

INFRAESTRUCTURA PARA BIG DATA

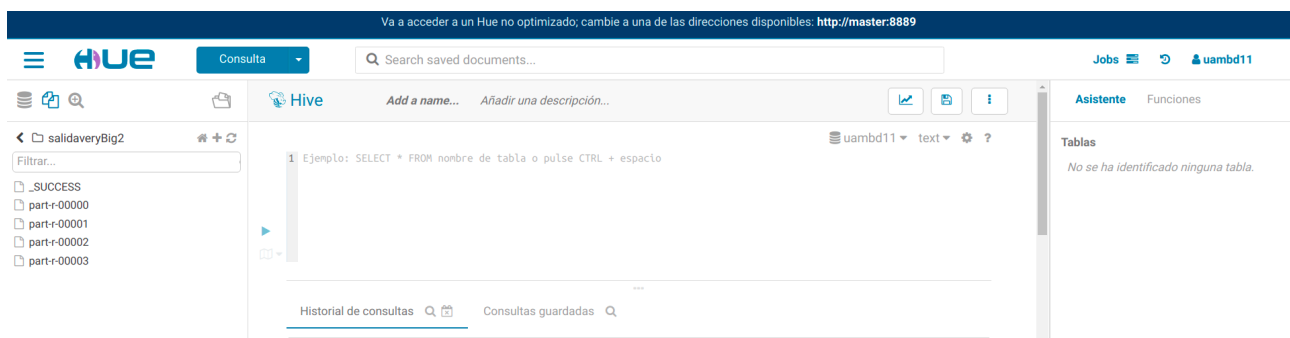
PRÁCTICA 1

Daniel Pérez Efremova
Diciembre de 2021

- Pregunta 1: ¿Cómo justificarías que el fichero de salida esté partido en varias partes (24 en el ejemplo)?

Al ejecutar la tarea se obtienen 4 partes. Revisando por consola el resumen del trabajo se ve que se ha realizado 1 map y 4 reducees.

Por tanto, la razón por la que el resultado se presenta en 4 bloques es la cantidad de reducees que se han aplicado.



```
[uambd11@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-examples.jar wordcount veryBig.txt salidaveryBig2/
21/12/18 13:41:13 INFO client.RMProxy: Connecting to ResourceManager at master/10.10.1.10:8032
21/12/18 13:41:13 INFO input.FileInputFormat: Total input paths to process : 1
21/12/18 13:41:13 INFO mapreduce.JobSubmitter: number of splits:1
21/12/18 13:41:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1639163889378_0362
21/12/18 13:41:14 INFO impl.YarnClientImpl: Submitted application application_1639163889378_0362
21/12/18 13:41:14 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1639163889378_0362/
21/12/18 13:41:14 INFO mapreduce.Job: Running job: job_1639163889378_0362
21/12/18 13:41:18 INFO mapreduce.Job: Job job_1639163889378_0362 running in uber mode : false
21/12/18 13:41:18 INFO mapreduce.Job: map 0% reduce 0%
21/12/18 13:41:34 INFO mapreduce.Job: map 46% reduce 0%
21/12/18 13:41:40 INFO mapreduce.Job: map 58% reduce 0%
21/12/18 13:41:46 INFO mapreduce.Job: map 67% reduce 0%
21/12/18 13:41:47 INFO mapreduce.Job: map 100% reduce 0%
21/12/18 13:41:52 INFO mapreduce.Job: map 100% reduce 25%
21/12/18 13:41:54 INFO mapreduce.Job: map 100% reduce 100%
21/12/18 13:41:54 INFO mapreduce.Job: Job job_1639163889378_0362 completed successfully
21/12/18 13:41:55 INFO mapreduce.Job: Counters: 49

File System Counters
  FILE: Number of bytes read=3046260
  FILE: Number of bytes written=4307559
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=125329020
  HDFS: Number of bytes written=1025190
  HDFS: Number of read operations=15
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=8

Job Counters
  Launched map tasks=1
  Launched reduce tasks=4
  Rack-local map tasks=1
  Total time spent by all maps in occupied slots (ms)=26578
  Total time spent by all maps in occupied slots (ms)=16848
  Total time spent by all map tasks (ms)=26578
  Total time spent by all reduce tasks (ms)=16848
  Total vcore-milliseconds taken by all map tasks=26578
  Total vcore-milliseconds taken by all reduce tasks=16848
```

- Pregunta 2:

1. ¿Cuál es el tamaño de bloque configurado para el HDFS del clúster? **128 MiB**
 2. ¿Cuál es el factor de replicación por defecto del HDFS? **3**
 3. ¿Cuál es el número de tareas MapReduce que se lanzarán por defecto al crear un nuevo trabajo? ¿Y para una tarea lanzada desde Hive?
- **4 Reducees, no hay número determinado de tareas map ya que depende de las particiones en el fichero input.**
- **Hive calcula por proceso de optimización interno el número de reducees. El número de maps viene determinado por el número de particiones en el fichero input.**

The top screenshot shows the Cloudera Manager interface for HDFS configuration. The 'Tamaño de bloque de HDFS' (HDFS block size) is set to 128 MiB. The bottom screenshot shows the 'Factor de replicación' (HDFS replication factor) set to 3.

- Pregunta 3: utilizar el interfaz de HUE para cargar un fichero en tu directorio del HDFS. Cópialo cambiándole el nombre, y después elimina la copia original.

Realizado para ejecutar el wordcount del quijote.

- Pregunta 4: Consultar los siguientes datos de configuración de la tarea “WordCount” de ejemplo lanzada al inicio de la práctica:

- Número de tareas Map lanzadas: **1**
- Número de tareas Reduce lanzadas: **4**
- Duración de la ejecución de la tarea: **36 s**
- Estado de terminación: **SUCCEEDED**

The screenshot shows the HUE interface for a WordCount job. The job is named 'word count' and is in the 'SUCCEEDED' state. The table shows 4 Reduce tasks and 1 Map task, all completed successfully.

Tipo	ID	Tiempo transcurrido	Progreso	Estado	Hora de inicio	Intento correcto
REDUCE	task_1639163889378_0362_r_000001	4.63s	100	SUCCEEDED	18 de diciembre de 2021 13:41	attempt_1639163889378_0362_r_000001
REDUCE	task_1639163889378_0362_r_000002	4.73s	100	SUCCEEDED	18 de diciembre de 2021 13:41	attempt_1639163889378_0362_r_000002
REDUCE	task_1639163889378_0362_r_000000	4.57s	100	SUCCEEDED	18 de diciembre de 2021 13:41	attempt_1639163889378_0362_r_000000
REDUCE	task_1639163889378_0362_r_000003	2.92s	100	SUCCEEDED	18 de diciembre de 2021 13:41	attempt_1639163889378_0362_r_000003
MAP	task_1639163889378_0362_m_000000	26.58s	100	SUCCEEDED	18 de diciembre de 2021 13:41	attempt_1639163889378_0362_m_000000

- Pregunta 4: vuelva a la lanzar la ejecución de una tarea “WordCount” sobre el fichero “veryBig.txt”. Acceda durante su ejecución al Job Browser de HUE y, utilizando la interfaz gráfica, termina (mata) la tarea. ¿Qué mensaje obtenemos en la consola desde la que lanzamos la tarea? ¿Y qué vemos en la configuración del trabajo matado en HUE?

1. La consola muestra un mensaje que contiene el Id del trabajo y que se ha matado.

```
(base) daniel@daniel:~$ ssh uambd11@150.244.65.34
uambd11@150.244.65.34's password:
Last login: Sat Dec 18 13:32:32 2021 from 172.30.192.124
uambd11@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-examples.jar wordcount veryBig.txt salidaveryBig3/
21/12/18 15:17:06 INFO client.RMProxy: Connecting to ResourceManager at master/10.10.1.10:8032
21/12/18 15:17:07 INFO Input.FileInputFormat: Total input paths to process : 1
21/12/18 15:17:07 INFO mapreduce.JobSubmitter: number of splits:1
21/12/18 15:17:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1639163889378_0367
21/12/18 15:17:07 INFO impl.YarnClientImpl: Submitted application application_1639163889378_0367
21/12/18 15:17:07 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1639163889378_0367/
21/12/18 15:17:07 INFO mapreduce.Job: Running job: job_1639163889378_0367
21/12/18 15:17:12 INFO mapreduce.Job: Job job_1639163889378_0367 running in uber mode : false
21/12/18 15:17:12 INFO mapreduce.Job: map 0% reduce 0%
21/12/18 15:17:29 INFO mapreduce.Job: map 31% reduce 0%
21/12/18 15:17:41 INFO mapreduce.Job: map 56% reduce 0%
21/12/18 15:17:47 INFO mapreduce.Job: map 58% reduce 0%
21/12/18 15:17:53 INFO mapreduce.Job: map 67% reduce 0%
21/12/18 15:17:55 INFO mapreduce.Job: map 68% reduce 0%
21/12/18 15:17:55 INFO mapreduce.Job: Job job_1639163889378_0367 failed with state KILLED due to: Application killed by user.
21/12/18 15:17:55 INFO mapreduce.Job: Counters: 0
uambd11@master ~]$
```

2. En HUE se ve que el trabajo se ha matado.

Va a acceder a un Hue no optimizado; cambie a una de las direcciones disponibles: <http://master:8889>

Consulta

Search saved documents...

Jobs Consultas Workflows Programas

user:uambd11

Satisfactorio En ejecución Erróneos

hace 2 horas - KILLED

MAPREDUCE

word count

application_1639163889378_0367

uambd11

hace 4 horas - SUCCEEDED

MAPREDUCE

word count

application_1639163889378_0362

2 trabajos

- Pregunta 5: Comprobar que el fichero se ha generado correctamente utilizando la línea de comandos de Hadoop, o la interfaz gráfica HUE ¿Cuántas tareas Map y Reduce se han lanzado para crear nuestro fichero?

Para crear el fichero se hacen 2 map.

```
uambd11@master:~$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-examples.jar teragen 40000000 terasort-input
21/12/18 17:45:55 INFO client.RMProxy: Connecting to ResourceManager at master/10.10.1.10:8032
21/12/18 17:45:56 INFO terasort.TeraGen: Generating 40000000 using 2
21/12/18 17:45:56 INFO mapreduce.JobSubmitter: number of splits:2
21/12/18 17:45:56 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1639163889378_0370
21/12/18 17:45:56 INFO impl.YarnClientImpl: Submitted application application_1639163889378_0370
21/12/18 17:45:56 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1639163889378_0370/
21/12/18 17:45:56 INFO mapreduce.Job: Running job: job_1639163889378_0370
21/12/18 17:46:01 INFO mapreduce.Job: Job job_1639163889378_0370 running in uber mode : false
21/12/18 17:46:01 INFO mapreduce.Job: map 0% reduce 0%
21/12/18 17:46:10 INFO mapreduce.Job: map 50% reduce 0%
21/12/18 17:46:25 INFO mapreduce.Job: map 79% reduce 0%
21/12/18 17:46:30 INFO mapreduce.Job: map 89% reduce 0%
21/12/18 17:46:31 INFO mapreduce.Job: map 100% reduce 0%
21/12/18 17:46:31 INFO mapreduce.Job: Job job_1639163889378_0370 completed successfully
21/12/18 17:46:31 INFO mapreduce.Job: Counters: 31
File System Counters
  FILE: Number of bytes read=0
  FILE: Number of bytes written=297762
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=170
  HDFS: Number of bytes written=4000000000
  HDFS: Number of read operations=0
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=4
Job Counters
  Launched map tasks=2
  Other local map tasks=2
  Total time spent by all maps in occupied slots (ms)=53496
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=53496
  Total vcore-milliseconds taken by all map tasks=53496
  Total megabyte-milliseconds taken by all map tasks=54779904
```

- Preguntas 6 y 7: compruebe el tamaño de los archivos de entrada y de salida de TeraSort. ¿Cómo están distribuidos los datos? ¿Sabrías explicar a qué se debe?

- Tamaño de entrada (input): 4 GB
- Tamaño de salida (output): 4GB

- Número de slots input: 2 slots de 2 GB
- Número de Slots output: 4 slots de 1 GB
- Para la tarea de ordenación se han realizado: 30 map y 4 reduces.

El número de slots se debe al número de reduces de las tareas.

```

uamdbd11@master:~$
21/12/18 17:50:51 INFO mapreduce.Job: Job job_1639163889378_0371 completed successfully
21/12/18 17:50:52 INFO mapreduce.Job: Counters: 50
File System Counters
  FILE: Number of bytes read=1779942442
  FILE: Number of bytes written=3533169446
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=4000000000
  HDFS: Number of bytes written=4000000000
  HDFS: Number of read operations=102
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=8
Job Counters
  Launched map tasks=30
  Launched reduce tasks=4
  Data-local map tasks=19
  Rack-local map tasks=11
  Total time spent by all maps in occupied slots (ms)=273946
  Total time spent by all reduces in occupied slots (ms)=100292
  Total time spent by all map tasks (ms)=273946
  Total time spent by all reduce tasks (ms)=100292
  Total vcore-milliseconds taken by all map tasks=273946
  Total vcore-milliseconds taken by all reduce tasks=100292
  Total megabyte-milliseconds taken by all map tasks=280520704
  Total megabyte-milliseconds taken by all reduce tasks=102699008
Map-Reduce Framework
  Map input records=400000000
  Map output records=400000000
  Map output bytes=4080000000
  Map output materialized bytes=1748111482
  Input split bytes=3720
  Combine input records=0
  Reduce input groups=400000000
  Reduce shuffle bytes=1748111482
  Reduce input records=400000000
  Reduce output records=400000000
  Spilled Records=800000000
  Shuffled Maps =120
  Failed Shuffles=0
  Merged Map outputs=120
  GC time elapsed (ms)=8229
  CPU time spent (ms)=269578

```

- Pregunta 8: variar el número de reducers lanzados, y tomar nota del tiempo de ejecución requerido.

- 4 Reduces (original): 46.46 seg
- 1 Reduce: 1 min 10 seg
- 2 Reduces: 57.3 seg
- 3 Reduces: 51.26 seg
- 5 Reduces: 44.41 seg
- 6 Reduces: 42.39 seg
- 7 Reduces: 38.77 seg
- 8 Reduces: 40.50 seg

A partir de 6 o 7 reduces, no se observan mejoras significativas en el rendimiento (con 24 reduces se tardan 35 seg en completar la tarea).

- Pregunta 9: Obtener datos de rendimiento en el sistema de ficheros distribuido del clúster. Obtener datos de rendimiento al pedir la escritura de 5,10,15 y 20 ficheros, y anótelos. ¿Qué tendencia se observa?

Se observa que al aumentar la cantidad de ficheros, la cantidad de trabajos por unidad de tiempo disminuye (Throughput mb/sec).

```
21/12/18 18:44:03 INFO fs.TestDFSIO: ----- TestDFSIO ----- : write
21/12/18 18:44:03 INFO fs.TestDFSIO:           Date & time: Sat Dec 18 18:44:03 CET 2021
21/12/18 18:44:03 INFO fs.TestDFSIO:           Number of files: 5
21/12/18 18:44:03 INFO fs.TestDFSIO: Total MBytes processed: 5000.0
21/12/18 18:44:03 INFO fs.TestDFSIO:           Throughput mb/sec: 26.01470351042409
21/12/18 18:44:03 INFO fs.TestDFSIO: Average IO rate mb/sec: 26.179065704345703
21/12/18 18:44:03 INFO fs.TestDFSIO: IO rate std deviation: 2.1645899176548693
21/12/18 18:44:03 INFO fs.TestDFSIO:           Test exec time sec: 57.143
21/12/18 18:44:03 INFO fs.TestDFSIO:
```

```
21/12/18 18:41:30 INFO fs.TestDFSIO: ----- TestDFSIO ----- : write
21/12/18 18:41:30 INFO fs.TestDFSIO:           Date & time: Sat Dec 18 18:41:30 CET 2021
21/12/18 18:41:30 INFO fs.TestDFSIO:           Number of files: 10
21/12/18 18:41:30 INFO fs.TestDFSIO: Total MBytes processed: 10000.0
21/12/18 18:41:30 INFO fs.TestDFSIO:           Throughput mb/sec: 21.35182686230634
21/12/18 18:41:30 INFO fs.TestDFSIO: Average IO rate mb/sec: 23.287153244018555
21/12/18 18:41:30 INFO fs.TestDFSIO: IO rate std deviation: 8.62275985064758
21/12/18 18:41:30 INFO fs.TestDFSIO:           Test exec time sec: 72.191
21/12/18 18:41:30 INFO fs.TestDFSIO:
```

```
21/12/18 18:46:48 INFO fs.TestDFSIO: ----- TestDFSIO ----- : write
21/12/18 18:46:48 INFO fs.TestDFSIO:           Date & time: Sat Dec 18 18:46:48 CET 2021
21/12/18 18:46:48 INFO fs.TestDFSIO:           Number of files: 15
21/12/18 18:46:48 INFO fs.TestDFSIO: Total MBytes processed: 15000.0
21/12/18 18:46:48 INFO fs.TestDFSIO:           Throughput mb/sec: 15.055857230324504
21/12/18 18:46:48 INFO fs.TestDFSIO: Average IO rate mb/sec: 16.095827102661133
21/12/18 18:46:48 INFO fs.TestDFSIO: IO rate std deviation: 4.876496217547487
21/12/18 18:46:48 INFO fs.TestDFSIO:           Test exec time sec: 95.606
21/12/18 18:46:48 INFO fs.TestDFSIO:
```

```
21/12/18 18:49:33 INFO fs.TestDFSIO: ----- TestDFSIO ----- : write
21/12/18 18:49:33 INFO fs.TestDFSIO:           Date & time: Sat Dec 18 18:49:33 CET 2021
21/12/18 18:49:33 INFO fs.TestDFSIO:           Number of files: 20
21/12/18 18:49:33 INFO fs.TestDFSIO: Total MBytes processed: 20000.0
21/12/18 18:49:33 INFO fs.TestDFSIO:           Throughput mb/sec: 10.848722698511285
21/12/18 18:49:33 INFO fs.TestDFSIO: Average IO rate mb/sec: 11.204983711242676
21/12/18 18:49:33 INFO fs.TestDFSIO: IO rate std deviation: 2.130820570860056
21/12/18 18:49:33 INFO fs.TestDFSIO:           Test exec time sec: 123.418
21/12/18 18:49:33 INFO fs.TestDFSIO:
```