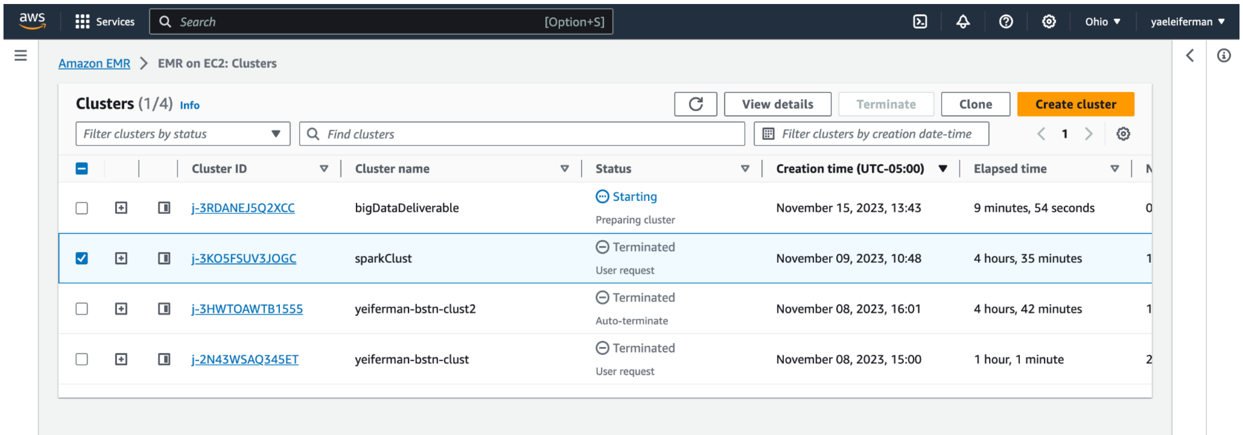
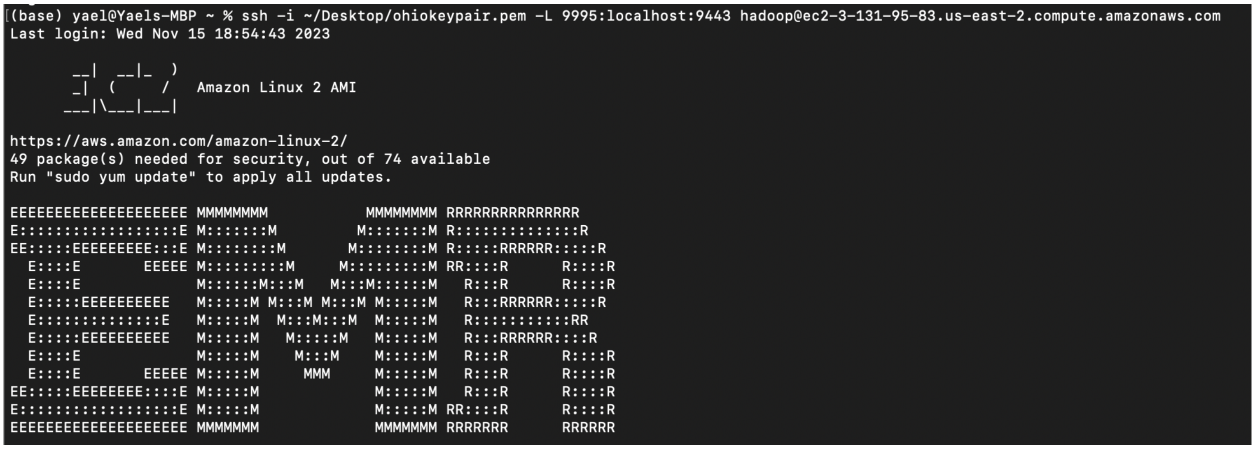
1. Created bigDataDeliverable cluster by checking off sparkClust and cloning (just changed the name):



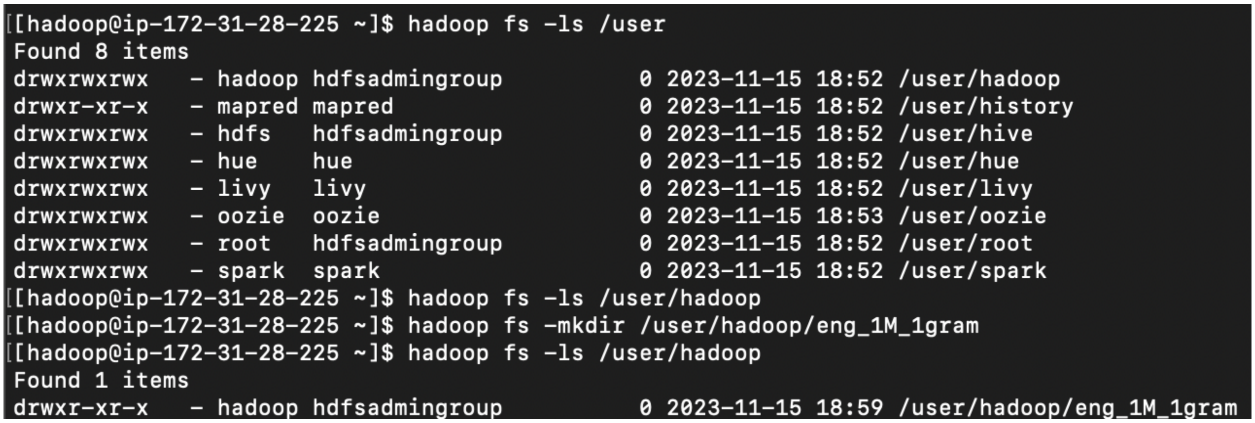
☐	Cluster ID	Cluster name	Status	Creation time (UTC-05:00)	Elapsed time	N
☐	j-3RDANEJ5Q2XCC	bigDataDeliverable	Starting Preparing cluster	November 15, 2023, 13:43	9 minutes, 54 seconds	0
☑	j-3KO5FSUV3JOGC	sparkClust	Terminated User request	November 09, 2023, 10:48	4 hours, 35 minutes	1
☐	j-3HWTOAWTB1555	yeiferman-bstn-clust2	Terminated Auto-terminate	November 08, 2023, 16:01	4 hours, 42 minutes	1
☐	j-2N43WSAQ345ET	yeiferman-bstn-clust	Terminated User request	November 08, 2023, 15:00	1 hour, 1 minute	2

2. SSH'd into the head node, making sure to include the -L option in order to access Jupyterhub from the web later:



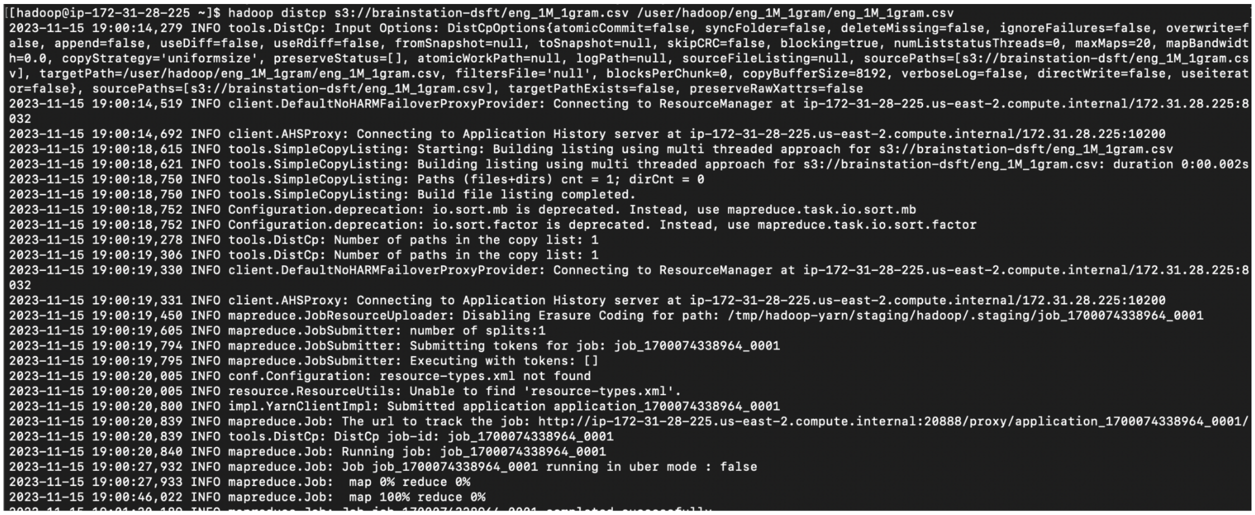
```
[(base) yael@Yaels-MBP ~ % ssh -i ~/Desktop/ohiokeypair.pem -L 9995:localhost:9443 hadoop@ec2-3-131-95-83.us-east-2.compute.amazonaws.com
Last login: Wed Nov 15 18:54:43 2023

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
49 package(s) needed for security, out of 74 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR
```

3. Checked what was in the /user/hadoop directory to begin with (nothing), then made a new directory called eng_1M_1gram within that directory:



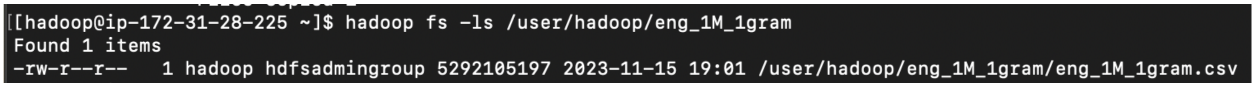
```
[[hadoop@ip-172-31-28-225 ~]$ hadoop fs -ls /user
Found 8 items
drwxrwxrwx   - hadoop hdfsadmingroup          0 2023-11-15 18:52 /user/hadoop
drwxr-xr-x   - mapred mapred                  0 2023-11-15 18:52 /user/history
drwxrwxrwx   - hdfs   hdfsadmingroup          0 2023-11-15 18:52 /user/hive
drwxrwxrwx   - hue    hue                     0 2023-11-15 18:52 /user/hue
drwxrwxrwx   - livy   livy                    0 2023-11-15 18:52 /user/livy
drwxrwxrwx   - oozie  oozie                   0 2023-11-15 18:53 /user/oozie
drwxrwxrwx   - root   hdfsadmingroup          0 2023-11-15 18:52 /user/root
drwxrwxrwx   - spark  spark                   0 2023-11-15 18:52 /user/spark
[[hadoop@ip-172-31-28-225 ~]$ hadoop fs -ls /user/hadoop
[[hadoop@ip-172-31-28-225 ~]$ hadoop fs -mkdir /user/hadoop/eng_1M_1gram
[[hadoop@ip-172-31-28-225 ~]$ hadoop fs -ls /user/hadoop
Found 1 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2023-11-15 18:59 /user/hadoop/eng_1M_1gram
```

Moved the CSV file from the BrainStation S3 bucket straight into this new directory using hadoop distcp:



```
[[hadoop@ip-172-31-28-225 ~]$ hadoop distcp s3://brainstation-dsft/eng_1M_1gram.csv /user/hadoop/eng_1M_1gram/eng_1M_1gram.csv
2023-11-15 19:00:14,279 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=f
alse, append=false, useDiff=false, useRdiff=false, fromSnapshot=null, toSnapshot=null, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidt
h=0.0, copyStrategy='uniformsize', preserveStatus=[], atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[s3://brainstation-dsft/eng_1M_1gram.cs
v], targetPath=/user/hadoop/eng_1M_1gram/eng_1M_1gram.csv, filtersFile='null', blocksPerChunk=0, copyBufferSize=8192, verboseLog=false, directWrite=false, useiterat
or=false}, sourcePaths=[s3://brainstation-dsft/eng_1M_1gram.csv], targetPathExists=false, preserveRawXattrs=false
2023-11-15 19:00:14,519 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at ip-172-31-28-225.us-east-2.compute.internal/172.31.28.225:8
032
2023-11-15 19:00:14,692 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-28-225.us-east-2.compute.internal/172.31.28.225:10200
2023-11-15 19:00:18,615 INFO tools.SimpleCopyListing: Starting: Building listing using multi threaded approach for s3://brainstation-dsft/eng_1M_1gram.csv
2023-11-15 19:00:18,621 INFO tools.SimpleCopyListing: Building listing using multi threaded approach for s3://brainstation-dsft/eng_1M_1gram.csv: duration 0:00.002s
2023-11-15 19:00:18,750 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
2023-11-15 19:00:18,750 INFO tools.SimpleCopyListing: Build file listing completed.
2023-11-15 19:00:18,752 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
2023-11-15 19:00:18,752 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
2023-11-15 19:00:19,278 INFO tools.DistCp: Number of paths in the copy list: 1
2023-11-15 19:00:19,306 INFO tools.DistCp: Number of paths in the copy list: 1
2023-11-15 19:00:19,330 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at ip-172-31-28-225.us-east-2.compute.internal/172.31.28.225:8
032
2023-11-15 19:00:19,331 INFO client.AHSProxy: Connecting to Application History server at ip-172-31-28-225.us-east-2.compute.internal/172.31.28.225:10200
2023-11-15 19:00:19,450 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1700074338964_0001
2023-11-15 19:00:19,605 INFO mapreduce.JobSubmitter: number of splits:1
2023-11-15 19:00:19,794 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1700074338964_0001
2023-11-15 19:00:19,795 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-11-15 19:00:20,005 INFO conf.Configuration: resource-types.xml not found
2023-11-15 19:00:20,005 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-11-15 19:00:20,800 INFO impl.YarnClientImpl: Submitted application application_1700074338964_0001
2023-11-15 19:00:20,839 INFO mapreduce.Job: The url to track the job: http://ip-172-31-28-225.us-east-2.compute.internal:20888/proxy/application_1700074338964_0001/
2023-11-15 19:00:20,839 INFO tools.DistCp: DistCp job-id: job_1700074338964_0001
2023-11-15 19:00:20,840 INFO mapreduce.Job: Running job: job_1700074338964_0001
2023-11-15 19:00:27,932 INFO mapreduce.Job: Job job_1700074338964_0001 running in uber mode : false
2023-11-15 19:00:27,933 INFO mapreduce.Job:  map 0% reduce 0%
2023-11-15 19:00:46,022 INFO mapreduce.Job:  map 100% reduce 0%
```

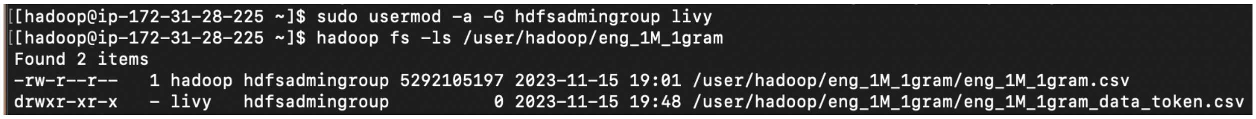
I can see that the file is showing within that directory:



```
[[hadoop@ip-172-31-28-225 ~]$ hadoop fs -ls /user/hadoop/eng_1M_1gram
Found 1 items
-rw-r--r--   1 hadoop hdfsadmingroup 5292105197 2023-11-15 19:01 /user/hadoop/eng_1M_1gram/eng_1M_1gram.csv
```

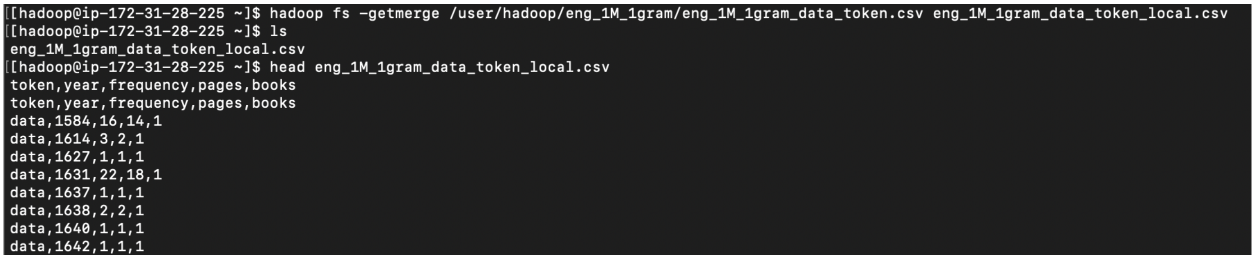
4. Refer to books_spark.ipynb notebook for steps taken to complete question 4.

   Needed to change permissions for user livy so that I had write access, then created a CSV from the filtered DataFrame in the pySpark notebook. I can see that it's now in this directory:



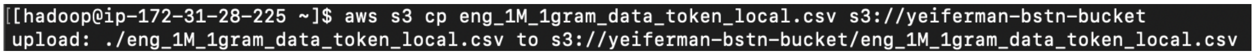
```
[[hadoop@ip-172-31-28-225 ~]$ sudo usermod -a -G hdfsadmingroup livy
[[hadoop@ip-172-31-28-225 ~]$ hadoop fs -ls /user/hadoop/eng_1M_1gram
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup 5292105197 2023-11-15 19:01 /user/hadoop/eng_1M_1gram/eng_1M_1gram.csv
drwxr-xr-x   - livy   hdfsadmingroup          0 2023-11-15 19:48 /user/hadoop/eng_1M_1gram/eng_1M_1gram_data_token.csv
```

5. Used getmerge command to merge all the files in that (apparently) directory I created into a single (actual) CSV file in the local node, then looked at the first five rows using the head command to make sure it worked:



```
[[hadoop@ip-172-31-28-225 ~]$ hadoop fs -getmerge /user/hadoop/eng_1M_1gram/eng_1M_1gram_data_token.csv eng_1M_1gram_data_token_local.csv
[[hadoop@ip-172-31-28-225 ~]$ ls
eng_1M_1gram_data_token_local.csv
[[hadoop@ip-172-31-28-225 ~]$ head eng_1M_1gram_data_token_local.csv
token,year,frequency,pages,books
token,year,frequency,pages,books
data,1584,16,14,1
data,1614,3,2,1
data,1627,1,1,1
data,1631,22,18,1
data,1637,1,1,1
data,1638,2,2,1
data,1640,1,1,1
data,1642,1,1,1
```

   Moved this CSV file from the head node to my s3 bucket:



```
[[hadoop@ip-172-31-28-225 ~]$ aws s3 cp eng_1M_1gram_data_token_local.csv s3://yeiferman-bstn-bucket
upload: ./eng_1M_1gram_data_token_local.csv to s3://yeiferman-bstn-bucket/eng_1M_1gram_data_token_local.csv
```

6. Refer to books_plot.ipynb notebook for steps taken to complete questions 6 and 7.
7. ^

8. Compare Hadoop and Spark as distributed file systems.
    a. What are the advantages/ differences between Hadoop and Spark? List two advantages for each.
        i. **Hadoop** - data replication is three-fold, making it fault tolerant and horizontally scalable for storing data
        ii. - compared to Spark, computation is less expensive since it’s done with mapReduce instead of using memory
        iii. **Spark** - primarily used for computing instead of storage, it stores intermediate data in memory to make computing faster
        iv. - provides a lot of functionality for data science workflow, such as SQL querying, machine learning, and streaming, and the ability to connect with python
    b. Explain how the HDFS stores the data.
        i. MapReduce. The task to be done is mapped to each block of data, then these outputs are aggregated, or reduced, into a single answer.