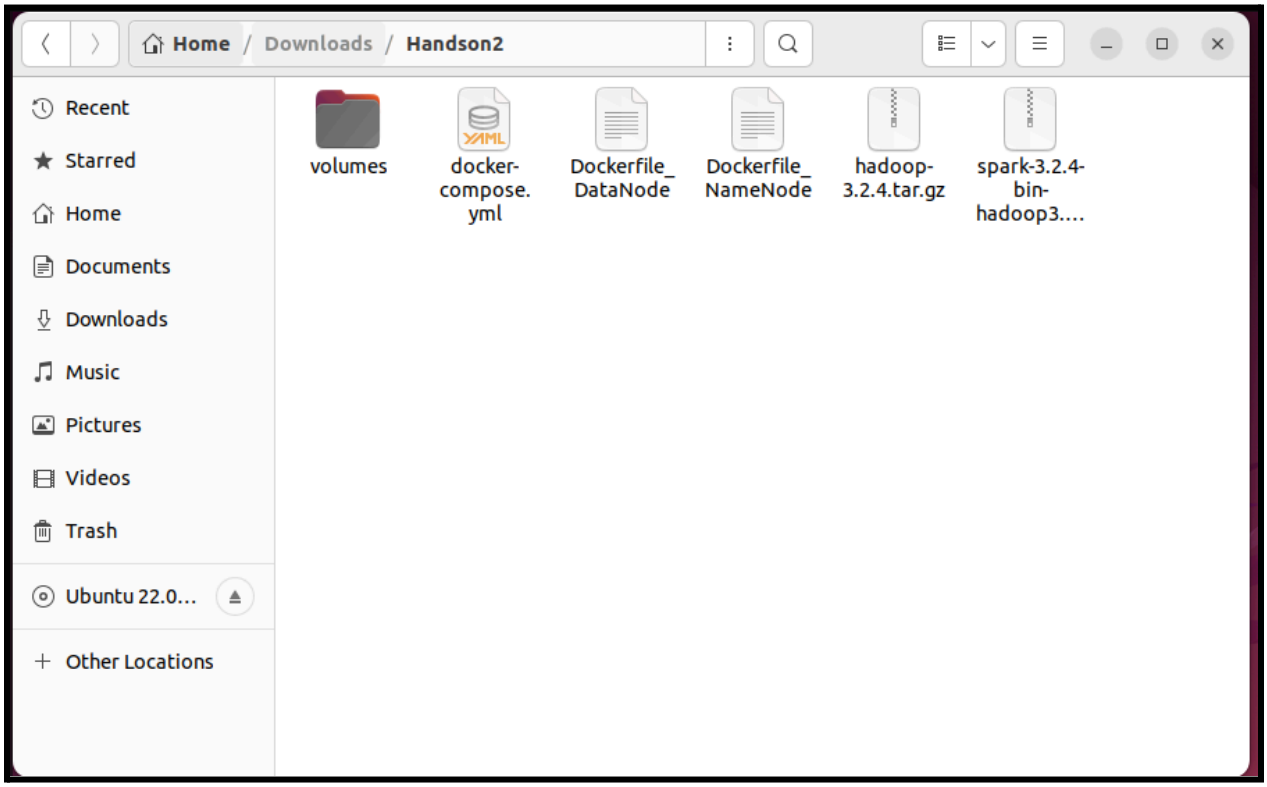
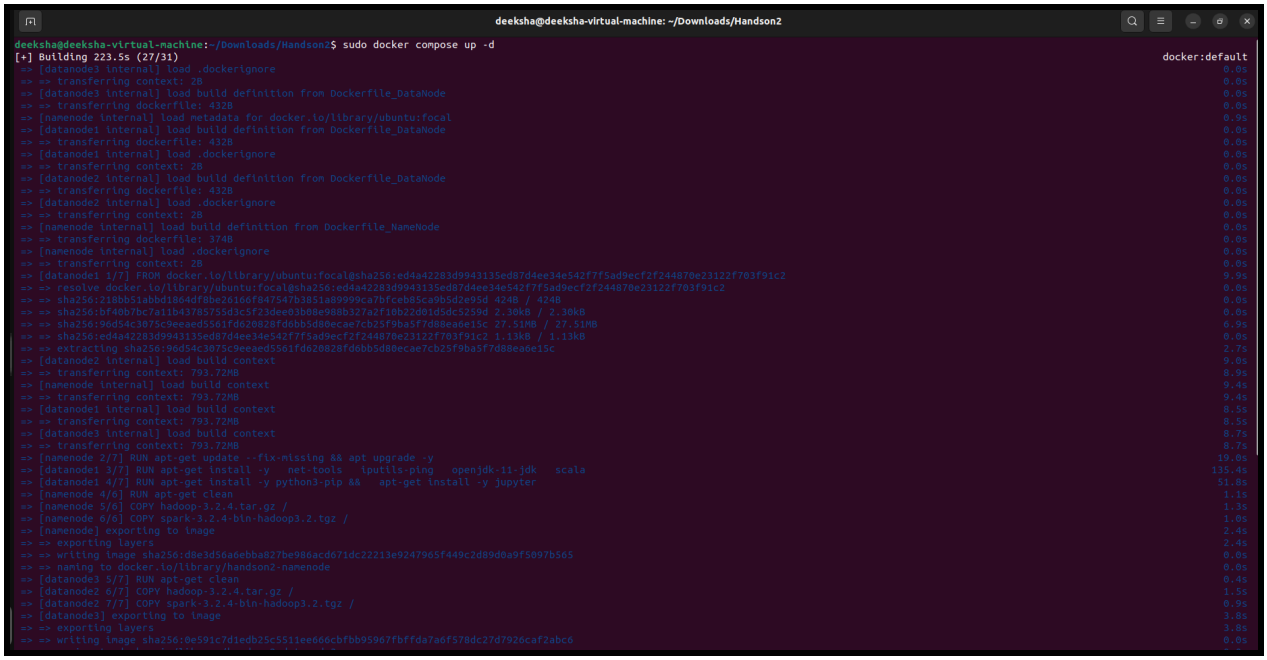
Electrical and Computer Engineering, Purdue University Northwest
Big Data (ECE49500/ECE59500)
Assignment 2

Task 1 [10 points] Hadoop and Spark set-up.

1. Download Handson2.zip file and extract the Handson2 folder from it.



2. Open the terminal to check-in to Handson2 folder and execute command: sudo docker compose up -d



3. Execute the command: sudo docker images

4. Open terminal tabs to access containers' shells. Run the 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

5. At namenode container, execute the following commands:

/volumes/setup.sh (Note: Run this command when you access a container first time)

su - hadoop hdfs namenode -format (Note: Run this command only when you access the *container first time)*

/volume/start.sh

```
deeksha@deeksha-virtual-machine:~/Downloads/Handson2$ sudo docker exec -it $(sudo docker ps --filter "name=namenode" --format "{{.ID}}") /bin/bash
root@e0992e40d6c1:/# /volumes/setup.sh
Created user account hadoop
Extracted hadoop files in /usr/local/hadoop/
Extracted spark files in /usr/local/spark/
Copied configuration files for hadoop
Set-up environmental variables for hadoop
Copied configuration files for hadoop
Set-up environmental variables for Spark
root@e0992e40d6c1:/# su - hadoop
hadoop@e0992e40d6c1:~$ hdfs namenode -format
WARNING: /usr/local/hadoop/logs does not exist. Creating.
2023-11-01 16:56:53,446 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = e0992e40d6c1/10.9.0.5
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/hadoop/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/hadoop/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/failureaccess-1.0.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-compress-1.21.jar:/usr/local/hadoop/share/hadoop/common/lib/kerb-core-1.0.1.jar:/usr/local/hadoop/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/hadoop/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/hadoop/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/share/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/hadoop/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/share/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/share/hadoop/common/lib/reload4j-1.2.18.3.jar:/usr/local/hadoop/share/hadoop/common/lib/asn-5.0.4.jar:/usr/local/hadoop/share/hadoop/common/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/share/hadoop/common/lib/animal-sniffer-annotations-1.17.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-io-2.8.0.jar:/usr/local/hadoop/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/local/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/hadoop/share/hadoop/common/lib/javax.activation-api-1.2.0.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-codec-1.11.jar:/usr/local/hadoop/share/hadoop/common/lib/guava-27.0-jre.jar:/usr/local/hadoop/share/hadoop/common/lib/jersey-core-1.19.jar:/usr/local/hadoop/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/common/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/common/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/common/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/lib/error_prone_annotations-2.2.0.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-xml-9.4.43.v20210629.jar:/usr/local/hadoop/share/hadoop/common/lib/jettison-1.1.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-collections-3.2.2.jar:/usr/local/hadoop/share/hadoop/common/lib/zookeeper-3.4.14.jar:/usr/local/hadoop/share/hadoop/common/lib/hadoop-annotations-3.2.4.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-http-9.4.43.v20210629.jar:/usr/local/hadoop/share/hadoop/common/lib/spotbugs-annotations-3.1.9.jar:/usr/local/hadoop/share/hadoop/common/lib/gson-2.9.0.jar:/usr/local/hadoop/share/hadoop/common/lib/listenablefuture-9999.0-empty-to-avoid-conflict-with-guava.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-annotations-2.10.5.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-util-ajax-9.4.43.v20210629.jar:/usr/local/hadoop/share/hadoop/common/lib/checker-qual-2.5.2.jar:/usr/local/hadoop/share/hadoop/common/lib/avro-1.7.7.jar:/usr/local/hadoop/share/hadoop/common/lib/netty-3.10.6.Final.jar:/usr/local/hadoop/share/hadoop/common/lib/paranamer-2.3.jar:/usr/local/hadoop/share/hadoop/common/lib/jsr305-3.0.2.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-server-9.4.43.v20210629.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-util-9.4.43.v20210629.jar:/usr/local/hadoop/share/hadoop/common/lib/curator-framework-2.13.0.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-webapp-9.4.43.v20210629.jar:/usr/local/hadoop/share/hadoop/common/lib/httpcore-4.4.13.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-beanutils-1.9.4.jar:/usr/local/hadoop/share/hadoop/common/lib/dnsjava-2.1.7.jar:/usr/local/hadoop/share/hadoop/common/lib/token-provider-1.0.1.jar:/usr/local/hadoop/share/hadoop/common/lib/audience-annotations-0.5.0.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-security-9.4.43.v20210629.jar:/usr/local/hadoop/share/hadoop/common/lib/kerb-simplekdc-1.0.1.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-io-9.4.43.v20210629.jar:/usr/local/hadoop/share/hadoop/common/lib/javax-annotations-1.1.jar:/usr/local/hadoop/share/hadoop/common/lib/slf4j-api-1.7.35.jar:/usr/local/hadoop/share/hadoop/common/lib/curator-client-2.13.0.jar:/usr/local/hadoop/share/hadoop/common/lib/kerby-xdr-1.0.1.jar:/usr/local/hadoop/share/hadoop/common/lib/jaxb-api-2.2.11.jar:/usr/local/hadoop/share/hadoop/common/lib/snappy-java-1.0.5.jar:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-3.2.4.jar:/usr/local/hadoop/share/hadoop/common/lib/kerb-util-1.0.1.jar:/usr/local/hadoop/share/hadoop/common/lib/woodstox-core-5.3.0.jar:/usr/local/hadoop/share/hadoop/common/lib/accessors-smart-2.4.7.jar:/usr/local/hadoop/share/hadoop/common/hadoop-common-3.2.4.jar:/usr/local/hadoop/share/hadoop/common/hadoop-nfs-3.2.4.jar:/usr/local/hadoop/share/hadoop/common/hadoop-kms-3.2.4.jar:/usr/local/hado
```

```
2023-11-01 16:56:53,942 INFO blockmanagement.BlockManager: maxReplicationStreams      = 2
2023-11-01 16:56:53,942 INFO blockmanagement.BlockManager: redundancyRecheckInterval  = 3000ms
2023-11-01 16:56:53,942 INFO blockmanagement.BlockManager: encryptDataTransfer        = false
2023-11-01 16:56:53,942 INFO blockmanagement.BlockManager: maxNumBlocksToLog          = 1000
2023-11-01 16:56:53,957 INFO namenode.FSDirectory: GLOBAL serial map: bits=29 maxEntries=536870911
2023-11-01 16:56:53,957 INFO namenode.FSDirectory: USER serial map: bits=24 maxEntries=16777215
2023-11-01 16:56:53,957 INFO namenode.FSDirectory: GROUP serial map: bits=24 maxEntries=16777215
2023-11-01 16:56:53,957 INFO namenode.FSDirectory: XATTR serial map: bits=24 maxEntries=16777215
2023-11-01 16:56:53,967 INFO util.GSet: Computing capacity for map INodeMap
2023-11-01 16:56:53,967 INFO util.GSet: VM type       = 64-bit
2023-11-01 16:56:53,968 INFO util.GSet: 1.0% max memory 1.9 GB = 19.8 MB
2023-11-01 16:56:53,968 INFO util.GSet: capacity      = 2^21 = 2097152 entries
2023-11-01 16:56:53,973 INFO namenode.FSDirectory: ACLs enabled? false
2023-11-01 16:56:53,974 INFO namenode.FSDirectory: POSIX ACL inheritance enabled? true
2023-11-01 16:56:53,974 INFO namenode.FSDirectory: XAttrs enabled? true
2023-11-01 16:56:53,975 INFO namenode.NameNode: Caching file names occurring more than 10 times
2023-11-01 16:56:53,978 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: false, skipCaptureAccessTimeOnlyChange: false, snapshotDiffAllowSnapRootDescendant: true, maxSnapshotLimit: 65536
2023-11-01 16:56:53,979 INFO snapshot.SnapshotManager: SkipList is disabled
2023-11-01 16:56:53,981 INFO util.GSet: Computing capacity for map cachedBlocks
2023-11-01 16:56:53,981 INFO util.GSet: VM type       = 64-bit
2023-11-01 16:56:53,982 INFO util.GSet: 0.25% max memory 1.9 GB = 4.9 MB
2023-11-01 16:56:53,982 INFO util.GSet: capacity      = 2^19 = 524288 entries
2023-11-01 16:56:53,989 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2023-11-01 16:56:53,989 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2023-11-01 16:56:53,989 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2023-11-01 16:56:53,992 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2023-11-01 16:56:53,992 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2023-11-01 16:56:53,994 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2023-11-01 16:56:53,994 INFO util.GSet: VM type       = 64-bit
2023-11-01 16:56:53,994 INFO util.GSet: 0.029999999329447746% max memory 1.9 GB = 607.0 KB
2023-11-01 16:56:53,995 INFO util.GSet: capacity      = 2^16 = 65536 entries
2023-11-01 16:56:54,010 INFO namenode.FSImage: Allocated new BlockPoolId: BP-150496094-10.9.0.5-1698857814005
2023-11-01 16:56:54,021 INFO common.Storage: Storage directory /usr/local/hadoop/data/nameNode has been successfully formatted.
2023-11-01 16:56:54,045 INFO namenode.FSImageFormatProtobuf: Saving image file /usr/local/hadoop/data/nameNode/current/fsimage.ckpt_0000000000000000000 using no compression
2023-11-01 16:56:54,104 INFO namenode.FSImageFormatProtobuf: Image file /usr/local/hadoop/data/nameNode/current/fsimage.ckpt_0000000000000000000 of size 401 bytes saved in 0 seconds .
2023-11-01 16:56:54,114 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2023-11-01 16:56:54,130 INFO namenode.FSNamesystem: Stopping services started for active state
2023-11-01 16:56:54,131 INFO namenode.FSNamesystem: Stopping services started for standby state
2023-11-01 16:56:54,134 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2023-11-01 16:56:54,135 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at e0992e40d6c1/10.9.0.5
************************************************************/
hadoop@e0992e40d6c1:~$ /volume/start.sh
-bash: /volume/start.sh: No such file or directory
hadoop@e0992e40d6c1:~$ /volumes/start.sh
240 ResourceManager
360 Jps
141 NameNode
hadoop@e0992e40d6c1:~$
```

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

/volume/start.sh

```
deeksha@deeksha-virtual-machine:~/Downloads/Handson2$ sudo docker exec -it $(sudo docker ps --filter "name=datanode1" --format "{{.ID}}") /bin/bash
root@9d26f7f1d36a:/# /volumes/setup.sh
Created user account hadoop
Extracted hadoop files in /usr/local/hadoop/
Extracted spark files in /usr/local/spark/
Copied configuration files for hadoop
Set-up environmental variables for hadoop
Copied configuration files for hadoop
Set-up environmental variables for Spark
root@9d26f7f1d36a:/# su - hadoop
hadoop@9d26f7f1d36a:~$ /volumes/start.sh
WARNING: /usr/local/hadoop/logs does not exist. Creating.
295 Jps
90 DataNode
190 NodeManager
hadoop@9d26f7f1d36a:~$
```

```
deeksha@deeksha-virtual-machine:~/Downloads/Handson2$ sudo docker exec -it $(sudo docker ps --filter "name=datanode2" --format "{{.ID}}") /bin/bash
root@80dd664ece0a:/# /volumes/setup.sh
Created user account hadoop
Extracted hadoop files in /usr/local/hadoop/
Extracted spark files in /usr/local/spark/
Copied configuration files for hadoop
Set-up environmental variables for hadoop
Copied configuration files for hadoop
Set-up environmental variables for Spark
root@80dd664ece0a:/# su - hadoop
hadoop@80dd664ece0a:~$ /volumes/start.sh
WARNING: /usr/local/hadoop/logs does not exist. Creating.
297 Jps
91 DataNode
191 NodeManager
hadoop@80dd664ece0a:~$
```

```
deeksha@deeksha-virtual-machine:~/Downloads/Handson2$ sudo docker exec -it $(sudo docker ps --filter "name=datanode3" --format "{{.ID}}") /bin/bash
root@403d6471c5d2:/# /volumes/setup.sh
Created user account hadoop
Extracted hadoop files in /usr/local/hadoop/
Extracted spark files in /usr/local/spark/
Copied configuration files for hadoop
Set-up environmental variables for hadoop
Copied configuration files for hadoop
Set-up environmental variables for Spark
root@403d6471c5d2:/# su - hadoop
hadoop@403d6471c5d2:~$ /volumes/start.sh
WARNING: /usr/local/hadoop/logs does not exist. Creating.
293 Jps
89 DataNode
189 NodeManager
hadoop@403d6471c5d2:~$
```

7. Go to any datanode's shell, (e.g. datanode3) and execute the following commands:

cd /volumes

pip3 install findspark

chmod +x start_jupyter.sh

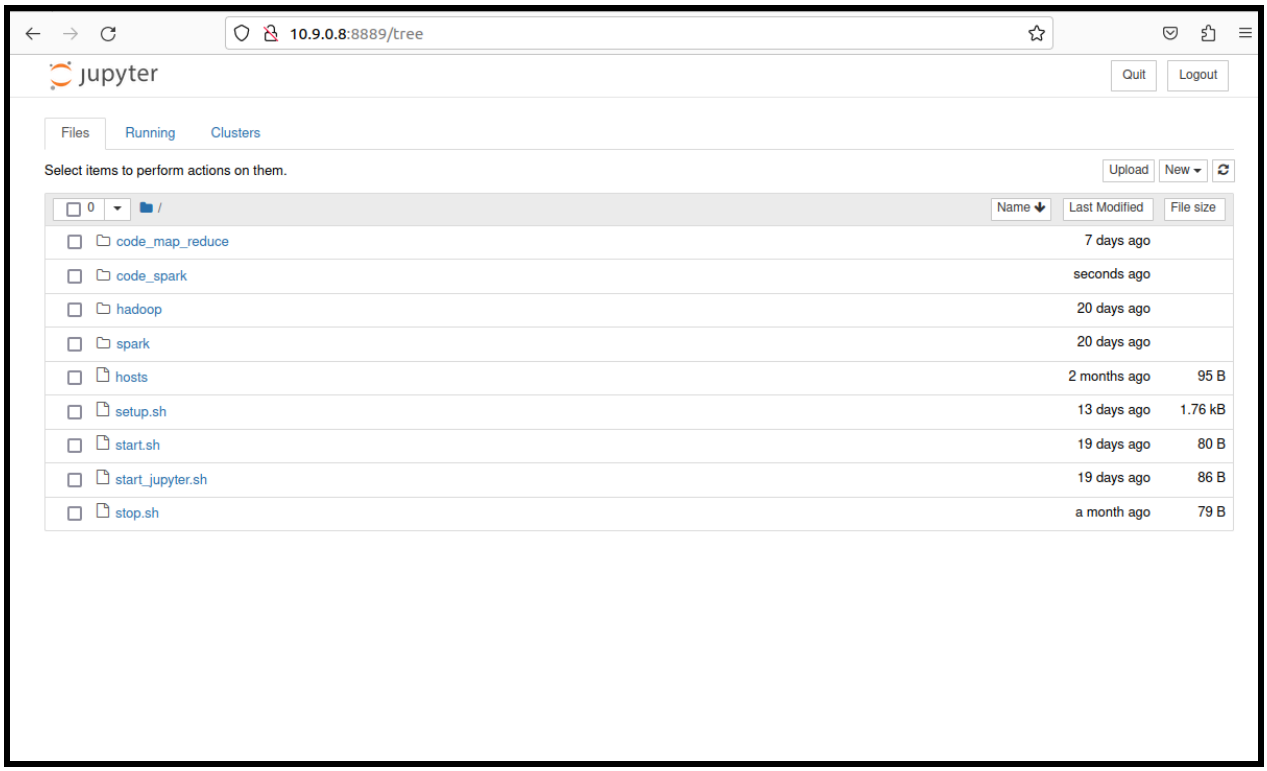
/volumes/start_jupyter.sh

```
deeksha@deeksha-virtual-machine:~/Downloads/Handson2$ sudo docker exec -it $(sudo docker ps --filter "name=datanode3" --format "{{.ID}}") /bin/bash
root@403d6471c5d2:/# /volumes/setup.sh
Created user account hadoop
Extracted hadoop files in /usr/local/hadoop/
Extracted spark files in /usr/local/spark/
Copied configuration files for hadoop
Set-up environmental variables for hadoop
Copied configuration files for hadoop
Set-up environmental variables for Spark
root@403d6471c5d2:/# su - hadoop
hadoop@403d6471c5d2:~$ /volumes/start.sh
WARNING: /usr/local/hadoop/logs does not exist. Creating.
293 Jps
89 DataNode
189 NodeManager
hadoop@403d6471c5d2:~$ cd /volumes
hadoop@403d6471c5d2:/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@403d6471c5d2:/volumes$ chmod +x start_jupyter.sh
hadoop@403d6471c5d2:/volumes$ /volumes/start_jupyter.sh
hadoop@403d6471c5d2:/volumes$ [I 17:06:05.001 NotebookApp] Writing notebook server cookie secret to /home/hadoop/.local/share/jupyter/runtime/notebook_cookie_secret
[I 17:06:05.209 NotebookApp] Serving notebooks from local directory: /volumes
[I 17:06:05.209 NotebookApp] The Jupyter Notebook is running at:
[I 17:06:05.209 NotebookApp] http://403d6471c5d2:8888/?token=0d46dcdc7c84675176ca1587903fdfcd6d2bc89443aa7937
[I 17:06:05.209 NotebookApp]  or http://127.0.0.1:8888/?token=0d46dcdc7c84675176ca1587903fdfcd6d2bc89443aa7937
[I 17:06:05.210 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 17:06:05.213 NotebookApp]

    To access the notebook, open this file in a browser:
        file:///home/hadoop/.local/share/jupyter/runtime/nbserver-319-open.html
    Or copy and paste one of these URLs:
        http://403d6471c5d2:8888/?token=0d46dcdc7c84675176ca1587903fdfcd6d2bc89443aa7937
     or http://127.0.0.1:8888/?token=0d46dcdc7c84675176ca1587903fdfcd6d2bc89443aa7937
[I 17:09:29.523 NotebookApp] 302 GET /?token=0d46dcdc7c84675176ca1587903fdfcd6d2bc89443aa7937 (10.9.0.1) 0.35ms
```

8. Copy the URL for Jupyter service 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 in Jupyter Notebook and complete the tasks in Hadoop_Task1.ipynb file and submit the Notebook.

IP of datanode - 10.9.0.8



Task 2 [10 points] Complete the tasks given in Hadoop_Task2.ipynb file and submit the Notebook file.

Hadoop_Task2.ipynb file submitted.

Task 3 [5 points] Go to code_spark directory in Jupyter Notebook and complete the tasks given in Spark_Task1.ipynb file and submit the Notebook file.

Spark_Task1.ipynb file submitted.

Task 4 [10 points] Go to code_spark directory in Jupyter Notebook and complete the tasks given in Spark_Task2.ipynb file and submit the Notebook file.

Spark_Task2.ipynb file submitted.