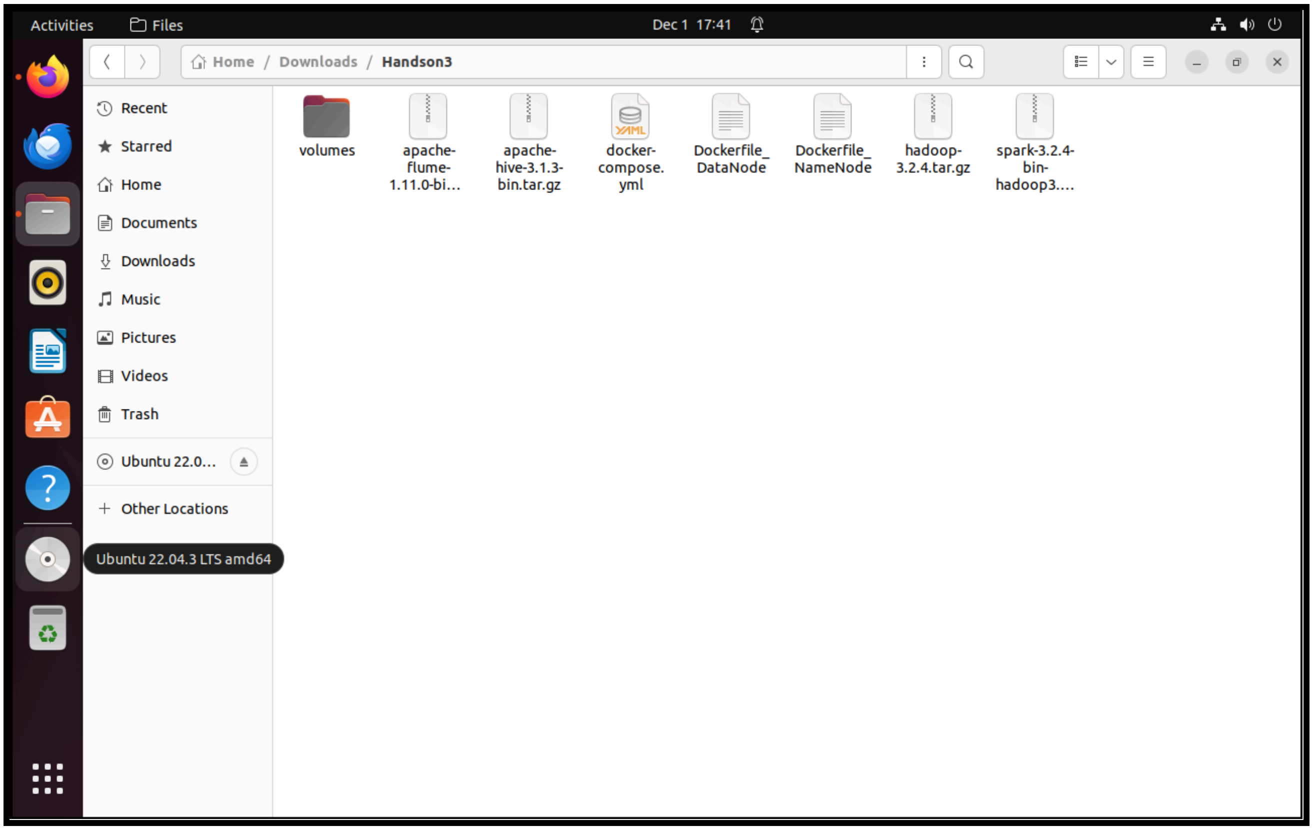
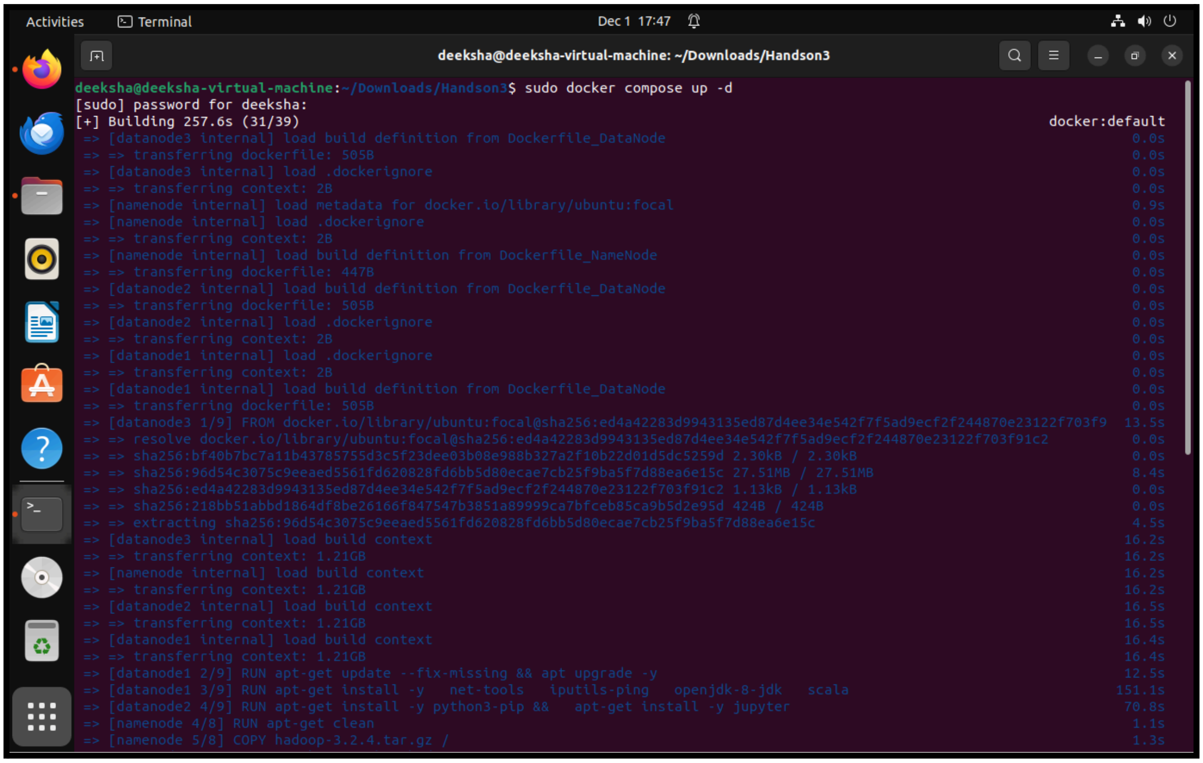
Electrical and Computer Engineering, Purdue University Northwest
Big Data (ECE49500/ECE59500)
Assignment 3
Name - Deeksha Hareesha Kulal

Task 1 [10 points] Hadoop, Spark, Hive, and Flume set-up. You must have to add relevant screenshots in the report for your work to get full credits for the task.

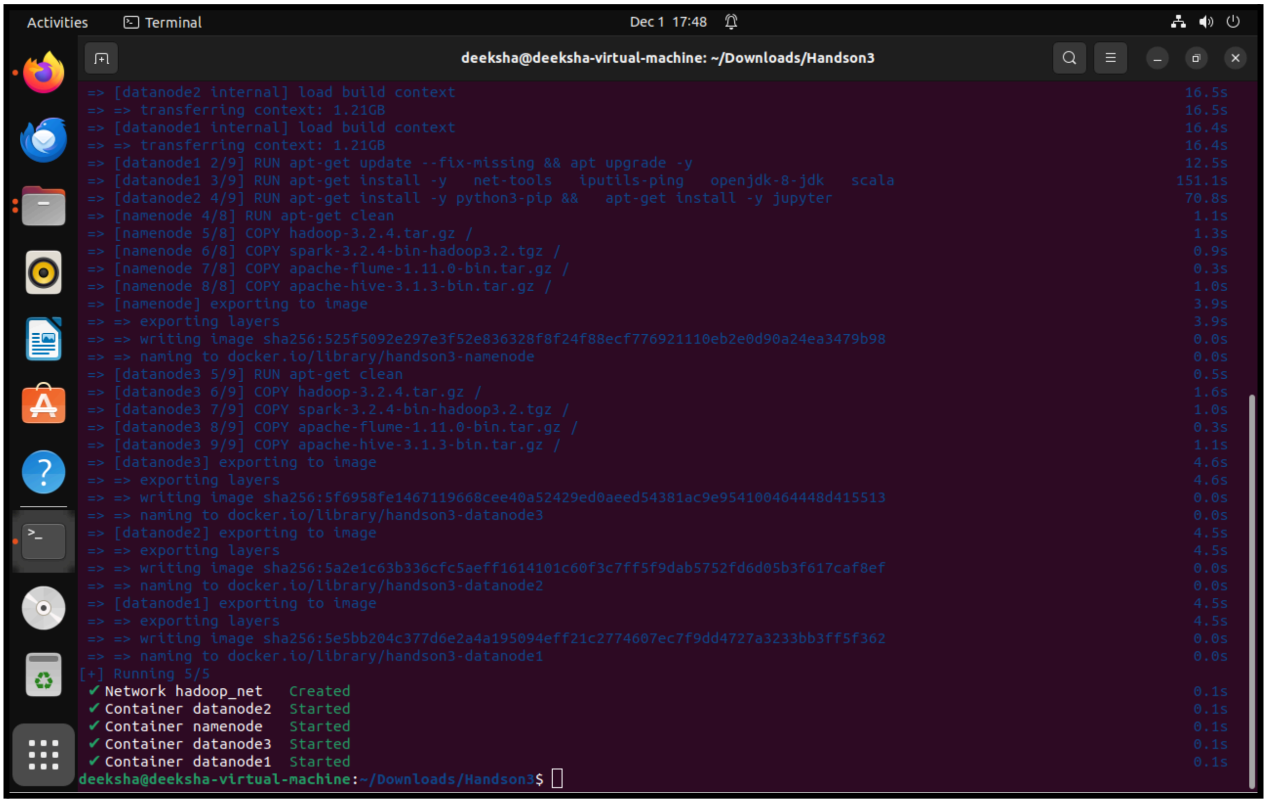
1. Download Handson3.zip file and extract the Handson3 folder from it.



2. Open the terminal to check-in to Handson3 folder and execute command: sudo docker compose up -d

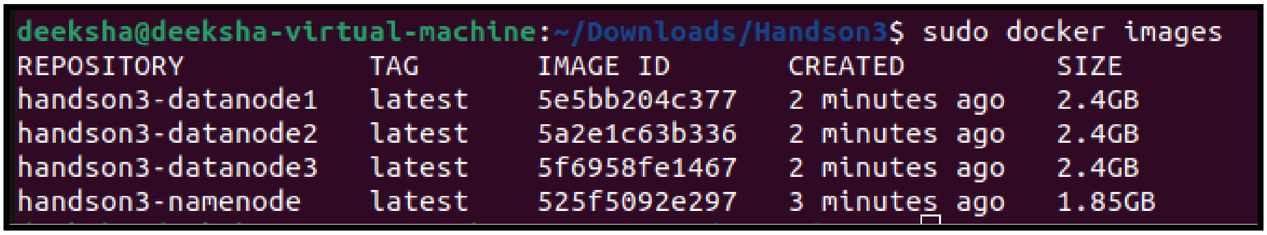


```
deeksha@deeksha-virtual-machine:~/Downloads/Handson3$ sudo docker compose up -d
[sudo] password for deeksha:
[+] Building 257.6s (31/39)                                                                                          docker:default
 => [datanode3 internal] load build definition from Dockerfile_DataNode                                                         0.0s
 => => transferring dockerfile: 505B                                                                                            0.0s
 => [datanode3 internal] load .dockerignore                                                                                     0.0s
 => => transferring context: 2B                                                                                                 0.0s
 => [namenode internal] load metadata for docker.io/library/ubuntu:focal                                                        0.9s
 => [namenode internal] load .dockerignore                                                                                      0.0s
 => => transferring context: 2B                                                                                                 0.0s
 => [namenode internal] load build definition from Dockerfile_NameNode                                                          0.0s
 => => transferring dockerfile: 447B                                                                                            0.0s
 => [datanode2 internal] load build definition from Dockerfile_DataNode                                                         0.0s
 => => transferring dockerfile: 505B                                                                                            0.0s
 => [datanode2 internal] load .dockerignore                                                                                     0.0s
 => => transferring context: 2B                                                                                                 0.0s
 => [datanode1 internal] load .dockerignore                                                                                     0.0s
 => => transferring context: 2B                                                                                                 0.0s
 => [datanode1 internal] load build definition from Dockerfile_DataNode                                                         0.0s
 => => transferring dockerfile: 505B                                                                                            0.0s
 => [datanode3 1/9] FROM docker.io/library/ubuntu:focal@sha256:ed4a42283d9943135ed87d4ee34e542f7f5ad9ecf2f244870e23122f703f9  13.5s
 => => resolve docker.io/library/ubuntu:focal@sha256:ed4a42283d9943135ed87d4ee34e542f7f5ad9ecf2f244870e23122f703f91c2          0.0s
 => => sha256:bf40b7bc7a11b43785755d3c5f23dee03b08e988b327a2f10b22d01d5dc5259d 2.30kB / 2.30kB                                  0.0s
 => => sha256:96d54c3075c9eeaed5561fd620828fd6bb5d80ecae7cb25f9ba5f7d88ea6e15c 27.51MB / 27.51MB                                8.4s
 => => sha256:ed4a42283d9943135ed87d4ee34e542f7f5ad9ecf2f244870e23122f703f91c2 1.13kB / 1.13kB                                  0.0s
 => => sha256:218bb51abbd1864df8be26166f847547b3851a89999ca7bfceb85ca9b5d2e95d 424B / 424B                                      0.0s
 => => extracting sha256:96d54c3075c9eeaed5561fd620828fd6bb5d80ecae7cb25f9ba5f7d88ea6e15c                                       4.5s
 => [datanode3 internal] load build context                                                                                    16.2s
 => => transferring context: 1.21GB                                                                                            16.2s
 => [namenode internal] load build context                                                                                     16.2s
 => => transferring context: 1.21GB                                                                                            16.2s
 => [datanode2 internal] load build context                                                                                    16.5s
 => => transferring context: 1.21GB                                                                                            16.5s
 => [datanode1 internal] load build context                                                                                    16.4s
 => => transferring context: 1.21GB                                                                                            16.4s
 => [datanode1 2/9] RUN apt-get update --fix-missing && apt upgrade -y                                                         12.5s
 => [datanode1 3/9] RUN apt-get install -y   net-tools   iputils-ping   openjdk-8-jdk   scala                                 151.1s
 => [datanode2 4/9] RUN apt-get install -y python3-pip &&   apt-get install -y jupyter                                         70.8s
 => [namenode 4/8] RUN apt-get clean                                                                                            1.1s
 => [namenode 5/8] COPY hadoop-3.2.4.tar.gz /                                                                                   1.3s
```



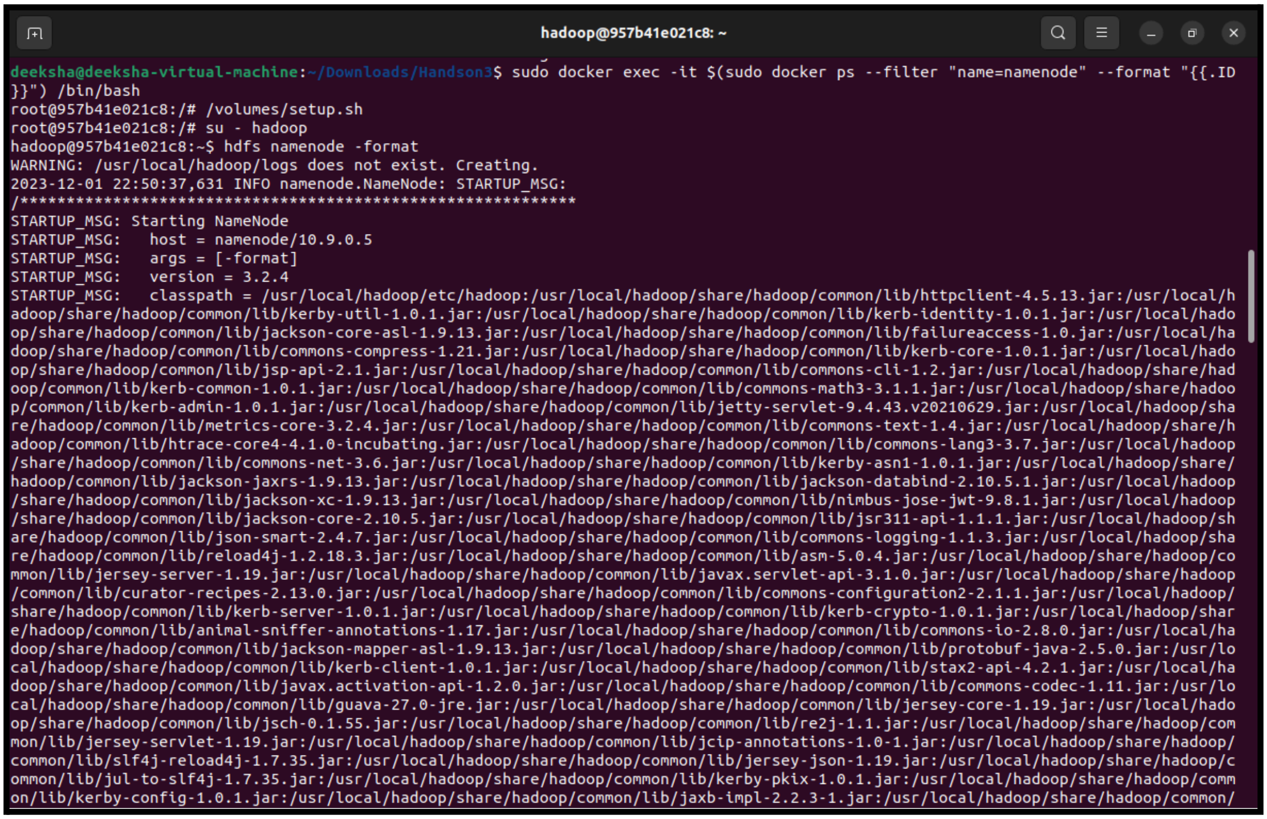
```
 => [datanode2 internal] load build context                                                                                    16.5s
 => => transferring context: 1.21GB                                                                                            16.5s
 => [datanode1 internal] load build context                                                                                    16.4s
 => => transferring context: 1.21GB                                                                                            16.4s
 => [datanode1 2/9] RUN apt-get update --fix-missing && apt upgrade -y                                                         12.5s
 => [datanode1 3/9] RUN apt-get install -y   net-tools   iputils-ping   openjdk-8-jdk   scala                                 151.1s
 => [datanode2 4/9] RUN apt-get install -y python3-pip &&   apt-get install -y jupyter                                         70.8s
 => [namenode 4/8] RUN apt-get clean                                                                                            1.1s
 => [namenode 5/8] COPY hadoop-3.2.4.tar.gz /                                                                                   1.3s
 => [namenode 6/8] COPY spark-3.2.4-bin-hadoop3.2.tgz /                                                                         0.9s
 => [namenode 7/8] COPY apache-flume-1.11.0-bin.tar.gz /                                                                        0.3s
 => [namenode 8/8] COPY apache-hive-3.1.3-bin.tar.gz /                                                                          1.0s
 => [namenode] exporting to image                                                                                               3.9s
 => => exporting layers                                                                                                         3.9s
 => => writing image sha256:525f5092e297e3f52e836328f8f24f88ecf776921110eb2e0d90a24ea3479b98                                    0.0s
 => => naming to docker.io/library/handson3-namenode                                                                            0.0s
 => [datanode3 5/9] RUN apt-get clean                                                                                           0.5s
 => [datanode3 6/9] COPY hadoop-3.2.4.tar.gz /                                                                                  1.6s
 => [datanode3 7/9] COPY spark-3.2.4-bin-hadoop3.2.tgz /                                                                        1.0s
 => [datanode3 8/9] COPY apache-flume-1.11.0-bin.tar.gz /                                                                       0.3s
 => [datanode3 9/9] COPY apache-hive-3.1.3-bin.tar.gz /                                                                         1.1s
 => [datanode3] exporting to image                                                                                              4.6s
 => => exporting layers                                                                                                         4.6s
 => => writing image sha256:5f6958fe1467119668cee40a52429ed0aeed54381ac9e954100464448d415513                                    0.0s
 => => naming to docker.io/library/handson3-datanode3                                                                           0.0s
 => [datanode2] exporting to image                                                                                              4.5s
 => => exporting layers                                                                                                         4.5s
 => => writing image sha256:5a2e1c63b336cfc5aeff1614101c60f3c7ff5f9dab5752fd6d05b3f617caf8ef                                    0.0s
 => => naming to docker.io/library/handson3-datanode2                                                                           0.0s
 => [datanode1] exporting to image                                                                                              4.5s
 => => exporting layers                                                                                                         4.5s
 => => writing image sha256:5e5bb204c377d6e2a4a195094eff21c2774607ec7f9dd4727a3233bb3ff5f362                                    0.0s
 => => naming to docker.io/library/handson3-datanode1                                                                           0.0s
[+] Running 5/5
 ✔ Network hadoop_net   Created                                                                                                 0.1s
 ✔ Container datanode2  Started                                                                                                 0.1s
 ✔ Container namenode   Started                                                                                                 0.1s
 ✔ Container datanode3  Started                                                                                                 0.1s
 ✔ Container datanode1  Started                                                                                                 0.1s
deeksha@deeksha-virtual-machine:~/Downloads/Handson3$ 
```

3. Execute the command: sudo docker images



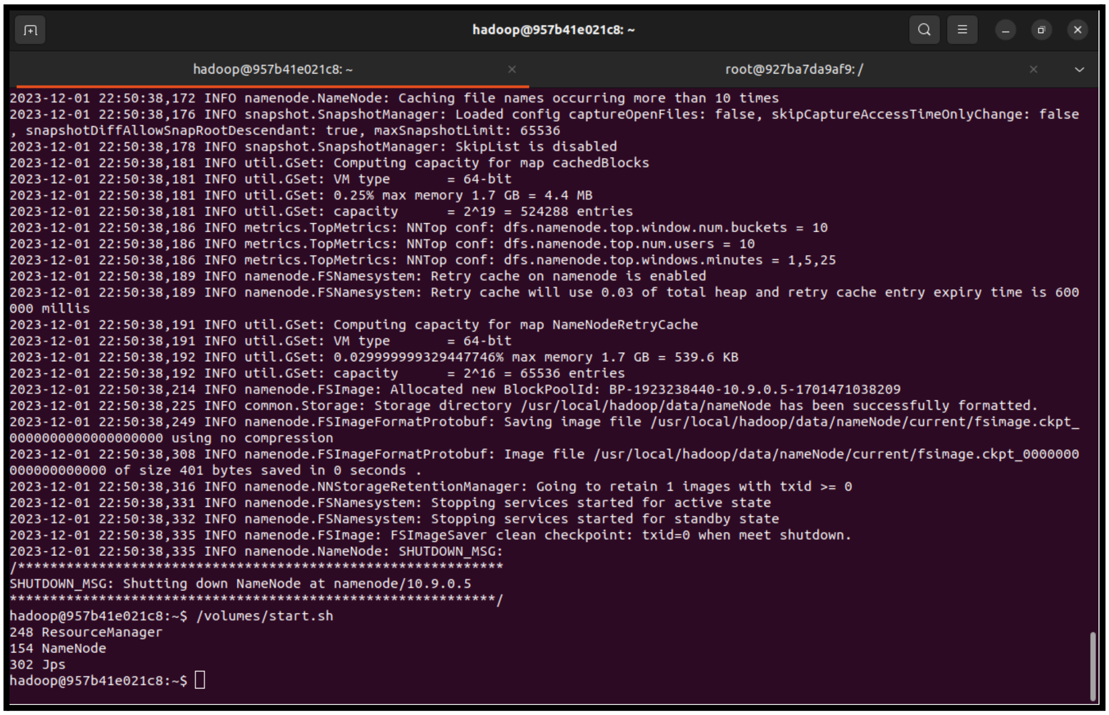
```
deeksha@deeksha-virtual-machine:~/Downloads/Handson3$ sudo docker images
REPOSITORY          TAG       IMAGE ID       CREATED         SIZE
handson3-datanode1  latest    5e5bb204c377   2 minutes ago   2.4GB
handson3-datanode2  latest    5a2e1c63b336   2 minutes ago   2.4GB
handson3-datanode3  latest    5f6958fe1467   2 minutes ago   2.4GB
handson3-namenode   latest    525f5092e297   3 minutes ago   1.85GB
```

4. Open terminal tabs to access containers' shells. Run following command at each tab to access a container's shell. Replace with the actual container's name. sudo docker exec -it $(sudo docker ps --filter "name=" --format "{{.ID}}") /bin/bash



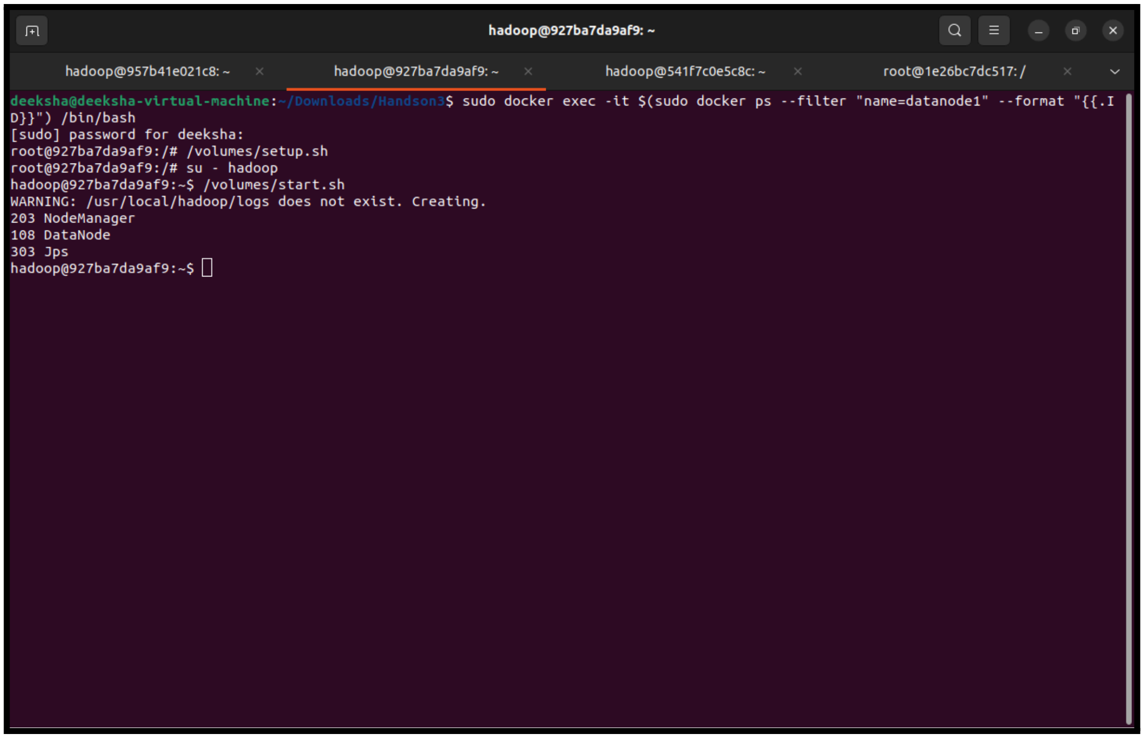
```
deeksha@deeksha-virtual-machine:~/Downloads/Handson3$ sudo docker exec -it $(sudo docker ps --filter "name=namenode" --format "{{.ID
}}") /bin/bash
root@957b41e021c8:/# /volumes/setup.sh
root@957b41e021c8:/# su - hadoop
hadoop@957b41e021c8:~$ hdfs namenode -format
WARNING: /usr/local/hadoop/logs does not exist. Creating.
2023-12-01 22:50:37,631 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = namenode/10.9.0.5
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.2.4
STARTUP_MSG:   classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/hadoop/common/lib/httpclient-4.5.13.jar:/usr/local/h
adoop/share/hadoop/common/lib/kerby-util-1.0.1.jar:/usr/local/hadoop/share/hadoop/common/lib/kerb-identity-1.0.1.jar:/usr/local/hado
op/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/failureaccess-1.0.jar:/usr/local/ha
doop/share/hadoop/common/lib/commons-compress-1.21.jar:/usr/local/hadoop/share/hadoop/common/lib/kerb-core-1.0.1.jar:/usr/local/hado
op/share/hadoop/common/lib/jsp-api-2.1.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-cli-1.2.jar:/usr/local/hadoop/share/had
oop/common/lib/kerb-common-1.0.1.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-math3-3.1.1.jar:/usr/local/hadoop/share/hadoo
p/common/lib/kerb-admin-1.0.1.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-servlet-9.4.43.v20210629.jar:/usr/local/hadoop/sha
re/hadoop/common/lib/metrics-core-3.2.4.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-text-1.4.jar:/usr/local/hadoop/share/h
adoop/common/lib/htrace-core4-4.1.0-incubating.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-lang3-3.7.jar:/usr/local/hadoop
/share/hadoop/common/lib/commons-net-3.6.jar:/usr/local/hadoop/share/hadoop/common/lib/kerby-asn1-1.0.1.jar:/usr/local/hadoop/share/
hadoop/common/lib/jackson-jaxrs-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-databind-2.10.5.1.jar:/usr/local/hadoop
/share/hadoop/common/lib/jackson-xc-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/nimbus-jose-jwt-9.8.1.jar:/usr/local/hadoop
/share/hadoop/common/lib/jackson-core-2.10.5.jar:/usr/local/hadoop/share/hadoop/common/lib/jsr311-api-1.1.1.jar:/usr/local/hadoop/sh
are/hadoop/common/lib/json-smart-2.4.7.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-logging-1.1.3.jar:/usr/local/hadoop/sha
re/hadoop/common/lib/reload4j-1.2.18.3.jar:/usr/local/hadoop/share/hadoop/common/lib/asm-5.0.4.jar:/usr/local/hadoop/share/hadoop/co
mmon/lib/jersey-server-1.19.jar:/usr/local/hadoop/share/hadoop/common/lib/javax.servlet-api-3.1.0.jar:/usr/local/hadoop/share/hadoop
/common/lib/curator-recipes-2.13.0.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-configuration2-2.1.1.jar:/usr/local/hadoop/
share/hadoop/common/lib/kerb-server-1.0.1.jar:/usr/local/hadoop/share/hadoop/common/lib/kerb-crypto-1.0.1.jar:/usr/local/hadoop/shar
e/hadoop/common/lib/animal-sniffer-annotations-1.17.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-io-2.8.0.jar:/usr/local/ha
doop/share/hadoop/common/lib/jackson-mapper-asl-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/protobuf-java-2.5.0.jar:/usr/lo
cal/hadoop/share/hadoop/common/lib/kerb-client-1.0.1.jar:/usr/local/hadoop/share/hadoop/common/lib/stax2-api-4.2.1.jar:/usr/local/ha
doop/share/hadoop/common/lib/javax.activation-api-1.2.0.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-codec-1.11.jar:/usr/lo
cal/hadoop/share/hadoop/common/lib/guava-27.0-jre.jar:/usr/local/hadoop/share/hadoop/common/lib/jersey-core-1.19.jar:/usr/local/hado
op/share/hadoop/common/lib/jsch-0.1.55.jar:/usr/local/hadoop/share/hadoop/common/lib/re2j-1.1.jar:/usr/local/hadoop/share/hadoop/com
mon/lib/jersey-servlet-1.19.jar:/usr/local/hadoop/share/hadoop/common/lib/jcip-annotations-1.0-1.jar:/usr/local/hadoop/share/hadoop/
common/lib/slf4j-reload4j-1.7.35.jar:/usr/local/hadoop/share/hadoop/common/lib/jersey-json-1.19.jar:/usr/local/hadoop/share/hadoop/c
ommon/lib/jul-to-slf4j-1.7.35.jar:/usr/local/hadoop/share/hadoop/common/lib/kerby-pkix-1.0.1.jar:/usr/local/hadoop/share/hadoop/comm
on/lib/kerby-config-1.0.1.jar:/usr/local/hadoop/share/hadoop/common/lib/jaxb-impl-2.2.3-1.jar:/usr/local/hadoop/share/hadoop/common/
```

5. At namenode container, execute the following commands: /volumes/setup.sh (Note: Run this command when you access the container first time) su - hadoop hdfs namenode -format (Note: Run this command only when you access the container first time) /volumes/start.sh



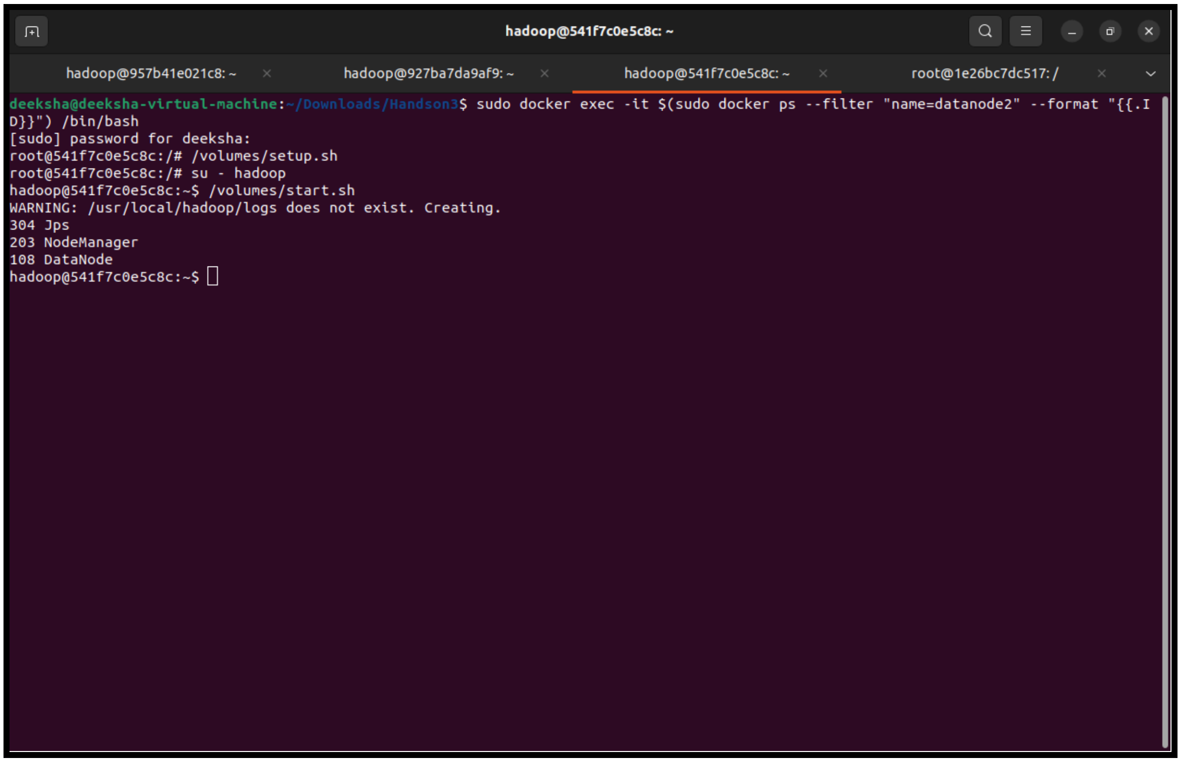
```
2023-12-01 22:50:38,172 INFO namenode.NameNode: Caching file names occurring more than 10 times
2023-12-01 22:50:38,176 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: false, skipCaptureAccessTimeOnlyChange: false
, snapshotDiffAllowSnapRootDescendant: true, maxSnapshotLimit: 65536
2023-12-01 22:50:38,178 INFO snapshot.SnapshotManager: SkipList is disabled
2023-12-01 22:50:38,181 INFO util.GSet: Computing capacity for map cachedBlocks
2023-12-01 22:50:38,181 INFO util.GSet: VM type       = 64-bit
2023-12-01 22:50:38,181 INFO util.GSet: 0.25% max memory 1.7 GB = 4.4 MB
2023-12-01 22:50:38,181 INFO util.GSet: capacity      = 2^19 = 524288 entries
2023-12-01 22:50:38,186 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2023-12-01 22:50:38,186 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2023-12-01 22:50:38,186 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2023-12-01 22:50:38,189 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2023-12-01 22:50:38,189 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600
000 millis
2023-12-01 22:50:38,191 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2023-12-01 22:50:38,191 INFO util.GSet: VM type       = 64-bit
2023-12-01 22:50:38,192 INFO util.GSet: 0.029999999329447746% max memory 1.7 GB = 539.6 KB
2023-12-01 22:50:38,192 INFO util.GSet: capacity      = 2^16 = 65536 entries
2023-12-01 22:50:38,214 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1923238440-10.9.0.5-1701471038209
2023-12-01 22:50:38,225 INFO common.Storage: Storage directory /usr/local/hadoop/data/nameNode has been successfully formatted.
2023-12-01 22:50:38,249 INFO namenode.FSImageFormatProtobuf: Saving image file /usr/local/hadoop/data/nameNode/current/fsimage.ckpt_
0000000000000000000 using no compression
2023-12-01 22:50:38,308 INFO namenode.FSImageFormatProtobuf: Image file /usr/local/hadoop/data/nameNode/current/fsimage.ckpt_0000000
000000000000 of size 401 bytes saved in 0 seconds .
2023-12-01 22:50:38,316 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2023-12-01 22:50:38,331 INFO namenode.FSNamesystem: Stopping services started for active state
2023-12-01 22:50:38,332 INFO namenode.FSNamesystem: Stopping services started for standby state
2023-12-01 22:50:38,335 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2023-12-01 22:50:38,335 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at namenode/10.9.0.5
************************************************************/
hadoop@957b41e021c8:~$ /volumes/start.sh
248 ResourceManager
154 NameNode
302 Jps
hadoop@957b41e021c8:~$ 
```

6. At each datanode container, execute the following commands: /volumes/setup.sh (Note: Run this command only when you access the container first time) su - hadoop /volumes/start.sh



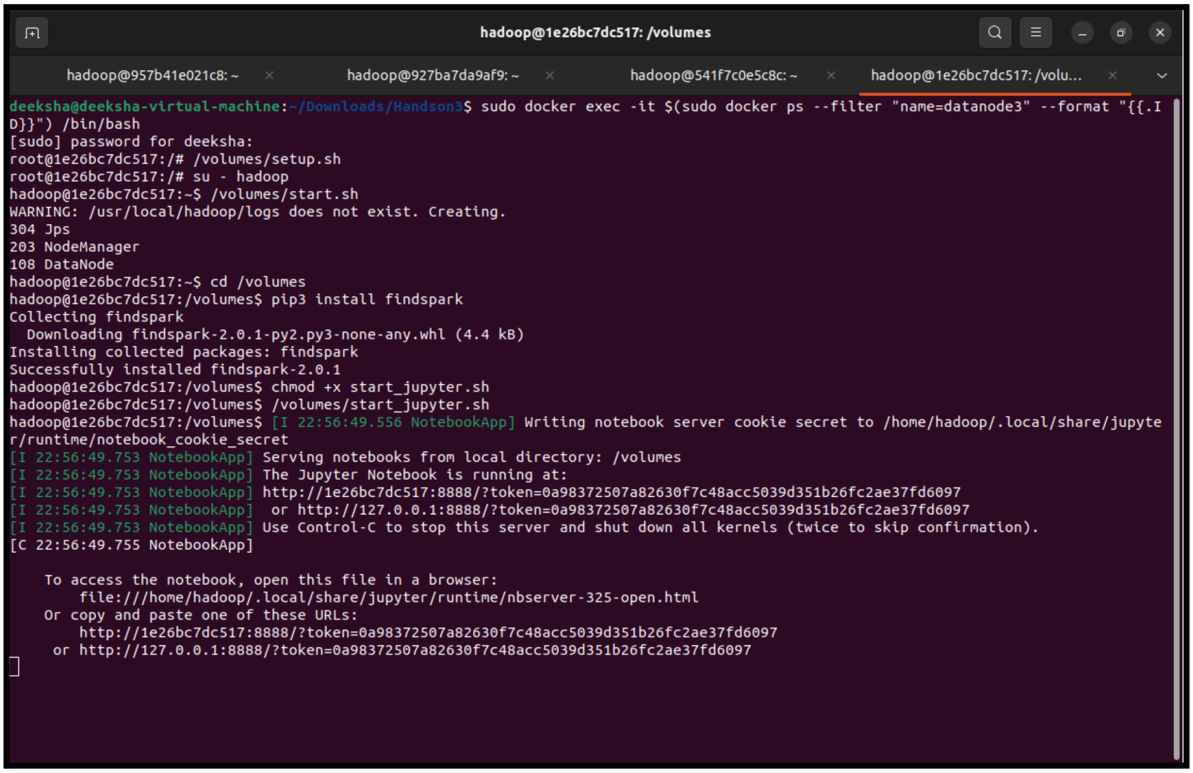
```
deeksha@deeksha-virtual-machine:~/Downloads/Handson3$ sudo docker exec -it $(sudo docker ps --filter "name=datanode1" --format "{{.I
D}}") /bin/bash
[sudo] password for deeksha:
root@927ba7da9af9:/# /volumes/setup.sh
root@927ba7da9af9:/# su - hadoop
hadoop@927ba7da9af9:~$ /volumes/start.sh
WARNING: /usr/local/hadoop/logs does not exist. Creating.
203 NodeManager
108 DataNode
303 Jps
hadoop@927ba7da9af9:~$ 
```

)



```
deeksha@deeksha-virtual-machine:~/Downloads/Handson3$ sudo docker exec -it $(sudo docker ps --filter "name=datanode2" --format "{{.I
D}}") /bin/bash
[sudo] password for deeksha:
root@541f7c0e5c8c:/# /volumes/setup.sh
root@541f7c0e5c8c:/# su - hadoop
hadoop@541f7c0e5c8c:~$ /volumes/start.sh
WARNING: /usr/local/hadoop/logs does not exist. Creating.
304 Jps
203 NodeManager
108 DataNode
hadoop@541f7c0e5c8c:~$ 
```

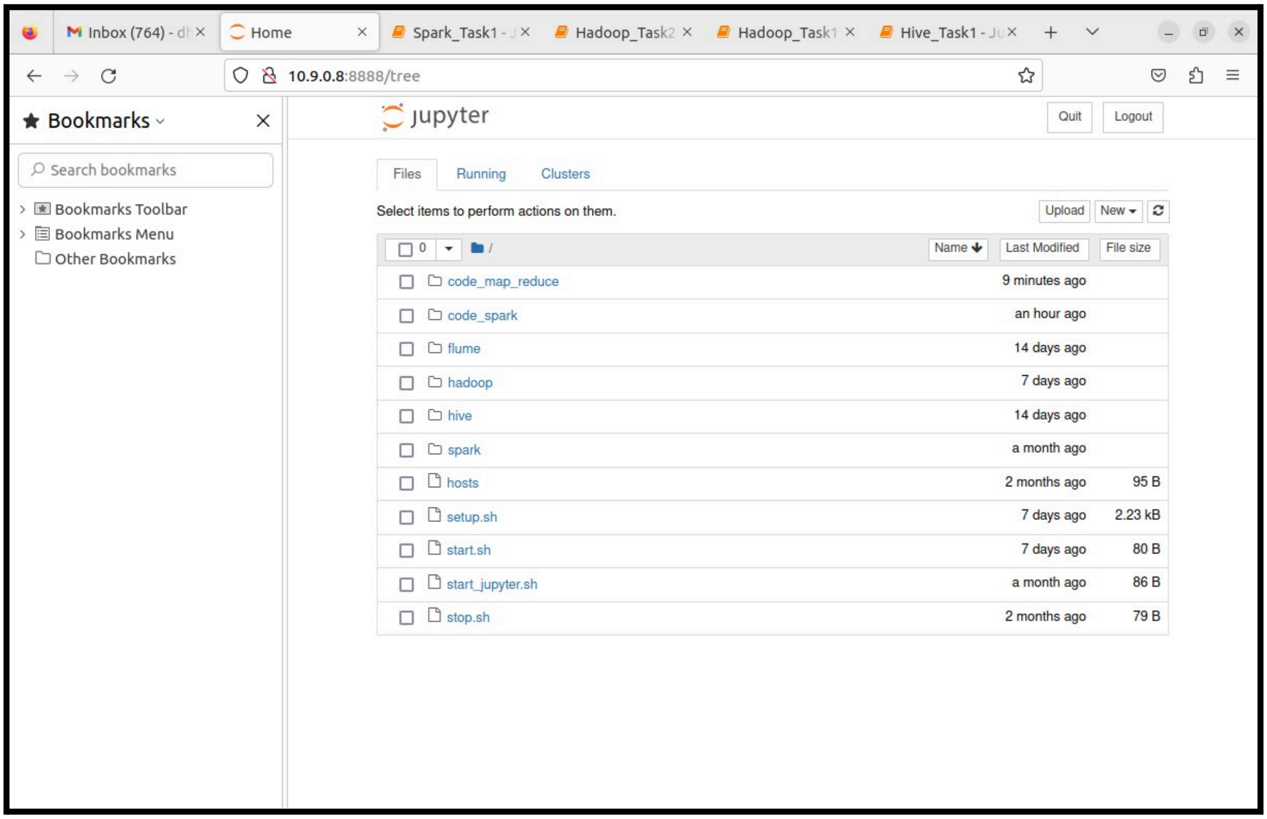
7. Go to any datanode's shell, (e.g. datanode1) and execute the following commands: cd /volumes pip3 install findspark chmod +x start_jupyter.sh /volumes/start_jupyter.sh



```
deeksha@deeksha-virtual-machine:~/Downloads/Handson3$ sudo docker exec -it $(sudo docker ps --filter "name=datanode3" --format "{{.I
D}}") /bin/bash
[sudo] password for deeksha:
root@1e26bc7dc517:/# /volumes/setup.sh
root@1e26bc7dc517:/# su - hadoop
hadoop@1e26bc7dc517:~$ /volumes/start.sh
WARNING: /usr/local/hadoop/logs does not exist. Creating.
304 Jps
203 NodeManager
108 DataNode
hadoop@1e26bc7dc517:~$ cd /volumes
hadoop@1e26bc7dc517:/volumes$ pip3 install findspark
Collecting findspark
  Downloading findspark-2.0.1-py2.py3-none-any.whl (4.4 kB)
Installing collected packages: findspark
Successfully installed findspark-2.0.1
hadoop@1e26bc7dc517:/volumes$ chmod +x start_jupyter.sh
hadoop@1e26bc7dc517:/volumes$ /volumes/start_jupyter.sh
hadoop@1e26bc7dc517:/volumes$ [I 22:56:49.556 NotebookApp] Writing notebook server cookie secret to /home/hadoop/.local/share/jupyte
r/runtime/notebook_cookie_secret
[I 22:56:49.753 NotebookApp] Serving notebooks from local directory: /volumes
[I 22:56:49.753 NotebookApp] The Jupyter Notebook is running at:
[I 22:56:49.753 NotebookApp] http://1e26bc7dc517:8888/?token=0a98372507a82630f7c48acc5039d351b26fc2ae37fd6097
[I 22:56:49.753 NotebookApp]  or http://127.0.0.1:8888/?token=0a98372507a82630f7c48acc5039d351b26fc2ae37fd6097
[I 22:56:49.753 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 22:56:49.755 NotebookApp]

    To access the notebook, open this file in a browser:
        file:///home/hadoop/.local/share/jupyter/runtime/nbserver-325-open.html
    Or copy and paste one of these URLs:
        http://1e26bc7dc517:8888/?token=0a98372507a82630f7c48acc5039d351b26fc2ae37fd6097
     or http://127.0.0.1:8888/?token=0a98372507a82630f7c48acc5039d351b26fc2ae37fd6097
```

8. Copy the Jupyter URL and paste it into a browser in your host machine or VM. Replace the IP 127.0.0.1 with the IP of the datanode. It will bring the Jupyter Notebook. Go to code_map_reduce directory, and complete Task1.pynb file and submit the Notebook file.



Select items to perform actions on them.

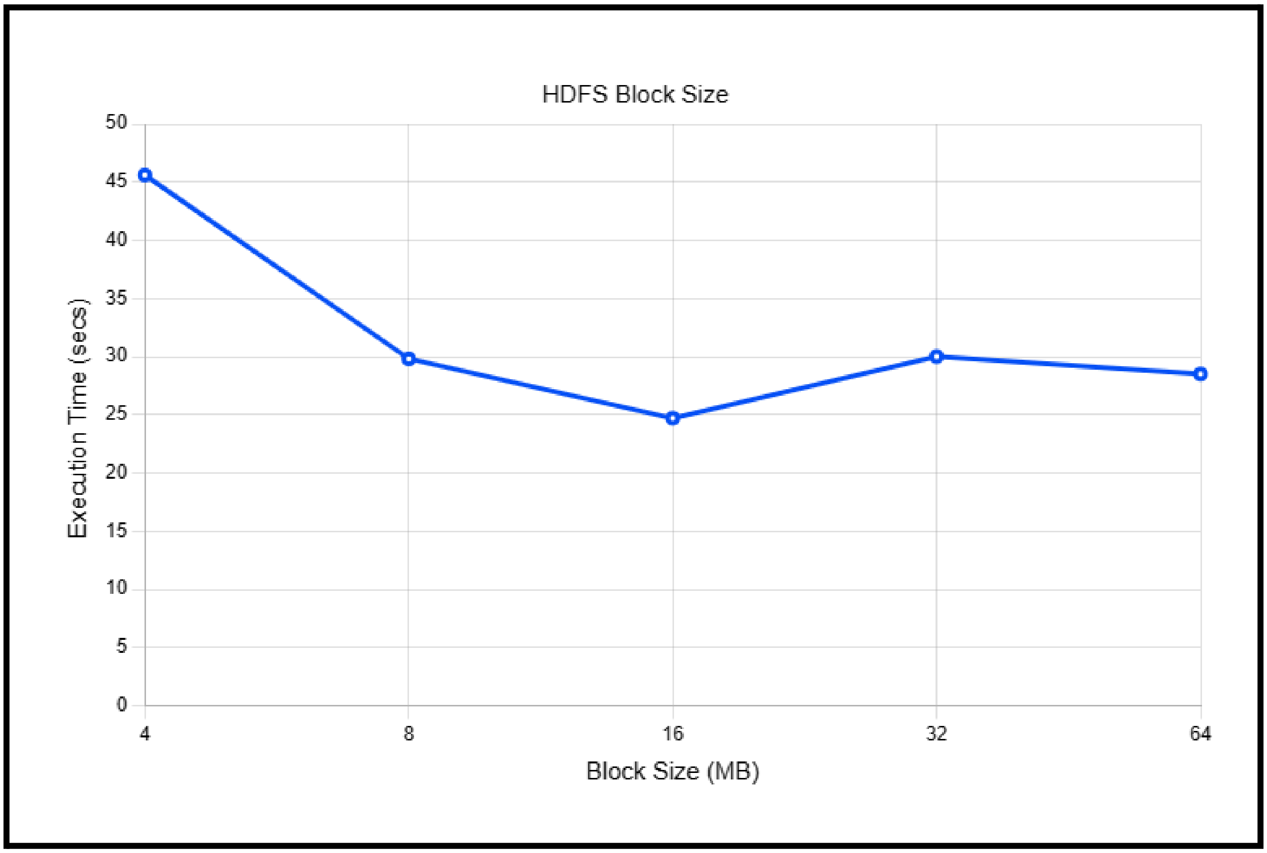
Name	Last Modified	File size
code_map_reduce	9 minutes ago	
code_spark	an hour ago	
flume	14 days ago	
hadoop	7 days ago	
hive	14 days ago	
spark	a month ago	
hosts	2 months ago	95 B
setup.sh	7 days ago	2.23 kB
start.sh	7 days ago	80 B
start_jupyter.sh	a month ago	86 B
stop.sh	2 months ago	79 B

Task 2 [15 points] Go to code_map_reduce directory and complete the tasks given in Hadoop_Task1.ipynb file and submit the Notebook file.
File submitted

Task 3 [15 points] Complete the tasks given in Hadoop_Task2.ipynb file, submit the Notebook file, and provide the comparison charts for execution times for following cases:
File submitted
a. Number of workers = 3, HDFS block size = 4MB, 8MB, 16MB, 32MB, and 64MB (File used - books64.txt)

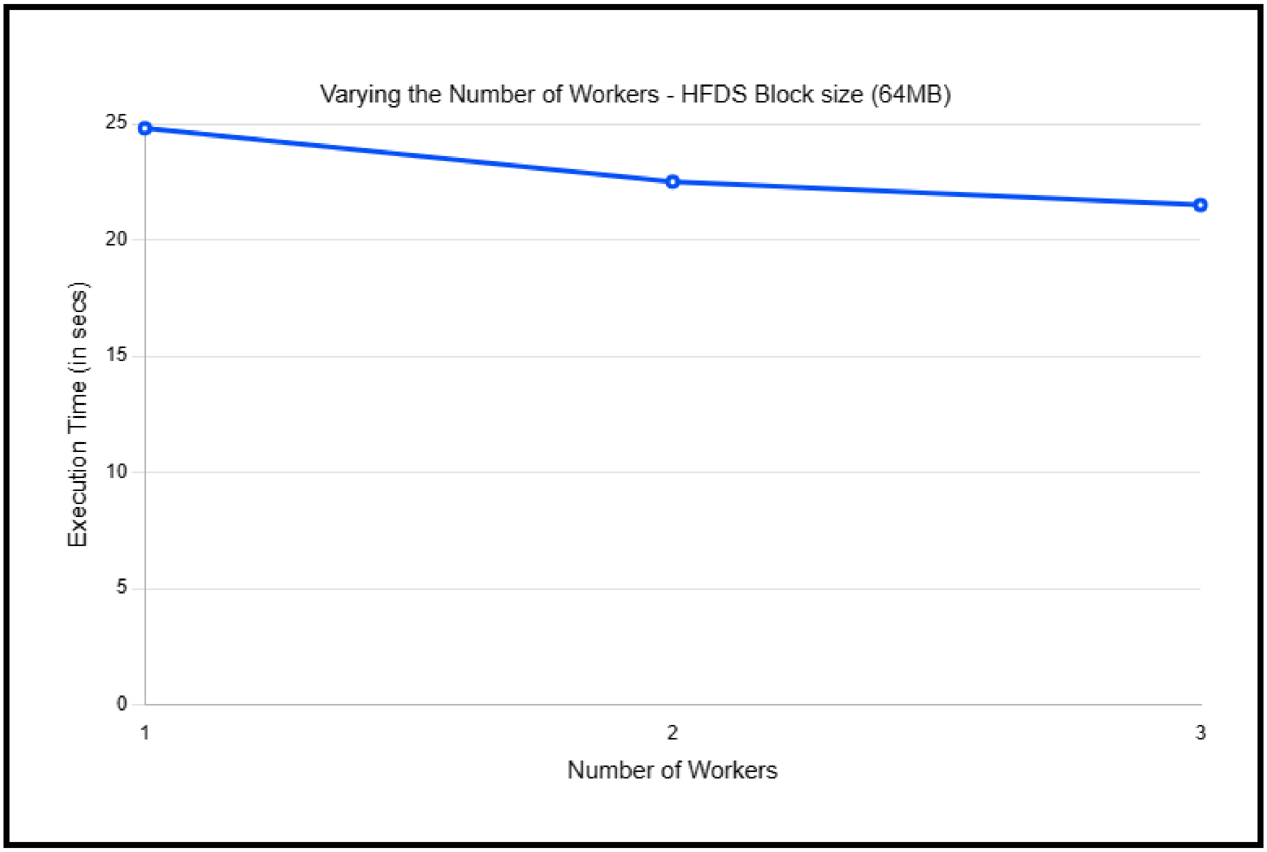
HDFS Block Size	**Execution time**
4MB	**45.6 secs**
8MB	**29.8 secs**
16MB	**24.7 secs**

32MB	**30 secs**
64MB	**28.5 secs**



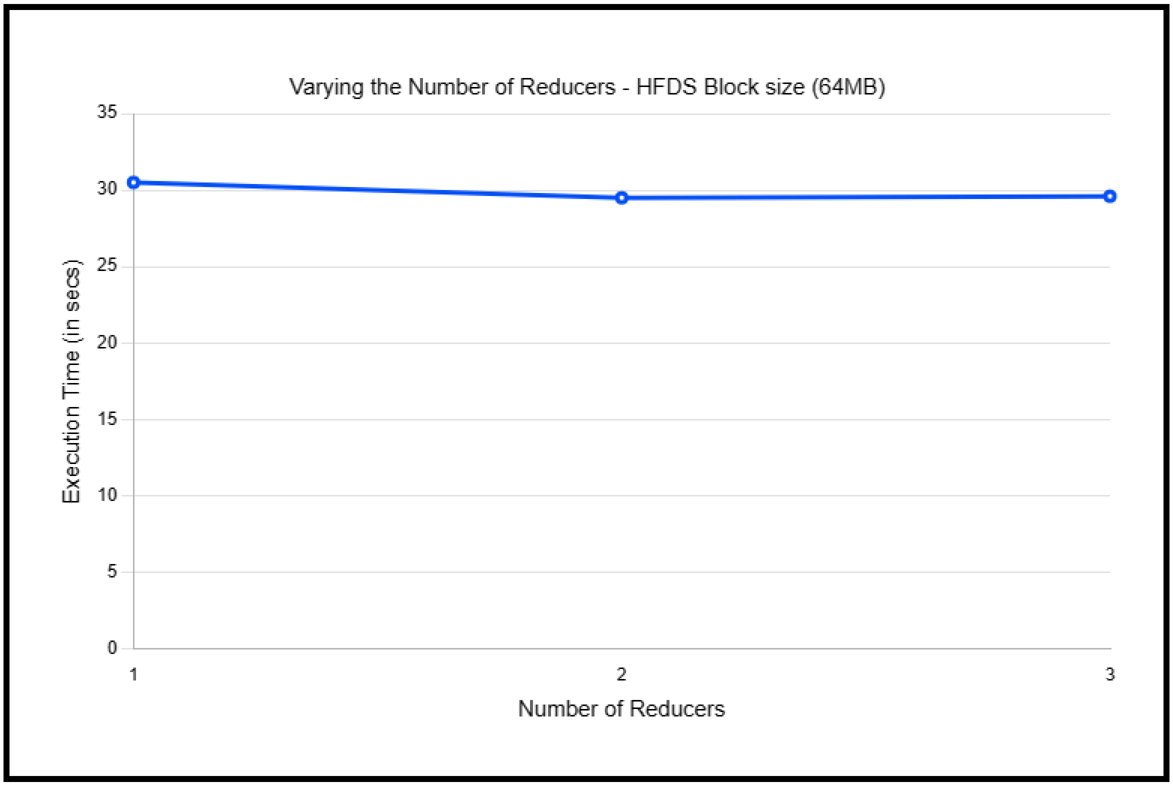
b. HDFS block size = 64MB, number of workers = 1, 2, and 3 (Optional for ECE 49500)
(File used - books32.txt hence block size of 32MB taken)

No of Workers	**Execution Time**
1	**24.8 secs**
2	**22.5 secs**
3	**21.5 secs**



c. HDFS block size = 64MB, number of workers = 3, number of reducers = 1, 2, and 3 (File used - books64.txt)

No of Reducers	**Execution Time**
1	**30.5 secs**
2	**29.5 secs**
3	**29.6 secs**



Task 4 [10 points] Complete the tasks given in Hive_Task1.ipynb file and submit the Notebook file.

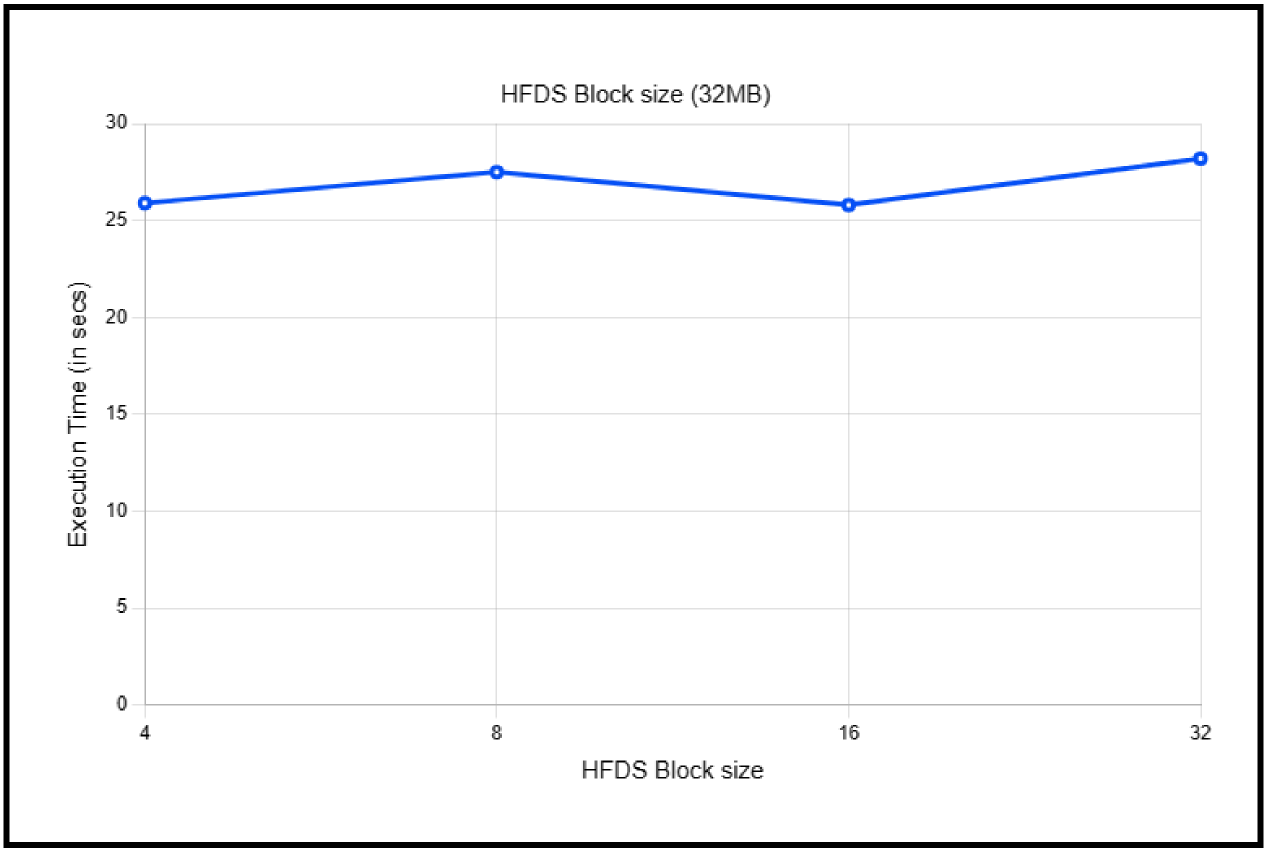
File submitted

Task 5 [10 points] Complete the tasks given in Spark_Task1.ipynb file, submit the Notebook file, and provide the comparison charts for execution times the following cases:

File submitted

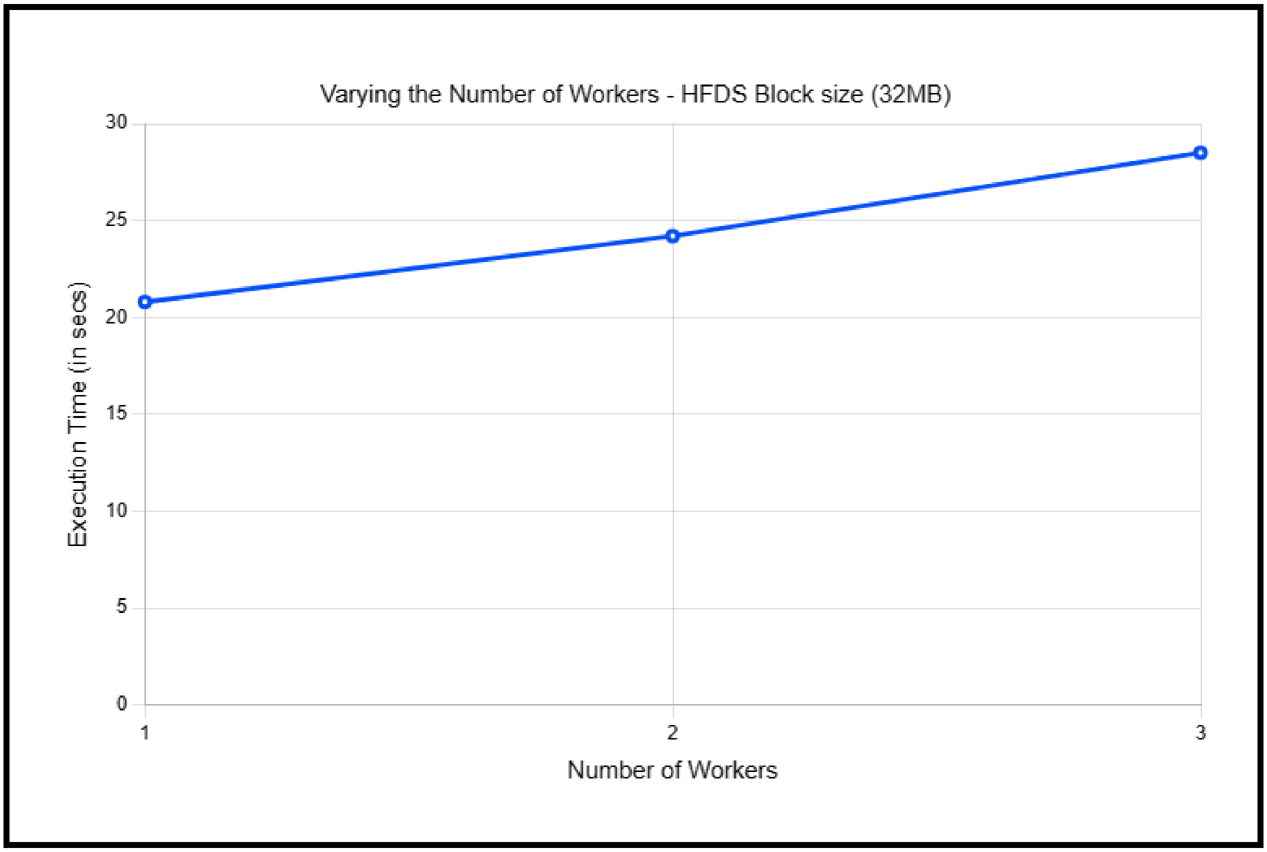
a. Number of workers = 3, HDFS block size = 4MB, 8MB, 16MB, 32MB, and 64MB (File used - books32.txt hence block size of 64 MB not considered)

HDFS Block Size	**Execution time**
4MB	**25.9 secs**
8MB	**27.5 secs**
16MB	**25.8 secs**
32MB	**28.2 secs**



b. HDFS block size = 64MB, number of workers = 1, 2, and 3 (Optional for ECE 49500) (HDFS block size of 32MB used with books32.txt file)

No of Workers	Execution Time
1	**20.8 secs**
2	**24.2 secs**
3	**28.5 secs**



Task 6 [10 points] Go to code_spark directory and complete the tasks given in Spark_Task2.ipynb file and submit the Notebook file.

File submitted