

Ejecutar una tarea de prueba: WordCount

En primer lugar, lanzamos el script creaFichero.bash y copiamos el fichero de veryBig.txt al nuestro clúster.

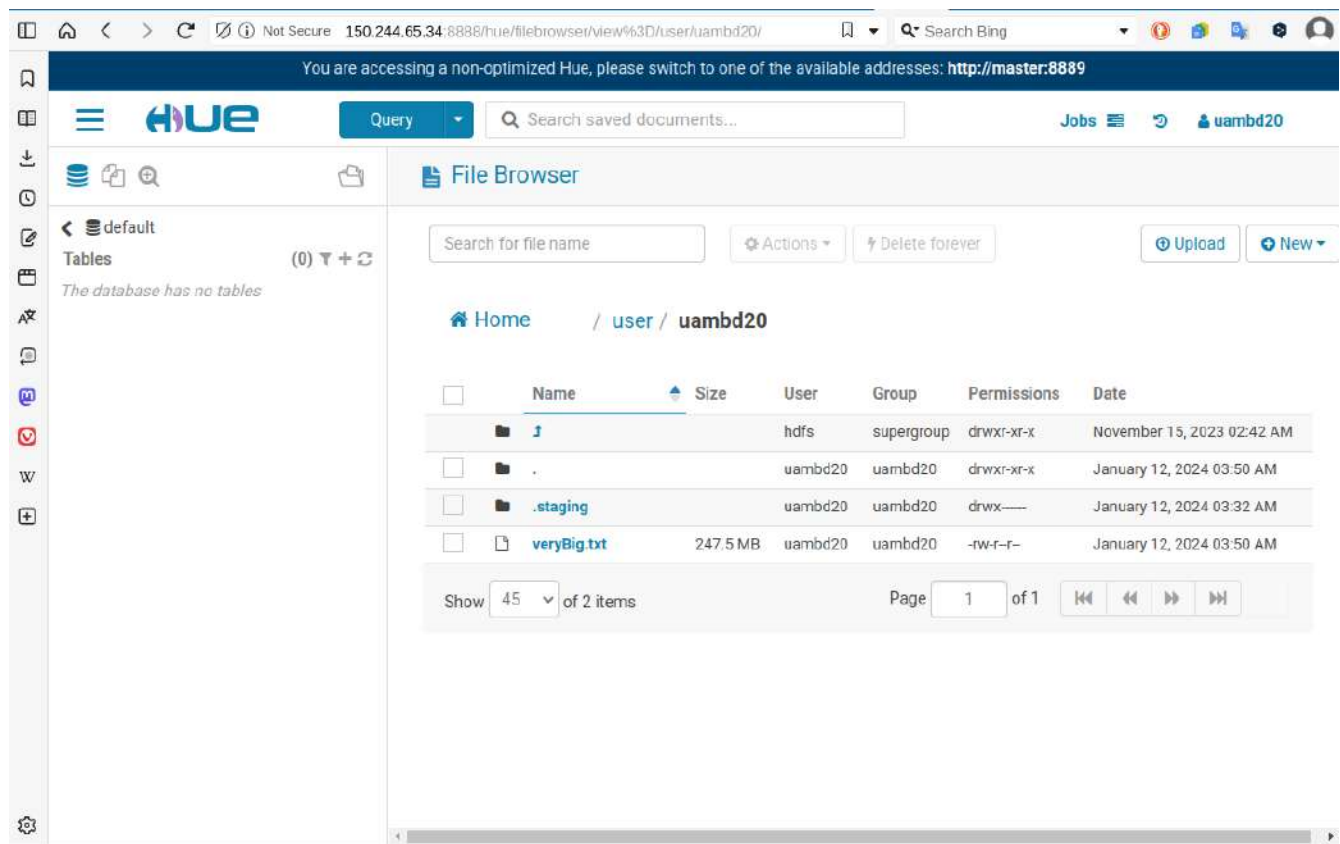
```
merve@onur-ideacenter: ~/Desktop/Infraestructura para big data/ficheros
File Edit View Search Terminal Help
merve@onur-ideacenter:~/Desktop/Infraestructura para big data/ficheros$ ./creaFichero.bash 40
merve@onur-ideacenter:~/Desktop/Infraestructura para big data/ficheros$ scp
scp scp-dbus-service
merve@onur-ideacenter:~/Desktop/Infraestructura para big data/ficheros$ scp veryBig.txt uamdb20@150.244.65.34:
uamdb20@150.244.65.34's password:
veryBig.txt 100% 248MB 10.3MB/s 00:23
merve@onur-ideacenter:~/Desktop/Infraestructura para big data/ficheros$
```

```
uamdb20@master:~
File Edit View Search Terminal Help
merve@onur-ideacenter:~/Desktop/Infraestructura para big data/ficheros$ ssh uamdb20@150.244.65.34
The authenticity of host '150.244.65.34 (150.244.65.34)' can't be established.
ED25519 key fingerprint is SHA256:UcBQSBjJjCvj34M30bf+Z2oPGHh0r+HASNLKhjPdBb24.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '150.244.65.34' (ED25519) to the list of known hosts.
uamdb20@150.244.65.34's password:
[uamdb20@master ~]$ ls
veryBig.txt
[uamdb20@master ~]$ pwd
/home/uamdb20
```

Subimos su contenido al sistema de ficheros distribuido (HDFS):

```
uamdb20@master:~
File Edit View Search Terminal Help
[uamdb20@master ~]$ hdfs dfs -put veryBig.txt
[uamdb20@master ~]$ hdfs dfs -ls
Found 2 items
drwx----- - uamdb20 uamdb20 0 2024-01-12 12:32 .staging
-rw-r--r-- 3 uamdb20 uamdb20 259546640 2024-01-12 12:50 veryBig.txt
[uamdb20@master ~]$
```

<http://150.244.65.34:8888> (Acceso a la consola de HUE)



Ejecutamos la aplicación WordCount sobre nuestro fichero de ejemplo:

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-examples.jar  
wordcount veryBig.txt salida-veryBig/  
24/01/12 12:55:06 INFO client.RMProxy: Connecting to ResourceManager at master/10.10.1.10:8032  
24/01/12 12:55:07 INFO input.FileInputFormat: Total input paths to process : 1  
24/01/12 12:55:07 INFO mapreduce.JobSubmitter: number of splits:2  
24/01/12 12:55:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917107_0012  
24/01/12 12:55:07 INFO impl.YarnClientImpl: Submitted application application_1704962917107_0012  
24/01/12 12:55:07 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_17  
04962917107_0012/  
24/01/12 12:55:07 INFO mapreduce.Job: Running job: job_1704962917107_0012  
24/01/12 12:55:12 INFO mapreduce.Job: Job job_1704962917107_0012 running in uber mode : false  
24/01/12 12:55:12 INFO mapreduce.Job: map 0% reduce 0%  
24/01/12 12:55:29 INFO mapreduce.Job: map 30% reduce 0%  
24/01/12 12:55:35 INFO mapreduce.Job: map 50% reduce 0%  
24/01/12 12:55:41 INFO mapreduce.Job: map 56% reduce 0%  
24/01/12 12:55:47 INFO mapreduce.Job: map 60% reduce 0%  
24/01/12 12:55:51 INFO mapreduce.Job: map 77% reduce 0%  
24/01/12 12:55:53 INFO mapreduce.Job: map 83% reduce 0%  
24/01/12 12:55:57 INFO mapreduce.Job: map 100% reduce 0%  
24/01/12 12:56:01 INFO mapreduce.Job: map 100% reduce 25%  
24/01/12 12:56:03 INFO mapreduce.Job: map 100% reduce 100%  
24/01/12 12:56:03 INFO mapreduce.Job: Job job_1704962917107_0012 completed successfully
```

Comprobamos la salida generada

Not Secure 150.244.65.34 8888/hue/filebrowser/view%3Duser/uamdb20/ Search Bing

You are accessing a non-optimized Hue, please switch to one of the available addresses: <http://master:8889>

Query Search saved documents... Jobs uamdb20

File Browser

Search for file name Actions Delete forever Upload New

default (0) Tables
The database has no tables

Home / user / uamdb20

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		hdfs	supergroup	drwxr-xr-x	November 15, 2023 02:42 AM
<input type="checkbox"/>	.		uamdb20	uamdb20	drwxr-xr-x	January 12, 2024 03:55 AM
<input type="checkbox"/>	.staging		uamdb20	uamdb20	drwx---	January 12, 2024 03:56 AM
<input type="checkbox"/>	salida-veryBig		uamdb20	uamdb20	drwxr-xr-x	January 12, 2024 03:56 AM
<input type="checkbox"/>	veryBig.txt	247.5 MB	uamdb20	uamdb20	-rw-r--	January 12, 2024 03:50 AM

Show 45 of 3 items Page 1 of 1

```
uamdb20@master:~  
File Edit View Search Terminal Help  
[uamdb20@master ~]$ hdfs dfs -ls salida-veryBig  
Found 5 items  
-rw-r--r--  3 uamdb20 uamdb20      0 2024-01-12 12:56 salida-veryBig/_SUCCESS  
-rw-r--r--  3 uamdb20 uamdb20 262635 2024-01-12 12:56 salida-veryBig/part-r-00000  
-rw-r--r--  3 uamdb20 uamdb20 259300 2024-01-12 12:56 salida-veryBig/part-r-00001  
-rw-r--r--  3 uamdb20 uamdb20 260171 2024-01-12 12:56 salida-veryBig/part-r-00002  
-rw-r--r--  3 uamdb20 uamdb20 256753 2024-01-12 12:55 salida-veryBig/part-r-00003  
[uamdb20@master ~]$
```

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

Hadoop es un framework para distribuir procesos a muchos ordenadores (y a muchas CPU). Cuando ejecutamos un trabajo MapReduce en Hadoop, la salida se divide en varias partes porque cada reducer en un trabajo MapReduce crea una salida. Así que si tenemos 4 archivos (en mi caso 4 partes) básicamente tenemos 4 reducers, si tenemos 24 archivos básicamente tenemos 24 reducers.

En otro lado, el fichero de salida esté partido en varias partes porque los ficheros en HDFS se almacenan de forma distribuida como partición en bloques, replicación. La razón por la que el archivo de salida se divide en varias partes en el cluster HUE de Hadoop es porque Hadoop está diseñado para trabajar con grandes conjuntos de datos que son demasiado grandes para caber en una sola máquina. Al dividir los datos en varias partes más pequeños, Hadoop puede distribuir el procesamiento entre varias máquinas, por eso permite procesar grandes conjuntos de datos mucho más rápidamente de lo que sería posible en una sola máquina.

¿Qué aplicaciones de ejemplo existen disponibles?

¿Qué aplicaciones soporta el jar con los ejemplos de nuestra distribución de Hadoop?

Ejecutamos siguiente código:

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-examples.jar  
An example program must be given as the first argument.  
Valid program names are:  
  aggregatewordcount: An Aggregate based map/reduce program that counts the words in the input files.  
  aggregatewordhist: An Aggregate based map/reduce program that computes the histogram of the words in  
the input files.  
  bbp: A map/reduce program that uses Bailey-Borwein-Plouffe to compute exact digits of Pi.  
  dbcount: An example job that count the pageview counts from a database.  
  distbbp: A map/reduce program that uses a BBP-type formula to compute exact bits of Pi.  
  grep: A map/reduce program that counts the matches of a regex in the input.  
  join: A job that effects a join over sorted, equally partitioned datasets  
  multifilewc: A job that counts words from several files.  
  pentomino: A map/reduce tile laying program to find solutions to pentomino problems.  
  pi: A map/reduce program that estimates Pi using a quasi-Monte Carlo method.  
  randomtextwriter: A map/reduce program that writes 10GB of random textual data per node.  
  randomwriter: A map/reduce program that writes 10GB of random data per node.  
  secondarysort: An example defining a secondary sort to the reduce.  
  sort: A map/reduce program that sorts the data written by the random writer.  
  sudoku: A sudoku solver.  
  teragen: Generate data for the terasort  
  terasort: Run the terasort  
  teravalidate: Checking results of terasort  
  wordcount: A map/reduce program that counts the words in the input files.  
  wordmean: A map/reduce program that counts the average length of the words in the input files.  
  wordmedian: A map/reduce program that counts the median length of the words in the input files.  
  wordstandarddeviation: A map/reduce program that counts the standard deviation of the length of the  
words in the input files.  
[uambd20@master ~]$
```

Vemos que las aplicaciones soporta el jar aggregatewordcount, aggregatewordhist, grep, join, teragen, terasort, teravalidate y wordcount etc.

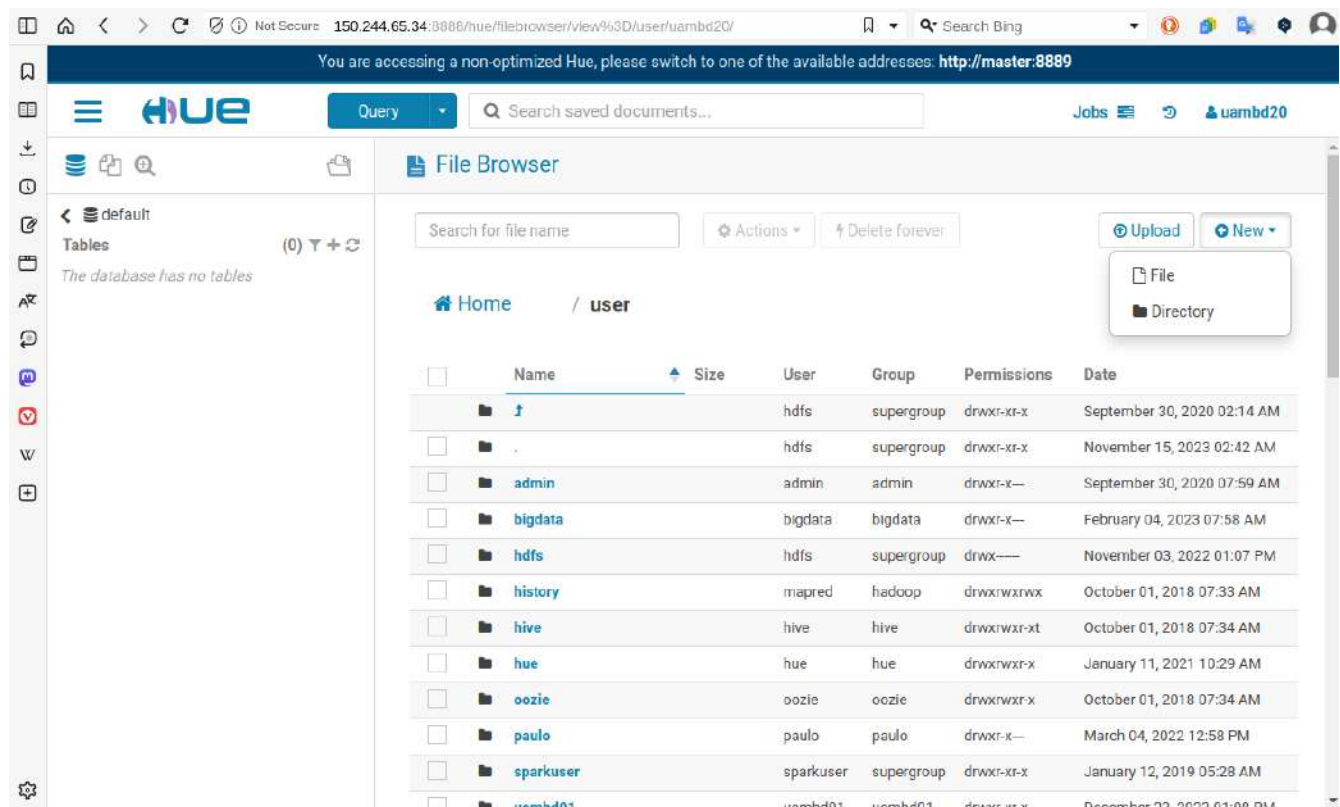
Para más información sobre los parámetros de E/S de una aplicación,

Ejecutamos siguiente código:

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-examples.jar wordcount  
Usage: wordcount <in> [<in>...] <out>  
[uambd20@master ~]$ hdfs dfs -ls /  
Found 4 items  
drwxr-xr-x - hbase hbase 0 2021-06-10 10:24 /hbase  
drwxr-xr-x - hdfs supergroup 0 2021-11-10 21:20 /system  
drwxrwxrwt - hdfs supergroup 0 2021-12-04 12:43 /tmp  
drwxr-xr-x - hdfs supergroup 0 2023-11-15 11:42 /user  
[uambd20@master ~]$ hdfs dfs -mkdir myTestDir  
[uambd20@master ~]$ hdfs dfs -ls  
Found 4 items  
drwx----- uambd20 uambd20 0 2024-01-12 12:56 .staging  
drwxr-xr-x uambd20 uambd20 0 2024-01-12 13:17 myTestDir  
drwxr-xr-x uambd20 uambd20 0 2024-01-12 12:56 salida-veryBig  
-rw-r--r-- 3 uambd20 uambd20 259546640 2024-01-12 12:50 veryBig.txt  
[uambd20@master ~]$
```

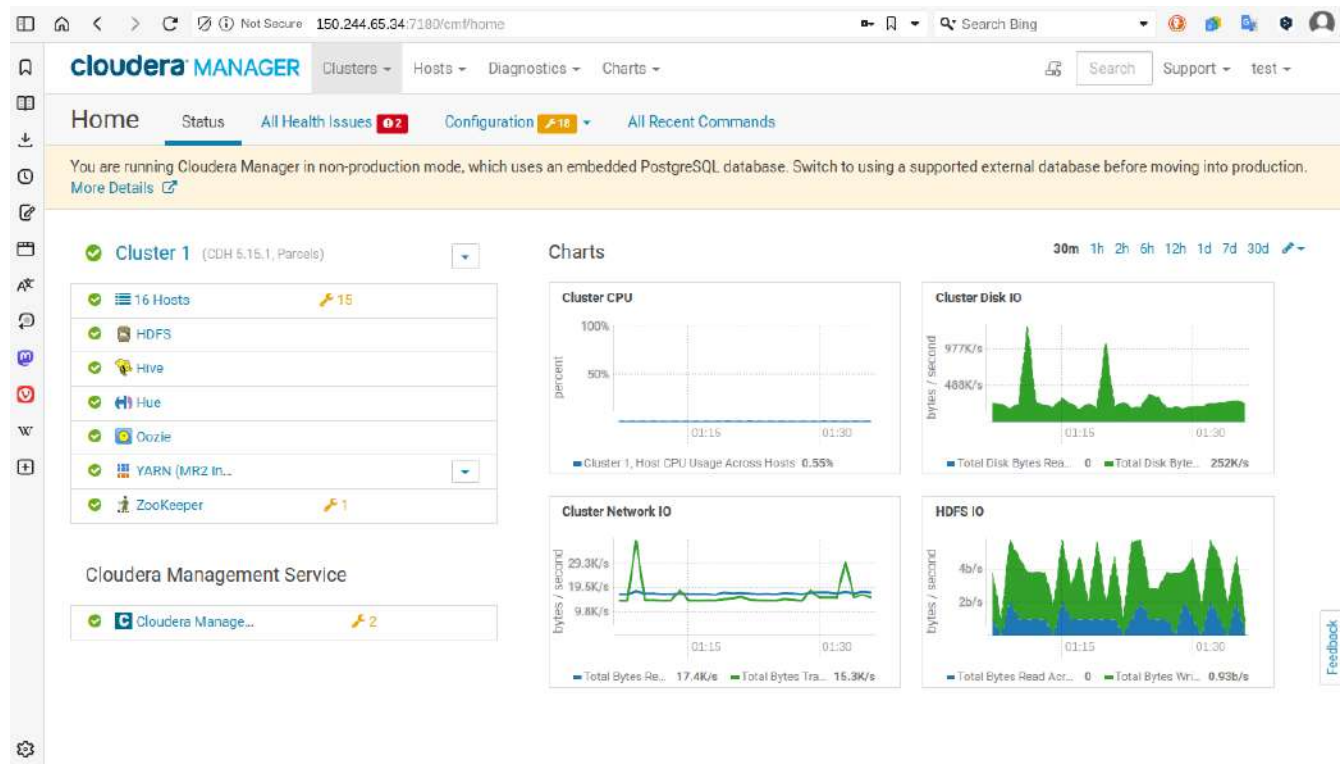
- Con el comando `hdfs dfs -ls`, podemos ver el contenido de los ficheros que estamos en hdfs.
- Con el comando `hdfs dfs -mkdir myTestDir` podemos crear fichero que se llama `myTestDir` en hdfs.

Además podemos crear los archivos mediante HUE pinchando el botón `New>Directory`



Acceso a la consola de Cloudera Manager de nuestro clúster

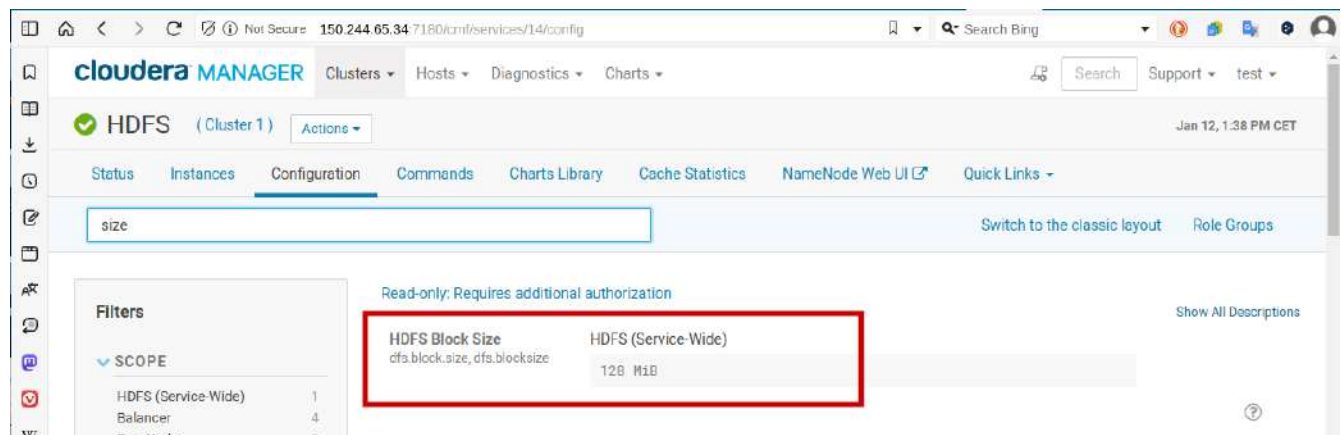
<http://150.244.65.34:7180>



Ejercicio: Utilizando la interfaz de web de Cloudera Manager, acceda a la siguiente información de configuración del clúster Hadoop:

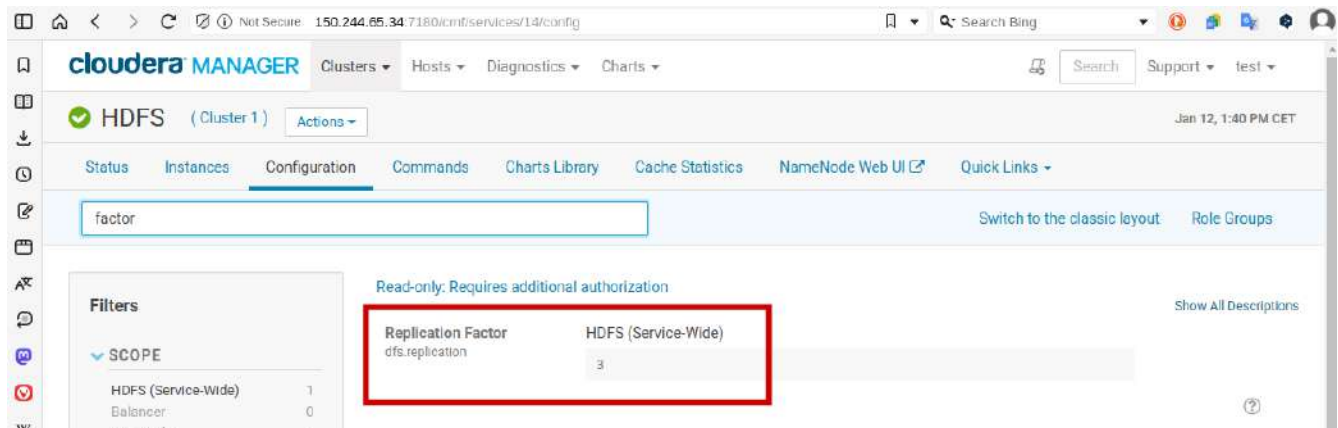
- ¿Cuál es el tamaño de bloque configurado para el HDFS del clúster?

El tamaño de bloque configurado para el HDFS del clúster es **128 MiB**



- ¿Cuál es el factor de replicación por defecto del HDFS?

El factor de replicación por defecto del HDFS es 3



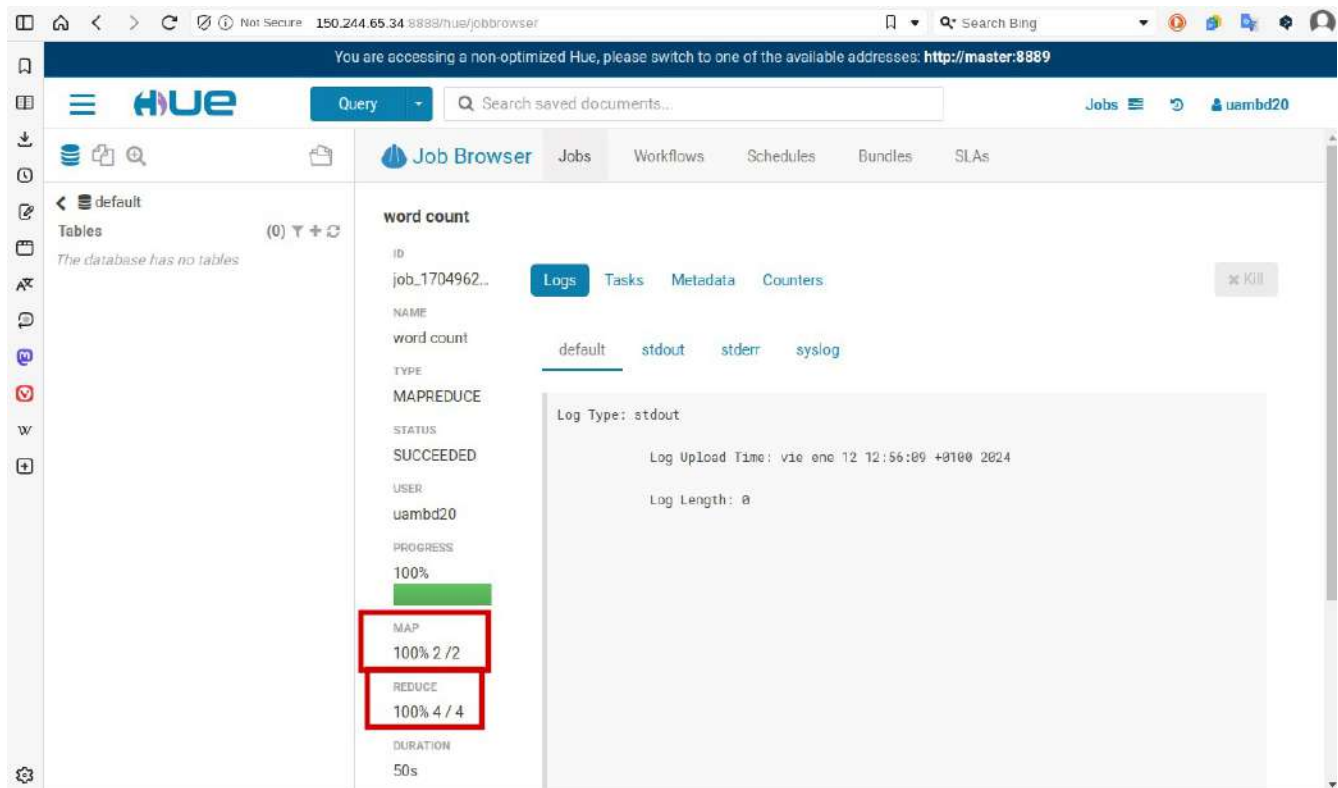
The screenshot shows the Cloudera Manager interface for the HDFS configuration. A search bar at the top contains the word 'factor'. On the left, under 'Filters', the 'SCOPE' is set to 'HDFS (Service-Wide)'. The main content area shows the 'Replication Factor' configuration, which is highlighted with a red box. It indicates a value of 3 for the HDFS (Service-Wide) scope. A note above the box states 'Read-only: Requires additional authorization'.

- ¿Cuál es el número de tareas MapReduce que se lanzarán por defecto al crear un nuevo trabajo? ¿Y para una tarea lanzada desde Hive?

En HUE

El número de tareas Map es 2

El número de tareas Reduce es 4



The screenshot shows the Hue Job Browser interface. A job named 'word count' is selected, and its details are displayed. The job is in a 'SUCCEEDED' state. The 'MAP' task is shown as '100% 2 / 2' and the 'REDUCE' task is shown as '100% 4 / 4', both highlighted with red boxes. The job duration is 50s. The interface also shows tabs for 'Logs', 'Tasks', 'Metadata', and 'Counters'.

En Cloudera Manager

El número de tareas Reduce es -1

The screenshot shows the Cloudera Manager interface for the Hive configuration of Cluster 1. The search filter 'map' is entered in the top search bar. The left sidebar shows filters for SCOPE and CATEGORY. The main content area displays configuration items related to MapReduce and Spark services.

Service	Configuration Item	Value
MapReduce Service	Hive (Service-Wide)	YARN (MR2 Included)
	Spark On YARN Service	Hive (Service-Wide)
Hive Reduce Tasks	mapred.reduce.tasks	-1
	HiveServer2 Enable	HiveServer2 Default Group

The screenshot shows the Cloudera Manager interface for the Hive configuration of Cluster 1. The search filter 'reducers' is entered in the top search bar. The left sidebar shows filters for SCOPE and CATEGORY. The main content area displays configuration items related to Hive reducers, with three items highlighted by red boxes.

Service	Configuration Item	Value
Hive Reduce Tasks	mapred.reduce.tasks	-1
	Hive Bytes Per Reducer	Hive (Service-Wide)
Hive Max Reducers	hive.exec.reducers.bytes.per.reducer	64 M1B
	hive.max.reducers	1099
Minimum Reducers for ReduceDeDuplication Optimization	hive.optimize.reducededuplication.min.reducer	4

Utilizando HUE para gestionar nuestra actividad en el clúster

<http://150.244.65.34:8888>

Podemos acceder al sitio de esta manera como lo hacíamos antes

El explorador de archivos de HUE

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

Cargamos el fichero de fichero2.tgz a user/uambd20 en HDFS mediante el interfaz de HUE.

The screenshot shows the HUE File Browser interface. The top navigation bar includes the HUE logo, a 'Query' dropdown, and a search bar. The left sidebar shows the 'default' database with no tables. The main area is titled 'File Browser' and shows the directory path 'Home / user / uambd20'. A search bar for file names is present, along with buttons for 'Actions', 'Delete forever', 'Extract', 'Upload', and 'New'. A table lists the files in the directory:

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	Home		hdfs	supergroup	drwxr-xr-x	November 15, 2023 02:42 AM
<input type="checkbox"/>	.		uambd20	uambd20	drwxr-xr-x	January 16, 2024 01:30 PM
<input type="checkbox"/>	.staging		uambd20	uambd20	drwx---	January 12, 2024 03:56 AM
<input checked="" type="checkbox"/>	ficheros2.tgz	2.3 MB	uambd20	uambd20	-rwxr-xr-x	January 16, 2024 01:23 PM
<input type="checkbox"/>	myTestDir		uambd20	uambd20	drwxr-xr-x	January 12, 2024 04:17 AM
<input type="checkbox"/>	salida-veryBig		uambd20	uambd20	drwxr-xr-x	January 12, 2024 03:56 AM
<input type="checkbox"/>	veryBig.txt	247.5 MB	uambd20	uambd20	-rwxr-xr-x	January 12, 2024 03:50 AM

At the bottom, there is a pagination bar showing 'Show 45 of 5 items' and 'Page 1 of 1'.

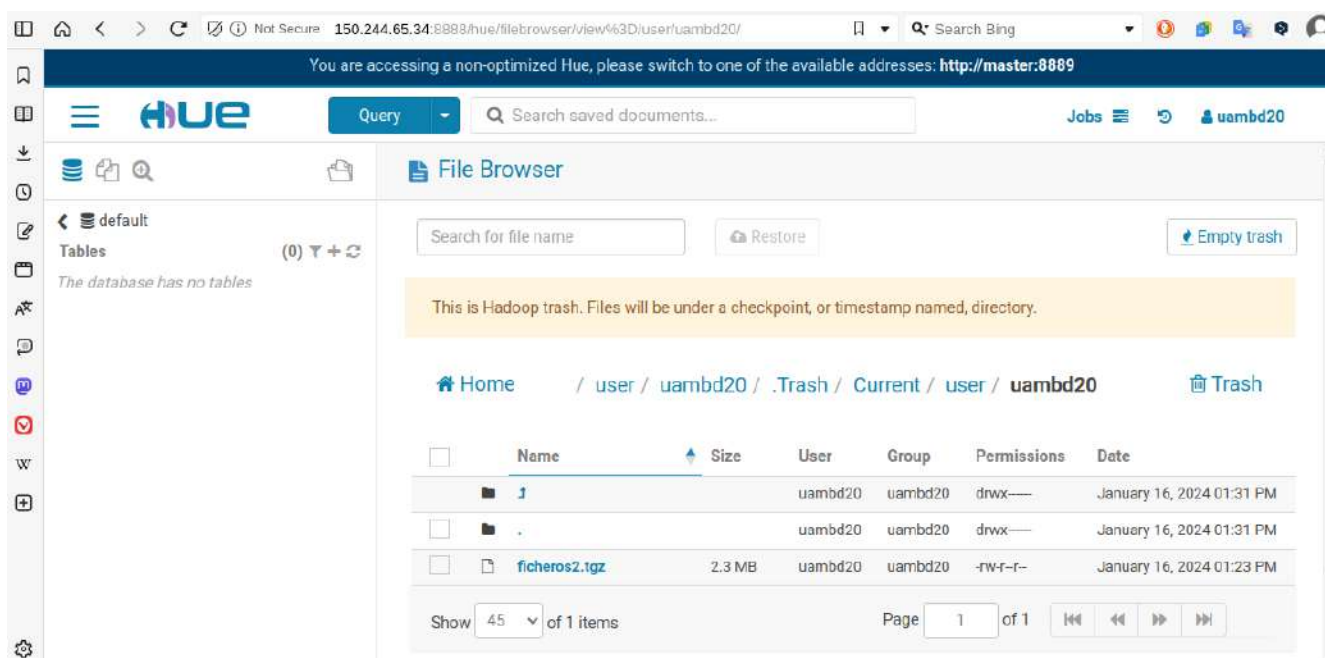
Lo copiamos cambiando el nombre y después eliminamos la copia original. Veremos que solo tenemos el fichero de ficheros.tgz

Lo copiamos cambiando el nombre a “ficheros.tgz” y después eliminamos la copia original. Con el código **hdfs dfs -ls** veremos que sólo tenemos el fichero “ficheros.tgz”.

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hdfs dfs -cp ficheros2.tgz ficheros.tgz  
[uambd20@master ~]$ hdfs dfs -ls  
Found 6 items  
drwx----- - uambd20 uambd20          0 2024-01-12 12:56 .staging  
-rw-r--r--  3 uambd20 uambd20    2385561 2024-01-16 22:30 ficheros.tgz  
-rw-r--r--  3 uambd20 uambd20    2385561 2024-01-16 22:23 ficheros2.tgz  
drwxr-xr-x  - uambd20 uambd20          0 2024-01-12 13:17 myTestDir  
drwxr-xr-x  - uambd20 uambd20          0 2024-01-12 12:56 salida-veryBig  
-rw-r--r--  3 uambd20 uambd20  259546640 2024-01-12 12:50 veryBig.txt  
[uambd20@master ~]$ hdfs dfs -rm ficheros2.tgz  
24/01/16 22:31:46 INFO fs.TrashPolicyDefault: Moved: 'hdfs://master:8020/user/uambd20/ficheros2.tgz' to trash at: hdfs://master:8020/user/uambd20/.Trash/Current/user/uambd20/ficheros2.tgz  
[uambd20@master ~]$ hdfs dfs -ls  
Found 6 items  
drwx----- - uambd20 uambd20          0 2024-01-16 22:31 .Trash  
drwx----- - uambd20 uambd20          0 2024-01-12 12:56 .staging  
-rw-r--r--  3 uambd20 uambd20    2385561 2024-01-16 22:30 ficheros.tgz  
drwxr-xr-x  - uambd20 uambd20          0 2024-01-12 13:17 myTestDir  
drwxr-xr-x  - uambd20 uambd20          0 2024-01-12 12:56 salida-veryBig  
-rw-r--r--  3 uambd20 uambd20  259546640 2024-01-12 12:50 veryBig.txt  
[uambd20@master ~]$
```

Podemos ver que el fichero que hemos borrado está en el directorio user/uamdb20/.Trash/Current/user/uamdb20/.

```
uamdb20@master:~  
File Edit View Search Terminal Help  
[uamdb20@master ~]$ hdfs dfs -ls .Trash/Current/user/uamdb20  
Found 1 items  
-rw-r--r--  3 uamdb20 uamdb20    2385561 2024-01-16 22:23 .Trash/Current/user/u  
amdb20/ficheros2.tgz  
[uamdb20@master ~]$
```



Navegador de trabajos en HUE

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

- Número de tareas Map lanzadas es **2**
- Número de tareas Reduce lanzadas **4**
- Duración de la ejecución de la tarea **50 segundos**
- Estado de terminación es **Succeeded (Ha tenido éxito)**

The screenshot shows the Hue web interface for a job named 'word count'. The job ID is 'job_170496291...'. The status is 'SUCCEEDED' (highlighted with a red box). The progress is 100% (highlighted with a green bar). The job was submitted on 01/12/24 at 03:55... The configuration table on the right lists various Hadoop and Yarn settings.

Name	Value
mapreduce.jobtracker.address	local
dfs.namenode.resource.check.interval	5000
hadoop.security.group.mapping.ldap.posix.attr.uid.name	uidNumber
mapreduce.jobhistory.client.thread-count	10
yarn.application.classpath	\$HADOOP_CLIENT_CONF_DIR,\$HADOOP_CONF_DIR,\$HADOOP_COMMON...
yarn.admin.acl	*
yarn.app.mapreduce.am.job.committer.cancel-timeout	60000
mapreduce.job.emit-timeline-data	false
dfs.journalnode.rpc-address	0.0.0.0:8485
dfs.disk.balancer.max.disk.throughputInMBperSec	10
mapred.mapper.new-api	true

Ejercicio: Utilice el script “creaFichero.bash” que se provee para crear un fichero de texto más grande:

```
./creaFichero.bash 80
```

Esta ejecución creará un fichero de texto “veryBig.txt”.

Ahora, ejecute el ejemplo WordCount de nuevo, y obtenga utilizando la interfaz de HUE los datos de configuración de la nueva tarea que se pedían anteriormente. ¿Ha cambiado algo? ¿A qué se debe?

- Creamos un nuevo archivo verBig.txt usando **./creaFichero.bash** y cambiamos el nombre a veryBig2.txt. Luego lo subimos a user/uamdb20 en HDFS.

```
merve@onur-ideacenter: ~/Desktop/Infraestructura para big data/ficheros
File Edit View Search Terminal Help
merve@onur-ideacenter:~/Desktop/Infraestructura para big data/ficheros$ scp veryBig2.txt uamdb20@150.244.65.34:
uamdb20@150.244.65.34's password:
veryBig2.txt 100% 495MB 10.6MB/s 00:46
merve@onur-ideacenter:~/Desktop/Infraestructura para big data/ficheros$
```

```
uamdb20@master:~
File Edit View Search Terminal Help
[uamdb20@master ~]$ hdfs dfs -put veryBig2.txt
[uamdb20@master ~]$ hdfs dfs -ls
Found 7 items
drwx----- - uamdb20 uamdb20 0 2024-01-16 23:00 .Trash
drwx----- - uamdb20 uamdb20 0 2024-01-12 12:56 .staging
-rw-r--r-- 3 uamdb20 uamdb20 2385561 2024-01-16 22:30 ficheros.tgz
drwxr-xr-x - uamdb20 uamdb20 0 2024-01-12 13:17 myTestDir
drwxr-xr-x - uamdb20 uamdb20 0 2024-01-12 12:56 salida-veryBig
-rw-r--r-- 3 uamdb20 uamdb20 259546640 2024-01-12 12:50 veryBig.txt
-rw-r--r-- 3 uamdb20 uamdb20 519093280 2024-01-16 23:24 veryBig2.txt
[uamdb20@master ~]$
```

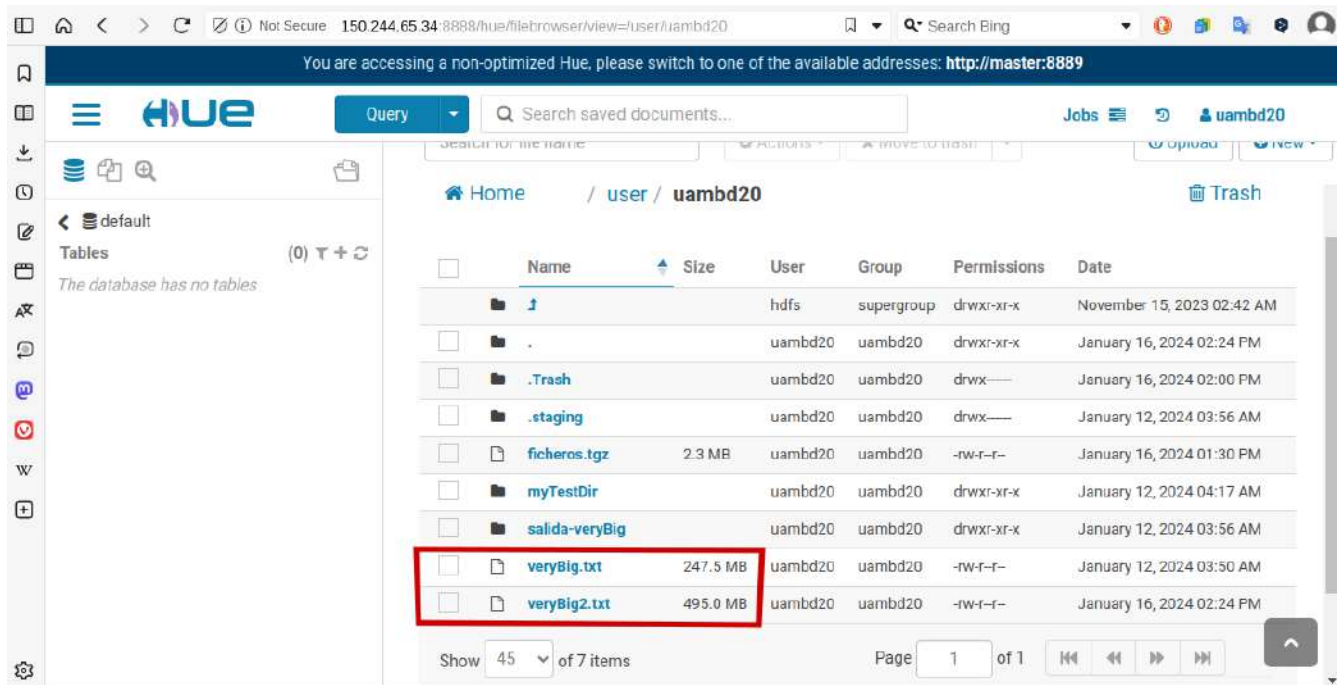
Ejecutamos **hdfs dfs -ls** para ver que el tamaño del archivo veryBig2.txt es mayor.

La razón por la que el archivo creado por **./creaFichero.bash 80** es mayor que el archivo creado por **./creaFichero.bash 40** es porque el script creaFichero.bash está creando un archivo con 80 líneas de texto, y cada línea de texto tiene asociada una cierta cantidad de datos. Cuantas más líneas de texto haya, más grande será el archivo.

Este código iterará 80 veces, y en cada iteración, creará una nueva línea de texto y la escribirá en el archivo veryBig.txt.

El archivo creado por **./creaFichero.bash 40** sólo tendrá 40 líneas de texto, por lo que será más pequeño que el archivo creado por **./creaFichero.bash 80**.

También podemos ver los tamaños utilizando la interfaz de HUE.



Ejecutamos la aplicación WordCount sobre nuestro fichero de ejemplo

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-examples.jar wordcount veryBig2.txt salida-veryBig2/  
24/01/16 23:35:24 INFO client.RMPProxy: Connecting to ResourceManager at master/10.10.1.10:8032  
24/01/16 23:35:25 INFO input.FileInputFormat: Total input paths to process : 1  
24/01/16 23:35:25 INFO mapreduce.JobSubmitter: number of splits:4  
24/01/16 23:35:25 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917107_0088  
24/01/16 23:35:26 INFO impl.YarnClientImpl: Submitted application application_1704962917107_0088  
24/01/16 23:35:26 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1704962917107_0088/  
24/01/16 23:35:26 INFO mapreduce.Job: Running job: job_1704962917107_0088  
24/01/16 23:35:31 INFO mapreduce.Job: Job job_1704962917107_0088 running in uber mode : false  
24/01/16 23:35:31 INFO mapreduce.Job: map 0% reduce 0%  
24/01/16 23:35:46 INFO mapreduce.Job: map 13% reduce 0%  
24/01/16 23:35:47 INFO mapreduce.Job: map 38% reduce 0%  
24/01/16 23:35:53 INFO mapreduce.Job: map 56% reduce 0%  
24/01/16 23:35:58 INFO mapreduce.Job: map 59% reduce 0%  
24/01/16 23:35:59 INFO mapreduce.Job: map 69% reduce 0%  
24/01/16 23:36:01 INFO mapreduce.Job: map 77% reduce 0%  
24/01/16 23:36:05 INFO mapreduce.Job: map 83% reduce 0%  
24/01/16 23:36:06 INFO mapreduce.Job: map 100% reduce 0%  
24/01/16 23:36:11 INFO mapreduce.Job: map 100% reduce 25%  
24/01/16 23:36:12 INFO mapreduce.Job: map 100% reduce 100%  
24/01/16 23:36:12 INFO mapreduce.Job: Job job_1704962917107_0088 completed successfully
```

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hdfs dfs -ls salida-veryBig2  
Found 5 items  
-rw-r--r--  3 uambd20 uambd20      0 2024-01-16 23:36 salida-veryBig2/_SUCCESS  
-rw-r--r--  3 uambd20 uambd20 266417 2024-01-16 23:36 salida-veryBig2/part-r-00000  
-rw-r--r--  3 uambd20 uambd20 262980 2024-01-16 23:36 salida-veryBig2/part-r-00001  
-rw-r--r--  3 uambd20 uambd20 263925 2024-01-16 23:36 salida-veryBig2/part-r-00002  
-rw-r--r--  3 uambd20 uambd20 260513 2024-01-16 23:36 salida-veryBig2/part-r-00003  
[uambd20@master ~]$
```

Consultamos los siguientes datos de configuración de la tarea “WordCount” de ejemplo lanzada ahora:

- Número de tareas Map lanzadas es **4**
- Número de tareas Reduce lanzadas **4**
- Duración de la ejecución de la tarea **41 segundos**
- Estado de terminación es **Succeeded (Ha tenido éxito)**

The screenshot shows the Hue web interface for a Hadoop job named "word count". The job ID is "job_17049629...". The status is "SUCCEEDED". The progress bar is at 100%. The map task is "100% 4 / 4" and the reduce task is "100% 4 / 4". The duration is "41s". The submit time is "01/16/24 14:3...". The right panel shows a list of configuration parameters and their values.

Name	Value
mapreduce.jobtracker.address	local
dfs.namenode.resource.check.interval	5000
hadoop.security.group.mapping.ldap.post.attr.uid.name	uidNumber
mapreduce.jobhistory.client.thread-count	10
yarn.application.classpath	\$HADOOP_CLIENT_CONF_DIR:\$HADOOP_CONF_DIR:\$HADOOP_COMMON...
yarn.admin.acl	*
yarn.app.mapreduce.am.job.committer.cancel-timeout	60000
mapreduce.job.emit-timeline-data	false
dfs.journalnode.rpc-address	0.0.0.0:8485
dfs.disk.balancer.max.disk.throughputInMBperSec	10

¿Ha cambiado algo? ¿A qué se debe?

El número de la tarea Map ha cambiado, pero el número de la tarea Reduce sigue siendo el mismo. Antes teníamos 2 tareas Map y 4 tareas Reduce, y ahora tenemos 4 tareas Map y también 4 tareas Reduce. El número de tareas Reduce no ha cambiado porque su cantidad se establece por el usuario. Si el archivo de entrada es más grande, Hadoop puede utilizar más tareas Map para procesar los datos en paralelo. Esto puede ayudar a acelerar el tiempo de procesamiento, pero también puede aumentar la cantidad de datos que necesitan ser procesados por los Reducers.

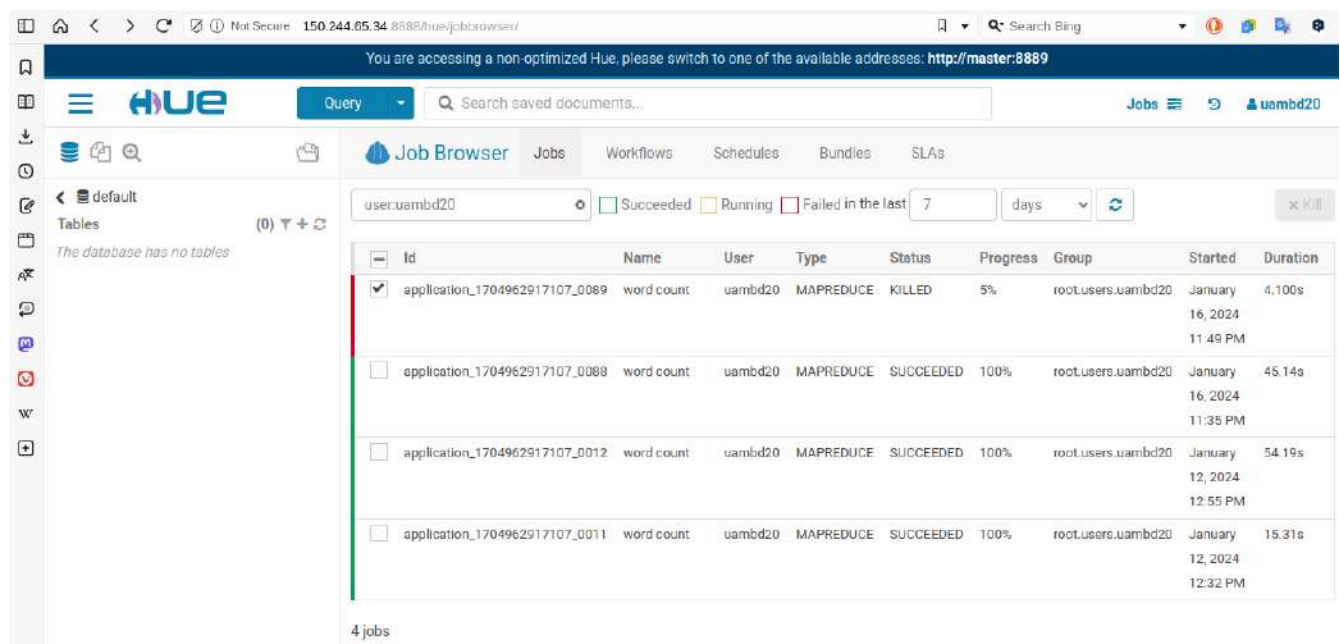
Por defecto, el número de tareas Reduce se establece en -1, lo que significa que el número de Reducers se determina automáticamente en función del tamaño de los datos de entrada. Si deseamos establecer el número de Reducers manualmente, podemos cambiar este valor al número deseado.

Ejercicio: 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?

Ejecutamos la aplicación WordCount sobre nuestro fichero de ejemplo otra vez.

Accedemos durante la ejecución al Job Browser de HUE

Utilizando la interfaz gráfica, **terminamos** (matamos) la tarea



The screenshot shows the Hue Job Browser interface. At the top, a message states: "You are accessing a non-optimized Hue, please switch to one of the available addresses: <http://master:8889>". The interface includes a sidebar with navigation icons and a main panel displaying a table of jobs. The table has columns: Id, Name, User, Type, Status, Progress, Group, Started, and Duration. There are four jobs listed, all of type 'MAPREDUCE' and user 'uambd20'. The first job is 'KILLED' (5% progress), while the others are 'SUCCEEDED' (100% progress). A 'Kill' button is visible in the top right of the job list area.

Id	Name	User	Type	Status	Progress	Group	Started	Duration
<input checked="" type="checkbox"/> application_1704962917107_0089	word count	uambd20	MAPREDUCE	KILLED	5%	root.users.uambd20	January 16, 2024 11:49 PM	4:100s
<input type="checkbox"/> application_1704962917107_0088	word count	uambd20	MAPREDUCE	SUCCEEDED	100%	root.users.uambd20	January 16, 2024 11:35 PM	45.14s
<input type="checkbox"/> application_1704962917107_0012	word count	uambd20	MAPREDUCE	SUCCEEDED	100%	root.users.uambd20	January 12, 2024 12:55 PM	54.19s
<input type="checkbox"/> application_1704962917107_0011	word count	uambd20	MAPREDUCE	SUCCEEDED	100%	root.users.uambd20	January 12, 2024 12:32 PM	15.31s

4 jobs

Obtenemos los mensajes en la consola después de matar la tarea. Dice que

Job job_1704962917107_0089 failed with state KILLED due to: Application killed by user.

“Aplicación muerta por el usuario.”

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-examples.jar wordcount veryBig2.txt salida-veryBig3/  
24/01/16 23:49:29 INFO client.RMPProxy: Connecting to ResourceManager at master/10.10.1.10:8032  
24/01/16 23:49:30 INFO input.FileInputFormat: Total input paths to process : 1  
24/01/16 23:49:30 INFO mapreduce.JobSubmitter: number of splits:4  
24/01/16 23:49:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917107_0089  
24/01/16 23:49:31 INFO impl.YarnClientImpl: Submitted application application_1704962917107_0089  
24/01/16 23:49:31 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1704962917107_0089/  
24/01/16 23:49:31 INFO mapreduce.Job: Running job: job_1704962917107_0089  
24/01/16 23:49:37 INFO mapreduce.Job: Job job_1704962917107_0089 running in uber mode : false  
24/01/16 23:49:37 INFO mapreduce.Job: map 0% reduce 0%  
24/01/16 23:49:52 INFO mapreduce.Job: map 14% reduce 0%  
24/01/16 23:49:53 INFO mapreduce.Job: map 38% reduce 0%  
24/01/16 23:49:55 INFO mapreduce.Job: map 0% reduce 0%  
24/01/16 23:49:55 INFO mapreduce.Job: Job job_1704962917107_0089 failed with state KILLED due to: Application killed by user.  
24/01/16 23:49:55 INFO mapreduce.Job: Counters: 0  
[uambd20@master ~]$
```

Consultamos los siguientes datos de configuración de la tarea “WordCount” de ejemplo lanzada ahora:

- Número de tareas Map lanzadas es **0**
- Número de tareas Reduce lanzadas **0**
- Duración de la ejecución de la tarea **22 segundos**
- Estado de terminación es **Killed (Matado)**

Y ahora no se muestra ningún avance, y observamos que los números de las tareas Map y Reduce son 0, y también el estado de terminación es "killed" porque hemos terminado la ejecución.

The screenshot displays the Hue Job Browser interface for a job named 'word count'. The job is in a 'KILLED' state, as indicated by the red box around the 'STATUS' field. The 'PROGRESS' bar is at 100%, and the 'DURATION' is 22s. The 'MAP' and 'REDUCE' counts are both 0/0. The 'LOGS' tab is selected, showing the message 'Application killed by user.'

Field	Value
ID	application_170...
NAME	word count
TYPE	MAPREDUCE
STATUS	KILLED
USER	uambd20
PROGRESS	100%
MAP	0% 0 / 0
REDUCE	0% 0 / 0
DURATION	22s
SUBMITTED	01/16/24 14:49...

Benchmarking de un clúster Hadoop

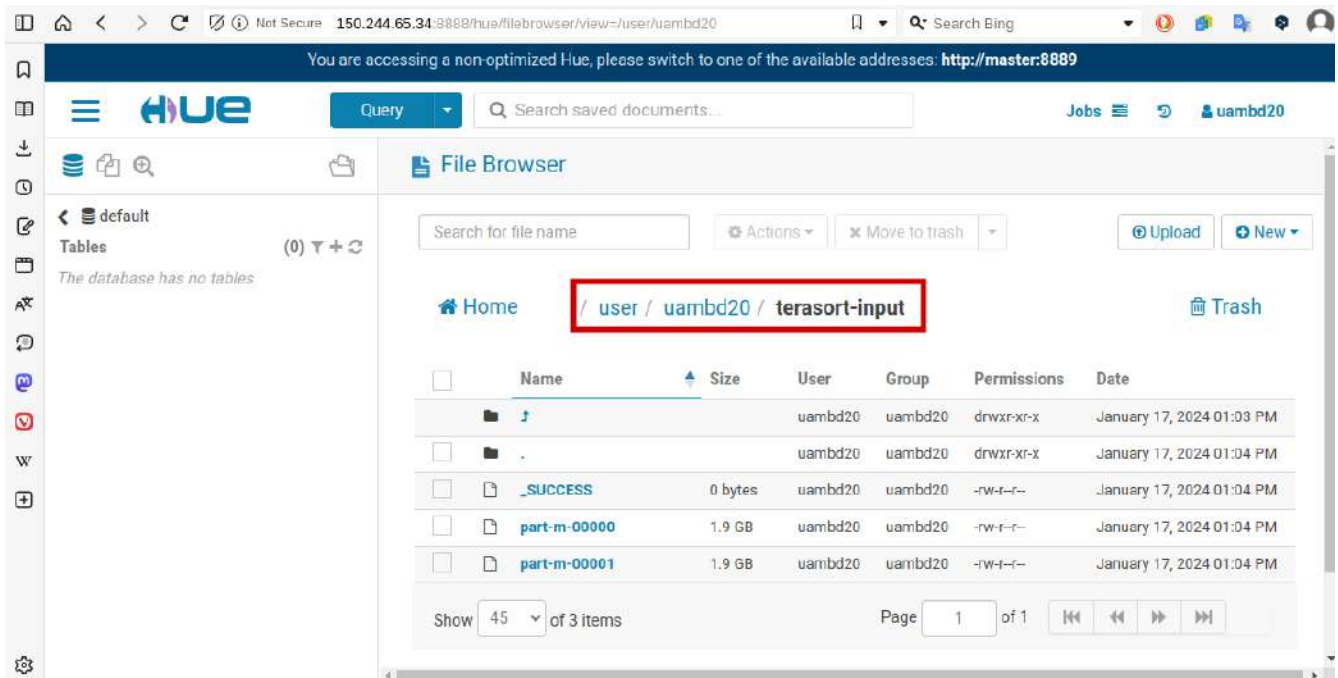
Para generar los datos, ejecutamos teragen:

```
uamdb20@master:~  
File Edit View Search Terminal Help  
[uamdb20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-examples.jar teragen 40000000 terasort-input  
24/01/17 22:03:54 INFO client.RMPProxy: Connecting to ResourceManager at master/10.10.1.10:8032  
24/01/17 22:03:55 INFO terasort.TeraGen: Generating 40000000 using 2  
24/01/17 22:03:55 INFO mapreduce.JobSubmitter: number of splits:2  
24/01/17 22:03:56 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917107_0113  
24/01/17 22:03:56 INFO impl.YarnClientImpl: Submitted application application_1704962917107_0113  
24/01/17 22:03:56 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1704962917107_0113/  
24/01/17 22:03:56 INFO mapreduce.Job: Running job: job_1704962917107_0113  
24/01/17 22:04:01 INFO mapreduce.Job: Job job_1704962917107_0113 running in uber mode : false  
24/01/17 22:04:01 INFO mapreduce.Job: map 0% reduce 0%  
24/01/17 22:04:17 INFO mapreduce.Job: map 34% reduce 0%  
24/01/17 22:04:23 INFO mapreduce.Job: map 52% reduce 0%  
24/01/17 22:04:29 INFO mapreduce.Job: map 69% reduce 0%  
24/01/17 22:04:35 INFO mapreduce.Job: map 87% reduce 0%  
24/01/17 22:04:38 INFO mapreduce.Job: map 91% reduce 0%  
24/01/17 22:04:39 INFO mapreduce.Job: map 100% reduce 0%  
24/01/17 22:04:39 INFO mapreduce.Job: Job job_1704962917107_0113 completed successfully  
24/01/17 22:04:40 INFO mapreduce.Job: Counters: 31
```

Ejercicio: 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?

Comprobamos que el fichero se ha generado correctamente:

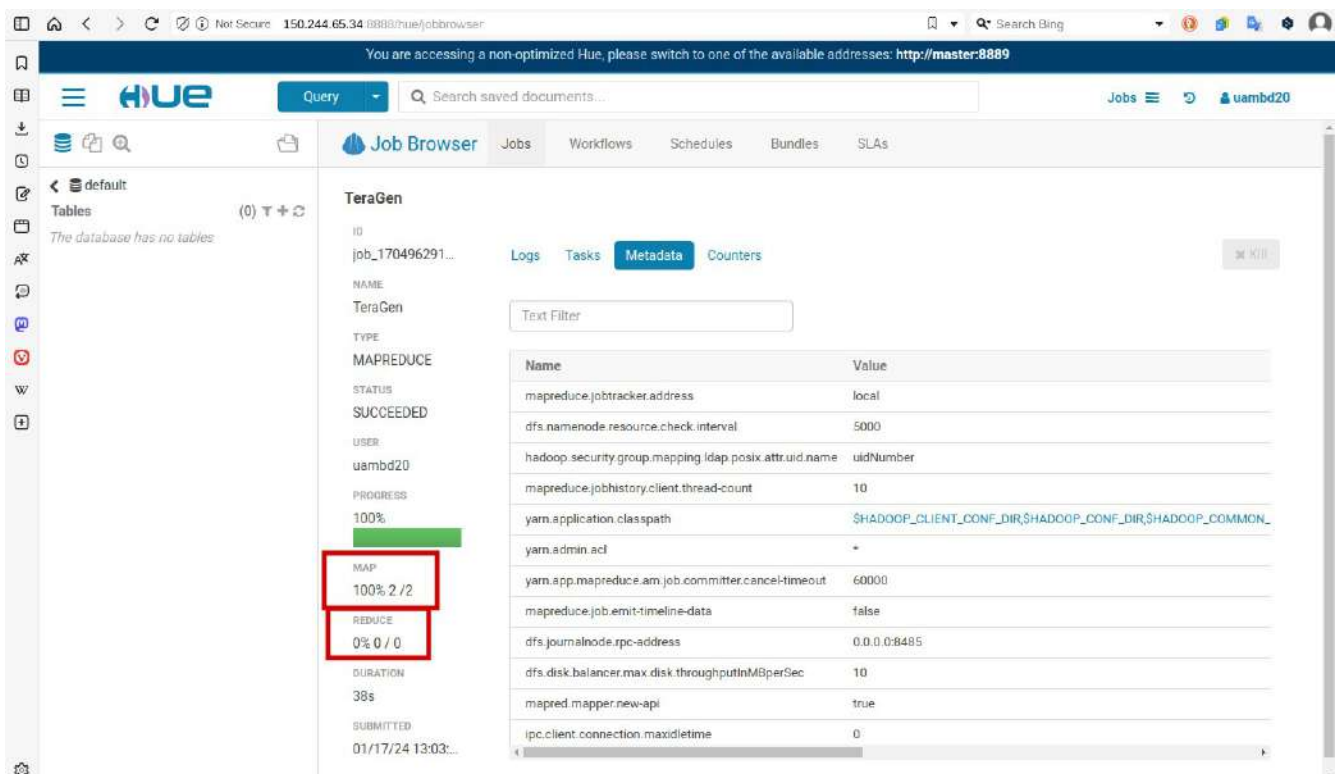
```
uamdb20@master:~  
File Edit View Search Terminal Help  
[uamdb20@master ~]$ hdfs dfs -ls -h terasort-input  
Found 3 items  
-rw-r--r-- 3 uamdb20 uamdb20 0 2024-01-17 22:04 terasort-input/_SUCCESS  
-rw-r--r-- 3 uamdb20 uamdb20 1.9 G 2024-01-17 22:04 terasort-input/part-m-00000  
-rw-r--r-- 3 uamdb20 uamdb20 1.9 G 2024-01-17 22:04 terasort-input/part-m-00001  
[uamdb20@master ~]$
```



¿Cuántas tareas Map y Reduce se han lanzado para crear nuestro fichero?

Número de tareas Map lanzadas es 2

Número de tareas Reduce lanzadas es 0



Ejecución de Terasort

Ejecutar nuestra aplicación de ordenación:

```
uamdb20@master:~  
File Edit View Search Terminal Help  
[uamdb20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop  
-examples.jar terasort terasort-input/ terasort-output  
24/01/17 22:26:29 INFO terasort.TeraSort: starting  
24/01/17 22:26:31 INFO input.FileInputFormat: Total input paths to process : 2  
Spent 256ms computing base-splits.  
Spent 3ms computing TeraScheduler splits.  
Computing input splits took 261ms  
Sampling 10 splits of 30  
Making 4 from 100000 sampled records  
Computing parititions took 748ms  
Spent 1012ms computing partitions.  
24/01/17 22:26:32 INFO client.RMPProxy: Connecting to ResourceManager at master/10.10.1.10  
:8032  
24/01/17 22:26:32 INFO mapreduce.JobSubmitter: number of splits:30  
24/01/17 22:26:33 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_17049629171  
07_0114  
24/01/17 22:26:33 INFO impl.YarnClientImpl: Submitted application application_17049629171  
07_0114  
24/01/17 22:26:33 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/  
application_1704962917107_0114/  
24/01/17 22:26:33 INFO mapreduce.Job: Running job: job_1704962917107_0114  
24/01/17 22:26:37 INFO mapreduce.Job: Job job_1704962917107_0114 running in uber mode : f  
alse  
24/01/17 22:26:37 INFO mapreduce.Job: map 0% reduce 0%  
24/01/17 22:26:45 INFO mapreduce.Job: map 3% reduce 0%  
24/01/17 22:26:46 INFO mapreduce.Job: map 30% reduce 0%  
24/01/17 22:26:49 INFO mapreduce.Job: map 43% reduce 0%  
24/01/17 22:26:50 INFO mapreduce.Job: map 67% reduce 0%  
24/01/17 22:26:51 INFO mapreduce.Job: map 90% reduce 0%  
24/01/17 22:26:52 INFO mapreduce.Job: map 100% reduce 0%  
24/01/17 22:27:08 INFO mapreduce.Job: map 100% reduce 73%  
24/01/17 22:27:13 INFO mapreduce.Job: map 100% reduce 83%  
24/01/17 22:27:14 INFO mapreduce.Job: map 100% reduce 93%  
24/01/17 22:27:20 INFO mapreduce.Job: map 100% reduce 97%  
24/01/17 22:27:25 INFO mapreduce.Job: map 100% reduce 98%  
24/01/17 22:27:26 INFO mapreduce.Job: map 100% reduce 99%  
24/01/17 22:27:27 INFO mapreduce.Job: map 100% reduce 100%
```


Ahora tenemos el directorio terasort-output una copia ordenada de los datos de entrada.

Comprobamos que el fichero se ha generado correctamente:

```
uamdb20@master:~  
File Edit View Search Terminal Help  
[uamdb20@master ~]$ hdfs dfs -ls -h terasort-output  
Found 6 items  
-rw-r--r-- 1 uamdb20 uamdb20 0 2024-01-17 22:27 terasort-output/_SUCCESS  
-rw-r--r-- 10 uamdb20 uamdb20 33 2024-01-17 22:26 terasort-output/_partition.lst  
-rw-r--r-- 1 uamdb20 uamdb20 956.4 M 2024-01-17 22:27 terasort-output/part-r-00000  
-rw-r--r-- 1 uamdb20 uamdb20 959.8 M 2024-01-17 22:27 terasort-output/part-r-00001  
-rw-r--r-- 1 uamdb20 uamdb20 946.1 M 2024-01-17 22:27 terasort-output/part-r-00002  
-rw-r--r-- 1 uamdb20 uamdb20 952.4 M 2024-01-17 22:27 terasort-output/part-r-00003  
[uamdb20@master ~]$
```

Pregunta: Compruebe el tamaño de los archivos de entrada y de salida de TeraSort. ¿Cómo están distribuidos los datos? ¿Sabrías explicar a qué se debe?

Comprobamos el tamaño de los archivos de entrada y de salida de TeraSort:

```
uamdb20@master:~  
File Edit View Search Terminal Help  
[uamdb20@master ~]$ hdfs dfs -du -h  
2.3 M 6.8 M .Trash  
547.6 K 1.6 M .staging  
2.3 M 6.8 M ficheros.tgz  
0 0 myTestDir  
1014.5 K 3.0 M salida-veryBig  
1.0 M 3.0 M salida-veryBig2  
0 0 salida-veryBig3  
3.7 G 11.2 G terasort-input  
3.7 G 3.7 G terasort-output  
247.5 M 742.6 M veryBig.txt  
495.0 M 1.5 G veryBig2.txt  
[uamdb20@master ~]$
```

El directorio terasort-input

- Ocupa 3.7 gigabytes en disco.
- El espacio total en disco consumido, incluidos sus subdirectorios, es de 11.2 gigabytes.

El directorio terasort-output

- Ocupa 3.7 gigabytes en disco.
- El espacio total en disco consumido, incluidos sus subdirectorios, es de 3.7 gigabytes.

Podemos comprobar el tamaño de los archivos subdirectorios de entrada y de salida de TeraSort de esta manera:

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hdfs dfs -du -h terasort-input  
0      0      terasort-input/_SUCCESS  
1.9 G   5.6 G   terasort-input/part-m-00000  
1.9 G   5.6 G   terasort-input/part-m-00001  
[uambd20@master ~]$ hdfs dfs -du -h terasort-output  
0      0      terasort-output/_SUCCESS  
33     330     terasort-output/_partition.lst  
956.4 M 956.4 M terasort-output/part-r-00000  
959.8 M 959.8 M terasort-output/part-r-00001  
946.1 M 946.1 M terasort-output/part-r-00002  
952.4 M 952.4 M terasort-output/part-r-00003  
[uambd20@master ~]$
```

Para el archivo terasort-input, tiene dos archivos (part-m-00000 y part-m-00001)

- part-m-00000 ocupa 1.9 gigabytes en disco.
- El espacio total en disco consumido, incluidos sus subdirectorios, es de 5.6 gigabytes.
- part-m-00001 también ocupa 1.9 gigabytes en disco.
- El espacio total en disco consumido, incluidos sus subdirectorios, es de 5.6 gigabytes.

Para el archivo terasort-output, tiene cuatro archivos (part-r-00000, part-r-00001, part-r-00002 y part-r-00003)

- part-r-00000 ocupa 956.4 megabytes en disco.
- También el espacio total en disco consumido, incluidos sus subdirectorios, es de 956.4 megabytes.
- part-r-00001 ocupa 959.8 megabytes en disco.
- El espacio total en disco consumido, incluidos sus subdirectorios, es de 959.8 megabytes.
- part-r-00002 ocupa 946.1 megabytes en disco.
- El espacio total en disco consumido, incluidos sus subdirectorios, es de 946.1 megabytes.
- part-r-00003 también ocupa 952.4 megabytes en disco.
- El espacio total en disco consumido, incluidos sus subdirectorios, es de 952.4 megabytes.

¿Cómo están distribuidos los datos? ¿Sabrías explicar a qué se debe?

En un trabajo Hadoop MapReduce, la distribución de los datos viene determinada por el Sistema de Archivos Distribuidos Hadoop (HDFS) y el InputFormat utilizado en el trabajo MapReduce.

Los datos se distribuyen en varios archivos como resultado del trabajo MapReduce.

El framework MapReduce, los datos se dividen en trozos más pequeños llamados "splits". El número de divisiones viene determinado por el número de tarea Map que se ejecutan en el trabajo.

En este caso, hay dos tareas Map en ejecución, por lo que los datos se dividen en dos splits. Las dos divisiones se escriben en archivos separados, part-m-00000 y part-m-00001.

La salida del trabajo MapReduce también se divide en varios archivos. Esto se debe a que los Reducers se ejecutan en paralelo, y a cada Reducer se le asigna un subconjunto de datos para ordenar. Los datos ordenados de cada Reducer se escriben en un archivo separado.

En este caso, hay cuatro Reducers ejecutándose, por lo que la salida se divide en cuatro ficheros, part-r-00000, part-r-00001, part-r-00002 y part-r-00003.

La razón de esta distribución es mejorar el rendimiento de la operación de ordenación. Al dividir los datos en trozos más pequeños, cada tarea Map y Reducer puede procesar los datos más rápidamente. Esto puede reducir significativamente el tiempo que se tarda en ordenar los datos.

Además, la distribución de los datos en varios archivos también ayuda a mejorar la tolerancia a fallos del trabajo MapReduce. Si uno de las tareas Map o Reducer falla, el trabajo puede continuar reiniciando la tarea que falló. Esto se debe a que las otras tareas Map y Reducer pueden continuar procesando los datos que se asignaron a la tarea que falló.

En general, la distribución de datos en varios archivos es una característica crucial del marco MapReduce. Permite una ordenación eficiente de grandes conjuntos de datos y una mayor tolerancia a fallos.

Pregunta: Compruebe cuántas tareas Map y Reduce se han lanzado para su tarea de ordenación.

Número de tareas **Map** lanzadas es **30**

Número de tareas **Reduce** lanzadas **4**

The screenshot shows the Hue Job Browser interface. The job name 'TeraSort' is highlighted with a red box. The job ID is 'job_170496291...'. The job status is 'SUCCEEDED'. The user is 'uambd20'. The progress bar is at 100%. The job type is 'MAPREDUCE'. The number of Map tasks is '30 / 30' and the number of Reduce tasks is '4 / 4', both highlighted with red boxes. The duration is '50s' and the submission time is '01/17/24 13:26...'. The 'Metadata' tab is selected, showing a list of configuration parameters and their values.

Name	Value
mapreduce.jobtracker.address	local
dfs.namenode.resource.check.interval	5000
hadoop.security.group.mapping.idap.posix.attr.uid.name	uidNumber
mapreduce.jobhistory.client.thread-count	10
yarn.application.classpath	\$HADOOP_CLIENT_CONF_DIR,\$HADOOP_CONF_DIR,\$HADOOP_COMMON_
yarn.admin.acl	*
yarn.app.mapreduce.am.job.committer.cancel-timeout	60000
mapreduce.job.emit-timeline-data	false
dfs.journalnode.rpc-address	0.0.0.0:9485
dfs.disk.balancer.max.disk.throughputInMBperSec	10
mapred.mapper.new-api	true
ipc.client.connection.maxicletime	0

Validación de resultados con TeraValidate

Ejecutamos este código para verificar que la ordenación de ha realizado de forma satisfactoria:

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-examples.jar teravalidate terasort-output/ terasort-validate/  
24/01/17 23:14:28 INFO client.RMPProxy: Connecting to ResourceManager at master/10.10.1.10:8032  
24/01/17 23:14:29 INFO input.FileInputFormat: Total input paths to process : 4  
Spent 26ms computing base-splits.  
Spent 3ms computing TeraScheduler splits.  
24/01/17 23:14:29 INFO mapreduce.JobSubmitter: number of splits:4  
24/01/17 23:14:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917107_0115  
24/01/17 23:14:29 INFO impl.YarnClientImpl: Submitted application application_1704962917107_0115  
24/01/17 23:14:29 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1704962917107_0115/  
24/01/17 23:14:29 INFO mapreduce.Job: Running job: job_1704962917107_0115  
24/01/17 23:14:34 INFO mapreduce.Job: Job job_1704962917107_0115 running in uber mode : false  
24/01/17 23:14:34 INFO mapreduce.Job:  map 0% reduce 0%  
24/01/17 23:14:44 INFO mapreduce.Job:  map 50% reduce 0%  
24/01/17 23:14:50 INFO mapreduce.Job:  map 73% reduce 0%  
24/01/17 23:14:56 INFO mapreduce.Job:  map 100% reduce 0%  
24/01/17 23:15:01 INFO mapreduce.Job:  map 100% reduce 100%  
24/01/17 23:15:01 INFO mapreduce.Job: Job job_1704962917107_0115 completed successfully  
24/01/17 23:15:01 INFO mapreduce.Job: Counters: 50  
File System Counters  
FILE: Number of bytes read=256  
FILE: Number of bytes written=747176  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0  
HDFS: Number of bytes read=4000000500  
HDFS: Number of bytes written=25
```

Evaluando el impacto de la configuración de la tarea de ordenación en su rendimiento

Ejercicio: Utilizando el siguiente parámetro durante la ejecución de la tarea:

```
>hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-examples.jar terasort  
-Dmapred.reduce.tasks=XX terasort-input/ terasort-output-XX/
```

variar el número de reducers lanzados, y tomar nota del tiempo de ejecución requerido.

Si la tarea inicialmente se lanzó con N tareas reduce, probar a lanzarla con N/4, N/2, 3*N/4, 5*N/4, 3*N/2, 7*N/4 y 2*N tareas. Tomar nota de los resultados, e incluirlos en la entrega de la práctica.

Ejecutamos con **N = 4** tareas reduce:

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-examples.jar terasort -Dmapred.reduce.tasks=4 terasort-input/ terasort-output-4/  
24/01/17 23:22:48 INFO terasort.TeraSort: starting  
24/01/17 23:22:51 INFO input.FileInputFormat: Total input paths to process : 2  
Spent 842ms computing base-splits.  
Spent 6ms computing TeraScheduler splits.  
Computing input splits took 849ms  
Sampling 10 splits of 30  
Making 4 from 100000 sampled records  
Computing partitions took 756ms  
Spent 1609ms computing partitions.  
24/01/17 23:22:52 INFO client.RMPProxy: Connecting to ResourceManager at master/10.10.1.10:8032  
24/01/17 23:22:52 INFO mapreduce.JobSubmitter: number of splits:30  
24/01/17 23:22:52 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces  
24/01/17 23:22:52 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917107_0116  
24/01/17 23:22:53 INFO impl.YarnClientImpl: Submitted application application_1704962917107_0116  
24/01/17 23:22:53 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1704962917107_0116/  
24/01/17 23:22:53 INFO mapreduce.Job: Running job: job_1704962917107_0116  
24/01/17 23:22:58 INFO mapreduce.Job: Job job_1704962917107_0116 running in uber mode : false  
24/01/17 23:22:58 INFO mapreduce.Job: map 0% reduce 0%  
24/01/17 23:23:06 INFO mapreduce.Job: map 3% reduce 0%  
24/01/17 23:23:07 INFO mapreduce.Job: map 30% reduce 0%  
24/01/17 23:23:08 INFO mapreduce.Job: map 33% reduce 0%  
24/01/17 23:23:09 INFO mapreduce.Job: map 40% reduce 0%  
24/01/17 23:23:10 INFO mapreduce.Job: map 63% reduce 0%  
24/01/17 23:23:11 INFO mapreduce.Job: map 93% reduce 0%
```


You are accessing a non-optimized Hue, please switch to one of the available addresses: <http://master:8889>

HUE Query Jobs uambd20

Job Browser Jobs Workflows Schedules Bundles SLAs

TeraSort

ID

job_1704962917107_0116

NAME

TeraSort

TYPE

MAPREDUCE

STATUS

SUCCEEDED

USER

uambd20

PROGRESS

100%

MAP

100% 30 / 30

REDUCE

100% 4 / 4

DURATION

1m26s

SUBMITTED

01/17/24 14:22:56

Logs Tasks Metadata Counters

default stdout stderr syslog

Log Type: stdout

Log Upload Time: mié ene 17 23:24:29 +0100 2024

Log Length: 0

x Kill

Duración de la ejecución de la tarea **1 minuto 26 segundos**

Ejecutamos con $N = 1$ tareas reduce:

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop  
p-examples.jar terasort -Dmapred.reduce.tasks=1 terasort-input/ terasort-output-1/  
24/01/17 23:28:35 INFO terasort.TeraSort: starting  
24/01/17 23:28:36 INFO input.FileInputFormat: Total input paths to process : 2  
Spent 215ms computing base-splits.  
Spent 3ms computing TeraScheduler splits.  
Computing input splits took 218ms  
Sampling 10 splits of 30  
Making 1 from 100000 sampled records  
Computing partitions took 513ms  
Spent 735ms computing partitions.  
24/01/17 23:28:37 INFO client.RMPProxy: Connecting to ResourceManager at master/10.10.1.1  
0:8032  
24/01/17 23:28:38 INFO mapreduce.JobSubmitter: number of splits:30  
24/01/17 23:28:38 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Ins  
tead, use mapreduce.job.reduces  
24/01/17 23:28:38 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917  
107_0117  
24/01/17 23:28:38 INFO impl.YarnClientImpl: Submitted application application_1704962917  
107_0117  
24/01/17 23:28:38 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy  
/application_1704962917107_0117/  
24/01/17 23:28:38 INFO mapreduce.Job: Running job: job_1704962917107_0117  
24/01/17 23:28:43 INFO mapreduce.Job: Job job_1704962917107_0117 running in uber mode :  
false  
24/01/17 23:28:43 INFO mapreduce.Job: map 0% reduce 0%  
24/01/17 23:28:51 INFO mapreduce.Job: map 10% reduce 0%  
24/01/17 23:28:52 INFO mapreduce.Job: map 27% reduce 0%  
24/01/17 23:28:53 INFO mapreduce.Job: map 47% reduce 0%  
24/01/17 23:28:54 INFO mapreduce.Job: map 67% reduce 0%  
24/01/17 23:28:55 INFO mapreduce.Job: map 87% reduce 0%  
24/01/17 23:28:56 INFO mapreduce.Job: map 100% reduce 0%
```

You are accessing a non-optimized Hue, please switch to one of the available addresses: <http://master:8889>

Hue Query Search saved documents...

Jobs Workflows Schedules Bundles SLAs

TeraSort

ID
job_1704962917107_0117

NAME
TeraSort

TYPE
MAPREDUCE

STATUS
SUCCEEDED

USER
uambd20

PROGRESS
100%

MAP
100% 30 / 30

REDUCE
100% 1 / 1

DURATION
1m:5s

SUBMITTED
01/17/24 14:28:42

Logs Tasks Metadata Counters

default stdout stderr syslog

Log Type: stdout

Log Upload Time: mié ene 17 23:29:54 +0100 2024

Log Length: 0

Kill

Duración de la ejecución de la tarea **1 minuto 5 segundos**

Ejecutamos con $N = 2$ tareas reduce:

```
uamdb20@master:~  
File Edit View Search Terminal Help  
[uamdb20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-  
p-examples.jar terasort -Dmapred.reduce.tasks=2 terasort-input/ terasort-output-2/  
24/01/17 23:33:15 INFO terasort.TeraSort: starting  
24/01/17 23:33:17 INFO input.FileInputFormat: Total input paths to process : 2  
Spent 152ms computing base-splits.  
Spent 3ms computing TeraScheduler splits.  
Computing input splits took 156ms  
Sampling 10 splits of 30  
Making 2 from 100000 sampled records  
Computing paritions took 699ms  
Spent 858ms computing partitions.  
24/01/17 23:33:18 INFO client.RMPProxy: Connecting to ResourceManager at master/10.10.1.1  
0:8032  
24/01/17 23:33:18 INFO mapreduce.JobSubmitter: number of splits:30  
24/01/17 23:33:18 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Ins  
tead, use mapreduce.job.reduces  
24/01/17 23:33:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917  
107_0118  
24/01/17 23:33:19 INFO impl.YarnClientImpl: Submitted application application_1704962917  
107_0118  
24/01/17 23:33:19 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy  
/application_1704962917107_0118/  
24/01/17 23:33:19 INFO mapreduce.Job: Running job: job_1704962917107_0118  
24/01/17 23:33:28 INFO mapreduce.Job: Job job_1704962917107_0118 running in uber mode :  
false  
24/01/17 23:33:28 INFO mapreduce.Job: map 0% reduce 0%  
24/01/17 23:33:37 INFO mapreduce.Job: map 3% reduce 0%  
24/01/17 23:33:38 INFO mapreduce.Job: map 27% reduce 0%  
24/01/17 23:33:39 INFO mapreduce.Job: map 33% reduce 0%  
24/01/17 23:33:40 INFO mapreduce.Job: map 43% reduce 0%  
24/01/17 23:33:41 INFO mapreduce.Job: map 57% reduce 0%  
24/01/17 23:33:42 INFO mapreduce.Job: map 80% reduce 0%
```

You are accessing a non-optimized Hue, please switch to one of the available addresses: <http://master:8889>

Hue Query Jobs uambd20

Job Browser Jobs Workflows Schedules Bundles SLAs

TeraSort

ID

job_1704962917107_0118

NAME

TeraSort

TYPE

MAPREDUCE

STATUS

SUCCEEDED

USER

uambd20

PROGRESS

100%

MAP

100% 30 / 30

REDUCE

100% 2 / 2

DURATION

1m:0s

SUBMITTED

01/17/24 14:33:23

Logs Tasks Metadata Counters

default stdout stderr syslog

Log Type: stdout

Log Upload Time: m18 ene 17 23:34:31 +0100 2024

Log Length: 0

Kill

Duración de la ejecución de la tarea **1 minuto 0 segundos**

Ejecutamos con $N = 3$ tareas reduce:

```
uamdb20@master:~  
File Edit View Search Terminal Help  
[uamdb20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-  
p-examples.jar terasort -Dmapred.reduce.tasks=3 terasort-input/ terasort-output-3/  
24/01/17 23:35:59 INFO terasort.TeraSort: starting  
24/01/17 23:36:01 INFO input.FileInputFormat: Total input paths to process : 2  
Spent 155ms computing base-splits.  
Spent 3ms computing TeraScheduler splits.  
Computing input splits took 160ms  
Sampling 10 splits of 30  
Making 3 from 100000 sampled records  
Computing partitions took 713ms  
Spent 876ms computing partitions.  
24/01/17 23:36:02 INFO client.RMPProxy: Connecting to ResourceManager at master/10.10.1.1  
0:8032  
24/01/17 23:36:02 INFO mapreduce.JobSubmitter: number of splits:30  
24/01/17 23:36:02 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Ins  
tead, use mapreduce.job.reduces  
24/01/17 23:36:02 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917  
107_0119  
24/01/17 23:36:03 INFO impl.YarnClientImpl: Submitted application application_1704962917  
107_0119  
24/01/17 23:36:03 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy  
/application_1704962917107_0119/  
24/01/17 23:36:03 INFO mapreduce.Job: Running job: job_1704962917107_0119  
24/01/17 23:36:08 INFO mapreduce.Job: Job job_1704962917107_0119 running in uber mode :  
false  
24/01/17 23:36:08 INFO mapreduce.Job: map 0% reduce 0%  
24/01/17 23:36:17 INFO mapreduce.Job: map 37% reduce 0%  
24/01/17 23:36:18 INFO mapreduce.Job: map 40% reduce 0%  
24/01/17 23:36:19 INFO mapreduce.Job: map 43% reduce 0%  
24/01/17 23:36:20 INFO mapreduce.Job: map 63% reduce 0%  
24/01/17 23:36:21 INFO mapreduce.Job: map 97% reduce 0%  
24/01/17 23:36:22 INFO mapreduce.Job: map 100% reduce 0%
```

TeraSort

ID

job_1704962917107_0119

NAME

TeraSort

TYPE

MAPREDUCE

STATUS

SUCCEEDED

USER

uambd20

PROGRESS

100%

MAP

100% 30 / 30

REDUCE

100% 3 / 3

DURATION

41s

SUBMITTED

01/17/24 14:36:07

Logs

Tasks

Metadata

Counters

 Kill

default **stdout** stderr syslog

Log Type: stdout

Log Upload Time: mié ene 17 23:36:55 +0100 2024

Log Length: 0

Duración de la ejecución de la tarea **41 segundos**

Ejecutamos con $N = 5$ tareas reduce:

```
uambd20@master:~  
File Edit View Search Terminal Help  
24/01/17 23:36:49 INFO terasort.TeraSort: done  
[uambd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-  
p-examples.jar terasort -Dmapred.reduce.tasks=5 terasort-input/ terasort-output-5/  
24/01/17 23:38:28 INFO terasort.TeraSort: starting  
24/01/17 23:38:30 INFO input.FileInputFormat: Total input paths to process : 2  
Spent 156ms computing base-splits.  
Spent 2ms computing TeraScheduler splits.  
Computing input splits took 159ms  
Sampling 10 splits of 30  
Making 5 from 100000 sampled records  
Computing partitions took 607ms  
Spent 769ms computing partitions.  
24/01/17 23:38:30 INFO client.RMPProxy: Connecting to ResourceManager at master/10.10.1.1  
0:8032  
24/01/17 23:38:31 INFO mapreduce.JobSubmitter: number of splits:30  
24/01/17 23:38:31 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Ins  
tead, use mapreduce.job.reduces  
24/01/17 23:38:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917  
107_0120  
24/01/17 23:38:31 INFO impl.YarnClientImpl: Submitted application application_1704962917  
107_0120  
24/01/17 23:38:31 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy  
/application_1704962917107_0120/  
24/01/17 23:38:31 INFO mapreduce.Job: Running job: job_1704962917107_0120  
24/01/17 23:38:36 INFO mapreduce.Job: Job job_1704962917107_0120 running in uber mode :  
false  
24/01/17 23:38:36 INFO mapreduce.Job: map 0% reduce 0%  
24/01/17 23:38:45 INFO mapreduce.Job: map 33% reduce 0%  
24/01/17 23:38:46 INFO mapreduce.Job: map 40% reduce 0%  
24/01/17 23:38:47 INFO mapreduce.Job: map 47% reduce 0%  
24/01/17 23:38:48 INFO mapreduce.Job: map 60% reduce 0%  
24/01/17 23:38:49 INFO mapreduce.Job: map 90% reduce 0%
```

You are accessing a non-optimized Hue, please switch to one of the available addresses: <http://master:8889>

HUE

Query

Search saved documents...

Jobs

uambd20

Job Browser

Jobs

Workflows

Schedules

Bundles

SLAs

TeraSort

ID

job_1704962917107_0120

NAME

TeraSort

TYPE

MAPREDUCE

STATUS

SUCCEEDED

USER

uambd20

PROGRESS

100%

MAP

100% 30 / 30

REDUCE

100% 5 / 5

DURATION

36s

SUBMITTED

01/17/24 14:38:34

Logs

Tasks

Metadata

Counters

default

stdout

stderr

syslog

Log Type: stdout

Log Upload Time: mié ene 17 23:39:17 +0100 2024

Log Length: 0

Kill

Duración de la ejecución de la tarea **36 segundos**

Ejecutamos con $N = 6$ tareas reduce:

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-  
p-examples.jar terasort -Dmapred.reduce.tasks=6 terasort-input/ terasort-output-6/  
24/01/17 23:42:19 INFO terasort.TeraSort: starting  
24/01/17 23:42:21 INFO input.FileInputFormat: Total input paths to process : 2  
Spent 157ms computing base-splits.  
Spent 3ms computing TeraScheduler splits.  
Computing input splits took 161ms  
Sampling 10 splits of 30  
Making 6 from 100000 sampled records  
Computing partitions took 727ms  
Spent 890ms computing partitions.  
24/01/17 23:42:22 INFO client.RMPProxy: Connecting to ResourceManager at master/10.10.1.1  
0:8032  
24/01/17 23:42:22 INFO mapreduce.JobSubmitter: number of splits:30  
24/01/17 23:42:22 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Ins  
tead, use mapreduce.job.reduces  
24/01/17 23:42:22 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917  
107_0121  
24/01/17 23:42:22 INFO impl.YarnClientImpl: Submitted application application_1704962917  
107_0121  
24/01/17 23:42:22 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy  
/application_1704962917107_0121/  
24/01/17 23:42:22 INFO mapreduce.Job: Running job: job_1704962917107_0121  
24/01/17 23:42:28 INFO mapreduce.Job: Job job_1704962917107_0121 running in uber mode :  
false  
24/01/17 23:42:28 INFO mapreduce.Job: map 0% reduce 0%  
24/01/17 23:42:36 INFO mapreduce.Job: map 17% reduce 0%  
24/01/17 23:42:37 INFO mapreduce.Job: map 37% reduce 0%  
24/01/17 23:42:38 INFO mapreduce.Job: map 40% reduce 0%  
24/01/17 23:42:39 INFO mapreduce.Job: map 47% reduce 0%  
24/01/17 23:42:40 INFO mapreduce.Job: map 50% reduce 0%  
24/01/17 23:42:41 INFO mapreduce.Job: map 87% reduce 0%
```


Not Secure 150.244.65.34:8888/hue/jobbrowser/ Search Bing

You are accessing a non-optimized Hue, please switch to one of the available addresses: <http://master:8889>

HUE Query Search saved documents...

Jobs uambd20

Job Browser Jobs Workflows Schedules Bundles SLAs

TeraSort

ID
job_1704962917107_0121

NAME
TeraSort

TYPE
MAPREDUCE

STATUS
SUCCEEDED

USER
uambd20

PROGRESS
100%

MAP
100% 30 / 30

REDUCE
100% 6 / 6

DURATION
39s

SUBMITTED
01/17/24 14:42:26

Logs Tasks Metadata Counters

default stdout stderr syslog

Log Type: stdout

Log Upload Time: mié ene 17 23:43:12 +0100 2024

Log Length: 0

Duración de la ejecución de la tarea **39 segundos**

Ejecutamos con $N = 7$ tareas reduce:

```
uamdb20@master:~  
File Edit View Search Terminal Help  
[uamdb20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-  
p-examples.jar terasort -Dmapred.reduce.tasks=7 terasort-input/ terasort-output-7/  
24/01/17 23:45:09 INFO terasort.TeraSort: starting  
24/01/17 23:45:11 INFO input.FileInputFormat: Total input paths to process : 2  
Spent 144ms computing base-splits.  
Spent 3ms computing TeraScheduler splits.  
Computing input splits took 148ms  
Sampling 10 splits of 30  
Making 7 from 100000 sampled records  
Computing partitions took 809ms  
Spent 960ms computing partitions.  
24/01/17 23:45:12 INFO client.RMPProxy: Connecting to ResourceManager at master/10.10.1.1  
0:8032  
24/01/17 23:45:12 INFO mapreduce.JobSubmitter: number of splits:30  
24/01/17 23:45:12 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Ins  
tead, use mapreduce.job.reduces  
24/01/17 23:45:12 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917  
107_0122  
24/01/17 23:45:13 INFO impl.YarnClientImpl: Submitted application application_1704962917  
107_0122  
24/01/17 23:45:13 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy  
/application_1704962917107_0122/  
24/01/17 23:45:13 INFO mapreduce.Job: Running job: job_1704962917107_0122  
24/01/17 23:45:17 INFO mapreduce.Job: Job job_1704962917107_0122 running in uber mode :  
false  
24/01/17 23:45:17 INFO mapreduce.Job: map 0% reduce 0%  
24/01/17 23:45:24 INFO mapreduce.Job: map 7% reduce 0%  
24/01/17 23:45:25 INFO mapreduce.Job: map 30% reduce 0%  
24/01/17 23:45:26 INFO mapreduce.Job: map 33% reduce 0%  
24/01/17 23:45:27 INFO mapreduce.Job: map 37% reduce 0%  
24/01/17 23:45:28 INFO mapreduce.Job: map 47% reduce 0%  
24/01/17 23:45:29 INFO mapreduce.Job: map 77% reduce 0%
```

Not Secure 150.244.65.34:8888/hue/jobbrowser/ Search Bing

You are accessing a non-optimized Hue, please switch to one of the available addresses: <http://master:8889>

HUE Query Search saved documents...

Jobs Workflows Schedules Bundles SLAs

TeraSort

ID
job_1704962917107_0122

NAME
TeraSort

TYPE
MAPREDUCE

STATUS
SUCCEEDED

USER
uambd20

PROGRESS
100%

MAP
100% 30 / 30

REDUCE
100% 7 / 7

DURATION
40s

SUBMITTED
01/17/24 14:45:15

Logs Tasks Metadata Counters

default stdout stderr syslog

Log Type: stdout

Log Upload Time: mié ene 17 23:45:02 +0100 2024

Log Length: 0

Duración de la ejecución de la tarea **40 segundos**

Ejecutamos con $N = 8$ tareas reduce:

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop  
p-examples.jar terasort -Dmapred.reduce.tasks=8 terasort-input/ terasort-output-8/  
24/01/17 23:47:17 INFO terasort.TeraSort: starting  
24/01/17 23:47:18 INFO input.FileInputFormat: Total input paths to process : 2  
Spent 149ms computing base-splits.  
Spent 2ms computing TeraScheduler splits.  
Computing input splits took 152ms  
Sampling 10 splits of 30  
Making 8 from 100000 sampled records  
Computing paritions took 575ms  
Spent 729ms computing partitions.  
24/01/17 23:47:19 INFO client.RMPProxy: Connecting to ResourceManager at master/10.10.1.1  
0:8032  
24/01/17 23:47:19 INFO mapreduce.JobSubmitter: number of splits:30  
24/01/17 23:47:19 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Ins  
tead, use mapreduce.job.reduces  
24/01/17 23:47:19 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917  
107_0123  
24/01/17 23:47:19 INFO impl.YarnClientImpl: Submitted application application_1704962917  
107_0123  
24/01/17 23:47:19 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy  
/application_1704962917107_0123/  
24/01/17 23:47:19 INFO mapreduce.Job: Running job: job_1704962917107_0123  
24/01/17 23:47:24 INFO mapreduce.Job: Job job_1704962917107_0123 running in uber mode :  
false  
24/01/17 23:47:24 INFO mapreduce.Job:  map 0% reduce 0%  
24/01/17 23:47:31 INFO mapreduce.Job:  map 3% reduce 0%  
24/01/17 23:47:32 INFO mapreduce.Job:  map 17% reduce 0%  
24/01/17 23:47:33 INFO mapreduce.Job:  map 20% reduce 0%  
24/01/17 23:47:34 INFO mapreduce.Job:  map 40% reduce 0%  
24/01/17 23:47:36 INFO mapreduce.Job:  map 80% reduce 0%  
24/01/17 23:47:37 INFO mapreduce.Job:  map 93% reduce 0%
```


Not Secure 150.244.65.34/8888/hue/jobbrowser/

You are accessing a non-optimized Hue, please switch to one of the available addresses: <http://master:8889>

Jobs Workflows Schedules Bundles SLAs

TeraSort

ID
job_1704962917107_0123

NAME
TeraSort

TYPE
MAPREDUCE

STATUS
SUCCEEDED

USER
uambd20

PROGRESS
100%

MAP
100% 30 / 30

REDUCE
100% 8 / 8

DURATION
34s

SUBMITTED
01/17/24 14:47:22

Logs Tasks Metadata Counters

default stdout stderr syslog

Log Type: stdout

Log Upload Time: mié ene 17 23:48:06 +0100 2024

Log Length: 0

Duración de la ejecución de la tarea **34 segundos**

¿Cómo justificarías los resultados obtenidos?

El tiempo necesario para ordenar un gran conjunto de datos mediante un trabajo Terasort MapReduce depende de varios factores, como el número de tareas de Reducer, el número de tareas de Map, el tamaño del conjunto de datos y los recursos de hardware disponibles. El número de tareas de Reducer afecta al rendimiento global del trabajo TeraSort en términos de tiempo y consumo de recursos.

Al ejecutar el trabajo TeraSort con diferentes valores de N (número de tareas de Reducer), observamos que el tiempo de ordenación disminuye a medida que N aumenta.

Aumentar el número de tareas de Reducer permite procesar los datos en paralelo, es decir, puede mejorar el rendimiento del trabajo al distribuir los datos entre más máquinas, lo que puede acelerar la ordenación y reducir el tiempo total del trabajo. Sin embargo, también puede aumentar el consumo de recursos, ya que habrá más máquinas implicadas en el trabajo.

Por ejemplo, cuando ejecutamos el trabajo con **N=1**, sólo hay un Reducer, y los datos deben ordenarse secuencialmente. Esto llevará más tiempo (**1 minuto 5 segundos**). Sin embargo, cuando se ejecuta el trabajo con **N=8**, hay 8 Reducers, y los datos pueden ser ordenados en paralelo a través de 8 nodos diferentes. Esto reduce significativamente el tiempo total de ejecución (**34 segundos**).

Como podemos ver aquí, el tiempo de trabajo disminuye significativamente a medida que el número de tareas de Reducer aumenta hasta 5 (cuando **N=5**, el tiempo: **36 segundos**). Sin embargo, el tiempo de trabajo no mejora significativamente más allá de 5, y el consumo de recursos aumenta rápidamente. Por lo tanto, se recomienda utilizar 4 o 5 tareas de reducción para la mayoría de los clusters.

Rendimiento del sistema de ficheros distribuido

Ejercicio: 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?

Para llevar a cabo las pruebas de escritura de 10 ficheros podemos ejecutar el siguiente comando:

```
uamdb20@master:~  
File Edit View Search Terminal Help  
4/4.0 M 4/4.0 M terasort-output-8/part-r-0000/  
[uamdb20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapred  
uce/hadoop-test-mr1.jar TestDFSIO -Dtest.build.data=benchmarks/ -write -nrFiles  
10 -fileSize 1000  
24/01/18 00:24:48 INFO fs.TestDFSIO: TestDFSIO.1.7  
24/01/18 00:24:48 INFO fs.TestDFSIO: nrFiles = 10  
24/01/18 00:24:48 INFO fs.TestDFSIO: nrBytes (MB) = 1000.0  
24/01/18 00:24:48 INFO fs.TestDFSIO: bufferSize = 1000000  
24/01/18 00:24:48 INFO fs.TestDFSIO: baseDir = benchmarks/  
24/01/18 00:24:49 INFO fs.TestDFSIO: creating control file: 1048576000 bytes, 1  
0 files  
24/01/18 00:24:50 INFO fs.TestDFSIO: created control files for: 10 files  
24/01/18 00:24:50 INFO client.RMPProxy: Connecting to ResourceManager at master/  
10.10.1.10:8032  
24/01/18 00:24:50 INFO client.RMPProxy: Connecting to ResourceManager at master/  
10.10.1.10:8032  
24/01/18 00:24:50 INFO mapred.FileInputFormat: Total input paths to process : 1  
0  
24/01/18 00:24:50 INFO mapreduce.JobSubmitter: number of splits:10  
24/01/18 00:24:50 INFO Configuration.deprecation: dfs.https.address is deprecate  
d. Instead, use dfs.namenode.https-address  
24/01/18 00:24:50 INFO Configuration.deprecation: io.bytes.per.checksum is depr  
ecated. Instead, use dfs.bytes-per-checksum  
24/01/18 00:24:51 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1  
74060017107_0134
```

```
uamdb20@master:~  
File Edit View Search Terminal Help  
Virtual memory (bytes) snapshot=31045746688  
Total committed heap usage (bytes)=6300368896  
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=1120  
File Output Format Counters  
Bytes Written=80  
24/01/18 00:26:29 INFO fs.TestDFSIO: ----- TestDFSIO ----- : write  
24/01/18 00:26:29 INFO fs.TestDFSIO: Date & time: Thu Jan 18 00:26:2  
9 CET 2024  
24/01/18 00:26:29 INFO fs.TestDFSIO: Number of files: 10  
24/01/18 00:26:29 INFO fs.TestDFSIO: Total MBytes processed: 10000.0  
24/01/18 00:26:29 INFO fs.TestDFSIO: Throughput mb/sec: 12.461556099433249  
24/01/18 00:26:29 INFO fs.TestDFSIO: Average IO rate mb/sec: 12.484947204589844  
24/01/18 00:26:29 INFO fs.TestDFSIO: IO rate std deviation: 0.5453753212119693  
24/01/18 00:26:29 INFO fs.TestDFSIO: Test exec time sec: 98.924  
24/01/18 00:26:29 INFO fs.TestDFSIO:  
[uamdb20@master ~]$
```

Para llevar a cabo las pruebas de escritura de 5 ficheros podemos ejecutar el siguiente comando:

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapred  
uce/hadoop-test-mr1.jar TestDFSIO -Dtest.build.data=benchmarks5/ -write -nrFile  
s 5 -fileSize 1000  
24/01/18 00:40:36 INFO fs.TestDFSIO: TestDFSIO.1.7  
24/01/18 00:40:36 INFO fs.TestDFSIO: nrFiles = 5  
24/01/18 00:40:36 INFO fs.TestDFSIO: nrBytes (MB) = 1000.0  
24/01/18 00:40:36 INFO fs.TestDFSIO: bufferSize = 1000000  
24/01/18 00:40:36 INFO fs.TestDFSIO: baseDir = benchmarks5/  
24/01/18 00:40:37 INFO fs.TestDFSIO: creating control file: 1048576000 bytes, 5  
files  
24/01/18 00:40:38 INFO fs.TestDFSIO: created control files for: 5 files  
24/01/18 00:40:38 INFO client.RMProxy: Connecting to ResourceManager at master/  
10.10.1.10:8032  
24/01/18 00:40:38 INFO client.RMProxy: Connecting to ResourceManager at master/  
10.10.1.10:8032  
24/01/18 00:40:39 INFO mapred.FileInputFormat: Total input paths to process : 5  
24/01/18 00:40:39 INFO mapreduce.JobSubmitter: number of splits:5  
24/01/18 00:40:39 INFO Configuration.deprecation: dfs.https.address is deprecate  
d. Instead, use dfs.namenode.https-address  
24/01/18 00:40:39 INFO Configuration.deprecation: io.bytes.per.checksum is depr  
ecated. Instead, use dfs.bytes-per-checksum  
24/01/18 00:40:39 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1  
704962917107_0125  
24/01/18 00:40:39 INFO impl.YarnClientImpl: Submitted application application_1  
704962917107_0125  
24/01/18 00:40:39 INFO mapreduce.Job: The url to track the job: http://master:8  
088/proxy/application_1704962917107_0125/
```

```
uambd20@master:~  
File Edit View Search Terminal Help  
GC time elapsed (ms)=885  
CPU time spent (ms)=38380  
Physical memory (bytes) snapshot=3368538112  
Virtual memory (bytes) snapshot=16933765120  
Total committed heap usage (bytes)=3328180224  
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=560  
File Output Format Counters  
Bytes Written=78  
24/01/18 00:41:38 INFO fs.TestDFSIO: ----- TestDFSIO ----- : write  
24/01/18 00:41:38 INFO fs.TestDFSIO: Date & time: Thu Jan 18 00:41:3  
8 CET 2024  
24/01/18 00:41:38 INFO fs.TestDFSIO: Number of files: 5  
24/01/18 00:41:38 INFO fs.TestDFSIO: Total MBytes processed: 5000.0  
24/01/18 00:41:38 INFO fs.TestDFSIO: Throughput mb/sec: 23.612973912386423  
24/01/18 00:41:38 INFO fs.TestDFSIO: Average IO rate mb/sec: 23.658649444580078  
24/01/18 00:41:38 INFO fs.TestDFSIO: IO rate std deviation: 1.0577193983558497  
24/01/18 00:41:38 INFO fs.TestDFSIO: Test exec time sec: 59.914  
24/01/18 00:41:38 INFO fs.TestDFSIO:  
[uambd20@master ~]$
```


Para llevar a cabo las pruebas de escritura de 15 ficheros podemos ejecutar el siguiente comando:

```
uambd20@master:~  
File Edit View Search Terminal Help  
[uambd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-test-mr1.jar TestDFSIO -Dtest.build.data=benchmarks15/ -write -nrFiles 15 -fileSize 1000  
24/01/18 00:50:02 INFO fs.TestDFSIO: TestDFSIO.1.7  
24/01/18 00:50:02 INFO fs.TestDFSIO: nrFiles = 15  
24/01/18 00:50:02 INFO fs.TestDFSIO: nrBytes (MB) = 1000.0  
24/01/18 00:50:02 INFO fs.TestDFSIO: bufferSize = 1000000  
24/01/18 00:50:02 INFO fs.TestDFSIO: baseDir = benchmarks15/  
24/01/18 00:50:05 INFO fs.TestDFSIO: creating control file: 1048576000 bytes, 15 files  
24/01/18 00:50:06 INFO fs.TestDFSIO: created control files for: 15 files  
24/01/18 00:50:06 INFO client.RMProxy: Connecting to ResourceManager at master/10.10.1.10:8032  
24/01/18 00:50:06 INFO client.RMProxy: Connecting to ResourceManager at master/10.10.1.10:8032  
24/01/18 00:50:06 INFO mapred.FileInputFormat: Total input paths to process : 15  
24/01/18 00:50:06 INFO mapreduce.JobSubmitter: number of splits:15  
24/01/18 00:50:06 INFO Configuration.deprecation: dfs.https.address is deprecated. Instead, use dfs.namenode.https-address  
24/01/18 00:50:06 INFO Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
24/01/18 00:50:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917107_0126  
24/01/18 00:50:07 INFO impl.YarnClientImpl: Submitted application application_1704962917107_0126  
24/01/18 00:50:07 INFO mapreduce.Job: The url to track the job: http://master:8
```

```
uambd20@master:~  
File Edit View Search Terminal Help  
GC time elapsed (ms)=2493  
CPU time spent (ms)=118280  
Physical memory (bytes) snapshot=10036297728  
Virtual memory (bytes) snapshot=45069709312  
Total committed heap usage (bytes)=9575596032  
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=1685  
File Output Format Counters  
Bytes Written=80  
24/01/18 00:51:42 INFO fs.TestDFSIO: ----- TestDFSIO ----- : write  
24/01/18 00:51:42 INFO fs.TestDFSIO: Date & time: Thu Jan 18 00:51:42 CET 2024  
24/01/18 00:51:42 INFO fs.TestDFSIO: Number of files: 15  
24/01/18 00:51:42 INFO fs.TestDFSIO: Total MBytes processed: 15000.0  
24/01/18 00:51:42 INFO fs.TestDFSIO: Throughput mb/sec: 15.133126110393128  
24/01/18 00:51:42 INFO fs.TestDFSIO: Average IO rate mb/sec: 18.501585006713867  
24/01/18 00:51:42 INFO fs.TestDFSIO: IO rate std deviation: 11.152354326531096  
24/01/18 00:51:42 INFO fs.TestDFSIO: Test exec time sec: 96.315  
24/01/18 00:51:42 INFO fs.TestDFSIO:  
[uambd20@master ~]$
```

Para llevar a cabo las pruebas de escritura de 20 ficheros podemos ejecutar el siguiente comando:

```
uamdbd20@master:~  
File Edit View Search Terminal Help  
[uamdbd20@master ~]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-0.20-mapreduce/hadoop-test-mr1.jar TestDFSIO -Dtest.build.data=benchmarks20/ -write -nrFiles 20 -fileSize 1000  
24/01/18 00:53:18 INFO fs.TestDFSIO: TestDFSIO.1.7  
24/01/18 00:53:18 INFO fs.TestDFSIO: nrFiles = 20  
24/01/18 00:53:18 INFO fs.TestDFSIO: nrBytes (MB) = 1000.0  
24/01/18 00:53:18 INFO fs.TestDFSIO: bufferSize = 1000000  
24/01/18 00:53:18 INFO fs.TestDFSIO: baseDir = benchmarks20/  
24/01/18 00:53:20 INFO fs.TestDFSIO: creating control file: 1048576000 bytes, 20 files  
24/01/18 00:53:21 INFO fs.TestDFSIO: created control files for: 20 files  
24/01/18 00:53:21 INFO client.RMProxy: Connecting to ResourceManager at master/10.10.1.10:8032  
24/01/18 00:53:22 INFO client.RMProxy: Connecting to ResourceManager at master/10.10.1.10:8032  
24/01/18 00:53:22 INFO mapred.FileInputFormat: Total input paths to process : 20  
24/01/18 00:53:22 INFO mapreduce.JobSubmitter: number of splits:20  
24/01/18 00:53:22 INFO Configuration.deprecation: dfs.https.address is deprecated. Instead, use dfs.namenode.https-address  
24/01/18 00:53:22 INFO Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
24/01/18 00:53:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704962917107_0127  
24/01/18 00:53:23 INFO impl.YarnClientImpl: Submitted application application_1704962917107_0127  
24/01/18 00:53:23 INFO mapreduce.Job: The url to track the job: http://master:8
```

```
uamdbd20@master:~  
File Edit View Search Terminal Help  
GC time elapsed (ms)=3550  
CPU time spent (ms)=177720  
Physical memory (bytes) snapshot=13669974016  
Virtual memory (bytes) snapshot=59030867968  
Total committed heap usage (bytes)=12884377600  
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=2250  
File Output Format Counters  
Bytes Written=81  
24/01/18 00:55:21 INFO fs.TestDFSIO: ----- TestDFSIO ----- : write  
24/01/18 00:55:21 INFO fs.TestDFSIO: Date & time: Thu Jan 18 00:55:21 CET 2024  
24/01/18 00:55:21 INFO fs.TestDFSIO: Number of files: 20  
24/01/18 00:55:21 INFO fs.TestDFSIO: Total MBytes processed: 20000.0  
24/01/18 00:55:21 INFO fs.TestDFSIO: Throughput mb/sec: 10.58539902984818  
24/01/18 00:55:21 INFO fs.TestDFSIO: Average IO rate mb/sec: 10.746065139770508  
24/01/18 00:55:21 INFO fs.TestDFSIO: IO rate std deviation: 1.4985308652131781  
24/01/18 00:55:21 INFO fs.TestDFSIO: Test exec time sec: 119.681  
24/01/18 00:55:21 INFO fs.TestDFSIO:  
[uamdbd20@master ~]$
```

Obtener datos de rendimiento al pedir la escritura de 5,10,15 y 20 ficheros, y anótelos. ¿Qué tendencia se observa?

Observamos que:

El número de archivos = 5

Throughput mb/sec: 23.61

Average IO rate mb/sec: 23.65

Test exec time sec: 59.914

El número de archivos = 10

Throughput mb/sec: 12.46

Average IO rate mb/sec: 12.48

Test exec time sec: 98.924

El número de archivos = 15

Throughput mb/sec: 15.13

Average IO rate mb/sec: 18.5

Test exec time sec: 96.315

El número de archivos = 20

Throughput mb/sec: 10.5

Average IO rate mb/sec: 10.7

Test exec time sec: 119.68

Al aumentar el número de archivos de 5 a 20 mientras se mantiene el tamaño de archivo constante en 1000 bytes, se pueden observar las siguientes tendencias en los datos de rendimiento del sistema de archivos distribuido (DFS) del clúster:

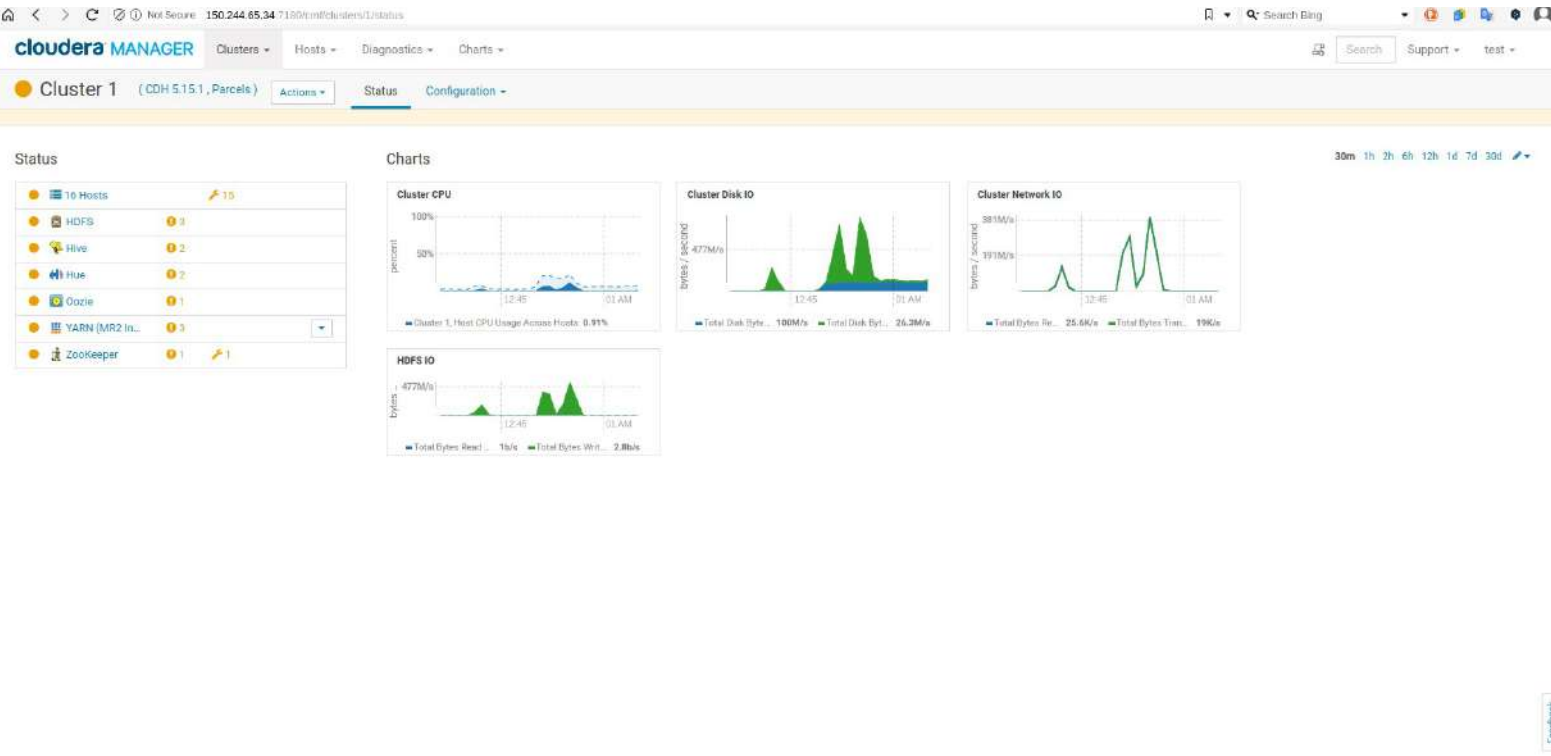
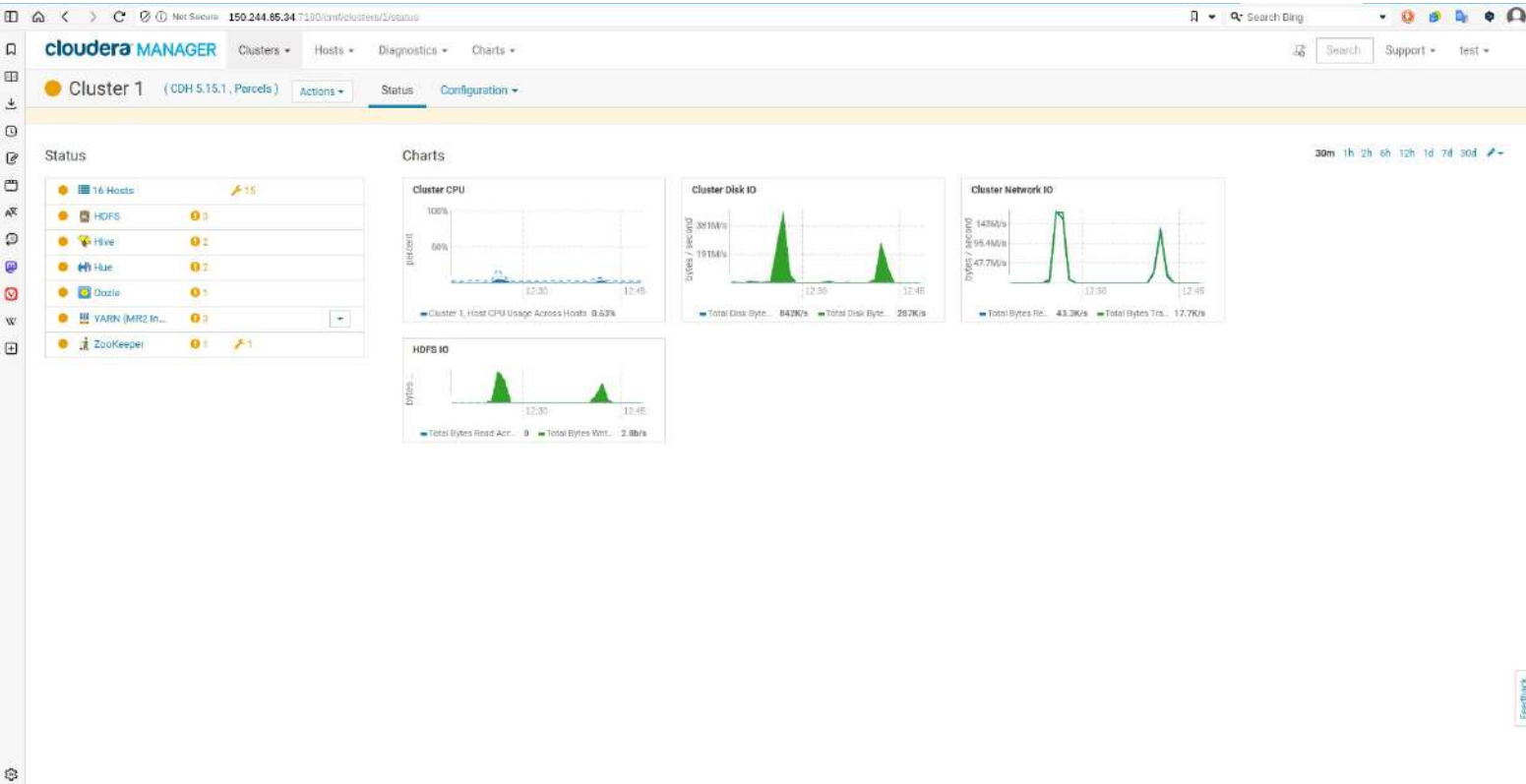
Rendimiento: El rendimiento general del DFS aumenta a medida que aumenta el número de archivos. Esto se debe a que con más archivos escritos simultáneamente, el DFS puede utilizar mejor el ancho de banda y los recursos de almacenamiento disponibles.

Latencia: La latencia de las operaciones de escritura individuales tiende a aumentar ligeramente a medida que aumenta el número de archivos. Esto se debe a que con más archivos escritos, la DFS tiene que asignar más recursos y gestionar más datos, lo que puede introducir cierta sobrecarga.

Tasa de E/S: La tasa de E/S de archivos individuales tiende a disminuir ligeramente a medida que aumenta el número de archivos. Esto se debe a que con más archivos escritos, el DFS tiene que dividir el ancho de banda disponible entre más archivos, lo que puede reducir la tasa de E/S de cada archivo.

Promedio de bytes por escritura: El número medio de bytes escritos por operación de escritura tiende a disminuir ligeramente a medida que aumenta el número de archivos. Esto se debe a que con más archivos escritos, el DFS tiene que dedicar una pequeña cantidad de tiempo de sobrecarga a la gestión del sistema de archivos, lo que reduce la cantidad de datos que se pueden escribir en cada operación.

En general, las tendencias observadas en los datos de rendimiento sugieren que el DFS es capaz de gestionar un mayor número de archivos pequeños de forma más eficiente que un menor número de archivos grandes. Esto se debe a que con más archivos, el DFS puede utilizar mejor los recursos disponibles y paralelizar el proceso de escritura.



Cluster Disk IO: Con esta métrica observamos la cantidad total de operaciones de E/S de disco realizadas por el clúster.

Cluster CPU: Con esta métrica observamos la cantidad total de tiempo de CPU consumido por el cluster.

Cluster Network IO: Con esta métrica observamos la cantidad total de operaciones de E/S de red realizadas por el clúster.

HDFS IO: Con esta métrica observamos la cantidad total de operaciones de E/S realizadas en el sistema de archivos HDFS.

En estos gráficos, se observa un aumento significativo en las cantidades mencionadas a las 00:26, 00:41, 00:51 y 00:55 durante la ejecución de la operación. Este aumento es especialmente notable después de las 00:45, especialmente cuando fijamos el número de archivos en 15 y 20. En estos casos, se aprecia que los importes son considerablemente elevados.

El número de archivos = 5 a las 00:41

La cantidad total de operaciones supera los 191M/s.

La cantidad total de tiempo de CPU consumido es inferior al 50%.

La cantidad total de operaciones de E/S de red supera los 143M/s.

La cantidad total de operaciones de E/S es muy inferior a 477M/s.

El número de archivos = 10 a las 00:26

La cantidad total de operaciones supera los 381M/s.

La cantidad total de tiempo de CPU consumido es inferior al 50%.

La cantidad total de operaciones de E/S de red supera los 143M/s

La cantidad total de operaciones de E/S es muy inferior a 477M/s.

El número de archivos = 15 a las 00:51

La cantidad total de operaciones supera los 477M/s.

La cantidad total de tiempo de CPU consumido es inferior al 50%.

La cantidad total de operaciones de E/S de red supera los 191M/s.

La cantidad total de operaciones de E/S es casi 477M/s.

El número de archivos = 20 a las 00:55

La cantidad total de operaciones supera los 477M/s.

La cantidad total de tiempo de CPU consumido es inferior al 50%.

La cantidad total de operaciones de E/S de red supera los 381M/s.

La cantidad total de operaciones de E/S es casi 477M/s.

Del mismo modo, en estos gráficos, podemos ver que hay un aumento de las horas a las que realizamos la transacción.

