# Hadoop Exercise

**Sabaragamuwa University of Sri Lanka**

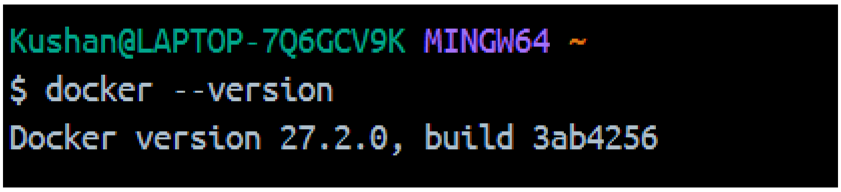
**Faculty of Computing**

**Department of Software Engineering**

**SE6103 - Parallel and Distributed Systems**

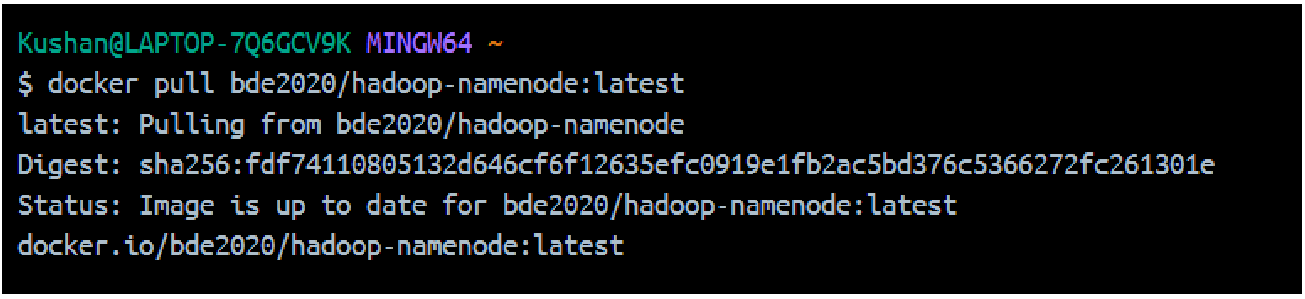
| | |
|---|---|
| Name | : K.M.Andarawewa |
| Reg. No | : 19APSE4269 |
| Academic Period | : 3rd Year 2nd Semester |
| Due Date | : 18/11/2024 |

## 1) Check the Docker Version

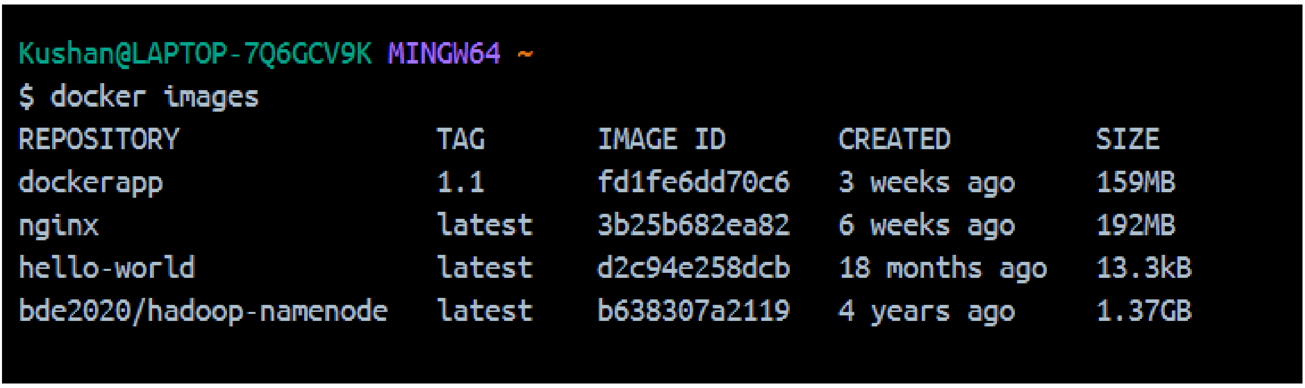


```
Kushan@LAPTOP-7Q6GCV9K MINGW64 ~
$ docker --version
Docker version 27.2.0, build 3ab4256
```

## 2) Pull the Hadoop image

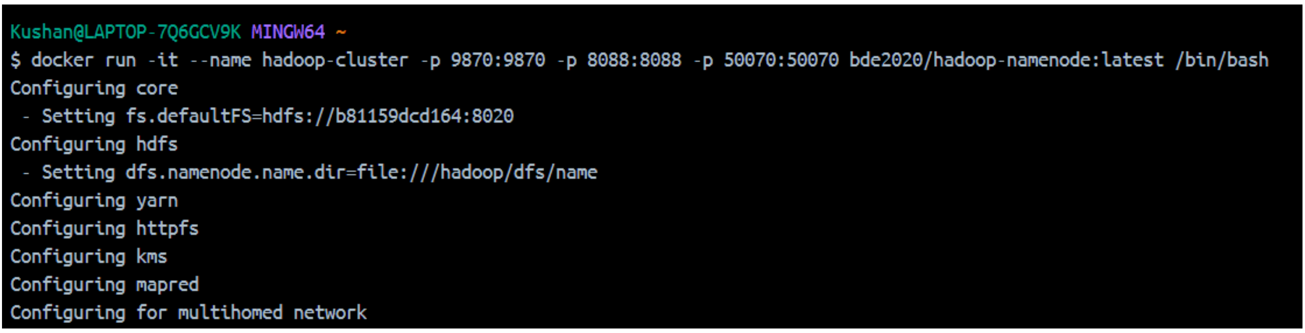


```
Kushan@LAPTOP-7Q6GCV9K MINGW64 ~
$ docker pull bde2020/hadoop-namenode:latest
latest: Pulling from bde2020/hadoop-namenode
Digest: sha256:fdf74110805132d646cf6f12635efc0919e1fb2ac5bd376c5366272fc261301e
Status: Image is up to date for bde2020/hadoop-namenode:latest
docker.io/bde2020/hadoop-namenode:latest
```



```
Kushan@LAPTOP-7Q6GCV9K MINGW64 ~
$ docker images
REPOSITORY                TAG       IMAGE ID       CREATED         SIZE
dockerapp                 1.1       fd1fe6dd70c6   3 weeks ago     159MB
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
```

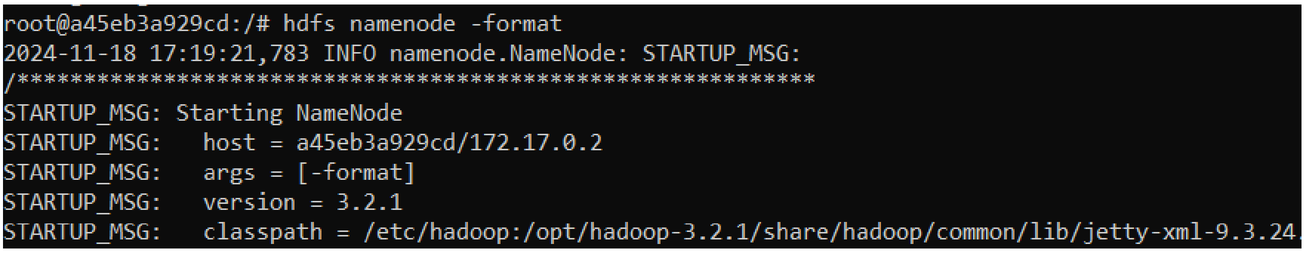
## 3) Run the Hadoop image



```
Kushan@LAPTOP-7Q6GCV9K MINGW64 ~
$ docker run -it --name hadoop-cluster -p 9870:9870 -p 8088:8088 -p 50070:50070 bde2020/hadoop-namenode:latest /bin/bash
Configuring core
 - Setting fs.defaultFS=hdfs://b81159dcd164:8020
Configuring hdfs
 - Setting dfs.namenode.name.dir=file:///hadoop/dfs/name
Configuring yarn
Configuring httpfs
Configuring kms
Configuring mapred
Configuring for multihomed network
```

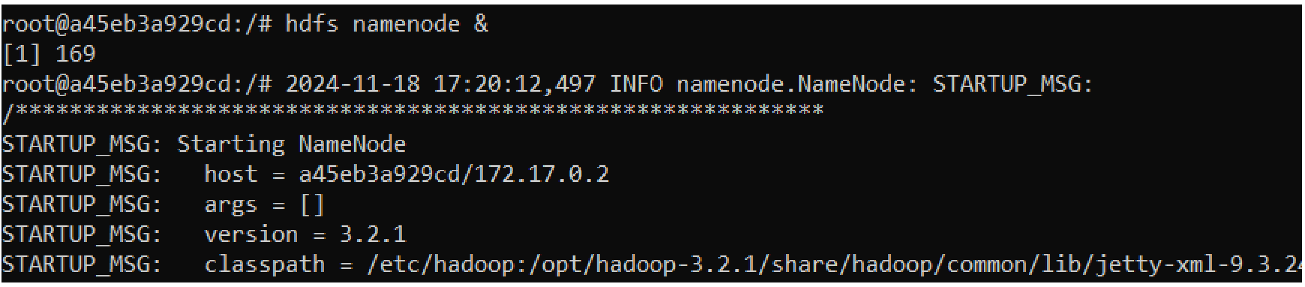
## 4) Configure the Hadoop file system

I. hdfs namenode -format



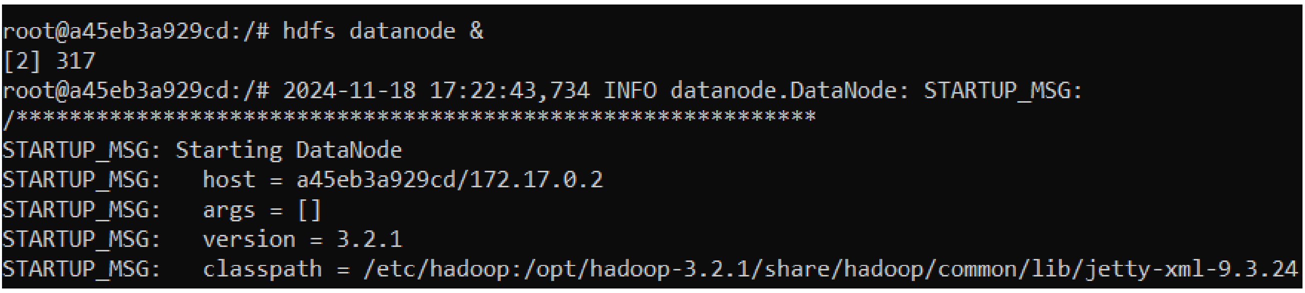
```
root@a45eb3a929cd:/# hdfs namenode -format
2024-11-18 17:19:21,783 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = a45eb3a929cd/172.17.0.2
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.2.1
STARTUP_MSG:   classpath = /etc/hadoop:/opt/hadoop-3.2.1/share/hadoop/common/lib/jetty-xml-9.3.24.
```

II. hdfs namenode &



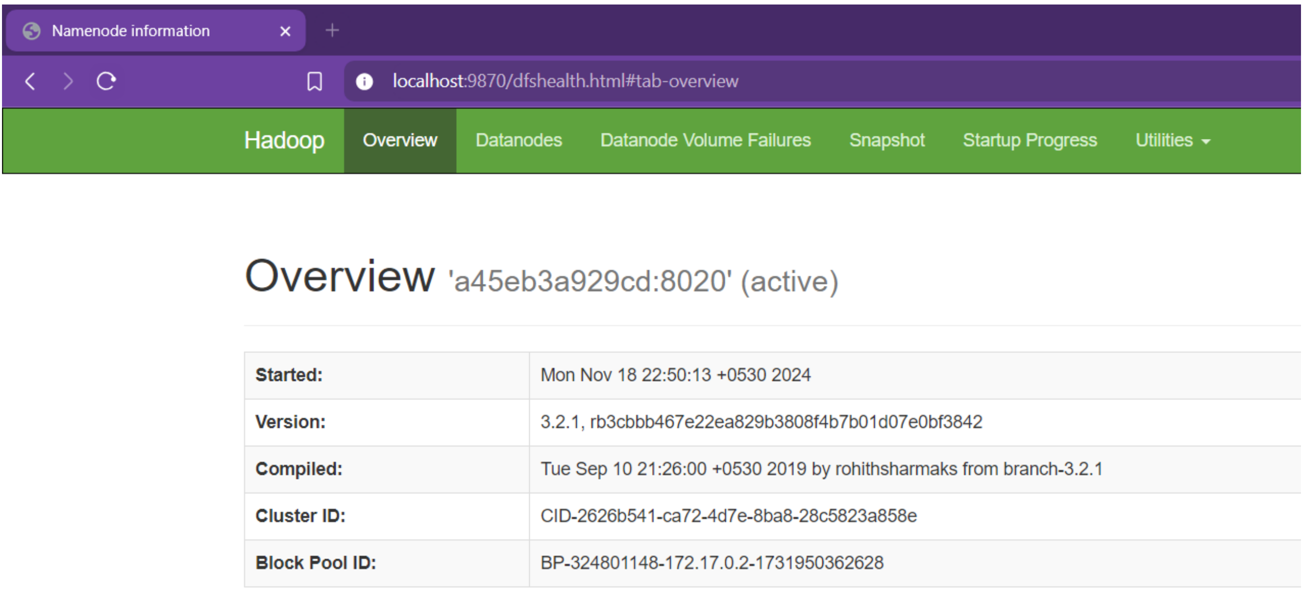
```
root@a45eb3a929cd:/# hdfs namenode &
[1] 169
root@a45eb3a929cd:/# 2024-11-18 17:20:12,497 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = a45eb3a929cd/172.17.0.2
STARTUP_MSG:   args = []
STARTUP_MSG:   version = 3.2.1
STARTUP_MSG:   classpath = /etc/hadoop:/opt/hadoop-3.2.1/share/hadoop/common/lib/jetty-xml-9.3.2
```

III. hdfs datanode &



```
root@a45eb3a929cd:/# hdfs datanode &
[2] 317
root@a45eb3a929cd:/# 2024-11-18 17:22:43,734 INFO datanode.DataNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting DataNode
STARTUP_MSG:   host = a45eb3a929cd/172.17.0.2
STARTUP_MSG:   args = []
STARTUP_MSG:   version = 3.2.1
STARTUP_MSG:   classpath = /etc/hadoop:/opt/hadoop-3.2.1/share/hadoop/common/lib/jetty-xml-9.3.24
```

## 5) Start the local host

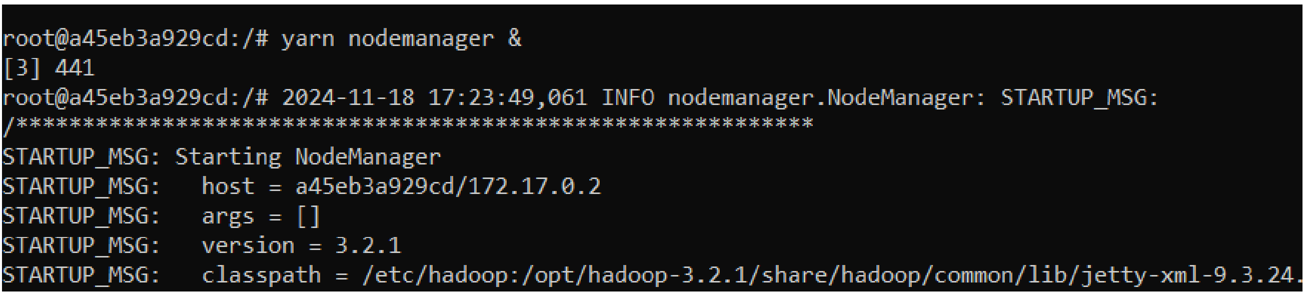


Overview 'a45eb3a929cd:8020' (active)

| | |
|---|---|
| **Started:** | Mon Nov 18 22:50:13 +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-2626b541-ca72-4d7e-8ba8-28c5823a858e |
| **Block Pool ID:** | BP-324801148-172.17.0.2-1731950362628 |

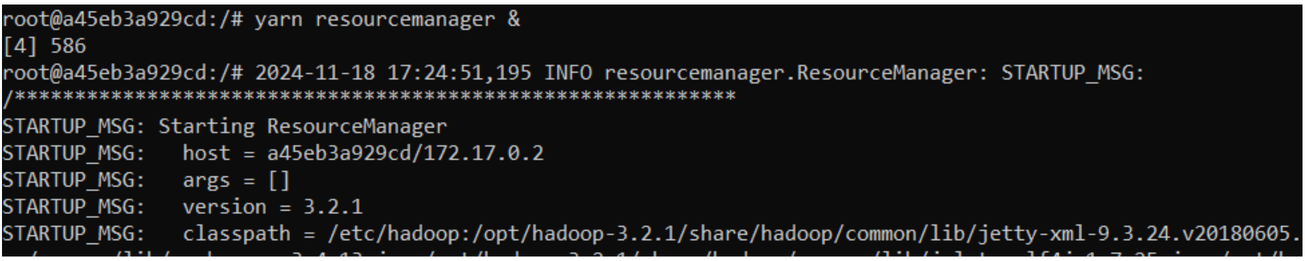
## 6) Start the node manager and resource manager

I. yarn nodemanager &



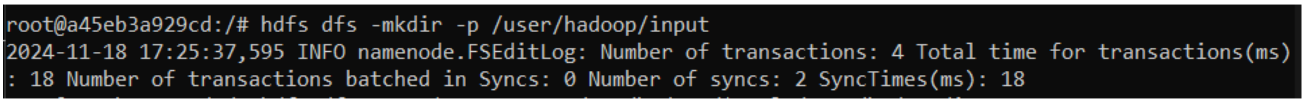
```
root@a45eb3a929cd:/# yarn nodemanager &
[3] 441
root@a45eb3a929cd:/# 2024-11-18 17:23:49,061 INFO nodemanager.NodeManager: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NodeManager
STARTUP_MSG:   host = a45eb3a929cd/172.17.0.2
STARTUP_MSG:   args = []
STARTUP_MSG:   version = 3.2.1
STARTUP_MSG:   classpath = /etc/hadoop:/opt/hadoop-3.2.1/share/hadoop/common/lib/jetty-xml-9.3.24.
```

II. yarn resourcemanager &

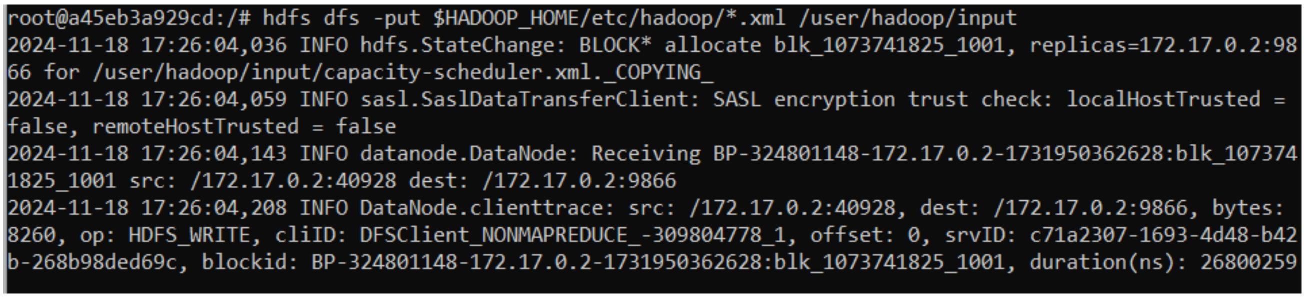


```
root@a45eb3a929cd:/# yarn resourcemanager &
[4] 586
root@a45eb3a929cd:/# 2024-11-18 17:24:51,195 INFO resourcemanager.ResourceManager: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting ResourceManager
STARTUP_MSG:   host = a45eb3a929cd/172.17.0.2
STARTUP_MSG:   args = []
STARTUP_MSG:   version = 3.2.1
STARTUP_MSG:   classpath = /etc/hadoop:/opt/hadoop-3.2.1/share/hadoop/common/lib/jetty-xml-9.3.24.v20180605.
```

## 7) Add sample data to HDFS.

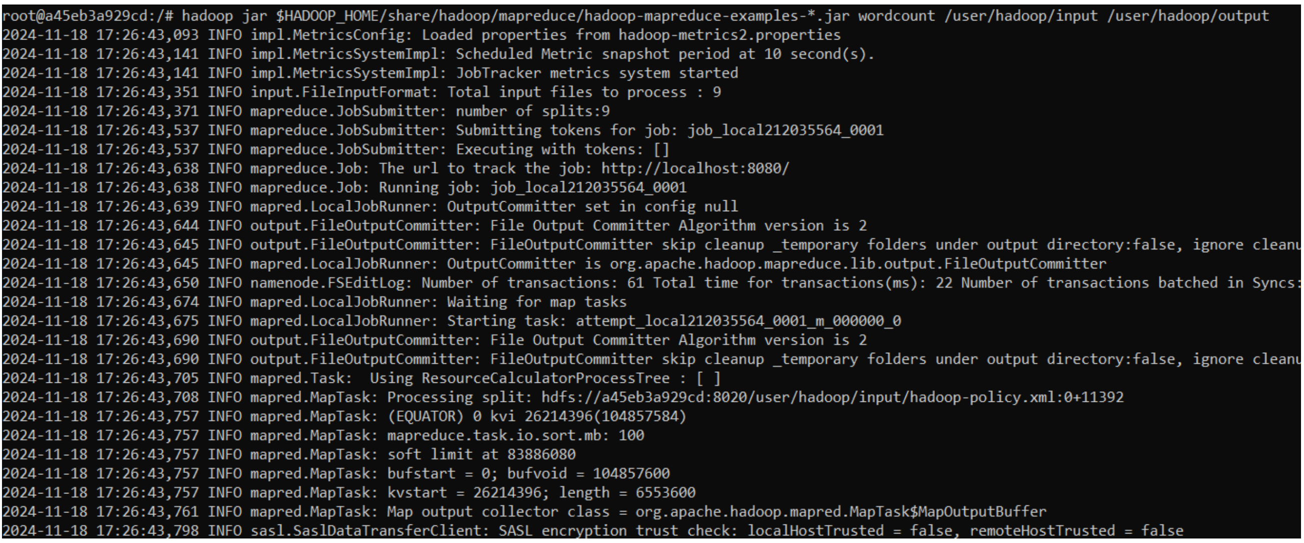


```
root@a45eb3a929cd:/# hdfs dfs -mkdir -p /user/hadoop/input
2024-11-18 17:25:37,595 INFO namenode.FSEditLog: Number of transactions: 4 Total time for transactions(ms)
: 18 Number of transactions batched in Syncs: 0 Number of syncs: 2 SyncTimes(ms): 18
```



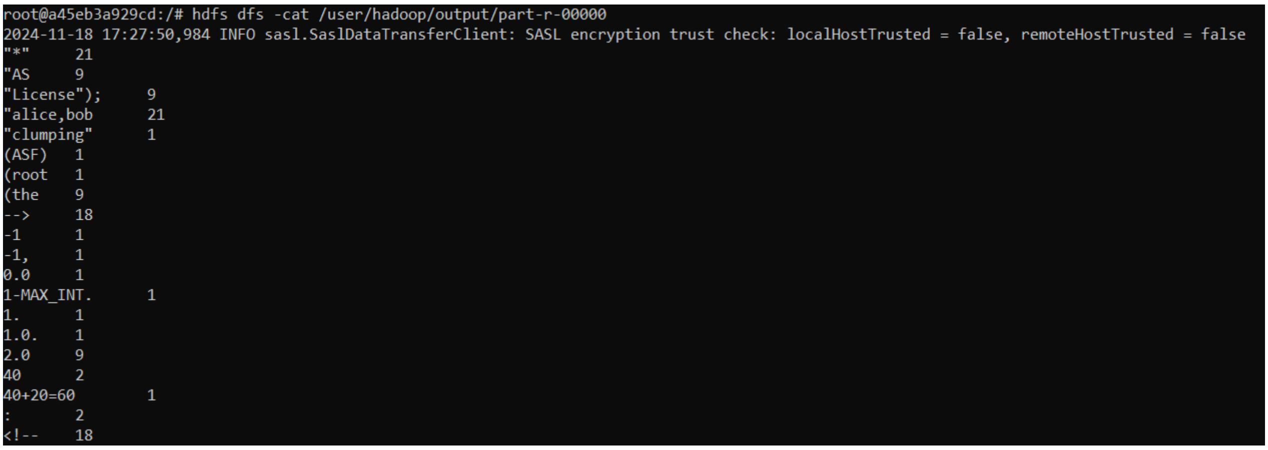
```
root@a45eb3a929cd:/# hdfs dfs -put $HADOOP_HOME/etc/hadoop/*.xml /user/hadoop/input
2024-11-18 17:26:04,036 INFO hdfs.StateChange: BLOCK* allocate blk_1073741825_1001, replicas=172.17.0.2:98
66 for /user/hadoop/input/capacity-scheduler.xml._COPYING_
2024-11-18 17:26:04,059 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted =
false, remoteHostTrusted = false
2024-11-18 17:26:04,143 INFO datanode.DataNode: Receiving BP-324801148-172.17.0.2-1731950362628:blk_107374
1825_1001 src: /172.17.0.2:40928 dest: /172.17.0.2:9866
2024-11-18 17:26:04,208 INFO DataNode.clienttrace: src: /172.17.0.2:40928, dest: /172.17.0.2:9866, bytes:
8260, op: HDFS_WRITE, cliID: DFSClient_NONMAPREDUCE_-309804778_1, offset: 0, srvID: c71a2307-1693-4d48-b42
b-268b98ded69c, blockid: BP-324801148-172.17.0.2-1731950362628:blk_1073741825_1001, duration(ns): 26800259
```

## 8) Execute the word count job.



```
root@a45eb3a929cd:/# hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-*.jar wordcount /user/hadoop/input /user/hadoop/output
2024-11-18 17:26:43,093 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-11-18 17:26:43,141 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-11-18 17:26:43,141 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-11-18 17:26:43,351 INFO input.FileInputFormat: Total input files to process : 9
2024-11-18 17:26:43,371 INFO mapreduce.JobSubmitter: number of splits:9
2024-11-18 17:26:43,537 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local212035564_0001
2024-11-18 17:26:43,537 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-11-18 17:26:43,638 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-11-18 17:26:43,638 INFO mapreduce.Job: Running job: job_local212035564_0001
2024-11-18 17:26:43,639 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-11-18 17:26:43,644 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-11-18 17:26:43,645 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanu
2024-11-18 17:26:43,645 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2024-11-18 17:26:43,650 INFO namenode.FSEditLog: Number of transactions: 61 Total time for transactions(ms): 22 Number of transactions batched in Syncs:
2024-11-18 17:26:43,674 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-11-18 17:26:43,675 INFO mapred.LocalJobRunner: Starting task: attempt_local212035564_0001_m_000000_0
2024-11-18 17:26:43,690 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-11-18 17:26:43,690 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanu
2024-11-18 17:26:43,705 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2024-11-18 17:26:43,708 INFO mapred.MapTask: Processing split: hdfs://a45eb3a929cd:8020/user/hadoop/input/hadoop-policy.xml:0+11392
2024-11-18 17:26:43,757 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-11-18 17:26:43,757 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-11-18 17:26:43,757 INFO mapred.MapTask: soft limit at 83886080
2024-11-18 17:26:43,757 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-11-18 17:26:43,757 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2024-11-18 17:26:43,761 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2024-11-18 17:26:43,798 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
```

## 9) Verify the results in the output.



```
root@a45eb3a929cd:/# hdfs dfs -cat /user/hadoop/output/part-r-00000
2024-11-18 17:27:50,984 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
```