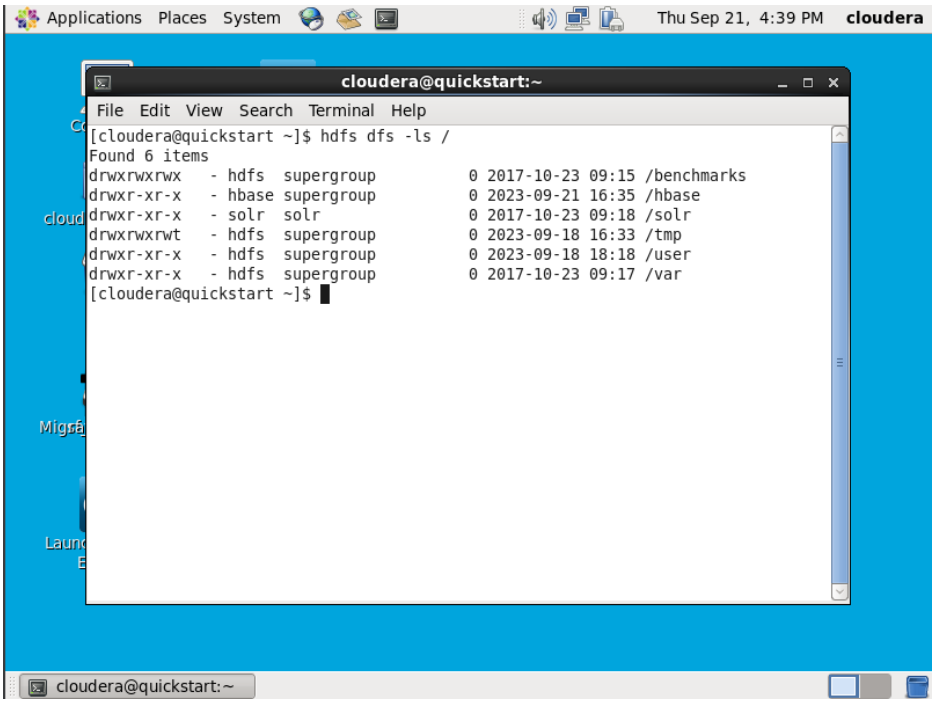
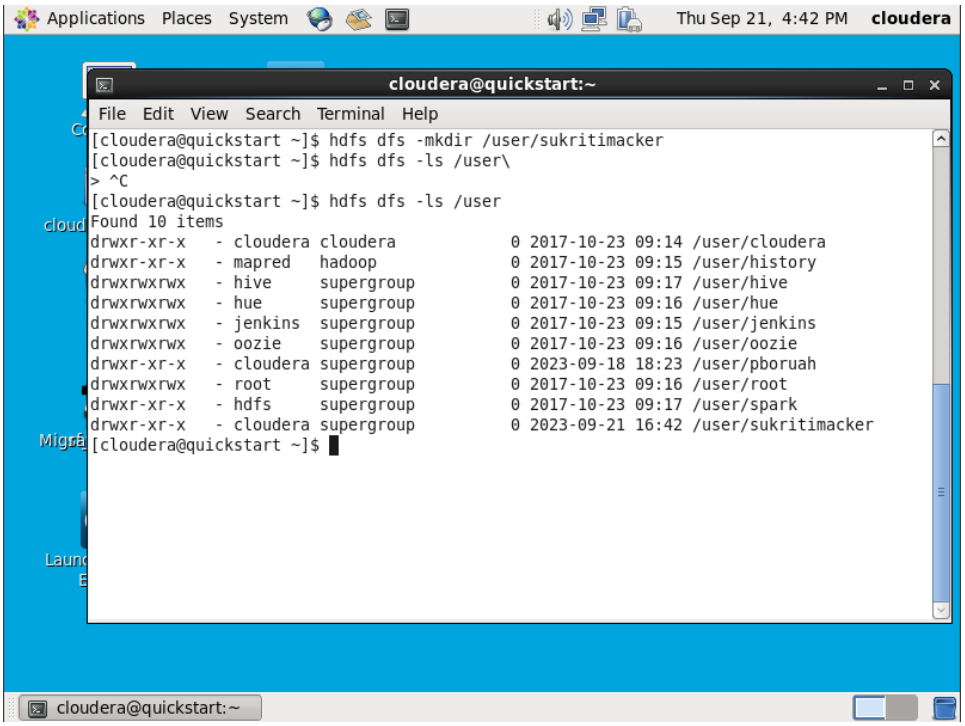
Big Data GHW – 1

Name: Sukriti Macker
NETID: sm11017
Course: Big Data Section – D
Semester: Fall 2023

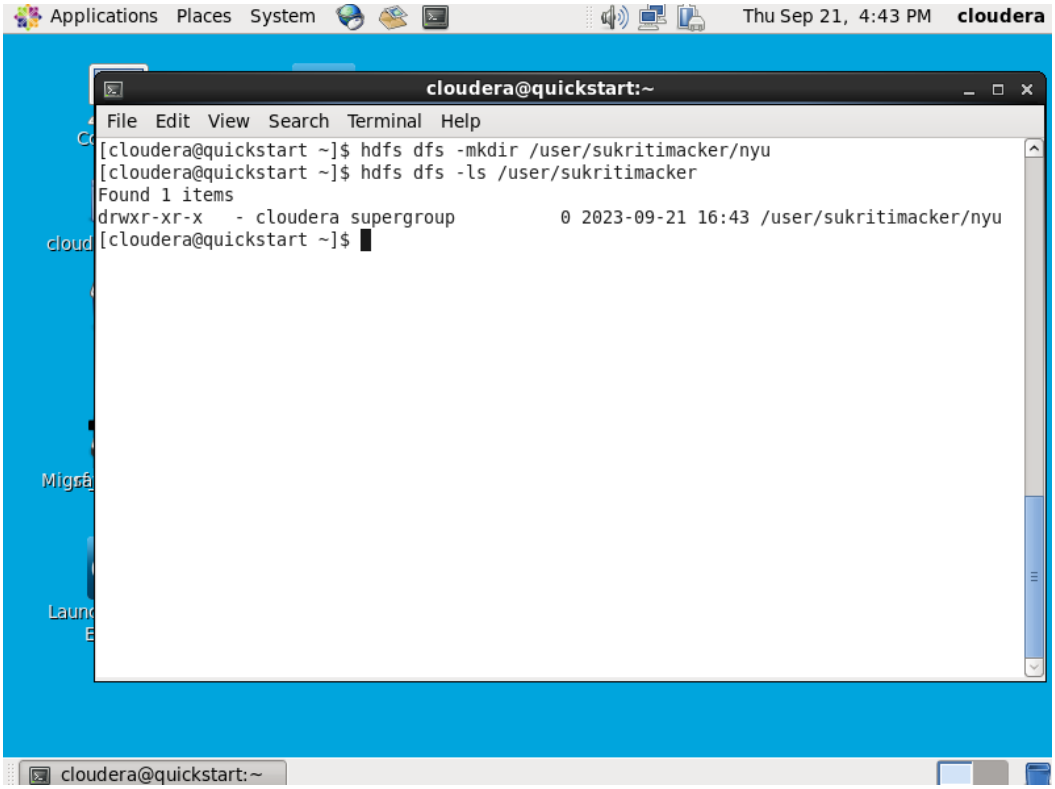
1. Listing directory and files: hdfs dfs -ls /



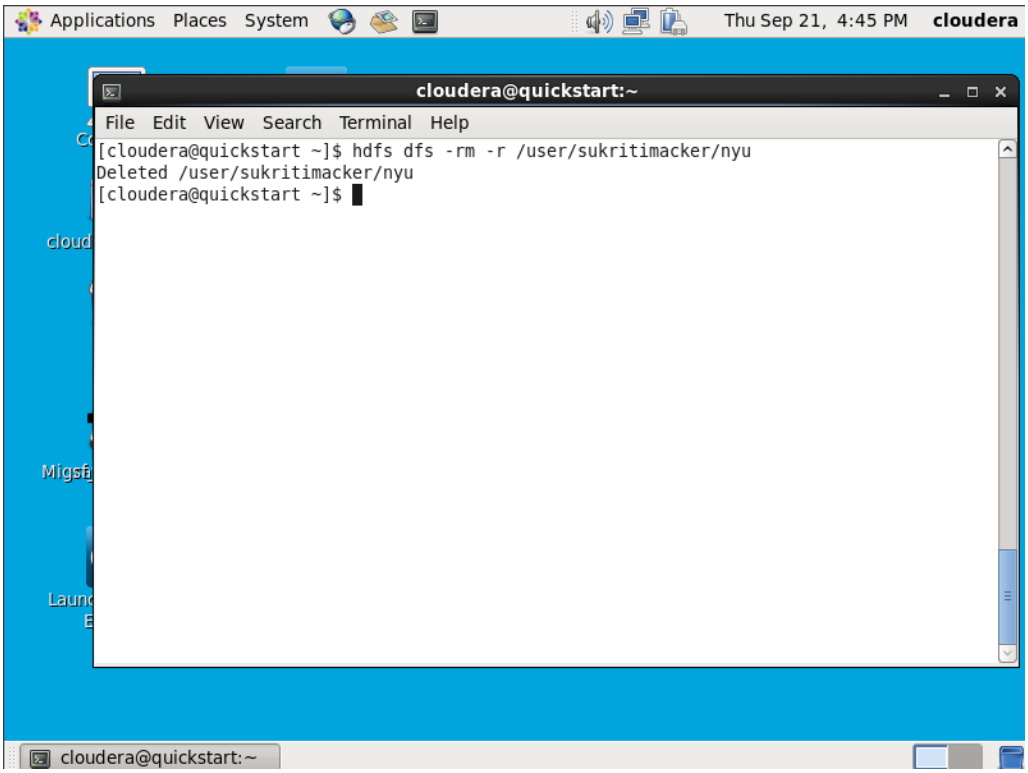
2. Create new directory under /user:-
hdfs dfs -mkdir
hdfs dfs -ls /user



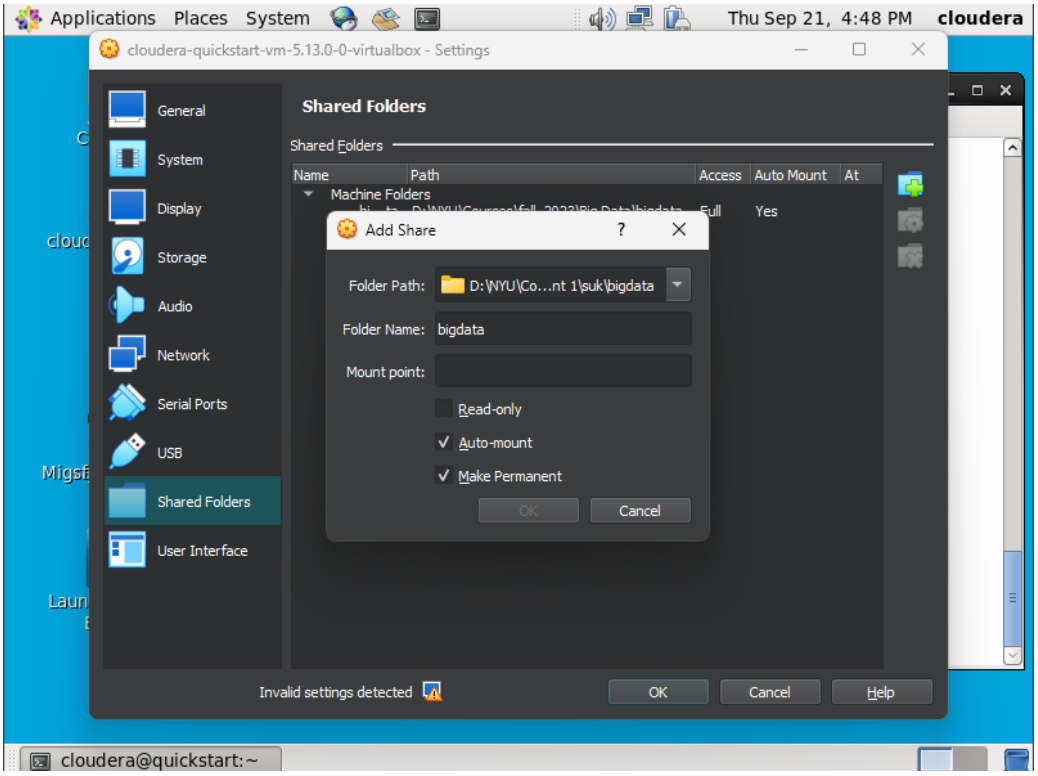
3) Nested destination directories:-
hdfs dfs -mkdir /user/sukritimacker/nyu
hdfs dfs -ls /user/sukritimacker



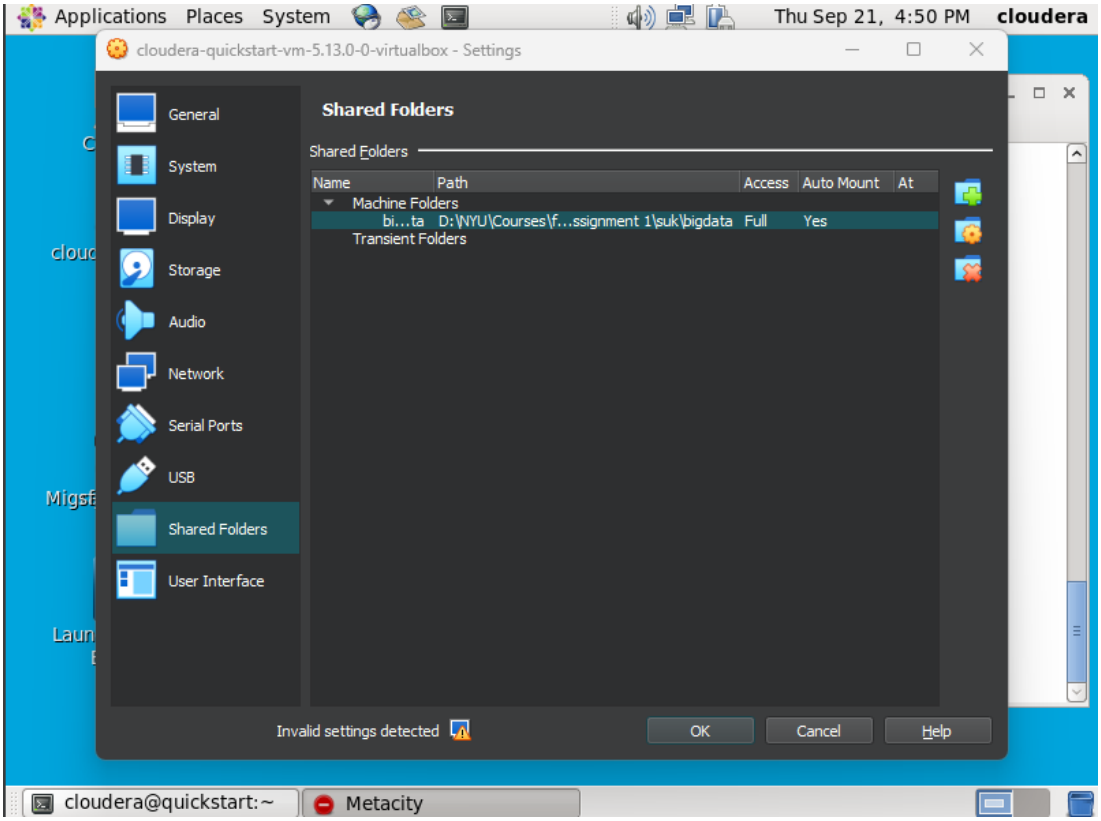
4) Remove dir:
hdfs dfs -rm -r /user/sukritimacker/nyu



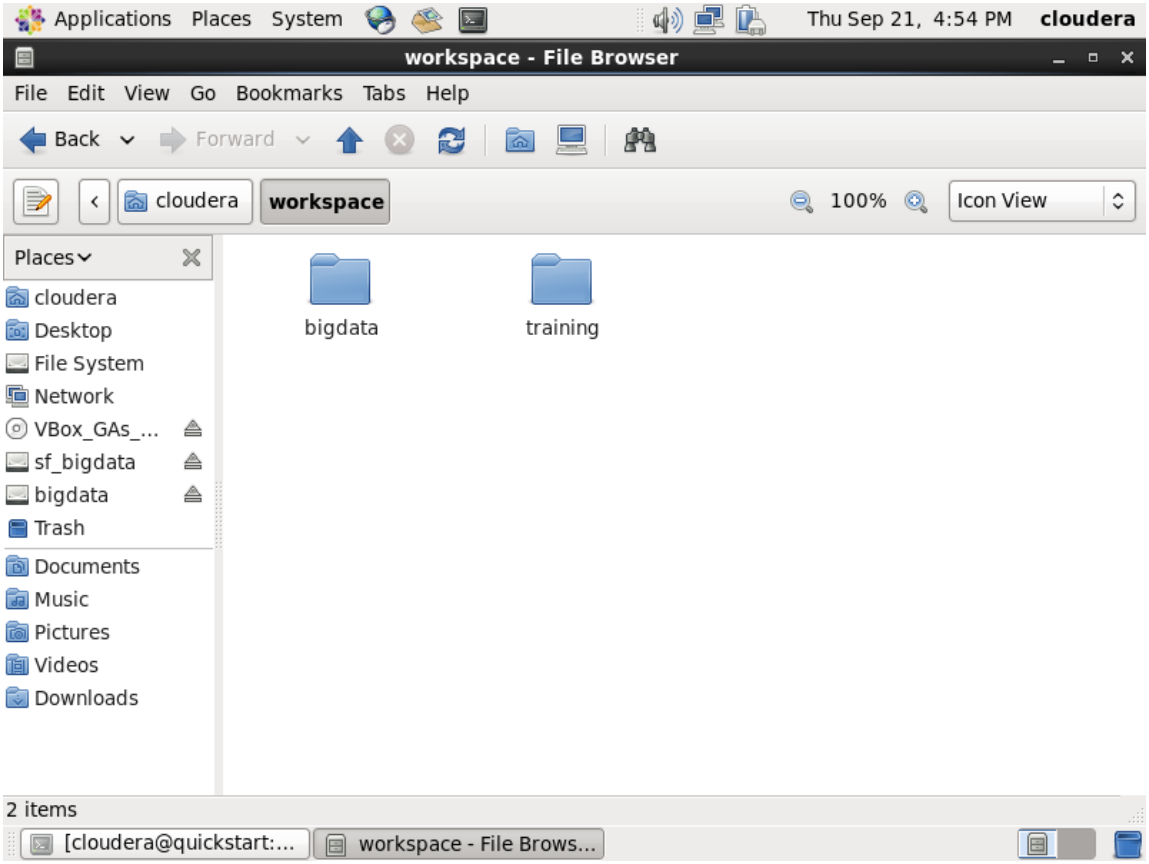
5) Adding shared folder:



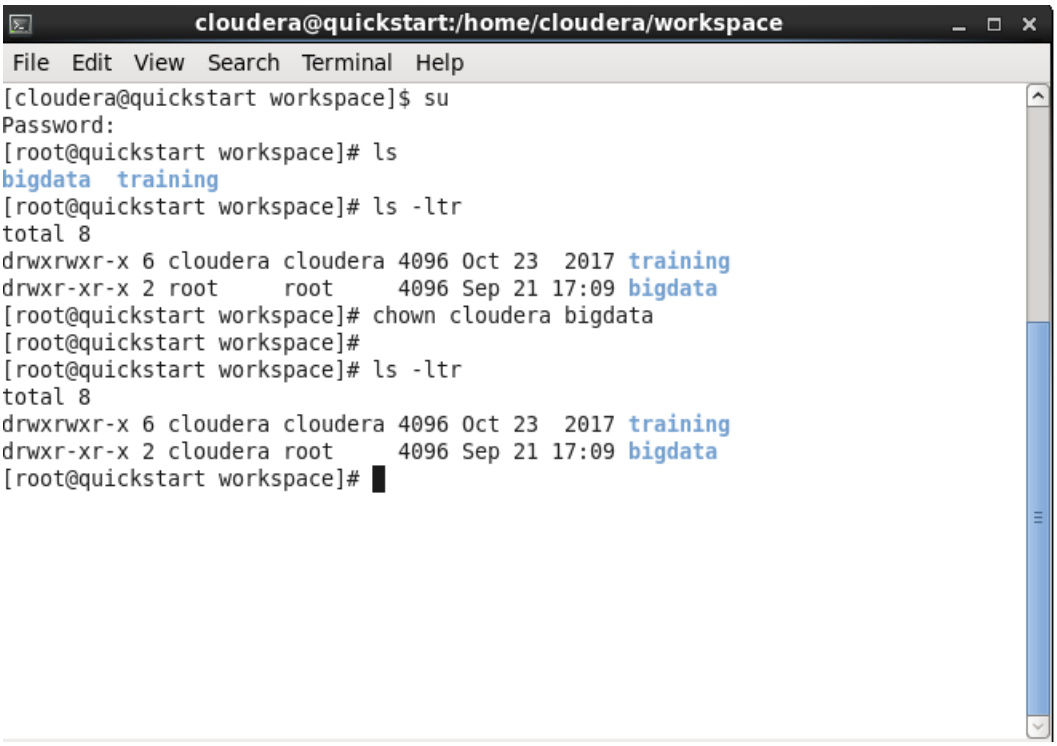
6) Shared Folder successfully added:



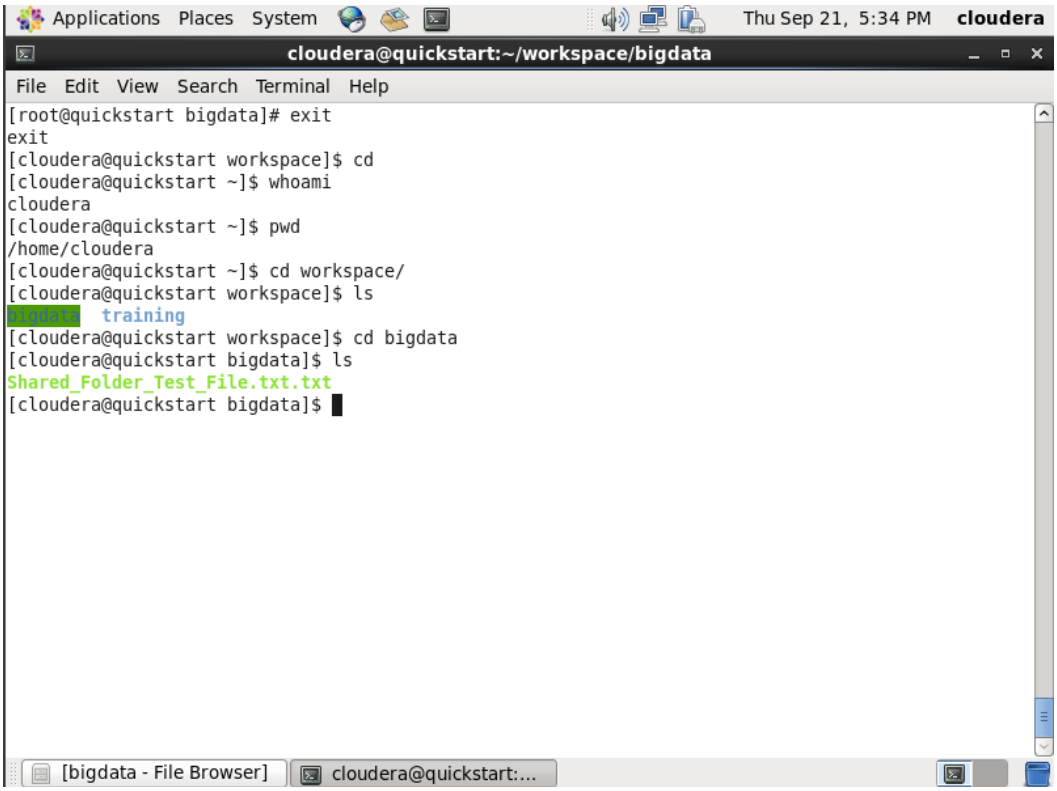
7) Verifying presence of shared folder on system:



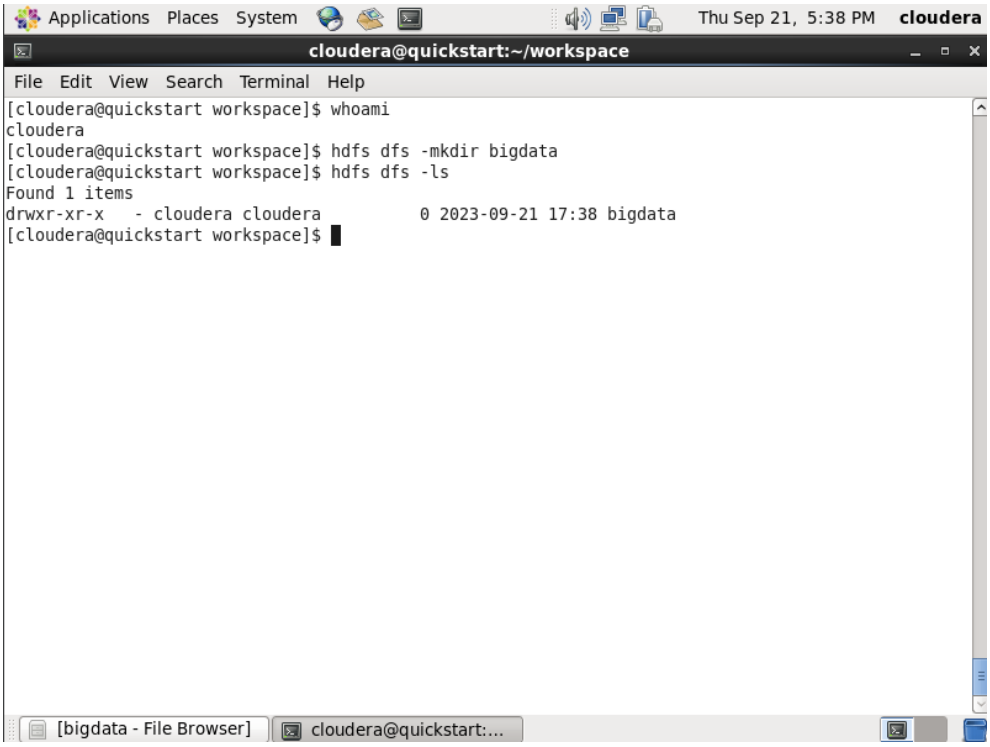
8) Changing ownership:-
chown cloudera bigdata
ls -ltr



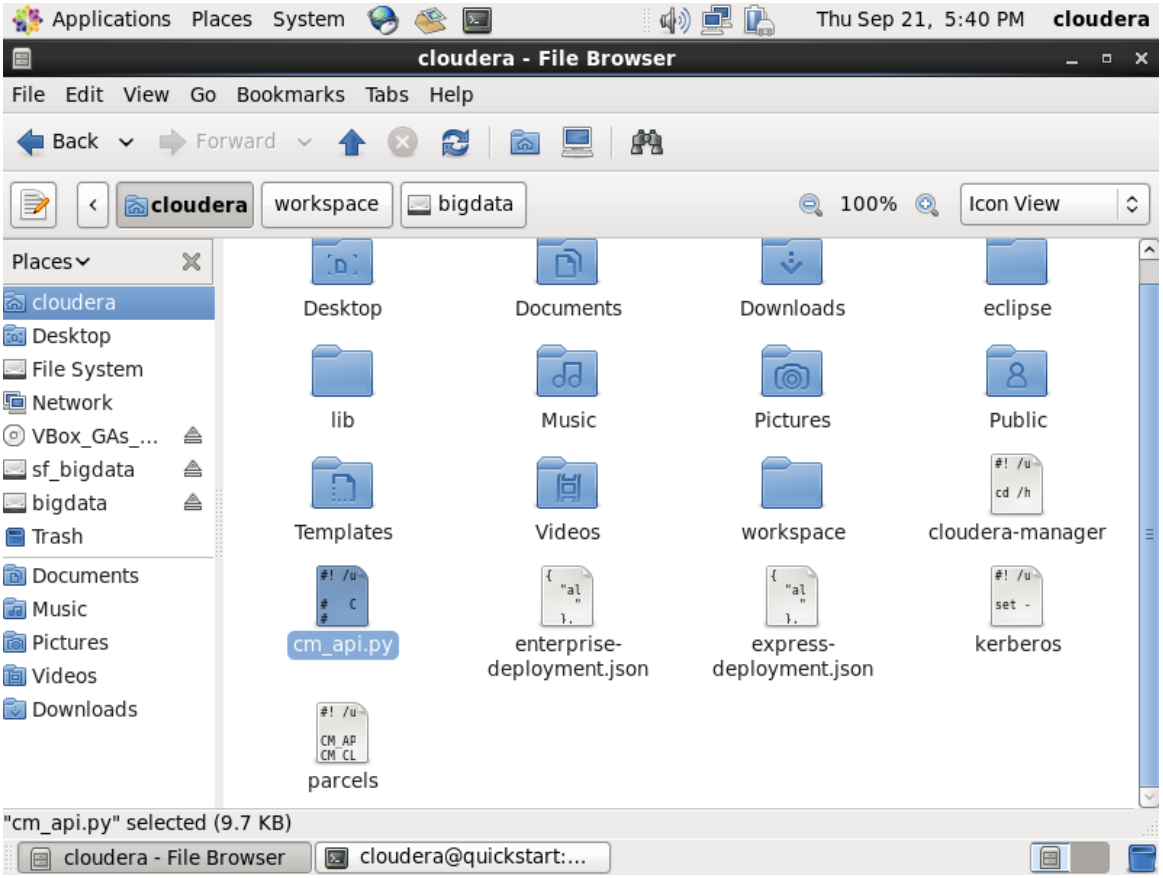
9) Looking into file within shared folder with “cloudera” user:



10) Created directories within hdfs:



11) Load access log to hdfs:-
hdfs dfs -put cm_api.py
hdfs dfs -ls input



```
[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 ~]$
```

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

```
[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 httplib
import json
import logging
import re
import sys
import time
import urllib
```

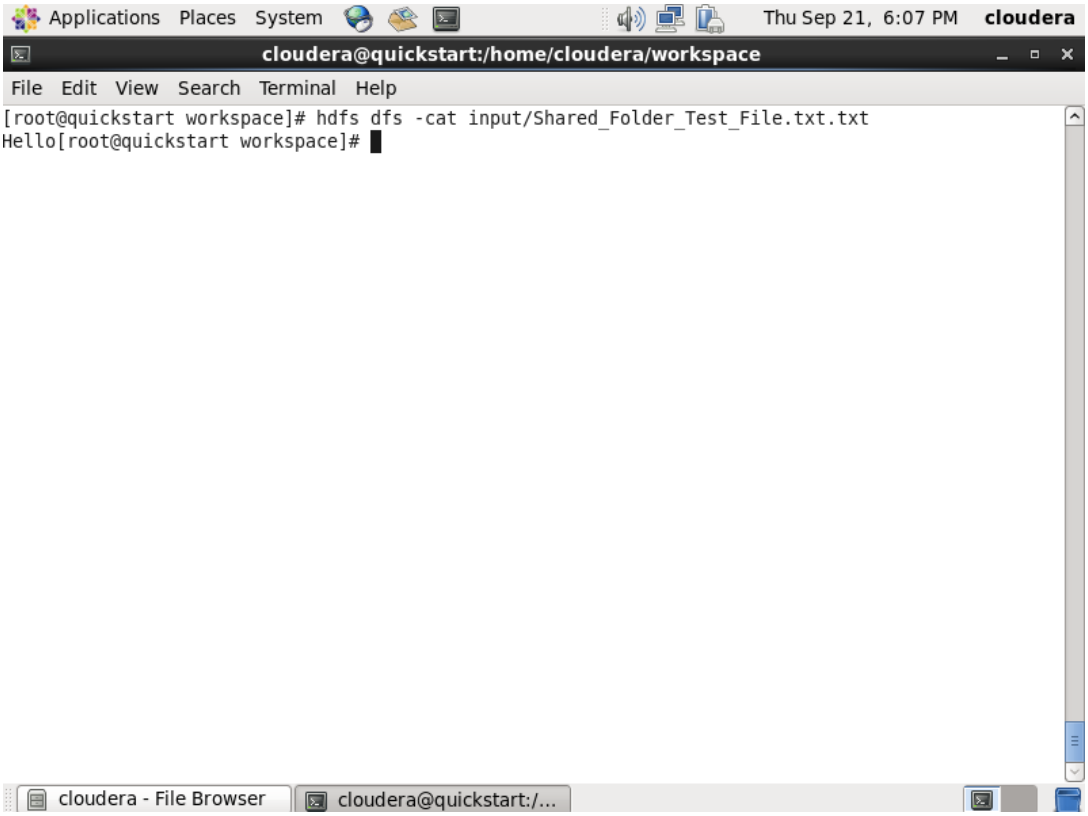
```
    if args['endpoint'] in ['live-trial', 'live-deployment', 'live-echo']:
        try:
            live = cm_live(args)
            if args['endpoint'] == 'live-trial':
                live.trial()
            elif args['endpoint'] == 'live-deployment':
                live.deployment()
            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

```
[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.