

CSCE 5300 INTRODUCTION TO BIG DATA AND DATA SCIENCE

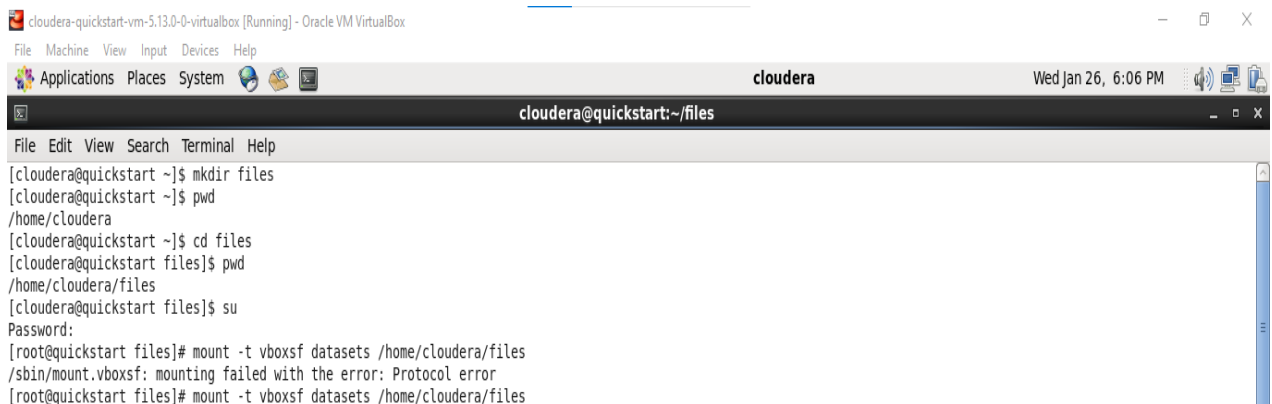
ASSIGNMENT-1

NAME : Eswara Reddy Thimmapuram

ID : 11506566

STEP 1 : Importing datasets

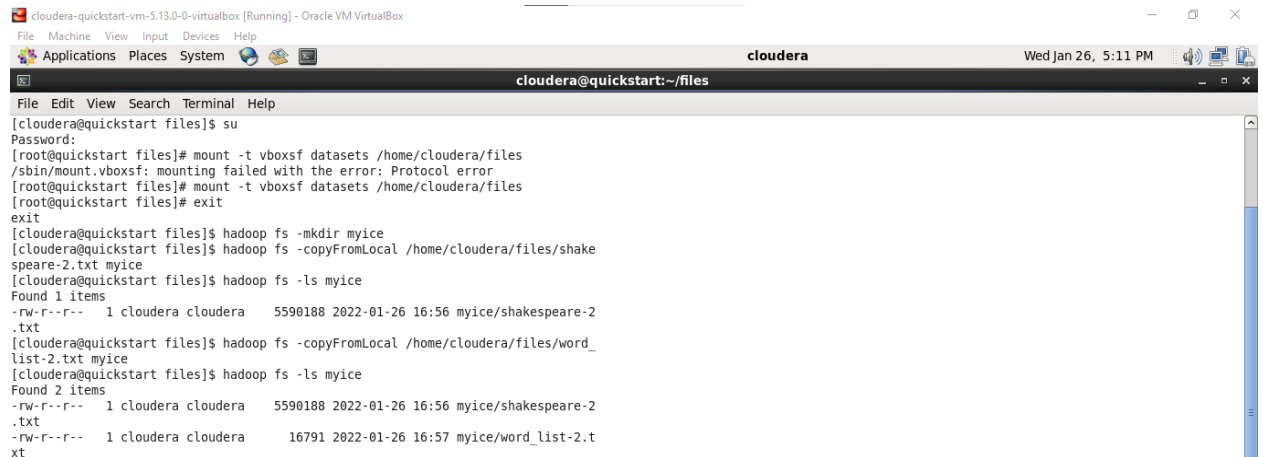
- In the first step, I have created a folder named 'files' in cloudera using the command **mkdir** and imported the datasets **shakespeare-2.txt** and **word_list-2.txt** into it using **mount -t vboxsf datasets /home/cloudera/files**.



```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System cloudera Wed Jan 26, 6:06 PM
cloudera@quickstart:~/files
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ mkdir files
[cloudera@quickstart ~]$ pwd
/home/cloudera
[cloudera@quickstart ~]$ cd files
[cloudera@quickstart files]$ pwd
/home/cloudera/files
[cloudera@quickstart files]$ su
Password:
[root@quickstart files]# mount -t vboxsf datasets /home/cloudera/files
/sbin/mount.vboxsf: mounting failed with the error: Protocol error
[root@quickstart files]# mount -t vboxsf datasets /home/cloudera/files
```

STEP 2 : Loading data into Hadoop hdfs

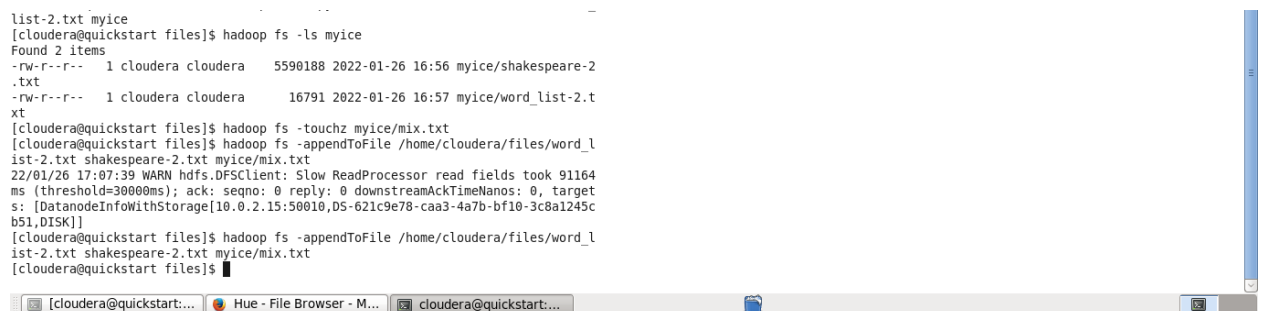
- In the second step, I have created a folder named 'myice' and loaded the datasets into Hadoop using **copyFromLocal** command.
- Also, I have checked the contents of 'myice' using **-ls** command.



```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera Wed Jan 26, 5:11 PM
cloudera@quickstart:~/files
File Edit View Search Terminal Help
[cloudera@quickstart files]$ su
Password:
[root@quickstart files]# mount -t vboxsf datasets /home/cloudera/files
/sbin/mount.vboxsf: mounting failed with the error: Protocol error
[root@quickstart files]# mount -t vboxsf datasets /home/cloudera/files
[root@quickstart files]# exit
exit
[cloudera@quickstart files]$ hadoop fs -mkdir myice
[cloudera@quickstart files]$ hadoop fs -copyFromLocal /home/cloudera/files/shakespeare-2.txt myice
[cloudera@quickstart files]$ hadoop fs -ls myice
Found 1 items
-rw-r--r-- 1 cloudera cloudera 5590188 2022-01-26 16:56 myice/shakespeare-2.txt
[cloudera@quickstart files]$ hadoop fs -copyFromLocal /home/cloudera/files/word_list-2.txt myice
[cloudera@quickstart files]$ hadoop fs -ls myice
Found 2 items
-rw-r--r-- 1 cloudera cloudera 5590188 2022-01-26 16:56 myice/shakespeare-2.txt
-rw-r--r-- 1 cloudera cloudera 16791 2022-01-26 16:57 myice/word_list-2.txt
```

STEP 3 : Using second file and appending the second file to the first file

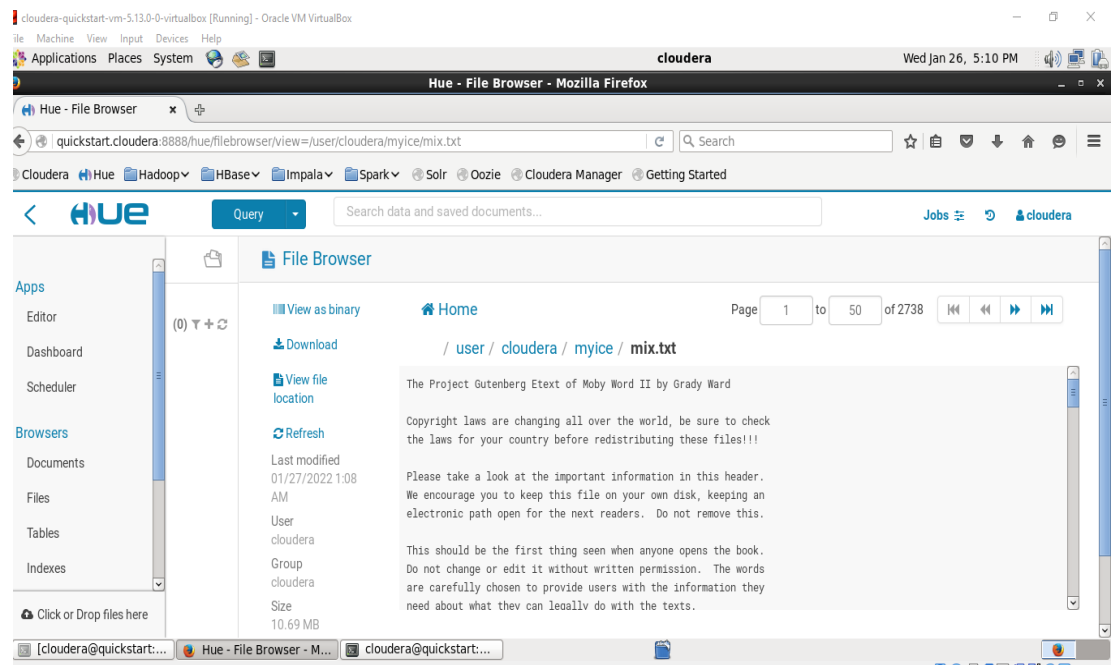
- In the third step, I have created a folder named 'mix' using **-touchz** to store the merged version of both the datasets.
- Then, I used the command **appendToFile** to append the second file to the first file and stored the appended version in mix.txt.



```
list-2.txt myice
[cloudera@quickstart files]$ hadoop fs -ls myice
Found 2 items
-rw-r--r-- 1 cloudera cloudera 5590188 2022-01-26 16:56 myice/shakespeare-2.txt
-rw-r--r-- 1 cloudera cloudera 16791 2022-01-26 16:57 myice/word_list-2.txt
[cloudera@quickstart files]$ hadoop fs -touchz myice/mix.txt
[cloudera@quickstart files]$ hadoop fs -appendToFile /home/cloudera/files/word_list-2.txt myice/mix.txt
22/01/26 17:07:39 WARN hdfs.DFSClient: Slow ReadProcessor read fields took 91164 ms (threshold=30000ms); ack: seqno: 0 reply: 0 downstreamAckTimeNanos: 0, target s: [DataNodeInfoWithStorage[10.0.2.15:50010,DS-621c9e78-caa3-4a7b-bf10-3c8a1245cb51,DISK]]
[cloudera@quickstart files]$ hadoop fs -appendToFile /home/cloudera/files/word_list-2.txt myice/mix.txt
[cloudera@quickstart files]$
```

STEP 4 : File Visualization with Hue

- Next I have opened Hue in the cloudera browser and opened dashboard and clicked on files.
- Then I visualized the 'mix.txt' file in which the appended files are present.



STEP 5 : Viewing the first five lines of merged dataset using head command

- Here I have used **cat** command to display the contents of the file.
- Then I have used **head** command and displayed the first five lines of merged dataset mix.txt.

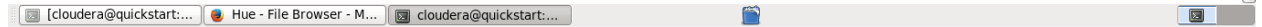
```
[cloudera@quickstart files]$ hadoop fs -cat /user/cloudera/myice/mix.txt | head -5
The Project Gutenberg Etext of Moby Word II by Grady Ward

Copyright laws are changing all over the world, be sure to check
the laws for your country before redistributing these files!!!

cat: Unable to write to output stream.
[cloudera@quickstart files]$ hadoop fs -cat /user/cloudera/myice/mix.txt | tail -7
or filename 24689 would be found at:
http://www.gutenberg.org/2/4/6/8/24689

An alternative method of locating eBooks:
http://www.gutenberg.org/GUTINDEX.ALL

*** END: FULL LICENSE ***[cloudera@quickstart files]$
```



STEP 6 : Viewing the last seven lines of merged dataset using tail command

- Here I have used the **tail** command and displayed the last seven lines of the merged dataset.

```
[cloudera@quickstart files]$ hadoop fs -cat /user/cloudera/myice/mix.txt | head -5
The Project Gutenberg Etext of Moby Word II by Grady Ward

Copyright laws are changing all over the world, be sure to check
the laws for your country before redistributing these files!!!

cat: Unable to write to output stream.
[cloudera@quickstart files]$ hadoop fs -cat /user/cloudera/myice/mix.txt | tail -7
or filename 24689 would be found at:
http://www.gutenberg.org/2/4/6/8/24689

An alternative method of locating eBooks:
http://www.gutenberg.org/GUTINDEX.ALL

*** END: FULL LICENSE ***[cloudera@quickstart files]$
```



STEP 7 : Creation of new file and loading it into Hadoop hdfs

- Here I have used **cat** command and **copyFromLocal** command and created a new file named newfile.txt and loaded it into Hadoop hdfs.

```
An alternative method of locating eBooks:
http://www.gutenberg.org/GUTINDEX.ALL

*** END: FULL LICENSE ***[cloudera@quickstart ~]$ cat > new.txt
this is my first assignment[cloudera@quickstart ~]$ hadoop fs -copyFromLocal new.txt newfile.txt
[cloudera@quickstart ~]$
```

STEP 8 : Appending three datasets

- Here I have used the command **-getmerge** to append all the three datasets by specifying the appropriate paths for each file.

```
An alternative method of locating eBooks:
http://www.gutenberg.org/GUTINDEX.ALL

*** END: FULL LICENSE ***[cloudera@quickstart ~]$ cat > new.txt
this is my first assignment[cloudera@quickstart ~]$ hadoop fs -copyFromLocal new.txt newfile.txt
[cloudera@quickstart ~]$ hdfs dfs -getmerge /user/cloudera/shakespeare.txt /user/cloudera/word_list.txt /user/cloudera/newfile.txt final.txt
[cloudera@quickstart ~]$ █
```

STEP 9 : Visualizing all datasets with Hue

- In the final step I have opened Hue and visualized all the three datasets.

