Ejercicios de HDFS

2. Comprobar si hadoop está instalado.

```
[cloudera@quickstart ~]$ hadoop version
Hadoop 2.6.0-cdh5.13.0
Subversion http://github.com/cloudera/hadoop -r 42e8860b182e55321bd5f5605264da4a
dc8882be
Compiled by jenkins on 2017-10-04T18:08Z
Compiled with protoc 2.5.0
From source with checksum 5e84c185f8a22158e2b0e4b8f85311
This command was run using /usr/lib/hadoop/hadoop-common-2.6.0-cdh5.13.0.jar
```

3. Abrir Shell Hadoop.

```
[cloudera@quickstart ~]$ hadoop fs -ls /
Found 6 items
drwxrwxrwx   - hdfs  supergroup          0 2017-10-23 18:15 /benchmarks
drwxr-xr-x   - hbase supergroup          0 2024-07-09 12:35 /hbase
drwxr-xr-x   - solr  solr                0 2017-10-23 18:18 /solr
drwxrwxrwt   - hdfs  supergroup          0 2022-02-17 12:04 /tmp
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 18:17 /user
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 18:17 /var
```

4. Enumera todos los archivos / directorios para la ruta de destino de hdfs.

```
[cloudera@quickstart ~]$ hadoop fs -ls /
Found 6 items
drwxrwxrwx   - hdfs  supergroup          0 2017-10-23 18:15 /benchmarks
drwxr-xr-x   - hbase supergroup          0 2024-07-09 12:35 /hbase
drwxr-xr-x   - solr  solr                0 2017-10-23 18:18 /solr
drwxrwxrwt   - hdfs  supergroup          0 2022-02-17 12:04 /tmp
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 18:17 /user
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 18:17 /var
```

5. Crear un directorio local de trabajo.

a. Dentro de tu nuevo directorio crea un árbol de directorios. /home/ejercicios/A/C /home/ejercicios/B

```
[cloudera@quickstart ~]$ hadoop fs -mkdir -p /user/cloudera/ejercicios/A/C
[cloudera@quickstart ~]$ hadoop fs -mkdir -p /user/cloudera/ejercicios/B
```

b. Comprueba que se han creado correctamente mostrando tanto los directorios como subdirectorios.

```
[cloudera@quickstart ~]$ hadoop fs -ls -R /user/cloudera/ejercicios
drwxr-xr-x   - cloudera cloudera          0 2024-07-09 13:31 /user/cloudera/ejer
cicios/A
drwxr-xr-x   - cloudera cloudera          0 2024-07-09 13:31 /user/cloudera/ejer
cicios/A/C
drwxr-xr-x   - cloudera cloudera          0 2024-07-09 13:31 /user/cloudera/ejer
cicios/B
```

6. Cuenta el número de directorios que contiene.

```
[cloudera@quickstart ~]$ find /user/cloudera/ejercicios
find: `/user/cloudera/ejercicios': No such file or directory
[cloudera@quickstart ~]$ hadoop fs -ls -R /user/cloudera/ejercicios | grep "^d"
| wc -l
3
```

a. ¿Qué significa la salida?

Significa que esta el directorio /A, el /B y el /A/C

7. Muestra el espacio libre en nuestro destino dfs –df hdfs:/

```
[cloudera@quickstart ~]$ hadoop fs -df -h hdfs://quickstart.cloudera:8020/
Filesystem                          Size     Used  Available  Use%
hdfs://quickstart.cloudera:8020  54.5 G  832.2 M     42.7 G    1%
```

8. Copia un archivo con mucho contenido desde el directorio local a hadoop en el directorio A creado anteriormente.

```
[cloudera@quickstart ~]$ hdfs dfs -put ~/Desktop/ejercicios_prueba.txt hdfs://qu
ickstart.cloudera:8020/user/cloudera/ejercicios/A/
```

9. Muestra el tamaño de los archivos y directorios contenidos en el directorio dado.

```
[cloudera@quickstart ~]$ hdfs dfs -du -h hdfs://quickstart.cloudera:8020/user/cl
oudera/ejercicios/A
0  0  hdfs://quickstart.cloudera:8020/user/cloudera/ejercicios/A/C
0  0  hdfs://quickstart.cloudera:8020/user/cloudera/ejercicios/A/ejercicios_prue
ba.txt
```

10. Elimina el directorio B.

a. Comprueba que el archivo se ha eliminado. ¿Se ha eliminado? Busca información sobre las opciones del comando de borrado. ¿Ahora?

```
[cloudera@quickstart ~]$ hdfs dfs -rm -r hdfs://quickstart.cloudera:8020/user/cl
oudera/ejercicios/B
Deleted hdfs://quickstart.cloudera:8020/user/cloudera/ejercicios/B
```

Vemos que no aparece, por lo tanto se ha eliminado:

```
[cloudera@quickstart ~]$ hdfs dfs -ls hdfs://quickstart.cloudera:8020/user/cloud
era/ejercicios/
Found 1 items
drwxr-xr-x   - cloudera cloudera          0 2024-07-09 13:53 hdfs://quickstart.c
loudera:8020/user/cloudera/ejercicios/A
```

11. Comprueba los permisos de archivo subido a hadoop con hdfs dfs –ls

```
[cloudera@quickstart ~]$ hdfs dfs -ls -d -h hdfs://quickstart.cloudera:8020/user
/cloudera/ejercicios/A/ejercicios_prueba.txt
-rw-r--r--   1 cloudera cloudera          0 2024-07-09 13:53 hdfs://quickstart.c
loudera:8020/user/cloudera/ejercicios/A/ejercicios prueba.txt
```

a. Modifica los permisos de ese archivo.

```
[cloudera@quickstart ~]$ hdfs dfs -chmod 755 hdfs://quickstart.cloudera:8020/use
r/cloudera/ejercicios/A/ejercicios_prueba.txt
```

b. Comprueba nuevamente que se ha modificado.

```
[cloudera@quickstart ~]$ hdfs dfs -ls -d -h hdfs://quickstart.cloudera:8020/user
/cloudera/ejercicios/A/ejercicios_prueba.txt
-rwxr-xr-x   1 cloudera cloudera          0 2024-07-09 13:53 hdfs://quickstart.c
loudera:8020/user/cloudera/ejercicios/A/ejercicios_prueba.txt
```

12. Copiar el archivo desde hadoop a tu escritorio.

```
[cloudera@quickstart ~]$ hdfs dfs -get hdfs://quickstart.cloudera:8020/user/clou
dera/ejercicios/A/ejercicios_prueba.txt ~/Desktop/prueba_a_local
```

```
dera/ejercicios/A/ejercicios_prueba.txt ~/Desktop/prueba_a_local
[cloudera@quickstart ~]$ cd Desktop/prueba_a_local/
[cloudera@quickstart prueba_a_local]$ ls
ejercicios_prueba.txt
```

13. Muestra el contenido del archivo de hadoop.

El que había metido estaba vacío, así que he metido otro con contenido dentro.

```
[cloudera@quickstart prueba_a_local]$ hdfs dfs -cat hdfs://quickstart.cloudera:8
020/user/cloudera/ejercicios/A/ejercicios2_prueba.txt
Hola, esto es un texto de prueba.
```

14. Muestra el contenido del archivo de hadoop por la salida estándar.

```
[cloudera@quickstart prueba_a_local]$ hdfs dfs -cat hdfs://quickstart.cloudera:8
020/user/cloudera/ejercicios/A/ejercicios2_prueba.txt > /dev/stdout
Hola, esto es un texto de prueba.
```

15. Crea un nuevo archivo en el directorio /home/ejercicios/A/C en hadoop.

```
[cloudera@quickstart prueba_a_local]$ hdfs dfs -put ~/Desktop/pruebaC.txt hdfs:/
/quickstart.cloudera:8020/user/cloudera/ejercicios/A/C/
```

16. Crea un archivo en local y cópialo al directorio A de hadoop sin usar el comando put.

```
[cloudera@quickstart Desktop]$ hadoop fs -copyFromLocal pruebaSinPut.txt hdfs://
quickstart.cloudera:8020/user/cloudera/ejercicios/A/
```

a. ¿En qué se diferencia del comando put?

hadoop fs -copyFromLocal es una forma explícita de copiar archivos locales a Hadoop, mientras que hdfs dfs -put es una abreviatura más directa pero funcionalmente equivalente.

17. Copia un archivo desde hadoop a local sin usar el comando get.

```
[cloudera@quickstart Desktop]$ hadoop fs -copyToLocal hdfs://quickstart.cloudera
:8020/user/cloudera/ejercicios/A/ejercicios_prueba.txt ~/Desktop/prueba_sin_get
[cloudera@quickstart Desktop]$ cd prueba_sin_get/
[cloudera@quickstart prueba_sin_get]$ ls
ejercicios prueba.txt
```

a. ¿En qué se diferencia del comando get?

En resumen, la diferencia principal entre hadoop fs -copyToLocal y hdfs dfs -get está en la sintaxis y la nomenclatura, pero ambos comandos son equivalentes en términos de funcionalidad para copiar archivos desde HDFS al sistema de archivos local.

18. Copia el fichero del directorio C al directorio A.

```
[cloudera@quickstart prueba_sin_get]$ hdfs dfs -cp hdfs://quickstart.cloudera:80
20/user/cloudera/ejercicios/A/C/pruebaC.txt hdfs://quickstart.cloudera:8020/user
/cloudera/ejercicios/A/
```

confirmándolo…

```
[cloudera@quickstart prueba_sin_get]$ hdfs dfs -ls hdfs://quickstart.cloudera:80
20/user/cloudera/ejercicios/A/
Found 5 items
drwxr-xr-x   - cloudera cloudera          0 2024-07-09 14:20 hdfs://quickstart.c
loudera:8020/user/cloudera/ejercicios/A/C
-rw-r--r--   1 cloudera cloudera         34 2024-07-09 14:17 hdfs://quickstart.c
loudera:8020/user/cloudera/ejercicios/A/ejercicios2_prueba.txt
-rwxr-xr-x   1 cloudera cloudera          0 2024-07-09 13:53 hdfs://quickstart.c
loudera:8020/user/cloudera/ejercicios/A/ejercicios_prueba.txt
-rw-r--r--   1 cloudera cloudera         57 2024-07-09 14:38 hdfs://quickstart.c
loudera:8020/user/cloudera/ejercicios/A/pruebaC.txt
-rw-r--r--   1 cloudera cloudera          0 2024-07-09 14:24 hdfs://quickstart.c
```

Se puede ver que aparece la última.

19. Muestra el ultimo kilobyte del fichero /home/ejercicios/A/

```
[cloudera@quickstart Desktop]$ hdfs dfs -tail hdfs://quickstart.cloudera:8020/us
er/cloudera/ejercicios/A/pruebaC.txt | tail -c 1024
Esto es una prueba que estará dentro de C, dentro de A.
```

20. Muestra la lista de control de acceso ACL del archivo en el directorio. /home/ejercicios/A/

He tenido problemas al meter el comando y para solucionarlo he tenido que modificar

```
[cloudera@quickstart ~]$ hdfs dfs -getfacl /user/cloudera/ejercicios/
# file: /user/cloudera/ejercicios
# owner: cloudera
# group: cloudera
user::rwx
group::r-x
other::r-x
```

21. Muestra el tamaño total del directorio /home de hadoop.

```
[cloudera@quickstart Desktop]$ hdfs dfs -du -s -h hdfs://quickstart.cloudera:802
0/user/cloudera/
148  148  hdfs://quickstart.cloudera:8020/user/cloudera
```

22. Elimina el directorio /home/ejercicios/A y su contenido.

```
[cloudera@quickstart Desktop]$ hdfs dfs -rm -r hdfs://quickstart.cloudera:8020/u
ser/cloudera/ejercicios/A
Deleted hdfs://quickstart.cloudera:8020/user/cloudera/ejercicios/A
```

23. Sube dos archivos con contenido en /home/ejercicios de hadoop.

```
[cloudera@quickstart Desktop]$ hdfs dfs -put ~/Desktop/archivo1 hdfs://quickstar
t.cloudera:8020/user/cloudera/ejercicios
```

```
[cloudera@quickstart Desktop]$ hdfs dfs -put ~/Desktop/archivo2 hdfs://quickstar
t.cloudera:8020/user/cloudera/ejercicios/
```

24. Investiga getmerge. Úsalo con los dos archivos subidos a hadoop del punto anterior.

Se usa para combinar archivos en HDFS en un solo archivo y descargarlo al sistema de archivos local. Útil cuando deseas consolidar múltiples archivos en HDFS en uno solo en tu máquina local.

```
[cloudera@quickstart Desktop]$ hdfs dfs -getmerge hdfs://quickstart.cloudera:802
0/user/cloudera/ejercicios/ archivo_combinado.txt
[cloudera@quickstart Desktop]$ cat archivo_combinado.txt
Primer archivo con contenido. Archivo 1.
Segundo archivo con contenido. Archivo 2.
```

Al haber solo dos archivos en la carpeta ejercicios, se hace directamente ahí.
Si hubiese mas de dos, crearemos una carpeta dentro de ejercicios, y copiaremos los dos archivos ahí, ya que el merge, mergea todos los archivos en uno.
Hay que especificar el nombre del nuevo archivo para que se combinen en él, e irá a local.

Investiga sobre los siguientes comandos

25. Comando checksum.

El comando checksum, lo que hace es devolver el checksum del archivo, que puedes usar para verificar la integridad del archivo en comparación con otro checksum calculado en otro

momento o lugar. Por lo tanto, si cambia el archivo, cambiará el checksum, y se podrá saber que el contenido del archivo ha cambiado.
Ejemplo para ver el checksum de un archivo:

```
hdfs://quickstart.cloudera:8020/user/cloudera/ejercicios/archivo1       MD5-of-0
MD5-of-512CRC32C        000002000000000000000000e460a294753cb220a69e966e8f8b0046
```

26. Comando stat.
Se utiliza para mostrar información detallada sobre un archivo o directorio en HDFS, como el tamaño del archivo, la última modificación, y los permisos.
Si ponemos solamente -stat, sale solo la hora de ultima modificación del archivo:

```
[cloudera@quickstart Desktop]$ hdfs dfs -stat hdfs://quickstart.cloudera:8020/us
er/cloudera/ejercicios/archivo1
2024-07-10 06:16:31
```

Pero podemos poner para que nos saque diferentes datos, como son;
%b - Tamaño del archivo en bytes
%F - Tipo de archivo (e.g., archivo, directorio)
%o - Permisos
%r - Factor de replicación
%u - Propietario
%g - Grupo
%y - Última modificación
%n - Nombre del archivo
Ejemplo con alguno de estos:

```
[cloudera@quickstart Desktop]$ hdfs dfs -stat "%n: %b bytes, %y, %o" hdfs://quic
kstart.cloudera:8020/user/cloudera/ejercicios/archivo1
archivo1: 41 bytes, 2024-07-10 06:16:31, 134217728
```

27. Comando setrep.
Se utiliza para cambiar el factor de replicación de archivos y directorios en HDFS. El factor de replicación define cuántas copias de un archivo se almacenarán en el clúster de HDFS.
Para ver el factor de replicación de nuestro archivo…

```
[cloudera@quickstart Desktop]$ hdfs dfs -stat "%r" hdfs://quickstart.cloudera:80
20/user/cloudera/ejercicios/archivo1
1
```

Vemos que es de 1. Ahora lo cambiaremos a 3 gracias al comando setrep:

```
[cloudera@quickstart Desktop]$ hdfs dfs -setrep 3 hdfs://quickstart.cloudera:802
0/user/cloudera/ejercicios/archivo1
Replication 3 set: hdfs://quickstart.cloudera:8020/user/cloudera/ejercicios/arch
ivo1
```

Ahora comprobamos de nuevo su factor de replicación:

```
[cloudera@quickstart Desktop]$ hdfs dfs -stat "%r" hdfs://quickstart.cloudera:80
20/user/cloudera/ejercicios/archivo1
3
```

Y se puede comprobar que es de 3.
Usar antes del número de replicación una -w, sirve para asegurar que el cambio se haya propagado completamente antes de continuar “hdfs dfs -stat -w 3 …”

28. Comando distcp.

Se utiliza para copiar grandes volúmenes de datos entre sistemas de archivos distribuidos de manera eficiente. Es especialmente útil para mover datos entre diferentes clústeres de Hadoop o entre HDFS y otros sistemas de almacenamiento compatibles.
Comando principal:

```
[cloudera@quickstart Desktop]$ hadoop distcp [opciones] <src_path> <dest_path>
```

En las opciones puede haber las siguientes:

`-update`: Copia solo los archivos que se han modificado.

`-delete`: Elimina los archivos en el destino que no están presentes en el origen.

`-overwrite`: Sobrescribe los archivos en el destino.

`-skipcrccheck`: Omite la verificación CRC.

`-bandwidth <BANDWIDTH>`: Establece el ancho de banda en MBps para la transferencia de datos.

`-i` o `-m <num_maps>`: Establece el número de trabajos de mapeo.

Ejercicios MapReduce (retocar y ampliar)

1. Ejecuta el ejemplo wordcount de MapReduce sobre el archivo generado por getmerge.

Como el archivo generado lo tengo en local, lo primero que hago es copiarlo en hadoop:

```
[cloudera@quickstart ~]$ hdfs dfs -put ~/Desktop/archivo_combinado.txt /user/clo
udera/ejercicios
```

Luego ejecuto el MapReduce sobre el archivo

```
[cloudera@quickstart ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-e
xamples-*.jar wordcount /user/cloudera/ejercicios ~/Desktop/wordcout.txt
24/07/10 11:58:50 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
```

Y el resultado final…

```
[cloudera@quickstart ~]$ hdfs dfs -cat ~/Desktop/wordcout.txt/part-r-00000
1.      2
2.      2
Archivo 4
Primer  2
Segundo 2
archivo 4
con     4
contenido.      4
```

2. Ejecuta el ejemplo pi de MapReduce con 16 10000000. ¿Qué significan estos 2 números?

```
[cloudera@quickstart ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-e
xamples-*.jar pi 16 10000000
Number of Maps  = 16
Samples per Map = 10000000
```

- Quedaria asi al final:

```
                Bytes Written=97
Job Finished in 92.147 seconds
Estimated value of Pi is 3.14159155000000000000
```

- **16**: Número de tareas de mapeo (map tasks). Este valor especifica cuántos mapas en paralelo se utilizarán para calcular Pi. Esto distribuye la carga de trabajo en 16 tareas de mapa.

- **10000000**: Número de muestras por mapa. Este valor indica cuántos puntos se generarán por cada tarea de mapa para utilizar en la estimación de Pi.

3. Ejecuta el ejemplo graysort de MapReduce con 10GB. 2

Primero ejecuto `teragen` para generar 10GB de datos. Aquí, el número de registros para 10GB sería aproximadamente 100,000,000 (cada registro es de 100 bytes).

```
[cloudera@quickstart ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-e
xamples-*.jar teragen 100000000 /user/cloudera/terasort-input
```

Luego ejecuto ejecuto `terasort` para ordenar los datos generados por `teragen`.

```
[cloudera@quickstart ~]$ hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-e
xamples-*.jar terasort /user/cloudera/terasort-input /user/cloudera/terasort-out
put
```

A continuación,