

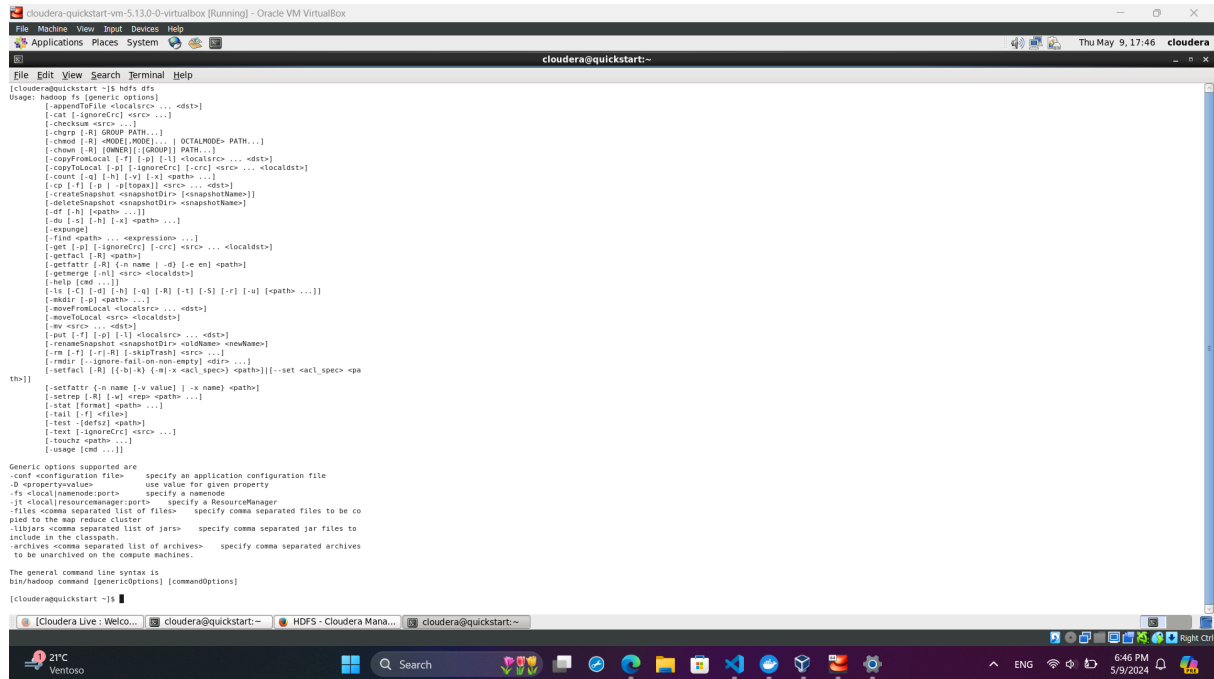
José Andrés Auyón Cóbar 201579  
Diego Alonzo 20172  
Base de datos 2  
Catedrático: Alejandra Mesalles  
Semestre 1, 2024  
10/5/2024

## Laboratorio 7 - Using Hadoop Storage

Repo: <https://github.com/DiggsPapu/DB2.git>

### Ejercicio 1:

hdfs dfs



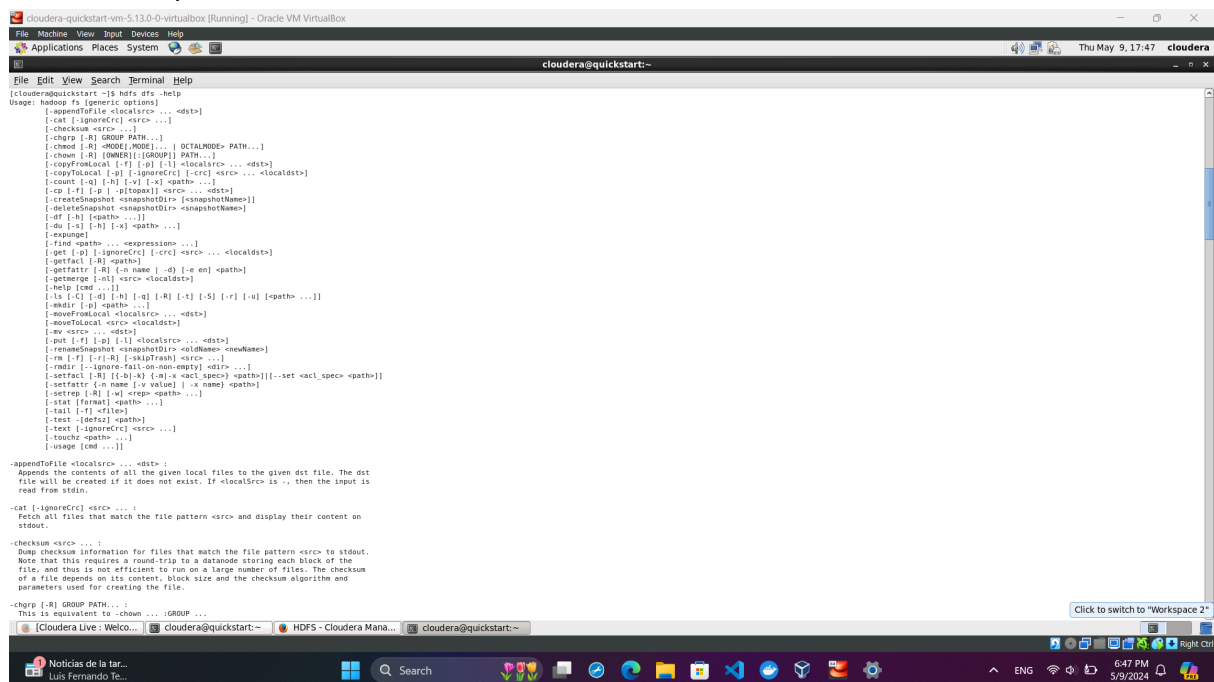
```
cloudera@quickstart:~$ hdfs dfs
Usage: hdfs fs [generic options]
       [-appendToFile <localsrc> ... <dst>]
       [-cat [-ignorecrc] <src> ...]
       [-checksum <src> ...]
       [-chgrp [-R] GROUP PATH ...]
       [-chmod [-R] <MODE,MODE>... [-OCTALMODE] PATH...]
       [-chown [-R] [OWNER][:GROUP] PATH...]
       [-copyFromLocal [-f] [-p] [-l] <localsrc> ... <dst>]
       [-copyFromLocal [-p] [-ignorecrc] [-crc] <src> ... <localdst>]
       [-count [-q] [-h] [-v] [-x] <paths> ...]
       [-cp [-f] [-p] [-ptopax] <src> ... <dst>]
       [-createSnapshot <snapshotDir> <snapshotName>]
       [-deleteSnapshot <snapshotDir> <snapshotName>]
       [-df [-h] <paths> ...]
       [-du [-s] [-h] [-x] <paths> ...]
       [-expunge]
       [-find <paths> ... <expressions> ...]
       [-get [-p] [-ignorecrc] [-crc] <src> ... <localdst>]
       [-getfacl [-R] <paths>]
       [-getfattr [-R] [-n name] [-d] [-e en] <paths>]
       [-getmerge [-n] <src> <localdst>]
       [-help [cmd ...]]
       [-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] <paths> ...]
       [-mkdir [-p] <paths> ...]
       [-moveFromLocal <localsrc> ... <dst>]
       [-moveFromLocal <src> <localdst>]
       [-mv <src> ... <dst>]
       [-put [-f] [-p] [-l] <localsrc> ... <dst>]
       [-renameSnapshot <snapshotDir> <oldName> <newName>]
       [-rm [-f] [-r] [-R] [-skipTrash] <src> ...]
       [-rmr [-ignore-fail-on-non-empty] <dir> ...]
       [-setfacl [-R] [-b] <h> [-n] <acl_spec> <paths>][...set <acl_spec> <pa
th>]

Generic options supported are
  -conf <configuration file>      specify an application configuration file
  -D <property=value>             use value for given property
  -fs <local|namenode|port>       specify a namenode
  -jt <local|resourcemanager:port> specify a ResourceManager
  -files <comma separated list of files> specify comma separated files to be co
ped to the map reduce cluster
  -libjars <comma separated list of jars> specify comma separated jar files to
include in the classpath
  -archives <comma separated list of archives> specify comma separated archives
to be unarchived on the compute machines.

The general command line syntax is
bin/hadoop command [genericOptions] [commandOptions]

cloudera@quickstart:~$
```

### hdfs dfs --help



```
cloudera@quickstart:~$ hdfs dfs --help
Usage: hdfs fs [generic options]
       [-appendToFile <localsrc> ... <dst>]
       [-cat [-ignorecrc] <src> ...]
       [-checksum <src> ...]
       [-chgrp [-R] GROUP PATH...]
       [-chmod [-R] <MODE,MODE>... [-OCTALMODE] PATH...]
       [-chown [-R] [OWNER][:GROUP] PATH...]
       [-copyFromLocal [-f] [-p] [-l] <localsrc> ... <dst>]
       [-copyFromLocal [-p] [-ignorecrc] [-crc] <src> ... <localdst>]
       [-count [-q] [-h] [-v] [-x] <paths> ...]
       [-cp [-f] [-p] [-ptopax] <src> ... <dst>]
       [-createSnapshot <snapshotDir> <snapshotName>]
       [-deleteSnapshot <snapshotDir> <snapshotName>]
       [-df [-h] <paths> ...]
       [-du [-s] [-h] [-x] <paths> ...]
       [-expunge]
       [-find <paths> ... <expressions> ...]
       [-get [-p] [-ignorecrc] [-crc] <src> ... <localdst>]
       [-getfacl [-R] <paths>]
       [-getfattr [-R] [-n name] [-d] [-e en] <paths>]
       [-getmerge [-n] <src> <localdst>]
       [-help [cmd ...]]
       [-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] <paths> ...]
       [-mkdir [-p] <paths> ...]
       [-moveFromLocal <localsrc> ... <dst>]
       [-moveFromLocal <src> <localdst>]
       [-mv <src> ... <dst>]
       [-put [-f] [-p] [-l] <localsrc> ... <dst>]
       [-renameSnapshot <snapshotDir> <oldName> <newName>]
       [-rm [-f] [-r] [-R] [-skipTrash] <src> ...]
       [-rmr [-ignore-fail-on-non-empty] <dir> ...]
       [-setfacl [-R] [-b] <h> [-n] <acl_spec> <paths>][...set <acl_spec> <paths>]
       [-setfattr [-n name] [-v value] [-x name] <paths>]
       [-setrep [-R] [-w] <rep> <paths> ...]
       [-stat [format] <paths> ...]
       [-tail [-f] <files>]
       [-test [-default] <paths>]
       [-text [-ignorecrc] <src> ...]
       [-touchz <paths> ...]
       [-usage [cmd ...]]

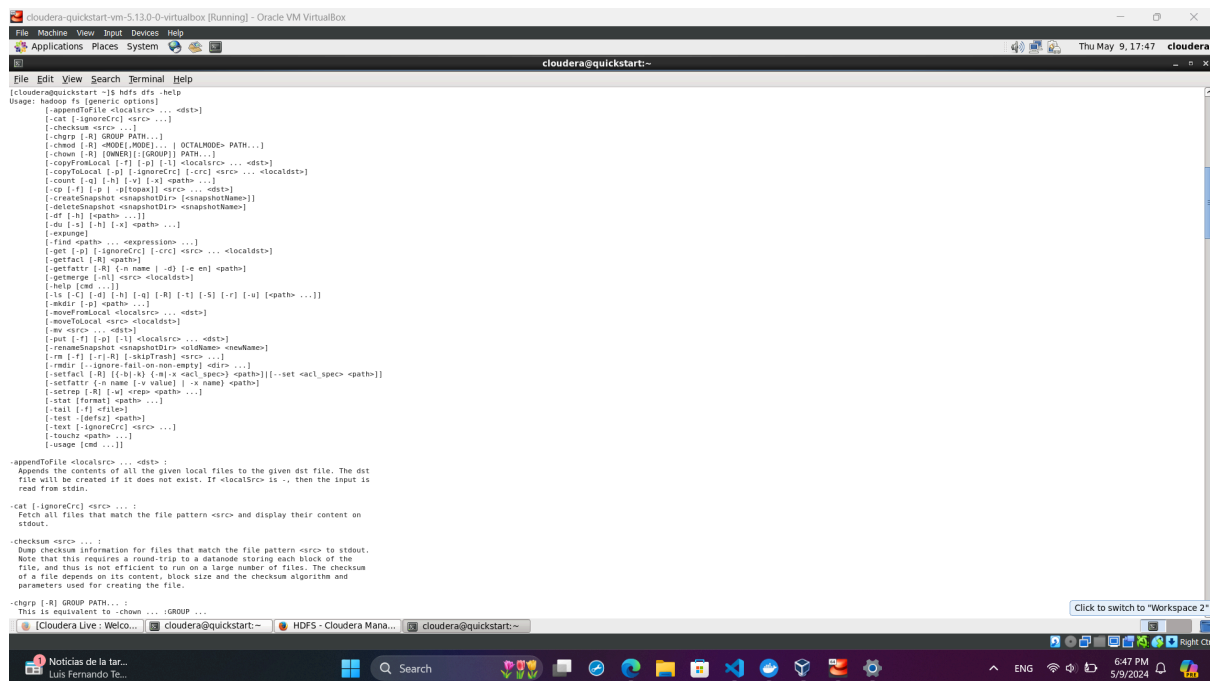
-appendToFile <localsrc> ... <dst> :
  Appends the contents of all the given local files to the given dst file. The dst
  file will be created if it does not exist. If <localSrc> is ., then the input is
  read from stdin.

-cat [-ignorecrc] <src> ... :
  Fetch all files that match the file pattern <src> and display their content on
  stdout.

-checksum <src> ... :
  Dump checksum information for files that match the file pattern <src> to stdout.
  Note that this requires a round-trip to a datanode storing each block of the
  file, and thus is not efficient to run on a large number of files. The checksum
  of a file depends on its content, block size and the checksum algorithm and
  parameters used for creating the file.

-chgrp [-R] GROUP PATH ... :
  This is equivalent to -chown ...:GROUP ...

cloudera@quickstart:~$
```



## hdfs dfs --help ls

```
[cloudera@quickstart ~]$ hdfs dfs -help ls
-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...] :
  List the contents that match the specified file pattern. If path is not
  specified, the contents of /user/<currentUser> will be listed. For a directory a
  list of its direct children is returned (unless -d option is specified).
```

Directory entries are of the form:

```
permissions - userId groupId sizeOfDirectory(in bytes)
modificationDate(yyyy-MM-dd HH:mm) directoryName
```

and file entries are of the form:

```
permissions numberOfReplicas userId groupId sizeOfFile(in bytes)
modificationDate(yyyy-MM-dd HH:mm) fileName
```

- C Display the paths of files and directories only.
- d Directories are listed as plain files.
- h Formats the sizes of files in a human-readable fashion rather than a number of bytes.
- q Print ? instead of non-printable characters.
- R Recursively list the contents of directories.
- t Sort files by modification time (most recent first).
- S Sort files by size.
- r Reverse the order of the sort.
- u Use time of last access instead of modification for display and sorting.

```
[cloudera@quickstart ~]$
```

## hdfs dfs -ls file:///usr

```
[cloudera@quickstart ~]$ hdfs dfs -ls file:///usr
Found 13 items
drwxrwxr-x - root root 36864 2017-10-23 16:20 file:///usr/bin
drwxr-xr-x - root root 4096 2011-09-23 04:50 file:///usr/etc
drwxr-xr-x - root root 4096 2011-09-23 04:50 file:///usr/games
drwxr-xr-x - root root 4096 2017-10-23 09:13 file:///usr/include
drwxr-xr-x - root root 4096 2017-10-23 09:12 file:///usr/java
dr-xr-xr-x - root root 4096 2017-10-23 09:19 file:///usr/lib
dr-xr-xr-x - root root 36864 2017-10-23 09:19 file:///usr/lib64
drwxr-xr-x - root root 12288 2017-10-23 09:13 file:///usr/libexec
drwxr-xr-x - root root 4096 2017-10-23 09:14 file:///usr/local
dr-xr-xr-x - root root 12288 2017-10-23 09:19 file:///usr/sbin
drwxrwxr-x - root root 4096 2017-10-23 09:19 file:///usr/share
drwxr-xr-x - root root 4096 2017-10-23 09:19 file:///usr/src
drwxrwxrwt - root root 4096 2024-05-09 17:50 file:///usr/tmp
[cloudera@quickstart ~]$
```

## hdfs dfs -ls hdfs:///user

```
[cloudera@quickstart ~]$ hdfs dfs -ls hdfs:///user
Found 8 items
drwxr-xr-x - cloudera cloudera 0 2017-10-23 09:14 hdfs:///user/cloudera
drwxr-xr-x - mapred hadoop 0 2017-10-23 09:15 hdfs:///user/history
drwxrwxrwx - hive supergroup 0 2017-10-23 09:17 hdfs:///user/hive
drwxrwxrwx - hue supergroup 0 2017-10-23 09:16 hdfs:///user/hue
drwxrwxrwx - jenkins supergroup 0 2017-10-23 09:15 hdfs:///user/jenkins
drwxrwxrwx - oozie supergroup 0 2017-10-23 09:16 hdfs:///user/oozie
drwxrwxrwx - root supergroup 0 2017-10-23 09:16 hdfs:///user/root
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 hdfs:///user/spark
[cloudera@quickstart ~]$
```

## hdfs dfs -ls /user

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user
Found 8 items
drwxr-xr-x - cloudera cloudera 0 2017-10-23 09:14 /user/cloudera
drwxr-xr-x - mapred hadoop 0 2017-10-23 09:15 /user/history
drwxrwxrwx - hive supergroup 0 2017-10-23 09:17 /user/hive
drwxrwxrwx - hue supergroup 0 2017-10-23 09:16 /user/hue
drwxrwxrwx - jenkins supergroup 0 2017-10-23 09:15 /user/jenkins
drwxrwxrwx - oozie supergroup 0 2017-10-23 09:16 /user/oozie
drwxrwxrwx - root supergroup 0 2017-10-23 09:16 /user/root
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /user/spark
[cloudera@quickstart ~]$
```

## hdfs dfs -ls /

```
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2024-05-09 17:36 /hbase
drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup 0 2024-05-09 17:45 /tmp
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
[cloudera@quickstart ~]$
```

## hdfs dfs -ls

```
[cloudera@quickstart ~]$ hdfs dfs -ls
[cloudera@quickstart ~]$
```

## hdfs dfs -mkdir -p dir1/dir2/dir3

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir -p dir1/dir2/dir3
[cloudera@quickstart ~]$
```

## hdfs dfs -ls dir1

```
[cloudera@quickstart ~]$ hdfs dfs -ls dir1
Found 1 items
drwxr-xr-x - cloudera cloudera 0 2024-05-09 17:56 dir1/dir2
[cloudera@quickstart ~]$
```

hdfs dfs -ls /user/cloudera/dir1

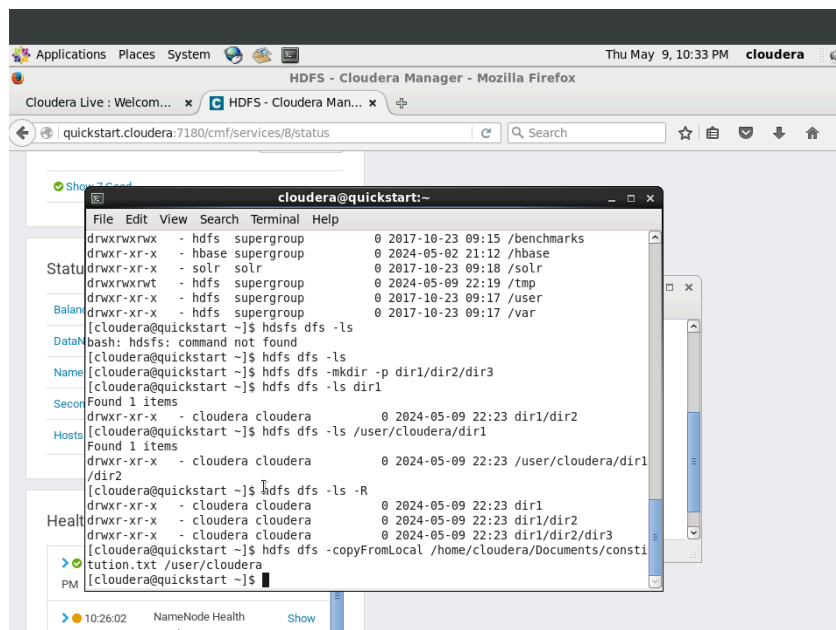
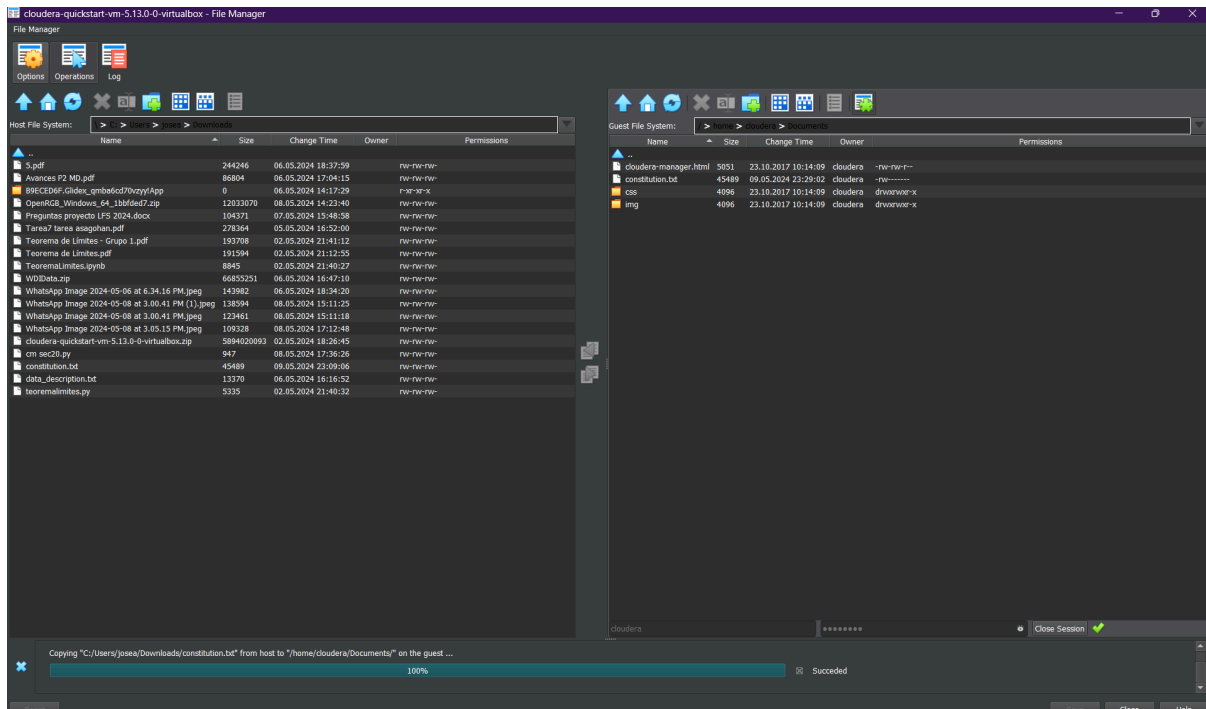
```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/dir1
Found 1 items
drwxr-xr-x - cloudera cloudera 0 2024-05-09 17:56 /user/cloudera/dir1/dir2
[cloudera@quickstart ~]$
```

hdfs dfs -ls -R

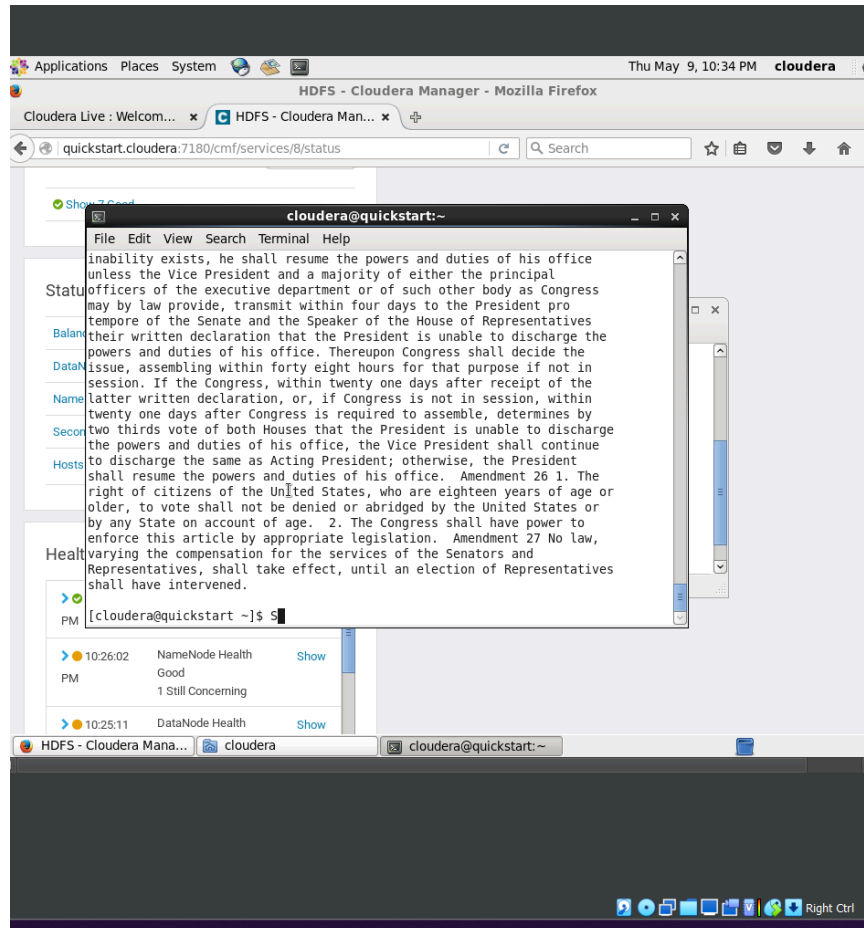
```
[cloudera@quickstart ~]$ hdfs dfs -ls -R
drwxr-xr-x - cloudera cloudera 0 2024-05-09 17:56 dir1
drwxr-xr-x - cloudera cloudera 0 2024-05-09 17:56 dir1/dir2
drwxr-xr-x - cloudera cloudera 0 2024-05-09 17:56 dir1/dir2/di3
[cloudera@quickstart ~]$
```

### Ejercicio 3:

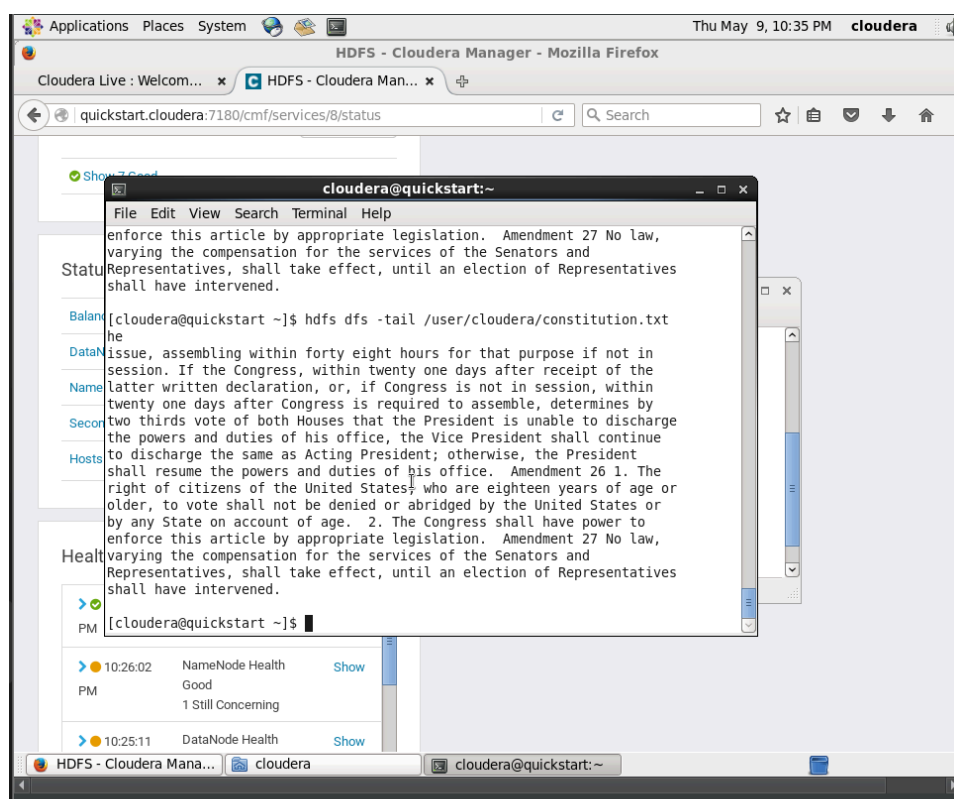
**Copiar el archivo (constitution.txt) del file system local hacia el home directory del usuario cloudera dentro de HDFS. Liste el directorio destino evidenciando que se haya copiado el archivo.**



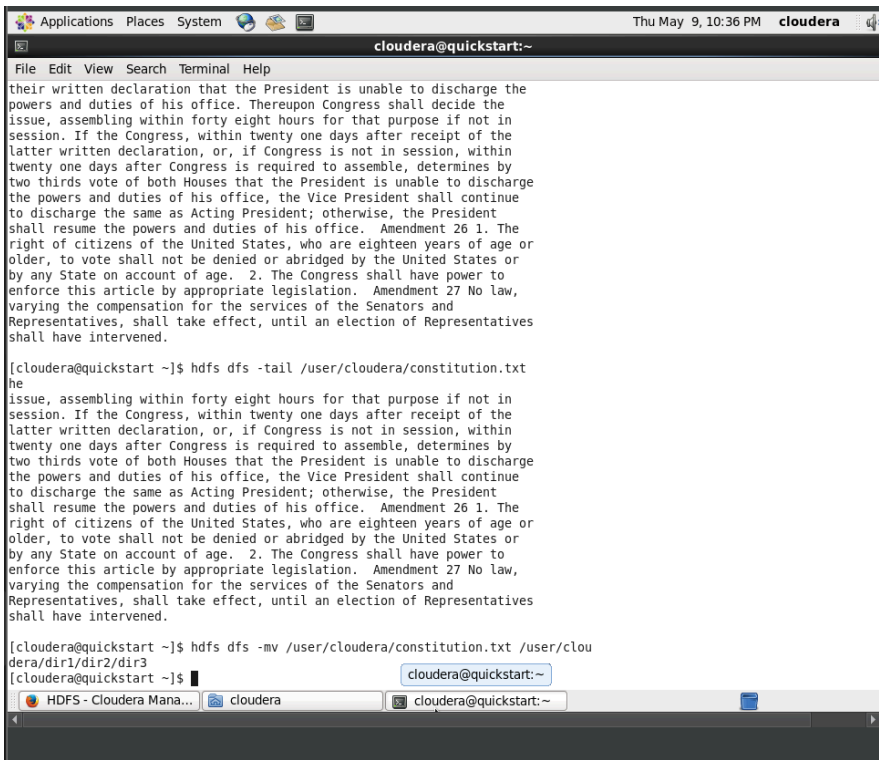
**Despliegue el contenido del archivo constitution.txt que se encuentra en HDFS**



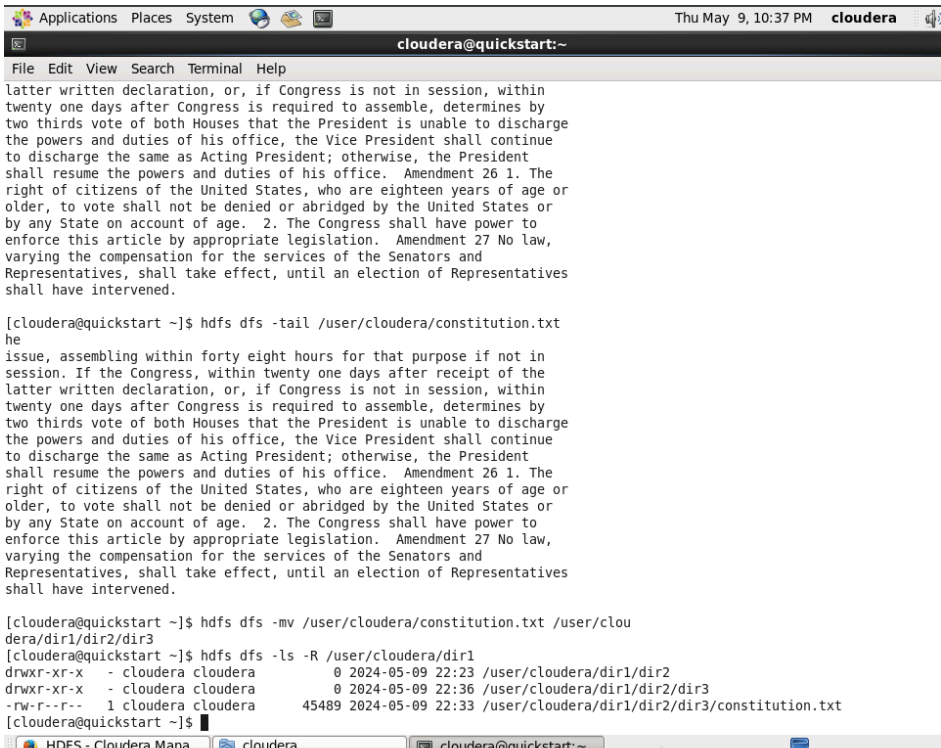
**Despliegue únicamente el final del archivo (1 KB). Use la función tail.**



Mueva el archivo desde su ubicación actual al dir3 que se creó previamente.



Liste el directorio padre de forma recursiva para evidenciar esto





**Ahora, traiga el archivo de regreso al file system local (el origen debería ser el directorio del paso anterior) bajo el nombre de constitution\_download.txt**

```
[cloudera@quickstart ~]$ hdfs dfs -mv /user/cloudera/constitution.txt /user/cloudera/dir1/dir2/dir3
[cloudera@quickstart ~]$ hdfs dfs -ls -R /user/cloudera/dir1
drwxr-xr-x  - cloudera cloudera      0 2024-05-09 22:23 /user/cloudera/dir1/dir2
drwxr-xr-x  - cloudera cloudera      0 2024-05-09 22:36 /user/cloudera/dir1/dir2/dir3
-rw-r--r--  1 cloudera cloudera    45489 2024-05-09 22:33 /user/cloudera/dir1/dir2/dir3/constitution.txt
[cloudera@quickstart ~]$ hdfs dfs -copyToLocal /user/cloudera/dir1/dir2/dir3/constitution.txt /home/cloudera/Documents/constitution_download.txt
[cloudera@quickstart ~]$
```

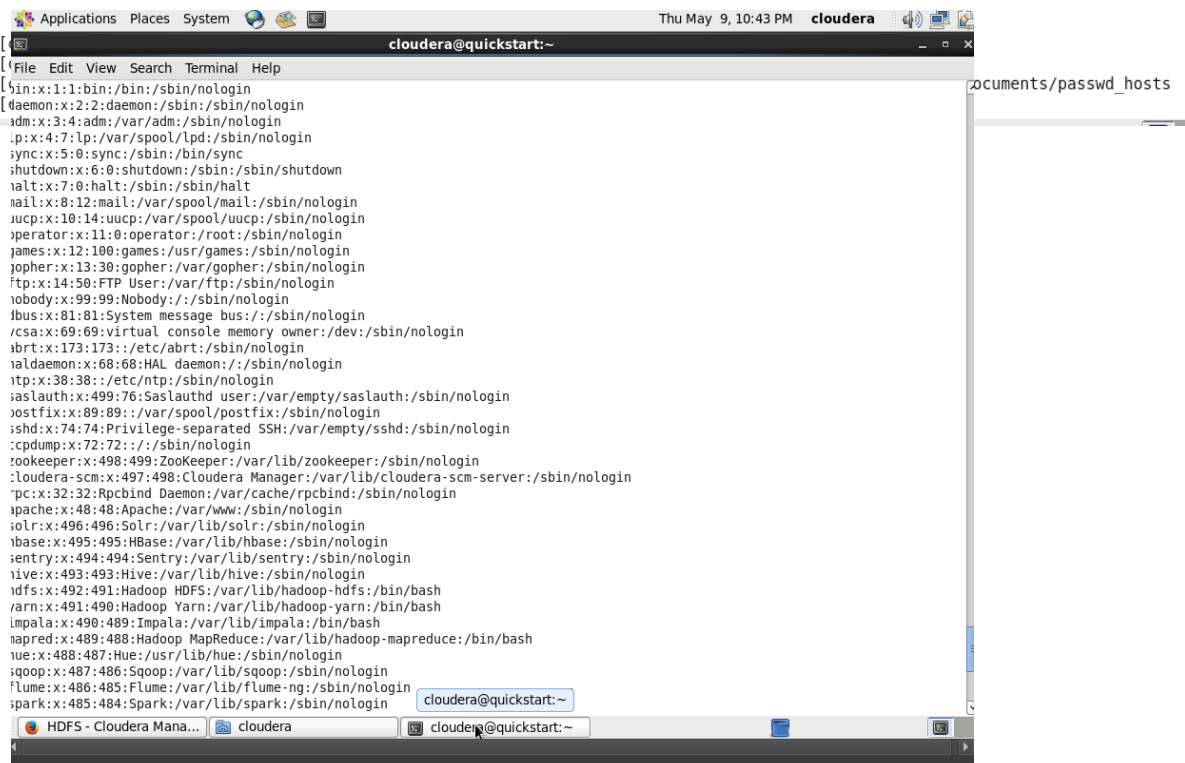
**Copie a HDFS el archivo /etc/passwd en la carpeta del directorio home del usuario cloudera.**

**Copie a HDFS el archivo /etc/hosts en la carpeta del directorio home del usuario cloudera**

tution\_download.txt

```
[cloudera@quickstart ~]$ hdfs dfs -copyFromLocal /etc/passwd /user/cloudera
[cloudera@quickstart ~]$ hdfs dfs -copyFromLocal /etc/hosts /user/cloudera
[cloudera@quickstart ~]$
```

**Utilice el comando -getmerge para copiarlos de vuelta al file system como un único archivo en la carpeta Documents bajo el nombrepasswd\_hosts**



```
Applications  Places  System  Thu May 9, 10:43 PM  cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
[bin:x:1:1:bin:/bin:/sbin/nologin
[daemon:x:2:2:daemon:/sbin:/sbin/nologin
adm:x:3:4:adm:/var/adm:/sbin/nologin
lp:x:4:7:lp:/var/spool/lpd:/sbin/nologin
sync:x:5:0:sync:/sbin:/bin/sync
shutdown:x:6:0:shutdown:/sbin:/sbin/shutdown
halt:x:7:0:halt:/sbin:/sbin/halt
mail:x:8:12:mail:/var/spool/mail:/sbin/nologin
uucp:x:10:14:uucp:/var/spool/uucp:/sbin/nologin
operator:x:11:0:operator:/root:/sbin/nologin
games:x:12:100:games:/usr/games:/sbin/nologin
gopher:x:13:30:gopher:/var/gopher:/sbin/nologin
ftp:x:14:50:FTP User:/var/ftp:/sbin/nologin
nobody:x:99:99:Nobody:/sbin/nologin
dbus:x:81:81:system message bus:/sbin/nologin
cvs:x:69:69:virtual console memory owner:/dev:/sbin/nologin
abrt:x:173:173:/etc/abrt:/sbin/nologin
haldaemon:x:68:68:HAL daemon:/sbin/nologin
ntp:x:38:38:/etc/ntp:/sbin/nologin
saslauthd:x:996:996:Saslauthd user:/var/empty/saslauthd:/sbin/nologin
postfix:x:89:89:/var/spool/postfix:/sbin/nologin
sshd:x:74:74:Privilege-separated SSH:/var/empty/ssh:/sbin/nologin
cpdump:x:72:72:/sbin/nologin
zookeeper:x:498:498:ZooKeeper:/var/lib/zookeeper:/sbin/nologin
cloudera-scm:x:497:498:Cloudera Manager:/var/lib/cloudera-scm-server:/sbin/nologin
rpcbind:x:32:32:Rpcbind Daemon:/var/cache/rpcbind:/sbin/nologin
apache:x:48:48:Apache:/var/www:/sbin/nologin
solr:x:496:496:Solr:/var/lib/solr:/sbin/nologin
hbase:x:495:495:HBase:/var/lib/hbase:/sbin/nologin
sentry:x:494:494:Sentry:/var/lib/sentry:/sbin/nologin
hive:x:493:493:Hive:/var/lib/hive:/sbin/nologin
hadoop:x:492:491:Hadoop HDFS:/var/lib/hadoop-hdfs:/bin/bash
hadoop-yarn:x:491:490:Hadoop Yarn:/var/lib/hadoop-yarn:/bin/bash
impala:x:490:489:Impala:/var/lib/impala:/bin/bash
mapred:x:489:488:Hadoop MapReduce:/var/lib/hadoop-mapreduce:/bin/bash
hue:x:488:487:Hue:/usr/lib/hue:/sbin/nologin
sqoop:x:487:486:Sqoop:/var/lib/sqoop:/sbin/nologin
flume:x:486:485:Flume-ng:/var/lib/flume-ng:/sbin/nologin
spark:x:485:484:Spark:/var/lib/spark:/sbin/nologin
Documents/passwd_hosts
```

```

127.0.0.1      localhost      localhost.domain
10.0.2.15     quickstart.cloudera  quickstart

[cloudera@quickstart ~]$ hdfs dfs -rm /user/cloudera/passwd
24/05/09 22:44:43 INFO fs.TrashPolicyDefault: Moved: 'hdfs://quickstart.cloudera:8020/user/cloudera/passwd' to trash at: hdfs
://quickstart.cloudera:8020/user/cloudera/.Trash/Current/user/cloudera/passwd
[cloudera@quickstart ~]$

```

## Despliegue el contenido del archivo en cuestión en el sistema operativo

```

cloudera-scm:x:497:498:Cloudera Manager:/var/lib/cloudera-scm-server:/sbin/nologin
rpc:x:32:32:Rpcbind Daemon:/var/cache/rpcbind:/sbin/nologin
apache:x:48:48:Apache:/var/www:/sbin/nologin
solr:x:496:496:Solr:/var/lib/solr:/sbin/nologin
hbase:x:495:495:HBase:/var/lib/hbase:/sbin/nologin
sentry:x:494:494:Sentry:/var/lib/sentry:/sbin/nologin
hive:x:493:493:Hive:/var/lib/hive:/sbin/nologin
hdfs:x:492:491:Hadoop HDFS:/var/lib/hadoop-hdfs:/bin/bash
yarn:x:491:490:Hadoop Yarn:/var/lib/hadoop-yarn:/bin/bash
impala:x:490:489:Impala:/var/lib/impala:/bin/bash
mapred:x:489:488:Hadoop MapReduce:/var/lib/hadoop-mapreduce:/bin/bash
hue:x:488:487:Hue:/usr/lib/hue:/sbin/nologin
sqoop:x:487:486:Sqoop:/var/lib/sqoop:/sbin/nologin
flume:x:486:485:Flume:/var/lib/flume-ng:/sbin/nologin
spark:x:485:484:Spark:/var/lib/spark:/sbin/nologin
sqoop2:x:484:486:Sqoop 2 User:/var/lib/sqoop2:/sbin/nologin
oozie:x:483:483:Oozie User:/var/lib/oozie:/bin/false
mysql:x:27:27:MySQL Server:/var/lib/mysql:/bin/bash
kms:x:482:482:Hadoop KMS:/var/lib/hadoop-kms:/bin/bash
llama:x:500:481:Llama:/var/lib/llama:/bin/bash
https:x:481:480:Hadoop HTTPS:/var/lib/hadoop-https:/bin/bash
gdm:x:42:42:./var/lib/gdm:/sbin/nologin
rtkit:x:480:477:RealtimeKit:/proc:/sbin/nologin
pulse:x:479:476:PulseAudio System Daemon:/var/run/pulse:/sbin/nologin
avahi-autoipd:x:170:170:Avahi IPv4LL Stack:/var/lib/avahi-autoipd:/sbin/nologin
cloudera:x:501:501:./home/cloudera:/bin/bash
vboxadd:x:478:1:./var/run/vboxadd:/bin/false
# This file is generated by /usr/bin/cloudera-quickstart-ip, which is invoked
# by /etc/init.d/cloudera-quickstart-init. If you wish to change the way that
# /etc/hosts is generated, you may edit /etc/init.d/cloudera-quickstart-init
# and hard-code a different IP address as a parameter to
# /usr/bin/cloudera-quickstart-ip, or you may comment out that line and manage
# /etc/hosts yourself.

127.0.0.1      localhost      localhost.domain
10.0.2.15     quickstart.cloudera  quickstart


[cloudera@quickstart ~]$

```

Ahora borremos el archivo passwd del directorio home del usuario cloudera dentro de Cloudera Express, luego listemos el directorio para verificar que se ha eliminado.



```
[cloudera@quickstart ~]$ hdfs dfs -rm /user/cloudera/passwd
24/05/09 22:44:43 INFO fs.TrashPolicyDefault: Moved: 'hdfs://quickstart.cloudera:8020/user/cloudera/passwd' to trash at: hdfs
://quickstart.cloudera:8020/user/cloudera/.Trash/Current/user/cloudera/passwd
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera
Found 3 items
drwx----- - cloudera cloudera          0 2024-05-09 22:44 /user/cloudera/.Trash
drwxr-xr-x - cloudera cloudera          0 2024-05-09 22:23 /user/cloudera/dir1
-rw-r--r-- 1 cloudera cloudera        473 2024-05-09 22:41 /user/cloudera/hosts
[cloudera@quickstart ~]$
```



**¿Qué mensaje recibió cuando eliminó el archivo passwd?, y cuando se listó el contenido del directorio home en HDFS ¿notó que ahora hay un nuevo directorio .Trash? Investigue qué significa este directorio y lo que almacena.**

Cuando eliminas un archivo en HDFS, recibirás un mensaje indicando que el archivo fue eliminado exitosamente. La aparición de un nuevo directorio .Trash en el directorio home en HDFS significa que se ha habilitado la papelera de reciclaje y que los archivos eliminados se mueven a este directorio en lugar de ser eliminados permanentemente. Puedes restaurar archivos eliminados de la papelera de reciclaje si es necesario.