

Big Data GHW – 1

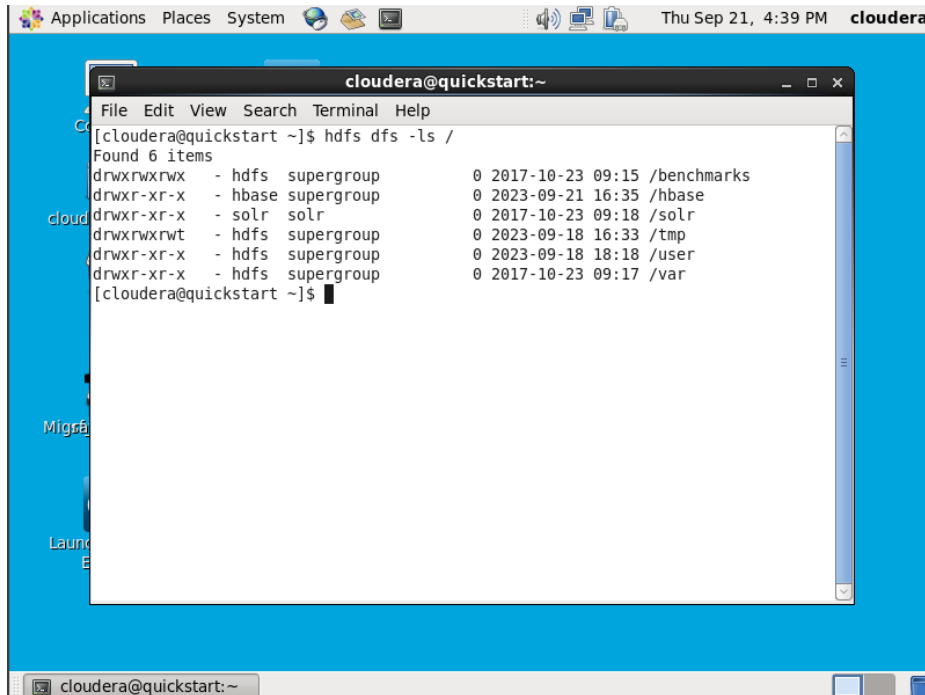
Name: Sukriti Macker

NETID: sm11017

Course: Big Data Section – D

Semester: Fall 2023

1. Listing directory and files: `hdfs dfs -ls /`

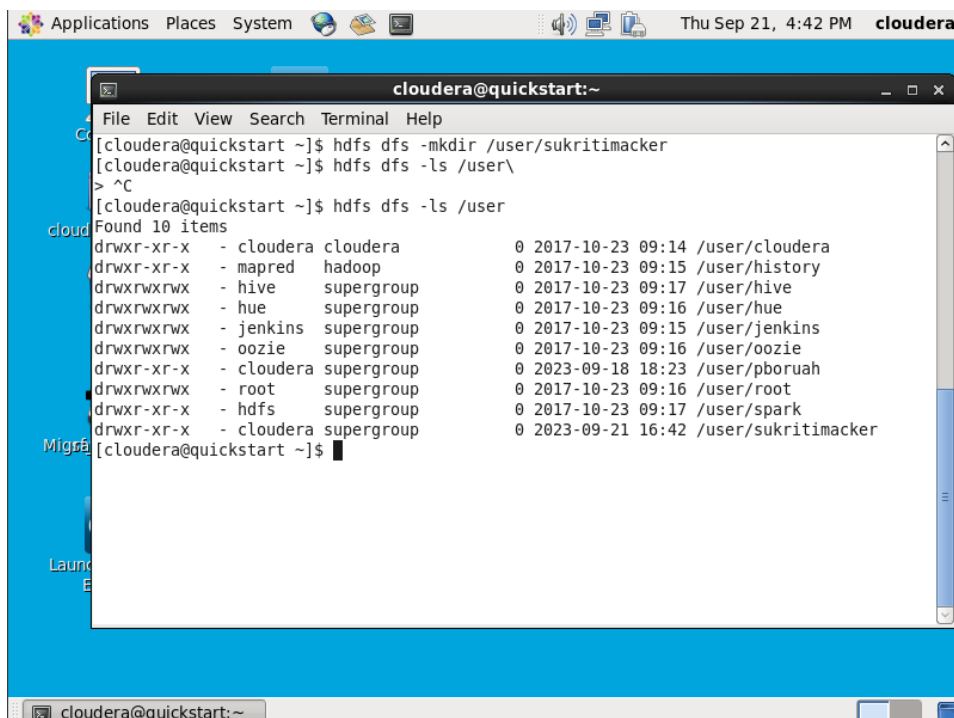


```
cloudera@quickstart:~$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx - hdfs supergroup      0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup      0 2023-09-21 16:35 /hbase
drwxr-xr-x - solr solr              0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup      0 2023-09-18 16:33 /tmp
drwxr-xr-x - hdfs supergroup      0 2023-09-18 18:18 /user
drwxr-xr-x - hdfs supergroup      0 2017-10-23 09:17 /var
cloudera@quickstart ~]$
```

2. Create new directory under /user:-

`hdfs dfs -mkdir`

`hdfs dfs -ls /user`

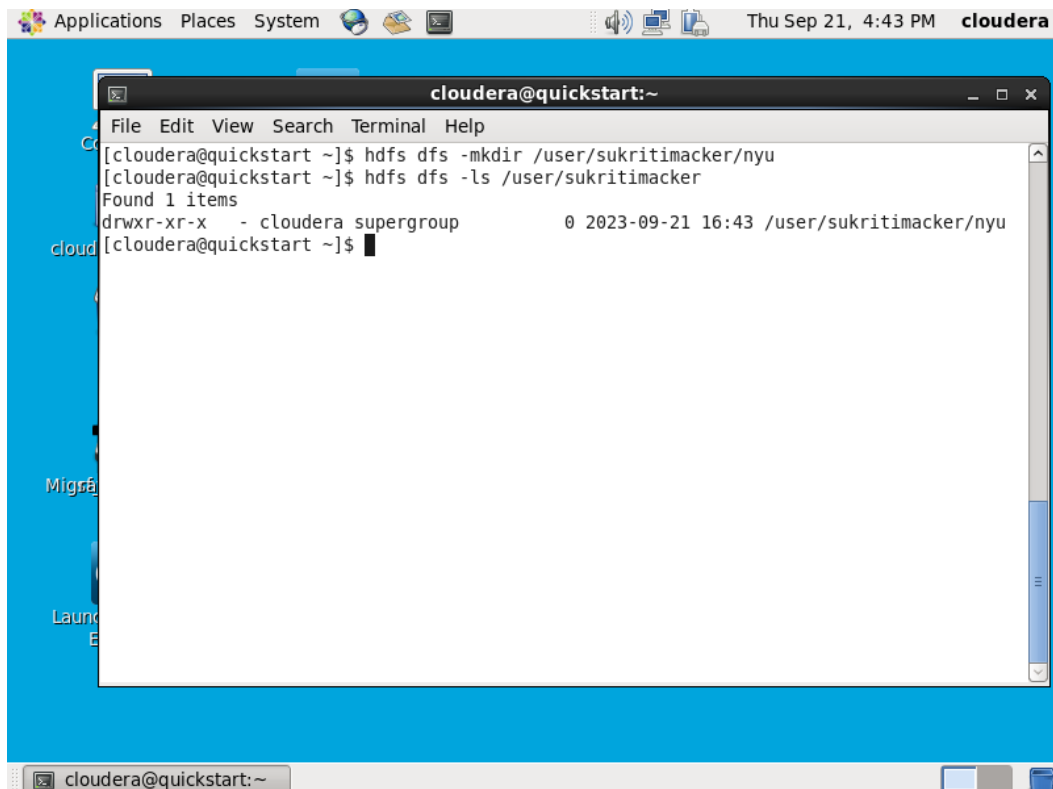


```
cloudera@quickstart:~$ hdfs dfs -mkdir /user/sukritimacker
cloudera@quickstart:~$ hdfs dfs -ls /user\
> ^C
cloudera@quickstart:~$ hdfs dfs -ls /user
Found 10 items
drwxr-xr-x - cloudera cloudera      0 2017-10-23 09:14 /user/cloudera
drwxr-xr-x - mapred hadoop          0 2017-10-23 09:15 /user/history
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
drwxr-xr-x - cloudera supergroup    0 2023-09-18 18:23 /user/pboruah
drwxrwxrwx - root supergroup        0 2017-10-23 09:16 /user/root
drwxr-xr-x - hdfs supergroup        0 2017-10-23 09:17 /user/spark
drwxr-xr-x - cloudera supergroup    0 2023-09-21 16:42 /user/sukritimacker
cloudera@quickstart ~]$
```

3) Nested destination directories:-

```
hdfs dfs -mkdir /user/sukritimacker/nyu
```

```
hdfs dfs -ls /user/sukritimacker
```

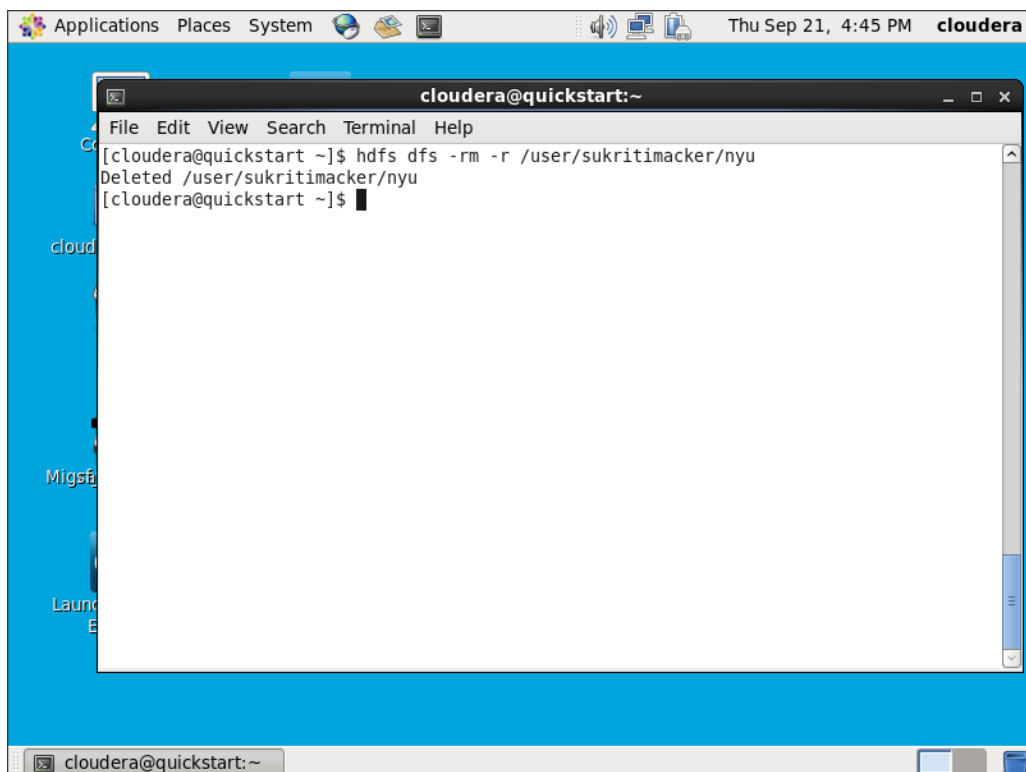


The screenshot shows a terminal window titled "cloudera@quickstart:~" with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal displays the following commands and output:

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/sukritimacker/nyu
[cloudera@quickstart ~]$ hdfs dfs -ls /user/sukritimacker
Found 1 items
drwxr-xr-x  - cloudera supergroup          0 2023-09-21 16:43 /user/sukritimacker/nyu
[cloudera@quickstart ~]$
```

4) Remove dir:

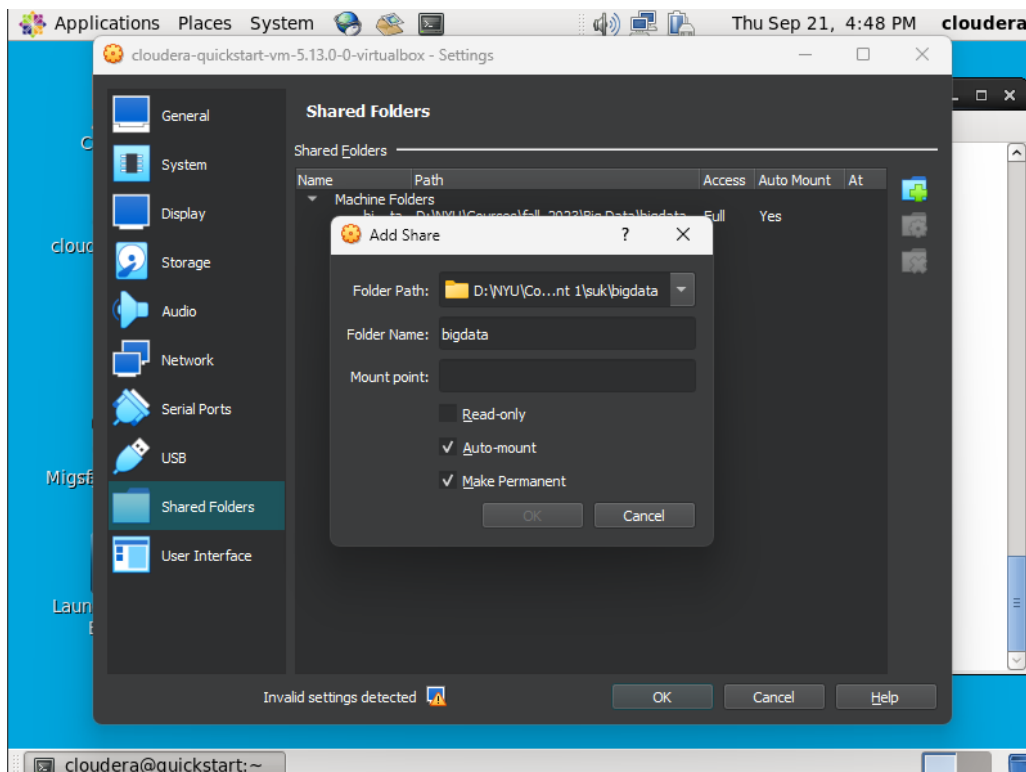
```
hdfs dfs -rm -r /user/sukritimacker/nyu
```



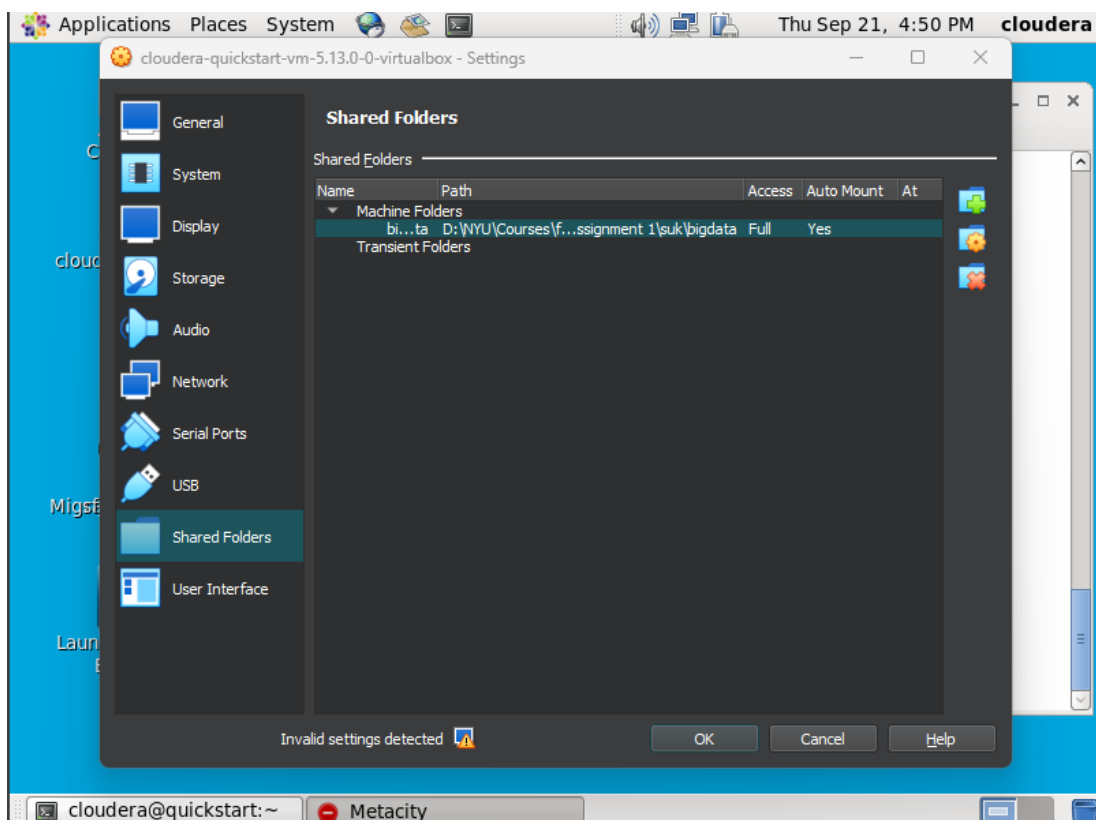
The screenshot shows a terminal window titled "cloudera@quickstart:~" with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal displays the following commands and output:

```
[cloudera@quickstart ~]$ hdfs dfs -rm -r /user/sukritimacker/nyu
Deleted /user/sukritimacker/nyu
[cloudera@quickstart ~]$
```

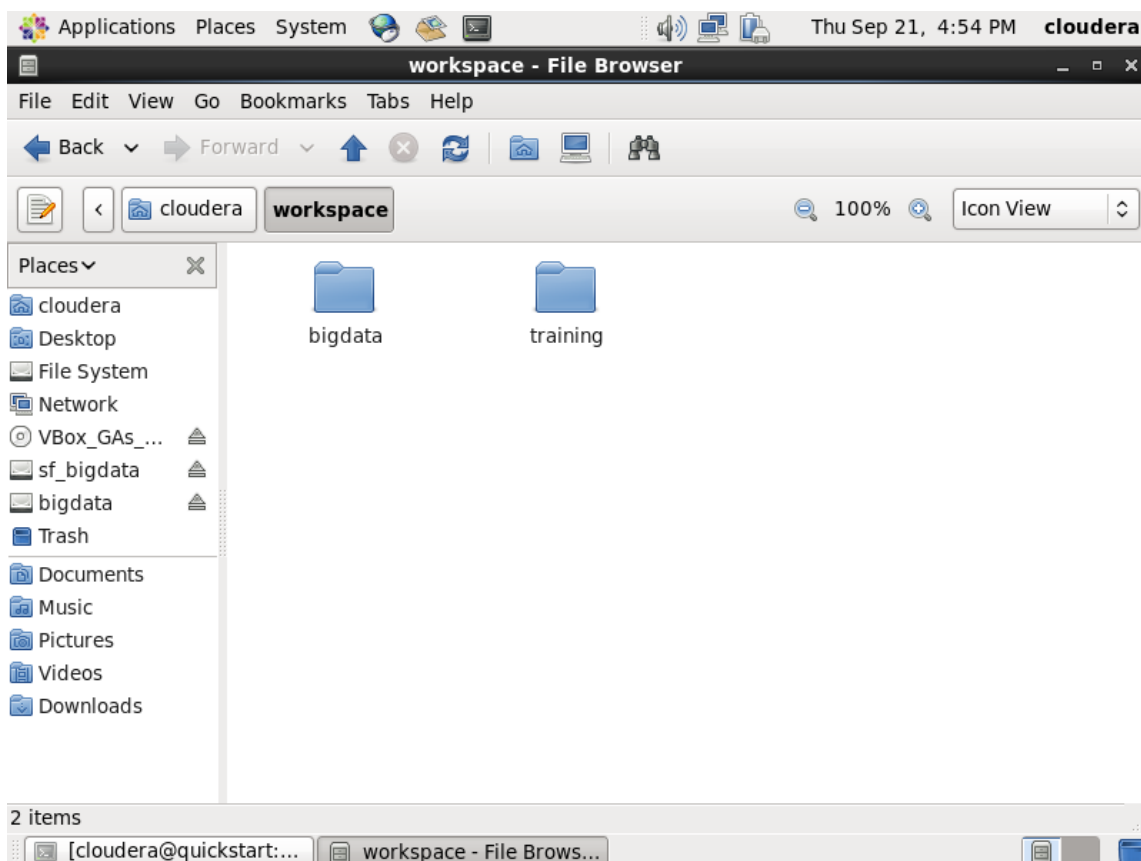
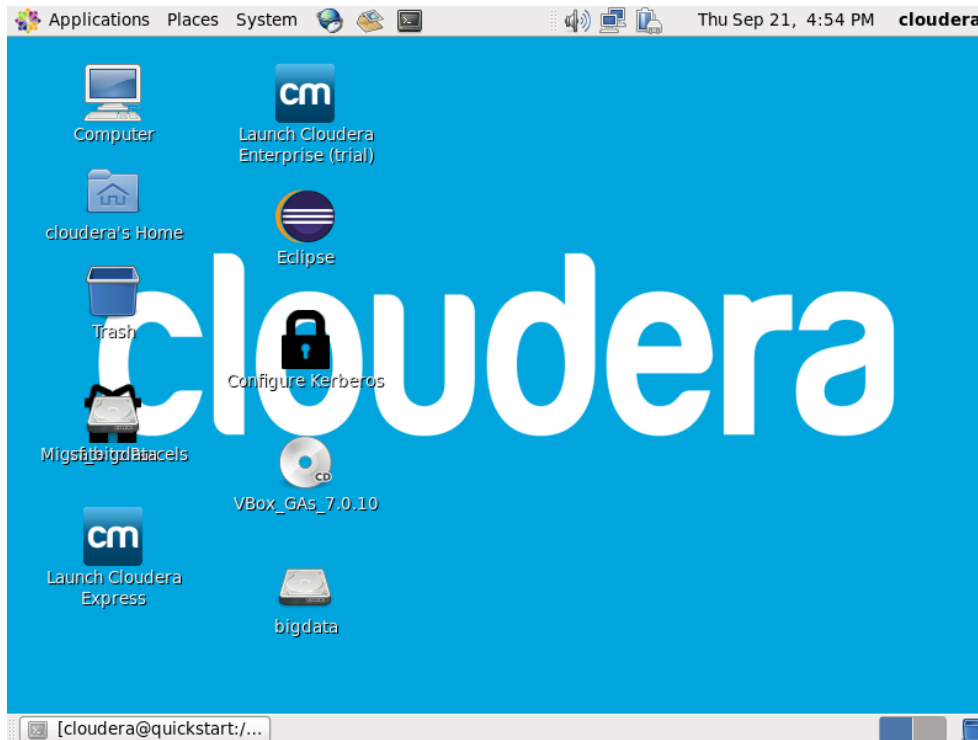
5) Adding shared folder:



6) Shared Folder successfully added:



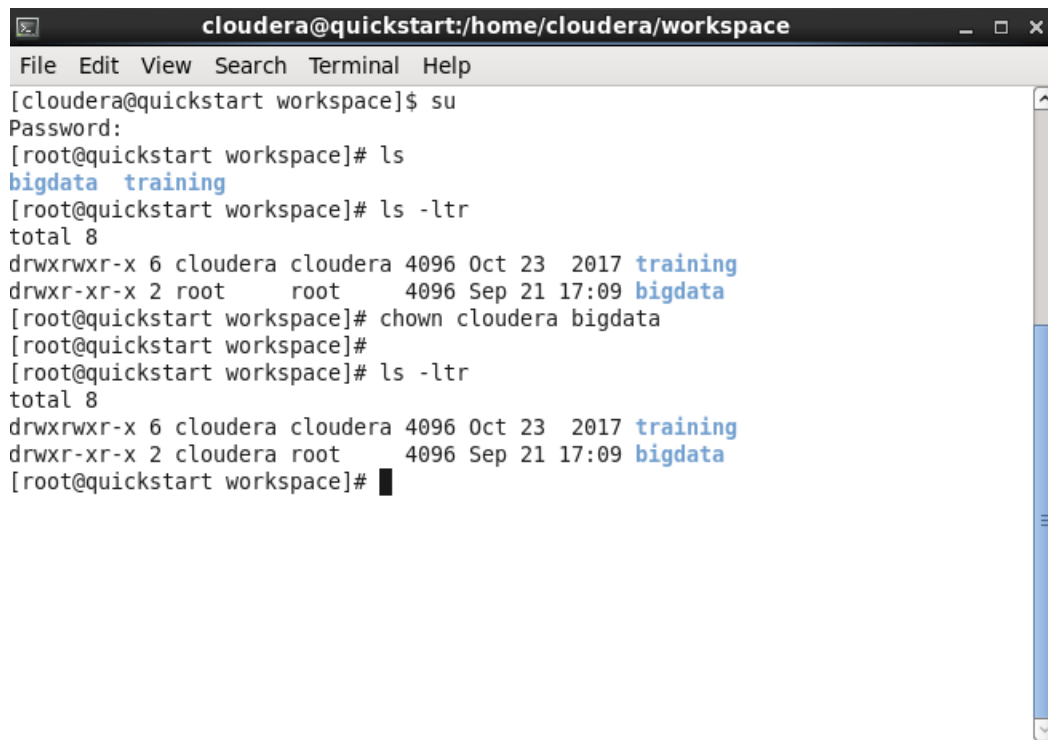
7) Verifying presence of shared folder on system:



8) Changing ownership:-

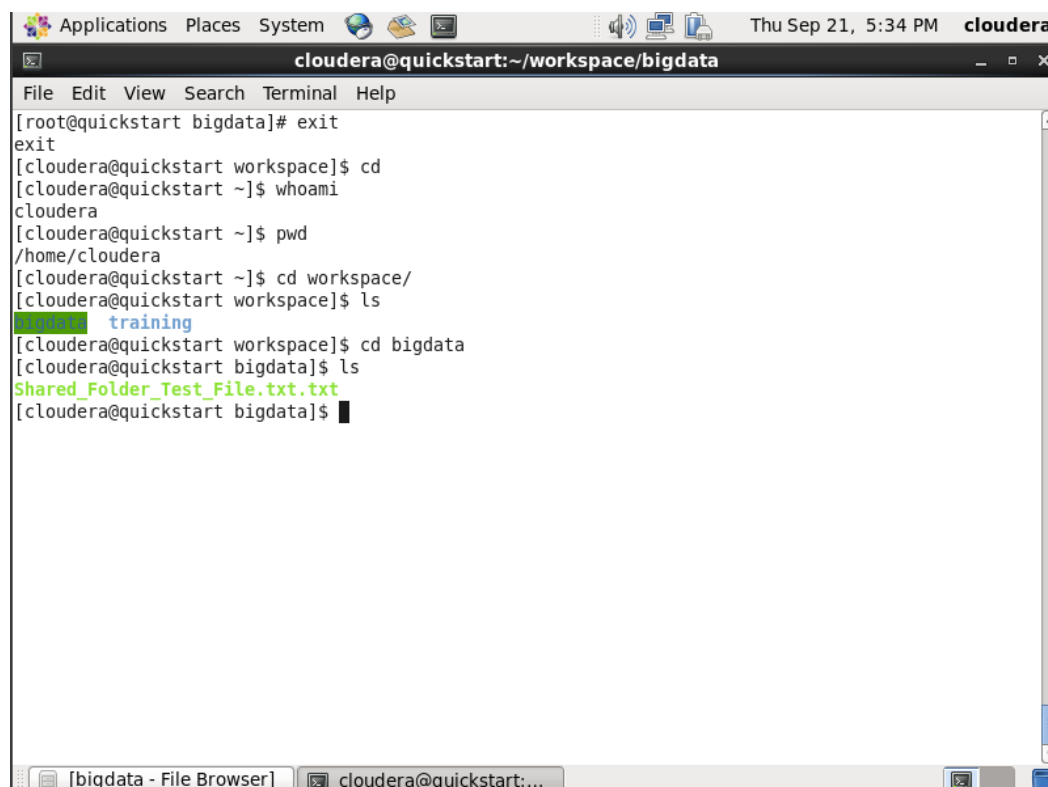
chown cloudera bigdata

ls -ltr



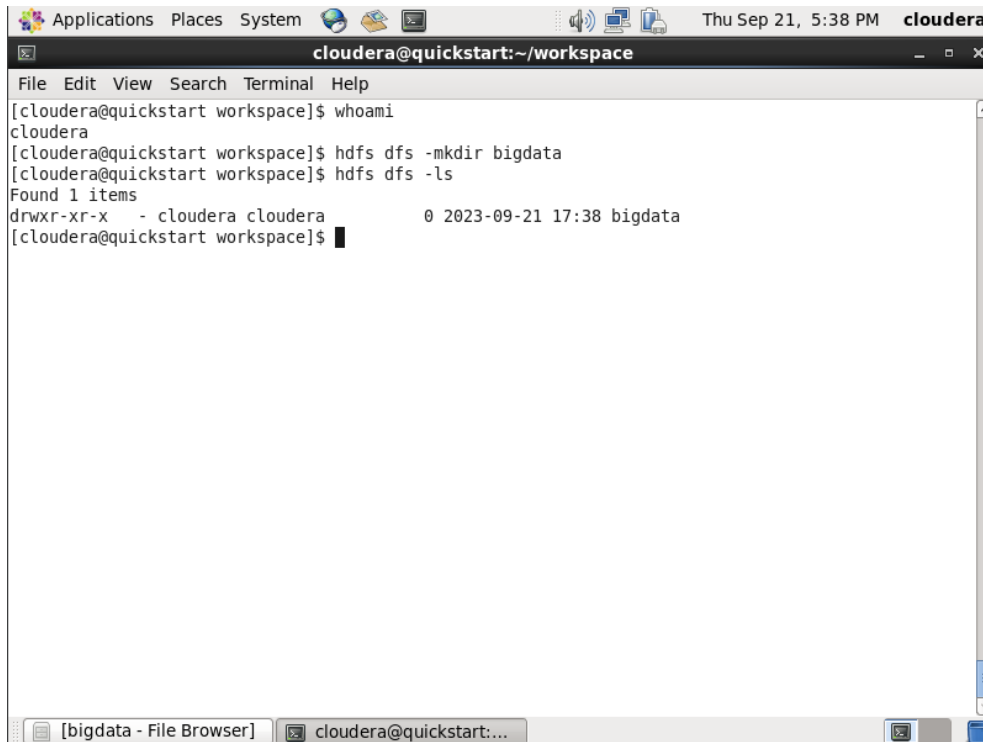
```
cloudera@quickstart:/home/cloudera/workspace
File Edit View Search Terminal Help
[cloudera@quickstart workspace]$ su
Password:
[root@quickstart workspace]# ls
bigdata training
[root@quickstart workspace]# ls -ltr
total 8
drwxrwxr-x 6 cloudera cloudera 4096 Oct 23 2017 training
drwxr-xr-x 2 root root 4096 Sep 21 17:09 bigdata
[root@quickstart workspace]# chown cloudera bigdata
[root@quickstart workspace]#
[root@quickstart workspace]# ls -ltr
total 8
drwxrwxr-x 6 cloudera cloudera 4096 Oct 23 2017 training
drwxr-xr-x 2 cloudera root 4096 Sep 21 17:09 bigdata
[root@quickstart workspace]#
```

9) Looking into file within shared folder with “cloudera” user:



```
Applications Places System Thu Sep 21, 5:34 PM cloudera
cloudera@quickstart:~/workspace/bigdata
File Edit View Search Terminal Help
[root@quickstart bigdata]# exit
exit
[cloudera@quickstart workspace]$ cd
[cloudera@quickstart ~]$ whoami
cloudera
[cloudera@quickstart ~]$ pwd
/home/cloudera
[cloudera@quickstart ~]$ cd workspace/
[cloudera@quickstart workspace]$ ls
bigdata training
[cloudera@quickstart workspace]$ cd bigdata
[cloudera@quickstart bigdata]$ ls
Shared_Folder_Test_File.txt.txt
[cloudera@quickstart bigdata]$
```

10) Created directories within hdfs:

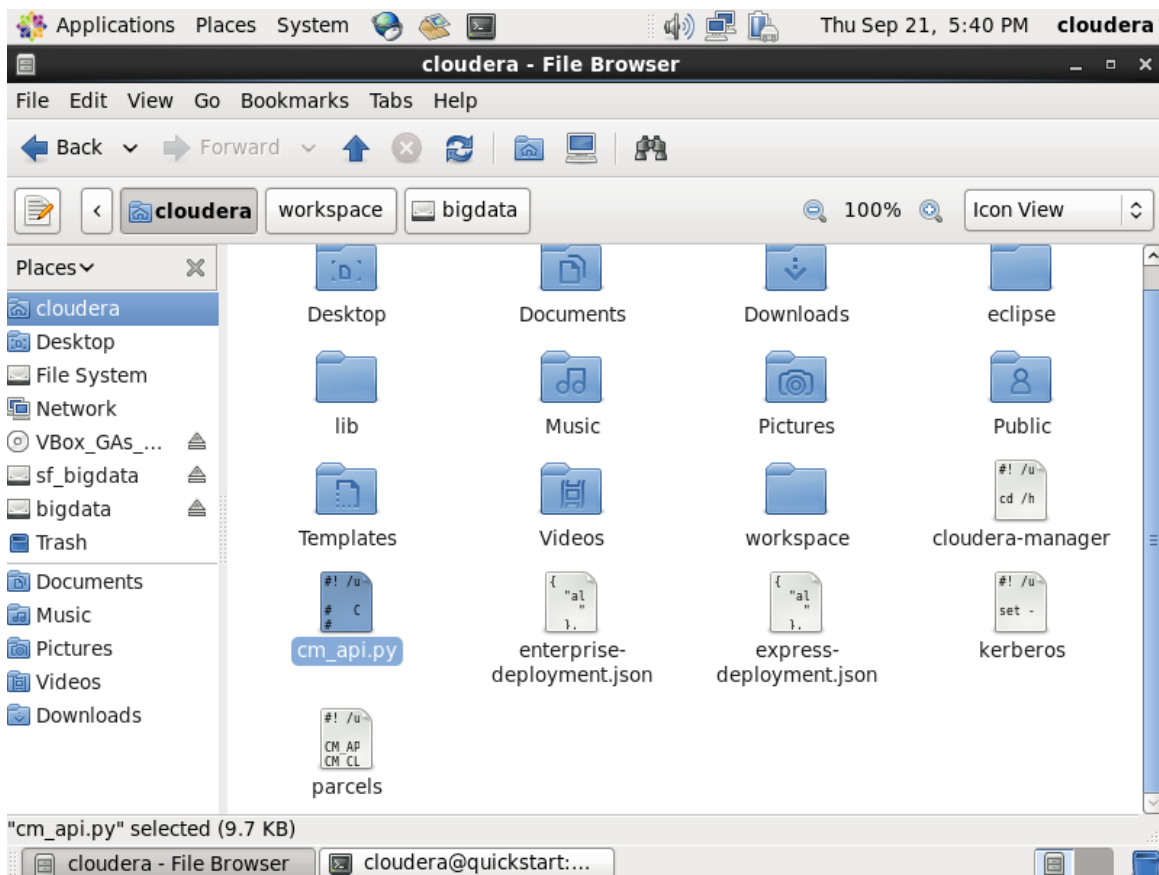


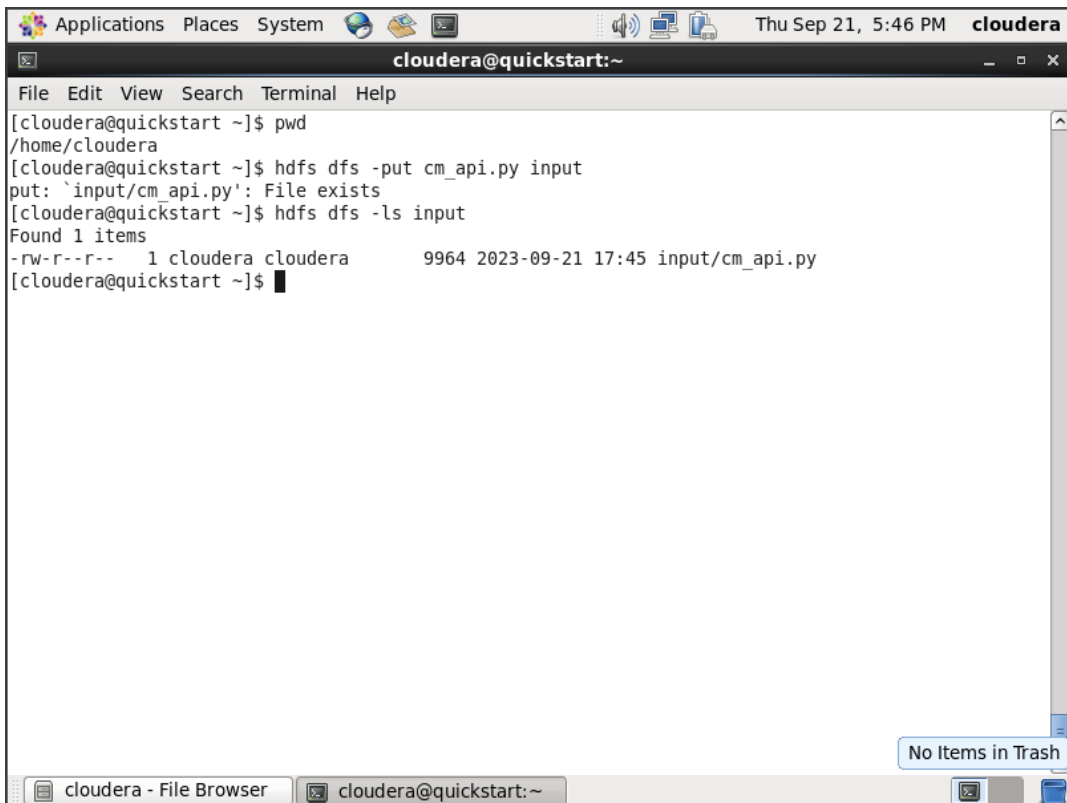
```
cloudera@quickstart:~/workspace
File Edit View Search Terminal Help
[cloudera@quickstart workspace]$ whoami
cloudera
[cloudera@quickstart workspace]$ hdfs dfs -mkdir bigdata
[cloudera@quickstart workspace]$ hdfs dfs -ls
Found 1 items
drwxr-xr-x  - cloudera cloudera      0 2023-09-21 17:38 bigdata
[cloudera@quickstart workspace]$
```

11) Load access log to hdfs:-

hdfs dfs -put cm_api.py

hdfs dfs -ls input

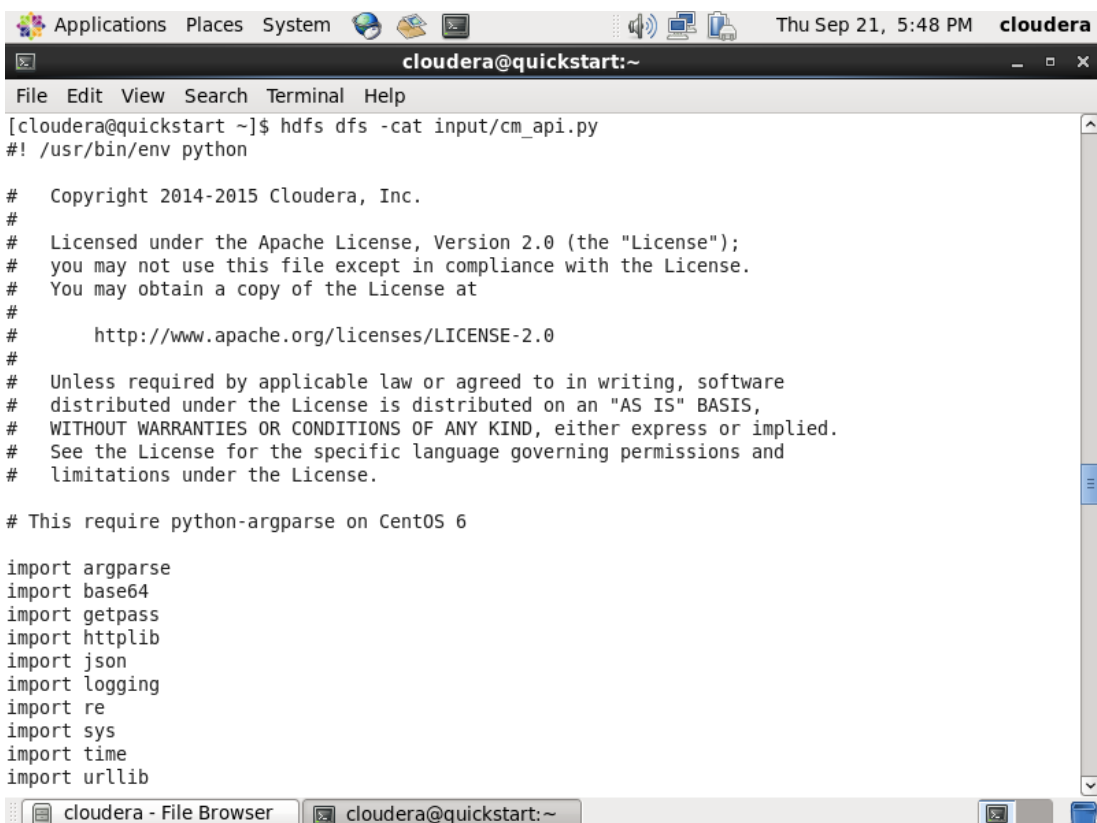


A terminal window titled 'cloudera@quickstart:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows the following commands and output:

```
[cloudera@quickstart ~]$ pwd
/home/cloudera
[cloudera@quickstart ~]$ hdfs dfs -put cm_api.py input
put: 'input/cm_api.py': File exists
[cloudera@quickstart ~]$ hdfs dfs -ls input
Found 1 items
-rw-r--r-- 1 cloudera cloudera      9964 2023-09-21 17:45 input/cm_api.py
[cloudera@quickstart ~]$
```

A 'No Items in Trash' notification bubble is visible in the bottom right corner. The taskbar at the bottom shows 'cloudera - File Browser' and 'cloudera@quickstart:~'.

12) View the contents of the log file:
`hdfs dfs -cat input/cm_api.py`

A terminal window titled 'cloudera@quickstart:~' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows the following commands and output:

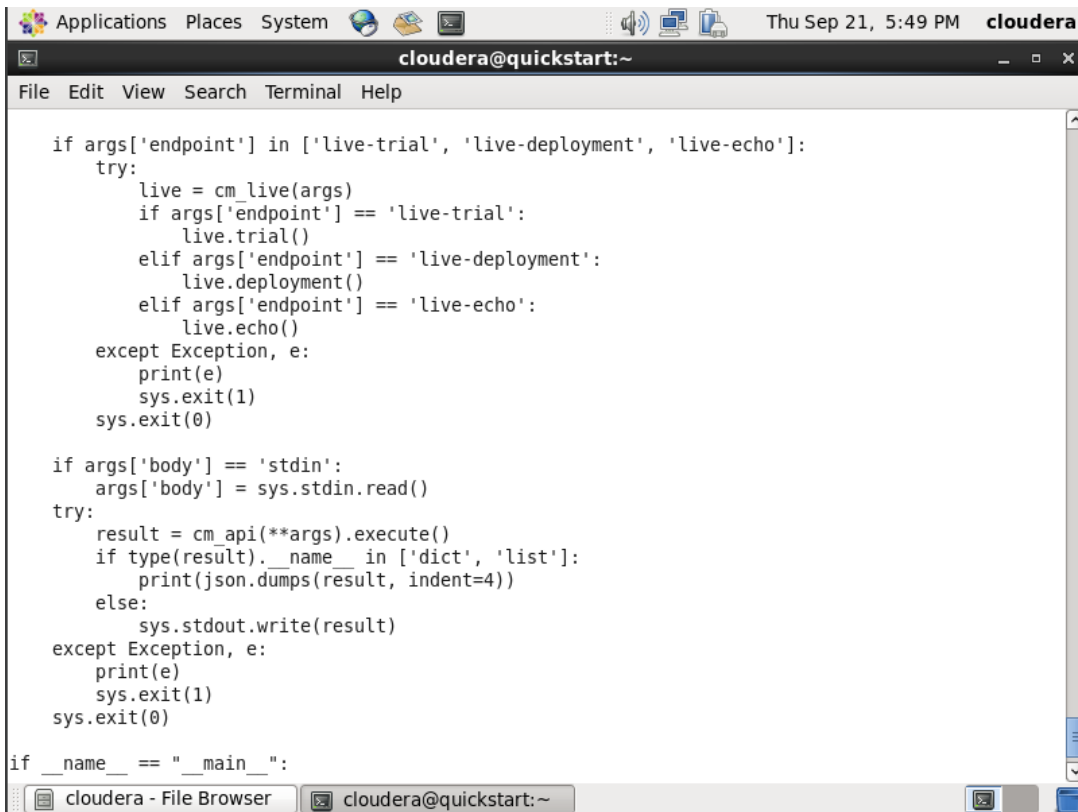
```
[cloudera@quickstart ~]$ hdfs dfs -cat input/cm_api.py
#!/usr/bin/env python

# Copyright 2014-2015 Cloudera, Inc.
#
# Licensed under the Apache License, Version 2.0 (the "License");
# you may not use this file except in compliance with the License.
# You may obtain a copy of the License at
#
#     http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.

# This require python-argparse on CentOS 6

import argparse
import base64
import getpass
import httpplib
import json
import logging
import re
import sys
import time
import urllib
```

The taskbar at the bottom shows 'cloudera - File Browser' and 'cloudera@quickstart:~'.



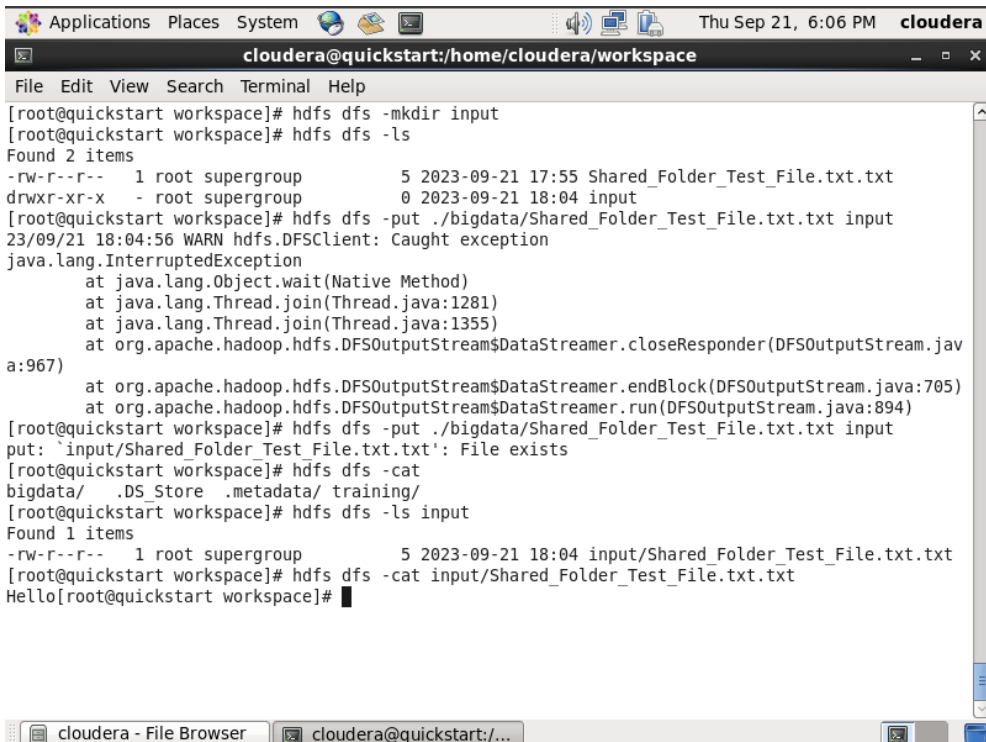
```
Applications  Places  System  Thu Sep 21, 5:49 PM  cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help

if args['endpoint'] in ['live-trial', 'live-deployment', 'live-echo']:
    try:
        live = cm_live(args)
        if args['endpoint'] == 'live-trial':
            live.trial()
        elif args['endpoint'] == 'live-deployment':
            live.deployment()
        elif args['endpoint'] == 'live-echo':
            live.echo()
    except Exception, e:
        print(e)
        sys.exit(1)
    sys.exit(0)

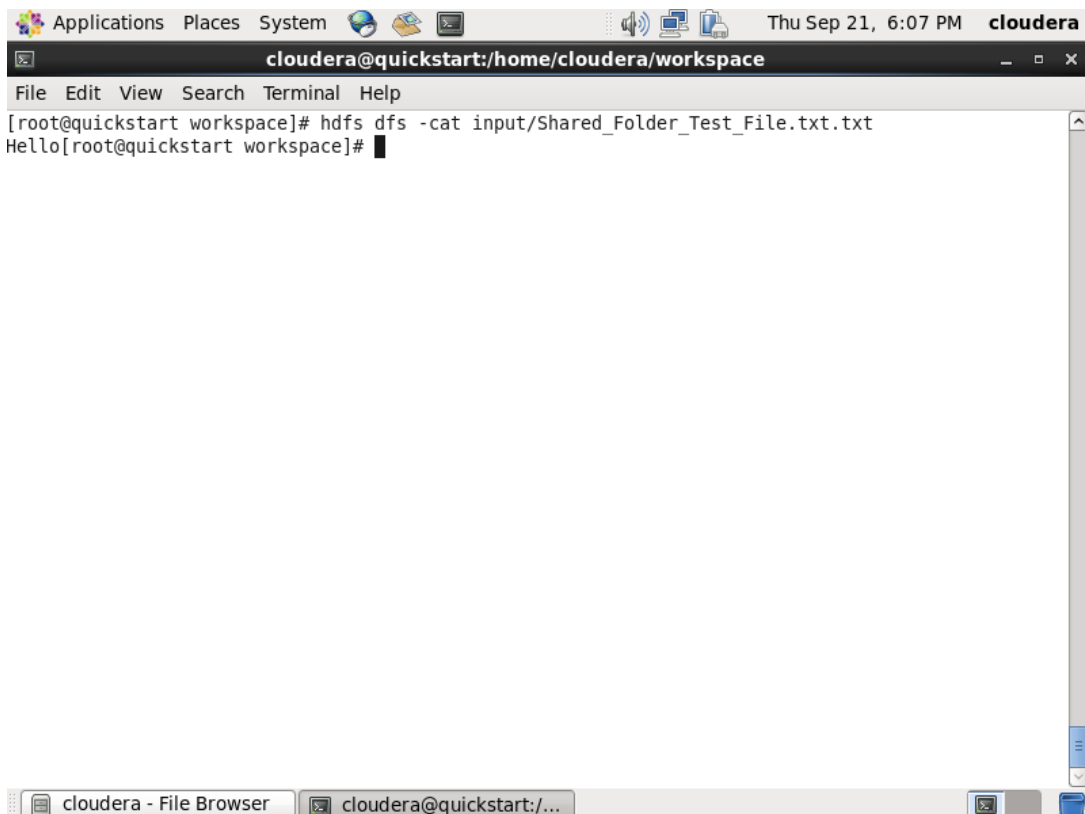
if args['body'] == 'stdin':
    args['body'] = sys.stdin.read()
try:
    result = cm_api(**args).execute()
    if type(result).__name__ in ['dict', 'list']:
        print(json.dumps(result, indent=4))
    else:
        sys.stdout.write(result)
except Exception, e:
    print(e)
    sys.exit(1)
    sys.exit(0)

if __name__ == "__main__":
```

13) Repeat step 11 and 12 to put the .txt file
hdfs dfs -put ./workspace/bigdata/Shared...txt input
hdfs dfs -ls input



```
Applications  Places  System  Thu Sep 21, 6:06 PM  cloudera
cloudera@quickstart:/home/cloudera/workspace
File Edit View Search Terminal Help
[root@quickstart workspace]# hdfs dfs -mkdir input
[root@quickstart workspace]# hdfs dfs -ls
Found 2 items
-rw-r--r--  1 root supergroup          5 2023-09-21 17:55 Shared_Folder_Test_File.txt.txt
drwxr-xr-x  - root supergroup          0 2023-09-21 18:04 input
[root@quickstart workspace]# hdfs dfs -put ./bigdata/Shared_Folder_Test_File.txt.txt input
23/09/21 18:04:56 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.jav
a:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
[root@quickstart workspace]# hdfs dfs -put ./bigdata/Shared_Folder_Test_File.txt.txt input
put: `input/Shared_Folder_Test File.txt.txt': File exists
[root@quickstart workspace]# hdfs dfs -cat
bigdata/  .DS_Store  .metadata/ training/
[root@quickstart workspace]# hdfs dfs -ls input
Found 1 items
-rw-r--r--  1 root supergroup          5 2023-09-21 18:04 input/Shared_Folder_Test_File.txt.txt
[root@quickstart workspace]# hdfs dfs -cat input/Shared_Folder_Test_File.txt.txt
Hello[root@quickstart workspace]#
```

Learning Outcomes: The main learning outcomes from this assignment were:

- Proficiency in using Hadoop Distributed File System (HDFS) commands to list directories, access files, and manage data, which is essential for effective data exploration and management in a Big Data environment.
- Skill in creating, organizing, and removing directories and files within HDFS, demonstrating the ability to structure and maintain data in a Big Data context.
- Competence in adding, verifying, and managing shared folders within HDFS, facilitating collaboration and data sharing among team members, which is critical for team-based Big Data projects.
- Understanding of changing ownership and permissions using Linux commands like "chown," ensuring secure data access and control, a fundamental aspect of data security and governance in Big Data systems.
- Proficiency in using "hdfs dfs -put" to load external data sources into HDFS and "hdfs dfs -cat" to view file contents, enabling data ingestion, and data inspection, which are key steps in the Big Data analytics pipeline.

Now that I have achieved these outcomes, I can navigate the Cloudera environment and use the Hadoop interface proficiently.

