

LAB SHEET 01

Sabaragamuwa University of Sri Lanka

Faculty of Computing

Software Engineering

SE6103 – Parallel and Distributed Systems

Name:	Disara Mapalagama
Reg. No:	19APSE4302
Academic Period:	3 rd Year 2 nd Semester
Degree Program:	BSc (Hons) in Software Engineering
Due Date:	18.11.2024

Lab Sheet: Single-Node Hadoop Cluster with Docker

1. Confirming Pre-requisites

Command:

```
docker --version
```

Output:

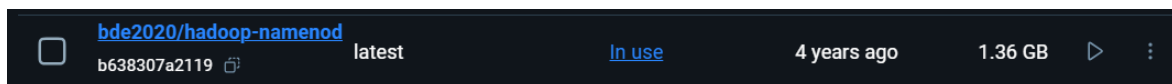
```
PS C:\Users\HP> docker --version  
Docker version 27.2.0, build 3ab4256
```

2. Step 1: Pull the Hadoop Docker image

Command:

```
docker pull bde2020/hadoop-namenode:latest
```

Output:



3. Verify the Download

Command:

```
docker images
```

Output:

```
PS C:\Users\HP> docker images  
REPOSITORY          TAG         IMAGE ID      CREATED        SIZE  
nginx                latest      3b25b682ea82  4 weeks ago   192MB  
alpine               latest      91ef0af61f39  8 weeks ago   7.8MB  
hello-world          latest      d2c94e258dcb  18 months ago 13.3kB  
bde2020/hadoop-namenode latest      b638307a2119  4 years ago   1.37GB
```

4. Step 2: Start the Hadoop Container

Command:

```
docker run -it --name hadoop-cluster-p 9870:9870 -p 8088:8088 -p 50070:50070 bde2020/hadoop-namenode:latest /bin/bash
```

Output:

```
Configuring core  
- Setting dfs.namenode.name.dir=file:///hadoop/dfs/name  
Configuring yarn  
Configuring httpfs  
Configuring kms  
Configuring mapred  
Configuring for multihomed network  
root@069201a88d8a:/#
```

5. Start Hadoop Services

Commands to run each service separately:

```
/opt/hadoop-3.2.1/bin/hdfs --daemon start namenode
```

```
/opt/hadoop-3.2.1/bin/hdfs --daemon start datanode
```

```
/opt/hadoop-3.2.1/bin/yarn --daemon start resourcemanager
```

```
/opt/hadoop-3.2.1/bin/yarn --daemon start nodemanager
```

6. Step 3: Access Hadoop Web Interfaces

- HDFS Web Interface (Resource Manager):

The screenshot shows the Hadoop All Applications web interface in a browser. The interface has a sidebar on the left with a 'Cluster' menu containing 'About', 'Nodes', 'Node Labels', 'Applications', and 'Scheduler'. The main content area is titled 'All Applications' and displays various metrics. At the top, there's a 'Cluster Metrics' table with columns: Apps Submitted, Apps Pending, Apps Running, Apps Completed, Containers Running, Memory Used, Memory Total, Memory Reserved, VCores Used, and VCores Total. Below this is a 'Cluster Nodes Metrics' table with columns: Active Nodes, Decommissioning Nodes, Decommissioned Nodes, Lost Nodes, Unhealthy Nodes, and Rebooted Nodes. Further down is a 'Scheduler Metrics' table with columns: Scheduler Type, Scheduling Resource Type, Minimum Allocation, Maximum Allocation, and Maximum Cluster App. At the bottom, there's a table with columns: ID, User, Name, Application Type, Queue, Application Priority, StartTime, LaunchTime, FinishTime, State, FinalStatus, Running Containers, Allocated CPU VCores, Allocated Memory MB, Reserved CPU VCores, Reserved Memory MB, % of Queue, % of Cluster, and Progress. The table currently shows 'No data available in table'.

- YARN Web Interface (NameNode Web UI):

The screenshot shows the YARN Web Interface (NameNode Web UI) Overview page. The page has a green header bar with a navigation menu: 'Hadoop', 'Overview', 'Datanodes', 'Datanode Volume Failures', 'Snapshot', 'Startup Progress', and 'Utilities'. The main content area is titled 'Overview '07e23cd8f647:8020' (active)'. Below the title is a table with the following information:

Started:	Mon Nov 18 14:16:20 +0530 2024
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 21:26:00 +0530 2019 by rohithsharmaks from branch-3.2.1
Cluster ID:	CID-46546643-960e-40ae-bd33-6027cac1fe18
Block Pool ID:	BP-398413681-172.17.0.2-1731919570021

Below the table is a 'Summary' section. It states: 'Security is off.', 'Safemode is off.', '1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).', 'Heap Memory used 39.83 MB of 236 MB Heap Memory. Max Heap Memory is 848 MB.', and 'Non Heap Memory used 44.72 MB of 46.15 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.' At the bottom, there's a table with the following information:

Configured Capacity:	1006.85 GB
Configured Remote Capacity:	0 B

7. Step 4: Running a Sample MapReduce Job

7.1. Upload Sample Data to HDFS

Command and Output:

```
root@07e23cd8f647:/# hdfs dfs -mkdir -p /user/hadoop/input  
root@07e23cd8f647:/#
```

Files uploaded successfully:

```
root@07e23cd8f647:/# hdfs dfs -ls /user/hadoop/input  
Found 9 items  
-rw-r--r--  3 root supergroup      8260 2024-11-18 08:53 /user/hadoop/input/capacity-scheduler.xml  
-rw-r--r--  3 root supergroup       860 2024-11-18 08:53 /user/hadoop/input/core-site.xml  
-rw-r--r--  3 root supergroup    11392 2024-11-18 08:53 /user/hadoop/input/hadoop-policy.xml  
-rw-r--r--  3 root supergroup     1385 2024-11-18 08:53 /user/hadoop/input/hdfs-site.xml  
-rw-r--r--  3 root supergroup       620 2024-11-18 08:53 /user/hadoop/input/https-site.xml  
-rw-r--r--  3 root supergroup     3518 2024-11-18 08:53 /user/hadoop/input/kms-acls.xml  
-rw-r--r--  3 root supergroup       682 2024-11-18 08:53 /user/hadoop/input/kms-site.xml  
-rw-r--r--  3 root supergroup       841 2024-11-18 08:53 /user/hadoop/input/mapred-site.xml  
-rw-r--r--  3 root supergroup     1031 2024-11-18 08:53 /user/hadoop/input/yarn-site.xml
```

7.2. Run the WordCount Job

```
root@07e23cd8f647:/# hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar word  
count /user/hadoop/input /user/hadoop/output  
2024-11-18 08:57:53,407 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties  
2024-11-18 08:57:53,515 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).  
2024-11-18 08:57:53,515 INFO impl.MetricsSystemImpl: JobTracker metrics system started  
2024-11-18 08:57:54,027 INFO input.FileInputFormat: Total input files to process : 9  
2024-11-18 08:57:54,096 INFO mapreduce.JobSubmitter: number of splits:9  
2024-11-18 08:57:54,278 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local382613832_0001  
2024-11-18 08:57:54,278 INFO mapreduce.JobSubmitter: Executing with tokens: []  
2024-11-18 08:57:54,442 INFO mapreduce.Job: The url to track the job: http://localhost:8080/  
2024-11-18 08:57:54,444 INFO mapreduce.Job: Running job: job_local382613832_0001  
2024-11-18 08:57:54,445 INFO mapred.LocalJobRunner: OutputCommitter set in config null  
2024-11-18 08:57:54,463 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2  
2024-11-18 08:57:54,464 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary fol
```

7.3. Check the output:

```
root@07e23cd8f647:/# hdfs dfs -cat /user/hadoop/output/part-r-00000  
2024-11-18 08:58:29,191 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
```

8. Step 5: Exiting the Container

8.1. Stop the container:

```
root@07e23cd8f647:/# exit
exit
PS C:\Users\HP> docker stop hadoop-cluster
hadoop-cluster
```

8.2. Restart the container:

```
PS C:\Users\HP> docker start -i hadoop-cluster
Configuring core
- Setting fs.defaultFS=hdfs://07e23cd8f647:8020
Configuring hdfs
- Setting dfs.namenode.name.dir=file:///hadoop/dfs/name
Configuring yarn
Configuring httpfs
Configuring kms
Configuring mapred
Configuring for multihomed network
root@07e23cd8f647:/#
```

9. Task 1: Customize HDFS Configurations

9.1. Edit core-site.xml and hdfs-site.xml

Inside the container, navigate to the Hadoop configuration directory:

Command:

```
cd $HADOOP_HOME/etc/hadoop
```

Output:

```
root@07e23cd8f647:/# cd $HADOOP_HOME/etc/hadoop
root@07e23cd8f647:/opt/hadoop-3.2.1/etc/hadoop#
```

Open the configuration files for editing using nano core-site.xml. Add or modify configurations such as:

```
root@07e23cd8f647:/opt/hadoop-3.2.1/etc/hadoop# sed -i '/<\configuration>/i \
> <property>\n\
>   <name>fs.defaultFS</name>\n\
>   <value>hdfs://localhost:9000</value>\n\
> </property>' core-site.xml
```

Customizing HDFS configuration

```
root@07e23cd8f647:/opt/hadoop-3.2.1/etc/hadoop# sed -i '/<\/configuration>/i \
> <property>\n\
>   <name>dfs.replication</name>\n\
>   <value>1</value>\n\
> </property>\n\
> <property>\n\
>   <name>dfs.blocksize</name>\n\
>   <value>134217728</value>\n\
> </property>' hdfs-site.xml
```

Restarting Hadoop:

```
root@07e23cd8f647:/opt/hadoop-3.2.1/etc/hadoop# $HADOOP_HOME/sbin/hadoop-daemon.sh stop namenode
WARNING: Use of this script to stop HDFS daemons is deprecated.
WARNING: Attempting to execute replacement "hdfs --daemon stop" instead.
root@07e23cd8f647:/opt/hadoop-3.2.1/etc/hadoop# $HADOOP_HOME/sbin/hadoop-daemon.sh start namenode
WARNING: Use of this script to start HDFS daemons is deprecated.
WARNING: Attempting to execute replacement "hdfs --daemon start" instead.
root@07e23cd8f647:/opt/hadoop-3.2.1/etc/hadoop# $HADOOP_HOME/sbin/hadoop-daemon.sh stop datanode
WARNING: Use of this script to stop HDFS daemons is deprecated.
WARNING: Attempting to execute replacement "hdfs --daemon stop" instead.
root@07e23cd8f647:/opt/hadoop-3.2.1/etc/hadoop# $HADOOP_HOME/sbin/hadoop-daemon.sh start datanode
WARNING: Use of this script to start HDFS daemons is deprecated.
WARNING: Attempting to execute replacement "hdfs --daemon start" instead.
root@07e23cd8f647:/opt/hadoop-3.2.1/etc/hadoop# $HADOOP_HOME/sbin/hadoop-daemon.sh stop secondarynamenode
WARNING: Use of this script to stop HDFS daemons is deprecated.
WARNING: Attempting to execute replacement "hdfs --daemon stop" instead.
```