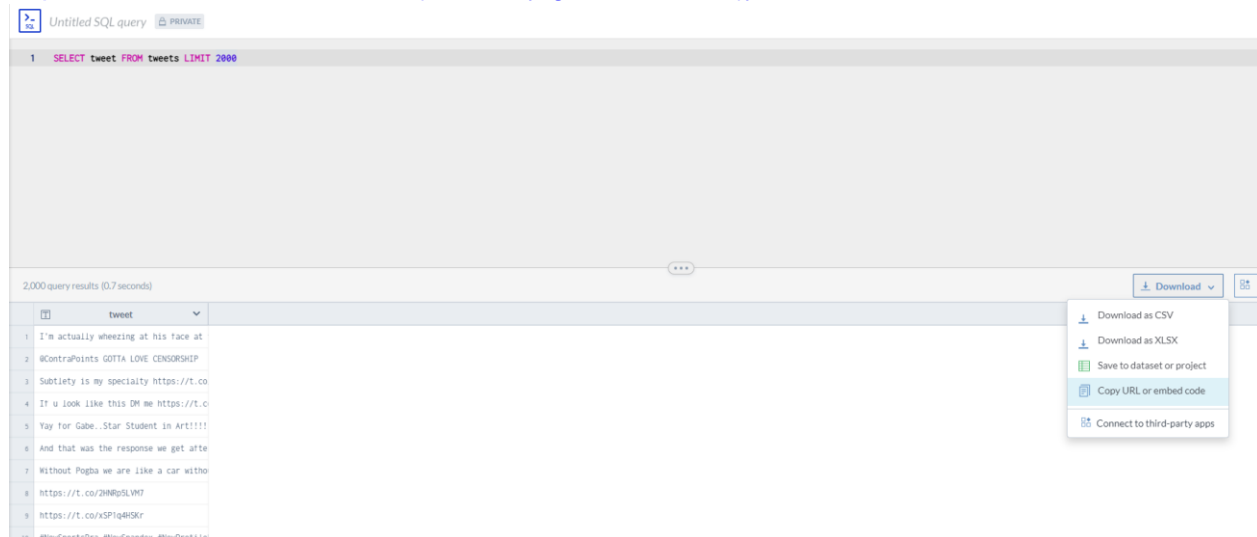
Al ejecutar MapReduce con La Celestina, vemos que :

```
somoza@cluster-somoza2-m:~$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount
2021-10-14 02:07:49,122 INFO client.RMProxy: Connecting to ResourceManager at cluster-somoza2-m/10.132.1.100
2021-10-14 02:07:49,401 INFO client.AHSProxy: Connecting to Application History server at cluster-somoza2-m/10.132.1.100
2021-10-14 02:07:49,727 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-somoza2-m/STAGING/tmp/hadoop-somoza2-m-2021-10-14_02-07-49_727
2021-10-14 02:07:50,283 INFO input.FileInputFormat: Total input files to process : 1
2021-10-14 02:07:50,287 WARN concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted
java.lang.InterruptedException
    at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
    at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
    at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:115)
    at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:115)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:748)
2021-10-14 02:07:50,855 INFO mapreduce.JobSubmitter: number of splits:1
2021-10-14 02:07:51,082 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1634177000902_0001
2021-10-14 02:07:51,085 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-14 02:07:51,379 INFO conf.Configuration: resource-types.xml not found
2021-10-14 02:07:51,379 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-14 02:07:51,838 INFO impl.YarnClientImpl: Submitted application application_1634177000902_0001
2021-10-14 02:07:51,984 INFO mapreduce.Job: The url to track the job: http://cluster-somoza2-m:8088/track/1634177000902_0001
2021-10-14 02:07:51,989 INFO mapreduce.Job: Running job: job_1634177000902_0001
2021-10-14 02:08:03,234 INFO mapreduce.Job: Job job_1634177000902_0001 running in uber mode : false
2021-10-14 02:08:03,236 INFO mapreduce.Job: map 0% reduce 0%
2021-10-14 02:08:14,385 INFO mapreduce.Job: map 100% reduce 0%
2021-10-14 02:08:20,424 INFO mapreduce.Job: map 100% reduce 9%
2021-10-14 02:08:23,442 INFO mapreduce.Job: map 100% reduce 18%
2021-10-14 02:08:24,447 INFO mapreduce.Job: map 100% reduce 27%
2021-10-14 02:08:25,451 INFO mapreduce.Job: map 100% reduce 36%
2021-10-14 02:08:26,458 INFO mapreduce.Job: map 100% reduce 100%
2021-10-14 02:08:27,470 INFO mapreduce.Job: Job job_1634177000902_0001 completed successfully
2021-10-14 02:08:27,574 INFO mapreduce.Job: Counters: 55
```

Ha empezado en el tiempo 02:07:51, y ha terminado en 02:08:27. Esto nos indica que ha tardado **36 segundos** en realizar el MapReduce con un master node y seis workers. Ha tardado más que con 3 workers dado que tarda más tiempo en repartir el trabajo entre los workers del que gana trabajando con 6 en lugar de con 3.

6. Parte 5

Para conseguir la colección de tweets, se ha utilizado la página <https://data.world/> en donde se ha encontrado un dataset de tweets. Gracias a la query **SELECT tweet FROM tweets LIMIT 2000** se han aislado 2000 tweets de esta y guardado para descargar en <https://download.data.world/s/ntjoxziox6ytlg62odcvdlttxojs2>.



Ya con el input disponible, se descargará al clúster con wget. También, se creará el fichero WordCount.java. Para hacer que cuente el número de “#” habrá que modificar dos líneas de código en el map:

```
public void map(Object key, Text value, Context context
                ) throws IOException, InterruptedException {
    StringTokenizer itr = new StringTokenizer(value.toString());
    while (itr.hasMoreTokens()) {
        String str = itr.nextToken();
        if(str.substring(0,1).equals("#")){
            word.set(str);
            context.write(word, one);
        }
    }
}
```

Para poder compilar el WordCount, primero hay que asignar el HADOOP_CLASSPATH con:

```
somozadev@cluster-da47-m:~$ export HADOOP_CLASSPATH=$(hadoop classpath)
```

Así, se podrá compilar con el siguiente comando:

```
somozadev@cluster-da47-m:~$ mkdir WordCount
somozadev@cluster-da47-m:~$ javac -classpath ${HADOOP_CLASSPATH} -d WordCount/ WordCount.java
somozadev@cluster-da47-m:~$ ls WordCount/
WordCount$IntSumReducer.class WordCount$TokenizerMapper.class WordCount.class
somozadev@cluster-da47-m:~$
```

Y finalmente:

```
somozadev@cluster-da47-m:~$ jar -cvf WordCount.jar -C WordCount/ .
added manifest
adding: WordCount$IntSumReducer.class(in = 1739) (out= 737) (deflated 57%)
adding: WordCount.class(in = 1491) (out= 817) (deflated 45%)
adding: WordCount$TokenizerMapper.class(in = 1835) (out= 806) (deflated 56%)
```


Para poder hacer el WordCount, habrá que copiar el fichero tweets.txt a HDFS

```
somozadev@cluster-da47-m:~$ hdfs dfs -put tweets.txt /
somozadev@cluster-da47-m:~$ hdfs dfs -ls /
Found 4 items
-rw-r--r--  2 somozadev hadoop      169815 2021-10-14 18:33 /datos
drwxrwxrwt  - hdfs      hadoop           0 2021-10-14 17:53 /tmp
-rw-r--r--  2 somozadev hadoop      169815 2021-10-14 18:36 /tweets.txt
drwxrwxrwt  - hdfs      hadoop           0 2021-10-14 17:53 /user
```

Y finalmente, hacer la llamada al jar mediante hadoop:

```
somozadev@cluster-da47-m:~$ hadoop jar /home/somozadev/WordCount.jar WordCount /tweets.txt /out
2021-10-14 18:51:24,944 INFO client.RMPProxy: Connecting to ResourceManager at cluster-da47-m/10.132.0.36:8032
2021-10-14 18:51:25,192 INFO client.AHSProxy: Connecting to Application History server at cluster-da47-m/10.132.0.36:10200
2021-10-14 18:51:25,419 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface with ToolRunner to remedy this.
2021-10-14 18:51:25,443 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/somozadev/.stage
2021-10-14 18:51:25,803 INFO input.FileInputFormat: Total input files to process : 1
2021-10-14 18:51:25,909 INFO mapreduce.JobSubmitter: number of splits:1
2021-10-14 18:51:26,116 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1634233981025_0002
2021-10-14 18:51:26,118 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-14 18:51:26,418 INFO conf.Configuration: resource-types.xml not found
2021-10-14 18:51:26,419 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-14 18:51:26,648 INFO impl.YarnClientImpl: Submitted application application_1634233981025_0002
2021-10-14 18:51:26,755 INFO mapreduce.Job: The url to track the job: http://cluster-da47-m:8088/proxy/application_1634233981025_0002/
2021-10-14 18:51:26,755 INFO mapreduce.Job: Running job: job_1634233981025_0002
2021-10-14 18:51:35,964 INFO mapreduce.Job: Job job_1634233981025_0002 running in uber mode : false
2021-10-14 18:51:35,965 INFO mapreduce.Job: map 0% reduce 0%
2021-10-14 18:51:42,042 INFO mapreduce.Job: map 100% reduce 0%
2021-10-14 18:51:49,090 INFO mapreduce.Job: map 100% reduce 20%
2021-10-14 18:51:51,103 INFO mapreduce.Job: map 100% reduce 40%
2021-10-14 18:51:52,110 INFO mapreduce.Job: map 100% reduce 80%
2021-10-14 18:51:53,143 INFO mapreduce.Job: map 100% reduce 100%
2021-10-14 18:51:55,167 INFO mapreduce.Job: Job job_1634233981025_0002 completed successfully
2021-10-14 18:51:55,279 INFO mapreduce.Job: Counters: 55
File System Counters
```

Como se puede ver, el proceso MapReduce empezó a las 18:51:26 y terminó a las 18:51:55, tardando **29 segundos** en hacer el conteo.

```
File System Counters
FILE: Number of bytes read=5658
FILE: Number of bytes written=1477939
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=169912
HDFS: Number of bytes written=4413
HDFS: Number of read operations=28
HDFS: Number of large read operations=0
HDFS: Number of write operations=15
HDFS: Number of bytes read erasure-coded=0
```

```
Job Counters
Killed reduce tasks=1
Launched map tasks=1
Launched reduce tasks=5
Rack-local map tasks=1
Total time spent by all maps in occupied slots (ms)=12429312
Total time spent by all reduces in occupied slots (ms)=93032448
Total time spent by all map tasks (ms)=1046
Total time spent by all reduce tasks (ms) 30284
Total vcore-milliseconds taken by all map tasks=1046
Total vcore-milliseconds taken by all reduce tasks=30284
Total megabyte-milliseconds taken by all map tasks=12429312
Total megabyte-milliseconds taken by all reduce tasks=93032448
```

```
Map-Reduce Framework
Map input records=2000
Map output records=454
Map output bytes=7173
Map output materialized bytes=5658
Input split bytes=91
Combine input records=454
Combine output records=305
Reduce input groups=305
Reduce shuffle bytes=5658
Reduce input records=305
Reduce output records=305
Spilled Records=610
Shuffled Maps=5
Failed shuffles=0
Merged Map outputs=5
GC time elapsed (ms)=789
CPU time spent (ms)=4108
Physical memory (bytes) snapshot=1766363136
Virtual memory (bytes) snapshot=26326405120
Total committed heap usage (bytes)=131286456
Peak Map Physical memory (bytes)=524333056
Peak Map virtual memory (bytes)=1380237824
Peak Reduce Physical memory (bytes)=257622016
Peak Reduce Virtual memory (bytes)=4392841216
```

```
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=169815
File Output Format Counters
Bytes Written=4413
```

```
somozadev@cluster-da47-m:~$ hdfs dfs -ls /out
Found 6 items
-rw-r--r--  2 somozadev hadoop           0 2021-10-14 18:51 /out/_SUCCESS
-rw-r--r--  2 somozadev hadoop        976 2021-10-14 18:51 /out/part-r-00000
-rw-r--r--  2 somozadev hadoop        980 2021-10-14 18:51 /out/part-r-00001
-rw-r--r--  2 somozadev hadoop        707 2021-10-14 18:51 /out/part-r-00002
-rw-r--r--  2 somozadev hadoop        822 2021-10-14 18:51 /out/part-r-00003
-rw-r--r--  2 somozadev hadoop        928 2021-10-14 18:51 /out/part-r-00004
```

```
#NewProfilePic 1
#NoodlesandCompany 1
#NowWeTrainedOurDragon 1
#PassTheFlask 1
#PetsOfInstagram 2
#RawrXD 1
#RespectTheBeard" 1
#SDSU 3
#Sam 1
#TBS. 1
#TeamOnsharp. 1
#TurkeyWhisperer 1
#WhosNext 1
#WhyDoILiveHere 1
#Winter 1
#almostparadise" 1
#bestjob 1
#bitteraboutdick 1
#christmascactus 1
#cobbercoachesshow" 1
#cordmn 3
#djaj 1
#dope 1
#entrepreneurship 1
#fmwfmembers 1
#freekbailey 1
#gfpsBF 1
#gofundme? 1
#homebound" 1
#keepemcomin 1
#killme" 1
#letthebeatdrawwp" 1
#lostinfargo" 1
#loveothers 1
#marathon"" 1
#mydayismade 1
#photo" 2
#playoffs 1
#saytheword 1
#singer 1
#smh2018" 1
#thereisnoboycott" 1
#winning 1
somozadev@cluster-da47-m:~$
```

Como apartado extra, me he tomado la libertad de realizar de nuevo el WordCounter pero asignando como salida un Google Cloud Storage, a fin de poder entregar los ficheros de salida en el zip. Para ello, he cambiado el directorio de salida /out a gs://digdata-bucket-somoza/out


```
somozadev@cluster-da47-m:~$ hadoop jar /home/somozadev/WordCount.jar WordCount /tweets.txt gs://digdata-bucket-somoza/out
```

Además de añadir también el fichero tweets.txt y el WordCounter.java.

digdata-bucket-somoza



Ubicación	Clase de almacenamiento	Acceso público	Protección
europa-west1 (Bélgica)	Standard	No público	Ninguna

OBJETOS CONFIGURACIÓN PERMISOS PROTECCIÓN CICLO DE VIDA

Depósitos > digdata-bucket-somoza 

[SUBIR ARCHIVOS](#) [SUBIR CARPETA](#) [CREAR CARPETA](#) [ADMINISTRAR CONSERVACIONES](#) [DESCARGAR](#) [BORRAR](#)


Filtrar solo por prefijo de nombre ▼  **Filtro** Filtrar objetos y carpetas

<input type="checkbox"/>	Nombre	Tamaño	Tipo	Fecha de creación ?	Clase de almacenamiento
<input type="checkbox"/>	 WordCount.java	2.1 KB	text/x-java	14 oct. 2021 21:05:42	Standard
<input type="checkbox"/>	 out/	—	Carpeta	—	—
<input type="checkbox"/>	 tweets.txt	165.8 KB	text/plain	14 oct. 2021 21:01:45	Standard

digdata-bucket-somoza







Ubicación	Clase de almacenamiento	Acceso público	Protección
europa-west1 (Bélgica)	Standard	No público	Ninguna

OBJETOS CONFIGURACIÓN PERMISOS PROTECCIÓN CICLO DE VIDA

Depósitos > digdata-bucket-somoza > out 

[SUBIR ARCHIVOS](#) [SUBIR CARPETA](#) [CREAR CARPETA](#) [ADMINISTRAR CONSERVACIONES](#) [DESCARGAR](#) [BORRAR](#)

Filtrar solo por prefijo de nombre ▼  **Filtro** Filtrar objetos y carpetas

<input type="checkbox"/>	Nombre	Tamaño	Tipo	Fecha de creación ?	Clase de almacenamiento
<input type="checkbox"/>	 _SUCCESS	0 B	application/octet-stream	14 oct. 2021 21:04:27	Standard
<input type="checkbox"/>	 part-r-00000	976 B	application/octet-stream	14 oct. 2021 21:04:24	Standard
<input type="checkbox"/>	 part-r-00001	980 B	application/octet-stream	14 oct. 2021 21:04:22	Standard
<input type="checkbox"/>	 part-r-00002	707 B	application/octet-stream	14 oct. 2021 21:04:25	Standard
<input type="checkbox"/>	 part-r-00003	822 B	application/octet-stream	14 oct. 2021 21:04:26	Standard
<input type="checkbox"/>	 part-r-00004	928 B	application/octet-stream	14 oct. 2021 21:04:26	Standard