# Practical#3

**Name:** Saurin Anilkumar Prajapati

**Roll No.:** 19BCE239

**Course Code and Name:** 2CS702 Big Data Analytics

**Batch:** D1

---

**AIM:**

Setup single node Hadoop cluster and apply HDFS commands on single node Hadoop Cluster.

---

## Hadoop Installation guide:

[NOTE: I am using MacBook M1, so the steps were a little more complicated than windows. I went through 2-3 installation guides on the Internet and saw some YouTube guide videos.]

I am attaching those guides:

Installing Hadoop on a Mac

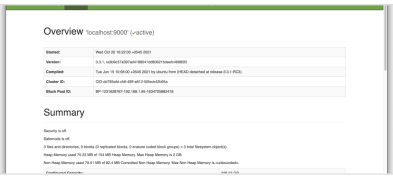
Is the only thing standing between you and Hadoop just trying to figure out how to install it on a Mac? A quick internet search will show you the lack of information about this fairly simple process.

https://towardsdatascience.com/installing-hadoop-on-a-mac-ec01c67b003c

Install Hadoop on MacOS

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. In this article I have shown how you can install hadoop on MacOS specific on Macbook M1 fixing the possible errors. So let's get started.

https://codewitharjun.medium.com/install-hadoop-on-macos-efe7c860c3ed

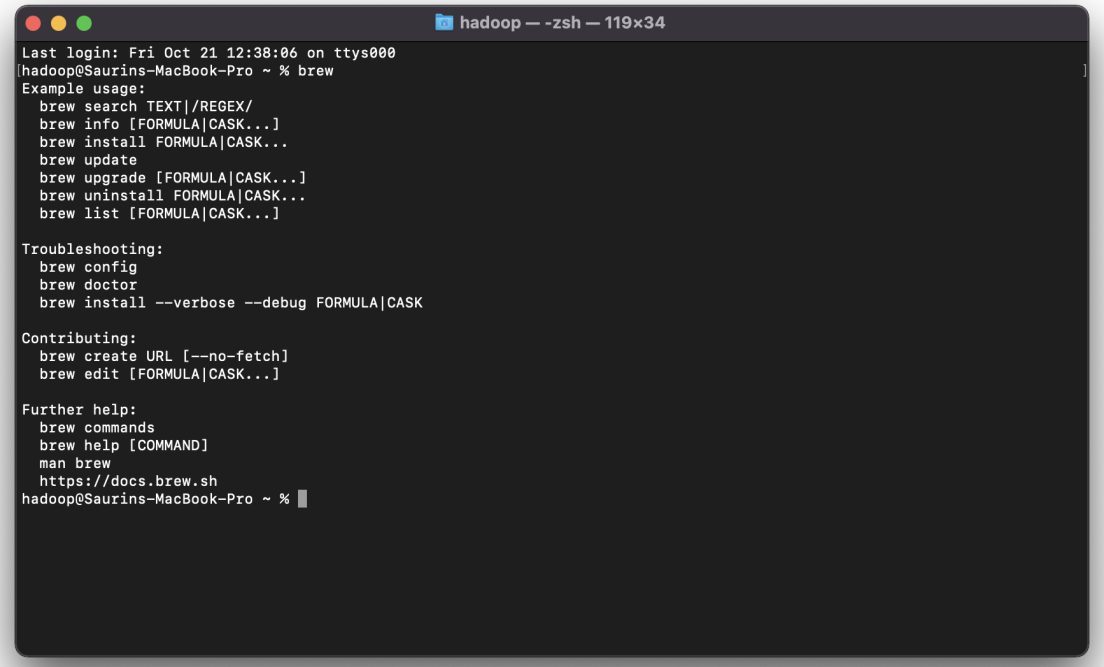



S**teps are as follows**:

[NOTE:  I have **NOT** attached screenshots of installation coz some middle steps were incorrect and would cause confusion. But at the end Hadoop was **successfully installed** and all nodes were working as expected.]

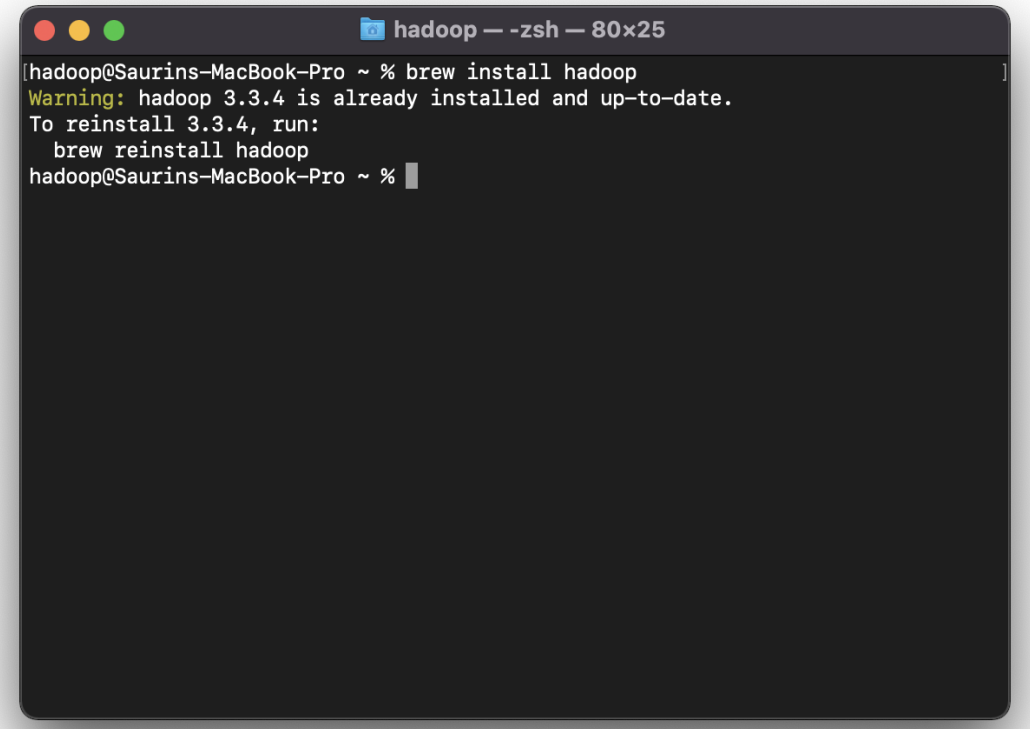
**Step 1:** Install home-brew on Mac if you do not already have it. (I already did)

Proof: (if home-brew was NOT installed ‘brew’ command would not have been recognised




**Step 2:** Install Hadoop using Home-brew

Proof: (hadoop already installed at time of screenshot)




**Step 3:**

Configuring Hadoop will take place over a few steps. A more detailed version can be found in the **Apache Hadoop documentation** for setting up a single node cluster. (Be sure to follow along with the correct version installed on your machine.)

1. Updating the **environment variable settings**
2. Make changes to **core**, **hdfs**, **mapred**and **yarn**site.xml files
3. Remove **password** requirement (if necessary)
4. Format **NameNode**

Open the document containing the environment variable settings :

```
$ cd /usr/local/cellar/hadoop/3.2.1/libexec/etc/hadoop
$ open hadoop-env.sh
```

Make the following changes to the document, save and close.

### Add the location for export JAVA_HOME

export JAVA_HOME= “/Library/Java/JavaVirtualMachines/adoptopenjdk-8.jdk/Contents/Home”

You can find this path by using the following code in the terminal window:

```
$ /usr/libexec/java_home
```

### Replace information for export HADOOP_OPTS

change export HADOOP_OPTS=”-Djava.net.preferIPv4Stack=true”

to export HADOOP_OPTS = ”-Djava.net.preferIPv4Stack=true -Djava.security.krb5.realm= -Djava.security.krb5.kdc=”

### Make changes to core files

```
$ open core-site.xml<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

### Make changes to hdfs files

```
$ open hdfs-site.xml<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

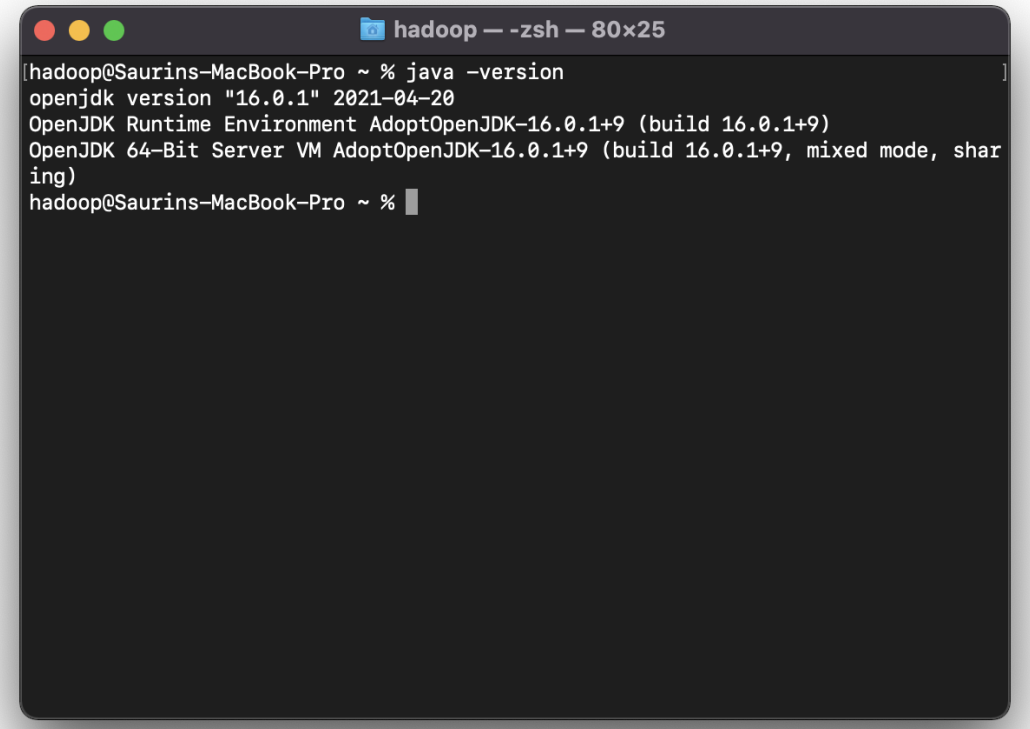
### Make changes to mapred files

```
$ open mapred-site.xml<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
    <name>mapreduce.application.classpath</name>   <value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/share/
hadoop/mapreduce/lib/*</value>
  </property>
</configuration>
```

### Make changes to yarn files

```
$ open yarn-site.xml
```

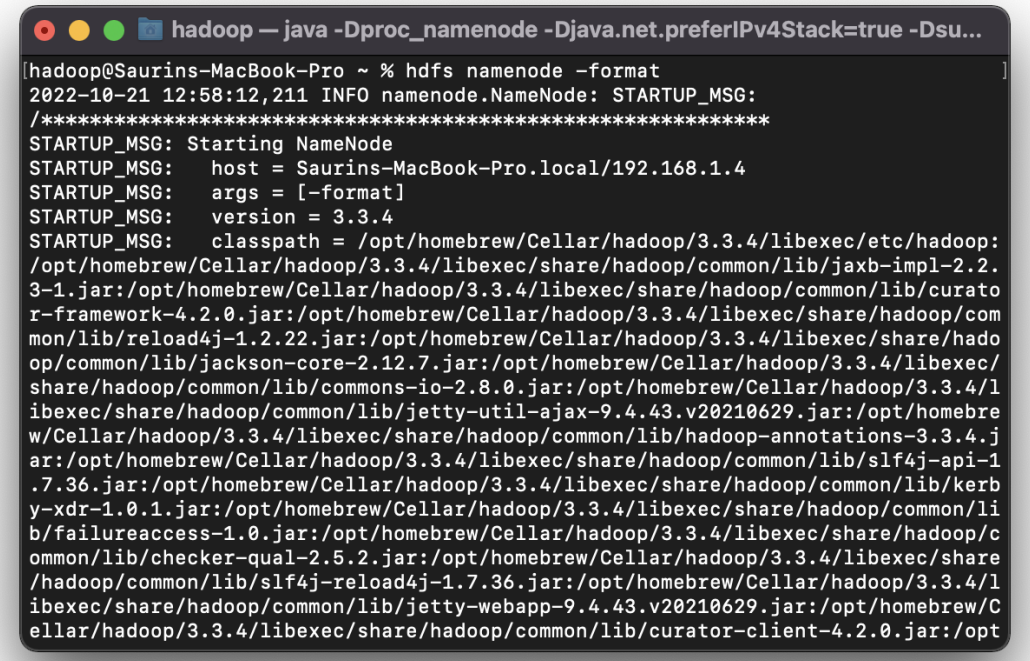
<configuration><property><name>yarn.nodemanager.aux-services</name><value>mapreduce_shuffle</value></property>
<property><name>yarn.nodemanager.env-whitelist</name>
<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PREPENI
</property></configuration>




---

So now we have successfully set-up everything we needed to do.

Next we just **format name node**, and **get started with hadoop** as we wish to

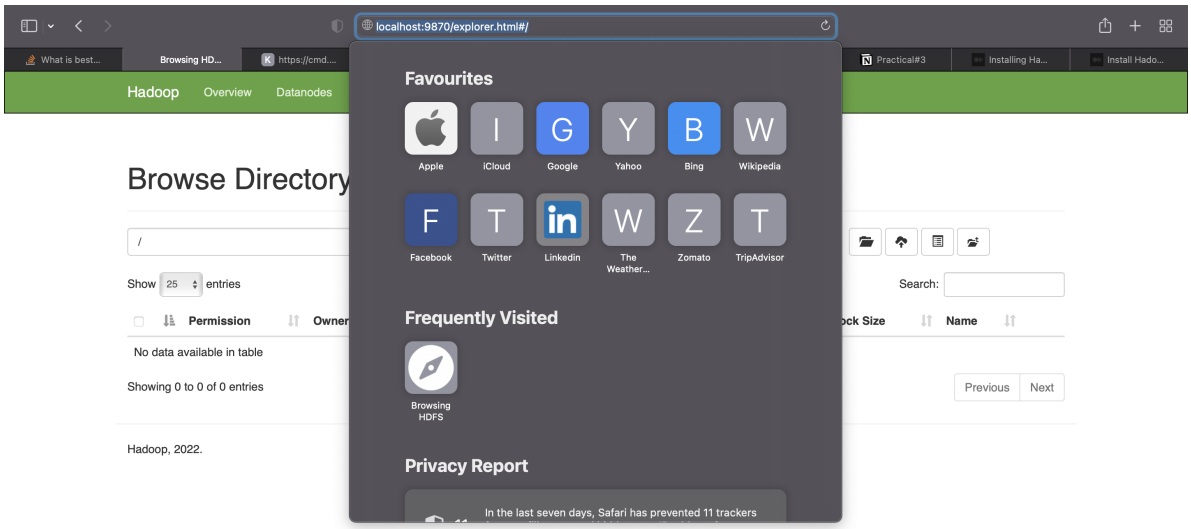



## Hadoop Commands

1. Command ‘start-all.sh’

```
hadoop@Saurins-MacBook-Pro ~ % start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [Saurins-MacBook-Pro.local]
Saurins-MacBook-Pro.local: Warning: Permanently added 'saurins-macbook-pro.local
' (ED25519) to the list of known hosts.
2022-10-21 12:59:43,831 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
Starting resourcemanager
Starting nodemanagers
hadoop@Saurins-MacBook-Pro ~ %
```

Now open: localhost:9870 on browser

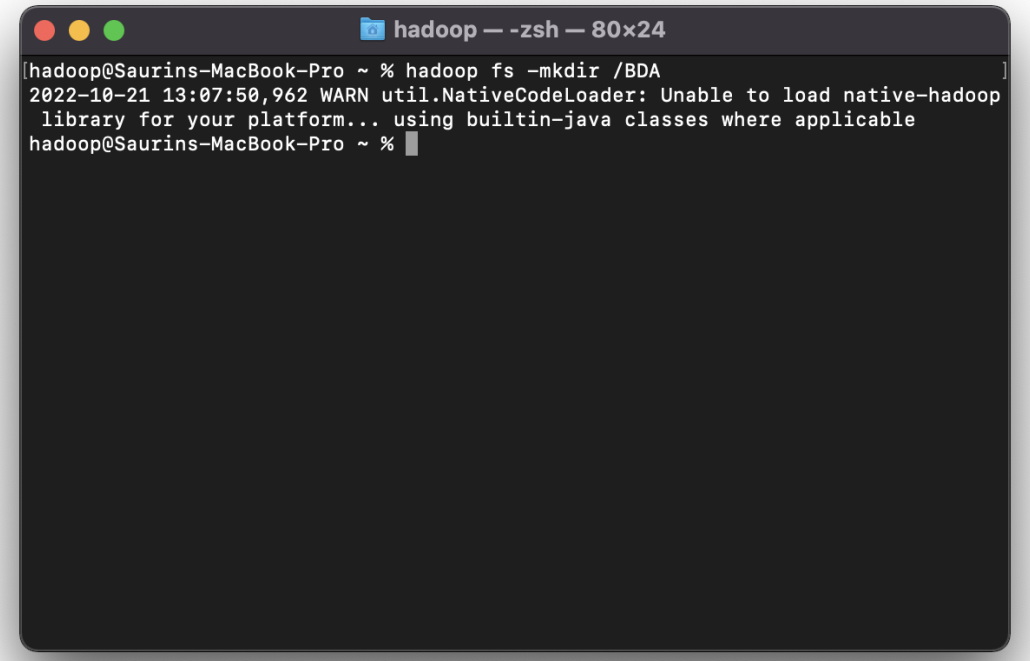


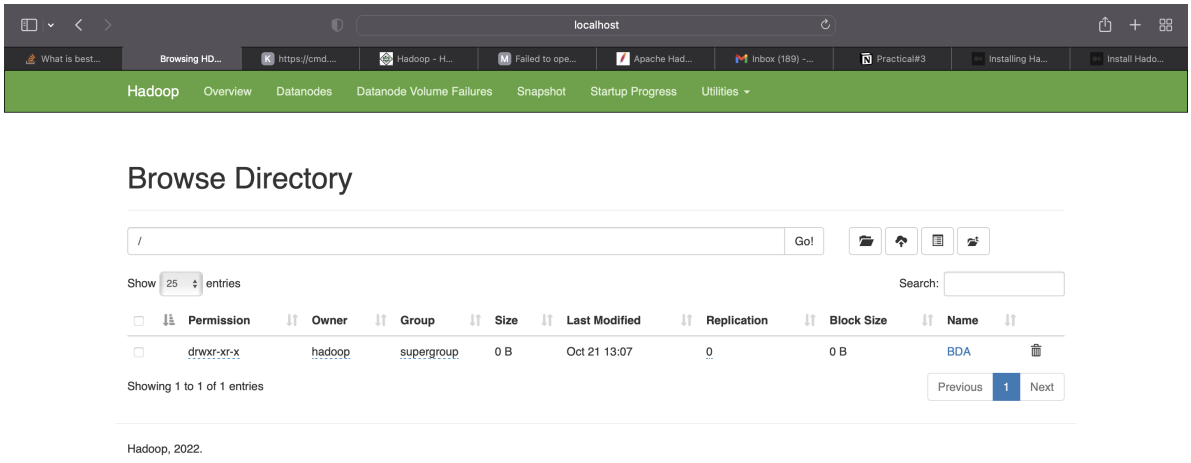

Currently empty

2. Command ‘jps’

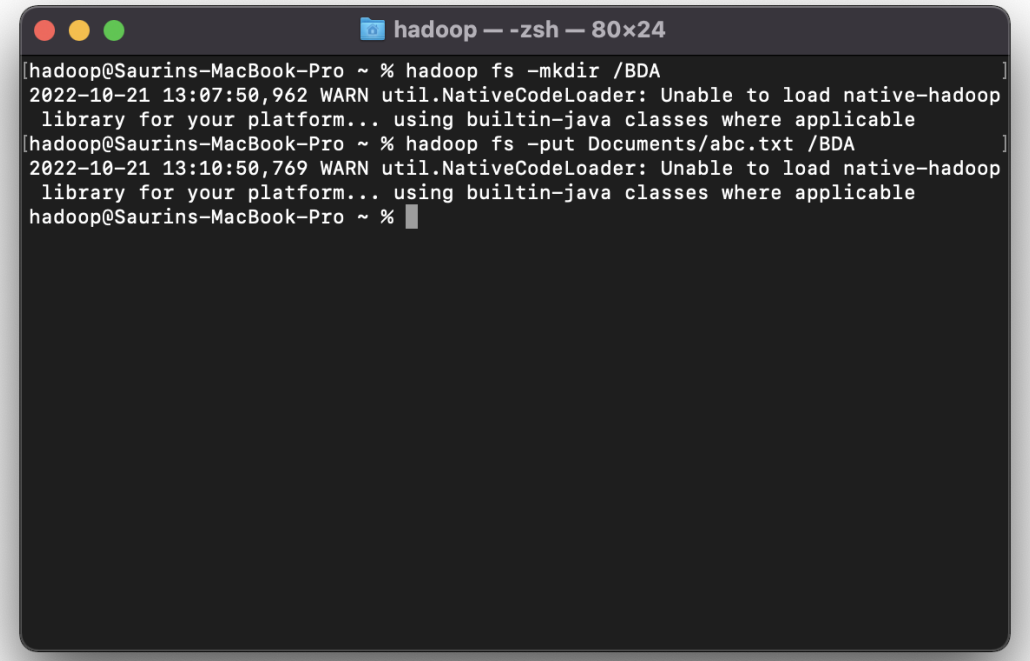
```
hadoop@Saurins-MacBook-Pro ~ % jps
64561 NameNode
64664 DataNode
65083 NodeManager
64988 ResourceManager
65182 Jps
64799 SecondaryNameNode
hadoop@Saurins-MacBook-Pro ~ %
```

3. Command ‘mkdir’



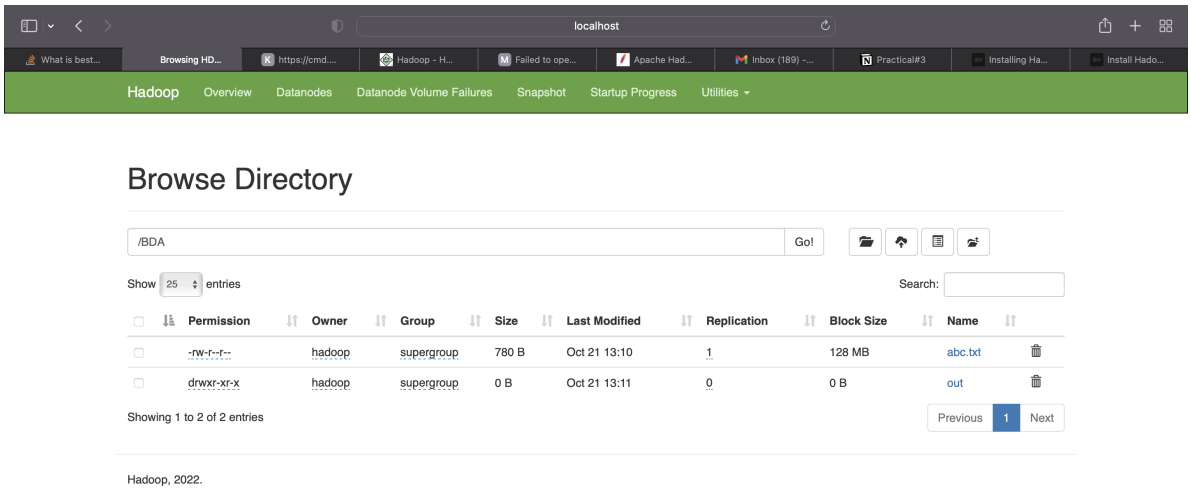

4. Command ‘put’

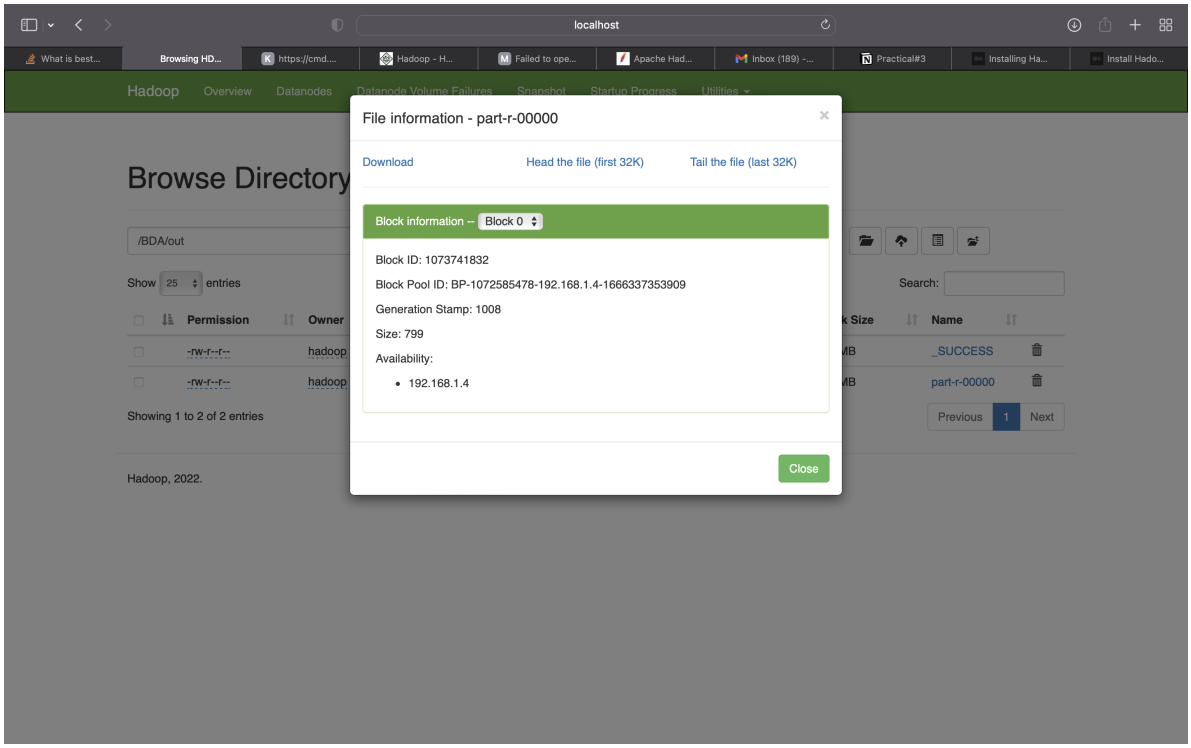


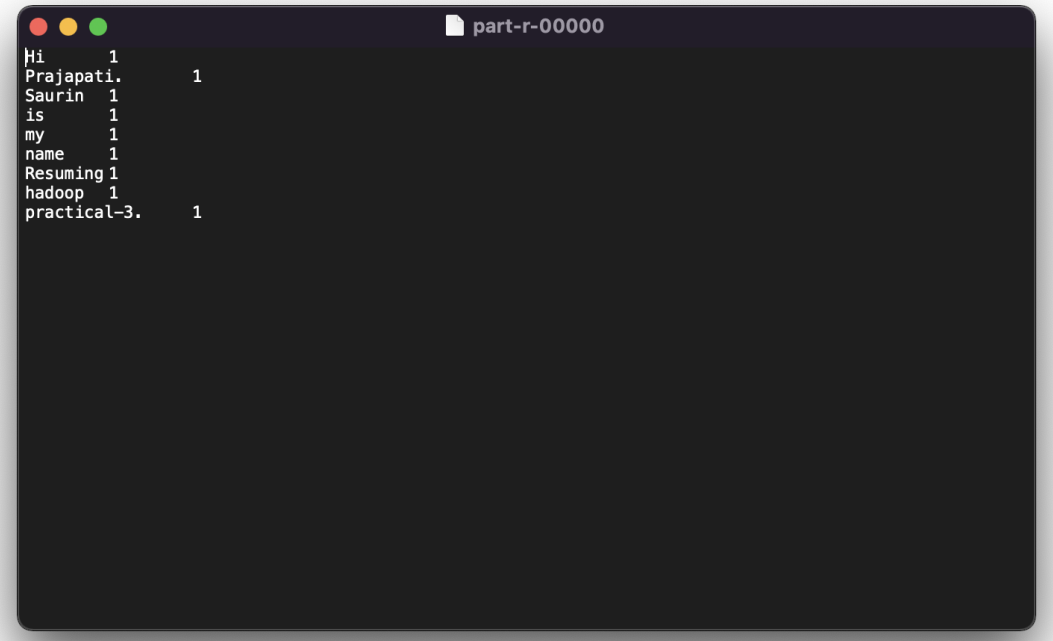

5. jar command to find word count of abc.txt we uploaded


hadoop — -zsh — 88×55
hadoop@Saurins-MacBook-Pro ~ % hadoop jar /opt/homebrew/Cellar/hadoop/3.3.4/libexec/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar wordcount /BDA/abc.txt /BDA/out
2022-10-21 13:11:18,230 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-10-21 13:11:18,880 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2022-10-21 13:11:19,591 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1666337389967_0001
2022-10-21 13:11:19,877 INFO input.FileInputFormat: Total input files to process : 1
2022-10-21 13:11:20,897 INFO mapreduce.JobSubmitter: number of splits:1
2022-10-21 13:11:21,582 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1666337389967_0001
2022-10-21 13:11:21,584 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-10-21 13:11:21,847 INFO conf.Configuration: resource-types.xml not found
2022-10-21 13:11:21,848 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-10-21 13:11:22,329 INFO impl.YarnClientImpl: Submitted application application_1666337389967_0001
2022-10-21 13:11:22,376 INFO mapreduce.Job: The url to track the job: http://Saurins-MacBook-Pro.local:8088/proxy/application_1666337389967_0001/
2022-10-21 13:11:22,377 INFO mapreduce.Job: Running job: job_1666337389967_0001
2022-10-21 13:11:31,710 INFO mapreduce.Job: Job job_1666337389967_0001 running in uber mode : false
2022-10-21 13:11:31,724 INFO mapreduce.Job:  map 0% reduce 0%
2022-10-21 13:11:37,014 INFO mapreduce.Job:  map 100% reduce 0%
2022-10-21 13:11:43,132 INFO mapreduce.Job:  map 100% reduce 100%
2022-10-21 13:11:45,260 INFO mapreduce.Job: Job job_1666337389967_0001 completed successfully
2022-10-21 13:11:45,422 INFO mapreduce.Job: Counters: 50
File System Counters
FILE: Number of bytes read=845
FILE: Number of bytes written=553111
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=878
HDFS: Number of bytes written=799
HDFS: Number of read operations=8
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=1
Launched reduce tasks=1
Data-local map tasks=1
Total time spent by all maps in occupied slots (ms)=2750
Total time spent by all reduces in occupied slots (ms)=3498
Total time spent by all map tasks (ms)=2750
Total time spent by all reduce tasks (ms)=3498
Total vcore-milliseconds taken by all map tasks=2750
Total vcore-milliseconds taken by all reduce tasks=3498
Total megabyte-milliseconds taken by all map tasks=2816000
Total megabyte-milliseconds taken by all reduce tasks=3581952
Map-Reduce Framework
Map input records=2


hadoop — -zsh — 88×55
FILE: Number of bytes written=553111
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=878
HDFS: Number of bytes written=799
HDFS: Number of read operations=8
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=1
Launched reduce tasks=1
Data-local map tasks=1
Total time spent by all maps in occupied slots (ms)=2750
Total time spent by all reduces in occupied slots (ms)=3498
Total time spent by all map tasks (ms)=2750
Total time spent by all reduce tasks (ms)=3498
Total vcore-milliseconds taken by all map tasks=2750
Total vcore-milliseconds taken by all reduce tasks=3498
Total megabyte-milliseconds taken by all map tasks=2816000
Total megabyte-milliseconds taken by all reduce tasks=3581952
Map-Reduce Framework
Map input records=2
Map output records=9
Map output bytes=819
Map output materialized bytes=845
Input split bytes=98
Combine input records=9
Combine output records=9
Reduce input groups=9
Reduce shuffle bytes=845
Reduce input records=9
Reduce output records=9
Spilled Records=18
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=112
CPU time spent (ms)=0
Physical memory (bytes) snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=489160704
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=780
File Output Format Counters
Bytes Written=799
hadoop@Saurins-MacBook-Pro ~ %


localhost
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/BDA
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r--
hadoop
supergroup
780 B
Oct 21 13:10
1
128 MB
abc.txt
drwxr-xr-x
hadoop
supergroup
0 B
Oct 21 13:11
0
0 B
out
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2022.


localhost
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
File information - part-r-00000
Download
Head the file (first 32K)
Tail the file (last 32K)
Block information -- Block 0
Block ID: 1073741832
Block Pool ID: BP-1072585478-192.168.1.4-1666337353909
Generation Stamp: 1008
Size: 799
Availability:
192.168.1.4
Close
Browse Directory
/BDA/out
Show 25 entries
Search:
Permission
Owner
Name
-rw-r--r--
hadoop
_SUCCESS
-rw-r--r--
hadoop
part-r-00000
Showing 1 to 2 of 2 entries
Previous
1
Next
Hadoop, 2022.


part-r-00000
Hi 1
Prajapati. 1
Saurin 1
is 1
my 1
name 1
Resuming 1
hadoop 1
practical-3. 1

Thank You.