

ECC Lab 1

Name: Samir Sanyal

Rajat Sawant

Objective:

- 1) Deploy a Single-Node Hadoop Cluster on Jetstream2.
- 2) Validate and enforce firewall rules via UFW and enable SSH key-pair authentication.
- 3) Install and validate Hadoop.
- 4) Configure the MapReduce framework.
- 5) Compile and run the WordStandardDeviation.java example to confirm cluster operability.

Procedure:

- To start the project, we setup our instance on Jetstream. We have used the image E516_Ubuntu22_UFW_Enabled with the flavor m3.small following the steps as shown in the video provided

Home > Allocation CIS240523 > Instances > Instance lab1_intro_team_14

Jetstream2 IU - CIS240523 (logged in as rsawant1@access-ci.org) Remove Allocation

Instance lab1_intro_team_14

Shelved

Info 0cb21d52-7989-4329-b068-67610b0f095f

created 8 days ago by user rsawant1@access-ci.org from image E516_Ubuntu22_UFW_Enabled flavor m3.small Burn rate 0.00 S

Interactions

- Web Shell
- Web Desktop
- Native SSH
- Console
- Workflow

Credentials

Public IP Address 149.165.171.44 Unassign

IP Details

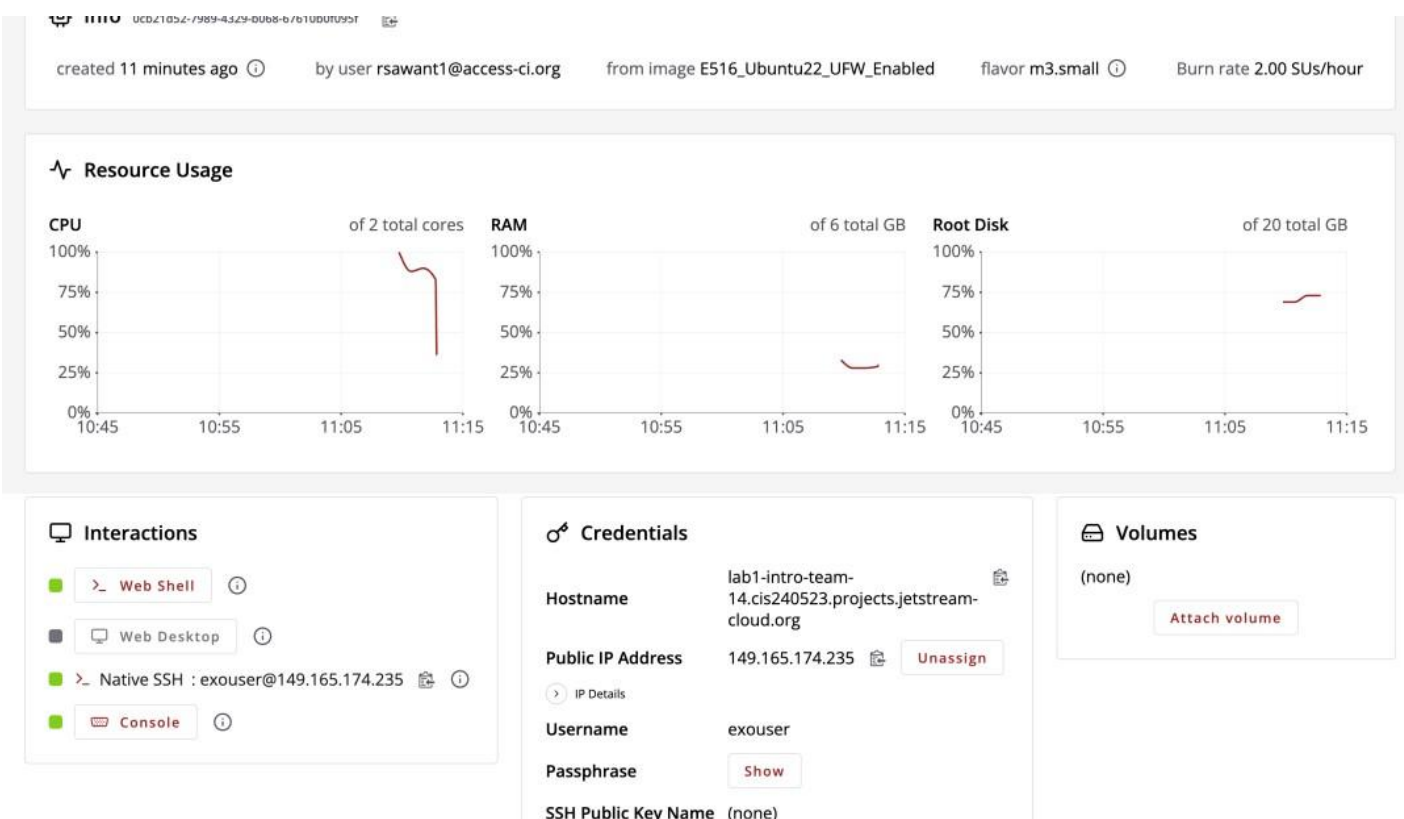
Username exouser

Passphrase Show

SSH Public Key Name (none)

Volumes

(none)



- After connecting to the instance, we first checked the firewall status. To do this we ran the command **sudo ufw status verbose**. Output showed active firewall with default deny (incoming) and allow (outgoing). Only SSH (port 22) was allowed. We then added rules to allow to permit all the incoming traffic from the IP's of our team and not just from the SSH

```

exouser@lab1-intro-team-14:~$ sudo ufw status verbose
Status: active
Logging: on (low)
Default: deny (incoming), allow (outgoing), deny (routed)
New profiles: skip

To Action From
--
22/tcp (OpenSSH) ALLOW IN Anywhere
22/tcp ALLOW IN Anywhere
22/tcp (OpenSSH (v6)) ALLOW IN Anywhere (v6)
22/tcp (v6) ALLOW IN Anywhere (v6)

exouser@lab1-intro-team-14:~$ sudo su -
root@lab1-intro-team-14:~# whoami
root
root@lab1-intro-team-14:~# sudo ufw allow from 68.50.13.204
Rule added
root@lab1-intro-team-14:~# sudo ufw allow from 68.50.22.130
Rule added
root@lab1-intro-team-14:~# sudo ufw status verbose
Status: active
Logging: on (low)
Default: deny (incoming), allow (outgoing), deny (routed)
New profiles: skip

To Action From
--
22/tcp (OpenSSH) ALLOW IN Anywhere
22/tcp ALLOW IN Anywhere
Anywhere ALLOW IN 68.50.13.204
Anywhere ALLOW IN 68.50.22.130
22/tcp (OpenSSH (v6)) ALLOW IN Anywhere (v6)
22/tcp (v6) ALLOW IN Anywhere (v6)

root@lab1-intro-team-14:~#

```

- To generate the public ssh key, we used the command shown in the figure below. Once we have our public key, we copied it into the `/.ssh/authorized_keys` directory

```

MINGW64:/c/Users/DELL
DELL@DESKTOP-ODNUD4G MINGW64 ~
$ ssh-keygen -t ed25519 -C "sasanya@iu.edu"

```

- Once we have the key in place, we attempted to reestablish our connection with the instance without a password which was successful as shown in the image below:

exouser@lab1-intro-team-14: ~

DELL@DESKTOP-ODNUD4G MINGW64 ~

\$ ssh exouser@149.165.174.235

System information as of Fri Oct 11 15:42:03 UTC 2024

System load:	0.52	Processes:	246
Usage of /:	53.8% of 19.20GB	Users logged in:	0
Memory usage:	40%	IPv4 address for ens3:	10.3.34.14
Swap usage:	0%		

<https://jetstream.status.io/>

Overall Jetstream2 Status: **Operational**

Active Status Items:

. **System-wide Networking Issues**

Last login: Sat Feb 8 04:17:01 2025 from 68.50.13.204

exouser@lab1-intro-team-14:~\$

- We proceeded to create a user called hadoop. We also tested if the hadoop user can perform ssh without password.

```

BAD PASSWORD: The password contains the user name in some form
Retype new password:
Sorry, passwords do not match.
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for hadoop
Enter the new value, or press ENTER for the default
    Full Name []:
    Room Number []:
    Work Phone []:
    Home Phone []:
    Other []:
Is the information correct? [Y/n] y
exouser@lab1-intro-team-14:~$ cat /etc/passwd
root:x:0:0:root:/root:/bin/bash
daemon:x:1:1:daemon:/usr/sbin:/usr/sbin/nologin
bin:x:2:2:bin:/bin:/usr/sbin/nologin
sys:x:3:3:sys:/dev:/usr/sbin/nologin
sync:x:4:65534:sync:/bin:/bin/sync
games:x:5:60:games:/usr/games:/usr/sbin/nologin
man:x:6:12:man:/var/cache/man:/usr/sbin/nologin
lp:x:7:7:lp:/var/spool/lpd:/usr/sbin/nologin
mail:x:8:8:mail:/var/mail:/usr/sbin/nologin
news:x:9:9:news:/var/spool/news:/usr/sbin/nologin
uucp:x:10:10:uucp:/var/spool/uucp:/usr/sbin/nologin
proxy:x:13:13:proxy:/bin:/usr/sbin/nologin
www-data:x:33:33:www-data:/var/www:/usr/sbin/nologin
backup:x:34:34:backup:/var/backups:/usr/sbin/nologin
list:x:38:38:Mail List Manager:/var/list:/usr/sbin/nologin
irc:x:39:39:ircd:/run/ircd:/usr/sbin/nologin
gnats:x:41:41:Gnats Bug-Reporting System (admin):/var/lib/gnats:/usr/sbin/nologin
nobody:x:65534:65534:nobody:/nonexistent:/usr/sbin/nologin
systemd-networkd:x:100:102:systemd Network Management,,:/run/systemd:/usr/sbin/nologin
systemd-resolve:x:101:103:systemd Resolver,,:/run/systemd:/usr/sbin/nologin
messagebus:x:102:105:/:/nonexistent:/usr/sbin/nologin
systemd-timesyncd:x:103:106:systemd Time Synchronization,,:/run/systemd:/usr/sbin/nologin
syslog:x:104:111:/:/home/syslog:/usr/sbin/nologin
_apt:x:105:65534:/:/nonexistent:/usr/sbin/nologin
tss:x:106:112:TPM software stack,,:/var/lib/tpm:/bin/false
uuidd:x:107:113:/:/run/uuidd:/usr/sbin/nologin
tcpdump:x:108:114:/:/nonexistent:/usr/sbin/nologin
sshd:x:109:65534:/:/run/sshd:/usr/sbin/nologin
pollinate:x:110:1:/:/var/cache/pollinate:/bin/false
landscape:x:111:116:/:/var/lib/landscape:/usr/sbin/nologin
fwupd-refresh:x:112:117:fwupd-refresh user,,:/run/systemd:/usr/sbin/nologin
ubuntu:x:1000:1000:Ubuntu:/home/ubuntu:/bin/bash
lxd:x:999:100:/:/var/snap/lxd/common/lxd:/bin/false
ceph:x:64045:64045:Ceph storage service:/var/lib/ceph:/usr/sbin/nologin
rtkit:x:113:122:RealtimeKit,,:/proc:/usr/sbin/nologin
dnsmasq:x:114:65534:dnsmasq,,:/var/lib/misc:/usr/sbin/nologin
kernoops:x:115:65534:Kernel Oops Tracking Daemon,,:/usr/sbin/nologin
cups-pk-helper:x:116:123:user for cups-pk-helper service,,:/home/cups-pk-helper:/usr/sbin/nologin
systemd-oomd:x:117:126:systemd Userspace OOM Killer,,:/run/systemd:/usr/sbin/nologin
whoopsie:x:118:127:/:/nonexistent:/bin/false
avahi-autoipd:x:119:128:Avahi autoip daemon,,:/var/lib/avahi-autoipd:/usr/sbin/nologin
usbmux:x:120:46:usbmux daemon,,:/var/lib/usbmux:/usr/sbin/nologin
avahi:x:121:129:Avahi mDNS daemon,,:/run/avahi-daemon:/usr/sbin/nologin
nm-openvpn:x:122:130:NetworkManager OpenVPN,,:/var/lib/openvpn/chroot:/usr/sbin/nologin
geoclue:x:123:131:/:/var/lib/geoclue:/usr/sbin/nologin
saned:x:124:133:/:/var/lib/saned:/usr/sbin/nologin
colord:x:125:134:colord colour management daemon,,:/var/lib/colord:/usr/sbin/nologin
sssd:x:126:135:SSSD system user,,:/var/lib/sss:/usr/sbin/nologin
pulse:x:127:136:PulseAudio daemon,,:/run/pulse:/usr/sbin/nologin
speech-dispatcher:x:128:29:Speech Dispatcher,,:/run/speech-dispatcher:/bin/false
hplip:x:129:7:HPLIP system user,,:/run/hplip:/bin/false
gnome-initial-setup:x:130:65534:/:/run/gnome-initial-setup:/bin/false
gdm:x:131:138:Gnome Display Manager:/var/lib/gdm3:/bin/false
exouser:x:1001:1001:/:/home/exouser:/bin/bash
hadoop:x:1002:1003:/:/home/hadoop:/bin/bash
exouser@lab1-intro-team-14:~$

```

- We used wget to download the apache files and packages using the command as shown below and unzipped it using the tar command.

```

hadoop@lab1-intro-team-14:~$ history
1  clear
2  ls -ltr
3  ls -ltr -a
4  wget https://d1cdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
5  tar -xvzf hadoop-3.3.6.tar.gz
6  history
hadoop@lab1-intro-team-14:~$ |

```

- We used the vi editor to add the following commands to .bashsrc file:
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

```

export HADOOP_HOME=/home/hadoop/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native export
PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"

```

- To load the configuration we used the command `source ~/.bashrc` By editing the `.bashrc` file we have added the correct path the packages and `openjdk`.

```

hadoop@lab1-intro-team-14:~$ ^C
hadoop@lab1-intro-team-14:~$ vi .bashrc
hadoop@lab1-intro-team-14:~$ source ~/.bashrc
hadoop@lab1-intro-team-14:~$ hadoop version
Hadoop 3.3.6
Source code repository https://github.com/apache/hadoop.git -r 1be78238728da9266a4f88195058f08fd012bf9c
Compiled by ubuntu on 2023-06-18T08:22Z
Compiled on platform linux-x86_64
Compiled with protoc 3.7.1
From source with checksum 5652179ad55f76cb287d9c633bb53bbd
This command was run using /home/hadoop/hadoop/share/hadoop/common/hadoop-common-3.3.6.jar
hadoop@lab1-intro-team-14:~$ |

```

- We checked the `openjdk` version and by default Ubuntu uses 11 so we uninstalled and downgraded it to 8 so that hive does not cause any issue.

```

hadoop@lab1-intro-team-14:~$ java -version
openjdk version "1.8.0_442"
OpenJDK Runtime Environment (build 1.8.0_442-8u442-b06~us1-0ubuntu1~22.04-b06)
OpenJDK 64-Bit Server VM (build 25.442-b06, mixed mode)
hadoop@lab1-intro-team-14:~$

```

- We also configured `JAVA_HOME` in `hadoop-env.sh` file. Edit the Hadoop environment variable file in the text editor to setup the variable environment. Also checked the hostname as changes needed to be made to configuration files.

```

hadoop@lab1-intro-team-14:~/hadoop/etc$ cd /hadoop/
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ ls -ltr
total 180
-rw-r--r-- 1 hadoop hadoop 10 Jun 18 2023 workers
-rw-r--r-- 1 hadoop hadoop 2697 Jun 18 2023 ssl-server.xml.example
-rw-r--r-- 1 hadoop hadoop 2316 Jun 18 2023 ssl-client.xml.example
drwxr-xr-x 2 hadoop hadoop 4096 Jun 18 2023 shellprofile.d
-rw-r--r-- 1 hadoop hadoop 13700 Jun 18 2023 log4j.properties
-rw-r--r-- 1 hadoop hadoop 3414 Jun 18 2023 hadoop-user-functions.sh.example
-rw-r--r-- 1 hadoop hadoop 11765 Jun 18 2023 hadoop-policy.xml
-rw-r--r-- 1 hadoop hadoop 3321 Jun 18 2023 hadoop-metrics2.properties
-rw-r--r-- 1 hadoop hadoop 3999 Jun 18 2023 hadoop-env.cmd
-rw-r--r-- 1 hadoop hadoop 682 Jun 18 2023 kms-site.xml
-rw-r--r-- 1 hadoop hadoop 1860 Jun 18 2023 kms-log4j.properties
-rw-r--r-- 1 hadoop hadoop 1351 Jun 18 2023 kms-env.sh
-rw-r--r-- 1 hadoop hadoop 3518 Jun 18 2023 kms-acls.xml
-rw-r--r-- 1 hadoop hadoop 2681 Jun 18 2023 user_ec_policies.xml.template
-rw-r--r-- 1 hadoop hadoop 620 Jun 18 2023 httpfs-site.xml
-rw-r--r-- 1 hadoop hadoop 1657 Jun 18 2023 httpfs-log4j.properties
-rw-r--r-- 1 hadoop hadoop 1484 Jun 18 2023 httpfs-env.sh
-rw-r--r-- 1 hadoop hadoop 683 Jun 18 2023 hdfs-rbf-site.xml
-rw-r--r-- 1 hadoop hadoop 2591 Jun 18 2023 yarnservice-log4j.properties
-rw-r--r-- 1 hadoop hadoop 6329 Jun 18 2023 yarn-env.sh
-rw-r--r-- 1 hadoop hadoop 2250 Jun 18 2023 yarn-env.cmd
-rw-r--r-- 1 hadoop hadoop 2567 Jun 18 2023 container-executor.cfg
-rw-r--r-- 1 hadoop hadoop 9213 Jun 18 2023 capacity-scheduler.xml
-rw-r--r-- 1 hadoop hadoop 4113 Jun 18 2023 mapred-queues.xml.template
-rw-r--r-- 1 hadoop hadoop 1764 Jun 18 2023 mapred-env.sh
-rw-r--r-- 1 hadoop hadoop 951 Jun 18 2023 mapred-env.cmd
-rw-r--r-- 1 hadoop hadoop 1335 Jun 18 2023 configuration.xsl
-rw-r--r-- 1 hadoop hadoop 16668 Feb 8 17:21 hadoop-env.sh
-rw-r--r-- 1 hadoop hadoop 893 Feb 8 17:25 core-site.xml
-rw-r--r-- 1 hadoop hadoop 1152 Feb 8 17:27 hdfs-site.xml
-rw-r--r-- 1 hadoop hadoop 1215 Feb 8 17:28 mapred-site.xml
-rw-r--r-- 1 hadoop hadoop 758 Feb 8 17:29 yarn-site.xml
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ |

```

```

hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ cat hadoop-env.sh | grep export
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
# export HADOOP_HOME=
# export HADOOP_CONF_DIR=${HADOOP_HOME}/etc/hadoop
# export HADOOP_HEAPSIZE_MAX=
# export HADOOP_HEAPSIZE_MIN=
# export HADOOP_JAAS_DEBUG=true
# export HADOOP_OPTS="-Djava.net.preferIPv4Stack=true"
# export HADOOP_OPTS="-Djava.net.preferIPv4Stack=true -Dsun.security.krb5.debug=true -Dsun.security.spnego.debug"
export HADOOP_OS_TYPE=${HADOOP_OS_TYPE:-$(uname -s)}
# export HADOOP_CLIENT_OPTS=""
# export HADOOP_CLASSPATH="/some/cool/path/on/your/machine"
# export HADOOP_USER_CLASSPATH_FIRST="yes"
# export HADOOP_USE_CLIENT_CLASSLOADER=true
# export HADOOP_CLIENT_CLASSLOADER_SYSTEM_CLASSES="-org.apache.hadoop.UserClass,java.,javax.,org.apache.hadoop."
# export HADOOP_OPTIONAL_TOOLS="hadoop-kafka,hadoop-aws,hadoop-azure-data-lake,hadoop-aliyun,hadoop-azure"

```

```

hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ hostname
lab1-intro-team-14
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$

```

- Next, we edited the core-site.xml file and updated it with instance name: **vi etc/hadoop/core-site.xml**

```
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ cat core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://lab1-intro-team-14:9000</value>
  </property>
</configuration>
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ |
```

- Then, we edited the hdfs-site.xml file: **vi etc/hadoop/hdfs-site.xml** and used vi to edit the hdfs-site.xml file


```

hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ cat hdfs-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:///home/hadoop/hadoopdata/hdfs/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>
  </property>
</configuration>
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$

```

- Then, we edited the hdfs-site.xml file: **vi etc/hadoop/mapred-site.xml** and used vi to edit the mapred-site.xml file

```

hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ cat mapred-site.xml
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>yarn.app.mapreduce.am.env</name>
    <value>HADOOP_MAPRED_HOME=$HADOOP_HOME/home/hadoop/hadoop/bin/hadoop</value>
  </property>
  <property>
    <name>mapreduce.map.env</name>
    <value>HADOOP_MAPRED_HOME=$HADOOP_HOME/home/hadoop/hadoop/bin/hadoop</value>
  </property>
  <property>
    <name>mapreduce.reduce.env</name>
    <value>HADOOP_MAPRED_HOME=$HADOOP_HOME/home/hadoop/hadoop/bin/hadoop</value>
  </property>
</configuration>
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$

```

- We then edited the the yarn-site.xml file: **vi etc/hadoop/yarn-site.xml**

```

hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ cat yarn-site.xml
<?xml version="1.0"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
</configuration>
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$

```

- Before starting the Hadoop cluster, we formatted the namenode as a hadoop user and ran the command **hdfs namenode -format**. By doing this we format the namenode and make it usable.

```

hadoop@lab1-intro-team-14: ~/hadoop/etc/hadoop
2025-02-08 05:17:15,576 INFO namenode.NameNode: Formatting using clusterid: CID-6970e893-Sdb0-41a4-a339-26edccb49ef
2025-02-08 05:17:15,645 INFO namenode.FSEditLog: Edit logging is async:true
2025-02-08 05:17:15,710 INFO namenode.FSNamesystem: KeyProvider: null
2025-02-08 05:17:15,712 INFO namenode.FSNamesystem: fsLock is fair: true
2025-02-08 05:17:15,712 INFO namenode.FSNamesystem: Detailed lock hold time metrics enabled: false
2025-02-08 05:17:15,750 INFO namenode.FSNamesystem: fsOwner = hadoop (auth:SIMPLE)
2025-02-08 05:17:15,750 INFO namenode.FSNamesystem: supergroup = supergroup
2025-02-08 05:17:15,750 INFO namenode.FSNamesystem: isPermissionEnabled = true
2025-02-08 05:17:15,750 INFO namenode.FSNamesystem: isStoragePolicyEnabled = true
2025-02-08 05:17:15,750 INFO namenode.FSNamesystem: HA Enabled: false
2025-02-08 05:17:15,789 INFO common.Util: dfs.datanode.fileio.profiling.sampling.percentage set to 0. Disabling file IO profiling
2025-02-08 05:17:16,062 INFO blockmanagement.DatanodeManager: dfs.block.invalidate.limit : configured=1000, counted=60, effected=1000
2025-02-08 05:17:16,063 INFO blockmanagement.DatanodeManager: dfs.namenode.datanode.registration.ip-hostname-check=true
2025-02-08 05:17:16,066 INFO blockmanagement.BlockManager: dfs.namenode.startup.delay.block.deletion.sec is set to 000:00:00:00.000
2025-02-08 05:17:16,067 INFO blockmanagement.BlockManager: The block deletion will start around 2025 Feb 08 05:17:16
2025-02-08 05:17:16,068 INFO util.GSet: Computing capacity for map BlocksMap
2025-02-08 05:17:16,069 INFO util.GSet: VM type = 64-bit
2025-02-08 05:17:16,070 INFO util.GSet: 2.0% max memory 1.4 GB = 29.6 MB
2025-02-08 05:17:16,070 INFO util.GSet: capacity = 2^22 = 4194304 entries
2025-02-08 05:17:16,094 INFO blockmanagement.BlockManager: Storage policy satisfier is disabled
2025-02-08 05:17:16,095 INFO blockmanagement.BlockManager: dfs.block.access.token.enable = false
2025-02-08 05:17:16,102 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.threshold-pct = 0.999
2025-02-08 05:17:16,102 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.min.datanodes = 0
2025-02-08 05:17:16,102 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safemode.extension = 30000
2025-02-08 05:17:16,103 INFO blockmanagement.BlockManager: defaultReplication = 1
2025-02-08 05:17:16,103 INFO blockmanagement.BlockManager: maxReplication = 512
2025-02-08 05:17:16,103 INFO blockmanagement.BlockManager: minReplication = 1
2025-02-08 05:17:16,103 INFO blockmanagement.BlockManager: maxReplicationStreams = 2
2025-02-08 05:17:16,103 INFO blockmanagement.BlockManager: redundancyRecheckInterval = 3000ms
2025-02-08 05:17:16,103 INFO blockmanagement.BlockManager: encryptDataTransfer = false
2025-02-08 05:17:16,103 INFO blockmanagement.BlockManager: maxNumBlocksToLog = 1000
2025-02-08 05:17:16,141 INFO namenode.FSDirectory: GLOBAL serial map: bits=29 maxEntries=536870911
2025-02-08 05:17:16,141 INFO namenode.FSDirectory: USER serial map: bits=24 maxEntries=16777215
2025-02-08 05:17:16,141 INFO namenode.FSDirectory: GROUP serial map: bits=24 maxEntries=16777215
2025-02-08 05:17:16,144 INFO namenode.FSDirectory: XATTR serial map: bits=24 maxEntries=16777215
2025-02-08 05:17:16,155 INFO util.GSet: Computing capacity for map INodeMap
2025-02-08 05:17:16,155 INFO util.GSet: VM type = 64-bit
2025-02-08 05:17:16,156 INFO util.GSet: 1.0% max memory 1.4 GB = 14.8 MB
2025-02-08 05:17:16,156 INFO util.GSet: capacity = 2^21 = 2097152 entries
2025-02-08 05:17:16,162 INFO namenode.FSDirectory: ACLs enabled? true
2025-02-08 05:17:16,162 INFO namenode.FSDirectory: POSIX ACL inheritance enabled? true
2025-02-08 05:17:16,162 INFO namenode.FSDirectory: XAttrs enabled? true
2025-02-08 05:17:16,162 INFO namenode.NameNode: Caching file names occurring more than 10 times
2025-02-08 05:17:16,167 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: false, skipCaptureAccessTimeOnlyChange: false, snap
2025-02-08 05:17:16,168 INFO snapshot.SnapshotManager: Skiplist is disabled
2025-02-08 05:17:16,172 INFO util.GSet: Computing capacity for map cachedBlocks
2025-02-08 05:17:16,172 INFO util.GSet: VM type = 64-bit
2025-02-08 05:17:16,172 INFO util.GSet: 0.25% max memory 1.4 GB = 3.7 MB
2025-02-08 05:17:16,172 INFO util.GSet: capacity = 2^19 = 524288 entries
2025-02-08 05:17:16,182 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2025-02-08 05:17:16,182 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2025-02-08 05:17:16,182 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2025-02-08 05:17:16,188 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2025-02-08 05:17:16,188 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 mi
2025-02-08 05:17:16,193 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2025-02-08 05:17:16,193 INFO util.GSet: VM type = 64-bit
2025-02-08 05:17:16,193 INFO util.GSet: 0.029999999329447746% max memory 1.4 GB = 455.3 KB
2025-02-08 05:17:16,193 INFO util.GSet: capacity = 2^16 = 65536 entries
2025-02-08 05:17:16,217 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1762509825-10.3.34.210-1738991836209
2025-02-08 05:17:16,248 INFO common.Storage: Storage directory /tmp/hadoop-hadoop/dfs/name has been successfully formatted.
2025-02-08 05:17:16,294 INFO namenode.FSImageFormatProtobuf: Saving image file /tmp/hadoop-hadoop/dfs/name/current/fsimage.ckpt_0000000000
2025-02-08 05:17:16,495 INFO namenode.FSImageFormatProtobuf: Image file /tmp/hadoop-hadoop/dfs/name/current/fsimage.ckpt_000000000000000000
2025-02-08 05:17:16,517 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2025-02-08 05:17:16,542 INFO namenode.FSNamesystem: Stopping services started for active state
2025-02-08 05:17:16,542 INFO namenode.FSNamesystem: Stopping services started for standby state
2025-02-08 05:17:16,546 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2025-02-08 05:17:16,547 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at lab1-intro-team-14/10.3.34.210
*****/
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$

```

```

/ *****
SHUTDOWN_MSG: Shutting down NameNode at lab1-intro-team-14/10.3.34.210
***** /
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ 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 [lab1-intro-team-14]
Starting datanodes
Starting secondary namenodes [lab1-intro-team-14]
Starting resourcemanager
Starting nodemanagers
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ jps
144676 SecondaryNameNode
145781 Jps
145050 ResourceManager
145243 NodeManager
144139 NameNode
144350 DataNode
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ hdfs dfs -mkdir /test1
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ hdfs dfs -mkdir /logs
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ hdfs dfs -ls /
Found 2 items
drwxr-xr-x - hadoop supergroup          0 2025-02-08 17:31 /logs
drwxr-xr-x - hadoop supergroup          0 2025-02-08 17:31 /test1
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$

```

- We then started the hadoop cluster using the command **start-all.sh** We also used **jps** command to check the status of all the hadoop services.

```

/ *****
SHUTDOWN_MSG: Shutting down NameNode at lab1-intro-team-14/10.3.34.210
***** /
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ 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 [lab1-intro-team-14]
Starting datanodes
Starting secondary namenodes [lab1-intro-team-14]
Starting resourcemanager
Starting nodemanagers
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ jps
144676 SecondaryNameNode
145781 Jps
145050 ResourceManager
145243 NodeManager
144139 NameNode
144350 DataNode
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ hdfs dfs -mkdir /test1
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ hdfs dfs -mkdir /logs
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ hdfs dfs -ls /
Found 2 items
drwxr-xr-x - hadoop supergroup          0 2025-02-08 17:31 /logs
drwxr-xr-x - hadoop supergroup          0 2025-02-08 17:31 /test1
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$

```

- At this point, the Hadoop cluster is installed and configured. Next we created some directories in the HDFS filesystem to test the Hadoop. We have successfully created the directories and confirmed that Hadoop has been installed correctly and its working fine.

- Now that we know everything is working as intended we copied the word count java code into our server. We directly copied the code and we used touch to create .java file on the instance. The process involved the steps as mentioned below:

1. Compile WordStandardDeviation.java

```
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ hdfs dfs -ls /
Found 2 items
drwxr-xr-x - hadoop supergroup 0 2025-02-08 17:31 /logs
drwxr-xr-x - hadoop supergroup 0 2025-02-08 17:31 /test1
hadoop@lab1-intro-team-14:~/hadoop/etc/hadoop$ cd ../../
hadoop@lab1-intro-team-14:~/hadoop$ touch WordStandardDeviation.java
hadoop@lab1-intro-team-14:~/hadoop$ vi WordStandardDeviation.java
hadoop@lab1-intro-team-14:~/hadoop$ javac -classpath $HADOOP_HOME/share/hadoop/common/hadoop-common-3.3.6.jar:$HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-core-3.3.6.jar WordStandardDeviation.java
hadoop@lab1-intro-team-14:~/hadoop$ ls -ltr
total 128
-rw-rw-r-- 1 hadoop hadoop 175 Jun 9 2023 README.txt
-rw-rw-r-- 1 hadoop hadoop 1541 Jun 9 2023 NOTICE.txt
-rw-rw-r-- 1 hadoop hadoop 29473 Jun 9 2023 NOTICE-binary
-rw-rw-r-- 1 hadoop hadoop 15217 Jun 9 2023 LICENSE.txt
-rw-rw-r-- 1 hadoop hadoop 24276 Jun 14 2023 LICENSE-binary
drwxr-xr-x 3 hadoop hadoop 4096 Jun 18 2023 sbin
drwxr-xr-x 3 hadoop hadoop 4096 Jun 18 2023 etc
drwxr-xr-x 2 hadoop hadoop 4096 Jun 18 2023 licenses-binary
drwxr-xr-x 3 hadoop hadoop 4096 Jun 18 2023 lib
drwxr-xr-x 2 hadoop hadoop 4096 Jun 18 2023 bin
drwxr-xr-x 2 hadoop hadoop 4096 Jun 18 2023 include
drwxr-xr-x 4 hadoop hadoop 4096 Jun 18 2023 libexec
drwxr-xr-x 4 hadoop hadoop 4096 Jun 18 2023 share
drwxrwxr-x 3 hadoop hadoop 4096 Feb 8 17:30 logs
-rw-rw-r-- 1 hadoop hadoop 7235 Feb 8 17:35 WordStandardDeviation.java
drwxrwxr-x 3 hadoop hadoop 4096 Feb 8 17:35 org
hadoop@lab1-intro-team-14:~/hadoop$ jar -cvf wordcount.jar -C org .
added manifest
adding: apache/(in = 0) (out= 0)(stored 0%)
adding: apache/hadoop/(in = 0) (out= 0)(stored 0%)
adding: apache/hadoop/examples/(in = 0) (out= 0)(stored 0%)
adding: apache/hadoop/examples/WordStandardDeviation$WordStandardDeviationReducer.class(in = 1866) (out= 761)(deflated 59%)
adding: apache/hadoop/examples/WordStandardDeviation$WordStandardDeviationMapper.class(in = 2112) (out= 922)(deflated 56%)
adding: apache/hadoop/examples/WordStandardDeviation.class(in = 4944) (out= 2399)(deflated 51%)
hadoop@lab1-intro-team-14:~/hadoop$ ls -ltr
total 136
-rw-rw-r-- 1 hadoop hadoop 175 Jun 9 2023 README.txt
-rw-rw-r-- 1 hadoop hadoop 1541 Jun 9 2023 NOTICE.txt
-rw-rw-r-- 1 hadoop hadoop 29473 Jun 9 2023 NOTICE-binary
-rw-rw-r-- 1 hadoop hadoop 15217 Jun 9 2023 LICENSE.txt
-rw-rw-r-- 1 hadoop hadoop 24276 Jun 14 2023 LICENSE-binary
drwxr-xr-x 3 hadoop hadoop 4096 Jun 18 2023 sbin
drwxr-xr-x 3 hadoop hadoop 4096 Jun 18 2023 etc
drwxr-xr-x 2 hadoop hadoop 4096 Jun 18 2023 licenses-binary
drwxr-xr-x 3 hadoop hadoop 4096 Jun 18 2023 lib
drwxr-xr-x 2 hadoop hadoop 4096 Jun 18 2023 bin
drwxr-xr-x 2 hadoop hadoop 4096 Jun 18 2023 include
drwxr-xr-x 4 hadoop hadoop 4096 Jun 18 2023 libexec
drwxr-xr-x 4 hadoop hadoop 4096 Jun 18 2023 share
drwxrwxr-x 3 hadoop hadoop 4096 Feb 8 17:30 logs
-rw-rw-r-- 1 hadoop hadoop 7235 Feb 8 17:35 WordStandardDeviation.java
drwxrwxr-x 3 hadoop hadoop 4096 Feb 8 17:35 org
-rw-rw-r-- 1 hadoop hadoop 5424 Feb 8 17:35 wordcount.jar
hadoop@lab1-intro-team-14:~/hadoop$ echo -e "hello world\nhello hadoop" > input.txt
hadoop@lab1-intro-team-14:~/hadoop$ hdfs dfs -mkdir /input
hadoop@lab1-intro-team-14:~/hadoop$ hdfs dfs -put input.txt /input/
hadoop@lab1-intro-team-14:~/hadoop$ hdfs dfs -ls /
Found 3 items
drwxr-xr-x - hadoop supergroup 0 2025-02-08 17:36 /input
drwxr-xr-x - hadoop supergroup 0 2025-02-08 17:31 /logs
drwxr-xr-x - hadoop supergroup 0 2025-02-08 17:31 /test1
hadoop@lab1-intro-team-14:~/hadoop$ hdfs dfs -rm -r /output
rm: '/output': No such file or directory
hadoop@lab1-intro-team-14:~/hadoop$
```

2. Run WordStandardDeviation.java
3. WordStandardDeviation.java is a simple map reduce function, it will calculate the number of total words in a distributed way.

```

hadoop@lab1-intro-team-14:~/hadoop$ hadoop jar wordcountj.jar org.apache.hadoop.examples.WordCount /input /output
2025-02-08 17:37:26,967 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-02-08 17:37:27,089 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-02-08 17:37:27,089 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-02-08 17:37:27,452 INFO input.FileInputFormat: Total input files to process : 1
2025-02-08 17:37:27,476 INFO mapreduce.JobSubmitter: number of splits:1
2025-02-08 17:37:27,716 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1050783815_0001
2025-02-08 17:37:27,716 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-02-08 17:37:27,823 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-02-08 17:37:27,824 INFO mapreduce.Job: Running job: job_local1050783815_0001
2025-02-08 17:37:27,825 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-02-08 17:37:27,842 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitter
2025-02-08 17:37:27,844 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-02-08 17:37:27,845 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory
2025-02-08 17:37:27,845 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-02-08 17:37:27,914 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-02-08 17:37:27,915 INFO mapred.LocalJobRunner: Starting task: attempt_local1050783815_0001_m_0000000_0
2025-02-08 17:37:27,959 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitter
2025-02-08 17:37:27,960 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-02-08 17:37:27,960 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory
2025-02-08 17:37:27,998 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-02-08 17:37:28,009 INFO mapred.MapTask: Processing split: hdfs://lab1-intro-team-14:9000/input/input.txt:0+25
2025-02-08 17:37:28,095 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-02-08 17:37:28,095 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-02-08 17:37:28,095 INFO mapred.MapTask: soft limit at 83886080
2025-02-08 17:37:28,096 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-02-08 17:37:28,096 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-02-08 17:37:28,101 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-02-08 17:37:28,245 INFO mapred.LocalJobRunner:
2025-02-08 17:37:28,247 INFO mapred.MapTask: Starting flush of map output
2025-02-08 17:37:28,247 INFO mapred.MapTask: Spilling map output
2025-02-08 17:37:28,247 INFO mapred.MapTask: bufstart = 0; bufend = 41; bufvoid = 104857600
2025-02-08 17:37:28,247 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214384(104857536); length = 13/6553600
2025-02-08 17:37:28,263 INFO mapred.MapTask: Finished spill 0
2025-02-08 17:37:28,276 INFO mapred.Task: Task:attempt_local1050783815_0001_m_0000000_0 is done. And is in the process of committing
2025-02-08 17:37:28,282 INFO mapred.LocalJobRunner: map
2025-02-08 17:37:28,282 INFO mapred.Task: Task 'attempt_local1050783815_0001_m_0000000_0' done.
2025-02-08 17:37:28,289 INFO mapred.Task: Final Counters for attempt_local1050783815_0001_m_0000000_0: Counters: 24
  File System Counters
    FILE: Number of bytes read=281529
    FILE: Number of bytes written=925346
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=25
    HDFS: Number of bytes written=0
    HDFS: Number of read operations=5
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=1
    HDFS: Number of bytes read erasure-coded=0
  Map-Reduce Framework
    Map input records=2
    Map output records=4
    Map output bytes=41
    Map output materialized bytes=43
    Input split bytes=111
    Combine input records=4

```

4. Run hadoop with map reduce **`hadoop jar wordcountj.jar org.apache.hadoop.examples.WordCount /input /output`**

```

        Reduce input groups=3
        Reduce shuffle bytes=43
        Reduce input records=3
        Reduce output records=3
        Spilled Records=3
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=16
        Total committed heap usage (bytes)=357040128
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Output Format Counters
        Bytes Written=25
2025-02-08 17:37:28,534 INFO mapred.LocalJobRunner: Finishing task: attempt_local1050783815_0001_r_000000_0
2025-02-08 17:37:28,534 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-02-08 17:37:28,833 INFO mapreduce.Job: Job job_local1050783815_0001 running in uber mode : false
2025-02-08 17:37:28,833 INFO mapreduce.Job: map 100% reduce 100%
2025-02-08 17:37:28,834 INFO mapreduce.Job: Job job_local1050783815_0001 completed successfully
2025-02-08 17:37:28,844 INFO mapreduce.Job: Counters: 36
    File System Counters
        FILE: Number of bytes read=563176
        FILE: Number of bytes written=1850735
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=50
        HDFS: Number of bytes written=25
        HDFS: Number of read operations=15
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=4
        HDFS: Number of bytes read erasure-coded=0
    Map-Reduce Framework
        Map input records=2
        Map output records=4
        Map output bytes=41
        Map output materialized bytes=43
        Input split bytes=111
        Combine input records=4
        Combine output records=3
        Reduce input groups=3
        Reduce shuffle bytes=43
        Reduce input records=3
        Reduce output records=3
        Spilled Records=6
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=16
        Total committed heap usage (bytes)=713555968
    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=25
    File Output Format Counters
        Bytes Written=25
hadoop@lab1-intro-team-14:~/hadoop$ hdfs dfs -cat /output/part-r-00000
hadoop 1
hello 2
world 1
hadoop@lab1-intro-team-14:~/hadoop$ |

```

5. Print the output result **hdfs dfs -cat /output/part-r-00000**

```

hadoop@lab1-intro-team-14:~/hadoop$ hdfs dfs -cat /output/part-r-00000
hadoop 1
hello 2
world 1
hadoop@lab1-intro-team-14:~/hadoop$ |

```

- As we can see, we have gotten the correct output. We then stopped the services and shelved the instance.

```
451 vi hadoop-env.sh
452 stop-all.sh
453 start-all.sh
454
```