Exp 9 Installation of Hadoop and execution of a MapReduce program
13/09/24

Aim :

To install Hadoop and execute a mapreduce program using it.

Description :

Open Oracle VM and start the ubuntu os.
Login and Go to terminal.
Update the OS.

```
tce@tce-VirtualBox:~$ sudo apt update
[sudo] password for tce:
Get:1 https://dl.google.com/linux/chrome/deb stable InRelease [1,825 B]
Hit:2 http://security.ubuntu.com/ubuntu bionic-security InRelease
Err:1 https://dl.google.com/linux/chrome/deb stable InRelease
  The following signatures couldn't be verified because the public key is not a
vailable: NO_PUBKEY E88979FB9B30ACF2
Hit:3 http://in.archive.ubuntu.com/ubuntu bionic InRelease
```

Install ssh

```
tce@tce-VirtualBox:~$ sudo apt install ssh
Reading package lists... Done
Building dependency tree
Reading state information... Done
ssh is already the newest version (1:7.6p1-4ubuntu0.7).
0 upgraded, 0 newly installed, 0 to remove and 272 not upgraded.
```

Install Hadoop

```
tce@tce-VirtualBox:~$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/
hadoop-3.3.6.tar.gz
--2024-09-29 22:36:07--  https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/ha
doop-3.3.6.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connect
ed.
HTTP request sent, awaiting response... 200 OK
Length: 730107476 (696M) [application/x-gzip]
Saving to: ‘hadoop-3.3.6.tar.gz.3’

hadoop-3.3.6.tar.gz 100%[===================>] 696.28M  1.03MB/s    in 14m 47s

2024-09-29 22:50:55 (803 KB/s) - ‘hadoop-3.3.6.tar.gz.3’ saved [730107476/73010
7476]
```

Unzip the downloaded zip file : tar -xzf Hadoop-3.3.6.tar.gz

Check if it is unzipped using ls

```
tce@tce-VirtualBox:~$ tar -xzf hadoop-3.3.6.tar.gz.3
tce@tce-VirtualBox:~$ ls
archiv        hadoop-3.3.6.tar.gz    hello.out  Pictures    Videos
Desktop       hadoop-3.3.6.tar.gz.1  loc        Public
Documents     hadoop-3.3.6.tar.gz.2  mpi        square.cpp
Downloads     hadoop-3.3.6.tar.gz.3  Music      square.out
hadoop-3.3.6  hello_omp.cpp          openmp     Templates
tce@tce-VirtualBox:~$
```

Rename the hadoop-3.3.6 to Hadoop: mv hadoop-3.3.6 hadoop

```
tce@tce-VirtualBox:~$ mv hadoop-3.3.6 hadoop
```

Install jdk version 11.

```
tce@tce-VirtualBox:~$ sudo apt install openjdk-11-jdk
[sudo] password for tce:
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  openjdk-11-jdk-headless openjdk-11-jre openjdk-11-jre-headless
Suggested packages:
  openjdk-11-demo openjdk-11-source visualvm fonts-ipafont-gothic
  fonts-ipafont-mincho fonts-wqy-microhei | fonts-wqy-zenhei
The following packages will be upgraded:
```

Check java version.

```
tce@tce-VirtualBox:~$ java -version
openjdk version "11.0.19" 2023-04-18
OpenJDK Runtime Environment (build 11.0.19+7-post-Ubuntu-0ubuntu118.04.1)
OpenJDK 64-Bit Server VM (build 11.0.19+7-post-Ubuntu-0ubuntu118.04.1, mixed mo
de, sharing)
```

Create /hadoop/data directory and create datanode and namenode directories on that directory.

```
tce@tce-VirtualBox:~$ cd hadoop
tce@tce-VirtualBox:~/hadoop$ ls
bin      lib             licenses-binary  NOTICE.txt  share
etc      libexec         LICENSE.txt      README.txt
include  LICENSE-binary  NOTICE-binary    sbin
tce@tce-VirtualBox:~/hadoop$ mkdir data
tce@tce-VirtualBox:~/hadoop$ cd data
tce@tce-VirtualBox:~/hadoop/data$ mkdir -p {datanode,namenode}
tce@tce-VirtualBox:~/hadoop/data$ ls
datanode  namenode
tce@tce-VirtualBox:~/hadoop/data$
```

Find java and Hadoop path.

```
tce@tce-VirtualBox:~/hadoop/data$ dirname  $(dirname $(readlink -f $(which java
)))
/usr/lib/jvm/java-11-openjdk-amd64
```

```
tce@tce-VirtualBox:~/hadoop$ pwd
/home/tce/hadoop
tce@tce-VirtualBox:~/hadoop$
```

It will be used to edit bashrc file.

Open bashrc file in text editor.

```
tce@tce-VirtualBox:~/hadoop/lib/native$ gedit ~/.bashrc
```

Include below contents to bashrc file and save file.

```
Open ▾   .bashrc  ~/   Save
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HADOOP_HOME=/home/tce/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=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
```

Apply the changes

```
tce@tce-VirtualBox:~/hadoop/lib/native$ source ~/.bashrc
```

Change the directory to /hadoop/etc/hadoop and list the files present in that directory.

```
tce@tce-VirtualBox:~/hadoop/lib/native$ cd ..
tce@tce-VirtualBox:~/hadoop/lib$ cd ..
tce@tce-VirtualBox:~/hadoop$ cd etc
tce@tce-VirtualBox:~/hadoop/etc$ cd hadoop
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ ls
capacity-scheduler.xml            kms-log4j.properties
configuration.xsl                 kms-site.xml
container-executor.cfg            log4j.properties
core-site.xml                     mapred-env.cmd
hadoop-env.cmd                    mapred-env.sh
hadoop-env.sh                     mapred-queues.xml.template
hadoop-metrics2.properties        mapred-site.xml
hadoop-policy.xml                 shellprofile.d
hadoop-user-functions.sh.example  ssl-client.xml.example
hdfs-rbf-site.xml                 ssl-server.xml.example
hdfs-site.xml                     user_ec_policies.xml.template
httpfs-env.sh                     workers
httpfs-log4j.properties           yarn-env.cmd
httpfs-site.xml                   yarn-env.sh
kms-acls.xml                      yarnservice-log4j.properties
kms-env.sh                        yarn-site.xml
```

Open hadoop-env.sh with gedit hadoop-env.sh. Go to Java Implementation part and paste JAVA_HOME.

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ gedit hadoop-env.sh
```

```
Open ▾   hadoop-env.sh
         ~/hadoop/etc/hadoop   Save
# such as in /etc/profile.d

# The java implementation to use. By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64

# Location of Hadoop.  By default, Hadoop will attempt to determine
# this location based upon its execution path.
# export HADOOP_HOME=
```

Configure below 4 files present in /hadoop/etc/hadoop directory.

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ gedit core-site.xml
```

```
<configuration>
<property>
  <name>fs.defaultFS</name>
  <value>hdfs://localhost:9000</value>
</property>
</configuration>
```

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ gedit hdfs-site.xml
```

```
<configuration>
<property>
  <name>dfs.replication</name>
  <value>3</value>
</property><property>
  <name>dfs.name.dir</name>
  <value>/home/tce/hadoop/data/namenode</value>
</property><property>
  <name>dfs.data.dir</name>
  <value>/home/tce/hadoop/data/datanode</value>
</property>
</configuration>
```

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ gedit yarn-site.xml
```

```
<property>
  <name>yarn.nodemanager.aux-services</name>
  <value>mapreduce_shuffle</value>
</property><property>
  <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
  <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
</configuration>
```

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ gedit mapred-site.xml
```

```
<configuration>
<property>
  <name>mapreduce.framework.name</name>
  <value>yarn</value>
</property>
</configuration>
```

Generate key and make changes to key location.

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/tce/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/tce/.ssh/id_rsa.
Your public key has been saved in /home/tce/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:urLihu+Gyok2WSMREvANnG1kV1uAhEEr/qLwaTmQ1qs tce@tce-VirtualBox
The key's randomart image is:
+---[RSA 2048]----+
|+o.==+ooo..      |
|.oo++o.  o       |
|. +.o   .        |
| o .             |
| .+     S        |
|oo =   .         |
|o+=.+ .          |
|**O+.  .         |
|*EOo.o.          |
+----[SHA256]-----+
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authori
zed_keys
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ chmod 640 ~/.ssh/authorized_keys
```

Format the name node.

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ hdfs namenode -format
2024-09-30 02:21:29,309 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = tce-VirtualBox/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.3.6
STARTUP_MSG:   classpath = /home/tce/hadoop/etc/hadoop:/home/tce/hadoop/share/h
adoop/common/lib/kerby-asn1-1.0.1.jar:/home/tce/hadoop/share/hadoop/common/lib/
guava-27.0-jre.jar:/home/tce/hadoop/share/hadoop/common/lib/hadoop-annotations-
3.3.6.jar:/home/tce/hadoop/share/hadoop/common/lib/commons-compress-1.21.jar:/h
ome/tce/hadoop/share/hadoop/common/lib/jetty-http-9.4.51.v20230217.jar:/home/tc
e/hadoop/share/hadoop/common/lib/netty-codec-stomp-4.1.89.Final.jar:/home/tce/h
adoop/share/hadoop/common/lib/commons-codec-1.15.jar:/home/tce/hadoop/share/had
oop/common/lib/kerb-simplekdc-1.0.1.jar:/home/tce/hadoop/share/hadoop/common/li
b/commons-text-1.10.0.jar:/home/tce/hadoop/share/hadoop/common/lib/gson-2.9.0.j
ar:/home/tce/hadoop/share/hadoop/common/lib/netty-codec-mqtt-4.1.89.Final.jar:/
home/tce/hadoop/share/hadoop/common/lib/jetty-util-9.4.51.v20230217.jar:/home/t
ce/hadoop/share/hadoop/common/lib/netty-resolver-dns-native-macos-4.1.89.Final-
osx-x86_64.jar:/home/tce/hadoop/share/hadoop/common/lib/netty-resolver-dns-4.1.
89.Final.jar:/home/tce/hadoop/share/hadoop/common/lib/kerb-util-1.0.1.jar:/home
/tce/hadoop/share/hadoop/common/lib/jackson-core-2.12.7.jar:/home/tce/hadoop/sh
are/hadoop/common/lib/jetty-server-9.4.51.v20230217.jar:/home/tce/hadoop/share/
hadoop/common/lib/netty-transport-4.1.89.Final.jar:/home/tce/hadoop/share/hadoo
```

Start the namenode using the command start-dfs.sh

If there is an error, try installing ssh with sudo apt install ssh

Restart.

Then type start-dfs.sh

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [tce-VirtualBox]
```

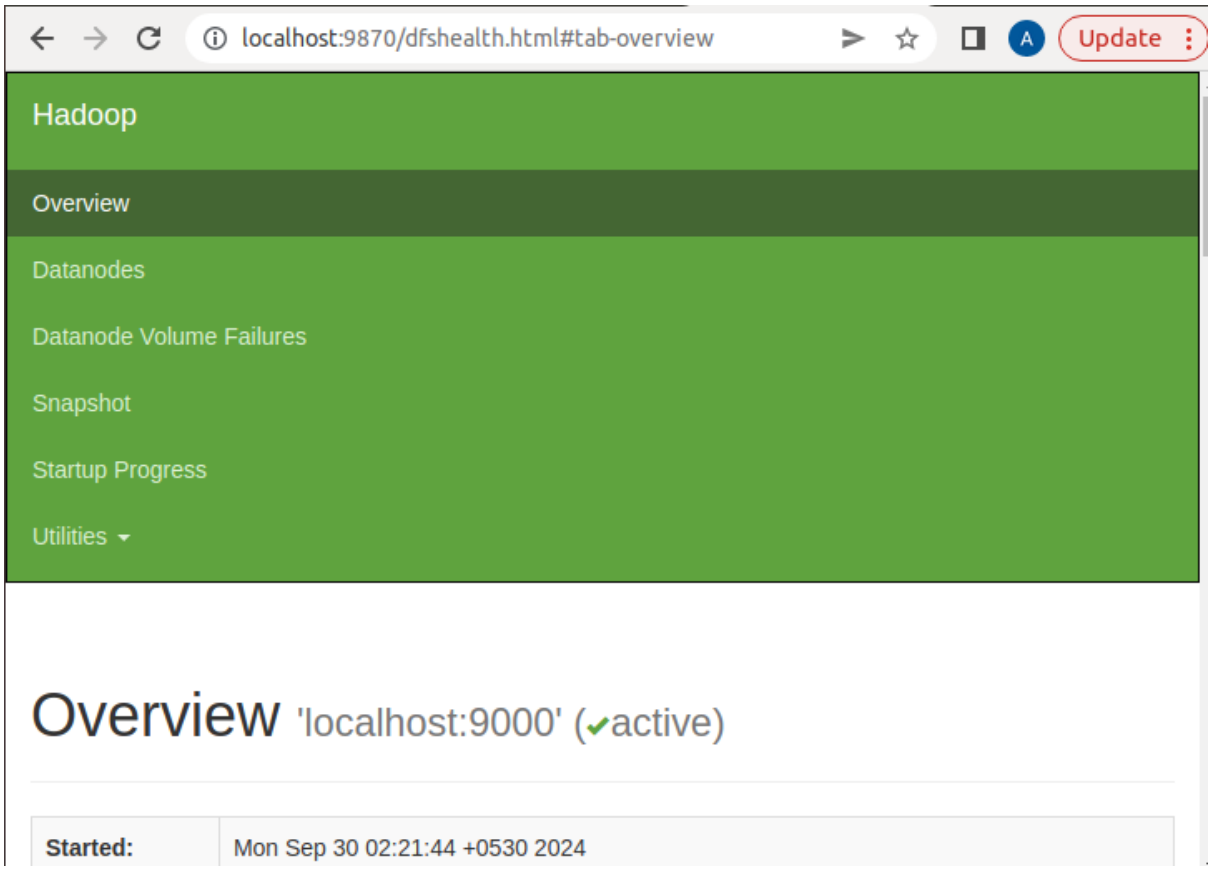
List the java processes running on the machine

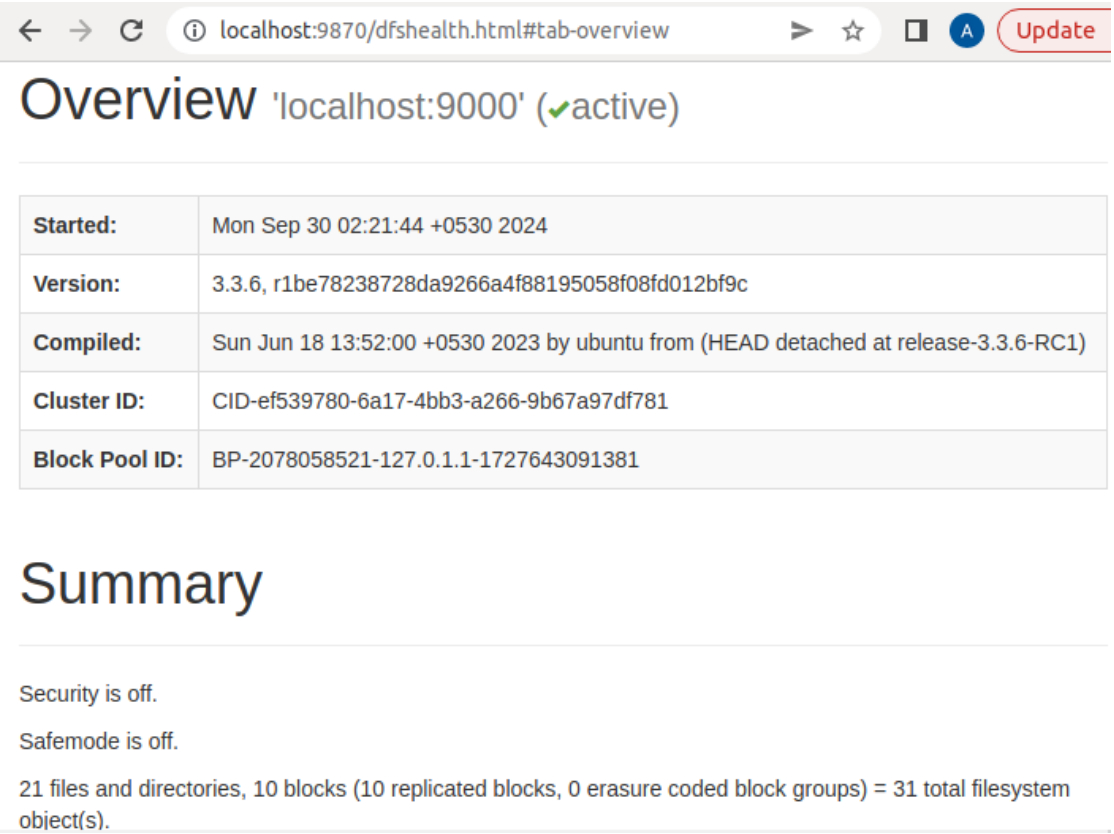
```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ jps
705 Jps
516 SecondaryNameNode
32725 DataNode
32571 NameNode
```

Start-yarn.sh and list the java processes running on the machine.

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ jps
1696 Jps
516 SecondaryNameNode
837 ResourceManager
32725 DataNode
32571 NameNode
1019 NodeManager
```

Open the browser and enter http://localhost:9870/





Execution of mapreduce program

Make input directory.

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ hdfs dfs -mkdir -p /user/hadoop/input
```

Put the text files to the input directory whose words will be counted.

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ hdfs dfs -put /home/tce/football.txt /u
ser/hadoop/input
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ hdfs dfs -put /home/tce/cricket.txt /us
er/hadoop/input
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ hdfs dfs -ls /user/hadoop/input
Found 2 items
-rw-r--r--   3 tce supergroup       3206 2024-09-30 02:38 /user/hadoop/input/cr
icket.txt
-rw-r--r--   3 tce supergroup       3208 2024-09-30 02:38 /user/hadoop/input/fo
otball.txt
```

Run the wordcount command.

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ hadoop jar ~/hadoop/share/hadoop/mapred
uce/hadoop-mapreduce-examples-*.jar wordcount /user/hadoop/input /user/hadoop/o
utput
2024-09-30 02:51:24,239 INFO client.DefaultNoHARMFailoverProxyProvider: Connect
ing to ResourceManager at localhost/127.0.0.1:8032
2024-09-30 02:51:25,449 INFO mapreduce.JobResourceUploader: Disabling Erasure C
oding for path: /tmp/hadoop-yarn/staging/tce/.staging/job_1727644856661_0001
2024-09-30 02:51:26,958 INFO input.FileInputFormat: Total input files to proces
s : 2
2024-09-30 02:51:27,552 INFO mapreduce.JobSubmitter: number of splits:2
2024-09-30 02:51:28,667 INFO mapreduce.JobSubmitter: Submitting tokens for job:
 job_1727644856661_0001
2024-09-30 02:51:28,668 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-09-30 02:51:29,133 INFO conf.Configuration: resource-types.xml not found
2024-09-30 02:51:29,134 INFO resource.ResourceUtils: Unable to find 'resource-t
ypes.xml'.
2024-09-30 02:51:29,913 INFO impl.YarnClientImpl: Submitted application applica
tion_1727644856661_0001
2024-09-30 02:51:30,085 INFO mapreduce.Job: The url to track the job: http://tc
e-VirtualBox:8088/proxy/application_1727644856661_0001/
2024-09-30 02:51:30,093 INFO mapreduce.Job: Running job: job_1727644856661_0001
2024-09-30 02:51:54,508 INFO mapreduce.Job: Job job_1727644856661_0001 running
2024-09-30 02:51:54,517 INFO mapreduce.Job:  map 0% reduce 0%
2024-09-30 02:52:30,260 INFO mapreduce.Job:  map 100% reduce 0%
2024-09-30 02:52:44,793 INFO mapreduce.Job:  map 100% reduce 100%
2024-09-30 02:52:45,814 INFO mapreduce.Job: Job job_1727644856661_0001 complete
d successfully
2024-09-30 02:52:46,026 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=8458
                FILE: Number of bytes written=846391
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=6647
                HDFS: Number of bytes written=5351
                HDFS: Number of read operations=11
                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=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=63848
                Total time spent by all reduces in occupied slots (ms)=11222
                Total time spent by all map tasks (ms)=63848
                Total time spent by all reduce tasks (ms)=11222
```

```
                Total vcore-milliseconds taken by all map tasks=63848
                Total vcore-milliseconds taken by all reduce tasks=11222
                Total megabyte-milliseconds taken by all map tasks=65380352
                Total megabyte-milliseconds taken by all reduce tasks=11491328
        Map-Reduce Framework
                Map input records=26
                Map output records=1013
                Map output bytes=10453
                Map output materialized bytes=8464
                Input split bytes=233
                Combine input records=1013
                Combine output records=627
                Reduce input groups=547
                Reduce shuffle bytes=8464
                Reduce input records=627
                Reduce output records=547
                Spilled Records=1254
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=324
                CPU time spent (ms)=4470
                Physical memory (bytes) snapshot=656539648
                Virtual memory (bytes) snapshot=8158330880
                Total committed heap usage (bytes)=347979776
                Peak Map Physical memory (bytes)=251584512
                Peak Map Virtual memory (bytes)=2719322112
                Peak Reduce Physical memory (bytes)=159768576
                Peak Reduce Virtual memory (bytes)=2720419840
        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=6414
        File Output Format Counters
                Bytes Written=5351
```

List the files in output directory.

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ hdfs dfs -ls /user/hadoop/output
Found 2 items
-rw-r--r--   3 tce supergroup          0 2024-09-30 02:52 /user/hadoop/output/_
SUCCESS
-rw-r--r--   3 tce supergroup       5351 2024-09-30 02:52 /user/hadoop/output/p
art-r-00000
```

Print the contents of part-r-00000

```
tce@tce-VirtualBox:~/hadoop/etc/hadoop$ hdfs dfs -cat /user/hadoop/output/part-
r-00000
"Hand   1
"gentleman’s    1
"soccer"        1
(EPL),  1
(Fédération     1
(IPL)   1
(ODIs)  1
(ODIs), 1
(T20)   1
(T20),  1
(between        2
(when   1
16th    1
1983    1
20      1
2019    1
22-yard-long    1
3-5-2   1
4-3-3,  1
4-4-2,  1
50      1
A       2
AB      1
Association)    1
```

It prints the count of each words on both files.

Result:

Thus ,the installation of Hadoop and execution of mapreduce program have been successfully completed.