

## Big Data and Hadoop - 01

**1. What is the default replication factor of Hadoop cluster?**

- a. 3
- b. 2
- c. 4
- d. 1

Ans: a

**2. Which component in Hadoop Cluster is responsible for serving read and write requests from the file system's clients?**

- a. Name Node
- b. Data Node
- c. Both a & b
- d. None of the above

Ans : b

**3. Which component of Hadoop Cluster manages the file system namespace and regulates access to files by clients?**

- a. Name Node
- b. Data Node
- c. Both a & b
- d. None of the above

Ans : a

**4. If a file size of size 100 MB is stored on HDFS, what would be the split size?**

- a. 64 MB & 64 MB
- b. 64 MB & 36 MB
- c. 100 MB
- d. None of the above

Ans : b

**5. State true or false: MR2 support various MPP modes for data processing?**

- a. FALSE
- b. True

Ans : a

**6. Which comand of HDFS helps copy files from HDFS to Local file system?**

- a. copyFromLocal
- b. copyToLocal
- c. put
- d. mv

Ans : c

**7. Which Eco system component of Hadoop is good for non sql programmers?**

- a. Hive
- b. Hbase
- c. Flume
- d. Pig

Ans : b

**8. Block size of a Hadoop cluster is configurable by Administrator?**

- a. TRUE
- b. FALSE

Ans : a

**9. The functions performed by DataNodes in Hadoop Cluster is/are?**

- a. Data Block Creation
- b. Data Block Deletion
- c. Data Block Replication
- d. All above

Ans : d

**10. Find error in below command:**

`hdfs dfs -put /home/user1/abc.txt`

- a. Target name missing
- b. Source name should include hdfs://
- c. No error

Ans : a

**11. Hadoop block size should be multiple of which unit?**

- a. 32 MB
- b. 50 MB
- c. 64 MB
- d. 70 MB

Ans : a

**12. Which component of the hadoop cluster manages data on slave nodes?**

- a. Name node
- b. Data Node
- c. Task Tracker
- d. Job Tracker

Ans : a

**13. MR1 and MR2 are two modes of processing in Hadoop?**

- a. TRUE
- b. FALSE

Ans : b

**14. What is Hadoop?**

- a. Open source software for reliable, scalable, distributed computing.
- b. A framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
- c. Both a & b
- d. None of the above

Ans : c

**15. Hadoop provides**

- a. A reliable distributed storage and processing system
- b. Only distributed storage
- c. Only processing system
- d. None of the above

Ans : a

1. What is Default replication factor and how will you change it at file level?

Ans: Default replication factor is 3

The replication factor is a property that can be set in the HDFS configuration file `hdfs-site.xml`

2. Why do we need replication factor > 1 in production Hadoop cluster?

Ans: For better data reliability and fault tolerance

3. How will you combine the 4 part-r files of a mapreduce job?

Ans : By using `-getmerge` command

4. What are the Compression techniques in HDFS and which is the best one and why?

Ans : GZIP, BZIP2, LZO, SNAPPY

5. How will you view the compressed files via HDFS command?

Ans :

6. What is Secondary Namenode and its Functionalities? why do we need it?

Ans : It periodically merges the namespace image with the edit log to prevent the edit log from becoming too large.

Requires similar hardware as Namenode machine

Not used for high-availability

Not a backup for Namenode

1. It gets the edit logs from the namenode in regular intervals and applies to fsimage
2. Once it has new fsimage, it copies back to namenode
3. Namenode will use this fsimage for the next restart, which will reduce the startup time

7. What is Backup node and how is it different from Secondary namenode?

Backup node as the name states its main role is to act as the dynamic Backup for the Filesystem Namespace(Metadata) in the Primary Namenode of the [Hadoop Ecosystem](#).

The Backup node implements the Checkpointing functionality along with the online streaming of the File system edits transaction in the Primary Namenode.

8. What is FSImage and editlogs and how they are related?

Ans : FsImage is a file stored on the OS filesystem that contains the complete directory structure (namespace) of the HDFS with details about the location of the data on the Data Blocks and which blocks are stored on which node. This file is used by the NameNode when it is started.

EditLogs is a transaction log that records the changes in the HDFS file system or any action performed on the HDFS cluster such as addition of a new block, replication, deletion etc. In short, it records the changes since the last FsImage was created.

9. what is default block size in HDFS? and why is it so large?

Ans : The default block size in HDFS is 128MB.

1. To minimize the cost of seek: For the large size blocks, time taken to transfer the data from disk can be longer as compared to the time taken to start the block. This results in the transfer of multiple blocks at the disk transfer rate.
2. If blocks are small, there will be too many blocks in Hadoop HDFS and thus too much metadata to store. Managing such a huge number of blocks and metadata will create overhead and lead to traffic in a network.

10. How will you copy a large file of 50GB into HDFS in parallel

Ans : By using `-cp` command

11. what is Balancing in HDFS?

Ans : HDFS provides a balancer utility. This utility analyzes block placement and balances data across the DataNodes. It keeps on moving blocks until the cluster is deemed to be balanced, which means that the utilization of every DataNode is uniform.

12. What is expunge in HDFS ?

Ans : This command is used to empty the trash available in an HDFS system.

Syntax: `$ hadoop fs -expunge`.

## HDFS Task -1

1. What is the Namenode's URI and which file is it configured in?

Ans : URI - Uniform Resource Identifier. It consists of Scheme, Authority and Path.  
format is `scheme://authority/path`

2. Where on a local file system will Namenode store its image and which file is it configured in?

Ans : FsImage

3. Where on a local file system will Datanode store its blocks and which file is it configured in?

Ans : The DataNode stores HDFS data in files in its local file system. The DataNode has no knowledge about HDFS files. It stores each block of HDFS data in a separate file in its local file system. The DataNode does not create all files in the same directory

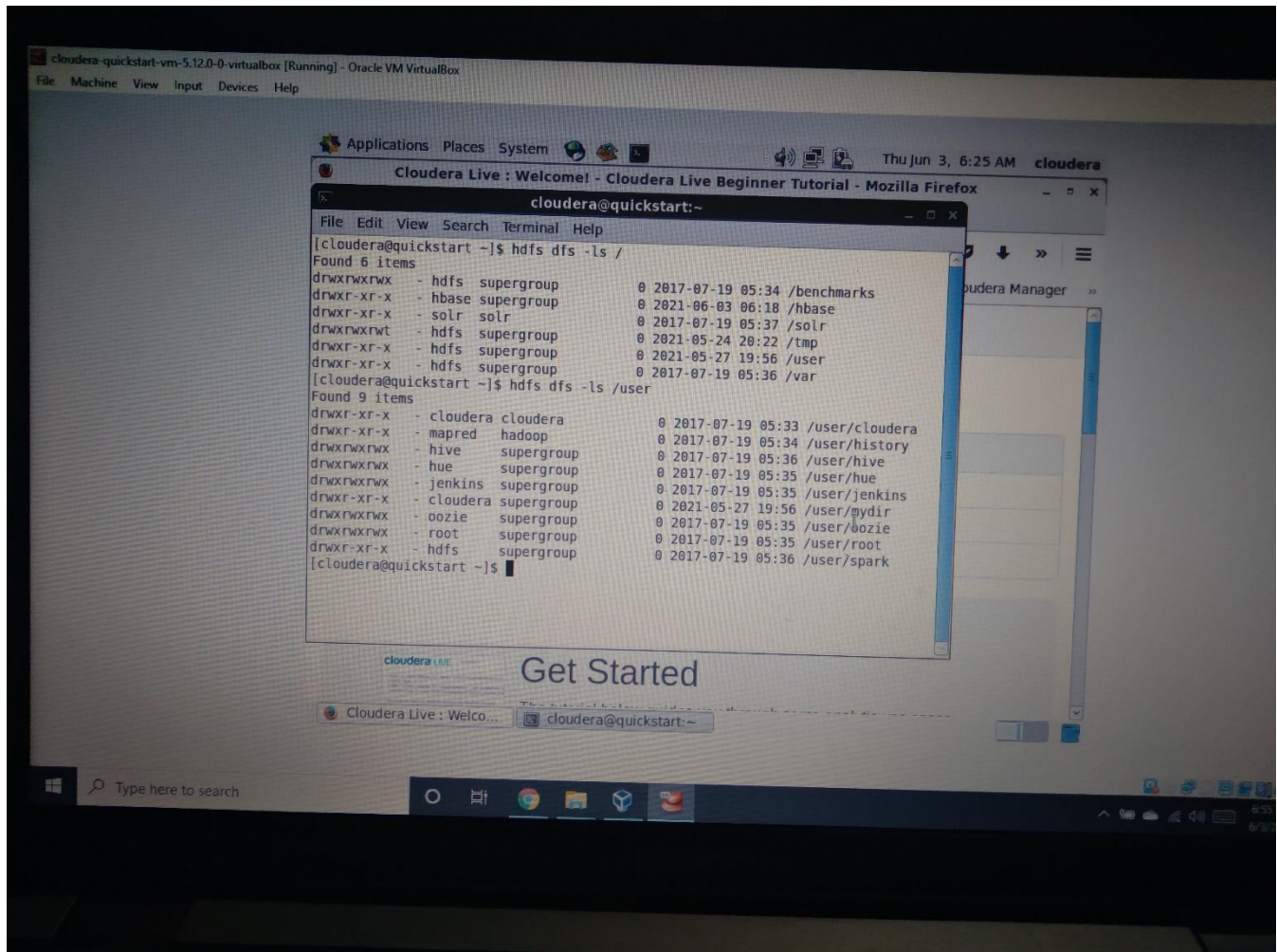
4. What is the block replication and which file is it configured in?

Ans : HDFS stores each file as a sequence of blocks. The blocks of a file are **replicated** for fault tolerance. The NameNode makes all decisions regarding replication of blocks. It periodically receives a Blockreport from each of the DataNodes in the cluster.

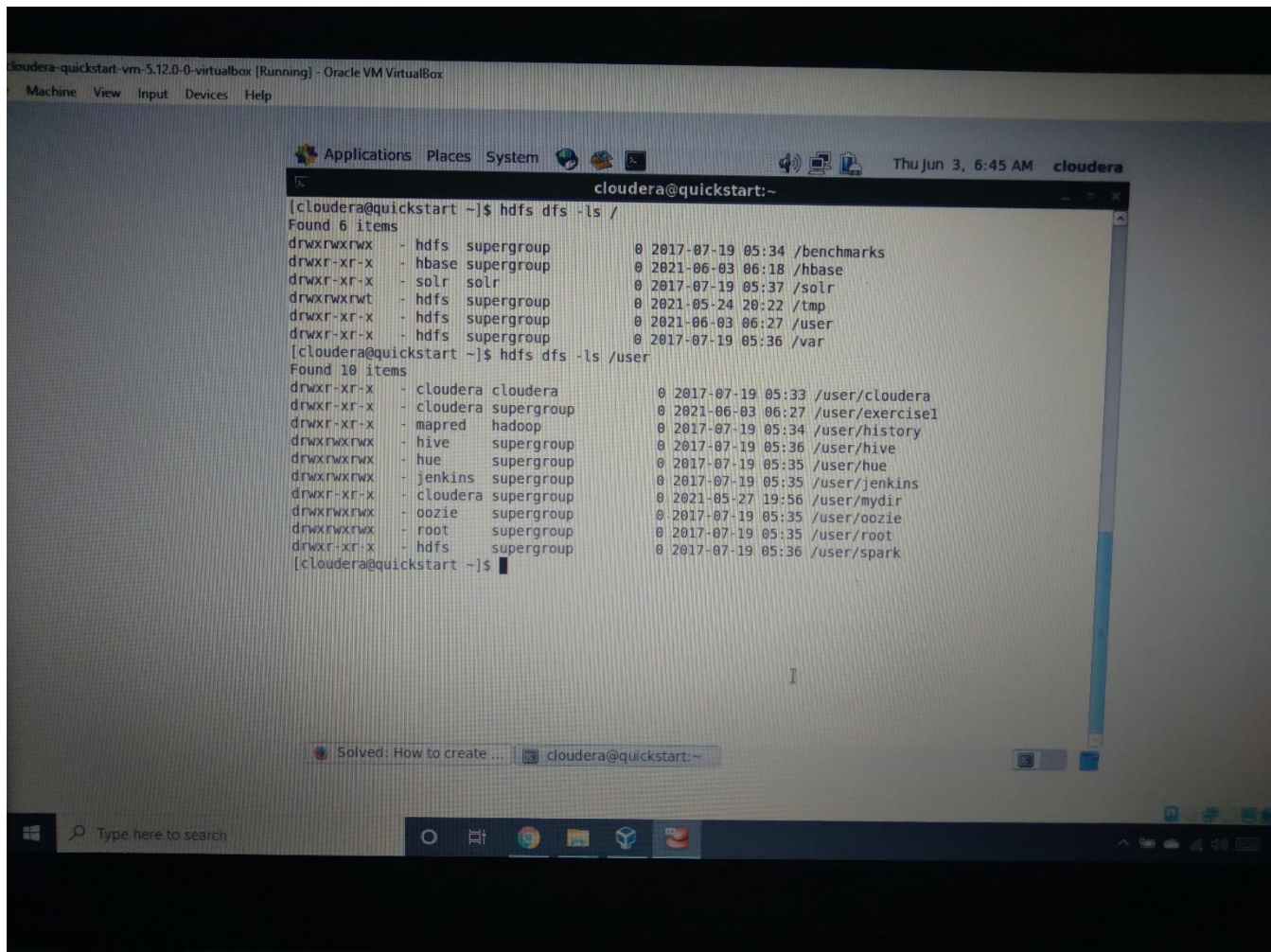
It is configured in hdfs-site.xml file

Task to perform

Verify HDFS is running



Created directory exercise1 and check under user



The screenshot shows a terminal window titled "cloudera@quickstart:~" within an Oracle VM VirtualBox environment. The terminal displays two HDFS commands and their outputs. The first command, `hdfs dfs -ls /`, lists the root directory contents. The second command, `hdfs dfs -ls /user`, lists the contents of the /user directory, which includes a subdirectory named "exercise1".

```
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx - hdfs supergroup          0 2017-07-19 05:34 /benchmarks
drwxr-xr-x - hbase supergroup          0 2021-06-03 06:18 /hbase
drwxr-xr-x - solr solr                  0 2017-07-19 05:37 /solr
drwxrwxrwt - hdfs supergroup          0 2021-05-24 20:22 /tmp
drwxr-xr-x - hdfs supergroup          0 2021-06-03 06:27 /user
drwxr-xr-x - hdfs supergroup          0 2017-07-19 05:36 /var

[cloudera@quickstart ~]$ hdfs dfs -ls /user
Found 10 items
drwxr-xr-x - cloudera cloudera          0 2017-07-19 05:33 /user/cloudera
drwxr-xr-x - cloudera supergroup        0 2021-06-03 06:27 /user/exercise1
drwxr-xr-x - mapred hadoop              0 2017-07-19 05:34 /user/history
drwxrwxrwx - hive supergroup            0 2017-07-19 05:36 /user/hive
drwxrwxrwx - hue supergroup             0 2017-07-19 05:35 /user/hue
drwxrwxrwx - jenkins supergroup         0 2017-07-19 05:35 /user/jenkins
drwxr-xr-x - cloudera supergroup        0 2021-05-27 19:56 /user/mydir
drwxrwxrwx - oozie supergroup           0 2017-07-19 05:35 /user/oozie
drwxr-xr-x - root supergroup            0 2017-07-19 05:35 /user/root
drwxr-xr-x - hdfs supergroup            0 2017-07-19 05:36 /user/spark

[cloudera@quickstart ~]$
```

Upload deckofcards.txt to directory exercise1 under user



```
cloudera@quickstart:~$ find /user -type d -ls
Found 6 items
drwxrwxrwx - hdfc supergroup          0 2017-07-19 05:34 /benchmarks
drwxr-xr-x - hbase supergroup         0 2021-06-03 06:18 /hbase
drwxr-xr-x - solr solr                 0 2017-07-19 05:37 /solr
drwxrwxrwt - hdfc supergroup          0 2021-05-24 20:22 /tmp
drwxr-xr-x - hdfc supergroup          0 2021-06-03 06:27 /user
drwxr-xr-x - hdfc supergroup          0 2017-07-19 05:36 /var
[cloudera@quickstart ~]$ hdfs dfs -ls /user
Found 10 items
drwxr-xr-x - cloudera cloudera          0 2017-07-19 05:33 /user/cloudera
drwxr-xr-x - cloudera supergroup        0 2021-06-03 06:27 /user/exercisel
drwxr-xr-x - mapred hadoop              0 2017-07-19 05:34 /user/history
drwxrwxrwx - hive supergroup           0 2017-07-19 05:36 /user/hive
drwxrwxrwx - hue supergroup            0 2017-07-19 05:35 /user/hue
drwxrwxrwx - jenkins supergroup        0 2017-07-19 05:35 /user/jenkins
drwxr-xr-x - cloudera supergroup        0 2021-05-27 19:56 /user/mydir
drwxrwxrwx - oozie supergroup          0 2017-07-19 05:35 /user/oozie
drwxrwxrwx - root supergroup           0 2017-07-19 05:35 /user/root
drwxr-xr-x - hdfc supergroup           0 2017-07-19 05:36 /user/spark
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/Documents/deckofcards.txt/user/exercisel
put: '/home/cloudera/Documents/deckofcards.txt/user/exercisel': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -ls /user/exercisel
[cloudera@quickstart ~]$ hdfs dfs -ls /user/mydir
[cloudera@quickstart ~]$ hdfs dfs -cat /user/exercisel/deckofcards.txt
cat: '/user/exercisel/deckofcards.txt': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/Documents/deckofcards.txt /user/exercisel
[cloudera@quickstart ~]$ hdfs dfs -ls /user/exercisel
Found 1 items
-rw-r--r-- 1 cloudera supergroup      693 2021-06-03 06:56 /user/exercisel/deckofcards.txt
[cloudera@quickstart ~]$
```

```
Applications Places System Thu Jun 3, 7:25 AM cloudera
cloudera@quickstart:~
[cloudera@quickstart ~]$ hdfs dfs -get /user/exercisel/deckofcards.txt /home/cloudera/Downloads/deckofcards.Copy.txt
[cloudera@quickstart ~]$ hdfs dfs fsck /user/exercisel/deckofcards.txt -blocks
fsck: Unknown command
[cloudera@quickstart ~]$ hdfs fsck /user/exercisel/deckofcards.txt -blocks
Connecting to namenode via http://quickstart.cloudera:50070/fsck?ugi=cloudera&blocks=1&path=%2Fuser%2Fexercisel%2Fdeckofcards.txt
FSCK started by cloudera (auth:SIMPLE) from /127.0.0.1 for path /user/exercisel/deckofcards.txt at Thu Jun 03 07:25:28 PDT 2021
Status: HEALTHY
Total size: 693 B
Total dirs: 0
Total files: 1
Total symlinks: 0
Total blocks (validated): 1 (avg. block size 693 B)
Minimally replicated blocks: 1 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Thu Jun 03 07:25:28 PDT 2021 in 7 milliseconds

The filesystem under path '/user/exercisel/deckofcards.txt' is HEALTHY
[cloudera@quickstart ~]$
```

Print first 25 lines from deckofcards.txt

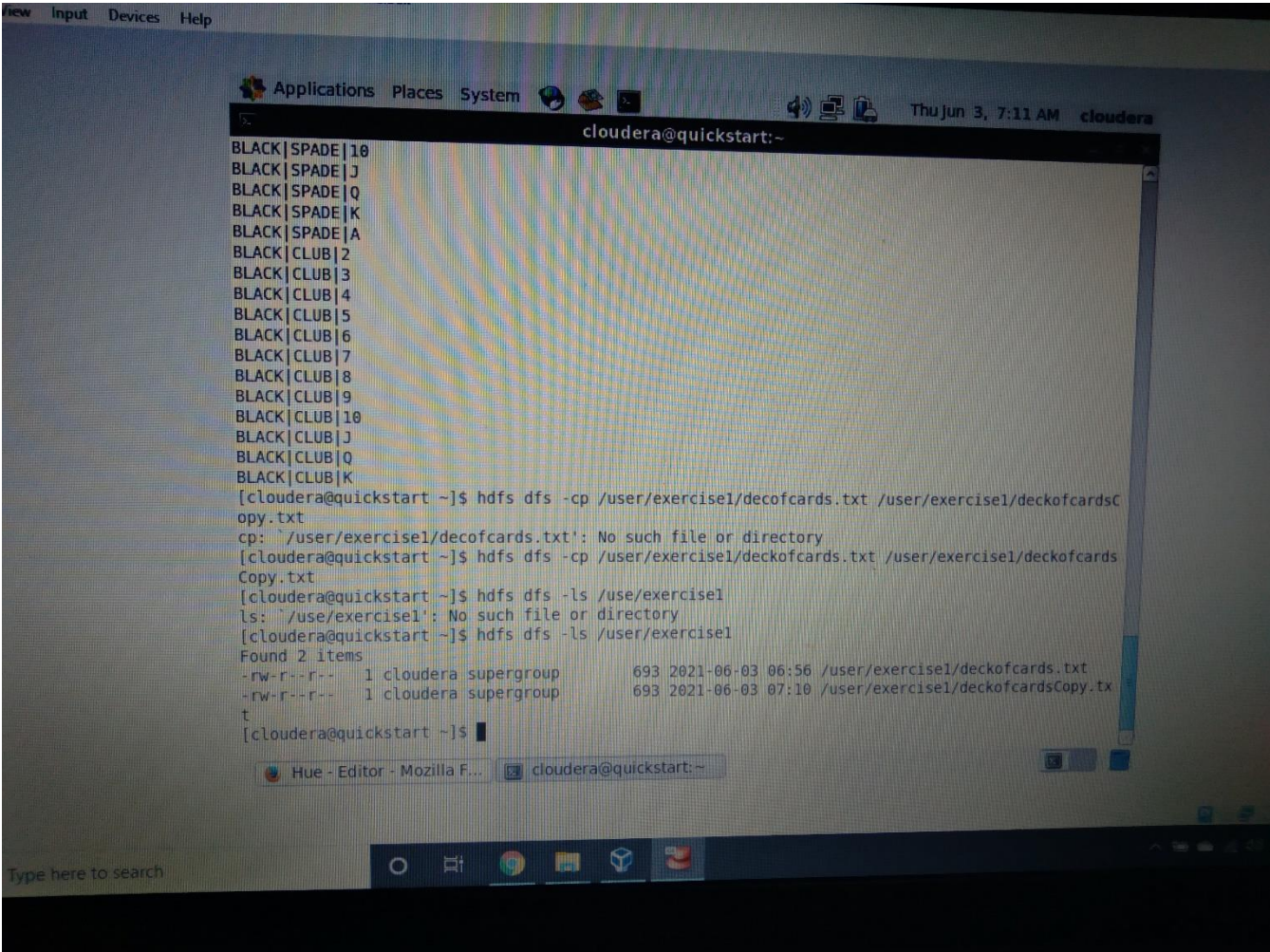


cloudera@quickstart:~

```
RED|HEART|Q
RED|HEART|K
RED|HEART|A
[cloudera@quickstart ~]$ hdfs dfs -cat /user/exercisel/deckofcards.txt | head -25
BLACK|SPADE|2
BLACK|SPADE|3
BLACK|SPADE|4
BLACK|SPADE|5
BLACK|SPADE|6
BLACK|SPADE|7
BLACK|SPADE|8
BLACK|SPADE|9
BLACK|SPADE|10
BLACK|SPADE|J
BLACK|SPADE|Q
BLACK|SPADE|K
BLACK|SPADE|A
BLACK|CLUB|2
BLACK|CLUB|3
BLACK|CLUB|4
BLACK|CLUB|5
BLACK|CLUB|6
BLACK|CLUB|7
BLACK|CLUB|8
BLACK|CLUB|9
BLACK|CLUB|10
BLACK|CLUB|J
BLACK|CLUB|Q
BLACK|CLUB|K
[cloudera@quickstart ~]$
```

Hue - Editor - Mozilla F... cloudera@quickstart:~

Copy file to new file

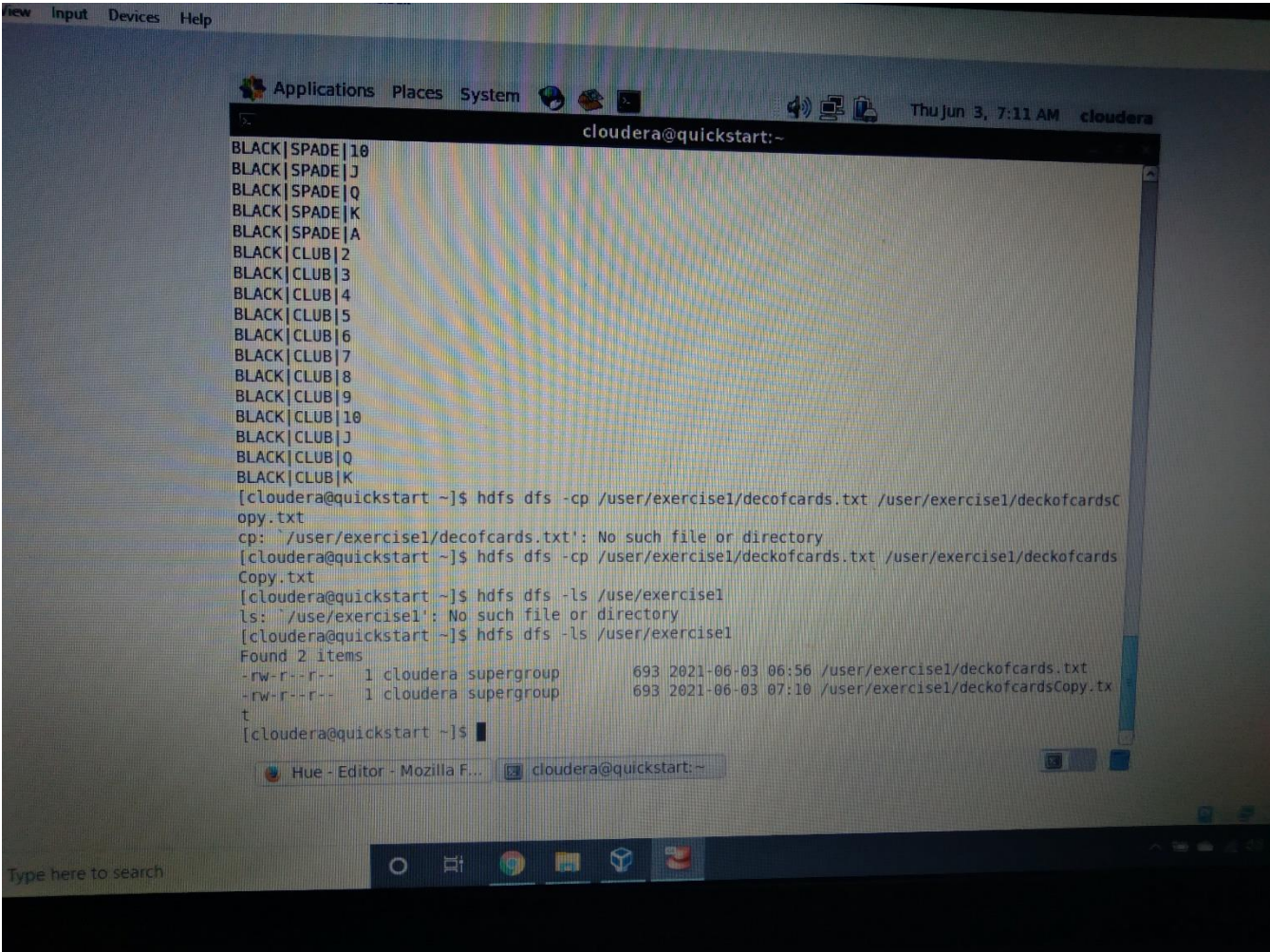


The screenshot shows a terminal window titled "cloudera@quickstart:~". The terminal displays a list of playing cards, followed by a series of commands and their outputs. The commands attempt to copy a file from "/user/exercisel/decofcards.txt" to "/user/exercisel/deckofcardsCopy.txt". The first attempt fails with the error "cp: '/user/exercisel/decofcards.txt': No such file or directory". The second attempt also fails with the same error. The third command is "hdfs dfs -ls /user/exercisel", which outputs "ls: '/use/exercisel': No such file or directory". The fourth command is "hdfs dfs -ls /user/exercisel", which outputs "Found 2 items" and a table of file information.

```
BLACK|SPADE|10
BLACK|SPADE|J
BLACK|SPADE|Q
BLACK|SPADE|K
BLACK|SPADE|A
BLACK|CLUB|2
BLACK|CLUB|3
BLACK|CLUB|4
BLACK|CLUB|5
BLACK|CLUB|6
BLACK|CLUB|7
BLACK|CLUB|8
BLACK|CLUB|9
BLACK|CLUB|10
BLACK|CLUB|J
BLACK|CLUB|Q
BLACK|CLUB|K
[cloudera@quickstart ~]$ hdfs dfs -cp /user/exercisel/decofcards.txt /user/exercisel/deckofcardsCopy.txt
cp: '/user/exercisel/decofcards.txt': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -cp /user/exercisel/deckofcards.txt /user/exercisel/deckofcardsCopy.txt
[cloudera@quickstart ~]$ hdfs dfs -ls /use/exercisel
ls: '/use/exercisel': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -ls /user/exercisel
Found 2 items
-rw-r--r-- 1 cloudera supergroup 693 2021-06-03 06:56 /user/exercisel/deckofcards.txt
-rw-r--r-- 1 cloudera supergroup 693 2021-06-03 07:10 /user/exercisel/deckofcardsCopy.txt
[cloudera@quickstart ~]$
```

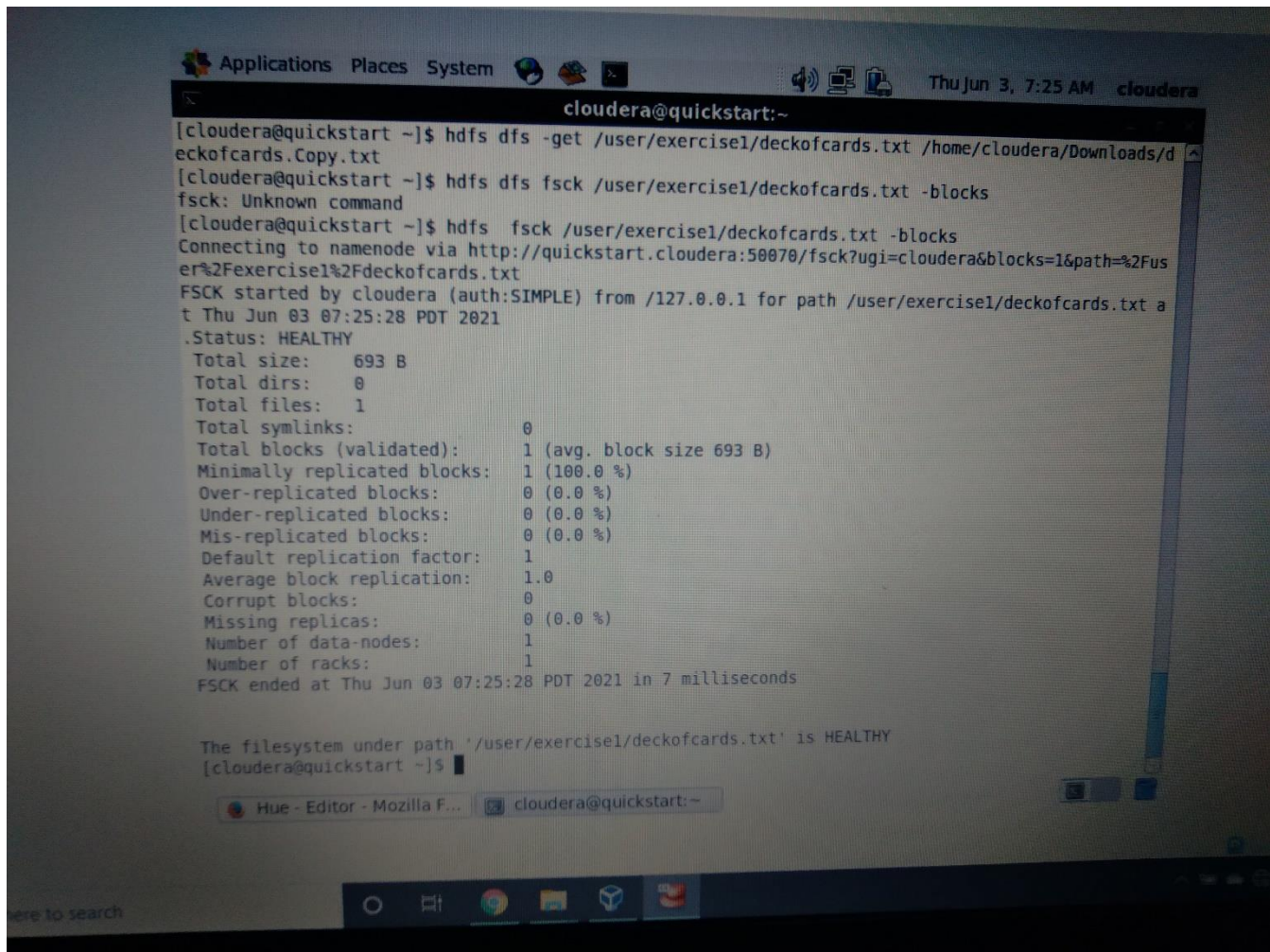


Copy deckofcards.txt to deckofcardsCopy.txt



```
new Input Devices Help
Applications Places System cloudera@quickstart:~
BLACK|SPADE|10
BLACK|SPADE|J
BLACK|SPADE|Q
BLACK|SPADE|K
BLACK|SPADE|A
BLACK|CLUB|2
BLACK|CLUB|3
BLACK|CLUB|4
BLACK|CLUB|5
BLACK|CLUB|6
BLACK|CLUB|7
BLACK|CLUB|8
BLACK|CLUB|9
BLACK|CLUB|10
BLACK|CLUB|J
BLACK|CLUB|Q
BLACK|CLUB|K
[cloudera@quickstart ~]$ hdfs dfs -cp /user/exercisel/decofcards.txt /user/exercisel/deckofcardsCopy.txt
cp: '/user/exercisel/decofcards.txt': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -cp /user/exercisel/deckofcards.txt /user/exercisel/deckofcardsCopy.txt
[cloudera@quickstart ~]$ hdfs dfs -ls /user/exercisel
ls: '/use/exercisel': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -ls /user/exercisel
Found 2 items
-rw-r--r-- 1 cloudera supergroup 693 2021-06-03 06:56 /user/exercisel/deckofcards.txt
-rw-r--r-- 1 cloudera supergroup 693 2021-06-03 07:10 /user/exercisel/deckofcardsCopy.txt
[cloudera@quickstart ~]$
```

Copy deckofcards.txt back to local file system and name it deckofcards.copy.txt and check using FSCK

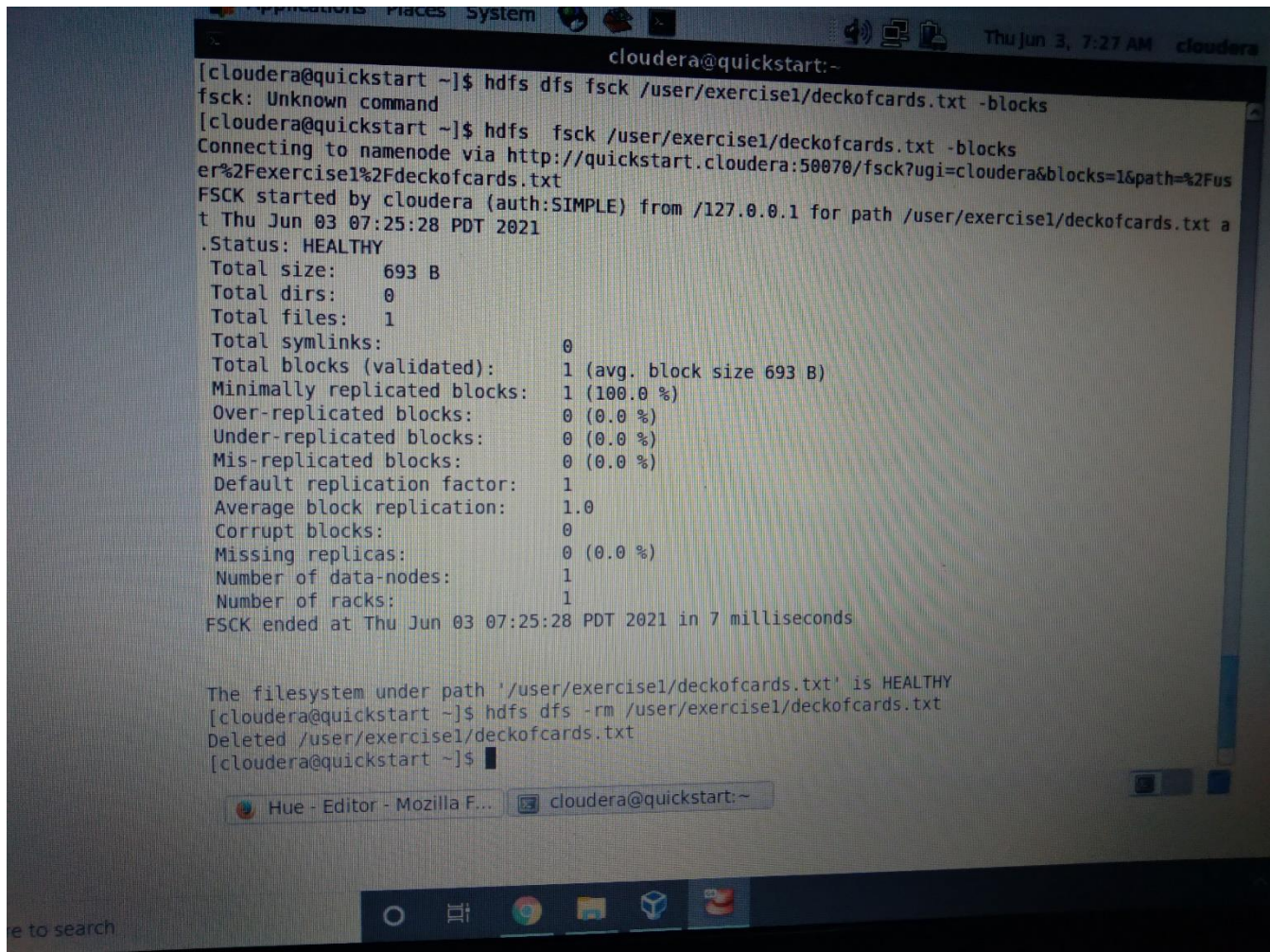


```
cloudera@quickstart:~$ hdfs dfs -get /user/exercisel/deckofcards.txt /home/cloudera/Downloads/deckofcards.Copy.txt
[cloudera@quickstart ~]$ hdfs dfs fsck /user/exercisel/deckofcards.txt -blocks
fsck: Unknown command
[cloudera@quickstart ~]$ hdfs fsck /user/exercisel/deckofcards.txt -blocks
Connecting to namenode via http://quickstart.cloudera:50070/fsck?ugi=cloudera&blocks=1&path=%2Fuser%2Fexercisel%2Fdeckofcards.txt
FSCK started by cloudera (auth:SIMPLE) from /127.0.0.1 for path /user/exercisel/deckofcards.txt at Thu Jun 03 07:25:28 PDT 2021
.Status: HEALTHY
Total size:      693 B
Total dirs:      0
Total files:      1
Total symlinks:    0
Total blocks (validated): 1 (avg. block size 693 B)
Minimally replicated blocks: 1 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Thu Jun 03 07:25:28 PDT 2021 in 7 milliseconds

The filesystem under path '/user/exercisel/deckofcards.txt' is HEALTHY
[cloudera@quickstart ~]$
```



Delete deckofcards.txt from HDFS



The screenshot shows a terminal window titled "cloudera@quickstart:~" with the following output:

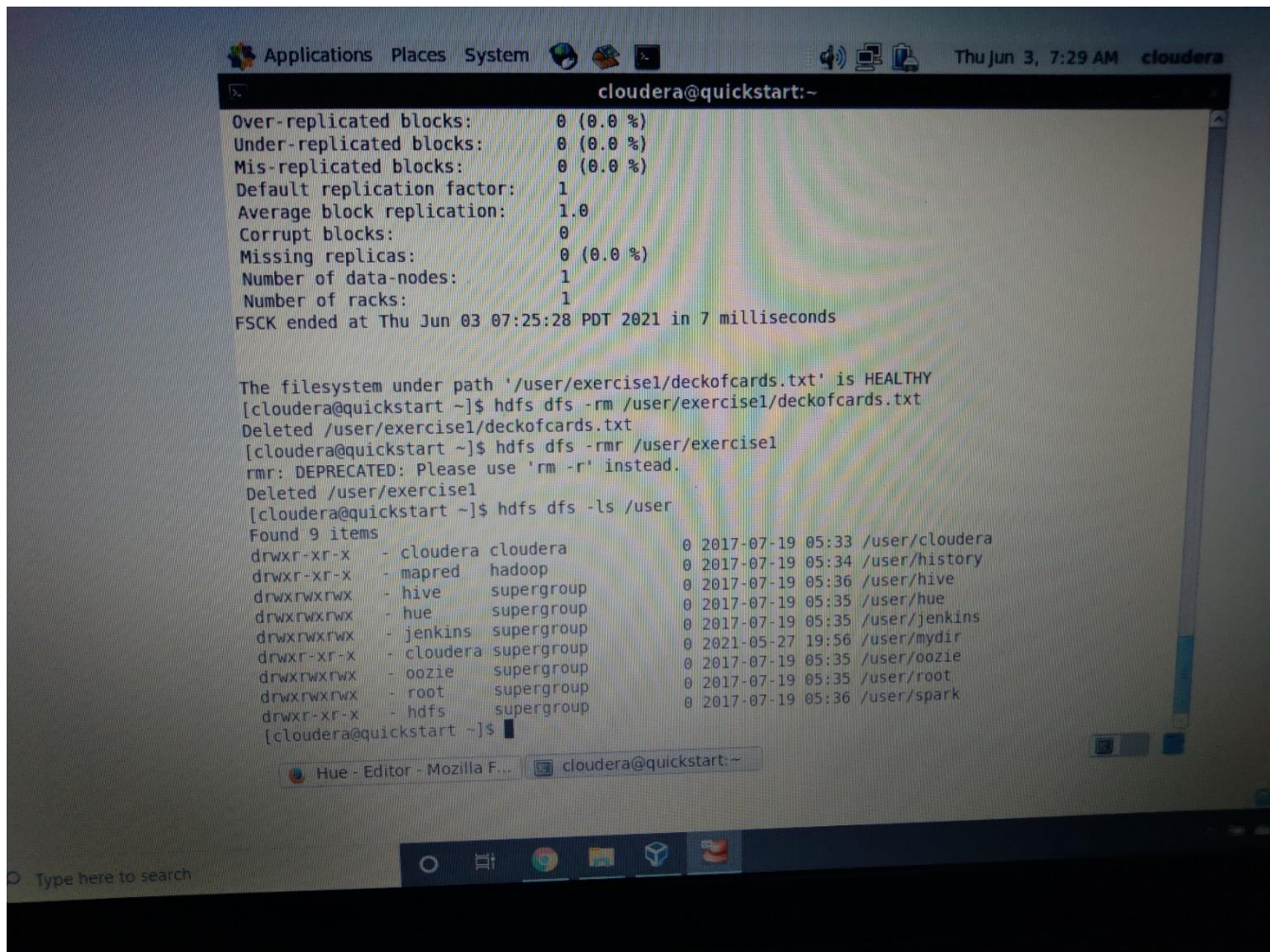
```
[cloudera@quickstart ~]$ hdfs dfs fsck /user/exercisel/deckofcards.txt -blocks
fsck: Unknown command
[cloudera@quickstart ~]$ hdfs fsck /user/exercisel/deckofcards.txt -blocks
Connecting to namenode via http://quickstart.cloudera:50070/fsck?ugi=cloudera&blocks=1&path=%2Fuser%2Fexercisel%2Fdeckofcards.txt
FSCK started by cloudera (auth:SIMPLE) from /127.0.0.1 for path /user/exercisel/deckofcards.txt at Thu Jun 03 07:25:28 PDT 2021
. Status: HEALTHY
Total size:      693 B
Total dirs:      0
Total files:     1
Total symlinks:   0
Total blocks (validated): 1 (avg. block size 693 B)
Minimally replicated blocks: 1 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Thu Jun 03 07:25:28 PDT 2021 in 7 milliseconds

The filesystem under path '/user/exercisel/deckofcards.txt' is HEALTHY
[cloudera@quickstart ~]$ hdfs dfs -rm /user/exercisel/deckofcards.txt
Deleted /user/exercisel/deckofcards.txt
[cloudera@quickstart ~]$
```

The terminal window has a taskbar at the bottom with icons for "Hue - Editor - Mozilla F..." and "cloudera@quickstart:~". The system clock in the top right corner shows "Thu Jun 3, 7:27 AM cloudera".



## Remove Directory exercise1 form HDFS



```
Applications Places System Thu Jun 3, 7:29 AM cloudera
cloudera@quickstart:~
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Thu Jun 03 07:25:28 PDT 2021 in 7 milliseconds

The filesystem under path '/user/exercise1/deckofcards.txt' is HEALTHY
[cloudera@quickstart ~]$ hdfs dfs -rm /user/exercise1/deckofcards.txt
Deleted /user/exercise1/deckofcards.txt
[cloudera@quickstart ~]$ hdfs dfs -rmr /user/exercise1
rmr: DEPRECATED: Please use 'rm -r' instead.
Deleted /user/exercise1
[cloudera@quickstart ~]$ hdfs dfs -ls /user
Found 9 items
drwxr-xr-x - cloudera cloudera 0 2017-07-19 05:33 /user/cloudera
drwxr-xr-x - mapred hadoop 0 2017-07-19 05:34 /user/history
drwxrwxrwx - hive supergroup 0 2017-07-19 05:36 /user/hive
drwxrwxrwx - hue supergroup 0 2017-07-19 05:35 /user/hue
drwxrwxrwx - jenkins supergroup 0 2017-07-19 05:35 /user/jenkins
drwxr-xr-x - cloudera supergroup 0 2021-05-27 19:56 /user/mydir
drwxrwxrwx - oozie supergroup 0 2017-07-19 05:35 /user/oozie
drwxrwxrwx - root supergroup 0 2017-07-19 05:35 /user/root
drwxr-xr-x - hdfs supergroup 0 2017-07-19 05:36 /user/spark
[cloudera@quickstart ~]$
```

