LAB SHEET 01

Sabaragamuwa University of Sri Lanka

Faculty of Computing

Department of Software Engineering

SE6103 – Parallel and Distributed Systems

Name	: K.O.K.S.Dayarathna
Reg. No	: 19APSE4299
Degree Program	: Software Engineering
Academic Period	: 3RD Year 2ND Semester

Step 1: Pull the Hadoop Docker Image

1. Choose a Hadoop Docker Image:

```
C:\Users\Kaumadi>docker pull bde2020/hadoop-namenode:latest
latest: Pulling from bde2020/hadoop-namenode
3192219afd04: Pull complete
7127a1d8cced: Pull complete
883a89599900: Pull complete
77920a3e82af: Pull complete
92329e81aec4: Pull complete
f373218fec59: Pull complete
aa53513fe997: Pull complete
8b1800105b98: Pull complete
c3a84a3e49c8: Pull complete
a65640a64a76: Pull complete
a29cc756d786: Pull complete
abf352b16046: Pull complete
dddd5a449e99: Pull complete
Digest: sha256:fdf74110805132d646cf6f12635efc0919e1fb2ac5bd376c5366272fc261301e
Status: Downloaded newer image for bde2020/hadoop-namenode:latest
docker.io/bde2020/hadoop-namenode:latest

What's Next?
  View a summary of image vulnerabilities and recommendations → docker scout quickview bde2020/hadoop-namenode:latest
```

2. Verify the Download:

```
C:\Users\Kaumadi>docker images
REPOSITORY                TAG       IMAGE ID       CREATED         SIZE
nginx                     latest    3b25b682ea82   6 weeks ago     192MB
hello-world               latest    d2c94e258dcb   18 months ago   13.3kB
bde2020/hadoop-namenode   latest    b638307a2119   4 years ago     1.37GB
```

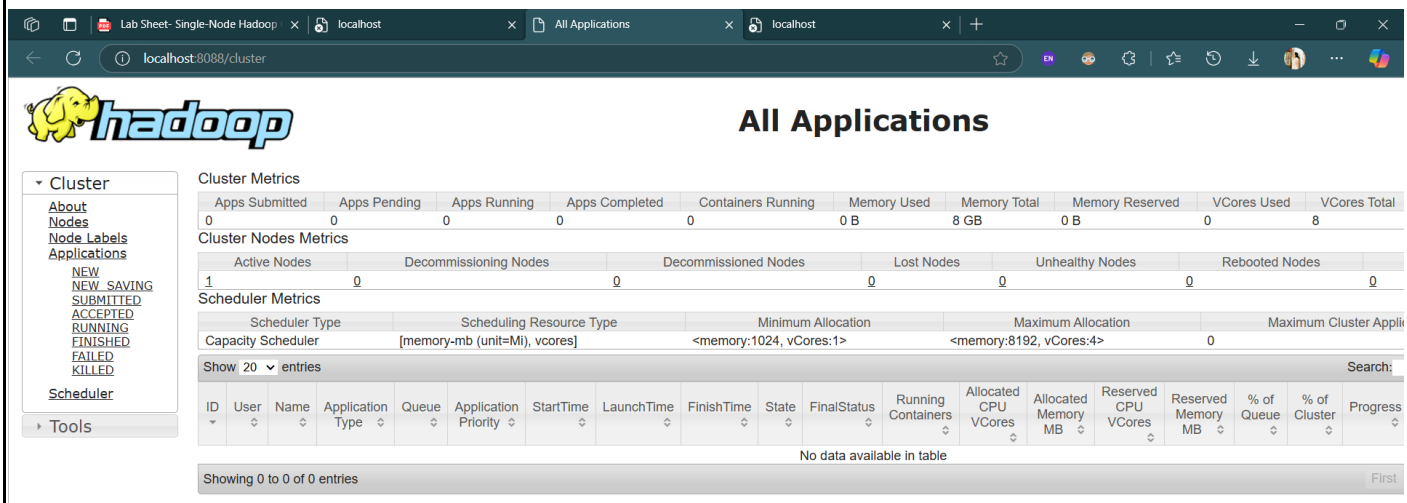
Step 2: Start the Hadoop Container

1. Run the Container:

```
C:\Users\Kaumadi>docker run -it --name hadoop-cluster -p 9870:9870 -p 8088 -p 50070:50070 bde2020/hadoop-namenode:latest /bin/bash
Configuring core
 - Setting fs.defaultFS=hdfs://7b7357074b48:8020
Configuring hdfs
 - Setting dfs.namenode.name.dir=file:///hadoop/dfs/name
Configuring yarn
Configuring httpfs
Configuring kms
Configuring mapred
Configuring for multihomed network
```

2. Start Hadoop Services:

```
root@7b7357074b48:/# /opt/hadoop-3.2.1/bin/hdfs --daemon start namenode
root@7b7357074b48:/# /opt/hadoop-3.2.1/bin/hdfs --daemon start datanode
root@7b7357074b48:/# /opt/hadoop-3.2.1/bin/yarn --daemon start resourcemanager
root@7b7357074b48:/# /opt/hadoop-3.2.1/bin/yarn --daemon start nodemanager
```



Step 4: Running a Sample MapReduce Job

1. Upload Sample Data to HDFS:

```
root@7b7357074b48:/# hdfs dfs -mkdir -p /user/hadoop/input
root@7b7357074b48:/# hdfs dfs -put $HADOOP_HOME/etc/hadoop/*.xml /user/hadoop/input
2024-11-18 09:38:23,221 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-11-18 09:38:23,860 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-11-18 09:38:23,921 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-11-18 09:38:24,410 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-11-18 09:38:24,859 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-11-18 09:38:25,301 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-11-18 09:38:25,332 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-11-18 09:38:25,387 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-11-18 09:38:25,428 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@7b7357074b48:/#
```

2. Run the WordCount Job:

```
root@7b7357074b48:/#  hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar wordcount /user/hadoop/input /user/hadoop/output
2024-11-18 09:50:02,739 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-11-18 09:50:02,869 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-11-18 09:50:02,869 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-11-18 09:50:03,220 INFO input.FileInputFormat: Total input files to process : 9
2024-11-18 09:50:03,265 INFO mapreduce.JobSubmitter: number of splits:9
2024-11-18 09:50:03,639 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local321594059_0001
2024-11-18 09:50:03,639 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-11-18 09:50:03,767 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-11-18 09:50:03,768 INFO mapreduce.Job: Running job: job_local321594059_0001
2024-11-18 09:50:03,771 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-11-18 09:50:03,782 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-11-18 09:50:03,782 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup fa
ilures: false
2024-11-18 09:50:03,783 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2024-11-18 09:50:03,830 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-11-18 09:50:03,830 INFO mapred.LocalJobRunner: Starting task: attempt_local321594059_0001_m_000000_0
2024-11-18 09:50:03,860 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-11-18 09:50:03,860 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup fa
ilures: false
2024-11-18 09:50:03,886 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2024-11-18 09:50:03,895 INFO mapred.MapTask: Processing split: hdfs://7b7357074b48:8020/user/hadoop/input/hadoop-policy.xml:0+11392
2024-11-18 09:50:05,265 INFO mapreduce.Job: Job job_local321594059_0001 running in uber mode : false
2024-11-18 09:50:05,360 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-11-18 09:50:05,360 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-11-18 09:50:05,360 INFO mapred.MapTask: soft limit at 83886080
2024-11-18 09:50:05,360 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-11-18 09:50:05,360 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
```

3. Check the Output:

```
root@7b7357074b48:/# hdfs dfs -cat /user/hadoop/output/part-r-00000
2024-11-18 09:53:15,796 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
"*"     21
"AS     9
"License");     9
"alice,bob      21
"clumping"      1
(ASF)   1
(root   1
(the    9
-->     18
-1      1
-1,     1
0.0     1
1-MAX_INT.      1
1.      1
1.0.    1
2.0     9
40      2
40+20=60        1
:       2
<!--    18
</configuration>        9
</description>  33
</name> 2
</property>     57
<?xml   8
<?xml-stylesheet        4
<configuration> 9
<description>   31
<description>ACL        25
<description>Default    1
<name>default.key.acl.DECRYPT_EEK</name>        1
<name>default.key.acl.GENERATE_EEK</name>       1
<name>default.key.acl.MANAGEMENT</name> 1
<name>default.key.acl.READ</name>       1
<name>hadoop.kms.acl.CREATE</name>      1
<name>hadoop.kms.acl.DECRYPT_EEK</name> 1
<name>hadoop.kms.acl.DELETE</name>      1
<name>hadoop.kms.acl.GENERATE_EEK</name>        1
<name>hadoop.kms.acl.GET</name> 1
```

Step 5: Exiting the Container

1. Stop the Container:

```
C:\Users\Kaumadi>docker stop hadoop-cluster
hadoop-cluster
```

2. Restart the Container:

```
C:\Users\Kaumadi>docker start -i hadoop-cluster
Configuring core
 - Setting fs.defaultFS=hdfs://7b7357074b48:8020
Configuring hdfs
 - Setting dfs.namenode.name.dir=file:///hadoop/dfs/name
Configuring yarn
Configuring httpfs
Configuring kms
Configuring mapred
Configuring for multihomed network
root@7b7357074b48:/#
```